Network Models for Materials and Biological Systems

by

Adam de Graff

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

Approved April 2011 by the
Graduate Supervisory Committee:

Michael F. Thorpe, Chair
Giovanna Ghirlanda
Dmitry Matyushov
Sefika B. Ozkan
Michael M. J. Treacy

ARIZONA STATE UNIVERSITY

August 2011

ABSTRACT

The properties of materials depend heavily on the spatial distribution and connectivity of their constituent parts. This applies equally to materials such as diamond and glasses as it does to biomolecules that are the product of billions of years of evolution. In science, insight is often gained through simple models with characteristics that are the result of the few features that have purposely been retained. Common to all research within in this thesis is the use of network-based models to describe the properties of materials.

This work begins with the description of a technique for decoupling boundary effects from intrinsic properties of nanomaterials that maps the atomic distribution of nanomaterials of diverse shape and size but common atomic geometry onto a universal curve. This is followed by an investigation of correlated density fluctuations in the large length scale limit in amorphous materials through the analysis of large continuous random network models. The difficulty of estimating this limit from finite models is overcome by the development of a technique that uses the variance in the number of atoms in finite subregions to perform the extrapolation to large length scales. The technique is applied to models of amorphous silicon and vitreous silica and compared with results from recent experiments.

The latter part this work applies network-based models to biological systems. The first application models force-induced protein unfolding as crack propagation on a constraint network consisting of interactions such as hydrogen bonds that cross-link and stabilize a folded polypeptide chain. Unfolding pathways generated by the model are compared with molecular dynamics simulation and experiment for a diverse set of proteins, demonstrating that the model is able to capture not only native state behavior but also partially unfolded intermediates far from the native state. This study concludes with the extension of the latter model in the development of an efficient algorithm for predicting protein structure through the flexible fitting of atomic models to low-resolution cryo-electron microscopy data. By optimizing the fit to synthetic data through directed sampling and context-dependent constraint removal, predictions are made with accuracies within the expected variability of the native state.

To my family for their love and constant support.

ACKNOWLEDGMENTS

I am very thankful to my advisor, Mike Thorpe, for his guidance in both in regards to materials science and protein conformational dynamics. His advice on how to give succinct talks will benefit me in my future pursuits. He has also managed to prove to one of the most adamant question-askers that there is still room for more.

I am thankful for the many interesting conversations that I have shared with Banu Ozkan on diverse topics, but in particular on protein allostery. I am grateful to Phil Williams and Gareth Shannon for their insightful collaboration. I appreciate the advice given to me by my committee that has helped improve my thesis.

Certainly some of the most positive experiences of my PhD were the countless conversations with colleagues and officemates that allowed me to understand the exciting work they were conducting. These conversations have given me a deeper appreciation for aspects of experimental and theoretical biophysics outside of my own research area.

I would also like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for their incredible support of my education abroad through PGS-M and PGS-D fellowships.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

APPENDIX

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER 1: INTRODUCTION

Common to all research within in this thesis is the use of network-based models to describe the properties of materials. To illustrate the importance of networks, imagine that you are in command of a large crowd and you ask each person to grab hold of two of their neighbors' jackets. If you then ask the whole crowd to move around as much as possible without letting go, you will find that the movement and distribution of the people in the crowd depends heavily on the properties of the network connecting them. This analogy applies equally well to the nanoscale, where properties of materials depend on the network of covalent bonds and electrostatic interactions between atoms. As with crowds, these networks generally do not have periodicity or long-ranged order, forming amorphous structures. The great importance of amorphous structures can be seen everywhere from state-of-the-art electronics (1) to the very building blocks of life itself (2, 3).

Amorphous materials are ubiquitous in our lives and widespread in industry. Every time you look out the window, you are peering through an amorphous material that allows visible light to travel through virtually unaffected. Amorphous silicon is used in photovoltaic cells (4) and is the material of choice for thin-film transistors (5) used in large-area electronics such as liquid-crystal displays. Amorphous forms of silica, other than being the major component in window glass, are used in optical fibers, providing good optical transmission while also being mechanically strong and chemically inert (6).

Our dependence on amorphous materials goes far beyond electronics and optical fibers, as our very lives depend on them. Each of the trillions of cells in our body uses millions of nanoscopic biological machines to keep themselves running (7). Most commonly they are made of proteins, amorphous structures composed of folded chains of amino acids. The manner in which these chains are folded and cross-linked determines their flexibility, mechanical resistance, and function (3, 8, 9). Efficient models focusing on networks of specific interactions can be used to characterize the ensemble of conformations that biomolecules are likely to possess (10), allowing a better understanding of their role in disease (11) and the engineering of proteins with enhanced function (12).

Materials research

One very important difference between crystalline and amorphous materials is the inability to determine the latter's atomic arrangement directly from experimental measurements such as X-ray or neutron diffraction experiments (13). This can be seen as resulting from the fact that whereas a crystal contains translational symmetry that allows it to be characterized by a finite set of numbers describing the positions of atoms within the unit cell and the geometry of the associated lattice, each atom in an amorphous material has a unique environment due to disorder (14). While not uniquely specifying the position of each atom within an amorphous material, a diffraction experiment does determine the distribution of atomic pair separations, characterized by the radial distribution function (RDF) (15). This distribution links microscopic atomic placements with macroscopic experimental observables such as pressure, compressibility, energy, and phase transitions (16). By comparing the experimentally derived RDF with that from a computational model, insight can be gained into the structural origin of such experimental observables. For example, RDFs have been used to probe the architecture of novel amorphous and porous materials (17), illustrate the phase transition across optimal doping of superconducting materials (18), and detect randomness in periodic superlattices (19).

In general, the RDF is affected by two types of structural properties. The first type relates to the intrinsic geometry of the atomic network, i.e., the average coordination of each atom, the distortion of bond lengths and angles, and the randomness of the atomic network. These properties determine the positions, intensities, widths, and overlaps of the peaks in the RDF. The second type relates to spatial confinement, i.e., the shape and size of the material sample. For bulk materials, the RDF is determined solely by the first effect reflecting the intrinsic properties of the bond network, but for nanomaterials the RDF is affected by both types of structural properties. Determination of the shape and size of a nanomaterial is not usually the goal of RDF analysis, as they can be obtained by other experimental techniques such as small-angle X-ray scattering (20) and transmission electron microscopy (21). The main research interest in this study is the determination of the RDF characteristic of the intrinsic atomic geometry of a nanomaterial and its deviations from the equivalent bulk material. The first part of this thesis (Chapter 5) describes a method for modifying the form of the RDF that decouples shape and size effects from intrinsic effects so that nanomaterials of any shape and size sharing a common atomic geometry fall onto a universal curve. This allows more subtle differences in the atomic geometry of nanomaterials due to effects such as surface relaxation to be directly compared.

After more than half a century of theoretical effort, determination of the atomic structure of amorphous materials from experimental observables such as the RDF remains one of the outstanding challenges of our day (22, 23). The difficulty of this problem is best exemplified by the extensive studies on vitreous silica (24). Significant progress has been made through the construction of physical and computational models, from which RDFs can be calculated and compared with experimental data to gain insight into features of the model that are likely correct and those associated with remaining discrepancies. Many of these models stem from Zachariasen’s famous proposal that the structure of glasses form a continuous random network (24), a view supported by the experiments such as those of Warren (25-27). Modern computers and efficient algorithms (28) have permitted the construction of very large computer models (29, 30) that are commonly validated against experiment through comparison of the position, shape, and area of peaks in either real space (RDF) or reciprocal space (structure factor, S(Q)), and as a result have focused on short and intermediate length scales. In contrast, properties on the largest length scales in the form of long-wavelength density fluctuations, described by the limiting behavior of the structure factor $S(Q \to 0)$, are rarely discussed in the context of amorphous modeling but are of considerable interest (31). The limiting value can be estimated from small angle elastic scattering experiments using either X-rays or neutrons (23). For a liquid in thermal equilibrium, $S(Q \to 0)$ is a linear function of the liquid's density, isothermal compressibility, and temperature. Upon cooling, the structural disorder of the liquid is frozen in at the glass transition, and therefore $S(Q \to 0)$ contains information about how far the system is from thermal equilibrium. Additionally, Florescu and coworkers (31) recently conjectured that a tetrahedrally coordinated continuous random network material with $S(Q \to 0) = 0$ have substantially larger photonic band gaps than those that do not, suggesting the commercial importance of such large length scale properties.

While providing important insight into the nature of amorphous materials, accurate determination of $S(Q \to 0)$ from finite models poses greater difficulty than the determination of more local properties, even from large models. The second part of this thesis (Chapter 6) describes a method that overcomes this difficulty by permitting accurate extrapolation to the limit $S(Q \to 0)$ using a general geometric principle true for any distribution of atoms, independent of thermal equilibrium (32). By calculating the variance in the number of atoms within finite regions as a function of the regions' volume, the method can be used not only to extrapolate to large length scales, but also as a metric for determining if a model is sufficiently large to make such an extrapolation accurate (32). The technique is applied to large models of both amorphous silicon (33,

34) and vitreous silica (30) and compared to recent experiments (35). Interesting implications of the results are discussed (32).

Biological research

Much like glasses, proteins are compact, cross-linked polymers stabilized by covalent bonds and weaker non-covalent interactions (10, 36). Similar in philosophy to the former network models of glasses, a simplified picture of folded proteins in terms of a discrete network of interactions can be used to gain insight into the sources of their underlying properties. One such property is the manner in which proteins respond under an applied force. This is of direct biological significance, as the physiological role of many proteins requires them to resist mechanical unfolding (8, 37-39). A complete understanding of the mechanical, regulatory, and signaling properties of many proteins depends not only on their native state conformations, but also on the nature of the intermediate states that become populated when subjected to an applied load. Unfolding behavior is studied experimentally using atomic force microscopy (39) and optical tweezers (40), but neither give direct atomistic descriptions of the unfolding pathways (41). Instead, they are sensitive to properties of the transition state and can identify the extension of partially unfolded intermediates (42). Computer-generated pathways from methods such as molecular dynamics simulation can then be compared to these observations to gain a better understanding of the atomistic identities of the transition state and intermediates (39).

While it is possible to study the unfolding behavior of proteins using detailed all-atom force fields, there is great interest in understanding simple yet general geometric principles underlying the mechanical anisotropy of protein stability. Recent work has come in the form of simplified coarse-grained dynamic models such as Gö-like models where each residue is represented by a bead (43, 44). More geometry-oriented approaches have been taken using elastic network models (45-48) in which the protein is modeled as a set of coarse-grained beads connected by springs. One such study correlated stability with the effective force constant along the pulling direction (49), while another used the equilibrium force distributions to determine mean fracture forces under the assumption that collective unfolding occurs upon fracture of the very first bond (50). While being very insightful models, their coarse-grained nature means that they lack the specific interactions such as hydrogen bonds that are largely responsible for a protein's mechanical stability (51, 52). The potential of Gö-like models are also intrinsically biased towards the native state (44, 53), making important conformational transitions requiring non-native interactions difficult or even impossible to model (54), whereas elastic network models represent an even more extreme case (49, 50), as they are unable to explore beyond the native basin.

To better understand the influence of the network of specific interactions on the mechanical anisotropy of protein stability and its role in determining unfolding pathways and the presence of metastable intermediates, I created a simple geometric model of protein unfolding that draws analogy to crack propagation in a solid. The algorithm builds on the all-atom constraint-based model developed in the Thorpe group, called FRODAN (10, 55). Within the protein unfolding model, non-covalent interactions such as hydrogen bonds, salt-bridges, and hydrophobic contacts are modeled as harmonic inequality constraints capable of supporting a finite load before breaking. Upon applying an external force and minimizing the constraint energy, an equilibrium strain distribution is produced that reflects the anisotropies of the underlying network of interactions. Complete unfolding pathways are generated by minimally overloading the network in an iterative fashion. By comparing the results for 12 proteins of diverse topology to both molecular dynamics simulation and experiment, it is demonstrated that for the majority of proteins studied (9/12), the simple model of protein unfolding as crack propagation on a constraint network is sufficient to capture both native state behavior as well as partially unfolded intermediates far from the native state.

Determination of a protein’s structure is essential in order to fully understand its functional properties, as structure determines the conformational ensemble necessary for such actions as ligand binding, signaling, or catalysis (3). While there exists several techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, that are able to determine the structure of proteins to atomic resolution, each have limitations in the systems that they can study. NMR spectroscopy is typically limited to small proteins due to the increased crowding of NMR spectra with peaks from an increasing number of atoms, whereas X-ray crystallography requires the formation of large protein crystals, which are difficult if not impossible to make for many proteins. Membrane proteins, highly flexible proteins, and loosely bound protein complexes are notoriously difficult to crystallize. In contrast, cryo-electron microscopy (cryo-EM) (56, 57) allows proteins to be imaged individually in conditions closer to their native environment by rapidly freezing them in thin aqueous samples. The process occurs so rapidly that the protein structure is not significantly perturbed by the freezing process. The higher scattering cross-section of electrons compared to X-rays allows single proteins to produce a sufficient amount of scattering to classify the resulting image. The ability to image a single layer of biomolecules is particularly advantageous with membrane proteins, for which the membrane serves as a natural template for one-dimensional crystals. Unfortunately, cryo-EM suffers from the major drawback that atomic resolutions cannot yet be reached due to effects such as radiation damage and sample charging (56, 58, 59).

While cryo-EM data typically possess resolutions of around 10 Å, when combined with the known structural constraints associated with the stereochemistry of a polypeptide chain, the cryo-EM data provides sufficient information to predict the underlying protein structure to near-atomic resolution (60). Current fitting techniques range from all-atom MD simulation to normal mode fitting (61, 62). While the computational requirements of MD techniques may be manageable for a single fit of a large complex to experimental data, the energy landscape must be heavily biased in order to complete the fit in a short amount of time and errors in the folded topology of the native state are likely to be retained. This is problematic, as starting structures are not generally in the same conformation as the one imaged (63) and are often the result of homology modeling for which hundreds of proposed models can be created with limited a priori knowledge of which starting model will result in the best fit to the experimental data (64). Coarse-grained normal mode techniques offer a means of rapidly fitting many models to the target data, but are limited to rather trivial conformational changes associated with normal modes. While the problem of structure determination from cryo-EM maps is of great importance, there is yet to be a technique that serves as a gold standard (65).

The constraint-based algorithm FRODAN used in the unfolding study possesses many properties that would be desired in an ideal fitting algorithm. Its constraints are sufficient to enforce high stereochemical quality comparable to that of an all-atom MD force field, whereas its efficient conformational sampling is similar to that of efficient coarse-grained methods, while allowing extensive conformational sampling for which normal mode techniques are incapable. The final work of my thesis consists of the development of a constraint-based fitting algorithm using a dynamic context-dependent constraint breaking criteria for the prediction of atomic protein conformations from cryo-EM data. By biasing the relatively flat FRODAN energy landscape towards conformations having high correlation with the target data, the algorithm iterates between phases of rapid conformational exploration and phases where constraints are sparsely removed based on equilibrium strain distributions. The automated method is tested on a set of seven proteins, each possessing synthetic cryo-EM target data with 10 Å resolution. In every case, after only two hours of computer time the fitting algorithm converges to a solution is closer to the known solution than the prediction made by the group that produced the benchmark set (64).

CHAPTER 2: REVIEW OF EXPERIMENTAL IMAGING TECHNIQUES

X-ray diffraction

Overview

X-ray diffraction is an imagining technique whereby a sample is imaged by recording the X-rays that are elastically scattered by the electron density of the sample as a function of the scattering angle. For a monoatomic sample containing atoms at positions $\boldsymbol{r}_i$, the scattered amplitude $\psi(\boldsymbol{Q})$ is given by

$$\psi(\boldsymbol{Q}) = \sum_i \exp(i\boldsymbol{Q}\cdot\boldsymbol{r}_i) \tag{2.1}$$

where $\boldsymbol{Q} = \boldsymbol{k}_i - \boldsymbol{k}_f$ is the difference between the incident and scattered wavevectors and the sum runs over all N atoms in the sample. Elastic scattering requires that $|\boldsymbol{k}_i| = |\boldsymbol{k}_f|$, allowing $|\mathbf{Q}|$ to be expressed in terms of $|\boldsymbol{k}_i|$ and the deflection angle 2θ according to

$$Q = 2k\sin\theta \tag{2.2}$$

The scattering amplitude is never measured directly, as the measured quantity is the intensity $I(\boldsymbol{Q})$, which is related to $\psi(\boldsymbol{Q})$ by

$$I(\boldsymbol{Q}) = \frac{f^2}{N}\psi^*(\boldsymbol{Q})\psi(\boldsymbol{Q}) \tag{2.3}$$

where f is the scattering factors of each of the N atoms. The most common quantity used to describe scattering data is a scaled version of the intensity called the structure factor $S(\boldsymbol{Q})$ which takes the form

$$S(\boldsymbol{Q}) = \frac{I(\boldsymbol{Q})}{f^2} \tag{2.4}$$

Samples can be classified as either resulting in an isotropic or anisotropic structure factor. Isotropic scattering can result from either a sample which is inherently isotropic, such as a bulk amorphous material for which all directions are equivalent, or one that contains many small anisotropic domains arranged in random orientations relative to one another, as in the case of powder diffraction (15). Spatial isotropy leads to isotropy in reciprocal space, such that in *S(**Q**)* = *S(Q)*. The spherically averaged structure factor can be related to a quantity characterizing the distribution of atomic pair separations called the reduced pair distribution function *G(r)* (15) by the equation

$$G(r) = \frac{2}{\pi}\int_0^\infty Q[S(Q)-1]\sin(Qr)\,dQ \qquad (2.5)$$

For crystalline samples, isotropy does not exist in either real space or reciprocal space. The atomic positions of all atoms in crystals can be described by their positions in the unit cell and by the lattice describing the set of translations of the unit cell needed to tile all of space. For a general scattering vector ***Q***, the scattering contributions of each atom, shown in Eq. (2.1), will add up in a fairly random manner, leading to a scattered amplitude that scales as $\sqrt{N}$ and therefore an intensity that scales as *N*. For very special scattering vectors, the translational symmetry of the crystal allows the phases from scattering events in different unit cells to add constructively, causing the scattered amplitude to scale as *N* and thus the intensities as N^2. For large crystals with many unit cells, the intensity of these constructive reflections, called Bragg reflections, completely dominate the scattered intensity pattern. The locations of these Bragg reflections can be understood by examining the condition necessary for constructive interference. If one imagines a crystal with a square lattice, the Bragg condition can be determined by examining parallel crystal planes, as shown in Figure 2.1. For X-rays with an incident

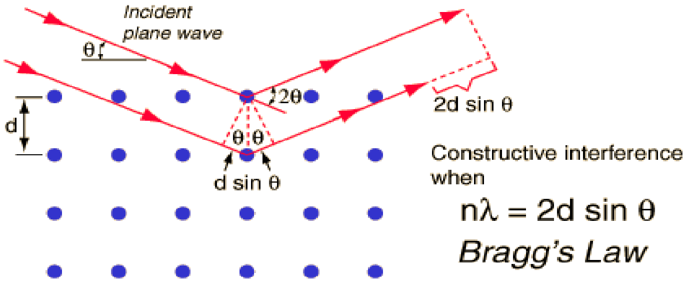


Figure 2.1 Illustration of Bragg's Law for constructive interference. Figure reproduced from (66).

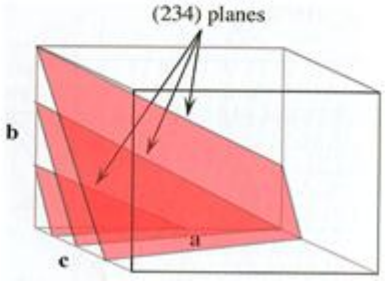


Figure 2.2 Illustration of lattice planes with Miller indices (234). Figure reproduced from (67).

angle θ relative to the planes, the path length difference between rays scattered from the top and bottom planes is $2dsin\theta$. Constructive interference requires that the path length difference be an integer number of wavelengths, described by the Bragg condition

$$n\lambda = 2d\sin\theta \tag{2.6}$$

A cubic lattice does not just contain the three crystalline planes parallel to the x, y, and z axes, but an infinite number described by the integer Miller indices *hkl*, as shown in Figure 2.2. By ensuring that the planes cut each edge of the unit cell an integer number of times, the corners of all unit cells in the lattice can be guaranteed to lie on a plane, causing the X-rays incident on each crystal cell to be in phase, and thus constructive interference to occur. By plotting all Bragg reflections in reciprocal space, one can show that they form a reciprocal lattice that can be related to the real space lattice of the crystal (67).

Application to experiment

A richer understanding of the prior mathematics can be gained by applying them to predict the results of a scattering experiment. A two-dimensional example will be used for ease of illustration. Imagine that a crystal sample is in the middle of a room in the path of an X-ray beam. For every orientation of the crystal, there is a unique orientation of the reciprocal lattice that one can imagine existing around the crystal. The two-

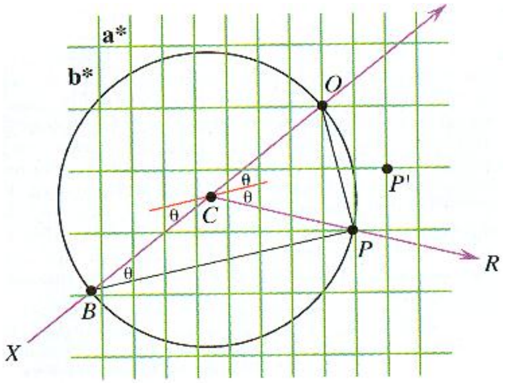


Figure 2.3 Diagram describing the necessary condition for a Bragg reflection. Figure reproduced from (67).

dimensional geometry of the scattering experiment is displayed in simplified form in Figure 2.3, where the crystal is located at point *O* in the path of the incident beam traveling from *B* to *O*. The grid represents the reciprocal lattice and the circle (called the sphere of reflection or the Ewald sphere in three dimensions) with radius $Q = |\boldsymbol{k}_i|$ represents the condition for elastic scattering. Figure 2.3 appears complicated, but it boils down to this: if the incident wavevector $\boldsymbol{k}_i$ is drawn from *C* to *O* and the scattered wavevector $\boldsymbol{k}_f$ is drawn from *C* to some point on the surface of the circle, the condition for a Bragg reflection is satisfied whenever the tip of $\boldsymbol{k}_f$ lies on a reciprocal lattice point, as it does at point *P*. This is necessarily true because the vector $\boldsymbol{Q} = \boldsymbol{k}_i - \boldsymbol{k}_f$ extends between two reciprocal lattice points *O* and *P*. High-resolution X-ray crystallography typically uses a beam with $Q \approx 1$ Å^{-1} on a crystal with a real space lattices spacing of 20 – 100 Å and therefore a reciprocal lattice spacing of $\Delta Q \approx 1/100 - 1/20$ Å. In practice, the reciprocal lattice spacing is therefore much smaller relative to the radius of the sphere of reflection than is displayed in Figure 2.3, causing many lattice points to lie on the sphere of reflection for a given crystal orientation. While a single crystal orientation results in a single two-dimensional slice through a three-dimensional intensity pattern, measurements can be performed for many crystal orientations in order to measure the intensity for all

Bragg reflections with $|\boldsymbol{Q}| < 2|\boldsymbol{k}_i|$, which serves as input for electron density reconstruction.

Electron density reconstruction

The ultimate goal of X-ray crystallography is to determine the atomic structure of the biological object being imaged. Ideally, an experiment would result in the measurement of the amplitude and phase of the diffracted waves, which could then be Fourier transformed to find the electron density of the sample and in turn be used to determine a likely atomic model, as shown in Figure 2.4. Unfortunately, detectors only measure the scattered intensity, causing all phase information to be lost.

This shortcoming can be overcome by several phase-recovering techniques. One of the most common techniques is called *isomorphous replacement* (67, 68) in which scattered intensities are collected from crystals with and without heavy atoms bound to the biomolecule. While these two scattering experiments both lack phases, the

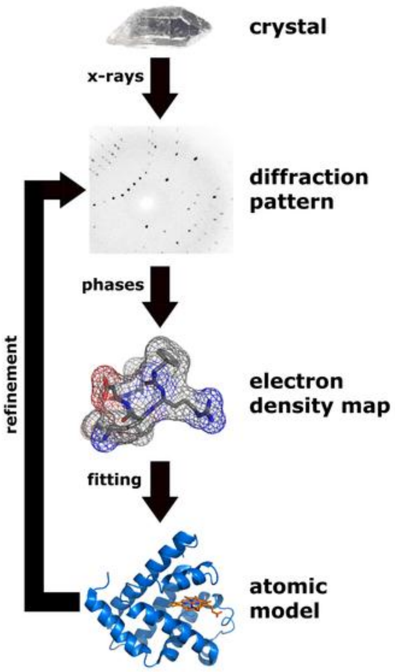


Figure 2.4 Determination of an atomic model from a protein crystal. Figure reproduced from (67).

information from the second scattering experiment with heavy atoms bound is sufficient to determine the phases for the first experiment. This amazing result can be understood by focusing on a single Bragg reflection from both experiments. Each of the two reflections is fully described by the incident wave's amplitude and phase and can therefore be represented by a vector in the complex plane confined to separate circles with radii |**PH**| and |**P**| equal to the amplitude of the scattered wave with and without the heavy atoms respectively. The two vectors are not independent, as the scattered wave of the system with heavy atoms (**PH**) is equal to the sum of the scattered wave from just the protein (**P**) and the scattered wave from just the heavy atoms (**H**), thus **PH** = **P** + **H**. This additional constraint, rearranged as **PH - P** = **H**, allows **H** to be phased using a Patterson map, as shown schematically in Figure 2.5*a-e*. This can in turn be used to constrain the vectors **P** and **PH** according to the relation **PH** = **P** + **H**, which in general contains two symmetry-related solutions for the desired phase of **P**, as shown in Figure 2.5. This ambiguity in the phase of each of the peaks in the structure factor is usually resolved by obtaining an additional constraint from a second heavy-atom experiment. A more detailed explanation can be found in (67).

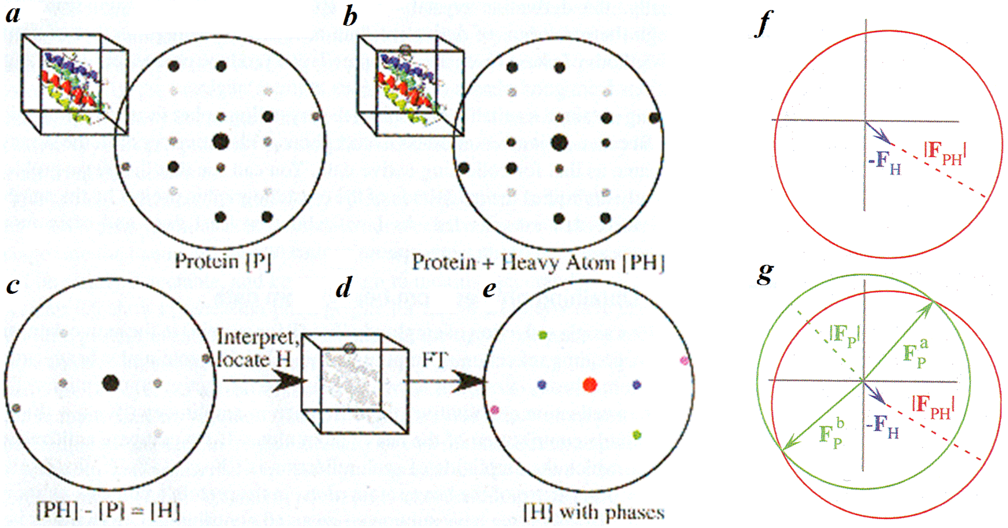


Figure 2.5 Determination of scattered wave from heavy atoms used as a constraint to determine the phases for a protein crystal. Figure reproduced from (67).

Once estimates of the phases are found, they can be combined with the amplitudes and Fourier transformed into real space. Even for perfect phases, the result is not a crisp electron density, but instead a resolution-limited distribution due to the finite number of structure factors that can be measured. The resulting distribution is similar to the correct “infinite resolution” density distribution blurred by the resolution function. As a first approximation, the resolution function is the Fourier transform of a sphere of uniform density in reciprocal space extending out to values of Q at which structure factors can be collected.

Cryo-electron microscopy

Brief history of electron microscopy

The beginnings of electron microscopy can arguably be traced back to the discovery by Julius Plücker in 1858 that cathode rays (electron beams) can be deflected with magnetic fields. This led to the realization by Eduard Rieke in 1891 that electron beams could be focused in a manner similar to a simple lens, although it was not until 1926 that it was shown theoretically by Hans Busch that under certain assumptions the lens maker’s equation could be applied to electron beams. These early efforts culminated in the construction of the first electron microscope by a German group led by Max Knoll and Ernst Ruska in 1931 (69). Interestingly, this occurred one year before the group became aware of the doctoral work of Louis de Broglie that described the wave-like nature of electrons, characterized by the de Broglie wavelength λ_e (70). The incredibly small wavelength of electrons (0.037 Å for a 100 kV accelerating voltage) was immediately recognized as offering the possibility of imaging atomic scale objects, thousands of times smaller than anything that could be imaged using visible light. By the late 1930’s, research at Siemens was already underway with the intent of imaging biological specimens (71), and by the 1950’s, work at Siemens by Ernst Ruska led to the first microscope with 100,000 times magnification, having a design similar to those in use today.

Transmission electron microscope

A modern transmission electron microscope (TEM) can be broken down functionally into five key parts: the creation and preparation of the electron beam, the sample, the objective lens, the intermediate and projector lenses, and the detector (56), as shown in Figure 2.6. Electrons are emitted by a thermally assisted field emission source and accelerated by a large electric field. Emitted electrons are pushed closer to the optical axis by a cup-shaped electrode containing a small opening, after which they travel through a set of condenser lenses that control the physical size of the beam and the beam convergence at the location of the specimen. Upon traveling through the specimen, the beam immediately travels through an objective lens. Unlike X-rays, which cannot be focused by any sort of lens, the use of magnetic fields to create an objective lens for electrons has the benefit of creating an image plane in addition to a back focal plane containing the diffraction pattern. Electron microscopy therefore has the advantage over

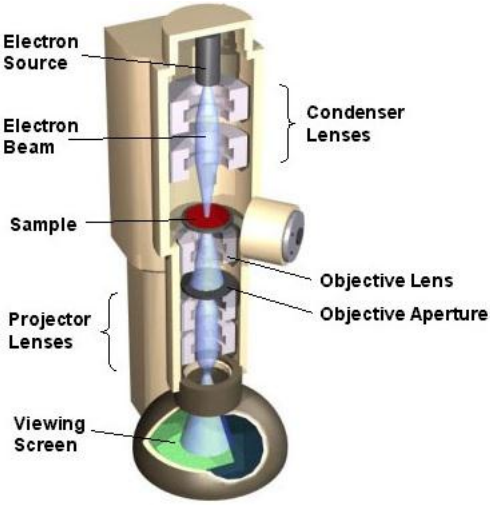


Figure 2.6 Schematic diagram of a transmission electron microscope (Web site: http://barrett-group.mcgill.ca/teaching/nanotechnology/nano02.htm). The intermediate lenses are grouped together with the projector lenses. The viewing screen is equivalent to the detector.

X-ray crystallography of not suffering from the phase problem, as phases can be extracted directly from the image plane. Next, intermediate lenses serve both to control the overall magnification of the microscope and to determine whether it is the real image or the diffraction pattern that is projected by the set of projector lenses onto the detector. Modern TEMs detect and record the intensity using either a photographic emulsion film or a charge-coupled device (CCD) camera consisting of a scintillator optically coupled to an array of photosensitive silicon diodes (56).

Theory

The traditional physics approach to electron scattering is to view a sample as a sum of individual scattering bodies, each described by a scattering amplitude $f_e(\boldsymbol{q})$. The total wave $\Psi(\boldsymbol{r})$ can be written as the sum of the incident plane wave and the scattered wave as

$$\Psi(\boldsymbol{r}) = e^{ik_o z} + f_e(\boldsymbol{q})\frac{e^{i\boldsymbol{q}\cdot\boldsymbol{r}}}{r} \qquad (2.7)$$

Insight into the form of the scattered wave can be found by beginning from the wave equation for an electron of energy *E* in the presence of a potential *V*(***r***), which takes the form

$$\nabla^2\Psi(\boldsymbol{r}) + \frac{2me}{\hbar^2}V(\boldsymbol{r})\Psi(\boldsymbol{r}) = -\frac{2me}{\hbar^2}E\Psi(\boldsymbol{r}) \qquad (2.8)$$

The general solution (57)

$$\Psi(\boldsymbol{r}) = \Psi_o(\boldsymbol{r}) + \mu\int\frac{\exp[-i\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}')]}{|\boldsymbol{r}-\boldsymbol{r}'|}V(\boldsymbol{r}')\Psi(\boldsymbol{r}')d\boldsymbol{r}' \qquad (2.9)$$

for an incident plane wave $\Psi_o(\boldsymbol{r})$ can be seen to have a form similar to that of Eq. (2.7), but unfortunately the desired solution $\Psi(\boldsymbol{r})$ appears on both the left and right hand sides. The general solution of Eq. (2.9) can be expressed as a Born series, with the first Born approximation sufficient for a weak enough potential that it can be assumed that the wave inside the sample is not significantly affected by the sample itself, allowing the substitution $\Psi(\boldsymbol{r}') = e^{ik_o z'}$ within the integral of Eq. (2.9). In addition, if the wave $\Psi(\boldsymbol{r})$ is observed a large distance R from the sample, the second term in Eq. (2.9) can be written as (57)

$$\Psi_1(\boldsymbol{R}) = \frac{\mu \exp(-i\boldsymbol{k}\cdot\boldsymbol{R})}{R}\int V(\boldsymbol{r}')\exp[i(\boldsymbol{k}-\boldsymbol{k}_o)\cdot\boldsymbol{r}']d\boldsymbol{r}' \tag{2.10}$$

By making the substitution $\boldsymbol{q} = \boldsymbol{k} - \boldsymbol{k}_o$, Eq. (2.10) can be written as

$$\Psi_1(\boldsymbol{R}) = \frac{\mu \exp(-i\boldsymbol{k}\cdot\boldsymbol{R})}{R}\sum_j f_{ej}(\boldsymbol{q}) \tag{2.11}$$

with

$$f_{ej}(\boldsymbol{q}) = \int V(\boldsymbol{r}')\exp[i\boldsymbol{q}\cdot\boldsymbol{r}']d\boldsymbol{r}' \tag{2.12}$$

called the first Born approximation for the scattering amplitude (57), equal to the Fourier transform of the screened Coulomb potential of an atom. As the potential for an atom satisfies Poisson's equation

$$\nabla^2 V_j(\boldsymbol{r}) = -\rho_j(\boldsymbol{r})/\varepsilon_o \tag{2.13}$$

the form of $f_{ej}(\boldsymbol{q}) = \mathrm{FT}\big[V_j(\boldsymbol{r})\big]$ for a given atom *j* can be found by taking the Fourier transform of Eq. (2.13) and solving for $f_{ej}(\boldsymbol{q})$, which gives

$$f_{ej}(\boldsymbol{q}) = \frac{e\big[Z - f_{Xj}(\boldsymbol{q})\big]}{\varepsilon_o q^2} \tag{2.14}$$

called the Mott formula (57), where *Z* is the atomic number of atom *j* (here representing the nuclear contribution), $f_{Xj}(\boldsymbol{q})$ the Fourier transform of the atom's electron distribution, *e* the unit charge, and ε_o the permittivity of free space. While the first Born approximation is the most popular estimate of $f_{ej}(\boldsymbol{q})$, a more detailed one called the Moliere approximation (72) leads to a complex scattering amplitude, in contrast to the purely real form of Eq. (2.12) resulting from the first Born approximation. The complex component of the Moliere approximation, which can be interpreted as contributing to absorption, is significant for heavier atoms, especially at large scattering angles.

A complimentary view to the one above begins with the realization that the de Broglie wavelength λ_e of an electron wave varies throughout a sample. Just as light undergoes a phase change when traveling through a lens relative to air due to differences in wavelength in the two materials, the de Broglie wavelength of an electron

$$\lambda_e = \frac{hc}{\sqrt{eV(2m_o c^2 + eV)}} \tag{2.15}$$

with kinetic energy eV due to the accelerating voltage of the TEM is further shortened by the generally positive screened Coulomb potential within a sample. If the screened potential is weak, the spatially varying phase shift $\Delta\varphi(x, y)$ that it induces can be expressed as

$$\Delta\varphi(x, y) = i\sigma V_s(x, y)z \tag{2.16}$$

where σ is the interaction parameter and $V_s(x, y) = \int V(x, y, z)dz$ is the potential projected along the $\hat{z}$ direction. In the weak-phase approximation (57), the transmitted wave exiting the sample can therefore be expressed as

$$\Psi_t(\boldsymbol{r}) \cong e^{ik_o z} e^{i\sigma V_s(x,y)z} \tag{2.17}$$

For a thin sample containing elements with low atomic number, $\Delta\varphi(x, y)$ is indeed small, allowing $e^{i\sigma V_s(x,y)z}$ to be written approximately as

$$e^{i\sigma V_s(x,y)z} \cong 1 + i\sigma V_s(x, y)z \tag{2.18}$$

by ignoring all terms higher than first order in $V_s(x, y)$. The transmitted wave

$$\Psi_t(\boldsymbol{r}) \cong \Psi_o(\boldsymbol{r})[1 + i\sigma V_s(x, y)] \tag{2.19}$$

is therefore the sum of an unscattered component and a term that depends linearly on the projected potential. While the phase component $i\sigma V_s(x, y)$ of the specimen transmission function is purely imaginary and to first order does not affect the magnitude of $\Psi_t(\boldsymbol{r})$, attenuation can be incorporated by the addition of an amplitude component $u(x, y)$ in the second exponent of Eq. (2.17). If the sample is a weak-amplitude object, a similar expansion to that of Eq. (2.18) can be performed, leading to

$$\Psi_t(\boldsymbol{r}) \cong \Psi_o(\boldsymbol{r})[1 + i\sigma V_s(x, y) - u(x, y)] \tag{2.20}$$

If the optics of a TEM were "perfect", the wave in the back focal plane would be the Fourier transform of the wave $\Psi_t(\boldsymbol{r})$ exiting the sample and the wave in the image plane would be a scaled version of $\Psi_t(\boldsymbol{r})$. It should be noted that it is never the wave itself that is measured experimentally, but the intensity $I(\boldsymbol{r}) = |\Psi(\boldsymbol{r})|^2$. Unfortunately the optics of real TEMs are not perfect and their effects on the wave and thus the intensity in the back focal plane and image plane are characterized by a set of functions. Among these is the effect of lens aberration and defocusing, which shifts the phase of the wave by an amount

$$\gamma(\boldsymbol{k}) = 2\pi\chi(\boldsymbol{k}) \tag{2.21}$$

where each $\boldsymbol{k} = (k_x, k_y)$ corresponds to a spatial position in the back focal plane (56). If this were the sole effect from the optics, the wave function in the back focal plane would have the form

$$\Psi_{bf}(\boldsymbol{k}) = FT\{\Psi_t(\boldsymbol{r})\}e^{i\gamma(\boldsymbol{k})} \tag{2.22}$$

Above a certain value of $\boldsymbol{k}$, $\gamma(\boldsymbol{k})$ increases rapidly, making data difficult to interpret. To remove this region from measurement, the objective lens is coupled with a finite aperture that blocks all wave vectors $\boldsymbol{k}$ having an angle greater than θ_{max} with respect to the optical axis. The effect of the aperture is modeled by a function $A(\boldsymbol{k})$ that is 1 for $\theta < \theta_{max}$ and zero otherwise. The wave function in the back field therefore becomes

$$\Psi_{bf}(\boldsymbol{k}) = FT\{\Psi_t(\boldsymbol{r})\}A(\boldsymbol{k})e^{i\gamma(\boldsymbol{k})} \tag{2.23}$$

from which the measured intensity in the image plane is

$$I_{image}(\boldsymbol{r}) = \left|\Psi_{image}(\boldsymbol{r})\right|^2 \tag{2.24}$$

where $\Psi_{image}(\boldsymbol{r}) = \mathrm{FT}^{-1}\left[\Psi_{bf}(\boldsymbol{k})\right]$.

The effects of the TEM on the transmitted wave $\Psi_t(\boldsymbol{r})$, which are conveniently expressed as products in $\boldsymbol{k}$-space, can be modeled as convolutions in $\boldsymbol{r}$-space. If we let $h(\boldsymbol{r}) = \mathrm{FT}^{-1}\left[A(\boldsymbol{k})e^{i\gamma(\boldsymbol{k})}\right]$, then $\Psi_{image}(\boldsymbol{r})$ can be written as the convolution

$$\Psi_{image}(\boldsymbol{r}) = \Psi_t(\boldsymbol{r}) \ \otimes \ h(\boldsymbol{r}) \tag{2.25}$$

The effect in real space is therefore equivalent to a convolution of the "ideal" wave by the point spread function $h(\boldsymbol{r})$. Inserting Eq. (2.25) into Eq. (2.24) and using the weak-phase approximation for $\Psi(\boldsymbol{r})$ expressed in Eq. (2.19) leads to

$$I_{image}(\boldsymbol{r}) = |[1 + i\sigma V_s(x,y)] \text{ o } h(\boldsymbol{r})|^2 \tag{2.26}$$

for which expansion to first order in $V_s(x,y)$ gives

$$I_{image}(\boldsymbol{r}) \cong 1 + 2\sigma V_s(x,y) \ \otimes \ h_{WP}(\boldsymbol{r}) \tag{2.27}$$

where $h_{WP}(\boldsymbol{r}) = \mathrm{FT}^{-1}[\sin\gamma(\boldsymbol{k})]$ is the weak-phase point spread function. Similar steps can be followed using Eq. (2.20) instead of Eq. (2.19) in order to include absorption in the weak-amplitude (small $u(x,y)$) limit, resulting in the additional term $-2u(x,y) \otimes h_{WA}(\boldsymbol{r})$ in Eq. (2.27), where $h_{WA}(\boldsymbol{r}) = \mathrm{FT}^{-1}[\cos\gamma(\boldsymbol{k})]$ is the weak-amplitude equivalent of $h_{WP}(\boldsymbol{r})$. In total, the image intensity can be conveniently written in reciprocal space as

$$I(\boldsymbol{k}) = \delta(k_z - k_o) + \Phi_r(\boldsymbol{k})A(\boldsymbol{k})\sin\gamma(\boldsymbol{k}) - \Phi_i(\boldsymbol{k})A(\boldsymbol{k})\cos\gamma(\boldsymbol{k}) \tag{2.28}$$

The functions $\sin\gamma(\boldsymbol{k})$ and $\cos\gamma(\boldsymbol{k})$ are called the phase contrast and amplitude contrast transfer functions (CTFs) and are of vital important in electron microscopy because of their effect on the image at low and high $\boldsymbol{k}$. For a particular choice of defocus and objective lens aberration, the form of $\gamma(\boldsymbol{k})$, $\sin\gamma(\boldsymbol{k})$, and $\cos\gamma(\boldsymbol{k})$ are displayed in Figure 2.7. Unique to $\sin\gamma(\boldsymbol{k})$ is its reduction of the small *k* (long wavelength) components of the contrast, while both $\sin\gamma(\boldsymbol{k})$ and $\cos\gamma(\boldsymbol{k})$ display rapid oscillation, or

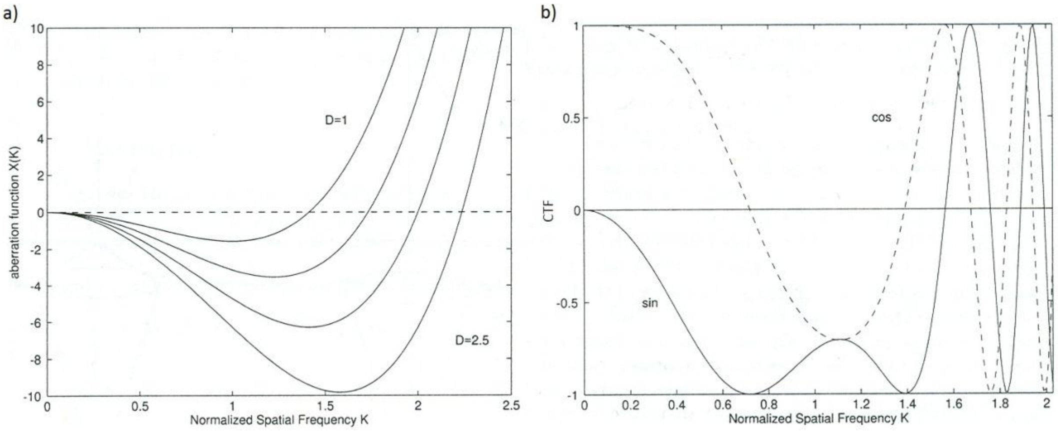


Figure 2.7 a) Form of the aberration function $\chi(\boldsymbol{k})$ for several values of the normalized defocus *D* and b) the associated phase (solid) and amplitude (dashed) contrast transfer functions for *D* = 1. Figure reproduced from (72).

contrast flipping, at large *k* (short wavelength). Finite variation in the energy of incident electrons has the effect of averaging the CTF over local regions of $\boldsymbol{k}$, causing the amplitude of the rapid oscillations at large *k* to be heavily attenuated. The phase CTF $\sin\gamma(\boldsymbol{k})$ thus acts as a band-pass filter, passing an intermediate range of frequencies with a flipped (negative) contrast. An example of the effects of such a filter is shown in Figure 2.8. For images of thin biological samples containing much less contrast than that of the frog in Figure 2.8, the contrast-reducing effects of $\sin\gamma(\boldsymbol{k})$ can make it extremely difficult to locate individual biomolecules in experimental data. This difficulty has had a large influence on the techniques used to prepare biological samples for imaging.

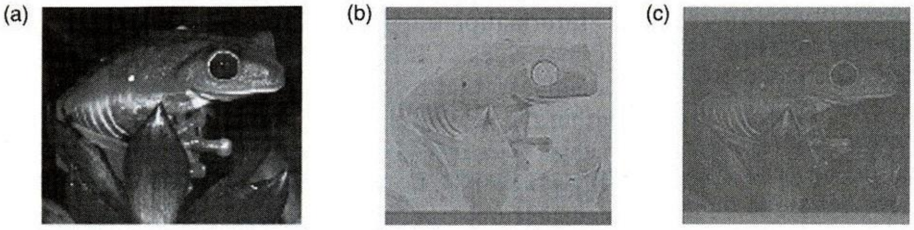


Figure 2.8 Image of a frog (a) before and (b) after being subjected to the contrast-reducing effects of the phase CTF $\sin\gamma(\boldsymbol{k})$, with $\gamma(\boldsymbol{k})$ possessing a form similar to those in Figure 2.7. Image (c) is identical to (b) upon flipping the contrast so that the reduced contrast can be directly compared with (a). Figure reproduced from (56).

Sample preparation

The properties of samples used to image biomolecules can be classified into two groups: those in which the biomolecules are organized in regular crystalline arrangements and those lacking such spatial order. The high atomic scattering cross section of electrons compared to X-rays allows sufficient contrast to be obtained from crystalline arrays that are a single layer thick in one of the dimensions. This makes TEM a powerful tool for imaging membrane proteins for which the plane of the membrane serves as a natural basis for the two-dimensional crystal. For such crystalline specimens, both the back focal plane and the image plane can be used to extract complimentary information, namely the amplitudes of the structure factor from the Bragg peaks in the diffraction pattern and the phases from the image. Many of the early high-resolution TEM images came from such crystalline samples (73).

Samples lacking spatial order do not result in the formation Bragg peaks and imaging of such samples can be viewed as an independent measurement of many individual (single) biomolecules. Single-particle measurements have the advantage that they do not require the molecules to be arranged in a crystalline array, which can be challenging for many biomolecules, but the disadvantage that they require the identification and classification of individual biomolecules in the image. Such identification is extremely difficult due to low contrast and poor signal-to-noise ratios. The low contrast is the result of the thinness of the sample, similarity in the scattering properties of protein to the surrounding substance, and aforementioned effects of the CTF, while the poor signal-to-noise ratio is due to sample exposure limitations of no more than a few $e^-/Å^2$ due to radiation damage from free radicals that form as a result, discussed later in greater detail.

Negative staining

Limitations on the contrast can be partially overcome through the use of negative staining, introduced in 1959 by Brenner and Horne (74), in which heavy metal salts such as uranyl acetate are added to the sample solution. These salts coat the solvent-accessible boundary of the biomolecules with atoms of high atomic number that have considerably stronger Coulomb potentials and proportionally even higher absorption that their low-atom number surroundings. The staining supplies much needed contrast, but only at the surface with very little internal detail. The aqueous solution with the sample and stain is never inserted directly into the high vacuum of the electron microscope, due to the high volatility of the solution. Instead, liquid is blotted away and the sample is allowed to dry. While the stain does provide a certain degree of protection, the removal of the aqueous environment during drying can cause significant distortion of the biomolecules, the degree of which depends on structural features such as the existence of internal cavities. Several alternatives to negative staining have been developed, such as glucose embedding. Glucose embedding was introduced by Unwin and Henderson in 1975 (75) as a means of replacing the aqueous medium with one having similar properties except with the additional benefit of being non-volatile. Unfortunately, the lack of heavy atoms in glucose causes poor contrast with biomolecules and was supplanted in the early 1980's by the much more successful technique of cryo-electron microscopy (cryo-EM).

Vitrified aqueous samples

An alternative method of maintaining the aqueous environment of biomolecules, developed by Taylor and Glaeser (73, 76, 77) and Dubochet (78), is to freeze the aqueous sample so rapidly as to avoid crystallization of the water, trapping the biomolecules in vitreous ice. The formation of vitreous ice is essential, as the volume change of water upon crystallizing damages samples. Like glucose embedding, vitrification has the advantage of not causing collapse or significant distortion of the sample, but with the added benefit of increasing the dose at which significant radiation damage occurs by a factor of two to six. An ice-embedded sample is prepared by first placing a small amount of aqueous specimen on a hydrophilic grid. The sample is then blotted to get rid of excess buffer until only a thin layer less than 1000 Å remains. The grid is immediately submerged into a cryogen such as liquid ethane cooled in a bath of liquid nitrogen, upon which it is transferred to the liquid nitrogen bath within the cryo-holder, which is in turn placed into the TEM and kept at temperatures of 100-115 K. Cryo-EM in vitreous ice is currently the most successful method of imaging biomolecules with electrons. The use of cryo-EM on samples of individual biomolecules in solution, called single-particle cryo-EM will be the focus of subsequent discussion.

Radiation damage

Presently, the resolution obtainable by cryo-EM is inferior to that of alternative techniques such as X-ray crystallography. One of the factors limiting the resolution of cryo-EM is the radiation damage incurred on a sample by the incident electron beam (56, 59). For typical energies of 100-300 keV used in modern TEMs, significant radiation damage begins to appear for exposures above 1 $e^{-}/\text{Å}^{2}$ at room temperature and roughly 2-6 $e^{-}/\text{Å}^{2}$ at the lower temperatures of liquid nitrogen (98-113 K) or liquid helium (10 K). Interactions of the electron beam with the sample produce free radicals that react with biomolecules, causing their gradual degradation (56, 59). As these effects are local, the high frequency components of the images are the first to be affected. This limitation on the incident flux in turn places a very low limit on the signal-to-noise ratio of the individual projections, making it very difficult to identify the small contrast differences between a biomolecule and its surroundings. Such identification is necessary in determining the location and orientation of each biomolecule, one of the first steps in image reconstruction.

Image reconstruction

The raw data consisting of the projections of individual biomolecules have a poor signal-to-noise ratio, low contrast, and measurement artifacts such as contrast flipping due to the oscillatory nature of the CTF, as shown in Figure 2.7. The poor signal-to-noise ratio can be improved by classifying projections based on the cross-correlation of their optimal alignment and averaging similar projections, as averaging reinforces the features due to the biomolecules while averaging out random noise. To collect the large amount of information required for high-resolution cryo-EM maps, tens of thousands of projections of individual biomolecules are needed. The effects of the CTF on the images can also be partially removed by characterizing the CTF and reversing the phase flipping and amplitude attenuation that it causes, while minimizing the amplification of noise (79). Images at several values of defocus can also be used to compensate for the effects of the zeros in the CTF.

One of the large challenges to determining a three-dimensional map is to determine the relative orientations of the averaged projections. They must be merged to form one or, in the case of samples containing multiple stable conformations, a few EM maps. It is not immediately obvious that all the information about the three-dimensional object is contained within a complete set of projections. Luckily, Radon's Theorem (80) and particularly one instance of it called the Fourier Projection Theorem (81) proves that a complete set of projections is sufficient for a complete three-dimensional reconstruction. The Fourier Projection Theorem states that for a three-dimensional distribution represented in ***r***-space by $f(x,y,z)$ and in reciprocal space by $F\left(k_x,k_y,k_z\right)$, that the act of taking a projection of $f(x,y,z)$ along a direction $\hat{\beta}$ is equivalent to retaining only the values of $F\left(k_x,k_y,k_z\right)$ on the plane normal to $\hat{\beta}$ that traverses the origin. From this theorem, the challenge of three-dimensional reconstruction can be viewed as one of filling Fourier space (out to a maximum wave vector corresponding to the limiting resolution) by a set of planes obtained from experimental projections.

A popular method for determining the relative orientations of the projections (or more directly their associated Fourier planes) is the “method of common lines” first proposed by Crowther (82). This method is based on the fact that the Fourier planes of any two projections intersect along a line, which in principle allows the determination of two of the three Euler angles relating one plane to the other by finding the pair of common lines with the greatest cross-correlation. A second common solution is the “random conical” data collection method (83) which takes advantage of the fact that many biomolecules have preferred orientations due to interactions with the surface of the sample. By tilting the sample in the TEM relative to the beam prior to measurement, a set of biomolecules in the sample sharing a common zero-tilt projection (but with random rotation angle) form a conical projection series in the tilted sample that fill Fourier space except for a conical section along the axis of the beam. The volume of this missing conical section can be minimized by choosing high tilt angles, typically 60-70°. While the method of common lines requires the biomolecule to occupy all orientations in the untilted sample, the random conical method works when either random or preferred orientations exist within the sample.

Once the orientation of a set of projections has been determined, a three-dimensional ***r***-space model can be created in a number of ways. A common method is weighted back-projection (56), whereby the projected potential is speared out uniformly over a distance *D* along the direction of the projection. If *D* is greater than the maximum diameter of the object, by adding up smears with appropriate weighting for all projections, one obtains a three-dimensional reconstruction of the biomolecule. A more intuitive method is to estimate the value of the Fourier components at a set of grid points by interpolating the data on the Fourier planes obtained from the experimental projections, upon which an inverse Fourier transform can be performed. It should be noted that the determination of projection orientations and three-dimensional reconstruction are not generally separate. Often a preliminary three-dimensional model is created in order for its simulated projections to be used to help classify the experimental data, which in turn can be used to create a better model. Such iterative refinement can be followed to convergence.

Resolution

Resolution is a vital characteristic of an imaging technique, as it determines the amount of information contained in the experimental measurement. A common method of assessing the resolution of an experimental map is to divide the measured projections into two sets of equal size and reconstruct the corresponding complex Fourier space values from each set. The similarity of the two reconstructions in Fourier space can then be determined by calculating the correlation of the two functions on shells of constant radius $|\boldsymbol{k}|$, called the Fourier shell correlation (56, 84). This correlation generally decreases with increasing wave vector, with the resolution typically defined as the reciprocal of the wave vector at which the correlation drops to 0.5, as Fourier components beyond this wave vector are dominated by noise. As discussed previously, one of the major factors limiting spatial resolution is the low exposure allowed due to radiation damage (56, 58, 59). The combination of low contrast and a poor signal-to-noise ratio increases the error in the angular assignment of each projection, which in turn leads to errors in the reconstruction of a three-dimensional model. This can be partially overcome by taking a large number of projections to improve averaging, with recent experiments using in excess of 100,000 projections of individual biomolecules for a single reconstruction. Another major factor that limits resolution is charging, whereby a positive charge is induced in the sample and carbon substrate due to the removal of electrons by the incident beam. The electric field associated with this charging affects the transmitted beam in an unpredictable and time-dependent way (59). Specimen movement and inelastic scattering can further reduce resolution (59). Lastly, as higher resolutions are reached, the heterogeneity of the sample being imaged becomes increasingly relevant. Biomolecules do not possess a single native conformation, but instead undergo conformational changes within the native ensemble. By wrongly assuming that all projections come from identical structures, a final model is obtained that resembles an average of this ensemble. Many biomolecules undergo conformational changes of several angstroms, and thus stabilization of a particular conformer through ligand binding or proper sorting of the projections are required if atomic resolutions are to be reached.

It is often the case that the atomic structure of a protein similar to the one being imaged by cryo-EM has been determined by other means such as X-ray crystallography, albeit not necessarily in the same conformation. By using such a structure or an associated homology model as a starting model (64), atomic structures can be predicted by flexibly fitting the model to the cryo-EM data. This can result in a structural prediction with much greater accuracy than one would infer from the resolution of the data itself (85). In Chapter 8, I will describe an efficient all-atom flexible fitting algorithm that I co-developed that performs this final step in the determination of atomic models from low-resolution experiments.

CHAPTER 3: REVIEW OF AMORPHOUS MATERIALS

Crystalline versus amorphous materials

The most immediately apparent difference between crystals and amorphous solids is the latter's lack of translational symmetry. In an infinite crystal, the positions of all atoms can be determined by knowing 1) the positions of a finite set of atoms in a local volume called the unit cell and 2) the lattice representing all translations of the unit cell necessary to tile all of space. A perfect infinite crystal therefore contains order extending to infinite length scales. An amorphous solid on the other hand lacks translational symmetry and long-ranged order (86). It contains short-ranged order due to chemical bonding and steric interactions, but correlations characterizing this order diminish with distance as the number of chemical bonds separating pairs of atoms increases. The difference between a crystalline solid and an amorphous glass can be seen from the two-dimensional models (24) displayed in Figure 3.1. Bond lengths and bond angles centered on the dark atoms are well-preserved, whereas bond angles centered on the light atoms are highly flexible; causing longer ranged behavior to be less predictable.

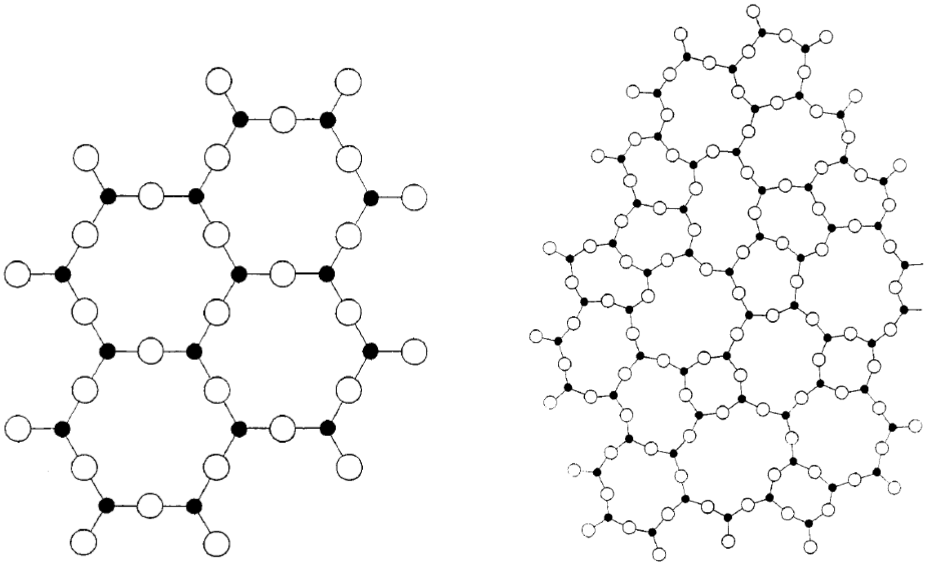


Figure 3.1 A two-dimensional crystalline solid and a corresponding continuous random network (CRN) model of a glass of composition A_2O_3. Figure reproduced from (24).

Supercooled liquids and glasses

The behavior of a liquid upon cooling below its melting temperature T_m differs greatly from substance to substance. In general, a liquid can either remain a liquid below T_m (becoming a supercooled liquid), form a crystal, or have the disorder in the liquid frozen in to form an amorphous glass (87). Below T_m there exists three competing timescales: the nucleation time, the relaxation time of the substance, and the cooling rate. Common substances like water typically freeze into a crystal upon cooling below T_m because impurities in the water help seed nucleation, causing the nucleation time to be short. If nucleation does not occur, cooling below T_m will result in a supercooled liquid, a metastable state having properties expected of a liquid, as can be seen by plotting the isobaric heat capacity C_p and entropy S as functions of temperature, shown in Figure 3.2a, and Figure 3.2b respectively. A supercooled liquid is metastable because below T_m the crystalline phase is the lowest free energy state and given enough time a nucleation event will occur that seeds the formation of a crystal. As a supercooled liquid is cooled,

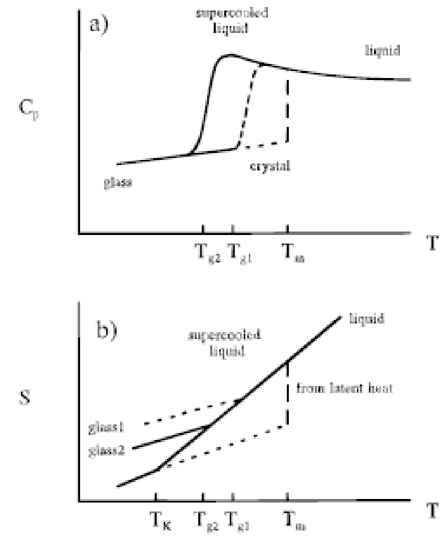


Figure 3.2 Behavior of the isobaric heat capacity and the entropy as a function of temperature, showing how the properties of the supercooled liquid behave like the liquid state before deviating near the glass transition temperature T_g. Figure reproduced from (87).

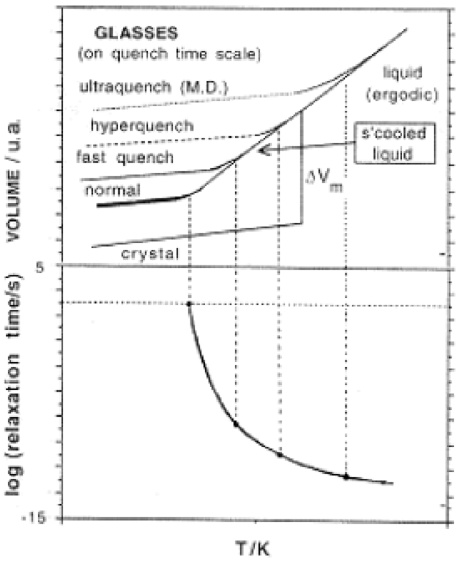


Figure 3.3 Schematic of the dependence of the glass transition temperature on the cooling rate due to increasingly long relaxation times at lower temperatures. Figure reproduced from (88).

the barriers in the energy landscape become increasingly high relative to typical thermal excitations on the order of k_BT, causing atomic rearrangement and relaxation to occur on increasingly long timescales (88), as shown in Figure 3.3.

Relaxations within the supercooled liquid can be categorized as being of two types: β-relaxations having Arrhenius relaxation time-temperature dependence and α-relaxations that depart from such a relation (89). The latter α-relaxations tend to occur on longer timescales and be of greater spatial extent than β-relaxations. The glass-forming properties of a liquid can be described in terms of the temperature dependence of their overall relaxation times, with liquids having a predominantly Arrhenius dependence, such as SiO_2, termed *strong* liquids, whereas those departing heavily from an Arrhenius dependence called *fragile* liquids (90).

For any fixed non-zero cooling rate, a temperature is eventually reached at which the relaxation times are longer than the time permitted by the cooling rate. The system will no longer be able to remain in metastable equilibrium and the bond topology of the supercooled liquid will be effectively frozen in, forming an amorphous glassy state. Thus in contrast to the melting temperature, which is a discrete point, the glass transition temperature T_g depends on the cooling rate. A very interesting consequence of this dependence is that it allows disorder in the atomic geometry of the supercooled liquid to be trapped at various temperatures and probed by X-ray and neutron scattering experiments (23). The temperature dependence of correlated density fluctuations over large length scales in vitreous silica will be discussed in Chapter 6 (32).

There are various ways of defining T_g, but a common one defines it as the point of intersection of the linear volume-temperature behavior in both the supercooled and glassy states, shown schematically in Figure 3.3. The act of freezing in the disorder is equivalent to trapping the liquid in a local minimum of the energy landscape, as displayed in Figure 3.4. Whereas the transition from a liquid to a crystal involves a first-order phase transition, the nature of the transition from a liquid to a glass is probably the "deepest and most interesting unsolved problem in solid state theory" (91). Due to the freezing in of the covalent bond network near T_g, the nature of the glass must be characterized by two distinct temperatures: the "standard" temperature describing the kinetic energy of the atoms and the *fictive* temperature T_f reflecting the strain energy of the system. It has long been observed that glasses possess structural heterogeneity in

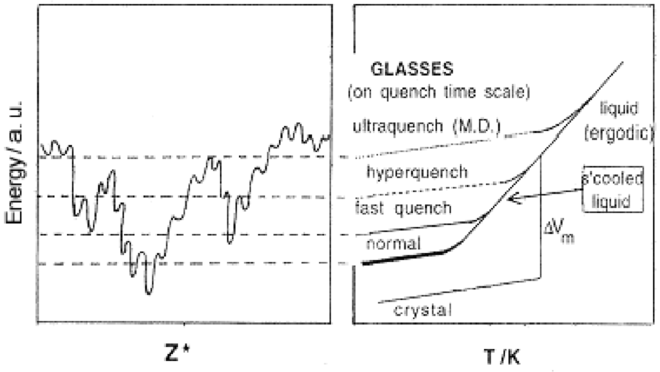


Figure 3.4 Schematic of glass formation by trapping a supercooled liquid in a metastable state. Figure reproduced from (88).

which certain spatial regions of a glass are more highly strained or “hot” than their surroundings, with such local variation having correlation lengths that define a characteristic length scale for the glass (92, 93).

Structure determination

Unlike crystalline materials, the position of all atoms in an amorphous material of substantial size cannot be described by a small amount of information and can therefore not be determined exactly by experiment (94). A great leap forward in the theory of the structure of glasses was made by Zachariasen in a landmark paper in 1932 (24). Zachariasen proposed that oxide glasses of the form AX_2 could be described by a random network of corner-sharing tetrahedra, with each tetrahedron having one X atom at each of the four corners surrounding a central A atom, as depicted in Figure 3.5. Such an arrangement is ideal energetically because it fills the valence shell of each atom while keeping the electronegative oxygen atoms separated from one another. A CRN requires that all four corner oxygens are shared with a neighboring tetrahedron such that all bonding needs are satisfied. In contrast to crystalline forms of AX_2 materials, where the orientations of the tetrahedra are specified by the particular

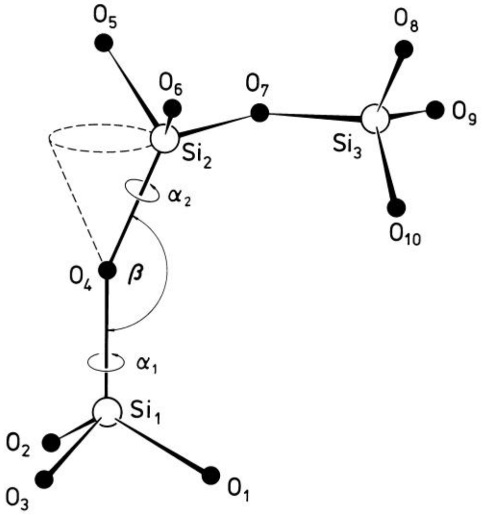


Figure 3.5 Local geometry of vitreous silica (SiO_2), showing three corner-sharing tetrahedra.

crystalline geometry, Zachariasen hypothesized that AX_2 glasses consist of alternative low-energy corner-sharing configurations of the tetrahedra represented by a CRN. His hypothesis soon gained support from the X-ray diffraction studies of Warren and co-workers (25-27).

It was not until the 1960's that questions were being asked as to what such glasses would look like in three dimensions. This led to the advent of structural modeling, the first models being of vitreous silica. While the initial models, such as the famous 614-atom hand-built CRN model of Bell and Dean (95), agreed well with experimental data, there were clear limitations to these physical models, including the difficulty of controlling the distribution of various structural parameters during construction.

With increasing computer power, it became possible to generate larger, less biased, and lower energy models than could be built by hand. Two classes of computational methods are commonly used to investigate glasses; one applies first-principles MD simulation (96), which due to computational requirements is limited to relatively small systems, and the other consists of iterative Monte Carlo (MC) algorithms similar to procedures one might follow in constructing a hand-built model (28). In addition to experimental validation, the quality of a model is generally measured by the amount of strain in the bond network, with the low spread in bond lengths and angles measured experimentally being ideal. According to this metric, higher quality models can be obtained by MC algorithms that allow greater relaxation of network strain than models produced by MD simulation (30), with some MC models having spreads in bond length and angle comparable to that of experimentally measured samples (34). While MD simulation permits the investigation of many properties that cannot be inferred from the static models produced by the iterative model-building algorithms, for the purposes of investigating the large length scale properties of interest in this thesis, very large models are essential. For this reason, the models of amorphous silicon and vitreous silica investigated in Chapter 6 are generated by methods specifically designed to efficiently produce large, low-energy CRN models. These methods all stem from the WWW algorithm, created by Wooten, Winer, and Weaire (28).

The WWW algorithm

The WWW algorithm begins from a crystalline structure and gradually introduces disorder into the bond network through numerous iterations of a "bond switching" step. This switch, illustrated in Figure 3.6, involves a rotation of approximately 90° of a chosen bond (in this case the bond between atoms B and C, or BC for short), followed by the swapping of two old covalent bonds for two new ones in order to maintain bond angles close to their ideal tetrahedral values. In this example, the covalent bond AB is replaced by bond AC, and likewise bond CD is replaced by bond BD. This bond switching step preserves the total number of bonds and the coordination of each atom.

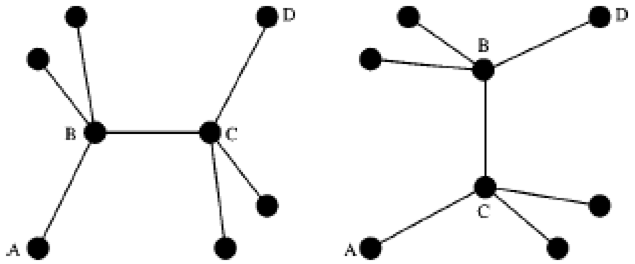


Figure 3.6 A schematic of a bond switch used in the WWW algorithm. Figure reproduced from (34).

After each iteration, the system is *geometrically* relaxed by minimizing the potential energy of the system while keeping the network of covalent bonds fixed. For monatomic materials such as silicon, the Keating potential (33) of the form

$$V = \frac{3\alpha}{16r_o^2}\sum_{l,i}(\boldsymbol{r}_{li}\cdot\boldsymbol{r}_{li} - r_o^2) + \frac{3\beta}{r_o^2}\sum_{l\{i,i'\}}\left(\boldsymbol{r}_{li}\cdot\boldsymbol{r}_{li'} + \frac{1}{3}r_o^2\right)^2 \qquad (3.1)$$

is commonly used, where r_o is the rest length of a covalent bond, $\boldsymbol{r_{li}}$ is the vector difference in the positions of bonded atoms l and i, and α and β are the parameters controlling the bond length and bond angle stiffness respectively. The Keating potential was used to construct the amorphous silicon models described in Chapter 6. Upon minimization, the new structure is accepted or rejected with a Boltzmann probability

$$P = \min\left[1, \exp\left(\frac{E_b - E_f}{k_B T}\right)\right] \tag{3.2}$$

that depends on the difference in the minimized energies before (E_b) and after (E_f) the bond switch, and the temperature T, given a value above the melting temperature T_m. If a sufficient number of iterations are performed, the model will lose all memory of the crystalline state and become fully amorphous.

Upon reaching a satisfactory amorphous state, the model is quenched by decreasing the temperature in small steps, each time allowing a new equilibrium to be reached through both *topological* relaxation (bond swapping) and *geometrical* relaxation (energy minimization within a fixed topology) . The temperature is decreased until an optimized amorphous structure is reached as the temperature approaches zero. A final relaxation can be performed by minimizing the total model energy with respect to the volume of the periodic cell, allowing the model to slightly expand or contract. Just as with experimental glass formation, the topology and structure of the model depends on the temperature at which the topology is frozen in. The models of amorphous silicon and vitreous silica built with modified WWW algorithm (30, 33, 34) and studied in Chapter 6 are among the largest and highest quality models built to date.

CHAPTER 4: REVIEW OF SIMULATION TECHNIQUES

Introduction

The human genome, encoding all the raw information needed to create a vibrant PhD student, is but a mere 3 billion nucleotide base pairs (7). From these nucleotide sequences, we can obtain protein sequences, and from protein sequences, folded structures (for intrinsically folded proteins), but the ultimate goal is to determine function. A complete understanding of such function, which can include catalysis of chemical reactions, chemical signaling, regulation of gene transcription, as well as mechanical infrastructure, requires more than just static structures; it requires knowledge about dynamics (3, 7). Since the determination of the first protein structures, those of hemoglobin and myoglobin by Max Perutz and Sir John Cowdery Kendrew by X-ray crystallography in 1958 (97), a protein's folded native state has typically been represented by a single conformation representing the best fit of an atomic model to the measured electron density. A major computational challenge has therefore been to determine, from a static structure, the ensemble of functionally relevant conformations in which a biomolecule can partake.

Molecular dynamics

Overview

The same tools that allowed Isaac Newton to understand the motion of the planets around the Sun form the basis of one of the most intuitive methods of conformational sampling. Newton's second law, which relates the translational acceleration of a body $\ddot{\boldsymbol{r}}_i$ to the total external force $\boldsymbol{F}_i$ via the equation

$$m_i\ddot{\boldsymbol{r}} = \boldsymbol{F}_i \tag{4.1}$$

was first used in conjunction with a realistic potential to iteratively update the positions and velocities of atoms by Aneesur Rahman in 1964 in his study of liquid argon (98).

The first simulation of a biomolecule came 13 years later with the 8.8 ps vacuum simulation of bovine pancreatic trypsin inhibitor by McCammon *et al.* (99).

Beginning from an initial set of atomic positions and velocities, a basic method for evolving a system forward in time is to discretize time into small intervals Δt (usually 1-2 fs) and use the velocity Verlet algorithm (100, 101) to update the positions, velocities, and accelerations from one time step to the next. Given the position $\boldsymbol{r}(t)$, velocity $\dot{\boldsymbol{r}}(t)$, and acceleration $\ddot{\boldsymbol{r}}(t)$ of each atom in the system at some time *t*, the velocity Verlet algorithm first finds the new positions at time *t*+Δ*t* using

$$\boldsymbol{r}(t+\Delta t)=\boldsymbol{r}(t)+\dot{\boldsymbol{r}}(t)\Delta t+\frac{1}{2}\ddot{\boldsymbol{r}}(t)\Delta t^2 \tag{4.2}$$

Next, accelerations are updated by finding the forces on each atom at its new position $\boldsymbol{r}(t+\Delta t)$. Lastly, the new positions and accelerations are used to update the velocity through the equation

$$\dot{\boldsymbol{r}}(t+\Delta t)=\dot{\boldsymbol{r}}(t)+\frac{1}{2}[\ddot{\boldsymbol{r}}(t)+\ddot{\boldsymbol{r}}(t+\Delta t)]\Delta t \tag{4.3}$$

which assumes that the average acceleration over the interval is equal to the mean of the end points.

Standard all-atom class 1 force fields used to characterize inter-atomic atomic forces have a potential energy *U(**r**)* of a form similar to (CHARMM (102))

$$\begin{aligned}U(\boldsymbol{r})={}&\sum_{bonds}K_b{(b-b_o)}^2+\sum_{angles}K_\theta{(\theta-\theta_o)}^2+\sum_{UB}K_{UB}{(s-s_o)}^2\\&+\sum_{bonds}\sum_{n}K_{\chi,n}(1+\cos(n\chi-\delta))+\sum_{impropers}K_\psi{(\psi-\psi_o)}^2\\&+\sum_{CMAP}U_{CMAP}(\varphi,\psi)+\sum_{ij}\varepsilon_{ij}\left(\left(\frac{R_{ij}}{r}\right)^{12}-2\left(\frac{R_{ij}}{r}\right)^{6}\right)+\frac{1}{4\pi\epsilon}\sum_{ij}\frac{q_iq_j}{r_{ij}}\end{aligned} \tag{4.4}$$

The potential can be divided into harmonic terms controlling the bond lengths *b*, the three-body angles *θ* formed by two covalent bonds meeting at a single atom, the Urey-

Bradley term between atoms separated by two covalent bonds, a sinusoidal term representing energy barriers between atoms separated by three covalent bonds as a function of the dihedral angle χ, the out-of-plane improper distortions ψ defined by a central atom and its three coplanar covalently bonded neighbors, and a term $U_{CMAP}(\varphi, \psi)$ serving as a correction to the backbone dihedral angle energy. Additionally, the energy function contains a pair of nonbonded terms, the first representing the van der Waals interaction in the form of a Lennard-Jones 12-6 potential and the second representing the Coulomb interaction between pairs of charges q_i and q_j. The degree to which the ideal geometry of each bonded interaction can be violated is dictated by a set of spring constants K_b, K_θ, K_{UB}, $K_{\chi,n}$, and K_ψ, the well depth of the van der Waals interaction ε_{ij} between atoms i and j, and the dielectric constant ϵ describing the extent of screening of the Coulomb interaction. The potential is individually parameterized for all atom types, typically resulting in a few thousand parameters, many of which are calibrated directly against quantum mechanics calculations (103, 104) and experimental measurements (104).

Solvation

Unlike the first MD simulation of a biomolecule, which was performed in vacuum (99), modern simulations account for the effects of the solvent environment either explicitly by surrounding the molecule with atomic water molecules, or implicitly through a continuum approximation (105). Due to the need to keep an explicitly solvated protein (in a periodic simulation volume) a sufficiently large distance from its periodic images, the box needs to be made so large that there are typically several times more solvent atoms than there are protein atoms, greatly increasing the computational cost of simulations with explicit solvent. Researchers tried to circumvent this problem by inventing implicit solvent models that mimic the properties of a solvent. Whereas explicit water molecules are constantly colliding with the protein, exchanging energy and creating a viscous environment, implicit solvent models use Langevin dynamics (100, 101) in which the force on each atom, given by

$$\boldsymbol{F}_i = -\nabla_{\mathbf{r}_\mathrm{i}} U(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) - m_i \gamma_i \dot{\boldsymbol{r}}_i + \boldsymbol{R}_i(t) \tag{4.5}$$

contains two force terms in addition to the gradient of the potential, one proportional to velocity that represents viscous drag, and a second term that mimics random collisions. Common implicit solvent models include the computationally expensive Poisson-Boltzmann model (105), which solves the Poisson-Boltzmann equation for the electrostatic environment of a solute in a solvent with ions, and the more efficient Generalized Born model (105), which is an approximation to the linearized Poisson-Boltzmann equation. A third, even less computationally demanding implicit model is the EEF1 model (106), which approximates the free energy of solvation by estimating how much of each atom's total possible solvent exposure is occluded away by surrounding atoms. The EEF1 model also includes a distance-dependent dielectric constant. Molecular dynamics simulations performed using the EEF1 implicit solvent model are discussed in Chapter 7.

Simplified methods

Normal mode analysis

Normal mode analysis (NMA) (46, 107-109) assumes that the global properties of the energy landscape can be estimated, to a first approximation, from the local curvature of the landscape about a native state conformation. This equates to determining the harmonic response to perturbations of each atom's position (the Hessian matrix) and diagonalizing it to re-express this curvature in the basis of normal modes. The functionally relevant motions, which tend to correspond to the low-frequency modes, are heavily determined by the overall shape of a protein and can be estimated from coarse-grained C_α-based network models. The most common such model is the Elastic Network Model (45-48) in which a protein is modeled as a set of infinitesimal beads (C_α)

connected by a harmonic spring to all other C$_\alpha$ atoms within a cutoff distance. The probability of a particular conformation in such a harmonic model is Gaussian along each individual normal mode directions with a width proportional to $1/\sqrt{k}$, where k is the curvature of the energy harmonic well along the normal mode. While computationally efficient, NMA and ENM have several drawbacks. One drawback is that non-globular proteins with significant conformational flexibility often have a high degree of anharmonicity along their most flexible directions of deformation, as they often involve hinge-like motions about the polypeptide backbone. Another drawback is that any significant displacement along an individual or linear combination of normal modes creates unphysical stereochemical distortions in all-atom models.

Essential dynamics

Essential dynamics (110, 111) is a method related to NMA in which the Hessian, which contained local curvature information in NMA, is replaced by a covariance matrix

$$\mathrm{cov}(\boldsymbol{X}) = \mathrm{E}[(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])^{\mathrm{T}}] \tag{4.6}$$

calculated using conformations $\boldsymbol{X}$ from sources such as molecular dynamics simulation. Diagonalization of the covariance matrix results in a set of eigenvectors, which in this case represents principal components instead of normal modes. A harmonic profile is still implicitly assumed, as the variance represents the second moment of the conformational distribution. Characterizing a distribution by its variance is equivalent to fitting it to a Gaussian distribution. Sorted in ascending order of their eigenvalues, principal components represent the directions of greatest variance remaining in the data after being projected along the directions of all lower principal components. Essential dynamics have the advantage that conformations outside of the native basin can contribute to the modes, but they only do so if such conformational variation is present in the input ensemble used to define the covariance matrix. The resulting principal components serve as an effective dimensional reduction technique by identifying directions of high flexibility and can in turn be used to help guide subsequent molecular dynamics simulations in directions that are likely to contain low energy conformational states. Such sampling methods are called *enhanced sampling techniques* (111).

Constrained geometric simulation

All of the previous techniques, including MD simulation, involve varying degrees of approximation regarding the true nature of the system that serve as trade-offs between realism and computational efficiency. For example, all-atom MD energy functions contain explicit dihedral angles but lack the quantum nature of the system, whereas ENM efficiently estimates large-scale motions but treats a biomolecule very much like a continuous elastic solid. There is plenty of room in the middle of these two extremes.

The constraint-based model FRODAN (10), used in the biological portion of this thesis, is one such example and shares a likeness to the very first ball-and-stick hand-built models used by pioneers in the field such as Watson and Crick, as shown in Figure 4.1. Whether the metal and plastic models were of DNA or proteins, these early models

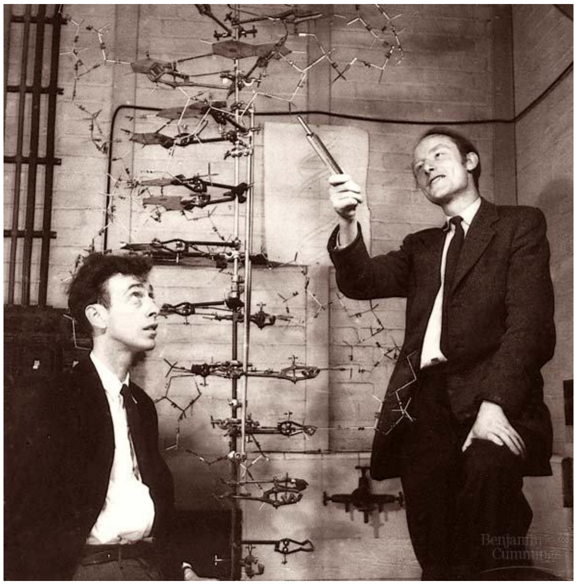


Figure 4.1 Watson (left) and Crick (right) analyzing a model of DNA.

implicitly assumed that bond lengths and angles are approximately fixed and that motion is confined to the subspace of torsion angles. Within this subspace, early model builders found conformations that allow DNA bases to pair up through the formation of hydrogen bonds, whereas in proteins certain backbone angles were found that cause the polypeptide to form helices, also possessing favorable hydrogen bonds. Treating non-covalent interactions as distance constraints between pairs of atoms, such as the hydrogen and oxygen atoms of a hydrogen bond for example, one could ask the following question: given the set of rigid bond lengths and angles and the set of non-covalent constraints, which parts of the model are flexible and which are rigid?

Rigidity analysis: FIRST

Such a question, viewing a mechanical system in terms of a set of fundamental bodies (i.e. nucleotide bases etc.) connected by distance constraints, can be expressed formally through the mathematical discipline of graph theory (112). The object of study in graph theory is called a “graph” and consists of a set of vertices and a set of edges that connect these vertices, as shown in Figure 4.2. While the uses of graph theory are very broad, of interest here is the subdiscipline of rigidity theory (113) and its application to biomolecules (114), where vertices represent objects with spatial degrees of freedom and edges represent distance constraints. Insight into the application of rigidity theory to more complex systems such as biomolecules can be gained from the history of rigidity analysis.

Perhaps the first person to use graph theory to infer the mechanical properties of a network was James Clerk Maxwell while studying the structural integrity of bridges. Treating the joints of a bridge as the set of vertices V and the beams as the set of edges E, he realized that he could approximate the number of degrees of freedom (dof) N of the framework (i.e., the number of independent motions that kept the lengths of all the beams

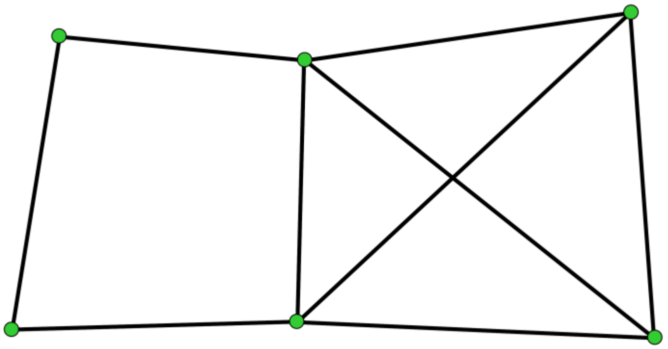


Figure 4.2 A graph containing 6 vertices (green dots) and 9 edges (black bars).

fixed) by the equation $N \cong d|V| - |E|$ where $|V|$ is the number of vertices, $|E|$ the number of edges, and d the dimensionality of the space (three in the case of bridges). The limitations of his equation, called a Maxwell count, can be seen by applying it to the two-dimensional graph in Figure 4.2. The 6 vertices, which have a total of 6 x 2 = 12 dof, are constrained by 9 edges, leading to a Maxwell count of 12 – 9 = 3 dof. As all bodies in two-dimensions have at least 3 dof (2 translational and 1 rotational), the Maxwell count would imply that the graph is completely rigid. This is clearly not the case, as the left side of the graph contains an internal dof, called a *floppy mode* that allows the rhombus to be sheared without changing the lengths of any of the edges. The failure of the Maxwell count is due to its inability to recognize that the right half of the graph contains more edges than are necessary to make it rigid (try removing one and see). One of the edges is therefore said to be redundant. A redundant constraint does not remove a dof from the system, as a rigid object already possesses the minimal number of dof. This problem can be avoided by removing redundant constraints prior to performing the Maxwell count, but identification of all redundant constraints in general requires Maxwell counts to be performed on all possible subgraphs, the number of which grows exponentially with the size of the system, making an exhaustive analysis of even moderately sized graphs infeasible.

Interestingly, the verbal expression of bodies "possessing" dof and constraints "taking them away" offers a subtle clue to an efficient algorithm for determining the rigidity of frameworks that requires at most order N^2 steps (in practice scaling closer to $N^{1.2}$). Such an algorithm, called *The Pebble Game*, was found by Jacobs and Thorpe in 1995 (115). In the Pebble Game, each vertex "possesses" one pebble for each of its dof (two for the example in Figure 4.2) which are "taken away" by the edges. As an edge (i.e. a distance constraint) can only remove one dof, the effect of the edge is to require one pebble from one of its terminal vertices to be placed on it *if and only if* a few specific rules are satisfied (115). The beauty of the algorithm is that the satisfaction of a few simple rules is sufficient to ensure that no redundant edge is ever covered and only one attempt to cover each edge is necessary for the algorithm to converge. From the final arrangement of pebbles, one can determine the regions that are rigid and those that are not. It might seem fortuitous that such a "pebble game" exists for characterizing rigidity. It can be shown that rigidity in three dimensions lacks an associated pebble game if the vertices are modeled as points with three dof. Luckily one does exist if the vertices are modeled as objects with six dof (116), which can be interpreted as the three translational and three rotational rigid-body dof. The Pebble Game algorithm is contained within the software package FIRST (Floppy Inclusions and Rigid Substructure Topography) (114).

When modeling the flexibility of biomolecules, each atom is treated as a generic body with six dof. Each stereochemical interaction can be modeled by a certain number of edges between pairs of bodies, each edge removing one dof from the system. For example, a single covalent bond is represented by 5 edges, as two bodies connected solely by a covalent bond possess only 7 of the 2 x 6 = 12 dof they would have in the absence of the covalent bond, as the connected bodies have the 6 rigid-body dof and one internal torsion angle. Similarly, a double bond is represented by 6 edges due to its lack of rotational freedom. It is more difficult to rigorously infer from mechanical behavior

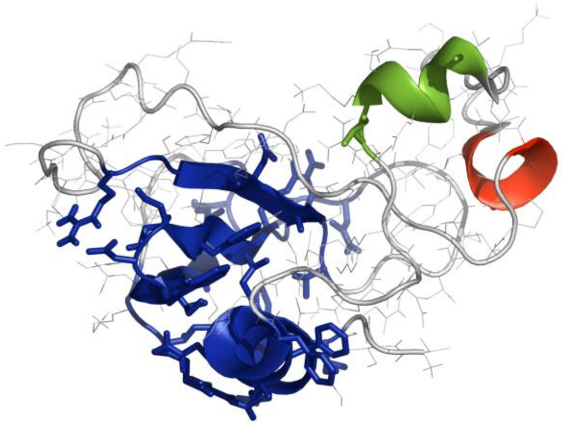


Figure 4.3 Large rigid regions within barnase. Figure reproduced from (117).

the number of dof removed by each of the non-bonded interactions (hydrogen bonds, salt-bridges, and hydrophobic contacts), but values of 5, 5, and 2 respectively result in flexibilities with optimal agreement to experiment and MD simulation (117).

When both stereochemical and non-bonded constraints are used to determine the rigid clusters, alpha helices and sufficiently large beta sheets possessing standard backbone hydrogen bonding pattern form single rigid clusters, as shown in Figure 4.3 for barnase (117). Interestingly, high densities of specific side-chain constraints can collectively cause rigidity to “percolate” between rigid secondary structures, forming the large rigid cluster in Figure 4.3. The net effect of non-bonded constraints on the polypeptide chain is to reduce the number of rotatable torsion angles. Rigidity analysis can therefore be viewed as an intuitive and chemically justified method of dimensional reduction.

Conformational sampling: FRODAN

Rigidity analysis can quickly determine which parts of a framework are rigid and which contain floppy modes that allow flexibility, but it does not determine the amplitude of these modes. This limitation is somewhat analogous to that of NMA and ENM: analysis is performed on a single static structure with a single geometric relationship between the atoms. Determining how far along a floppy mode a protein can move before encountering steric clashes or limitations due to the complex network of stereochemical constraints is an incredibly difficult problem that can only be approximately solved by building a computational model, similar in spirit to those of Watson and Crick, and exploring the accessible conformational space.

Such a model, called FRODAN (10), was created by Daniel Farrell and is based on an older version called FRODA (117) developed by Stephen Wells. The first step within FRODAN is to determine the rigid units (RUs) that will serve as the fundamental mobile components of the model. Unlike FRODA, which used both covalent and non-covalent constraints to determine RUs resulting in large clusters such as those in Figure 4.3, FRODAN uses only covalent bond and angle constraints to perform the rigidity analysis. Non-covalent interactions in the form of hydrogen bonds, salt bridges, and hydrophobic contacts are instead modeled as tethers, acting as upper limit “less than” distance constraints between pairs of atoms in different RUs, discussed in more detail later this section. The RUs defined using only covalent constraints have a very intuitive property: for fixed bond lengths and angles, the relative positions of all atoms within a RU are fixed, while the relative position between pairs of atoms in different RUs can vary through the rotation of torsion angles. This is made clearer by looking at the RU decomposition of phenylalanine shown in Figure 4.4A. Imagining that you are one of the carbon atoms of the aromatic ring, your position is fixed relative to the other atoms in the aromatic ring as well as the atoms sharing a covalent bond with the ring. Likewise, taking the perspective of the C_α and C_β atoms, your position relative to your four covalent neighbors is fixed, but no others. An interesting consequence of the rigid unit decomposition is that a single atom can be shared by more than one RU. This is a natural consequence of the fact that torsional rotation preserves a bond’s length while creating relative motion between the rigid bodies that it connects. Analyzing a full polypeptide

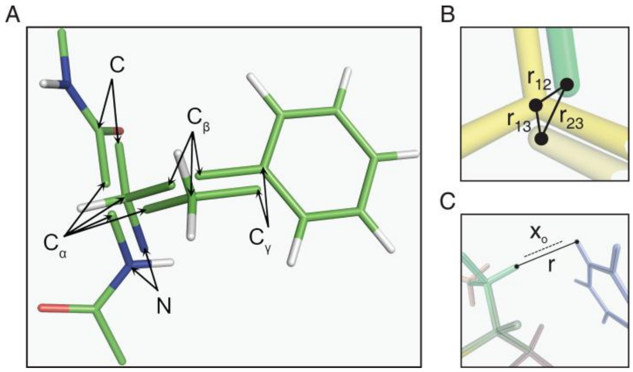


Figure 4.4 (A) Decomposition of phenylalanine into rigid units. The shared atoms are labeled and are non-overlapping simply for clarity. (B) Demonstration of the “shared atom” constraints that connect rigid units. (C) Demonstration of a “greater than” constraint enforcing steric repulsion between two atoms. Figure modified from (10).

chain, one would find that all torsion angles are represented by shared edges between pairs of rigid units.

Now that RU decompositions with and without non-covalent interactions are better understood, an explanation is needed for their absence from the rigidity analysis in FRODAN. Both choices lead to models that are equally simple conceptually and roughly equal in terms of computational efficiency, but whereas hydrogen bonds and hydrophobic interactions located in large RUs have both their lengths and angles locked, in FRODAN they always possess a small window of possible lengths and angles (no angular restrictions exist for hydrophobic contacts). Collectively, these small windows allow alpha helices and beta sheets to have a small amount of flexibility instead of being strictly rigid blocks, resulting in protein models with conformational subspaces that agree much more closely with those explored during MD simulation. It should be stressed that regions that are rigid in FRODA still have very rigid behavior in FRODAN due to the dense array of upper distance constraints imposed on the smaller RUs by the set of non-covalent interactions.

It may have come to your attention that whether the RUs be the larger ones of FRODA or the smaller ones of FRODAN, the relative motions of the RUs must be limited by more than just non-covalent constraints, otherwise phenylalanine shown in Figure 4.4A would fall apart. In FRODAN, which will be the sole focus of discussion from this point forward, these additional constraints are similar to those for the non-covalent interactions in that they are not constraints in the graph theory sense, but act to limit the relative motion of the RUs. These constraints can be categorized into three types: equality constraints, “less than” constraints and “greater than” constraints. The constraints are enforced by minimizing an objective function that is zero if the constraints are met and rises quadratically as constraints are violated. Equality constraints have potentials of the form

$$E_{equality} = \frac{1}{2} k \Delta x^2 \tag{4.7}$$

where Δx is the separation between two atoms. The only examples of these are the “shared atom” constraints that force copies of the same atom in different rigid units to be located at the same place, as shown in Figure 4.4B. “Greater than” constraints are half-harmonic springs with potentials of the form

$$E_{gt} = \begin{cases} \frac{1}{2} k (x - x_0)^2, & x < x_0 \\ 0, & otherwise \end{cases} \tag{4.8}$$

that try to keep the distance x between two atoms greater than some bound x_o. These include the steric interactions, shown in Figure 4.4C, that prevent atoms from overlapping, as well as the constraints that enforce proper Ramachandran angles and torsion angles. The use of distance constraints to maintain proper Ramachandran angles follows the work of Ho *et al*. (118) and Farrell *et al*. (10). Similarly, defining minimum allowable distances between all pairs of 1-4 atoms on either side of a rotatable bond can be used to enforce low-energy staggered conformations. Lastly, “less than” constraints are half-harmonic springs with potentials of the form

$$E_{lt} = \begin{cases} \frac{1}{2}k(x - x_0)^2, & x > x_0 \\ 0, & otherwise \end{cases} \tag{4.9}$$

that try to keep the distance x between two atoms less than some bound x_o and include hydrogen bonds and salt bridges, as well as hydrophobic interactions. Hydrogen bonds and salt bridges are identified as those having an energy E < -1.0 kcal/mol according to a modified Mayo potential (119, 120). Hydrophobic interactions occur between pairs of non-polar carbon or sulphur atoms separated by less than 3.9 Å that belong to the side-chains of hydrophobic residues. Overall, the constraint-enforcing energy function within FRODAN can be summarized as

$$E_{prot} = E_{equality} + E_{lt} + E_{gt} \tag{4.10}$$

Conformations that satisfy all bonded and non-bonded constraints have zero energy. There is no consideration of electrostatic interactions or solvation effects other than those implicit in the non-bonded constraints. With the exception of small torsional energy barriers, the energy landscape is therefore flat within the allowed floppy mode subspace, outside of which it rises harmonically.

The allowed subspace is explored by independently perturbing the positions and orientations of all RUs, followed by a conjugate gradient minimization of the constraint

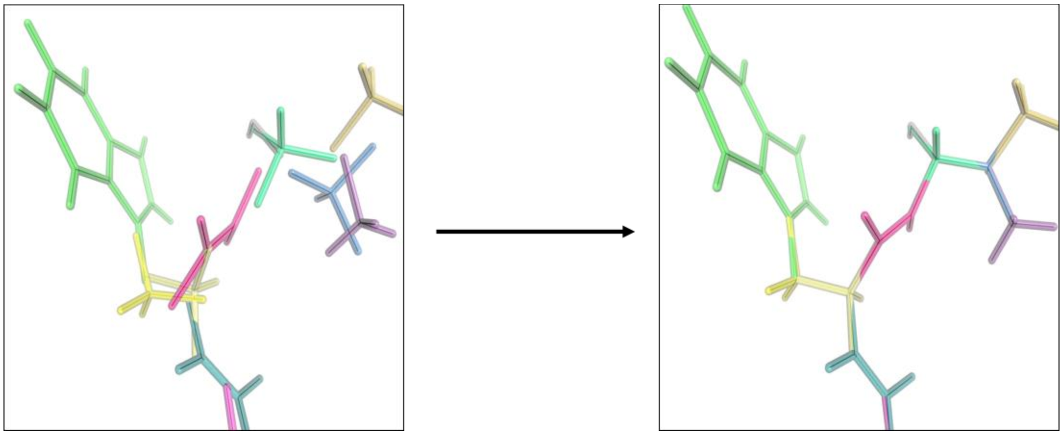


Figure 4.5 Example iteration in FRODAN involving an initial perturbation of the rigid units, followed by re-enforcement of the constraints (10). Figure courtesy of Daniel Farrell.

energy E_{prot} that enforces the set of constraints, as shown in Figure 4.5 These perturbations can be quite large, involving translations of up to 2 Å and rotations of up to 180^{o}. If upon perturbation, the constraints cannot be satisfied to within a strictly chosen tolerance, the structure is reverted to its last good conformation and a new perturbation is performed. The ability to traverse torsional barriers in a single perturbation step and the flat energy landscape of the FRODAN model can be seen as two of FRODAN's greatest strengths, as it allows the conformational subspace to be extensively sampled far more rapidly than with MD techniques (55). In fact, many MD sampling techniques exist that attempt to flatten out the landscape (121), whereas this is implicit in the constraint-based model.

CHAPTER 5: FINITE SIZE CORRECTION FOR SCATTERING FROM NANOMATERIALS

Introduction

The atomic pair distribution function describes the distance-dependent density of a material as viewed from an average atom. It links microscopic atomic positions with macroscopic experimental observables such as pressure, compressibility, energy, and phase transitions (16). It can be determined either experimentally by taking the Fourier transform of neutron or X-ray diffraction data or from computer-generated structure models (13). We will focus on the radial distribution function (RDF) that is closely related to the pair distribution function. A comparison between the measured and computed RDFs provides insight into the structural origin of experimental observables. For example, RDFs have been used to probe the architecture of novel amorphous and porous materials (17), illustrate the phase transition across the optimal doping of superconducting materials (18), and detect randomness in periodic superlattices (19).

The computation of an RDF consists essentially of counting the number of atoms within a thin shell a given distance away from an average atom. In general, it is affected by two types of structural properties. The first type relates to the intrinsic geometry of the atomic network, i.e., the average coordination of each atom, the distortion of bond lengths and bond angles, and the randomness of the atomic network. These properties influence how atoms are placed with respect to each other. They determine the positions, intensities, widths, and overlaps of the peaks in the RDF. The second type relates to spatial confinement, i.e., the shape and the size of the material sample. They determine the envelope of the RDF. Infinite in all directions, a bulk material has neither shape nor size. Thus the RDF of a bulk material is only determined by the intrinsic geometry of its atomic network. In contrast, the RDF of a nanomaterial is a function of its shape and size in addition to the atomic geometry (122). A nanomaterial, by definition, is smaller than 1 μm in at least one dimension and thus a non-negligible fraction of the atoms are on or close to the surface of the material. These surface atoms are surrounded partially by the material and partially by vacuum. The density distributions viewed from these atoms differ from those viewed from the deeply buried atoms. Since the RDF of a nanomaterial is the average of the density distributions viewed from all atoms, the RDF entangles the contributions from both the intrinsic atomic geometry and the spatial confinement.

The main research interest here is to describe a method for removing the effects of the finite nature of nanomaterials on scattering data so that the RDFs for all materials sharing a common atomic geometry fall on a single universal curve. The determination of the shape and size of a nanomaterial are usually not the goal of RDF analysis, as they can be obtained from experimental techniques such as small-angle X-ray scattering (20) and transmission electron microscopy (21). Most conventional forms of RDFs discussed in textbooks and the literature, however, do not take spatial confinement factors into consideration. This is not a surprise, as most of the RDF theory was developed in the days when bulk materials were the main if not the sole research subjects in condensed-matter physics and materials science. With ever-growing interest in nanomaterials, it is desirable to have a form of RDF that is free of the spatial confinement effects so deviations in the intrinsic atomic geometry can be more easily compared.

As my contribution to this work, I derived the shape factor for an infinite cylindrical rod (see Appendix A) and corrected the code that creates nanomaterial RDFs and maps them to the universal curve. I also had a significant role in writing the resulting paper (123) and determining its logical flow.

Theory and methodology

Under the general name of pair distribution functions, several sets of functions are used in the powder-diffraction community (124). The nomenclature used here follows that of the book by Warren (13). The most intuitive of the distribution functions is the RDF, defined as

$$R(r) = \frac{1}{N}\sum_{ij}\frac{w_i w_j}{\langle w\rangle^2}\delta\left(r - r_{ij}\right) \tag{5.1}$$

where r_{ij} is the interatomic distance between atoms i and j, δ is the Dirac delta function, the w_i's are the atomic weight factors suitable for X-ray or neutron scattering, and $<w> = \Sigma_i w_i/N$, where N is the number of atoms in the material. The sum in Eq. (5.1) is over all atom pairs. The function that is found directly from the structure factor $S(Q)$ measured experimentally is the reduced pair distribution function $G(r)$. The form of $G(r)$ is found from $S(Q)$ according to

$$G(r) = \frac{2}{\pi}\int_0^\infty Q[S(Q) - 1]\sin(Qr)\,dQ \tag{2.5}$$

This commonly used equation is somewhat misleading, as it assumes that $S(Q)$ does not contain contributions from small-angle scattering (SAS), a point discussed in more detail later in this section. For bulk materials, $G(r)$ can be related to $R(r)$ through

$$G(r) = 4\pi r\rho_o\left(\frac{R(r)}{4\pi r^2\rho_o} - 1\right) \tag{5.2}$$

The derivation of a universal function characteristic of any given material free of finite-size effects will be performed using $R(r)$ due to its intuitive nature. For a bulk material, the RDF $R(r)$ approaches $4\pi r^2\rho_o$ at large distances, where ρ_o is the average density of the material. A reduced RDF (RRDF) $P(r)$ is frequently used in the literature to normalize out the long-distance trend

$$P_b(r) = \frac{R(r)}{4\pi r^2\rho_o} \tag{5.3}$$

so that at large distances, this function approaches 1 for bulk systems. The subscript b indicates that this form of the RRDF only possesses the desired normalization behavior for bulk systems. For bounded systems, $P(r)$ would decrease asymptotically as $1/r$ and

1/r^{2} for nanomaterials finite in 1 and 2 dimensions respectively, and would be exactly zero beyond some maximum distance for nanomaterials finite in all three dimensions. For such nanomaterials, it is also desirable to have a similarly defined distribution function that has the same flat baseline of unity at large distances and be independent of the shape of the material, depending only on the intrinsic atomic geometry of the material.

The reason the RRDF as defined by Eq. (5.3) trends away from unity for a material bounded in one or more dimensions is that a spherical shell of radius *r* placed about a typical atom can have part its surface outside of the bounded material, whereas this can never occur for a bulk material. The RDF of a bounded material is therefore always less than its bulk equivalent, causing the RRDF to trend below unity. For the infinite sheet and infinite rod, their long-distance trends indicate that the average fraction of the spherical shell lying outside the boundary of the material decreases as *1/r* and *1/r*^{2} respectively. The function that describes the distance dependence of this fraction has been called the characteristic function of the shape (122) or the nanoparticle form factor (125) in the literature, but for clarity I will simply refer to it as the shape factor, $\alpha(r)$.

Shape factor

The shape factor $\alpha(r)$ is equal to 1 at all *r* for a bulk material. For materials with boundaries, $\alpha(r)$ can be written as a Taylor series expansion about small *r* as

$$\alpha(r) = 1 + c_1 r + O(r^2) \tag{5.4}$$

The value of c_1 can be shown to be *–S/4V*, where *S/V* is the surface to volume ratio of the nanomaterial. The argument is based on consideration of length scales such that the surface is approximately locally flat on small enough scales. Let us consider an atom lying at a distance *a* inside a surface. When we construct $R(r)$ for $r > a$, part of the spherical shell extending from *r* to $r + dr$ centered at atom will lie in empty space rather than within the material, and thus the contribution of the atom to $R(r)$ will be less than that of an atom in the bulk. The lost contribution can be quantified in terms of the fraction of the surface area of the sphere of radius r that lies outside of the boundary of the nanomaterial. This "missing" contribution is that of a spherical cap, equal to $2\pi r(r - a)$, while the remaining surface area is $4\pi r^2 - 2\pi r(r - a)$. We now consider that there will be a missing area contribution in $R(r)$ from all points lying within $0 < a < r$ of the surface. We therefore integrate the missing and remaining contributions. For the missing contribution we have

$$\int_0^r 2\pi(r^2 - ar)da = \pi r^3 \tag{5.5}$$

and for the remaining contribution,

$$\int_0^r [4\pi r^2 - 2\pi(r - a)]da = 3\pi r^3 \tag{5.6}$$

The net effect is that we are missing 1/4 of the total contribution to $R(r)$ from points lying within a distance r of the surface. For a nanomaterial of volume V and surface area S, the volume lying within a very small distance r of the surface is rS, which is a fraction rS/V of the total volume of the nanomaterial. Therefore, $R(r)$ for the nanomaterial at small r will be equal to $R_b(r)$, the value for the infinite bulk material, less 1/4 of the contribution from the "surface volume;" so the shape factor $\alpha(r)$ to first order in r is

$$\alpha(r) \cong 1 - \frac{S}{4V}r \tag{5.7}$$

At larger values of r, $\alpha(r)$ will deviate from this linear form as the assumption that the surface is locally flat begins to break down. We can confirm that $\alpha(r)$ for all the shapes we list in this work behave as Eq. (5.7) at low r. This indicates that $\alpha(r)$ can be similar for solids of different shapes, *e.g.*, different ellipsoids, as the leading term in $\alpha(r)$ depends only on the surface-to-volume ratio S/V. *This suggests a limitation on the amount of shape information that can be obtained from RDF studies on nanomaterials*.

While this prior derivation of the first order term in $\alpha(r)$ was made more intuitive by considering the limit of small values of r, a similar procedure can be followed for all values of *r* to derive a complete expression for the shape factor of any object. The steps followed for small *r* can be written as a double integral, one corresponding to integrating over the shell centered around a fixed “observer” atom, located somewhere within the material, and the second integral acting to average this result over all possible observer positions. Written mathematically, this is equivalent to a density-density autocorrelation function $c(\boldsymbol{r})$ of an object of the desired shape and uniform density, namely

$$c(\boldsymbol{r}) = \frac{1}{\rho_o V} \int_0^\infty \rho(\boldsymbol{R})\rho(\boldsymbol{R} + \boldsymbol{r}) d^3 R \tag{5.8}$$

where $\rho(\boldsymbol{r})$ is the three-dimensional density distribution of the object of interest and V is its volume. The autocorrelation is normalized here so as to have a maximum density of 1 at $c(\boldsymbol{0})$. Note that $c(\boldsymbol{r})$ is proportional to the probability of finding two units of density within the object with separation $\mathbf{r}$. The RDF is by its very nature spherically averaged, depends only on the magnitude of $\boldsymbol{r}$, and can therefore be found by performing a spherical integration of $c(\boldsymbol{r})$ about the origin. For objects of uniform density, $R_u(r) = 4\pi r^2 \rho_o\, \alpha(r)$, allowing $\alpha(r)$ to be found directly from the spherical average of $c(\boldsymbol{r})$.

Not all shapes have shape factors that can be solved in closed form and must be solved numerically. As a simple example, applying Eq. (5.8) for a sphere of radius *R* and dividing by $4\pi r^2 \rho_o$ produces the shape factor

$$\alpha_{sphere}(r) = \begin{cases} \left(1 - \frac{3}{4a} r + \frac{1}{16a^3} r^3\right) & r \leq 2a \\ 0 & r > 2a \end{cases} \tag{5.9}$$

which can be seen to possess the same first order term derived earlier.

Shape factors that are commonly used are Gaussian [$\exp(-r^2/\sigma^2)$] and exponential [$\exp(-r/\sigma)$], where σ is the length scale describing the nanomaterial. In Figure 5.1 we show $\alpha(r)$ for a sphere of radius a = 10 Å and compare it with two Gaussian and two exponential shape factors. The Gaussian shape factors are shown with $\sigma = a$ and $\sigma = 2a$, and of the two exponential shape factors, one has a length scale $\sigma = a$ and the other has

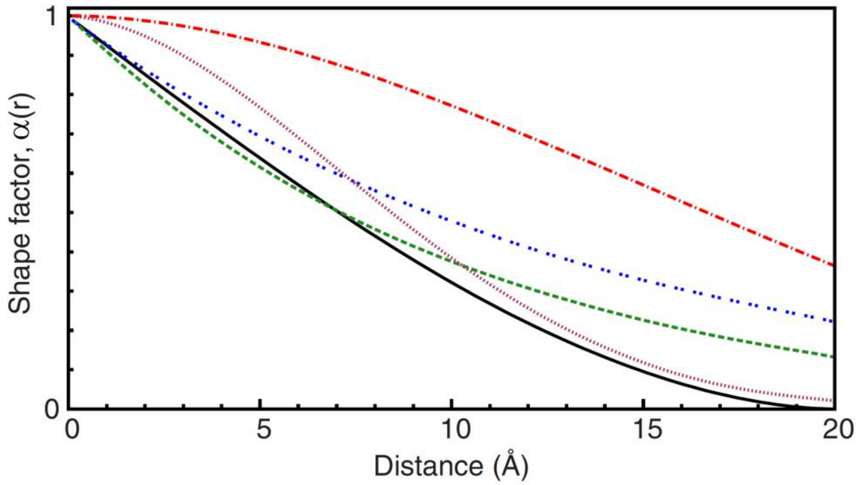


Figure 5.1 The shape factor for a sphere of radius a =10 Å (black solid) compared with four commonly used shape factors. The Gaussian shape factors have length scales σ = a (purple dot) and σ = 2a (red dot dash) while the exponential shape factors have σ = a (green dash) and σ such that the gradient at r = 0 Å matches the gradient of α(r) for the sphere (blue double dot).

the gradient at $r = 0$ matched to the gradient of $\alpha(r)$ for the sphere. None of these functions is a good match to the actual shape of $\alpha(r)$ for the sphere, with the Gaussian even lacking the proper linear behavior at small r. The Gaussian and exponential shape factors can also be shown to fail at modeling $\alpha(r)$ of spheroids and other simple geometric shapes. Great caution should therefore be taken when using Gaussian or exponential shape factors in the interpretation of RDF data on nanomaterials. For films, cylinders, etc., the volume of the nanomaterial would of course be infinite. By combining Eq. (5.7) with the general sum rule $V_n = \int_0^\infty 4\pi r^2 \alpha(r)\, dr$ relating the volume V_n to the shape factor of a nanomaterial and applying it to an exponential form factor $\alpha(r) = \exp(-r/\sigma)$ leads to a volume of $8\pi\sigma^3$ and a surface area of $32\pi\sigma^2$ with at least one dimension being infinite in extent. It is unlikely that such a shape of uniform density exists. Thus we recommend that in the absence of any information concerning the shape of the nanomaterial, it is better to use the form for a sphere given in Eq. (5.9), with an appropriate choice of the radius a.

Boundary corrections

If the RRDF in Eq. (5.3) is instead defined as

$$P(r) = \frac{R(r)}{4\pi r^2 \rho_o \alpha(r)} \tag{5.10}$$

the distance dependence of *R*(*r*) is matched by that of *α(r)*, causing this more general RRDF to fluctuate about unity for a nanomaterial of any shape and size, as desired. The general RRDF *P*(*r*) is shape-invariant, depending only on the intrinsic atomic geometry of the bulk material and not on the shape and size of possible boundaries.

Rewriting Eq. (5.3) as $R(r) = 4\pi r^2 \rho_o P(r)\alpha(r)$ and recognizing that $4\pi r^2 \rho_o P(r)$ is the RDF $R_b(r)$ of the bulk material, Eq. (5) is equivalent to

$$R(r) = R_b(r)\alpha(r) \tag{5.11}$$

The RDF of an undistorted nanomaterial can therefore be expressed as the product of two independent distributions, one containing only information regarding the intrinsic atomic geometry of the material and the other describing only the effects of spatial confinement. From Eq. (5.8), *α(r)* for a nanomaterial can also be interpreted as describing the probability that two randomly chosen points from the bulk material separated by a distance *r* will be found within the boundary of the nanomaterial. The nanomaterial can be imagined to have been cut from the bulk material without undergoing deformation or

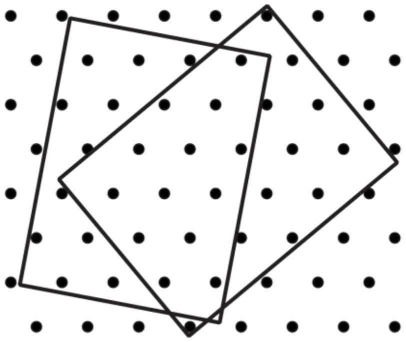


Figure 5.2 The statement that $R(r) = R_b(r)\alpha(r)$ is equivalent to averaging the RDFs of the ensemble of nanomaterials cut from the bulk at all locations and orientations with equal probability. Two such random locations and orientations are displayed for the case of rectangular nanomaterials cut from a triangular lattice.

reconstruction. Unfortunately, knowing the bulk material and the boundary describing the shape and size does not uniquely specify the nanomaterial, as it could be cut from the bulk at any location and orientation, each giving a different realization of the nanomaterial with a different RDF, as shown in Figure 5.2. The RDFs of the two cuts in Figure 5.2 not only contain peaks of different amplitude but the rightmost cut contains a peak due to the atoms in the upper and lower corners, that is, entirely absent in the first cut. The apparent dilemma is due to the fact that $\alpha(r)$ depends only on a nanomaterial's shape and size and is defined for an object of uniform density, while here the density is inhomogeneous at the atomic level. The problem is resolved and Eq. (5.11) made exact if cuts at *all locations and orientations* are sampled with equal probability and $R(r)$ is the average RDF of the ensemble. For any sample of nanomaterials that does not contain an equal representation of all boundary locations and orientations, as is the case for nonspherical nanomaterials with preferential directions of growth that correlate with the underlying atomic geometry (126), Eq. (5.11) is only an approximation for the average RDF of the sample. As the shape factor $\alpha(r)$ is independent of the density distribution within the boundary, it can be calculated by finding the RDF $R^u(r)$ (the superscript *u* stands for uniform density) of a material of uniform density ρ_o of the desired shape and size, and dividing it by the RDF of the uniform bulk, $R_b{}^u(r) = 4\pi r^2\rho_o$. The general RRDF for infinite or bounded materials can therefore be written as

$$P(r) = \frac{R(r)}{R^u(r)} = \frac{R(r)}{R_b^u(r)\alpha(r)} = \frac{R(r)}{4\pi r^2 \rho_o \alpha(r)} \qquad (5.12)$$

For bulk materials, $\alpha(r) = 1$ and the RRDF given by Eq. (5.12) reduces to Eq. (5.3). Therefore, Eq. (5.12) is an extension of an already widely used distribution function. *The RRDF P*(r) is a means *of plotting RDF data such that data for nanomaterials of all shapes and sizes with a common atomic arrangement fall on a single curve, allowing differences in their intrinsic atomic geometry to be more readily* *compared*. Although strictly speaking, the shape independence of *P*(*r*) in Eq. (5.11) is only true after averaging over nanomaterials with all possible locations and orientations with respect to the bulk material, in practice it can be used to approximate the RDF of a single realization of the nanomaterial, except at the very largest spanning distances within the nanomaterial, discussed further in the Results section. Care is needed when the nanomaterials are highly non-spherical, as for example in needles for which the deviations from spherical symmetry are strongly correlated with asymmetries of the atomic lattice (126). For nanomaterials, where all the spanning lengths are at the same length scales, deviations in *P*(*r*) from that of the bulk material can be ascribed to structural changes from the bulk due to surface relaxation and structural rearrangement (126).

While *R*(*r*) is intuitive, *G*(*r*) is the distribution determined directly from experimental data. The form of *G*(*r*) in Eq. (2.5) assumes that the structure factor *S*(*Q*) is measured down to a $Q_{\text{min}} > 0$, that is, large enough to exclude contributions from small angle scattering (SAS) (20) in either X-ray or neutron-scattering experiments, as the second term in *G*(*r*), namely, $4\pi r\rho_o$, is the contribution from $S(Q < Q_{\text{min}})$ for bulk materials (127). Within the small Q region containing the SAS data, scattering is unaffected by the atomic granularity of the density and is thus equal to that for a material of uniform density. The general form of $S(Q < Q_{\text{min}})$ for materials of any shape and size (127) is

$$S(Q) - 1 = \rho_o \int_0^\infty 4\pi r^2 \alpha(r) \frac{\sin(Qr)}{Qr} dr \tag{5.13}$$

Transforming the SAS data to *r* space using Eq. (2.5) gives $4\pi r\rho_o\alpha(r)$. Knowing that $\alpha(0) = 1$, the form of $\alpha(r)$ can thus be found directly from SAS data. If the second term in Eq. (5.2) is replaced by $4\pi r\rho_o\alpha(r)$, one finds the general form of *G*(*r*) that fluctuates about zero at large *r* for materials of all shapes and sizes, namely,

$$G(r) = 4\pi r\rho_o\left[\frac{R(r)}{4\pi r^2\rho_o} - \alpha(r)\right] = 4\pi r\rho_o\alpha(r)[P(r) - 1] \tag{5.14}$$

Solving Eq. (5.14) for *R*(*r*) gives

$$R(r) = rG(r) + 4\pi r^2\rho_o\alpha(r) \tag{5.15}$$

Inserting this expression for *R*(*r*) into Eq. (5.12), one gets the expression for the universal RRDF *P*(*r*) of the material that can be found directly from experimental data, namely,

$$P(r) = 1 + \frac{G(r)}{4\pi r\rho_o\alpha(r)} \tag{5.16}$$

The first term describes the baseline that represents the homogeneous density limit and the second term describes the fluctuations due to atomic geometry and granularity.

In addition to the finite extent of a material, another experimental limitation that affects the amplitude of the peaks in *G*(*r*) is the finite *Q*-space resolution of the instrument. The finite resolution has the effect of convoluting the true structure factor *S*(*Q*) by a resolution function, causing the true *G*(*r*) to be multiplied by an envelope function equal to the Fourier transform of the resolution function, thus dampening the peak amplitudes. For example, a Gaussian resolution function causes *G*(*r*) to be multiplied by the corresponding Gaussian envelope function and more complex functions can also be used (94). The finite resolution of the instrument acts on data from bulk materials and nanomaterials alike.

This raises the question of the best way to analyze experimental data and the key decision as to whether to compare theory (including computer simulations) in real space or reciprocal space (15). There are advantages to both approaches. If the resolution function of the instrument is unknown and has a significant effect on the structure factor *S*(*Q*), then there is little choice than to do the comparison in reciprocal space. One way to do this would be to use a RRDF *P*(*r*) as in the bulk material and obtain the reduced pair distribution function *G*(*r*) via Eq. (5.14). This requires some assumed form for the shape factor *α(r)* to be used, which will have to be obtained from microstructural information, small-angle scattering, a plausible guess, etc. Then the structure factor *S*(*Q*) can be obtained from the back sine Fourier transform of Eq. (2.5) and compared to the experiment. The fact that *P*(*r*) oscillates about unity at large values of *r* provides a very useful consistency check on procedures.

If sufficient knowledge of the experimental resolution is available, then the experimental structure factor *S*(*Q*) can be resolution corrected, and the reduced pair distribution function *G*(*r*) obtained via Eq. (2.5). One way this can be done is provided by a parameterization scheme given in (94), which is particularly straightforward if a single Gaussian convolution is involved. The RRDF *P*(*r*) is then obtained via Eq. (5.16), where the resolution function is removed as a multiplicative Gaussian. The form factor *α(r)* used should be such that at large distances *r*, the RRDF goes to unity as shown, for example, in the lower panel of Figure 5.4. This is a rather strong constraint. The determination of an appropriate shape factor *α(r)* is facilitated if independent data is available via microstructural studies, small-angle scattering, etc. If there is a distribution of shape factors, due to differences in the sizes and shapes of the nanomaterials, then an ensemble averaged *α(r)* can be used (127) because *G*(*r*) is linear in *α(r)* from Eq. (5.14). It should be noted that all this analysis assumes that there are no correlations between the orientation of the nanomaterial boundaries with that of the atomic lattice, that is, the individual nanomaterials act independently and are uncorrelated, and also that there is no matrix material between the nanomaterials. Further refinements to the theory are needed to incorporate such effects.

Results

Three amorphous silica models were built as part of a study on noncrystalline networks using a modified Wooten, Winer, and Weaire (WWW) approach (28, 33, 128): a bulk, nanofilm, and nanorod model. In the bulk model the cubic supercell is periodic in

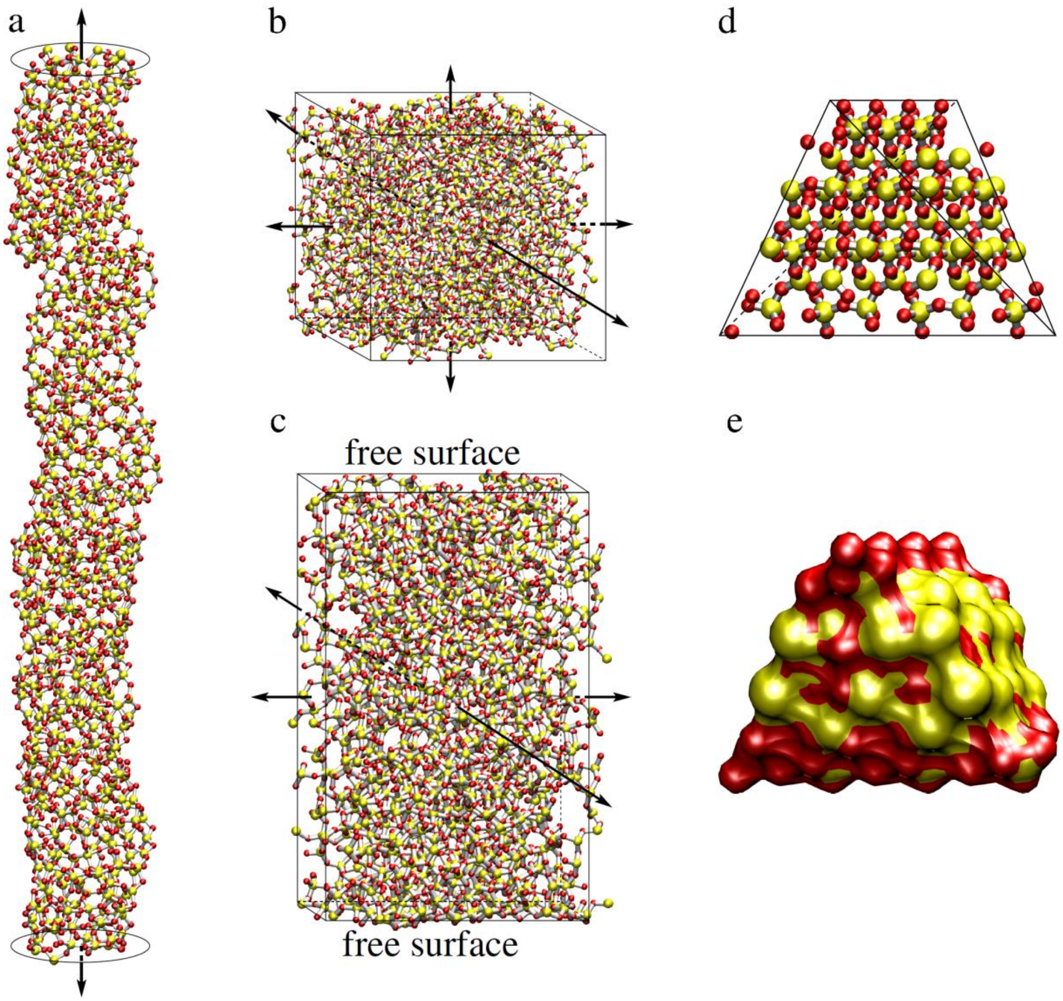


Figure 5.3 Network models of amorphous silica are shown for (a) a nanorod, (b) bulk, and (c) a nanofilm. These models are fully coordinated everywhere, including at the surface. A crystalline silica network in the shape of a nanotetrahedron is shown in both (d) and (e). In all five figures, silicon and oxygen atoms are colored yellow and red, respectively. Those surfaces subject to periodic boundary conditions are indicated by their normal vectors. In the first four figures, the supercells are outlined with black lines, while (e) shows a more accurate space-filling representation of (d).

all three dimensions (Figure 5.3b). In the nanofilm model the rectangular supercell is periodic in two dimensions while having two free surfaces along the remaining dimension (Figure 5.3c). The model represents an infinitely wide nanomaterial that has a finite thickness. In the rod model the supercell is periodic in only one dimension (Figure 5.3a) and represents an infinitely long nanomaterial with a roughly circular cross section. In all three models, each silicon and oxygen atom, including those at the surface is, respectively, bonded to four nearest-neighboring oxygen atoms and two nearest-neighboring silicon atoms.

The three amorphous silica models differ significantly from each other in shape and size. Their RDFs as defined in Eq. (5.1) differ considerably, as shown in the top panel of Figure 5.4. At large distances, the RDFs of the bulk, nanofilm, and nanorod models are proportional to r^2, r, and a constant respectively, as expected. We apply Eq. (5.12) to decouple the intrinsic atomic geometry of the three amorphous silica models from the shape and size effects. The denominator $R^u(r)$ for each model, namely, the RDF of the medium of uniform density having the same shape and size, has an analytical form for the three models. As discussed previously, $R^u(r) = 4\pi r^2 \rho_o$ for the bulk model. To the best of my knowledge, the RDF of an infinite uniform cylindrical rod has not previously been found in the concise form derived in the Appendix A. By dividing the raw RDF data of the three models displayed in the top panel of Figure 5.4 by the appropriate $R^u(r)$, we obtain the RRDFs of the three models, as shown in the bottom panel of Figure 5.4.

Independent of the shape and size of the network model, the RRDF reveals the underlying intrinsic atomic geometry with great accuracy. As shown in the bottom panel of Figure 5.4, the RRDFs of the bulk, nanofilm, and nanorod are essentially the same. This correctly represents the fact that the three models are virtually indistinguishable from each other in terms of local topology with minor differences due to surface reconstruction and distortion from the ideal geometric shape (the nanorod is not a perfect cylinder, etc.). In all three models, atoms are fully coordinated; bonding networks are amorphous; distortions in bond lengths and bond angles are within narrow ranges. The nanorod model has the widest second peak in its RRDF due to the high fraction of surface atoms that have had their bond angles distorted due to surface reconstruction.

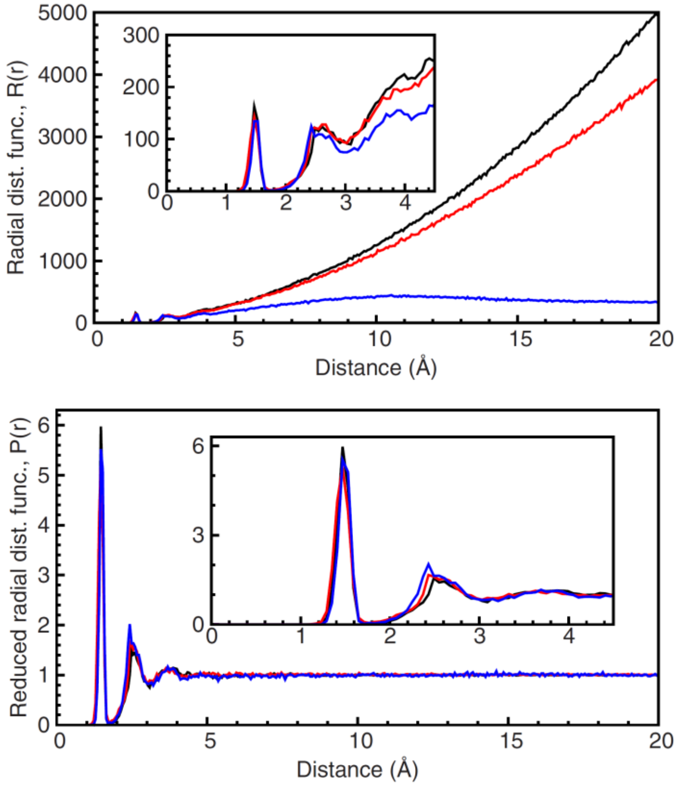


Figure 5.4 The distance distributions computed according to the RDF (top) and the RRDF (bottom) of the bulk (black), nanofilm (red), and nanorod (blue) amorphous silica network models. The inset figures show closeups at short distances 0 - 4.5 Å. The atomic numbers are used as weight factors in the computation of the RDF and RRDF.

The nanofilm and cylindrical nanorod models are two of the few fortunate cases for which the RDFs $R^u(r)$ of the corresponding uniform media have analytical expressions. For nanomaterials of most shapes, analytical expressions for $R^u(r)$ are not available. In fact it is quite challenging to derive the analytical form of the RDF of almost any geometrical shape, and to date it has not been possible for any shape whose surface contains a singularity, such as an edge or vertex. For example, even for the simplest case, the RDF of a uniform medium in the shape of a cube has not been derived in closed form, although it is easy to write in terms of a double integral that does the spherical averaging.

The computation of the RRDF according to Eq. (5.12), however, is not hindered by the lack of analytical expressions for RDFs of uniform media. No matter how complicated the shape of a nanomaterial, the RDF of the correspondingly shaped uniform

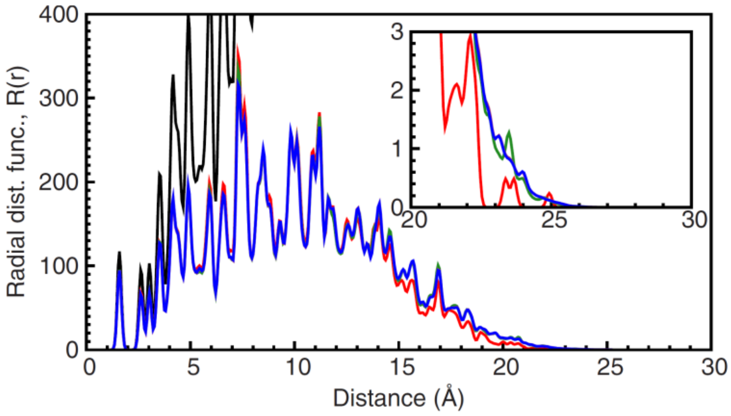


Figure 5.5 The RDF of a bulk crystalline-quartz network model (black) compared to the RDF of a single tetrahedral silica network model (red). Also included is the average RDF of one million tetrahedral silica network models with random locations but fixed orientation (green), fixed location but random orientations (purple), and random locations and orientations (blue). The purple and blue curves are indistinguishable at the resolution plotted. The inset figure shows a close-up over distances from 20 to 30 Å.

medium can be calculated numerically. As long as the definitions of the "inside" and "outside" of a material are programmable, a large number of distances can be computed between randomly generated pairs of points that lie within the boundary of the shape. The histogram of pair separations is proportional to the RDF of the uniform medium of the same shape and size as the real material. The RRDF of the nanomaterial is then computed according to Eq. (5.12).

To demonstrate this numerical procedure, the RRDF is computed for a crystalline silica network model in the shape of a regular tetrahedron (Figure 5.3d and e). The nanotetrahedron model is cut out of a bulk crystalline-quartz network model without further optimization, creating dangling bonds at the surfaces. The edge length of the tetrahedron is chosen to be 28.3 Å. The model is used in this study to exemplify the numerical calculation of an RRDF for an object bounded in all three dimensions. The RRDF is well defined up to the maximum possible separation within the object. To the best of the authors' knowledge, the analytical form of the RDF of a regular tetrahedron of uniform density has not been derived. We therefore numerically compute the RDF of a

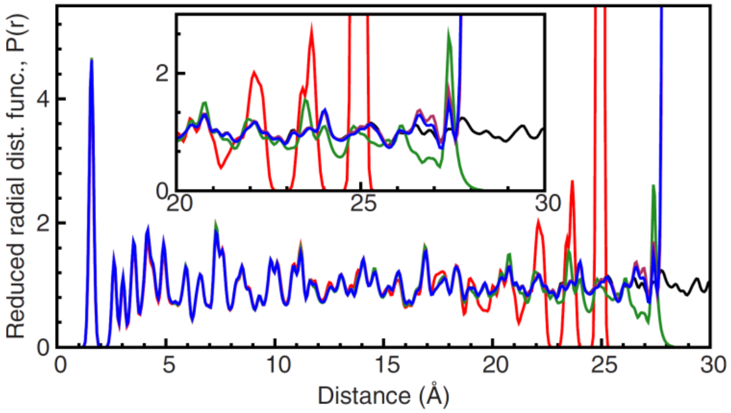


Figure 5.6 The same distribution as Figure 5.5 but plotted as RRDFs. The inset figure shows a close-up over distances from 20 to 30 Å. The largest distance within the nanotetrahedron is the edge length, 28.3 Å.

uniform tetrahedron using one billion pairs of points to achieve a smooth and well-converged distance distribution. The RRDF of the tetrahedral silica network model is then computed according to Eq. (5.12).

As discussed previously, expressing $R(r)$ as $R_b(r)\alpha(r)$ is exact only when $R(r)$ is the RDF averaged over nanomaterials representing cuts in all possible locations and orientations with respect to the bulk material, as shown in Figure 5.2. If the set of nanomaterials does not represent all possible locations and rotations, the use of the shape factor through Eq. (5.12) gives only an approximation to the average RDF of the set. The robustness of this approximation is shown in Figure 5.5 and Figure 5.6 by comparing the RDF and RRDF of the bulk material with the average RDF and RRDF of several sets of tetrahedra. These sets include a single tetrahedron, tetrahedra with a single fixed orientation but all possible locations, tetrahedra with a single fixed location but all possible orientations, and tetrahedra with all possible locations and orientations.

The RRDFs of all four sets show good agreement with the RRDF of the bulk material except at distances that approach the maximum possible pair distance contained within the tetrahedral boundaries. All peaks in the RRDF represent genuine interatomic distance contributions, as the numerically determined RDF of the tetrahedron of uniform density is smooth and nonzero over the relevant distances. The deviations from unity at large distances are amplified in the RRDF relative to the RDF, as $\alpha(r)$ in the denominator of Eq. (5.12) becomes small at these distances. The small disparity between the average RRDF of the set of tetrahedra with all possible locations and orientations and the RRDF of the bulk (below 28.3 Å) is due only to the computational limitations of sampling a finite number of tetrahedra in the calculation of $R(r)$ and a finite number of pairs in the calculation of $R^u(r)$. Otherwise the agreement would be perfect, as this set represents the complete ensemble of possible tetrahedra. For the other three sets, additional deviations in peak amplitude are due to differences in the frequency that a pair of atoms of a given separation appears in the sets relative to the frequency in the complete ensemble. Some atom pairs from the bulk may be completely absent within a given set of tetrahedral despite having separations below 28.3 Å due to constraints on location and orientation. Averaging over orientation alone results in an RDF that more closely resembles the RRDF of the bulk than does the average over location, although this may not be a general result for nanomaterials of all shapes and sizes, and for materials of all atomic geometries.

Summary

In this work it has been demonstrated that a shape factor can be used to transform the RDF of finite and bulk material onto a more general function, the RRDF depending only on the intrinsic atomic geometry of the material and not on the shape and size of the nanomaterials. The RRDF will be affected by surface reconstruction and other changes, such as voids, for example, when compared to bulk material with nominally similar atomic structure. The RRDF has a baseline of unity for materials of all atomic geometries and of any shape and size, as illustrated in Figure 5.4 and Figure 5.6, and this is a particularly useful constraint on the data at large r where the oscillations in the RRDF decay. The RRDF keeps the information describing the vital atomic geometry intact so that differences between nanomaterials of various shapes and sizes due to surface relaxation and structural rearrangement can be directly observed, independent of the main size and shape effects.

CHAPTER 6: LONG-WAVELENGTH LIMIT OF THE STRUCTURE FACTOR IN AMORPHOUS MATERIALS

Introduction

Correlated density fluctuations over large length scales can be determined from the small Q limit of the static structure factor $S(Q)$, and thus can be obtained directly from diffraction experiments (15). As discussed in Chapter 5, the structure factor can be defined in terms of the real-space pair density $\rho(r)$ via the sine Fourier transform

$$\begin{aligned} Q[S(Q)-1] &= \int_0^\infty 4\pi r[\rho(r)-\rho_0]\sin Qr\,\mathrm{d}r \\ &= \int_0^\infty G(r)\sin Qr\,\mathrm{d}r \end{aligned} \tag{6.1}$$

where ρ_0 is the average density and $G(r) = 4\pi r[\rho(r)-\rho_0]$ is the reduced pair distribution function. This is also a convenient way to obtain $S(Q)$ from computer generated structural models, as $\rho(r)$ and hence $G(r)$ is rather straightforward to compute.

Of interest here is the structure factor (15) in the small Q limit $S(Q \to 0)$ describing correlated density fluctuations over large length scales. This limit has rarely been discussed in the context of amorphous modeling but which is of considerable interest. We will refer to this limit as the static structure factor, which can be measured by small angle elastic scattering (i.e. diffraction) experiments using either X-rays or neutrons, and it is of considerable interest theoretically as it contains information about how far the system is from thermal equilibrium, which will be discussed later. We note that it is $S(Q)$ in the limit as $Q \to 0$ that is of interest, and not $S(0)$ itself, as $Q = 0$ is a singular point. For brevity, this limit will henceforth be implied whenever $S(0)$ is used.

In order to obtain any kind of reliable estimate of $S(0)$ from computer generated models, it is necessary for the model to be large, and I will focus on the excellent models of amorphous silicon and vitreous silica developed by Mousseau, Barkema and Vink (30, 34). I will describe an extrapolation technique that I developed, building upon concepts from Chapter 5, which removes much of the finite size effects at low Q that would otherwise make accurate extrapolation difficult. I will also show that the method provides a quantitative measure of how large a model must be in order to make a reliable estimate of $S(0)$. I performed all analyses, wrote the majority of the resulting paper (32), and was involved in its submission.

Theory and methodology

In Chapter 5, I showed how nanomaterials sharing the same shape and size could differ from one another due to differences in the position and orientation of the boundary relative to the atoms in the bulk material from which the nanomaterials can be envisioned to have been "cut." If out of simplicity we imagine this material to be composed of a single type of scatterer (ie. a single isotope of a single element), density fluctuations can be straightforwardly associated with differences in the number density of atoms in different regions. For such a material, a non-vanishing limit $S(0)$ would therefore correspond to correlations in fluctuations of atom number density that extend over very large length scales. If we were to create many nanomaterials by taking an ensemble of cuts from the bulk (let them be spherical cuts for simplicity) and count the number of atoms in each, we would observe fluctuations due to two sources. The first is due to differences in the number of atoms very close to the boundary that are included or excluded. Slight changes in the origin of a sphere will cause a small number of surface atoms to pop in and out of the boundary, with this number being proportional to the surface area. The second source of variation is due to differences in the local density that are being sampled by cuts sampling different regions of the bulk material and scales as the volume of the nanomaterial. It is this latter source of variation, which is independent of boundary effects and generally dominates as the size of the nanomaterial is increased, that is of interest here. In the limit of very large nanomaterials, it is density fluctuations on the size scale of the system that are the main source of number variance within the ensemble of nanomaterials.

From general considerations (129), there is a sum rule relating the limit $S(0)$ to the variance in the number of atoms N within a volume V, namely

$$S(0) = [\langle N^2\rangle - \langle N\rangle^2]/\langle N\rangle \tag{6.2}$$

in the thermodynamic limit as $V \rightarrow \infty$. We demonstrate that the static structure factor in the small Q limit is small but non-zero for realistic and large enough models of amorphous silicon and vitreous silica that numerical values can be obtained with some confidence. For crystals, with no variance in the density due to their periodicity, Eq. (6.2) gives $S(0) = 0$. Note that there are no assumptions about thermal equilibrium in the derivation of Eq. (6.2) which is of purely geometrical origin (130).

If further assumptions about thermal equilibrium and ergodicity are made, there is the additional result, well known in liquid theory (129), that relates number fluctuations to the isothermal compressibility χ_T, namely

$$[\langle N^2\rangle - \langle N\rangle^2]/\langle N\rangle = \rho_0 k_{\mathrm{B}} T \chi_{\mathrm{T}} \tag{6.3}$$

This relation assumes that all the states of a system at temperature T governed by a potential are sampled according to Boltzmann statistics. Hence for liquids (and other thermodynamic, ergodic systems in thermal equilibrium), we have

$$S(0) = \rho_0 k_{\mathrm{B}} T \chi_{\mathrm{T}} \tag{6.4}$$

Eq. (6.4) is also true for multi-component systems if ρ_0 is interpreted as the atomic number density, causing $S(0)$ to become $S_{NN}(0)$, a Bhatia-Thornton structure factor (22, 131, 132) where N refers to the total number of atoms.

Background on amorphous materials

Amorphous silicon is perhaps the furthest from equilibrium of all amorphous materials. This is because it is highly strained, with most of the strain being taken up by deviations of the bond angles from their ideal tetrahedral value of 109.5°. Each silicon atom has 3 degrees of freedom. The important terms in the potential are the bond stretching and angle bending forces around each atom. There are 4 covalent bonds at each silicon atom, each of which is shared, giving a net of 2 bond stretching constraints per atom. Of the 6 angles at each silicon atom, 5 are independent, giving a total of 7 constraints per atom. As there are considerably more constraints than degrees of freedom, the network is highly over-constrained (133). In thermal equilibrium, silicon cycles between crystalline solid and liquid forms. There is no glass transition. However, amorphous silicon can be prepared by various techniques involving very fast cooling and provides an extreme example of a non-equilibrium state.

Vitreous silica is a bulk glass, which contains very little strain, as can be seen as follows. The important constraints are the bond stretching and angle bending forces associated with the silicon atoms, as in amorphous silicon. The angular forces at the oxygen ions (β) are weak (134). The total number of constraints per SiO_2 unit is 4 Si-O bond stretching constraints plus 5 angular forces at the Si giving a total of 9 constraints. However, the number of degrees of freedom per SiO_2 is also 9 (3 per atom). The system is therefore isostatic and not over-constrained (133). Thus, the strong Si-O bond stretching and 0-Si-O angle bending forces are well accommodated (although the weaker angular distribution at the oxygen atom less so), so that vitreous silica is closer to thermal equilibrium than amorphous silicon, although not close enough that Eq. (6.4) can be used. However, Eq. (6.4) is much more likely to be obeyed *if* the fixative temperature T_f at which the glass was formed is used instead of T (including for the compressibility). A much slower decrease in $S(0)$ is observed as the temperature is decreased below T_f due to the freezing out of thermal vibrations about a fixed topology, as shown in the extensive and informative experiments of Levelut *et. al* (23, 135, 136).

Computer models

In this study, I analyzed two high-quality periodic computer generated models of amorphous silicon. The first is a small model with 4096 atoms (henceforth called the 4096 atom model) (33), built within a cubic super-cell with sides of length L = 43.42Å. The average bond length is a = 2.35Å, equal to the known value for crystalline silicon, and the model has the same density as crystalline silicon, which is about right for structurally good samples of amorphous silicon containing few voids, defects, etc. The network was constructed using the Wooten, Winer, and Weaire (WWW) technique (28, 33), based on locally restructuring the topology of crystalline silicon, while keeping the number of atoms and covalent bonds fixed, until the ring statistics settle down and there are no Bragg peaks apparent in the diffraction pattern.

The second model contains 100,000 atoms (referred to as the 100k model) within a cubic super-cell of sides L = 124.05Å, with an average bond length of a = 2.31Å, and was built using a modified WWW technique (34) based on previous work by Barkema and Mousseau (29). We note that the models of Mousseau and Barkema have the narrowest angular variance (~9°) at the silicon atoms ever achieved in a non-crystalline tetrahedral network, and they also avoid the issue of possible crystal memory effects in WWW type models, as they use a non-crystalline atomic arrangement initially. The 100k model, like other models built by Barkema and Mousseau (29), has a density ~5% above that of crystalline silicon, which is too large for amorphous silicon. The reason why this model has a higher density, while being excellent in other aspects is not entirely clear, but it may be necessary to let the angular variance increase back up to ~11° in order to get the experimental density of amorphous silicon. The correlation between this angular spread and the density needs further study. This difference should not affect the limit $S(0)$ to first order, as an isotropic compression or expansion of the whole structure leaves the relative number fluctuations invariant in the thermodynamic limit.

A very large model of vitreous silica (300k model) has been produced by the same group (30) by first decorating the 100k amorphous silicon model with an oxygen ion between each silicon ion and relaxing appropriately. The covalent bond network was then modified using the WWW technique. With only a few exceptions, all silicon atoms maintain only oxygen atoms as covalently bonded neighbours and vice versa. An important difference between the 100k amorphous silicon and the 300k vitreous silica models is that by effectively changing the fundamental unit from a silicon atom to a corner sharing SiO_2 tetrahedron, the system is no longer overconstrained but instead isostatic (133), a point that was discussed in the Amorphous materials section. One might expect the greater number of degrees of freedom and the lower internal stress of the vitreous silica model to affect the static structure factor, as vitreous silica is closer than amorphous silicon to thermal equilibrium. We will return to this point later.

Methods for calculating the structure factor in the limit $\mathbf{Q} \rightarrow \mathbf{0}$

Directly from the set of pair distances

The static structure factor $\mathrm{S(Q)}$ can be calculated in a number of ways, some of which are more useful (i.e. smoother) than others when extrapolating $\mathrm{Q} \rightarrow 0$. We focus first on amorphous silicon, a material with a single atomic species. The structure factor can be computed directly from the set of atom coordinates by taking the spherical average of

$$S(\boldsymbol{Q}) = 1 + \frac{1}{N\langle f\rangle^2} \sum_{i\neq j} f_j^* f_i exp(i\boldsymbol{Q} \cdot \boldsymbol{r}_{ij}) \tag{6.5}$$

where f_i is the scattering factor of atom i. A spherical average yields

$$S(Q) = 1 + \frac{1}{N\langle f\rangle^2}\sum_{i\neq j} f_j^* f_i \frac{\sin Qr_{ij}}{Qr_{ij}} \tag{6.6}$$

where the sum $i \neq j$ goes over all pairs of atoms (excluding the self terms) in the periodic cubic super-cell of size L, and is evaluated at $Q_{lmn} = \frac{2\pi}{L}\sqrt{l^2 + m^2 + n^2}$ where *l*, *m*, and *n* are integers. For a finite model with periodic boundary conditions, this means that it does not matter if the distances r_{ij} are measured within the unit super-cell or across unit super-cells, as long as all $N(N-1)$ terms are computed in Eq. (6.6).

This computational approach using Eq. (6.6) suffers from the problem that there are finite size effects at small Q, even with periodic boundary conditions, creating a peak in $\mathrm{S(Q)}$ at the origin of finite width $\sim 1/\mathrm{L}$ and amplitude N. The peak at small Q, studied by small angle X-ray or neutron scattering, is given by the convolution of the delta function that would exist at the origin if the model were infinite, with a function related to the shape of the box in which the model exists, as discussed in Chapter 5. This problem at small Q could in principle be alleviated by subtracting the peak at the origin due to the finite size of the model (or sample), but the form of the peak is only known algebraically for a limited set of shapes (137) which do not include the cube for which a double angular integration is needed. The numerical subtraction of two large numbers of $O(N)$ would lead to errors of $O(1)$, which is the order of the answer required. A better approach to finding the form of $\mathrm{S(Q)}$ in a form suitable for extrapolation to small Q is described below.

Fourier transform approach

As a way to circumvent issues associated with the finite size of the sample that affect small Q, the structure factor $\mathrm{S(Q)}$ can be obtained from $\mathrm{G(r)}$ via the sine Fourier transform given in Eq. (6.1). It appears from the form of Eq. (6.1) as though the limit $\mathrm{S(Q \to 0)}$ depends upon the sine transform of $\mathrm{G(r)}$ alone, and thus the behavior of $\mathrm{G(r)}$

at large r does not contribute much to the limit $\mathrm{S}(0)$ [see Figure 6.2 for an example of $\mathrm{G}(\mathrm{r})$]. This can be shown to be false by expanding Eq. (6.1) in powers of Q and keeping only the lowest order terms that would dominate in the small Q limit. To the lowest order in Q

$$S(0) = 1 + \int_{0}^{\infty} 4\pi r^2[\rho(r) - \rho_0]dr = 1 + \int_{0}^{\infty} rG(r)\,dr \tag{6.7}$$

which depends on the integral of $\mathrm{rG}(\mathrm{r})$, not $\mathrm{G}(\mathrm{r})$. This factor of r increases the sensitivity of the limit $\mathrm{S}(0)$ to the details of the decay in $\mathrm{G}(\mathrm{r})$ at large distances. Oscillations in $\mathrm{G}(\mathrm{r})$ associated with a single reference atom are known to persist out to large distances (14) and are a serious concern when computing $S(0)$ from a model. Even upon averaging over all reference atoms, the use of Eq. (6.7) to find the limit $\mathrm{S}(0)$ suffers from poor convergence at small Q, as $\mathrm{rG}(\mathrm{r})$ amplifies the ripples that persist at large distances because of the finite nature of the model, although it is superior to using Eq. (6.6).

At this point, one might wonder why we do not use the tools from Chapter 5 for a cubic cell to reduce finite size effects from $\mathrm{G}(\mathrm{r})$ to compute $S(0)$. The reason is that the periodic boundary conditions of the model causes there to be no boundaries beyond which atoms do not exist. While a direct application of tools from Chapter 5 to the cubic models is not appropriate, I will now show that they can be applied as an excellent tool for extrapolation.

Sampling volume method

Quite generally, even in the absence of thermal equilibrium, the small Q limit $S(0)$ is related in the thermodynamic limit to number (or density) fluctuations within sub-regions of volume V according to Eq. (6.2). As we only have models of finite size, even with periodic boundary conditions it is not possible to determine the limit directly and it is necessary to extrapolate to the $N \rightarrow \infty$ limit as best we can. The approach of extrapolating $S(Q)$ as $Q \rightarrow 0$ suffers from finite size effects that cause oscillations about the ideal $S(Q)$ which would be obtained for an infinite model. It is difficult to disentangle the finite size effects from the underlying ideal $S(Q)$, making accurate extrapolation always challenging.

A more accurate determination of $S(0)$ can be achieved through Eq. (6.2). The equality states that the relative variance in the number of atoms within an ensemble of randomly placed, bounded, convex volumes (130) is equal to $S(0)$ in the limit that the sampling volume goes to infinity. For a finite sampling volume of fixed shape, the variance in the number of atoms within the enclosed volume, which samples all possible positions and orientations equally, can be divided into terms that scale as the volume, those that scale as surface area, and those with lower order dependencies on the length scale of the enclosed volume (130). If R describes such a sampling length scale, then the relative variance, which divides the variance by the average number of atoms within the sampling volume, can be expressed as the sum of a volume term of order R^0, a surface term of order R^{-1}, and lower order terms.

Atomic structures for which the number variance does not depend on volume are called *hyperuniform*, examples of which are materials with a periodic lattice, as their unit cells have well defined volume and density. The number variance for such systems is related to the Gauss circle problem (14, 130, 138). The static structure factor for crystals is zero, as the structure factor $S(Q)$ is zero for all values of Q smaller than that associated with the first Bragg peak, leading to the result $S(0) = 0$. Also the relative variance of the number fluctuations is clearly zero on length scales that are much greater than the size of the unit super-cell. This result is only strictly true in the absence of diffuse scattering at a temperature of absolute zero. Note it is important to take the limit $Q \rightarrow 0$ so as to avoid the peak at the origin. For all periodic models at large enough length scales (corresponding to small enough Q), the static structure factor will go to zero as the static limit is approached due to the hyperuniformity associated with the crystallinity. Nevertheless we can get meaningful results if we restrict ourselves to distances less than the size of the super-cell, and Q values that are small ($\sim 1/L$ where L is the linear dimension of the supercell) but not too small.

For non-crystalline systems, like amorphous silicon and vitreous silica, we will show that determining the relative variance of $N(R)$ for various sampling radii R and extrapolating the result as $R \to \infty$ provides a much more precise method of extracting the limit $S(Q \to 0)$ from a finite model. Indeed it is the optimal procedure. The relative variance has been thoroughly described by Torquato and Stillinger (130) and Eq. (58) from their paper can be written for spherical sampling volumes as

$$\frac{\langle N(R)^2\rangle - \langle N(R)\rangle^2}{\langle N(R)\rangle} = 1 - \rho_0 \frac{4\pi}{3} R^3 + \frac{1}{n} \sum_{i \neq j}^{n} \alpha\left(r_{ij}; R\right) \tag{6.8}$$

where n is the number of atoms in the model, and the function $\alpha\left(r_{ij}; R\right)$ is the fractional intersection volume of two (continuum) spheres, with radii R and centers separated by r_{ij}. The function $\alpha\left(r_{ij}; R\right)$ is proportional to the probability of two points, separated by r_{ij}, both being contained within a randomly placed sphere of radius R, and has a form given by Torquato and Stillinger in Eq. (A11) of their paper as

$$\alpha(r; R) = \begin{cases} \left(1 - \frac{3}{4R} r + \frac{1}{16R^3} r^3\right) = \left(1 - \frac{r}{2R}\right)^2 \left(1 + \frac{r}{4R}\right) & r \leq 2R \\ 0 & r > 2R \end{cases} \tag{6.9}$$

and zero if $r > 2R$. This is just the shape function that is widely used in describing scattering from spherical micro-crystallites (137), but is used in quite a different context here, as it is merely an arbitrary but convenient sampling volume. Using the real-space pair density $\rho(r)$ to convert the sum in Eq. (6.8) into an integral, we can write

$$\frac{\langle N(R)^2\rangle - \langle N(R)\rangle^2}{\langle N(R)\rangle} = 1 - \rho_0 \frac{4\pi}{3} R^3 + \int_0^\infty 4\pi r^2\, \rho(r)\alpha(r;R)\mathrm{d}r \tag{6.10}$$

Using the identity

$$\frac{\langle N(R)^2\rangle - \langle N(R)\rangle^2}{\langle N(R)\rangle} = 1 - \rho_0 \frac{4\pi}{3} R^3 + \int_0^\infty 4\pi r^2\, \rho(r)\alpha(r;R)\mathrm{d}r \tag{6.11}$$

we obtain the following result

$$\frac{\langle N(R)^2\rangle - \langle N(R)\rangle^2}{\langle N(R)\rangle} = 1 + \int_0^\infty 4\pi r^2\, [\rho(r) - \rho_0]\alpha(r;R)\mathrm{d}r \tag{6.12}$$

which can be conveniently re-written as

$$\frac{\langle N(R)^2\rangle - \langle N(R)\rangle^2}{\langle N(R)\rangle} = 1 + \int_0^\infty rG(r)\alpha(r;R)\, \mathrm{d}r \tag{6.13}$$

Comparing Eq. (6.13) to Eq. (6.7), they are clearly equivalent as $R \to \infty$ and $Q \to 0$, as the integrand in Eq. (6.13) contains $\alpha(r;R)$ which tends to unity for all r as $R \to \infty$. The presence of $\alpha(r;R)$ arises due to the finite nature of the sampling volume, and acts as a natural convergence factor for the integral in Eq. (6.7). Notice that the relative variance of $N(R)$ is *not* related to $S(Q)$ except in the limit as both $R \to \infty$ and $Q \to 0$. The sampling volume factor $\alpha(r;R)$ for a sphere can be written as a Taylor expansion in integer powers of $1/R$, allowing the relative variance to be written in the form

$$\frac{\langle N(R)^2\rangle - \langle N(R)\rangle^2}{\langle N(R)\rangle} = a + b/R + O(1/R^2) \tag{6.14}$$

where $a = S(0)$ describes the volume dependence, and b describes the surface dependence associated with the sampling volume. In conjunction with Eq. (6.2), Eq. (6.14) is therefore a simple but exact relation that allows one to obtain the static structure factor $S(Q \to 0)$ from a large model structure, contained within a super-cell that periodically repeats, and avoids problems associated with extrapolating an oscillating function. We have found this to be the best possible procedure.

Results

Amorphous silicon models

One major focus of this work is to determine the limit $S(0)$ for amorphous silicon from computer models which serves as a prediction for this important material. As discussed earlier, there is more than one way to find the limit $S(0)$, and we will explain the numerical results obtained with all of them here.

The first approach is shown in Figure 6.1, where we show the most direct calculation of $S(Q)$ using Eq. (6.6) at the points $Q_{lmn} = \frac{2\pi}{L}\sqrt{l^2 + m^2 + n^2}$ determined by the super-cell. While this gives a good overall description of $S(Q)$, it is very limited at small Q and extrapolation or analytic continuation to $Q = 0$ is not possible, even for the much larger 100k model. This is because the finite size oscillations are too severe. Note that the higher density of the 100k model leads to a shift of the peaks to slightly larger Q values. Note also that the structure factor approaches unity at large Q as it must, which sets the scale for comparison for the limit $S(0)$. No harmonic phonons (or zero point motion) were added to any of the results in this work. The inclusion of phonons would

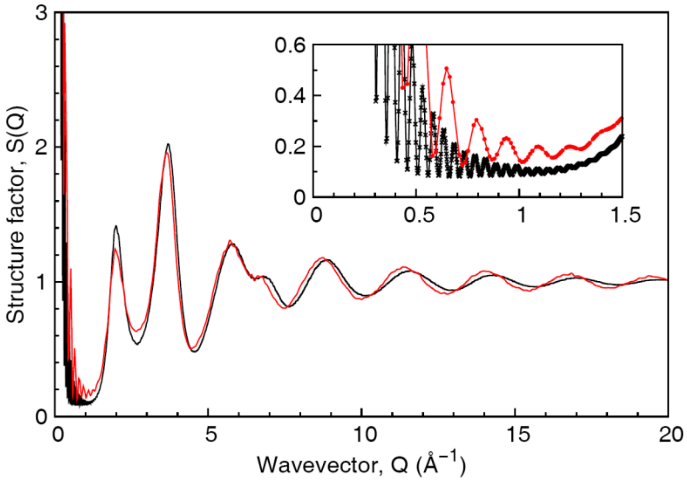


Figure 6.1 The structure factor for amorphous silicon is calculated directly using Eq. (6.6) at the super-cell values Q_{lmn}, shown in the inset as red circles for the 4096 atom model and black crosses for the 100k model.

have the effect of adding a term that goes as Q^2 at small Q, but this would vanish as $\mathrm{Q} \rightarrow 0$.

The second method is the Fourier transform method in which $S(Q)$ is determined from the sine transform using Eq. (6.1) with $G(r)$ input from the model. Both models of amorphous silicon, with 4096 and 100k atoms are used in Figure 6.2 which shows the distribution $G(r) = 4\pi r[\rho(r) - \rho_0]$. Notice the differences in the two silicon models. The difference of 5% in the densities is apparent at small r, where $G(r) = -4\pi r\rho_0$, and by the small shift in the peak positions. For comparison, the average separation of bonded silicon atoms determined from the first peak is 2.35Å in the 4096 atom model but only 2.31Å in the 100k model. An isotropic contraction of the whole system does not affect the limit $S(0)$, so to first order, the overly dense 100k model should give appropriate values in the limit, as there is no length metric in the limit $Q \rightarrow 0$.

The structure factor can be found by applying Eq. (6.1) using $G(r)$ for each model. Only the structure factor of the 100k model is shown in Figure 6.3, where, even here, the difficulty of trying to extrapolate to $Q = 0$ is again apparent, although the

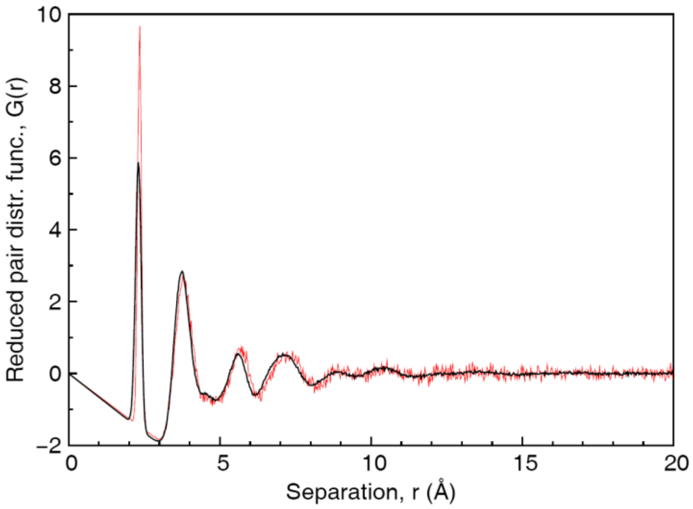


Figure 6.2 The pair distribution function $G(r)$ for amorphous silicon for the 4096 atom model (rough red line) and the 100k model (smooth black line).

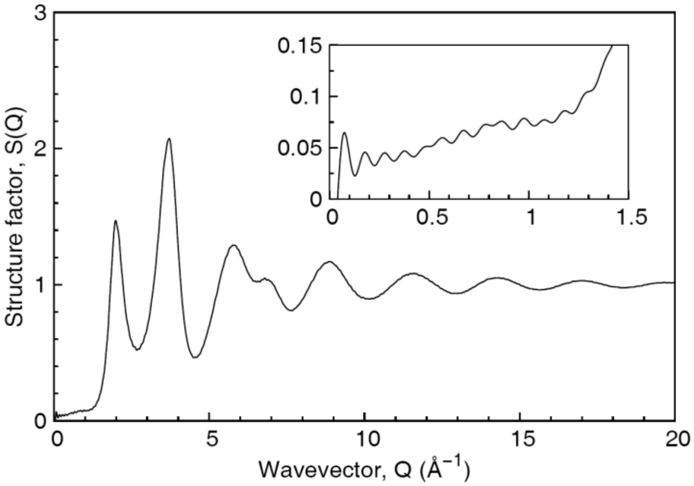


Figure 6.3 The structure factor $S(Q)$ for amorphous silicon obtained from Eq. (6.1) for the 100k model. The insert shows the small Q region expanded.

situation is improved somewhat from the direct method shown in Figure 6.1. From the inset of Figure 6.3 that displays $S(Q)$ at small Q, the structure factor of the 100k model still displays significant oscillations due to finite size effects. Of course these oscillations are even more pronounced for the 4096 atom model, which is not shown. These effects arise from the truncation of $G(r)$ beyond $L/2$ (half the width of the cubic super-cell), beyond which $G(r)$ is almost but not quite zero. The source of the oscillations is apparent from their wavelength of $2\pi/(L/2)$. A very approximate limit of $S(0) \cong 0.03$ can be extrapolated by eye for the 100k model from Figure 6.3, through the ripples in the insert, but the uncertainty is almost as large as the value itself. For the smaller 4096 atom model, the oscillations are even larger, making any attempt to extrapolate $S(Q)$ quite hopeless. Smoothing techniques can be used to attenuate the oscillations, but are not very convincing. There is a better approach.

An alternative to the Fourier transform approach involves finding the relative variance within finite sampling volumes of increasing size (but identical shape- we have used spheres) and extrapolating to the thermodynamic limit. This has the great operational advantage of avoiding oscillations. The relative variance in the number of

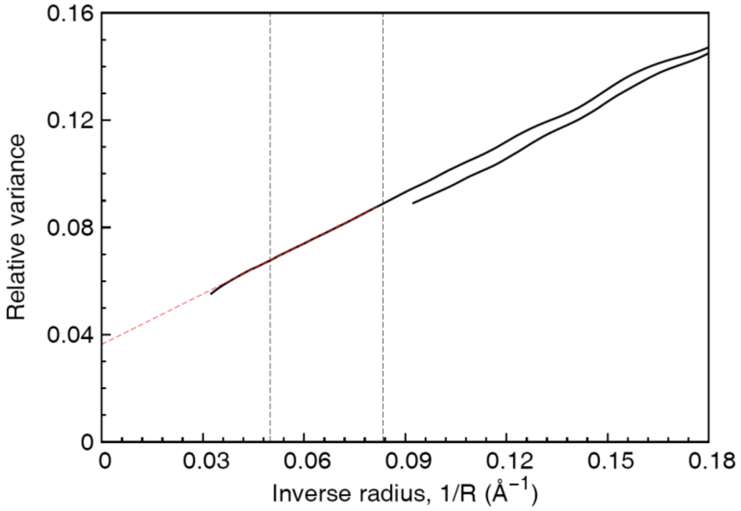


Figure 6.4 The relative variance in the number fluctuations in amorphous silicon is computed within spheres of various radii R using the sampling volume method. The extrapolated value of S(0), which is just the limit of the relative variance for small 1/R, is given by $S(Q \rightarrow 0) = 0.035 \pm 0.001$ for the 100k mode using Eq. (6.2). The vertical dashed lines indicate the range over which the linear fit was performed. It can be seen that the smallest value of 1/R for the 4096 atom model is larger than the upper limit of the range over which the relative variance is linear and therefore a reliable extrapolation cannot be made.

atoms within spheres of different radii is plotted in Figure 6.4 for both silicon models. The distribution $G(r)$ can only be computed safely out to $r = L/2$ due to the periodic nature of the model. As the sampling volume factor $\alpha(r; R)$ for a sphere is non-zero out to $r = 2R$, the relative variance should only be computed using Eq. (6.13) out to $R = L/4$, causing the curve for the 4096 model to terminate at a larger value of $1/R = 4/L$ than that for the 100k model. The relative variance for the 100k model shows a definite linear region within the interval $12\text{Å} < R < 20\text{Å}$ or $0.05\text{Å}^{-1} < 1/R < 0.083\text{Å}^{-1}$. From Figure 6.2, the lower limit $R_{min} = 12\text{Å}$ corresponds to the distance at which strong correlations in atom pair separations all but vanish. The upper limit $R_{max} = 20\text{Å}$ corresponds to the radius at which the relative variance within the spherical volumes begins to deviate noticeably from its linear behaviour due to the finite size of the periodic model. The maximum possible radius given the sampling volume argument above is $L/4 = 31\text{Å}$, so $20\text{Å} \approx L/6$ represents a conservative and safe cut-off. If the largest sampling volume for which the relative variance maintains linear behaviour is assumed to be determined by the ratio of the width of the sampling volume to the width of the model, we would expect the linear region to be entirely absent for the 4096 atom model, as $R_{max} = L/6 \cong 7.2\text{Å}$ is less than the lower limit $R_{min} = 12\text{Å}$. Indeed this is what is observed in Figure 6.4 for the 4096 atom model, as the oscillations at large values of $1/R$ are still significant by the time the lower limit of $4/L$ is reached. These observations would imply that there is a critical size that a model should be in order for a good extrapolation to $S(0)$ in the thermodynamic limit to be possible. At a bare minimum, the width of the box (or for general shapes, the minimum diameter) should be greater than six times the distance over which strong correlations in atom pair separations persist in order for a linear fitting window to exist. For amorphous silicon, this bare minimum would correspond to a periodic super-cell with sides of length 70Å containing ~18,000 atoms. To get a window of decent size for the linear fit, it would be very difficult to work with a model of less than ~50,000 atoms. Triple this amount, ~150,000 atoms, is needed for vitreous silica.

The value of the limit $S(0)$ found from linear extrapolation over the linear region of the 100k model is $S(0) = 0.035 \pm 0.001$, where the uncertainty represents the spread in the values of the intercept that result for different choices of the fitting interval. Applying the same extrapolation technique for all Q values, according to (15), results in a structure factor similar to that of Figure 6.3 but without the oscillations due to the finite size of the model, as shown in Figure 6.5. The large Q values are unaffected by using the convergence factor in (15), but there is a significant effect at small values of Q. In order to compare with experiment, the Q values of the structure factor for the 100k model shown in Figure 6.5 were decreased to 0.985 of their original

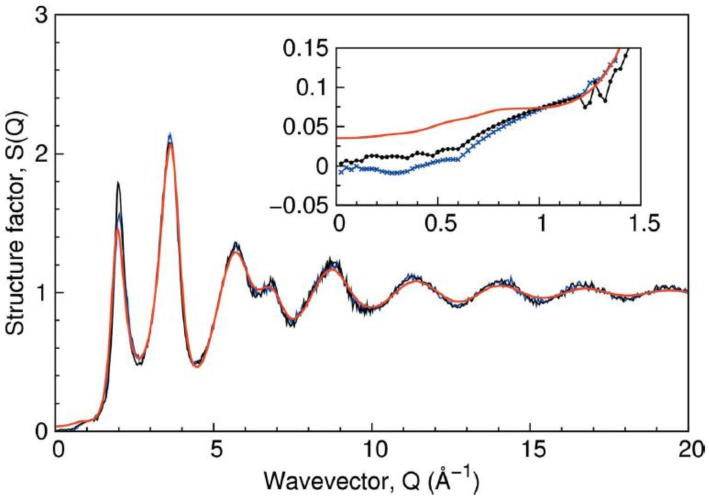


Figure 6.5 Comparison of the structure factor for amorphous silicon (as implanted, blue crosses, and annealed, black circles) experimentally determined by Laaziri *et al*. (1999b) with the structure factor for the 100k model (red curve, no points), rescaled to make the density of the model match that of crystalline silicon and hence void-free amorphous silicon.

value to account for the fact that the model has a higher density than that of crystalline silicon and hence void-free amorphous silicon. The rescaled structure factor shows good agreement with the experimental results of Laaziri *et al.* (35) (whom we thank for providing original data points used in Figure 6.5) except for differences in the low Q region and in the amplitude of the oscillations. This requires further modeling to determine the effects of the angular spread at the silicon atom, ring statistics etc. on the structure factor.

Vitreous silica model

In general for polyatomic systems, it is useful to define partial pair distribution functions (PPDFs) and their corresponding Faber-Ziman partial structure factors (139). For vitreous SiO_2, the three PPDFs are $G_{SiSi}(r)$, $G_{OO}(r)$, and $G_{SiO}(r)$, where the PPDFs are computed using the subsets of atom types specified by their respective subscripts. Vitreous silica can be viewed as a network of corner sharing tetrahedral SiO_2 subunits that are very rigid compared to the flexibility of the Si-O-Si angle at their shared corners.

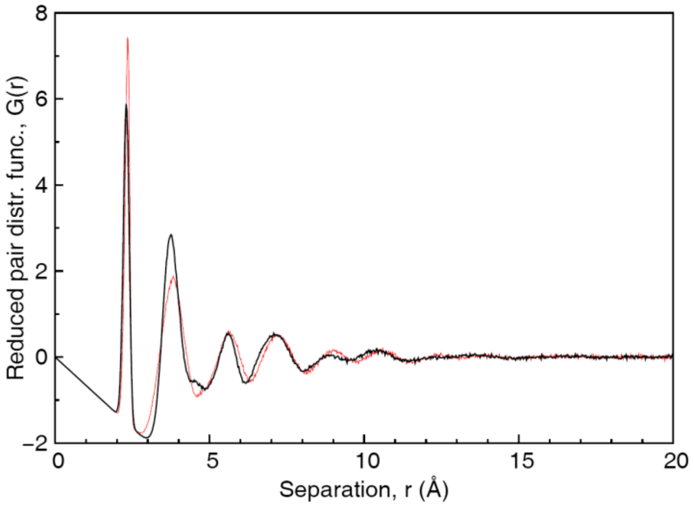


Figure 6.6 The pair distribution function $\mathrm{G_{SiSi}(r)}$ for the 300k vitreous silica model (thin red) rescaled by a length factor of 1/1.33 and superimposed on the same distribution from the 100k amorphous silicon model (thick black, as in Figure 6.2).

To a first approximation, the density fluctuations of the 300k vitreous silica model produced by Vink and Barkema (30) captured by $G_{SiSi}(r)$ can be compared to $G(r)$ of the 100k model of amorphous silicon produced by the same group. Figure 6.4 displays the PPDF $G_{SiSi}(r)$ superimposed on $G(r)$ from the 100k silicon model, where the silicon distances in the 300k model have been decreased by a factor of 1.33 to make the silicon atom densities the same. The two distributions are not the same, nor should they be, but are quite surprisingly close. Using the *rescaled* PPDF $G_{SiSi}(r)$ of vitreous silica as an example of a highly distorted model for amorphous silicon leads to $S_{SiSi}(0) = 0.039 \pm 0.001$ by applying the volume sampling method, and is remarkably close to the value of $S_{SiSi}(0) = 0.035 \pm 0.001$ for the 100k model obtained in the previous section. Thus it appears that the fourfold tetrahedral coordination of the amorphous network is the most important factor in determining $S(0)$.

The three associated Faber-Ziman partial structure factors $S'_{SiSi}(Q)$, $S'_{OO}(Q)$, and $S'_{SiO}(Q)$ can be found from their respective PPDFs through the sine Fourier transform

$$Q\left[S'_{\alpha\beta}(Q)-1\right]=\rho_o\int_{0}^{\infty}4\pi r\left[g_{\alpha\beta}(r)-1\right]\sin Qr\,\mathrm{d}r \tag{6.15}$$

where ρ_o is the number density associated with *all* the atoms in the system, $g(r)$ is the reduced pair distribution function, a scaled version of $\rho(r)$ such that it oscillates about unity at large r, and α and β define the atom pairs used in the distribution function. This definition of the partial structure factor differs from the intuitive definition that would be obtained if atoms of each type were isolated. This more intuitive definition (for which we use unprimed notation) is represented by partial structure factors of the form

$$Q[S_{\alpha\alpha}(Q)-1]=\rho_\alpha\int_{0}^{\infty}4\pi r[g_{\alpha\alpha}(r)-1]\sin Qr\,\mathrm{d}r \tag{6.16}$$

where $\rho_\alpha = c_\alpha\rho_o$, c_α being the fraction of atoms of type α. These two distributions are simply related by

$$S_{\alpha\alpha}(Q)-1=(\rho_\alpha/\rho_o)[S'_{\alpha\alpha}(Q)-1]=c_\alpha[S'_{\alpha\alpha}(Q)-1] \tag{6.17}$$

Three Bhatia-Thornton structure factors (22, 131, 132, 140) that describe correlations between atom number and concentration can be defined in terms of the three $S'_{\alpha\beta}(Q)$ according to

$$\begin{gathered}S'_{NN}(Q)=c_{Si}^2S'_{SiSi}(Q)+c_O^2S'_{OO}(Q)+2c_{Si}c_OS'_{SiO}(Q)\\ S'_{CC}(Q)=\mathrm{c_{Si}c_O}\left[1+\mathrm{c_{Si}c_O}\left(S'_{SiSi}(Q)+S'_{OO}(Q)-2S'_{SiO}(Q)\right)\right]\\ S'_{NC}(Q)=c_{Si}c_O[c_{Si}(S'_{SiSi}(Q)-S'_{SiO}(Q))-c_O(S'_{OO}(Q)-S'_{SiO}(Q))]\end{gathered} \tag{6.18}$$

Three of the six unknowns in Eq. (6.18) can be found in the limit as $Q\to 0$ by applying the sampling volume method [Eq. (6.13)] to $G_{SiSi}(r)$, $G_{OO}(r)$, and $G_{NN}(r)$ (avoiding terms of type $G_{\alpha\beta}(r)$ with $\alpha\neq\beta$). Using the same fitting interval as that for the silicon model results in the limiting values $S_{SiSi}(0)=0.039\pm0.001$, $S_{OO}(0)=0.078\pm0.002$, and $S_{NN}(0)=0.116\pm0.003$, as shown in Figure 6.6. Inserting these values into the three Bhatia-Thornton relations (18) and solving for the remaining three unknowns, one

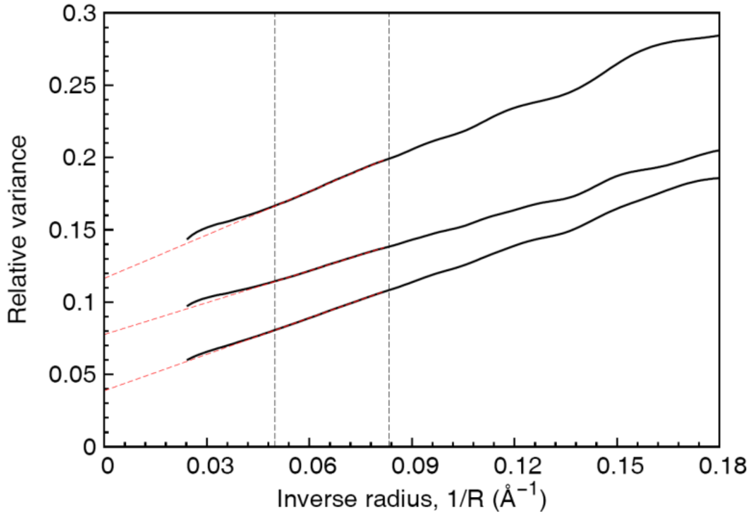


Figure 6.7 The relative variance of the number fluctuations in vitreous silica within sampling spheres of radii R. The variance is computed using the sampling volume method and plotted against 1/R. The extrapolated values of $S(0)$, which are just the limits of the relative variances of the number fluctuations for small 1/R, are given by $\mathrm{S_{SiSi}}(0) = 0.039 \pm 0.001$, $\mathrm{S_{OO}}(0) = 0.078 \pm 0.002$, and $\mathrm{S_{NN}}(0) = 0.116 \pm 0.003$. The position and size of the sampling window is determined in a similar way to that described for amorphous silicon.

finds $S_{SiO}(0) = 1.116$, $S_{CC}(0) = -1.5 \times 10^{-5}$, and $S_{NC}(0) = 0.96 \times 10^{-5}$. Within the uncertainty of the extrapolation, and remembering that there are ~10^5 atoms in the model, the limits of the last two Bhatia-Thornton structure factors are consistent with zero, i.e. $S_{CC}(0) = S_{NC}(0) = 0$. This reflects the fact that the chemical disorder is virtually zero, as only several out of the 100,000 silicon atoms in the model are bonded to another silicon atom instead of to an oxygen atom.

If the two quantities $S_{CC}(0)$ and $S_{NC}(0)$ are exactly zero, which we will take to be true from now on, the relationship between the limiting values of the other structure factors simplify greatly, and can all be expressed in terms of a single structure factor rather than the original three. Eqs. (6.18) can be rewritten as

$$\begin{gathered} S'_{SiSi}(0) = S'_{NN}(0) - \frac{c_O}{c_{Si}} \\ S'_{OO}(0) = S'_{NN}(0) - \frac{c_{Si}}{c_O} \\ S'_{SiO}(0) = S'_{NN}(0) + 1 \end{gathered} \tag{6.19}$$

From Eq. (6.17), one can write down the relation

$$\begin{gathered} S^{'}_{SiSi}(Q) = \frac{1}{c_{Si}} S_{SiSi}(Q) - \frac{c_O}{c_{Si}} \\ S^{'}_{OO}(Q) = \frac{1}{c_O} S_{OO}(Q) - \frac{c_{Si}}{c_O} \end{gathered} \tag{6.20}$$

In the thermodynamic limit, the previous six equations relate the limiting values of the seven structure factors, and thus there is only one independent quantity. The limiting value of the other structure factors that one would find if each atom type *was taken in isolation* can be expressed along with the Bhatia-Thornton number correlation as

$$\begin{gathered} S_{SiSi}(0) = c_{Si} S_{NN}(0) \\ S_{OO}(0) = c_O S_{NN}(0) \end{gathered} \tag{6.21}$$

The scaling factors that exist between these three values when there is no chemical disorder in the system explains why the values found from the sampling volume method follow a 1:2:3 ratio [$S_{SiSi}(0) = 0.039 \pm 0.001$, $S_{OO}(0) = 0.078 \pm 0.002$, and $S_{NN}(0) = 0.116 \pm 0.003$], as $c_{Si} = 1/3$ and $c_{Si} = 2/3$. Notice that this scaling is only present as $Q \rightarrow 0$ and of course is not true at a general Q. All the analysis of the 300k vitreous silica model can therefore be summarized in a single number by there being virtually no chemical disorder and $S_{NN}(0) = 0.116 \pm 0.003$.

The expression for the limiting value of the differential scattering cross section per atom obtained from scattering experiments also simplifies if no chemical disorder is present. The general form of differential cross section per atom (22, 132, 140), namely

$$\sum_{\alpha\beta} c_\alpha c_\beta b_\alpha b_\beta \left[S'_{\alpha\beta}(Q) - 1\right] + \sum_{\alpha} c_\alpha b_\alpha^2 \tag{6.22}$$

where b_α is the scattering length of atoms of type α, can be simplified in the limit $Q \to 0$ by writing the three partial structure factors $S'_{SiSi}(0)$, $S'_{OO}(0)$, and $S'_{SiO}(0)$ in expression (6.22) in terms of $S_{NN}(0)$ using Eqs. (6.20) and (6.21). Performing the substitutions, one finds that the differential cross section per atom simplifies to

$$[c_{Si}b_{Si} + c_{Si}b_{Si}]^2 S(0) \tag{6.23}$$

Eq. (6.23) is often used to interpret experimental data (23, 135, 136, 141, 142) under the assumption that the AX$_2$ units can be considered as the basic entity, with an associated scattering factor $(c_{Si}b_{Si} + c_{Si}b_{Si})$. It was not clear to us until doing the present analysis that this was justified, as two out of the four neighboring X atoms are arbitrarily associated with an A atom, and in addition, this AX$_2$ unit may straddle the perimeter of the sampling volume, leading to partial counting. Nevertheless, the above derivation shows that this widely used phenomenological assumption (23) is indeed justified and correct, subject to there being no chemical concentration fluctuations, so that each A atom is bonded to four X atoms and each X atom is bonded to two A atoms.

Experiments to determine the absolute value of $S(0)$ are not easy because the scattering has to be normalized to a standard, and also because of multiple scattering corrections that are best determined by measuring a number of samples of varying thickness and extrapolating to zero thickness. This complicated procedure has been done recently by Wright (141, 142), who using Eq. (6.23) obtains a value for vitreous silica of $S(0) = 0.0300 \pm 0.0016$ per formula unit, which by incorporating the factor of three leads to a value for the static structure factor of $S(0) = 0.0900 \pm 0.0048$. Note that Wright was able to get down to $Q \approx 0.02\text{Å}^{-1}$, which is about a factor of 10 better than can be obtained with the 300k model. The model value of $S(0) = 0.116$ is about 20% higher than the experimental value, which we comment on below. Nevertheless, this is the first calculation of $S(0)$ from a model of vitreous silica and is gratifyingly close to the experimental value.

We note that Salmon (22, 132, 143) has made measurements of structure factors on a number of AX_2 glasses using isotopes so that the partial structure factors can be found, and hence $S_{NN}(Q)$. These experiments are a real tour de force but not specifically designed to measure the $Q \rightarrow 0$ limit. Not being performed at very small Q (down to $Q \approx 0.5\text{Å}^{-1}$) and they are only indicative, but approximate values can be extrapolated from the plots of the partial structure factors at small Q, giving values between ~ 0.1 and ~ 0.15 for $Ge0_2$, $GeSe_2$, and $ZnCl_2$ (22, 132, 143). These are very close to the more accurate value for vitreous silica obtained by Wright *et al*. and to the model calculation performed here, suggesting *perhaps* that this value, $S(0)$ ~ 0.10 is a general feature of AX_2 glasses, as a value ~ 0.035 is characteristic of single component tetrahedral glasses.

Discussion

For a system in thermal equilibrium, like a liquid, we expect Eq. (6.4) to hold. It is useful to use this relation to access how far amorphous silicon, as well other amorphous materials and glasses, are from equilibrium. The compressibility χ_T of amorphous silicon is between 2 x 10^{-11} m^2/N and 3 x 10^{-11} m^2/N, obtained from silicon-aluminum alloy data extrapolated to zero aluminum doping (144). Using $\rho_0 = 0.05$ atoms/Å^3 (35, 145), and using room temperature of 300 K, we find from Eq. (6.4) that 0.004 < $S(0)$ < 0.006, which is an order of magnitude less than the computer model value of 0.035. If we use the *melting temperature* of crystalline silicon of roughly T = 1685 K (146), this estimate increases to 0.023 < $S(0)$ < 0.035, where we note that both the density ρ_0 and the compressibility χ_T are only weakly dependent on temperature so that almost all of the temperature dependence in Eq. (6.4) comes through the temperature factor T itself. Nevertheless, the figures based on high temperatures are in the general area of the value of $S(0)$ = 0.035 determined from the 100k model, which is not unreasonable. Note that the comparison is a little less favorable if we use the melting temperatures of 1220 K to 1420 K for amorphous silicon (146, 147), which leads to 0.017 < $S(0)$ < 0.030.

When comparing $S(0)$ to experimental results, one must also consider the possibility for structural heterogeneity in the experimental sample that may depart from that present in continuous random network (CRN) models. For example, using electron correlograph analysis, Treacy and coworkers (148) demonstrated that measurements were more closely reproduced by modeling the sample as being at least 65% paracrystalline by volume, with the remaining 35% a CRN. If this were indeed the case, one would expect $S(0)$ = 0.035 for the 100k model to serve as an upper bound. It is also interesting to note that Treacy and coworkers observed ordering on length scales of 10-20 Å, similar to the distances of 12-15 Å over which strong correlations persist in the CRN models studied here, as shown in Figure 6.2, and comparable to estimated length scales for dynamic and structural heterogeneity in glasses (92). Previous studies of CRN (29) and paracrystalline models (149, 150) have shown structural and electronic properties in strong agreement with experiments. Hypotheses regarding paracrystalline regions have a long history, existing even in the time of Zachariasen (24), and future coupling of experiment and theory are needed to resolve this debate.

The most extensive data on the static structure factor for liquid and vitreous silica have been assembled by Levelut and co-workers (23, 135, 136). They used small angle X-ray scattering with wavevectors down to $Q \approx 0.027\text{Å}^{-1}$, which is comparable to that obtained from the 300k vitreous silica model used in this work. Absolute measurements are difficult in this region [a notable exception being the work of Wright *et al*. (2005)] and so it was necessary to normalize to the assumed liquid behavior at high temperatures

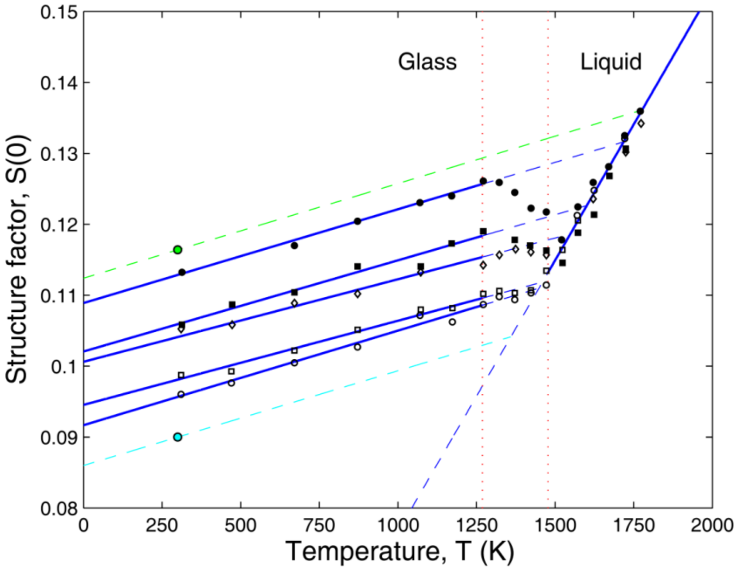


Figure 6.8 The points and fitted blue solid lines in both the glass and liquid region of silica are digitized from Figure (2) of Levelut (23) multiplied by a factor of 1.43 as described in the text. The five lines in the glass phase correspond to fictive temperatures of 1373 K (open circles), 1473 K (open squares), 1533 K (solid squares), 1573 K (open diamonds), and 1773 K (solid squares). The lower isolated point (cian) is from Wright (141) and the upper isolated point (green) is from the computer model used in this study.

using (4). However, there are discrepancies between compressibility values and so there is some uncertainty as to what values to take (Levelut *et al*., 2005). Note that there is a factor of 900 = $(30)^2$ between the data of Wright and Levelut, due to the electron units used by Levelut, which in turn differs by a factor of three from the conventional definition of the structure factor as used here and by Salmon (22, 132, 140).

To try and gain some perspective, we have used another set of compressibility measurements (141, 142, 151) and assumed (6.4) to be true in order to renormalize the Levelut data *upward* by a factor of 1.43, which is now re-plotted in Figure 6.7. This scale factor is the ratio of the liquid compressibility value quoted by Bucaro (151) to the average of the two liquid compressibility values quoted by Levelut (23), i.e. 1.43 = 8.50/[(6.16+5.69)/2]. Figure 6.7 raises many interesting questions relating to glass structure and the fictive temperature (152). It is clear from the data of Levelut *et al*. that the fictive temperature is very close to where the extrapolated straight lines from the glass phase intersect with the liquid structure factor. Note that the temperature dependence is considerably lower in the glass phase and is due to the thermal vibrations about a fixed network topology (141, 142, 153). A most important and intriguing question is *how is information about the fictive temperature embedded in the glass at room temperature?* The information presumably involves ring statistics and possibly the oxygen angle distribution, but it is subtle and will require careful modeling to resolve. All models used will have to be as large as those used in this study to get reliable values for $S(0)$, as discussed earlier. The dashed lines drawn through the two isolated points in Figure 6.7, parallel to the Levelut *et al*. lines, suggest a fictive temperature of ~1360 K for the Wright sample and a fictive temperature of ~1780 K for the 300k vitreous silica model of Vink and Barkema (30), which is close to the value of 1740 K used for the start of the quench in their computer model. Note that while this close agreement is promising, one must not forget that the computer model is quenched at a much more rapid rate than an actual sample, and it is not clear how close the values of the experimental temperature and the "computer" temperature should be. One might argue that the quench rate is of secondary importance to the fictive temperature in determining the glass structure, but this is very speculative and requires further study.

A final note regards the strange behavior of $S(0)$ in Figure 6.8 between 1250 K and the intersection with the liquid line. Instead of following the linear trend due to thermal vibrations about a fixed topology, the values of $S(0)$ noticeably descend prematurely towards the liquid line. This is a common behavior in glasses that can also be seen in the behavior of the volume as a function of temperature, as seen in Figure 6.9. The faster cooling necessary to obtain the higher fictive temperatures freezes in disorder that would have otherwise been able to relax on relatively short timescales. As the samples were heated and analyzed to obtain Figure 6.8, these relaxations started to

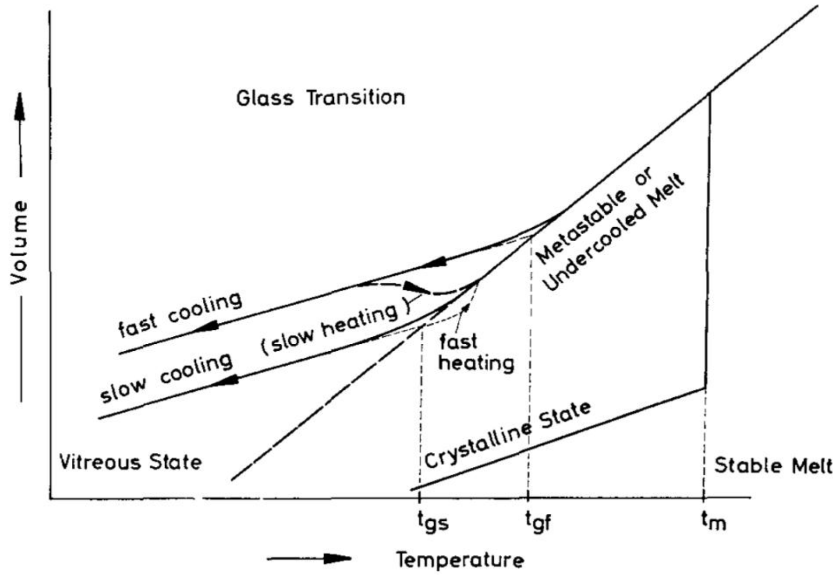


Figure 6.9 Dependence of the glass transition temperature on the cooling rate, showing relaxation upon reheating. Figure reproduced from (87).

occur, causing the structure and therefore $S(0)$ to approach that of the metastable supercooled liquid.

Concluding remarks

The static structure factor $S(0)$ for two non-crystalline materials, amorphous silicon and vitreous silica, lie between that of a crystalline solid (where it is close to zero) and that of a liquid. From the 100k amorphous silicon model of Mousseau, Barkema, and Vink, the static structure factor is computed to be $S(0)$ = 0.035±0.001. This non-zero value is caused by density fluctuations, similar to those found in a liquid, even though the system is far from thermal equilibrium, and seems to be determined largely by the tetrahedral coordination in the amorphous material. This result awaits experimental confirmation, for which it will also be interesting to measure the temperature dependence, caused by thermal fluctuations about the network structure.

For vitreous silica, the situation is richer as the results depend on both the actual temperature and the fictive temperature, as demonstrated clearly by the experimental results of Levelut *et al*. The large periodic computer model of Vink and Barkema gives a reasonable value $S(0)$ = 0.116±0.003 for vitreous silica at room temperature which corresponds to an experimental fictive temperature of about 1780 K, close to 1740 K used computationally to achieve the quenched structure. The intriguing question that remains unanswered is how the information about the fictive temperature is encoded within the network structure, and we speculate that it is in the distinct ring statistics, but this remains to be seen.

CHAPTER 7: CRACK PROPAGATION IN A NETWORK: A MODEL FOR PROTEIN UNFOLDING UNDER FORCE

Introduction

The manner in which proteins respond under an applied force is of direct biological significance, as the physiological role of many proteins requires them to resist mechanical unfolding. A complete understanding of the mechanical, regulatory and signaling properties of many proteins depends not only on their native state conformations, but also on the nature of intermediate states that become populated when subjected to an applied load. Well studied cases include the A2 domain of von Willebrand factor in which a cleavage site is exposed upon unfolding (154-157) and the 10th domain of type III fibronectin for which it has been suggested that partial unfolding reveals an otherwise hidden, so-called cryptic binding site, that could signal extracellular matrix assembly (158).

Mechanical unfolding can be studied experimentally using atomic force microscopy (AFM) in which the two ends are stressed between the tip of a cantilever and a substrate. For polymeric tandem repeats of identical protein domains, this results in saw-tooth or plateau patterns for constant velocity and constant force experiments respectively. Measurements by AFM are typically limited to constant pulling speeds of between 10 nm/s and 1000 nm/s equaling force loading rates of order 10 pN/s to 10,000 pN/s. An overview of this technique can be found in a recent review (159). Lower forces and loading rates, for which standard AFM is ill-suited, can be probed through the use of optical tweezers (154, 160). These experiments have been used in conjunction with ϕ-value analysis on mutants (54, 161) to determine the regions of a protein that are structured in the transition state(s) (162).

While current experimental techniques provide unfolding force distributions, extensions of stable intermediates and the regions that are non-native in structure at the transition state, they do not provide atomistic detail of the underlying events. All-atom molecular dynamics (MD) simulations, typically either at constant force (54, 161, 163) or constant velocity (54, 156), have proved insightful by providing possible intermediate structures and, moreover, unfolding pathways (54, 164), but this comes at a high computational cost. As a result, they must be performed at pulling velocities that are roughly six orders of magnitude greater than those probed experimentally.

Coarse-grained techniques such as Gö-like models (165) narrow the gap in timescales at the cost of representing each residue by a bead and tend to use potentials that favor native interactions while disfavoring non-native ones. Computational cost can also be decreased by employing methods that do not rely on the integration of Newton's equations of motion, such as Monte Carlo based methods (166). There also exists coarse-grained techniques that focus explicitly on the topology of proteins and residue connectivity, such as the work by Eyal *et al*. (167) and Dietz *et al*. (50) in which the effective spring constants and stress distributions in elastic network models were correlated with the mean unfolding force of a given pulling geometry.

My work uses stress distributions within constraint networks to determine unfolding properties, similar in spirit to the study by Dietz *et al*. (50), but differs in many respects. In recognition of the importance of non-native states to the functional roles of many proteins, this study probes beyond the native state, in contrast to previous elastic network models (50, 167). Unlike these former coarse-grained studies, an all-atom representation is used that maintains proper stereochemistry and contains specific interactions such as hydrogen bonds and salt bridges that are vital to a protein's resistance to force. The premise of this work is that structural heterogeneities of the bond network affect how stress is distributed and in turn determine the order in which different regions unfold, as some bonding patterns bear the load in series and others in parallel, as shown in Figure 7.1. By performing this work, the influence of geometry and topology on the

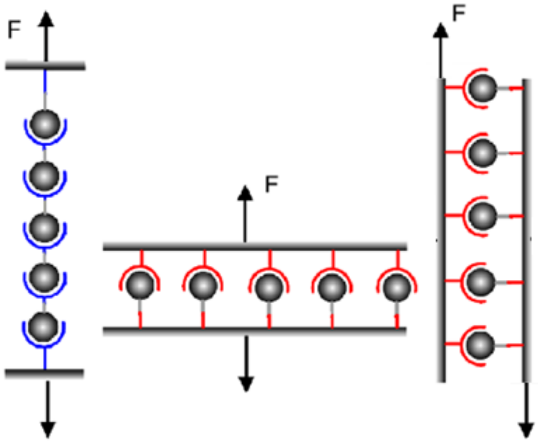


Figure 7.1 Different bond arrangements resisting a force. Figure reproduced from (50).

complete unfolding pathways of proteins is explored and it is found that the simple and intuitive model of protein unfolding as crack propagation on a constraint network is sufficient to capture the unfolding pathways of a diverse set of proteins far from their native state.

My contribution to this study involved the customization of FRODAN to the unfolding problem, which included the creation of modules for determining hydrogen bond burial and bond breaking. I calibrated the model's spring constants and other parameters to improve agreement with MD pathways, as well as performing full analysis of the constrain-based unfolding pathways. I also wrote the paper, which is currently in press.

Model and methods

Constraint-based model

The model used in this work is an extension of the FRODAN model described in Chapter 4. Standard applications of FRODAN treat the constraints as either fixed throughout a simulation or selectively removed from knowledge of the constraints in a final target state. The energy function serves only to re-enforce the constraints upon random perturbation, allowing a universal spring constant for all constraints to be sufficient for this role. For the problem of modeling protein unfolding under force, the FRODAN model was extended in several ways.

The first extension stems from the fact that a constraint network under tension is not able to satisfy all of its constraints, as the distribution of tension in the network acts to balance the external force. For general networks, the tension distribution depends on the individual spring constants, which must therefore be set to realistic values. For this work, k_{sh} was assigned a large value such that all distances between copies of a shared atom rarely exceed 0.02 Å after minimization. The value of k_{st} was chosen such that atoms rarely approach more than 0.2 Å closer than their pair-specific constraint distance (168). The value of k_{rm} was chosen to roughly reproduce the barrier height associated with the O_{i-1}-C^{β} clash in the Ramachandran plot of alanine dipeptide, following the work of Ho *et al*. (118) and Maragakis *et al*. (169), and k_{tr} was calibrated to match the anti/gauche barrier of *n*-butane (170). The remaining spring constants, k_{hb} and k_{ph}, were free parameters that were chosen so as to best match the MD unfolding pathways for the set of 12 proteins in this study. Hydrogen bonds were originally divided into backbone (bbhb) and side-chain (hb) types with k_{bbhb} assigned a fixed value (based on a Mayo potential with a well depth 2 kcal/mol (119)) and k_{hb} allowed to vary, but it was found that agreement with MD was best when k_{hb} possessed the same fixed value that was assigned to k_{bbhb}. Once chosen, the same values for all spring constants were used for all 12 proteins. These values are

$$\begin{aligned} k_{sh} &= 1000\lambda & k_{st} &= 100\lambda \\ k_{rm} &= 28\lambda & k_{tr} &= 18\lambda \\ k_{hb} &= 30\lambda & k_{ph} &= 5\lambda \end{aligned} \tag{7.1}$$

where λ = 1 kcal/(mol·Å^2). Hydrogen bond, salt bridge and hydrophobic constraints are intrinsically different from the others, as they break during the unfolding process. To account for this, these constraints have a maximum load that they can bear, beyond which they break and are removed. As the load across a constraint is equal to the product of its spring constant and the extent of violation, the maximum load was set by giving the constraints a default maximum extension x_{max} = 0.15 Å, a value chosen to be small enough to prevent significant distortion of the protein structure.

The strength of hydrogen bonds and salt bridges depend on the effective dielectric properties of their environment, with the dielectric constant of the solvent being much greater than that inside of a protein. To account for this, we scale the maximum load of hydrogen bonds and salt bridges by multiplying the breaking extension x_{max} by a factor Ω describing the extent of burial. The factor is simply the number of non-hydrogen atoms within a distance of 7.2 Å (four water layers, as used in (171)) of the geometric center of the interaction of interest (i.e. the hydrogen and acceptor atoms), normalized such that the maximum value of Ω is 2 for the set of 12 proteins in their native states. To ensure that fully exposed hydrogen bonds and salt bridges maintained a finite load-bearing capacity, values of Ω below 0.5 were set to 0.5.

Constraint-based unfolding algorithm

The second extension to the standard FRODAN model involved the development of a force-induced constraint breaking algorithm. To induce unfolding, three backbone atoms of each of the terminal residues are targeted to an equal number of target atoms placed on either side of the protein at a separation much larger than the length of the unfolded protein. The targeted atoms are then pulled apart by using a biasing potential of the form

$$E_{RMSD} = \begin{cases} \frac{1}{2} k_{RMSD}(RMSD - C)^2 & RMSD \geq C \\ 0 & RMSD < C \end{cases} \tag{7.2}$$

and decreasing the desired RMSD, *C*, in steps of 0.05 Å, where the RMSD is measured between the six targeted atoms and their target values. The value k_{RMSD} = 300λ was chosen to be as low as possible without having the difference *RMSD* – *C* ever exceed x_{max}. Throughout this work, the extension of a structure is defined as the N-to-C distance between the backbone nitrogen of the first residue and the backbone carbonyl carbon of the last residue. The complete potential is given by

$$E = E_{prot} + E_{RMSD} \tag{7.3}$$

where

$$E_{prot} = E_{equality} + E_{lt} + E_{gt} \tag{7.4}$$

The algorithm for mapping out an unfolding pathway can be summarized as:

1) Decrease the desired RMSD to the target *C* by 0.05 Å.
2) Minimize the energy (6) within the constraint network, resulting in an equilibrium stress distribution.
3) If one or more of the breakable constraints exceed their maximum allowed extension, remove the constraint with the greatest fractional excess and return to step 2. Otherwise, return to step 1.

Upon each iteration, any new hydrogen bonds, salt bridges and hydrophobic interactions that have arisen are identified and added using the same criteria as previously described. This allows non-native interactions to form along the pathway. The algorithm is followed until the protein is completely unfolded.

This model of protein unfolding is similar to crack propagation in a solid material, being deterministic and force-driven. The results share many characteristics with unfolding at low temperature and high force, but differ in that the force distribution is always in equilibrium and the constraint network is minimally overloaded to cause unfolding to proceed. The constraints could have been broken probabilistically, but rather

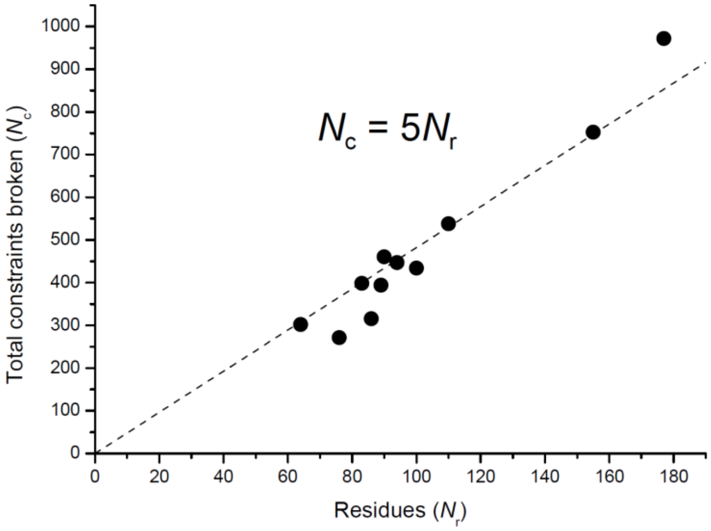


Figure 7.2 Total constraints broken as a function of the number of residues for the 12 proteins in this study. On average, each residue is involved in $2 \times 5 = 10$ constraints, as each constraint has two end points. Figure courtesy of Phil Williams.

variation in the pathways is allowed to arise solely from variation in the starting structures.

Choice of model proteins

The set of 12 single-domain proteins in this study were chosen from those that have been experimentally characterized while selecting for a broad range of topologies, as described in Table 1. The set contains immunoglobulin-like β-sandwich proteins I27 (I27, PDB ID 1tit (172)), fibronectin (FNfn10, 1fnf (173)), tenascin (TNfn3, 1ten (174)), PKD (ArPKD , 1loq (175)), and filamin (DDFLN4, 1ksr (176)), as well as proteins containing both α-helical and β-sheet regions, like the β-grasp proteins ubiquitin (1ubq (177)) and protein L (1hz6 (178)), and the larger proteins ribonuclease H (RNase H , 2rn2 (179)) and von Willebrand factor (vWF A2, 3gxb (180)). The diversity of folds is rounded out with the non-mechanical α and β protein barnase (1bni (181)) and two all-helical proteins spectrin (1aj3 (182)) and acyl-coenzyme A binding protein (ACA , 2abd (183)), the latter being the only protein in the set not to have been studied experimentally.

Molecular dynamics

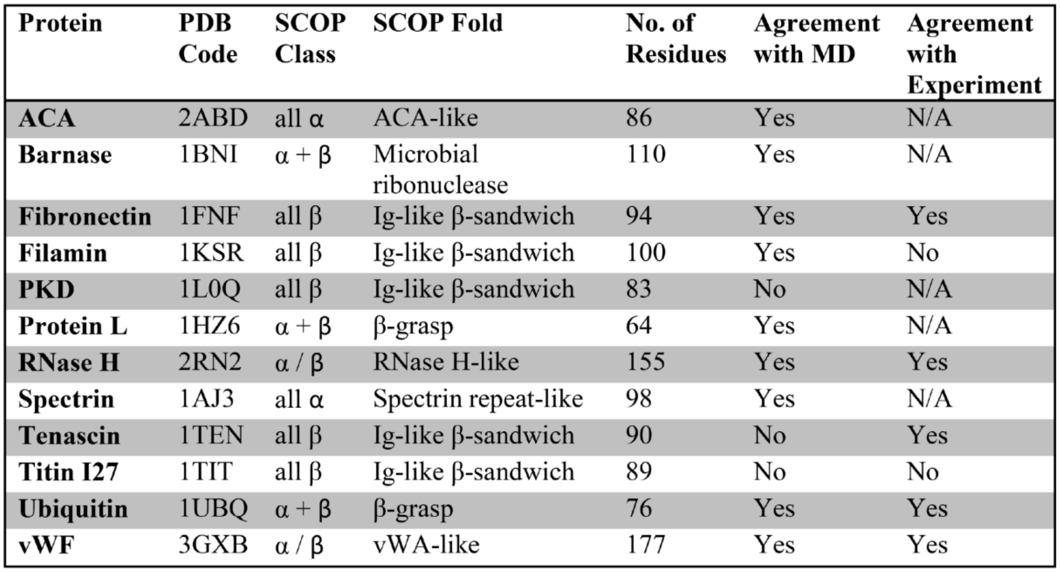
Molecular dynamics simulations were performed using CHARMM (102) and an implicit solvation model (EEF1) (106, 184). Starting structures from the PDB were minimized, heated and then equilibrated for at least 1 ns (100 ps for ACA, as described previously (185)). For each protein a further equilibration of 5 ns was performed from which 10 pairs of coordinates and velocities were extracted, each spaced 500 ps apart (1 ns equilibration and 100 ps spacing for ACA). Ten constant force molecular dynamics simulations were then performed for each protein. Force was applied to both the main-chain nitrogen of the N-terminus and the main-chain carbonyl carbon of the C-terminus, in the direction of the vector between the two atoms such that the protein was pulled apart. A constant force of 265 pN was applied to I27, 220 pN to TNfn3, 300 pN to vWF A2, 205 pN to barnase, 525 pN to protein L, 150 pN to FNfn10, 375 pN to ubiquitin, 250 pN to RNase H, 190 pN to DDFLN4, 250 pN to ArPKD, 250 pN to spectrin and 125 pN to ACA. Ten constant velocity MD simulations were also performed starting from the same set of coordinates and velocities used to begin the constant force simulations. A spring constant of 1 kcal/(mol·Å^2) was used to enforce the constant pulling velocity, which was such that each protein would unfold fully in the 10 ns simulations. All simulations were performed at a temperature of 298.15 K using the Nosé-Hoover thermostat with a 2 fs timestep.

Results

The constraint-based method, which models protein unfolding as crack propagation on a constraint network is compared with both constant force and constant velocity MD simulations. Within the constraint-based model, the mechanical stability of a given structure can be inferred from the amount of force required to minimally overload the network and cause unfolding to proceed. At some stages of the unfolding pathway a protein may be well braced with the load shared in parallel over many constraints, whereas at other stages the constraints act more in series, causing the network to be more easily overloaded. The relative extensions of such states will be determined from the position of force peaks along the unfolding pathway.

Comparison to constant force MD simulation

Despite the constraint-based algorithm being more akin to constant velocity MD simulation, a comparison is made to constant force MD unfolding pathways, as the latter method allows the protein to spend more time in conformations with high stability and thus results in pathways that may be closer to those probed experimentally. For the purposes of characterizing and comparing unfolding pathways, 10 constraint-based and 10 constant force MD pathways were generated beginning from equilibrated structures and a set of critical structures were identified to act as check points in a flow diagram connecting the native state to the unfolded ensemble, as shown in Figure 7.3 for barnase. Both constraint-based and MD pathways were used to select these check points, with those from the constraint-based method being mechanically strong structures that require large forces to unfold and those from constant force MD pathways being structures that result in stable N-to-C distances for prolonged periods of time. From both methods, frequently occurring structures were also selected for which whole units of secondary structure such as α-helices or β-strands were detached. The dominant constraint-based pathways are those with large numbers of trajectories passing between each pair of nodes in the flow diagram. While the constraint-based algorithm unfolded all proteins completely, unfolding was not usually completed within the 10 ns duration of the constant force MD simulations. The flow into and out of nodes is thus only conserved for the constraint-based simulations. The results for 4 of the 12 proteins are described in the next section, with the remaining results described in Appendix B. A summary of the results is shown in Table 7.1.



Protein	PDB Code	SCOP Class	SCOP Fold	No. of Residues	Agreement with MD	Agreement with Experiment
ACA	2ABD	all α	ACA-like	86	Yes	N/A
Barnase	1BNI	α + β	Microbial ribonuclease	110	Yes	N/A
Fibronectin	1FNF	all β	Ig-like β-sandwich	94	Yes	Yes
Filamin	1KSR	all β	Ig-like β-sandwich	100	Yes	No
PKD	1L0Q	all β	Ig-like β-sandwich	83	No	N/A
Protein L	1HZ6	α + β	β-grasp	64	Yes	N/A
RNase H	2RN2	α / β	RNase H-like	155	Yes	Yes
Spectrin	1AJ3	all α	Spectrin repeat-like	98	Yes	N/A
Tenascin	1TEN	all β	Ig-like β-sandwich	90	No	Yes
Titin I27	1TIT	all β	Ig-like β-sandwich	89	No	No
Ubiquitin	1UBQ	α + β	β-grasp	76	Yes	Yes
vWF	3GXB	α / β	vWA-like	177	Yes	Yes

Table 7.1 Summary of the results for the 12 proteins in the data set.

Barnase

Barnase is a bacterial protein with ribonuclease activity that can kill a cell when expressed in the absence of its inhibitor barstar, which binds over the active site to prevent barnase from damaging the cell's RNA. The primary experimental and theoretical unfolding studies of barnase were performed by Best *et al*. (186) using AFM and MD.

Unfolding in the constraint-based pathways begin predominantly from the C-terminus via the detachment of the terminal β_5 (8/10), although the first unfolding event does depend on the constraint distribution in the starting conformation, as α_1 detaches first on 2 occasions, as shown in Figure 7.3. Among the 8 constraint-based pathways for which β_5 is the first to detach, α_1 and β_4 are equally likely to be the next to unfold. In 6 of the constraint-based pathways, α_1 and β_5 are the first two structural units to detach, and half of pathways lead to a core lacking α_1, β_4 and β_5 protected at each end by a pair of β-

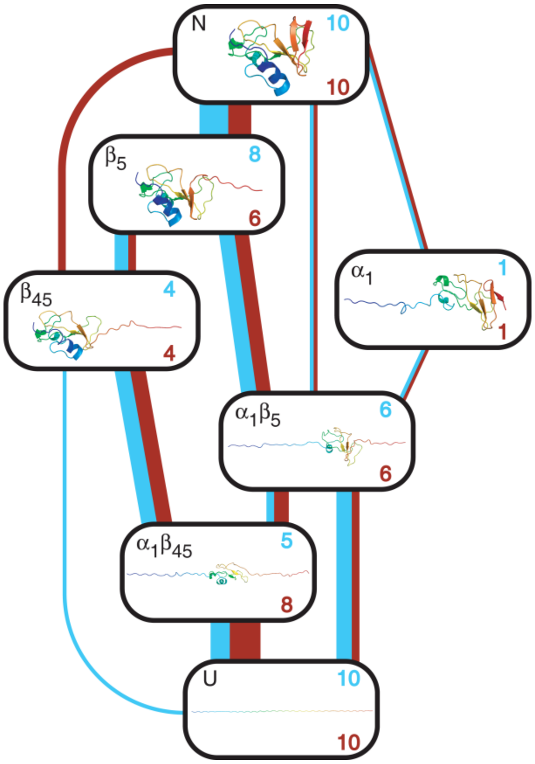


Figure 7.3 Unfolding pathways of barnase. Boxes serve as check points, with the label at the top left corner indicating the secondary structure that has been lost. The boxes are connected by lines; colored blue (left) and red (right) for constraint-based pathways and MD trajectories respectively, and have thicknesses proportional to the number of pathways that transit between the two end states. The numbers at the upper right and lower right of each box indicate the number of incoming constraint-based pathways and MD pathways respectively.

clamps, one consisting of β_2 and β_3, and the other of β_1 and residues of the α_1-α_2 loop. These features are in excellent agreement with the MD pathways. The same fraction of unfolding events begin from the C-terminus (8/10), detachment of α_1 and β_4 are both observed to follow the detachment of β_5, and detachment of α_1 and β_5 share the same likelihood to be the first two unfolding steps (6/10). Although extreme for this protein, these similarities highlight the ability of the simple deterministic model of crack propagation on a constraint network to capture the diversity of pathways found from MD simulations.

von Willebrand factor

von Willebrand factor forms long tandem arrays which function within blood vessels to promote blood clotting. Unfolding of the A2 domain exposes a cleavage site that allows the body to regulate the length of the tandem arrays and consequently the extent of clotting at a wound. Mutation to the gene coding for these domains is the most common cause of genetic blood clotting disorders. Unfolding of the A2 domain has recently been studied by Zhang *et al*. (154) using optical tweezers, in which an intermediate was observed with an N-to-C distance 40% that of the fully unfolded domain, with complimentary MD studies performed by Baldauf *et al*. (157) and Chen *et al*. (156).

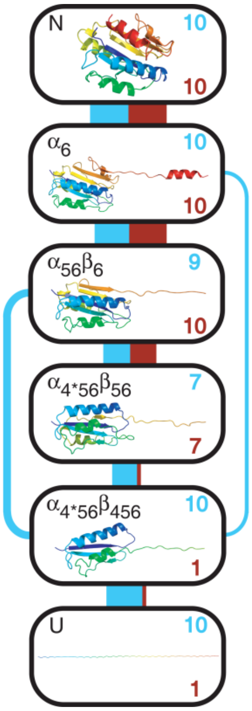


Figure 7.4 Unfolding pathways of the A2 domain of von Willebrand factor. Boxes serve as check points, with the label at the top left corner indicating the secondary structure that has been lost. The boxes are connected by lines; colored blue (left) and red (right) for constraint-based pathways and MD trajectories respectively, and have thicknesses proportional to the number of pathways that transit between the two end states. The numbers at the upper right and lower right of each box indicate the number of incoming constraint-based pathways and MD pathways respectively.

All constraint-based pathways began with the detachment of the C-terminal helix. In the majority of the pathways, unfolding continued through the sequential detachment of β-strands β_6, β_5 and β_4 and unfolding of the α-helices in between, although in 3 pathways multiple β-strands detached as a unit prior to separating from one another. Two states were identified as having the highest mechanical stabilities along the unfolding pathways, one lacking α_5, α_6 and β_6, and the second lacking the α_4–less loop, α_5, α_6, β_5 and β_6. The latter state possesses an extension approximately 40% that of the fully unfolded A2 domain and is thought to correspond to the intermediate observed experimentally by Zhang *et al*. (154) using optical tweezers. Interestingly, the cleavage site on β_4 becomes highly exposed for the first time in this latter state, as it is no longer protected by β_5. Again, the dominant pathway from the constraint-based model is able to reproduce the pathways observed in the constant force MD simulations, as shown in Figure 7.4.

Titin I27

Titin I27 is the 27th immunoglobulin domain within the I-band region of the giant muscle protein titin. The mechanical unfolding of I27 has been studied by AFM (8, 162, 187), combined AFM and MD studies on mutants (161), as well as by various coarse-grained models (165, 188). At forces above about 100 pN (8, 189), I27 is believed to unfold via a force-stabilized intermediate I_1 with an extension roughly 6 Å greater than that of the native state (164, 189). The results of a mutational study (161) suggest that I_1 lacks native contacts between Val4 and β_G, while MD simulations at forces of 300 pN from the same study predict the loss of some contacts between β_A and β_B, including the backbone hydrogen bond between Glu3(O) and Ser26(H) (161).

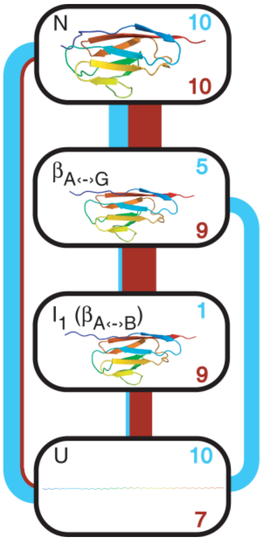

N
10
10
$\beta_{A\leftrightarrow G}$
5
9
I_1 ($\beta_{A\leftrightarrow B}$)
1
9
U
10
7

Figure 7.5 Unfolding pathways of titin I27. Boxes serve as check points, with the label at the top left corner indicating the secondary structure that has been lost. A box possesses the additional label “I” if the state has been identified as an intermediate in previous studies. The boxes are connected by lines; colored blue (left) and red (right) for constraint-based pathways and MD trajectories respectively, and have thicknesses proportional to the number of pathways that transit between the two end states. The numbers at the upper right and lower right of each box indicate the number of incoming constraint-based pathways and MD pathways respectively.

In the constraint-based pathways, represented in Figure 7.5 along with the MD pathways, the first constraints to break were predominantly hydrophobic interactions in two regions, one at the N-terminus and the other beneath the C-clamp. In all pathways, breaking of constraints at the N-terminus allowed residues 1 and 2 to separate from the hydrophobic core and extend along the direction of force. In only half of the pathways did this lead to the separation of residues 1 to 4 from β_G prior to the shearing apart of the C-clamp, as the ability of the C-clamp to resist the load may have been compromised by the aforementioned breaking of hydrophobic interactions at the C-terminus. Despite having an average N-to-C extension indistinguishable from the average extension of 52.5 Å for I_1 found from the MD simulations of this study, the constraint-based state in which residues 1 to 4 have separated from β_G are not considered to be I_1 due to the presence of the backbone hydrogen bond between Glu3(O) and Ser26(H), which is absent in I_1 in the

MD simulations of the present work. To distinguish between these two nearby states, they were each assigned to different nodes in the flow diagram, one labelled $\beta_{A\leftrightarrow G}$ and the other $I_1(\beta_{A\leftrightarrow B})$, where $\beta_{A\leftrightarrow B}$ implies that Glu3(O) and Ser26(H) have separated. It should be noted that the relative stability of the two termini are finely balanced, as demonstrated by the experimental observation that the single mutation of Val86 to Ala86 is sufficient to cause I27 to no longer unfold via I_1 (8). Despite this, I_1 is not included in a dominant constraint-based unfolding pathway and is considered a failure of the model.

Fibronectin

Fibronectin forms part of the extracellular matrix and is likely under frequent tension. It has been proposed that the stretching and partial unfolding of fibronectin may expose a hidden binding site that could signal extracellular matrix assembly (158). In an AFM study by Li *et al*. (190), an intermediate was observed. By unfolding mutations, they concluded that the intermediate likely involved the unfolding of stands β_A and β_B. Several computational studies on fibronectin type III domains (163, 191) have suggested the presence of multiple energy barriers along the unfolding pathway.

The first unfolding event in all 10 constraint-based pathways involves the breaking of hydrogen bonds at the N-terminus between stands β_A and β_G. The external force, which runs along the axis going through the N- and C-terminal residues, applies a torque to the two β-sheets, causing many hydrophobic constraints to break and non-native ones to form as the two sheets rotate relative to one another. This rotation increases the N-to-C distance from roughly 48 Å to 60 Å upon which the β-strands become closely aligned. In the majority (6/10) of the constraint-based pathways, unfolding proceeds through the detachment of strands β_A and β_B, forming the intermediate observed experimentally by Li *et al*. Interestingly, this structure was found to possess the highest mechanical stability among all structures along the unfolding pathways and thus from a

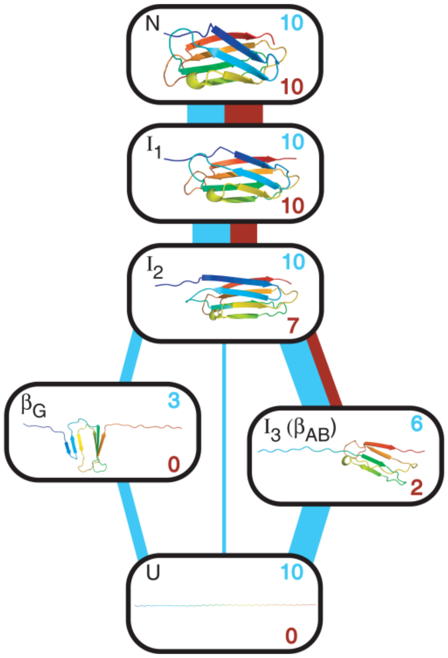


Figure 7.6 Unfolding pathways of fibronectin. Boxes serve as check points, with the label at the top left corner indicating the secondary structure that has been lost. A box possesses the additional label "I" if the state has been identified as an intermediate in previous studies. The boxes are connected by lines; colored blue (left) and red (right) for constraint-based pathways and MD trajectories respectively, and have thicknesses proportional to the number of pathways that transit between the two end states. The numbers at the upper right and lower right of each box indicate the number of incoming constraint-based pathways and MD pathways respectively.

purely mechanical perspective would be the best candidate for the intermediate, in agreement with the conclusions drawn from mutation analysis (190) and the constant force MD simulations, as summarized in Figure 7.6.

Comparison to constant velocity MD simulation

For a given protein, the mechanical strength of the corresponding constraint network changes along the unfolding pathway. This variation in mechanical strength is expressed by plotting the force required to minimally overload the network as a function of the N-to-C distance. The resulting profile is compared with results from constant velocity MD simulation in which a variable force is applied to the two terminal residues to cause them to separate at a constant velocity. Ten constant velocity MD simulations were performed for each of the 12 proteins in this study.

von Willebrand factor

The force profiles for vWF from both methods, compared in Figure 7.7, show a small peak at an N-to-C distances below 20 Å that corresponds to the detachment of the C-terminal end of α_6 from the remainder of the protein. Both models display a second small peak at 45 Å corresponding to the breaking of strong side-chain interactions, allowing α_6 to become completely free. The remaining peaks observed from the constraint-based method, located at approximately 100 Å, 175 Å and 250 Å, correspond to the detachment of β-strands β_6, β_5 and β_4 respectively and agree well with peaks due to the same unfolding events observed in the constant velocity MD simulations. The two dominant peaks in the constraint-based force profiles suggest the presence of mechanically stable structures at extensions slightly less than the peaks at 175 Å and 250 Å. The stability of the latter structure is supported by experiment, as it is thought to correspond to the intermediate observed by Zhang *et al*. (154) using optical tweezers.

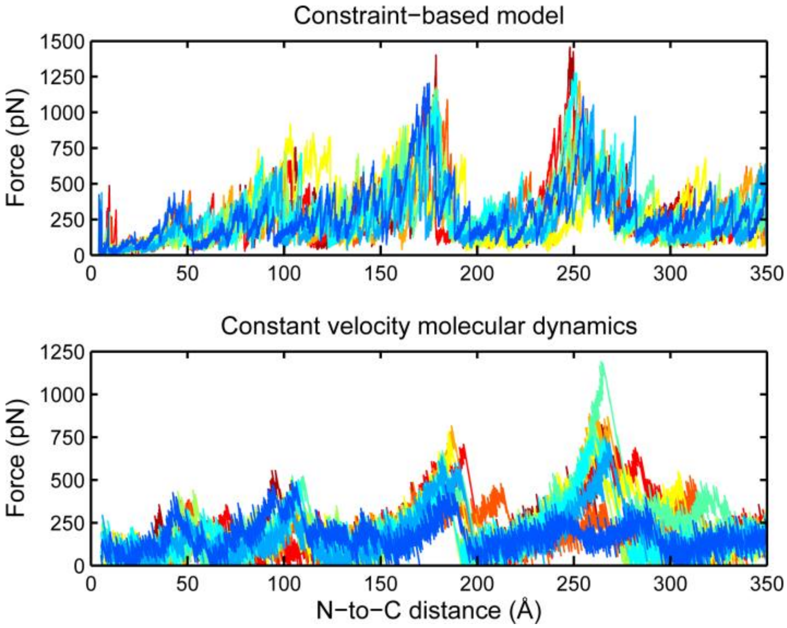


Figure 7.7 Force profiles for the mechanical unfolding of the A2 domain of von Willebrand factor obtained from crack propagation on the constraint network of 10 starting structures compared with those from constant velocity MD simulations.

Fibronectin

The constraint-based profiles for fibronectin, displayed in Figure 7.8, show an initial broad peak that corresponds to the transitions to I_1 and I_2, as well as the initial loss of β-strands. The majority of runs from both sets of pathways form a partially unfolded state known as I_3 in which β_A and β_B have detached. Consistent with the profiles from constant velocity MD simulations, the constraint-based model identifies I_3 as the most mechanically stable structure along the unfolding pathway.

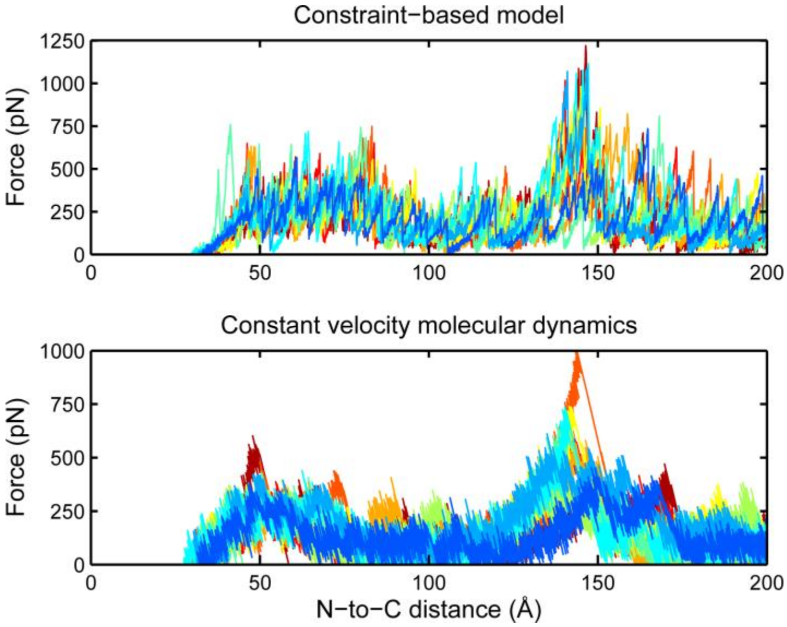


Figure 7.8 Force profiles for the mechanical unfolding of fibronectin obtained from crack propagation on the constraint network of 10 starting structures compared with those from constant velocity MD simulations.

Discussion

The main goal of this work has been to demonstrate that unfolding pathways using MD simulation can be described as crack propagation in a constraint network. Dominant unfolding pathways from the constraint-based approach agree with those from constant force MD simulation for 9 out of the 12 proteins in this study which is impressive considering the simplicity of the model. As experiment is the true metric of comparison for theory, it is valuable to compare the results with those of past experimental studies. Despite the inability of experiments to give an atomic-level picture of the unfolding pathways, mutation analysis can be used to probe the regions of a protein that are altered in a transition state or intermediate. There exists sufficient experimental data to characterize the nature of the intermediates with some confidence for fibronectin, filamin, tenascin, titin I27 and the A2 domain of von Willebrand factor. The dominant constraint-based pathways are consistent with all known intermediates except for those of titin I27 and filamin. For filamin, both constraint-based and MD pathways agree with one another but fail to predict the unfolding of β_A and β_B observed in the mutant study of Schwaiger *et al*. (192). It should be noted that the dominant constraint-based pathway of tenascin was considered to disagree with those from MD simulation solely because of the former's lack of I_3, a state that has never been observed experimentally. Two of the other proteins, namely ubiquitin and RNase H, possess intermediates that have been observed experimentally, but their structures are less conclusively known. For ubiquitin, a mechanically stable state is observed in the constraint-based pathways at 78±11 Å, in impressive agreement with the extension of 81±7 Å found experimentally by Schlierf *et al*. (193). For RNase H, the main constraint-based pathway began with the unfolding of α_5 followed by the detachment of β_1, β_2 and β_3, leaving the stable core observed in bulk studies (194, 195).

Unlike previous studies that use coarse-grained networks to predict properties of the native state alone (50, 167), this simple and intuitive model is sufficient to capture intermediates far from the native state when benchmarked against MD simulations. For example, as discussed in detail in Appendix B, in only a single case (PKD) was the dominant MD pathway not captured by at least one constraint-based pathway. Whilst the applied force caused the A-A' loop in PKD to approach the G strand, the loop did not approach closely enough to form non-native hydrogen bonds, highlighting a limitation of this simple model. Lacking thermal motion and electrostatics, strands are incapable of being electrostatically attracted to one another from a distance to create significant non-native secondary structure. Despite this deficiency, the addition of non-native constraints did improve the unfolding pathways of the majority of proteins in this study. Without them, fibronectin and tenascin do not form I_2, as the two β-sheets are unable to rotate relative to one another prior to separating.

The success of the constraint-based model is surprising, as unfolding is based purely on strain and does not sample the free energy of the states along the pathway. One might expect strain-based pathways to rapidly deviate from those of MD simulations for which the protein is allowed to diffuse in a detailed energy landscape. Instead sequential strain-based breaking events analogous to crack propagation in a solid can be followed far from the native state and reproduce the order of loss of many secondary structure units. This provides further evidence that the high forces used in typical MD simulations tilt the energy landscape to such an extent that the unfolding is not the thermally driven process that occurs under experimentally and physiologically relevant forces.

In principle the constraint-based method could be sensitive to mutations due to its all-atom representation. Of the mutations attempted, namely Ile88→Pro88 and Tyr92→Pro92 in fibronectin and Ile8→Ala8 in tenascin, only Tyr92→Pro92 has been shown experimentally to change the unfolding pathway, causing it to no longer traverse the stable intermediate lacking β_A and β_B (190). None of the constraint-based pathways were affected, including the Tyr92→Pro92 mutant, as the backbone hydrogen bonds disrupted by the latter proline mutation were replaced by side-chain hydrogen bonds of similar strength in the initial structures. Probing mutant sensitivity offers a challenge for future work.

Generating constraint-based pathways is computationally less demanding than the 10 ns MD simulations with which they are compared. A full unfolding pathway of fibronectin requires only 42 minutes on a single HP DL120 3.0 Gz Intel E5472 core, roughly 1/20th that for a constant velocity MD pathway and less than 1/20th that for a constant force MD pathway, as unfolding was often not completed within the 10 ns simulation time. Potential applications of this technique include the study of the dependence of the unfolding pathway on pulling direction, in which force is applied between many pairs of residues [see (159)]. The constraint-based method, which rapidly produces pathways that often possess variability greater than those from MD simulation, can also be used to generate a vast number of stereochemically acceptable all-atom starting structures for milestoning calculations using MD (196).

CHAPTER 8: FLEXIBLE FITTING USING CONSTRAINED GEOMETRIC SIMULATION

Overview

As an imaging technique, cryo-electron microscopy (cryo-EM) has some significant advantages over X-ray crystallography, including the ability to image biomolecules in different conformations without the need for crystallization, as discussed earlier in Chapter 2. The major drawback of cryo-EM is the lower spatial resolution, typically between 4 - 25 Å. Such maps possess relatively little information compared to their X-ray counterparts, the amount of which can be estimated by dividing the volume of a map into cubic elements with sides equal to the resolution. Even the relatively high 7.7 Å resolution map of GroEL (a 192 Å cube) (197) only possesses roughly 25^3 = 15,000 pieces of information, far less than that desired to specify the atomic positions of a complex containing 110,000 atoms. Determining an atomic structure from a cryo-EM map would be hopeless were it not for the large number of constraints that we know from the stereochemistry of polypeptide chains in the form of known bond lengths and angles, favorable Ramachandran and torsion angles, as well as electrostatic interactions such as hydrogen bonds and salt bridges. What is more, the structures of biomolecular complexes imaged by cryo-EM have often been solved by alternative techniques in other conformations, contributing information about the general structural features of their native fold (197).

The ever-increasing number of low and medium resolution cryo-EM maps has spurred the development of techniques that try to predict the atomic coordinates of biomolecules from these maps by making use of a priori knowledge such X-ray or NMR structures, typically of homologous sequences in different conformations to those imaged by cryo-EM. These fitting techniques span a broad range of complexity, from computationally inexpensive rigid-body docking and coarse-grained normal mode fitting to all-atom biased molecular dynamics. Due to the large amount of effort that has gone into determining models from experimental densities, a thorough review of current methods is crucial to a proper assessment of our method.

It should be noted that my work is an extension of previous cryo-EM fitting work by Craig Jolley (198). He created modules for handling density maps and perturbing the rigid units based on the gradient of the density and the gradient of the correlation coefficient. FRODAN had evolved since Jolley's work, and one of my contributions was to create a new cryo-EM fitting module that both accounted for these changes and allowed for greater flexibility for future enhancements. I also added a density-based energy term given by Eq. (8.1) that allows selective bond breaking to occur. Lastly, I created and applied a breaking routine to a benchmark set of 7 proteins, as described in the Results section.

Review of currently used techniques

Rigid-body docking

The first step in most fitting algorithms is finding the optimal position and orientation of the initial atomic model with respect to the cryo-EM density. Treating the entire atomic model as a single rigid body reduces the dimensionality of the search to the three translational and three rotational rigid-body degrees of freedom (60). This is commonly solved by first representing the model and the target in a simplified representation. Borrowing techniques developed in the image processing field for creating maximally informative reduced representations of objects, a cryo-EM density map can be coarse-grained into a set of feature points that, generally speaking, identify mutually exclusive regions of high density by a method called vector quantization (60). The number of such features per map will depend on its resolution and information content, but on the order of a dozen are often sufficient to capture a map's shape and structure. The same coarse-graining can be performed on an initial atomic model by first creating an approximate cryo-EM density map for the model by placing diffuse spherically symmetric densities about each atom, often with a Gaussian profile with a width proportional to the resolution of the target map. Once two coarse-grained representations have been created, they can be optimally aligned by various techniques such as anchor-point matching (199) where a mapping is created between the two sets of feature points $\{i\}$ and $\{j\}$ and minimizing a distance metric between sets. Seeing as how rigid-body docking does not alter the internal structure of the atomic model, rigid-body docking is followed by optimization techniques that allow conformational changes to occur during the fitting process. Such processes of optimization are referred to under the umbrella term “flexible fitting.”

Interpolation techniques

Several flexible fitting techniques exist that are based on a first-order approximation that the atomic model collectively deforms as a single body without an immediate regard to the details of the underlying stereochemistry. Possibly the most intuitive “first-order” method of approximate is that of interpolation. In the method of Rusu *et al.* (200), the initial model and target density map are first represented by a set of feature points, as described in the Rigid-body docking section, upon which an optimal matching of the feature points is determined. In general, the two sets of feature points will possess different relative geometries, reflecting the conformational differences between the atomic model and the structure underlying the cryo-EM map. The fundamental assumption of the interpolation technique is that the position of a given atom in the final structure relative to nearby feature points is preserved from the initial structure. For example, if an atom lies at the center of a tetrahedron formed by four feature vectors, the position of the atom is predicted to lie at the center of the matching set of four feature points for the target map. Various interpolation methods can be employed, but in the study of Rusu *et al.* it was found that the best results were obtained when the importance of a nearby feature point in determining a final atom's position is scales inversely with its distance to that atom in the initial structure.

This technique has the advantage that it is computationally inexpensive, but generally leads to only large-scale conformational changes, as it possesses no means of sampling such things as alternative loop conformations. The final structure formed upon interpolating all atoms will also contain distorted bond lengths and angles. These can be alleviated by running the structure through a refinement tool such as RefMac (201), but this is external to the method itself and could equally well be used on the output of any flexible fitting technique.

Normal mode fitting

One of the most common methods of flexible fitting, normal mode techniques allow an initial model to deform along a subset of low frequency normal modes. Given an energy function, the energy landscape surrounding a minimized conformation can be approximated by a Hessian describing the local harmonic curvature under all possible deformations. The Hessian can be diagonalized to find a set of eigenfunctions (normal modes) and corresponding eigenvalues (describing the stiffness of the corresponding normal modes). It has been observed for a broad class of proteins that most of a protein's conformational variability exists within the subspace of the lowest frequency modes (61).

The correlation coefficient *C*, given by Eq. (8.3), is a frequently used metric describing the quality of the fit between the target density and an atomic model (specifically the simulated density that it would possess, as described earlier in the Rigid-body docking section). The initial atomic model can be flexibly fit to the target map by taking the gradient of *C* with respect to the set of low frequency modes and choosing to deform along a linear combination of normal modes that allows the greatest increase in the *C* per unit energy of elastic distortion. The amplitude of the motion is limited to prevent significant distortion to occur on any given step, although this does not prevent distortion from accumulating over many such iterations of normal mode-based perturbations (61). While in practice the system could be relaxed according to an atomistic energy function after each iteration, this is not typically done in practice, as it would severely reduce the computational efficiency that is a major strength of the method.

Hybrid elastic network-atomic model flexible fitting

A method implemented in the program DireX (202) developed in the group of Axel Brunger and Michael Levitt combines an atomistic restraint-based model (203) with a deformable elastic network. Proper stereochemistry of the polypeptide chain is maintained by a set of distance interval restraints similar in nature to those in FRODAN, but with a different treatment of non-bonded interactions such as hydrogen bonds, salt-bridges and hydrophobic contacts. Generally speaking, their atomistic model is under-restrained relative to that within FRODAN. As a result, when atoms are iteratively perturbed along directions based on the simulated and target densities, the structural integrity is reduced causing the RMSD to increase with increasing correlation in a process called overfitting. To counteract this effect, a large set of atom pairs are randomly chosen to be connected by harmonic restraints, forming an elastic network over top of the atomic model. By allowing the rest length of these additional restraints to slowly adapt during the fitting process, overfitting is greatly reduced, allowing the atomic model to converge to a stable structure over many iterations possessing a high correlation and low RMSD (for theoretical target maps for which the answer is known).

Threading techniques

Threading consists of two steps; a sequence alignment (superposition) of a starting sequence with a template sequence, followed by the spatial overlaying of the starting sequence (typically a polypeptide chain) on the template model. There exists at least two large groups for which threading is a significant component of their methods, the first developed in the group of Andrej Sali (204). The technique begins from three pieces of information: a known sequence, a template structure, and a cryo-EM map. The starting sequence generally shares relatively little sequence homology with the template, with the percentage of conserved residues often ranging between only 10-30%. With such little sequence overlap, no single alignment is vastly better than all others and it is appropriate to create a large family (~300) of the highest scoring alignments. In the second step, they use each alignment as a guide for spatial threading over the template structure while simultaneously attempting to satisfy spatial restraints as implemented in the program MODELLER (205). Each member of this set of structures is then evaluated according to a scoring function consisting of a weighted average of a “structural integrity” score and a “density fitting” score. The scores of this population are used by a generic algorithm to produce a new set of alignments for the next iteration of threading, details of which can be found in (204). This initial work, which contains no explicit flexible fitting, was extended in a recent work (64). Beginning from the best structure from the aforementioned threading method, the biomolecule is divided into large rigid domains, connected by flexible linkers in cases where a single chain contains more than one domain. A conjugate gradient minimization of a scoring metric consisting of a linear combination of stereochemical, non-bonded, and density fitting scores is then performed with respect to the rigid-body degrees of freedom. The top five scoring structures are then divided into smaller rigid bodies consisting of their secondary structures and the conjugate gradient minimization is repeated. Finally, the top scoring structure is subjected to several iterations of simulated annealing using the same secondary structure rigid bodies, followed by a final conjugate gradient minimization.

A second threading technique has been recently developed in the group of David Baker (206) that uses functionality build into the software package ROSETTA (207). Beginning from a set of sequence alignments, threaded structures are built. A local correlation metric is then used to identify regions of the threaded models that disagree most with the cryo-EM density map. Using the sequence information of these regions, alternative conformations for 3- to 9-residue fragments within these regions are determined and inserted in place of the former fragment, using a Monte Carlo algorithm to ensure loop closure. These new structures are then subject to torsion angle optimization using a combination of an all-atom energy function and a density-matching score. Regions of high disagreement with the cryo-EM map are then re-evaluated and the refinement continues until satisfactory convergence.

Molecular dynamics

Unlike the former techniques that apply an atomistic energy function sparingly to improve computational efficiency, it is possible to perform flexible fitting entirely with molecular dynamics simulation. Information about the target map is introduced by adding a biasing potential to the atomistic force field. One such biasing potential used by Trabuco *et al.* (62) has the form

$$V_{EM} = \begin{cases} \sum_{i=1}^{N} V_{max}\left[1 - \frac{\rho(\boldsymbol{r}_i) - \rho_{thr}}{\rho_{max} - \rho_{thr}}\right] & \text{if } \varphi(\boldsymbol{r}_i) \geq \varphi_{thr} \\ V_{max} & \text{if } \varphi(\boldsymbol{r}_i) < \varphi_{thr} \end{cases} \qquad (8.1)$$

where $\boldsymbol{r}_i$ is the location of an atom, V_{max} sets the overall scale of the energy, ρ_{max} is the maximum map density, and ρ_{thr} is the lowest allowed map density. A lower bound on the density is used because it is often the case in experimental maps that density drops below zero due to negative staining to enhance contrast. While minimizing the combined potential $V_{TOT} = V_{MD} + V_{EM}$ does not strictly minimize the correlation between the model and target density, it does act to pull the biomolecule into favorable high density regions. A drawback of the method is that the biasing potential often has to be applied, making it more than a small perturbation, in order for structural rearrangements to take place in the short amount of time that one is able to simulate due to the high computational costs of all-atom molecular dynamics. Additional restraints are frequently added to $V_{MD} + V_{EM}$ to stabilize secondary structure under the forces resulting from the strong biasing potential V_{EM}. It is true that CPU time is becoming increasingly cheap, but the computational gap is likely to remain due to growth in the size and number of systems being imaged by cryo-EM.

Scoring metrics

The standard metric for assessing the quality of the fit of a model to a target density is the real-space correlation coefficient between the model's theoretical density and that of the target map. As the model consists of a set of atomic coordinates, the theoretical density must be built onto the model. This can be done at several levels of accuracy (198), but for typical cryo-EM densities, the width of the experimental resolution factor (typically > 4 Å) is sufficiently large compared to the width of an atom's Coulomb potential that the "density" about each atom can be approximated by a Gaussian distribution of width equal to the resolution of the target map. The total theoretical density is therefore

$$\rho^{sim}(\boldsymbol{r}) = \sum_{\alpha} Z_{\alpha} \exp\left[-\frac{3(\boldsymbol{r}-\boldsymbol{r}_{\alpha})^2}{2\sigma^2}\right] \tag{8.2}$$

where $\boldsymbol{r}_{\alpha}$ is the position of atom α with atomic number Z_{α}. The real-space correlation coefficient C can then be expressed as

$$C = \frac{\sum_{ijk} \rho^{sim}(\boldsymbol{r}_{ijk})\rho^{exp}(\boldsymbol{r}_{ijk})}{\sqrt{\sum_{ijk}\left[\rho^{sim}(\boldsymbol{r}_{ijk})\right]^2 \sum_{ijk}\left[\rho^{exp}(\boldsymbol{r}_{ijk})\right]^2}} \tag{8.3}$$

where densities are evaluated at the discrete set of points at which the target density has been measured, typically on a cubic lattice (198).

The target density does not have to come from experiment. Especially when developing a fitting technique, it is helpful to have a target map for which the structure is known. For this reason, theoretical target maps are often generated from known atomic models in a different conformation than the starting model. In this case, the quality of the fit can be determined directly from the RMSD between the two structures.

Fitting methods

Perturbation-based fitting

The methods of fitting the atomic model to the target density can be divided into two categories: perturbation-based fitting and fitting based on an energy bias. There are four ways of perturbing the rigid units (RUs), the first of which, random perturbation, was discussed in Chapter 4. This is not ideal as a generator of new conformations, as it does not direct the protein towards the target. For this purpose, one can throw each atom based on either the gradient of the target density at its location or the gradient of the correlation coefficient as a function of each atom's location.

Gradient-based perturbation

Perturbing according to the local gradient of the target density has the advantage that it is fast to compute, as only the target densities at the nearest grid points are needed to approximate the gradient, but the disadvantage that it does not strictly minimize the correlation coefficient. While the perturbations try to make all of the atoms to move towards the highest local density region, in practice this does not happen because of volume exclusion and stereochemical constraints.

Correlation-based perturbation

Perturbation according to the gradient of the correlation coefficient has the advantage that it is based on the metric that one wishes to optimize, but the disadvantage that it is slower to calculate because it is less local in nature. Due to resolution factor being several times larger than the spacing of the lattice points at which the correlation coefficient is evaluated, calculation of the gradient of this correlation requires more computational time than the method based solely on the local gradient of the target density. This increase in computational time can be minimized by sparsening the target map so that the lattice spacing is only two to three times smaller than the resolution factor. Very little information is actually lost by this sparsening, as the densities at nearby lattice points are correlated due to the smoothness of the density itself.

Momentum-based perturbation

The last method of perturbation is meant to be used in conjunction with one of the prior three. If two or more iterations of FRODAN have been run and the conformations of the RUs in the two previous steps are represented by $\mathbf{q}_1$ and $\mathbf{q}_2$, with $\mathbf{q}_2$ being the more recent of the two, use of the momentum perturbation serves to perturb the RUs by an amount $\Delta\mathbf{q}_{\text{mom}} = \mathbf{q}_2 - \mathbf{q}_1$ in the current perturbation step. The total perturbation would therefore be $\Delta\mathbf{q}_{\text{mom}} + \Delta\mathbf{q}_{\text{other}}$, where $\Delta\mathbf{q}_{\mathbf{other}}$ is the perturbation due to one of the three other perturbation methods. If one randomly perturbed and minimized a biomolecule many times, one would notice that the majority of any given perturbation is in a direction orthogonal to the allowed subspace. This component of the perturbation is wasted because is it negated upon enforcing the constraints. The momentum perturbation serves as memory of the component of the last perturbation that was beneficial and assumes that the same direction will be beneficial in the next perturbation step. It has the effect of finding “soft” directions in the available subspace, somewhat analogous to following low-frequency normal modes. Momentum perturbations also require much less computational time in the subsequent minimization step, as constraints tend not to be violated as severely as for random perturbations. The use of momentum-based perturbation has helped to drastically reduce fitting times, particularly for very large biomolecules displaying large conformational changes during the fitting process.

Energy-based fitting

Biasing FRODAN conformations through the addition of an energy term to the constraint-enforcing energy was used previously to model protein unfolding under force.

In that case, the RMSD energy served to increase the distance between pairs of residues. The same general concept can be applied to flexible fitting, in which an energy term of the form Eq. (8.1) is added that favors atoms being in high density regions. This has been applied previously to bias MD simulations towards a target density (62).

Constraint removal

The benefit of E_{EM} is not so much to direct a biomolecule towards the region of the allowed subspace that maximizes the fit to the target density, as perturbation-based optimization is more efficient, but as a means of determining which constraints are most likely impeding further progress. In general, the ideal fit to the density lies outside the subspace defined by the initial set of constraints and thus one of the challenges is to remove as few constraints as possible in order to extend the subspace to include the desired conformation. Unlike previous applications such as geometric targeting, which finds pathways between two known structures (10), in cryo-EM fitting the set of constraints in the final structure is not known a priori. As in the protein unfolding problem, the biasing potential serves to create an equilibrium stress distribution due to the conflicting desire to simultaneously minimize the constraint energy and the biasing energy.

To demonstrate that the equilibrium stress distribution is a good metric for determining constraint removal, we need to understand what the constraints and the equilibrium stress represent. The problem of flexible fitting provides two sources of information: 1) the starting model provides information about favorable contacts of the native fold, represented in a quantized form by the set of initial constraints, and 2) the target density provides information about the shape and arrangement of secondary structure of the target conformation. I specifically say secondary structure because at the resolutions of typical cryo-EM maps, the only internal heterogeneities are due to the high atomic number backbone of secondary structure. As most conformational changes mostly involve relatively minor changes in local contacts, information in the form of initial constraints should only be removed if there is sufficient evidence from the target map. In the limit that you weigh the initial information much more heavily that the target information, constraint removal would never be warranted. In the opposite limit, all initial constraints should be removed, as it is generally true that the correlation coefficient can be improved by removing more constraints, resulting in overfitting. The ideal balance of information is somewhere in between and one can interpret the static stress distribution as representing this compromise.

The spring constants used to model protein unfolding under force were also used in the cryo-EM fitting, with the same breaking extension of 0.15 Å. The relative importance of the initial and target information was controlled through the scaling factor E_{EM}, for which a value of 0.15 was observed to limit breaking to those constraints that hindered “necessary” motions. This latter judgment is by necessity subjective. No constraints were added during the fitting process.

Fitting protocol

Cryo-EM fitting was performed in three stages, namely:

1) 200 steps of both correlation- and momentum-based perturbation (breaking disallowed)
2) 20 steps of gradient-based perturbation with the biasing potential (breaking allowed)
3) 200 steps of both correlation- and momentum-based perturbation (breaking disallowed)

The 200 steps in the first stage are sufficient to converge to the region of the subspace with a high correlation coefficient. This number is greatly reduced through the use of momentum-based perturbation. The second step uses gradient-based perturbations and adds the biasing energy in order to create the constraint violations required for breaking to occur, expanding the subspace as much as is justified by the information in the target map. Twenty steps are sufficient for all breaking events to occur. The final 200 steps allow a final optimization of the correlation coefficient within the expanded subspace.

Benchmark set

A benchmark set of seven monomeric proteins was taken from the work of Topf *et al*. (64) from the group of Andrej Sali. One of the difficulties in proving the value of a new method is that each group applies their method to a different set of proteins. It is therefore appropriate to apply the FRODAN fitting algorithm to a test set that has already been fit by a competing method, called Flex-EM (64). For each protein in the benchmark set, the target map was theoretically generated using a resolution of 10 Å. The initial atomic model was generated by threading the target sequence into a homologous protein whose structure has been determined in a different conformation by X-ray crystallography, as described in the Threading techniques section.

Results

Conformational changes required during flexible fitting can be classified into several categories ranging from global to local scales, namely:

1) relative motion between secondary structure domains (and between monomers for multimeric proteins),
2) Changes in the secondary structure itself, such as register shifts of α-helices and β-sheets,
3) loop rearrangement, and
4) side-chain rotamer changes.

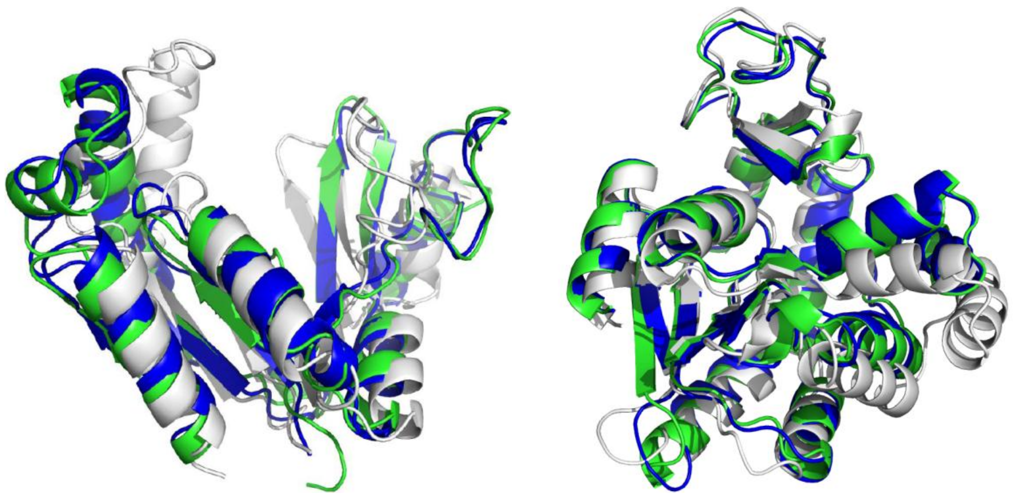


Figure 8.1 Homology models (white) are used to generate flexible fits (blue) to 10 Å-resolution target densities generated from known structures (green) for the SH3 and guanylate kinase domains of PSD-95 (PDB: 1jxm, left) and adenylate kinase (PDB: 1ake, right).

Global motions of type 1 often connect proteins occupying different conformational states induced, for example, by ligand binding. These motions frequently form part of the subspace available to the constraint model, as these motions commonly involve hinges which FRODAN models explicitly. The main impedance to capturing these global motions is that they can involve changes in interfaces as two domains either separate or slide relative to one another, requiring constraints to break. The breaking phase of the flexible fitting protocol allows such motions to occur, as observed for the proteins 1jxm, 1ake, and 1cll in benchmark set, the former two shown in Figure 8.1. In all three cases, the force due to E_{EM} caused both hydrogen bonds and hydrophobic interactions to be broken, allowing the subspace to expand in the direction of a better fit. Less than 5% of hydrogen bonds and salt bridges and 15% of hydrophobic constraints were lost in all three cases. This asymmetry can be understood by the weaker nature of the hydrophobic interactions, which are physically less specific than hydrogen bonds. Motions that resulted from this loss include the separation of an α-helix and β-sheet domain in 1jxm (see Figure 8.1), the closure of domains in 1ake, and the twisting of a

Probe PDB ID	Template PDB ID	Fold	Seq. Identity (%)	Initial C_α RMSD (Å)	Final C_α RMSD (Å) Flex-EM	Final C_α RMSD (Å) FRODAN	Improv. relative to Flex-EM (Å)
1akeA	1dvrB	α/β	46	4.5	2.2	**1.2**	**1.0**
1c1xA	1gtmA	α/β	30	6.6	4.6	**4.1**	**0.5**
1cll	2ggmB	α	52	5.0	3.1	**1.7**	**1.4**
1g5yD	3erdA	α	30	5.4	5.1	**2.8**	**2.3**
1jxmA	1ex7A	α/β	33	5.4	3.3	**2.6**	**0.7**
1uwoA	1k9pA	α	41	4.7	4.0	**3.0**	**1.0**
1cczA	1hnf	β	37	5.2	5.1	**4.7**	**0.4**

Table 8.1 Comparison of FRODAN results to those of Flex-EM for the benchmark set.

terminal domain relative to a central helix in 1cll. By measuring the RMSD between the final fit and the known target structure, shown in Table 8.1, we see that two of the three conformational changes involved significant improvement compared with Flex-EM, in each case bringing the final model almost twice as close to the correct solution. Both structures are fit to within 2 Å of the target structure and thus almost indistinguishable from typical variation in the native ensemble.

Motion of type 2, which involve changes to the secondary structure such as register shifts in α-helices and β-sheets is much more difficult, if not impossible, for FRODAN to capture. Such conformational changes are often necessary due to errors in the homology model that serves as input to the flexible fitting. The level of complexity required to address these problems are simply not feasible to try to address with the current implementation of FRODAN, as register shifts can result in very little difference to the density distribution while requiring many constraints to be broken. This problem is not unique to FRODAN, as MD flexible fitting would have similar trouble due to the long timescales required for partial unfolding and refolding of secondary structure. Much of the residual RMSD observed in the benchmark set, including 1jxm discussed earlier, is due to errors in the homology model. Luckily, homology models can be rapidly generated using many different sequence alignments. The computational efficiency of FRODAN therefore makes it feasible to perform flexible fitting on hundreds if not thousands of homology models, some of which may lack errors in the secondary structural elements.

Motion of type 3 exclusively involves loop rearrangement. Loops can possess multiple low-energy conformations possessing different specific interactions, but of those a single one is typically chosen and submitted to the PDB data bank. It is therefore possible that the information content of a constraint in a loop region is less than that in α-helices or β-sheets. Unfortunately this lack of a unique conformation also causes the gradient of the cryo-EM density to be weaker at the location of loops, as cryo-EM density represents an average over an ensemble of individual proteins in the sample, and thus averages the densities of the various loop conformations. In the fitting discussed here, loops were treated the same as every other part of the protein, causing much of the residual RMSD in structures such as 1g5y and 1uwo are due to poor loop conformations. Possible solutions include making it easier to break constraints in loop regions or their removal altogether. It is unlikely that their removal would lead to overfitting, but this deserves future investigation.

Lastly, motions of type 4 involve the most local conformational changes, namely changes in the rotameric states of side-chains. For models converging to less than 2 Å of the solution (for theoretical maps), a significant fraction of the residual RMSD can be due to poor side-chain conformations. While improvement can be gained by a final refinement stage in which random perturbations are made and conformations accepted or rejected in a Metropolis fashion, this level of fitting can be unjustified for experimental maps. As previously mentioned, experimental maps represent an average over an ensemble of molecules, causing specific side-chain rotamers to be washed out. This is contrary to theoretical maps which are typically generated from a single conformation. Both for clarity and because most proteins in the benchmark set displayed larger scale disagreements that would make rotamer optimization difficult, a final refinement stage was not performed in the fitting process discussed here.

Concluding remarks

A simple and computationally efficient constraint-based fitting algorithm based on the FRODAN package (10) was developed and shown to be effective at determining the atomic structure underlying cryo-EM maps for a set of 7 proteins (64). The RMSD of all 7 fitted structures were closer, in multiple cases significantly so, to the known conformation that those from the group that created the benchmark set. A few of the models were not significantly improved due to errors in the starting structures determined from homology modeling, as well as unwanted constraints in loop regions not allowing the necessary conformational rearrangements. Performing flexible fitting on hundreds of candidate homology models and selective weakening or removing loop constraints should allow for further improved fitting. The method’s ability to capture large scale conformational changes without sacrificing local stereochemistry, its conceptual simplicity, and its computational efficiency makes me hopeful that this can become a standard tool in building all-atom models from low-resolution data.

CHAPTER 9: OUTLOOK

Having had the opportunity to immerse myself in an area and encountered both the strengths and weaknesses of various models, it is beneficial to reflect on the promising avenues that could be taken in the future. In the area of amorphous materials, exciting experiments have very recently been conducted on amorphous silicon that probe density fluctuations on large length scales, allowing values of the limit $S(Q \rightarrow 0)$ to be estimated. Intriguingly, the presently unpublished value lays almost exactly half way between the results derived in this thesis for the 100,000 atom model and that of a crystal. The growing interest in amorphous materials with very small density fluctuations at large length scales (31) is further reason to understand the disparity between the experimental samples and theoretical network models. Future studies could investigate the dependence of density fluctuations on the temperature at which bond transpositions are made in the WWW algorithm, as well as on the details of the energy function.

In the biological realm, both the constraint-based protein unfolding model and the flexible fitting model have room for future improvement. The protein unfolding model uses a rather crude estimate of the hydrogen bond burial factor that could be replaced by more realistic models. These include computationally efficient models for estimating solvent exposed surface areas, as well as future models derived from the excellent experimental work of Jeffery W. Kelly (208) in which he selectively knocks out backbone hydrogen bonds and determines their effect on protein stability. The deterministic breaking algorithm could also be extended to make constraint breaking probabilistic and dependent on the height of the bond’s energy barrier. Lastly, modifications could be made that improve the ability to find non-native contacts. The constraint network for a given protein conformation possesses flexibility that allows pairs of atoms to possess a range of pair distances. These biomolecular motions, which maintain a fixed topology, occur on short timescales compared with more complex protein conformational changes requiring of considerable alteration of the bond network. One could therefore conceive of a search method within this flexibility window for non-native bond formation. Such searches would necessarily make the algorithm more complex, but could allow for more accurate unfolding predictions, especially when compared with either experiment or MD simulation at slower pulling speeds.

Finally, I have great optimism regarding future improvements to the flexible fitting algorithm used to predict atomic structures from low-resolution cryo-EM density maps. This problem is an optimization problem and there exist a large number of potentially fruitful techniques to explore, especially given the computational efficiency and flexibility (no pun intended) of the present model. All of the methods for optimization described in this thesis can be categorized as local in nature, as perturbations are based on gradients, whether it is the gradient of the target density or the gradient of the correlation coefficient. Local techniques can often get trapped at local optima that may not be the best global solution. Feature points extracted by the method of vector quantization discussed previously could be used to locate geometrically unique features in both the atomic model and the target density and associate the two sets through a mapping. This has the potential, particularly for maps with better resolutions, of allowing associated regions to be identified even if they are not overlapping or similarly oriented in the initial alignment and used to guide regions of the constraint-based network model towards associated regions in the target density. This is likely to be particularly useful for large biomolecular complexes consisting of many smaller domains or individual subunits that display a great deal of flexibility that would otherwise be difficult to properly fit by local gradient-based techniques. I truly believe that the constraint-based model within FRODAN has the necessary properties to become a very successful fitting technique that can provide more accurate fitting than normal mode-based techniques while being far more computationally efficient than MD techniques, allowing fits to be performed for a large number of starting structures derived from homology modeling.

REFERENCES

1. Nalwa, H. S. 2001. Silicon-based material and devices. Academic Press, San Diego, Calif. London.

2. Brändén, C.-I., and J. Tooze. 1999. Introduction to protein structure. Garland Pub., New York.

3. James, L. C., and D. S. Tawfik. 2003. Conformational diversity and protein evolution - a 60-year-old hypothesis revisited. Trends in Biochemical Sciences 28:361-368.

4. Takahashi, K., and M. Konagai. 1986. Amorphous silicon solar cells. North Oxford Academic, London.

5. Powell, M. J. 1989. The physics of amorphous-silicon thin-film transistors. Ieee Transactions on Electron Devices 36:2753-2763.

6. Stone, J. 1987. Interactions of hydrogen and deuterium with silica optical fibers - a review. Journal of Lightwave Technology 5:712-733.

7. Alberts, B. 2008. Molecular biology of the cell. Garland Science, New York.

8. Williams, P. M., S. B. Fowler, R. B. Best, J. L. Toca-Herrera, K. A. Scott, A. Steward, and J. Clarke. 2003. Hidden complexity in the mechanical properties of titin. Nature 422:446-449.

9. Menor, S. A., A. M. R. de Graff, and M. F. Thorpe. 2009. Hierarchical plasticity from pair distance fluctuations. Physical Biology 6.

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

11. Gebhardt, J. C. M., and M. Rief. 2009. Force Signaling in Biology. Science 324:1278-1280.

12. Ng, S. P., K. S. Billings, L. G. Randles, and J. Clarke. 2008. Manipulating the stability of fibronectin type III domains by protein engineering. Nanotechnology 19.

13. Warren, B. E. 1969. X-ray diffraction. Addison-Wesley Pub. Co., Reading, Mass.,.

14. Levashov, V. A., S. J. L. Billinge, and M. F. Thorpe. 2005. Density fluctuations and the pair distribution function. Physical Review B 72:024111.

15. Egami, T., and S. J. L. Billinge. 2003. Underneath the Bragg peaks : structural analysis of complex materials. Pergamon, Amsterdam ; Boston.

16. Attard, P. 2002. Thermodynamics and statistical mechanics : equilibrium by entropy maximisation. Academic Press, San Diego, CA.

17. Wachhold, M., K. K. Rangan, M. Lei, M. F. Thorpe, S. J. L. Billinge, V. Petkov, J. Heising, and M. G. Kanatzidis. 2000. Mesostructured metal germanium sulfide and selenide materials based on the tetrahedral Ge4S10 (4-) and Ge4Se10 (4-) units: Surfactant templated three-dimensional disordered frameworks perforated with worm holes. Journal of Solid State Chemistry 152:21-36.

18. Bozin, E. S., G. H. Kwei, H. Takagi, and S. J. L. Billinge. 2000. Neutron diffraction evidence of microscopic charge inhomogeneities in the CuO2 plane of superconducting La2-xSrxCuO4 (0 <= x <= 0.30). Physical Review Letters 84:5856-5859.

19. Petkov, V., S. J. L. Billinge, J. Heising, and M. G. Kanatzidis. 2000. Application of atomic pair distribution function analysis to materials with intrinsic disorder. Three-dimensional structure of exfoliated-restacked WS2: Not just a random turbostratic assembly of layers. Journal of the American Chemical Society 122:11571-11576.

20. Glatter, O., and O. Kratky. 1982. Small angle x-ray scattering. Academic Press, London ; New York.

21. Williams, D. B., and C. B. Carter. 1996. Transmission electron microscopy : a textbook for materials science. Plenum Press, New York.

22. Salmon, P. S. 2006. Decay of the pair correlations and small-angle scattering for binary liquids and glasses. Journal of Physics-Condensed Matter 18:11443-11469.

23. Levelut, C., A. Faivre, R. Le Parc, B. Champagnon, J. L. Hazemann, and J. P. Simon. 2005. In situ measurements of density fluctuations and compressibility in silica glasses as a function of temperature and thermal history. Physical Review B 72:224201.

24. Zachariasen, W. H. 1932. The atomic arrangement in glass. Journal of the American Chemical Society 54:3841-3851.

25. Warren, B. E. 1937. X-ray determination of the structure of liquids and glass. Journal of Applied Physics 8:645-654.

26. Warren, B. E., and J. Biscoe. 1938. The structure of silica glass by X-ray diffraction studies. Journal of the American Ceramic Society 21:49-54.

27. Warren, B. E. 1940. X-ray diffraction study of the structure of glass. Chemical Reviews 26:237-255.

28. Wooten, F., K. Winer, and D. Weaire. 1985. Computer-generation of structural models of amorphous Si and Ge. Physical Review Letters 54:1392-1395.

29. Barkema, G. T., and N. Mousseau. 2000. High-quality continuous random networks. Physical Review B 62:4985-4990.

30. Vink, R. L. C., and G. T. Barkema. 2003. Large well-relaxed models of vitreous silica, coordination numbers, and entropy. Physical Review B 67:245201.

31. Florescu, M., S. Torquato, and P. J. Steinhardt. 2009. Designer disordered materials with large, complete photonic band gaps. Proceedings of the National Academy of Sciences of the United States of America 106:20658-20663.

32. de Graff, A. M. R., and M. F. Thorpe. 2010. The long-wavelength limit of the structure factor of amorphous silicon and vitreous silica. Acta Crystallographica Section A 66:22-31.

33. Djordjevic, B. R., M. F. Thorpe, and F. Wooten. 1995. Computer-model of tetrahedral amorphous diamond. Physical Review B 52:5685-5689.

34. Vink, R. L. C., G. T. Barkema, M. A. Stijnman, and R. H. Bisseling. 2001. Device-size atomistic models of amorphous silicon. Physical Review B 64:245214.

35. Laaziri, K., S. Kycia, S. Roorda, H. Chicoine, J. L. Robertson, J. Wang, and S. C. Moss. 1999. High resolution radial distribution function of pure amorphous silicon. Physical Review Letters 82:3460-3463.

36. Rader, A. J., B. M. Hespenheide, L. A. Kuhn, and M. F. Thorpe. 2002. Protein unfolding: Rigidity lost. Proceedings of the National Academy of Sciences of the United States of America 99:3540-3545.

37. Baldauf, C., R. Schneppenheim, W. Stacklies, T. Obser, A. Pieconka, S. Schneppenheim, U. Budde, J. Zhou, and F. Grater. 2009. Shear-induced unfolding activates von Willebrand factor A2 domain for proteolysis. 7:2096-2105.

38. Forman, J. R., and J. Clarke. Mechanical unfolding of proteins: insights into biology, structure and folding. 17:58-66.

39. Fowler, S. B., R. B. Best, J. L. T. Herrera, T. J. Rutherford, A. Steward, E. Paci, M. Karplus, and J. Clarke. Mechanical unfolding of a titin Ig domain: Structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering. 322:841-849.

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

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

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

43. West, D. K., D. J. Brockwell, P. D. Olmsted, S. E. Radford, and E. Paci. Mechanical resistance of proteins explained using simple molecular models. 90:287-297.

44. Taketomi, H., Y. Ueda, and N. Go. 1975. Studies on protein folding, unfolding and fluctuations by computer-simulation .1. effect of specific amino-acid sequence represented by specific inter-unit interactions. International Journal of Peptide and Protein Research 7:445-459.

45. Atilgan, A. R., S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophysical Journal 80:505-515.

46. Hinsen, K. 1998. Analysis of domain motions by approximate normal mode calculations. Proteins-Structure Function and Genetics 33:417-429.

47. Tirion, M. M. 1996. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Physical Review Letters 77:1905-1908.

48. Tama, F., W. Wriggers, and C. L. Brooks. 2002. Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. Journal of Molecular Biology 321:297-305.

49. Eyal, E., and I. Bahar. 2008. Toward a molecular understanding of the anisotropic response of proteins to external forces: Insights from elastic network models. 94:3424-3435.

50. Dietz, H., and M. Rief. 2008. Elastic bond network model for protein unfolding mechanics. Physical Review Letters 100:4.

51. Lu, H., B. Isralewitz, A. Krammer, V. Vogel, and K. Schulten. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. 75:662-671.

52. Li, P. C., and D. E. Makarov. 2004. Simulation of the mechanical unfolding of ubiquitin: Probing different unfolding reaction coordinates by changing the pulling geometry. Journal of Chemical Physics 121:4826-4832.

53. Hills, R. D., and C. L. Brooks. 2009. Insights from Coarse-Grained Go Models for Protein Folding and Dynamics. International Journal of Molecular Sciences 10:889-905.

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

55. Farrell, D. W., M. Lei, and M. F. Thorpe. 2011. Comparison of pathways from the geometric targeting method and targeted molecular dynamics in nitrogen regulatory protein C. Physical Biology 8.

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

57. Buseck, P., J. M. Cowley, and L. Eyring. 1988. High-resolution transmission electron microscopy and associated techniques. Oxford University Press, New York.

58. Henderson, R., and R. M. Glaeser. 1985. Quantitative-analysis of image-contrast in electron-micrographs of beam-sensitive crystals. Ultramicroscopy 16:139-150.

59. Henderson, R. 1992. Image-contrast in high-resolution electron-microscopy of biological macromolecules - tmv in ice. Ultramicroscopy 46:1-18.

60. Wriggers, W., and P. Chacon. 2001. Using Situs for the registration of protein structures with low-resolution bead models from X-ray solution scattering. Journal of Applied Crystallography 34:773-776.

61. Tama, F., O. Miyashita, and C. L. Brooks. 2004. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. Journal of Structural Biology 147:315-326.

62. Trabuco, L. G., E. Villa, K. Mitra, J. Frank, and K. Schulten. 2008. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. Structure 16:673-683.

63. Dubochet, J., M. Adrian, J. J. Chang, J. C. Homo, J. Lepault, A. W. McDowall, and P. Schultz. 1988. Cryo-electron microscopy of vitrified specimens. Quarterly Reviews of Biophysics 21:129-228.

64. Topf, M., K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali. 2008. Protein structure fitting and refinement guided by cryo-EM density. Structure 16:295-307.

65. Fabiola, F., and M. S. Chapman. 2005. Fitting of high-resolution structures into electron microscopy reconstruction images. Structure 13:389-400.

66. Taylor, J. R., and C. D. Zafiratos. 1991. Modern physics for scientists and engineers. Prentice Hall, Englewood Cliffs, N.J.

67. Rhodes, G. 2006. Crystallography made crystal clear : a guide for users of macromolecular models. Elsevier/Academic Press, Amsterdam ; Boston.

68. Terwilliger, T. C., and J. Berendzen. 1999. Automated MAD and MIR structure solution. Acta Crystallographica Section D-Biological Crystallography 55:849-861.

69. Ruska, E. 1987. The development of the electron-microscope and of electron-microscopy. Reviews of Modern Physics 59:627-638.

70. de Broglie, L. 1926. The possibility of interconnecting the phenomena of interference and diffraction with the theory of light quanta. Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences 183:447-448.

71. Hawkes, P. W. 1985. The Beginnings of electron microscopy. Academic Press, Orlando.

72. Kirkland, E. J. 1998. Advanced computing in electron microscopy. Plenum Press, New York.

73. Glaeser, R. M. 1985. Electron crystallography of biological macromolecules. Annual Review of Physical Chemistry 36:243-275.

74. Brenner, S., and R. W. Horne. 1959. A negative staining method for high resolution electron microscopy of viruses. Biochimica Et Biophysica Acta 34:103-110.

75. Unwin, P. N. T., and R. Henderson. 1975. Molecular-structure determination by electron-microscopy of unstained crystalline specimens. Journal of Molecular Biology 94:425-&.

76. Taylor, K. A., and R. M. Glaeser. 1974. Electron-diffraction of frozen, hydrated protein crystals. Science 186:1036-1037.

77. Taylor, K. A., and R. M. Glaeser. 1976. Electron-microscopy of frozen hydrated biological specimens. Journal of Ultrastructure Research 55:448-456.

78. Dubochet, J., M. Adrian, J. Lepault, and A. W. McDowall. 1985. Cryo-electron microscopy of vitrified biological specimens. Trends in Biochemical Sciences 10:143-146.

79. Kubler, O., M. Hahn, and J. Seredynski. 1978. Optical and digital spatial-frequency filtering of electron-micrographs .1. theoretical considerations. Optik 51:171-188.

80. Tverberg, H. 1966. A generalization of radons theorem. Journal of the London Mathematical Society 41:123-&.

81. Bracewell, R. N., and G. W. Preston. 1956. Radio reflection and refraction phenomena in the high solar corona. Astrophysical Journal 123:14-29.

82. Crowther, R. A., D. J. Derosier, and A. Klug. 1970. Reconstruction of 3 dimensional structure from projections and its application to electron microscopy. Proceedings of the Royal Society of London Series a-Mathematical and Physical Sciences 317:319-&.

83. Frank, J., W. Goldfarb, D. Eisenberg, and T. S. Baker. 1978. Reconstruction of glutamine-synthetase using computer averaging. Ultramicroscopy 3:283-290.

84. Radermacher, M., T. Wagenknecht, A. Verschoor, and J. Frank. 1987. 3-Dimensional structure of the large ribosomal-subunit from escherichia-coli. Embo Journal 6:1107-1114.

85. Wriggers, W., and S. Birmanns. 2001. Using Situs for flexible and rigid-body fitting of multiresolution single-molecule data. Journal of Structural Biology 133:193-202.

86. Uchino, T., J. D. Harrop, S. N. Taraskin, and S. R. Elliott. 2005. Real and reciprocal space structural correlations contributing to the first sharp diffraction peak in silica glass. Physical Review B 71.

87. Ediger, M. D., C. A. Angell, and S. R. Nagel. 1996. Supercooled liquids and glasses. Journal of Physical Chemistry 100:13200-13212.

88. Angell, C. A. 2002. Calorimetric studies of the energy landscapes of glassformers by hyperquenching methods. Journal of Thermal Analysis and Calorimetry 69:785-794.

89. Angell, C. A., K. L. Ngai, G. B. McKenna, P. F. McMillan, and S. W. Martin. 2000. Relaxation in glassforming liquids and amorphous solids. Journal of Applied Physics 88:3113-3157.

90. Angell, C. A. 2002. Liquid fragility and the glass transition in water and aqueous solutions. Chemical Reviews 102:2627-2649.

91. Weintraub, H., M. Ashburner, P. N. Goodfellow, H. F. Lodish, C. J. Arntzen, P. W. Anderson, T. M. Rice, T. H. Geballe, A. R. Means, H. M. Ranney, T. R. Cech, R. R. Colwell, H. R. Bourne, B. Richter, I. M. Singer, P. Marrack, D. T. Fearon, A. Penzias, A. J. Bard, W. F. Brinkman, P. A. Marks, B. Vogelstein, K. W. Kinzler, J. M. Bishop, R. N. Zare, G. Schatz, S. J. Benkovic, H. B. Gray, J. S. Valentine, P. J. Crutzen, D. W. Choi, S. Nakanishi, S. M. Kosslyn, J. I. Brauman, D. C. Rees, W. J. Brill, J. Schell, R. Luhrmann, C. L. Will, W. Wulf, G. J. Vermeij, K. J. Arrow, N. J. Smelser, D. L. Anderson, and P. H. Abelson. 1995. Through the glass lightly. Science 267:1609-1618.

92. Bohmer, R. 1998. Nanoscale heterogeneity of glass-forming liquids: experimental advances. Current Opinion in Solid State & Materials Science 3:378-385.

93. Sillescu, H. 1999. Heterogeneity at the glass transition: a review. Journal of Non-Crystalline Solids 243:81-108.

94. Billinge, S. J. L., and M. F. Thorpe. 2002. From semiconductors to proteins : beyond the average structure. Kluwer Academic/Plenum, New York.

95. Bell, R. J., and P. Dean. 1966. Properties of vitreous silica - analysis of random network models. Nature 212:1354-&.

96. Woodcock, L. V., C. A. Angell, and P. Cheeseman. 1976. Molecular-dynamics studies of vitreous state - simple ionic systems and silica. Journal of Chemical Physics 65:1565-1577.

97. Kendrew, J. C., G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. 1958. 3-Dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181:662-666.

98. Rahman, A. 1964. Correlations in motion of atoms in liquid argon. Physical Review a-General Physics 136:A405-&.

99. McCammon, J. A., B. R. Gelin, and M. Karplus. 1977. Dynamics of folded proteins. Nature 267:585-590.

100. Scheraga, H. A., M. Khalili, and A. Liwo. 2007. Protein-folding dynamics: Overview of molecular simulation techniques. Annual Review of Physical Chemistry 58:57-83.

101. Frenkel, D., and B. Smit. 2002. Understanding molecular simulation : from algorithms to applications. Academic, San Diego, Calif. ; London.

102. Brooks, B. R., C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. 2009. CHARMM: The Biomolecular Simulation Program. Journal of Computational Chemistry 30:1545-1614.

103. Ponder, J. W., and D. A. Case. 2003. Force fields for protein simulations. Protein Simulations 66:27-+.

104. MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. Journal of Physical Chemistry B 102:3586-3616.

105. Cramer, C. J., and D. G. Truhlar. 1999. Implicit solvation models: Equilibria, structure, spectra, and dynamics. Chemical Reviews 99:2161-2200.

106. Lazaridis, T., and M. Karplus. 1999. Effective energy function for proteins in solution. Proteins-Structure Function and Genetics 35:133-152.

107. Go, N., T. Noguti, and T. Nishikawa. 1983. Dynamics of a small globular protein in terms of low-frequency vibrational-modes. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences 80:3696-3700.

108. Brooks, B., and M. Karplus. 1983. Harmonic dynamics of proteins - normal-modes and fluctuations in bovine pancreatic trypsin-inhibitor. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences 80:6571-6575.

109. Case, D. A. 1994. Normal-mode analysis of protein dynamics. Current Opinion in Structural Biology 4:285-290.

110. Kitao, A., and N. Go. 1999. Investigating protein dynamics in collective coordinate space. Current Opinion in Structural Biology 9:164-169.

111. Amadei, A., A. B. M. Linssen, B. L. deGroot, D. M. F. vanAalten, and H. J. C. Berendsen. 1996. An efficient method for sampling the essential subspace of proteins. Journal of Biomolecular Structure & Dynamics 13:615-625.

112. Trudeau, R. J. 1993. Introduction to graph theory. Dover Pub., New York.

113. Thorpe, M. F., and P. M. Duxbury. 1999. Rigidity theory and applications / edited by M.F. Thorpe and P.M. Duxbury. Kluwer Academic/Plenum, New York.

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

115. Jacobs, D. J., and M. F. Thorpe. 1995. Generic rigidity percolation - the pebble game. Physical Review Letters 75:4051-4054.

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

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

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

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

120. Hespenheide, B. M., A. J. Rader, M. F. Thorpe, and L. A. Kuhn. 2002. Identifying protein folding cores from the evolution of flexible regions during unfolding. Journal of Molecular Graphics & Modelling 21:195-207.

121. Bartels, C., and M. Karplus. 1997. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. Journal of Computational Chemistry 18:1450-1462.

122. Guinier, A. 1963. X-ray diffraction in crystals, imperfect crystals, and amorphous bodies. W.H. Freeman, San Francisco,.

123. Lei, M., A. M. R. de Graff, M. F. Thorpe, S. A. Wells, and A. Sartbaeva. 2009. Uncovering the intrinsic geometry from the atomic pair distribution function of nanomaterials. Physical Review B 80.

124. Keen, D. A. 2001. A comparison of various commonly used correlation functions for describing total scattering. Journal of Applied Crystallography 34:172-177.

125. Kodama, K., S. Iikubo, T. Taguchi, and S. Shamoto. 2006. Finite size effects of nanoparticles on the atomic pair distribution functions. Acta Crystallographica Section A 62:444-453.

126. Gilbert, B. 2008. Finite size effects on the real-space pair distribution function of nanoparticles. Journal of Applied Crystallography 41:554-562.

127. Farrow, C. L., and S. J. L. Billinge. 2009. Relationship between the atomic pair distribution function and small-angle scattering: implications for modeling of nanoparticles. Acta Crystallographica Section A 65:232-239.

128. Wooten, F., and D. Weaire. 1987. Modeling tetrahedrally bonded random networks by computer. Solid State Physics-Advances in Research and Applications 40:1-&.

129. Hansen, J. P., and I. R. McDonald. 1986. Theory of simple liquids. Academic, London ; New York.

130. Torquato, S., and F. H. Stillinger. 2003. Local density fluctuations, hyperuniformity, and order metrics. Physical Review E 68:041113.

131. Bhatia, A. B., and D. E. Thornton. 1970. Structural aspects of the electrical resistivity of binary alloys. Physical Review B-Solid State 2:3004-3012.

132. Salmon, P. S. 2007. The structure of tetrahedral network glass forming systems at intermediate and extended length scales. Journal of Physics-Condensed Matter 19:455208.

133. Thorpe, M. F. 1983. Continuous deformations in random networks. Journal of Non-Crystalline Solids 57:355-370.

134. Sartbaeva, A., S. A. Wells, M. M. J. Treacy, and M. F. Thorpe. 2006. The flexibility window in zeolites. Nature Materials 5:962-965.

135. Levelut, C., A. Faivre, R. Le Parc, B. Champagnon, J. L. Hazemann, L. David, C. Rochas, and J. P. Simon. 2002. Influence of thermal aging on density fluctuations in oxide glasses measured by small-angle X-ray scattering. Journal of Non-Crystalline Solids 307:426-435.

136. Levelut, C., R. Le Parc, A. Faivre, R. Bruning, B. Champagnon, V. Martinez, J. P. Simon, F. Bley, and J. L. Hazemann. 2007. Density fluctuations in oxide glasses investigated by small-angle X-ray scattering. Journal of Applied Crystallography 40:S512-S516.

137. Lei, M., A. M. R. de Graff, M. F. Thorpe, S. A. Wells, and A. Sartbaeva. 2009. Uncovering the intrinsic geometry from the atomic pair distribution function of nanomaterials. Physical Review B 80:024118.

138. Bleher, P. M., F. J. Dyson, and J. L. Lebowitz. 1993. Non-gaussian energy-level statistics for some integrable systems. Physical Review Letters 71:3047-3050.

139. Faber, T. E., and J. M. Ziman. 1965. A theory of electrical properties of liquid metals .3. resistivity of binary alloys. Philosophical Magazine 11:153.

140. Fischer, H. E., A. C. Barnes, and P. S. Salmon. 2006. Neutron and x-ray diffraction studies of liquids and glasses. Reports on Progress in Physics 69:233-299.

141. Wright, A. C., R. A. Hulme, and R. N. Sinclair. 2005. A small angle neutron scattering study of long range density fluctuations in vitreous silica. Physics and Chemistry of Glasses 46:59-66.

142. Wright, A. C. 2008. Longer range order in single component network glasses? Physics and Chemistry of Glasses-European Journal of Glass Science and Technology Part B 49:103-117.

143. Salmon, P. S., A. C. Barnes, R. A. Martin, and G. J. Cuello. 2007. Structure of glassy GeO2. Journal of Physics-Condensed Matter 19:415110.

144. Keita, N. M., and S. Steinemann. 1978. Compressibility and structure factors at zero wave-vector of liquid aluminum-silicon alloys. Journal of Physics C-Solid State Physics 11:4635-4641.

145. Custer, J. S., M. O. Thompson, D. C. Jacobson, J. M. Poate, S. Roorda, W. C. Sinke, and F. Spaepen. 1994. Density of amorphous Si. Applied Physics Letters 64:437-439.

146. Grimaldi, M. G., P. Baeri, and M. A. Malvezzi. 1991. Melting temperature of unrelaxed amorphous-silicon. Physical Review B 44:1546-1553.

147. Donovan, E. P., F. Spaepen, J. M. Poate, and D. C. Jacobson. 1989. Homogeneous and interfacial heat releases in amorphous-silicon. Applied Physics Letters 55:1516-1518.

148. Gibson, J. M., M. M. J. Treacy, T. Sun, and N. J. Zaluzec. 2010. Substantial Crystalline Topology in Amorphous Silicon. Physical Review Letters 105.

149. Nakhmanson, S. M., P. M. Voyles, N. Mousseau, G. T. Barkema, and D. A. Drabold. 2001. Realistic models of paracrystalline silicon. Physical Review B 63:art. no.-235207.

150. Voyles, P. M., N. Zotov, S. M. Nakhmanson, D. A. Drabold, J. M. Gibson, M. M. J. Treacy, and P. Keblinski. 2001. Structure and physical properties of paracrystalline atomistic models of amorphous silicon. Journal of Applied Physics 90:4437-4451.

151. Bucaro, J. A., and H. D. Dardy. 1976. Equilibrium compressibility of glassy sio2 between transformation and melting temperature. Journal of Non-Crystalline Solids 20:149-151.

152. Geissberger, A. E., and F. L. Galeener. 1983. Raman studies of vitreous sio2 versus fictive temperature. Physical Review B 28:3266-3271.

153. Weinberg, D. L. 1963. X-ray scattering measurements of long range thermal density fluctuations in liquids. Physics Letters 7:324-325.

154. Zhang, X. H., K. Halvorsen, C. Z. Zhang, W. P. Wong, and T. A. Springer. 2009. Mechanoenzymatic Cleavage of the Ultralarge Vascular Protein von Willebrand Factor. Science 324:1330-1334.

155. Auton, M., M. A. Cruz, and J. Moake. 2007. Conformational stability and domain unfolding of the Von Willebrand factor A domains. Journal of Molecular Biology 366:986-1000.

156. Chen, W., J. Z. Lou, and C. Zhu. 2009. Molecular Dynamics Simulated Unfolding of von Willebrand Factor A Domains by Force. Cellular and Molecular Bioengineering 2:75-86.

157. Baldauf, C., R. Schneppenheim, W. Stacklies, T. Obser, A. Pieconka, S. Schneppenheim, U. Budde, J. Zhou, and F. Grater. 2009. Shear-induced unfolding activates von Willebrand factor A2 domain for proteolysis. Journal of Thrombosis and Haemostasis 7:2096-2105.

158. Vogel, V. 2006. Mechanotransduction involving multimodular proteins: Converting force into biochemical signals. Annual Review of Biophysics and Biomolecular Structure 35:459-488.

159. Forman, J. R., and J. Clarke. 2007. Mechanical unfolding of proteins: insights into biology, structure and folding. Current Opinion in Structural Biology 17:58-66.

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

161. Fowler, S. B., R. B. Best, J. L. T. Herrera, T. J. Rutherford, A. Steward, E. Paci, M. Karplus, and J. Clarke. 2002. Mechanical unfolding of a titin Ig domain: Structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering. Journal of Molecular Biology 322:841-849.

162. Best, R. B., S. B. Fowler, J. L. T. Herrera, A. Steward, E. Paci, and J. Clarke. 2003. Mechanical unfolding of a titin Ig domain: Structure of transition state revealed by combining atomic force microscopy, protein engineering and molecular dynamics simulations. Journal of Molecular Biology 330:867-877.

163. Gao, M., D. Craig, V. Vogel, and K. Schulten. 2002. Identifying unfolding intermediates of FN-III10 by steered molecular dynamics. Journal of Molecular Biology 323:939-950.

164. Lu, H., B. Isralewitz, A. Krammer, V. Vogel, and K. Schulten. 1998. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. Biophysical Journal 75:662-671.

165. West, D. K., D. J. Brockwell, P. D. Olmsted, S. E. Radford, and E. Paci. 2006. Mechanical resistance of proteins explained using simple molecular models. Biophysical Journal 90:287-297.

166. Mitternacht, S., S. Luccioli, A. Torcini, A. Imparato, and A. Irback. 2009. Changing the Mechanical Unfolding Pathway of FnIII(10) by Tuning the Pulling Strength. Biophysical Journal 96:429-441.

167. Eyal, E., and I. Bahar. 2008. Toward a molecular understanding of the anisotropic response of proteins to external forces: Insights from elastic network models. Biophysical Journal 94:3424-3435.

168. Farrell, D. W., T. Mamonova, M. Kurnikova, and M. F. Thorpe. In press.

169. Maragakis, P., A. van der Vaart, and M. Karplus. 2009. Gaussian-Mixture Umbrella Sampling. Journal of Physical Chemistry B 113:4664-4673.

170. Murcko, M. A., H. Castejon, and K. B. Wiberg. 1996. Carbon-carbon rotational barriers in butane, 1-butene, and 1,3-butadiene. Journal of Physical Chemistry 100:16162-16168.

171. Fernandez, A., and R. S. Berry. 2002. Extent of hydrogen-bond protection in folded proteins: A constraint on packing architectures. Biophysical Journal 83:2475-2481.

172. Improta, S., A. S. Politou, and A. Pastore. 1996. Immunoglobulin-like modules from titin I-band: Extensible components of muscle elasticity. Structure 4:323-337.

173. Leahy, D. J., I. Aukhil, and H. P. Erickson. 1996. 2.0 angstrom crystal structure of a four-domain segment of human fibronectin encompassing the RGD loop and synergy region. Cell 84:155-164.

174. Leahy, D. J., W. A. Hendrickson, I. Aukhil, and H. P. Erickson. 1992. Structure of a fibronectin type-III domain from tenascin phased by MAD analysis of the selenomethionyl protein. Science 258:987-991.

175. Jing, H., J. Takagi, J. H. Liu, S. Lindgren, R. G. Zhang, A. Joachimiak, J. H. Wang, and T. A. Springer. 2002. Archaeal surface layer proteins contain beta propeller, PKD, and beta helix domains and are related to metazoan cell surface proteins. Structure 10:1453-1464.

176. Fucini, P., C. Renner, C. Herberhold, A. A. Noegel, and T. A. Holak. 1997. The repeating segments of the F-actin cross-linking gelation factor (ABP-120) have an immunoglobulin-like fold. Nature Structural Biology 4:223-230.

177. Vijaykumar, S., C. E. Bugg, and W. J. Cook. 1987. Structure of ubiquitin refined at 1.8 A resolution. Journal of Molecular Biology 194:531-544.

178. O'Neill, J. W., D. E. Kim, D. Baker, and K. Y. J. Zhang. 2001. Structures of the B1 domain of protein L from Peptostreptococcus magnus with a tyrosine to tryptophan substitution. Acta Crystallographica Section D-Biological Crystallography 57:480-487.

179. Katayanagi, K., M. Miyagawa, M. Matsushima, M. Ishikawa, S. Kanaya, H. Nakamura, M. Ikehara, T. Matsuzaki, and K. Morikawa. 1992. Structural details of ribonuclease-h from escherichia-coli as refined to an atomic resolution. Journal of Molecular Biology 223:1029-1052.

180. Zhang, Q., Y. F. Zhou, C. Z. Zhang, X. H. Zhang, C. F. Lu, and T. A. Springer. 2009. Structural specializations of A2, a force-sensing domain in the ultralarge vascular protein von Willebrand factor. Proceedings of the National Academy of Sciences of the United States of America 106:9226-9231.

181. Buckle, A. M., K. Henrick, and A. R. Fersht. 1993. Crystal structural-analysis of mutations in the hydrophobic cores of barnase. Journal of Molecular Biology 234:847-860.

182. Pascual, J., M. Pfuhl, D. Walther, M. Saraste, and M. Nilges. 1997. Solution structure of the spectrin repeat: A left-handed antiparallel triple-helical coiled-coil. Journal of Molecular Biology 273:740-751.

183. Andersen, K. V., and F. M. Poulsen. 1993. The 3-dimensional structure of acyl-coenzyme a binding-protein from bovine liver - structural refinement using heteronuclear multidimensional nmr-spectroscopy. Journal of Biomolecular Nmr 3:271-284.

184. Neria, E., S. Fischer, and M. Karplus. 1996. Simulation of activation free energies in molecular systems. Journal of Chemical Physics 105:1902-1921.

185. Paci, E., and M. Karplus. 2000. Unfolding proteins by external forces and temperature: The importance of topology and energetics. Proceedings of the National Academy of Sciences of the United States of America 97:6521-6526.

186. Best, R. B., B. Li, A. Steward, V. Daggett, and J. Clarke. 2001. Can non-mechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation. Biophysical Journal 81:2344-2356.

187. Rief, M., M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub. 1997. Reversible unfolding of individual titin immunoglobulin domains by AFM. Science 276:1109-1112.

188. Klimov, D. K., and D. Thirumalai. 2000. Native topology determines force-induced unfolding pathways in globular proteins. Proceedings of the National Academy of Sciences of the United States of America 97:7254-7259.

189. Marszalek, P. E., H. Lu, H. B. Li, M. Carrion-Vazquez, A. F. Oberhauser, K. Schulten, and J. M. Fernandez. 1999. Mechanical unfolding intermediates in titin modules. Nature 402:100-103.

190. Li, L. W., H. H. L. Huang, C. L. Badilla, and J. M. Fernandez. 2005. Mechanical unfolding intermediates observed by single-molecule force spectroscopy in a fibronectin type III module. Journal of Molecular Biology 345:817-826.

191. Craig, D., A. Krammer, K. Schulten, and V. Vogel. 2001. Comparison of the early stages of forced unfolding for fibronectin type III modules. Proceedings of the National Academy of Sciences of the United States of America 98:5590-5595.

192. Schwaiger, I., A. Kardinal, M. Schleicher, A. A. Noegel, and M. Rief. 2004. A mechanical unfolding intermediate in an actin-crosslinking protein. Nature Structural & Molecular Biology 11:81-85.

193. Schlierf, M., H. B. Li, and J. M. Fernandez. 2004. The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques. Proceedings of the National Academy of Sciences of the United States of America 101:7299-7304.

194. Raschke, T. M., and S. Marqusee. 1997. The kinetic folding intermediate of ribonuclease H resembles the acid molten globule and partially unfolded molecules detected under native conditions. Nature Structural Biology 4:298-304.

195. Chamberlain, A. K., T. M. Handel, and S. Marqusee. 1996. Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH. Nature Structural Biology 3:782-787.

196. Faradjian, A. K., and R. Elber. 2004. Computing time scales from reaction coordinates by milestoning. Journal of Chemical Physics 120:10880-10889.

197. Ranson, N. A., D. K. Clare, G. W. Farr, D. Houldershaw, A. L. Horwich, and H. R. Saibil. 2006. Allosteric signaling of ATP hydrolysis in GroEL-GroES complexes. Nature Structural & Molecular Biology 13:147-152.

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

199. Birmanns, S., and W. Wriggers. 2007. Multi-resolution anchor-point registration of biomolecular assemblies and their components. Journal of Structural Biology 157:271-280.

200. Rusu, M., S. Birmanns, and W. Wriggers. 2008. Biomolecular pleiomorphism probed by spatial interpolation of coarse models. Bioinformatics 24:2460-2466.

201. Murshudov, G. N., A. A. Vagin, and E. J. Dodson. 1997. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallographica Section D-Biological Crystallography 53:240-255.

202. Schroder, G. F., A. T. Brunger, and M. Levitt. 2007. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. Structure 15:1630-1641.

203. deGroot, B. L., D. M. F. vanAalten, R. M. Scheek, A. Amadei, G. Vriend, and H. J. C. Berendsen. 1997. Prediction of protein conformational freedom from distance constraints. Proteins-Structure Function and Genetics 29:240-251.

204. Topf, M., M. L. Baker, M. A. Marti-Renom, W. Chiu, and A. Sali. 2006. Refinement of protein structures by iterative comparative modeling and cryoEM density fitting. Journal of Molecular Biology 357:1655-1668.

205. Fiser, A., and A. Sali. 2003. MODELLER: Generation and refinement of homology-based protein structure models. Macromolecular Crystallography, Pt D 374:461-+.

206. DiMaio, F., M. D. Tyka, M. L. Baker, W. Chiu, and D. Baker. 2009. Refinement of Protein Structures into Low-Resolution Density Maps Using Rosetta. Journal of Molecular Biology 392:181-190.

207. Rohl, C. A., C. E. M. Strauss, K. M. S. Misura, and D. Baker. 2004. Protein structure prediction using rosetta. Numerical Computer Methods, Pt D 383:66-+.

208. Gao, J. M., D. A. Bosco, E. T. Powers, and J. W. Kelly. 2009. Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. Nature Structural & Molecular Biology 16:684-U681.

209. Muller, J. J., S. Hansen, and H. V. Purschel. 1996. The use of small-angle scattering and the maximum-entropy method for shape-model determination from distance-distribution functions. Journal of Applied Crystallography 29:547-554.

210. Pickart, C. M., and R. E. Cohen. 2004. Proteasomes and their kin: Proteases in the machine age. Nature Reviews Molecular Cell Biology 5:177-187.

211. Carrion-Vazquez, M., H. B. Li, H. Lu, P. E. Marszalek, A. F. Oberhauser, and J. M. Fernandez. 2003. The mechanical stability of ubiquitin is linkage dependent. Nature Structural Biology 10:738-743.

212. Schmitt, T. J., J. E. Clark, and T. A. Knotts. 2009. Thermal and mechanical multistate folding of ribonuclease H. Journal of Chemical Physics 131:9.

213. Clementi, C., H. Nymeyer, and J. N. Onuchic. 2000. Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. Journal of Molecular Biology 298:937-953.

214. Kouza, M., C. K. Hu, H. Zung, and M. S. Li. 2009. Protein mechanical unfolding: Importance of non-native interactions. Journal of Chemical Physics 131.

215. Li, M. S., A. M. Gabovich, and A. I. Voitenko. 2008. New method for deciphering free energy landscape of three-state proteins. Journal of Chemical Physics 129.

216. Li, M. S., and M. Kouza. 2009. Dependence of protein mechanical unfolding pathways on pulling speeds. Journal of Chemical Physics 130.

217. Altmann, S. M., R. G. Grunberg, P. F. Lenne, J. Ylanne, A. Raae, K. Herbert, M. Saraste, M. Nilges, and J. K. H. Horber. 2002. Pathways and intermediates in forced unfolding of spectrin repeats. Structure 10:1085-1096.

218. Forman, J. R., S. Qamar, E. Paci, R. N. Sandford, and J. Clarke. 2005. The remarkable mechanical strength of polycystin-1 supports a direct role in mechanotransduction. Journal of Molecular Biology 349:861-871.

219. Forman, J. R., Z. T. Yew, S. Qamar, R. N. Sandford, E. Paci, and J. Clarke. 2009. Non-Native Interactions Are Critical for Mechanical Strength in PKD Domains. Structure 17:1582-1590.

220. Baneyx, G., L. Baugh, and V. Vogel. 2002. Fibronectin extension and unfolding within cell matrix fibrils controlled by cytoskeletal tension. Proceedings of the National Academy of Sciences of the United States of America 99:5139-5143.

221. Oberhauser, A. F., P. E. Marszalek, H. P. Erickson, and J. M. Fernandez. 1998. The molecular elasticity of the extracellular matrix protein tenascin. Nature 393:181-185.

222. Brockwell, D. J., G. S. Beddard, E. Paci, D. K. West, P. D. Olmsted, D. A. Smith, and S. E. Radford. 2005. Mechanically unfolding the small, topologically simple protein L. Biophysical Journal 89:506-519.

223. West, D. K., P. D. Olmsted, and E. Paci. 2006. Mechanical unfolding revisited through a simple but realistic model. Journal of Chemical Physics 124:8.

224. Oberhauser, A. F., C. Badilla-Fernandez, M. Carrion-Vazquez, and J. M. Fernandez. 2002. The mechanical hierarchies of fibronectin observed with single-molecule AFM. Journal of Molecular Biology 319:433-447.

225. Rief, M., J. Pascual, M. Saraste, and H. E. Gaub. 1999. Single molecule force spectroscopy of spectrin repeats: Low unfolding forces in helix bundles. Journal of Molecular Biology 286:553-561.

226. Carrion-Vazquez, M., A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke, and J. M. Fernandez. 1999. Mechanical and chemical unfolding of a single protein: A comparison. Proceedings of the National Academy of Sciences of the United States of America 96:3694-3699.

227. Ying, J. Y., Y. C. Ling, L. A. Westfield, J. E. Sadler, and J. Y. Shao. 2010. Unfolding the A2 Domain of Von Willebrand Factor with the Optical Trap. Biophysical Journal 98:1685-1693.

APPENDIX A

RDF OF UNIFORM MEDIA

For a handful of simple geometrical shapes, the analytical forms of the RDFs of uniform continuous media $R^{u}(r)$ have been presented in the literature. For the sake of convenience these analytical expressions are listed here and some new expressions added. To save space, all those efforts that express RDFs in integral forms that need further numerical computations are not listed. In all the expressions below, the symbol ρ_o represents the three-dimensional density.

The RDF of single objects can be found by using the fact that the RDF is the average distribution seen by the units of density within it. Each unit of density observes the same three-dimensional density distribution as does the unit of density at the predefined origin, except that the distribution appears translated due to the difference in viewing locations. Averaging over the observed distributions is equivalent to finding the density-density autocorrelation of the object (126). The density-density autocorrelation $c(\mathbf{r})$is a three-dimensional ρ density distribution given by

$$c(\mathbf{r}) = \frac{1}{\rho_0 V}\int_{-\infty}^{\infty} \rho(\mathbf{R})\rho(\mathbf{R}+\mathbf{r})d^3R \tag{A.1}$$

where $\rho(\boldsymbol{r})$ is the three-dimensional density distribution of the object of interest and V is its volume. The autocorrelation is normalized here so as to have a maximum density of ρ_o at $c(\mathbf{0})$. Note that $c(\boldsymbol{r})$ is proportional to the probability of finding two units of density within the object at a separation $\boldsymbol{r}$. The RDF depends only on the magnitude of $\boldsymbol{r}$ and can be found by performing a spherical integration of $c(\boldsymbol{r})$ about the origin. For objects of uniform density, $R^{u}(r) = 4\pi r^2\rho_o\alpha(r)$, allowing $\alpha(r)$ to be found directly from the spherical average of $c(\boldsymbol{r})$.

Applying Eq. (A.1) to a single uniform sphere of density ρ_o and radius a, and dividing by $4\pi r^2$ produces the shape factor (209)

$$\alpha_{sphere}(r) = \begin{cases} \left(1 - \frac{r}{2a}\right)^2 \left(1 + \frac{r}{4a}\right) & r < 2a \\ 0 & r > 2a \end{cases} \tag{A.2}$$

Similarly, the shape factor of a single uniform infinitely wide film of thickness *d* has the form (125)

$$\alpha_{film}(r) = \begin{cases} 1 - \frac{r}{2d} & r < d \\ \frac{d}{2r} & r > d \end{cases} \tag{A.3}$$

The advantage of using the density-density autocorrelation to obtain the RDF over the method used by Kodama *et al.* (125) can be seen, for example, in an infinitely long cylinder of radius a. Calculating the autocorrelation with the proper normalization, one obtains a three-dimensional distribution with cylindrical symmetry and a radial dependence given by

$$c(p) = \begin{cases} \frac{2}{\pi}\left[\sin^{-1}\sqrt{1 - \left(\frac{p}{2a}\right)^2} - \frac{p}{2a}\sqrt{1 - \left(\frac{p}{2a}\right)^2} \right] & r < 2a \\ 0 & r > 2a \end{cases} \tag{A.4}$$

where *p* is the distance from the axis of symmetry. By choosing a point along the axis as the center for the spherical averaging, p can be expressed as $p = r\sin\theta$, where θ is the angle between ***r*** and the axis of the cylinder. The spherical average can be expressed as

$$\begin{aligned} R^u_{cyl}(r) = 8r^2\rho_0 \int_0^{\theta_{\max}} \sin\,\theta \sin^{-1}\sqrt{1 - \left(\frac{r}{2a}\right)^2 \sin^2\,\theta}\, d\theta \\ - \frac{4r^3\rho_0}{a}\int_0^{\theta_{\max}} \sin^2\,\theta \sqrt{1 - \left(\frac{r}{2a}\right)^2 \sin^2\,\theta}\, d\theta, \end{aligned} \tag{A.5}$$

where θ_{max} is the angle at which *p* is maximal for a given *r* while remaining within the region $p < 2a$ where the effective density is larger than zero. For $r < 2a$, $\theta_{max} = \pi/2$, otherwise $\theta_{max} = \sin^{-1}(2a/r)$. By applying integration by parts to the first term, it becomes

$$8r^2\rho_0\left[\frac{\pi}{2}-\frac{r}{2a}\int_0^{\theta_{\max}}\frac{1-\sin^2\theta}{\sqrt{1-\left(\frac{r}{2a}\right)^2\sin^2\theta}}d\theta\right] \tag{A.6}$$

Substituting Eq. (A.6) into Eq. (A.5), the RDF of an infinite cylinder can be expressed as the sum of elliptical integrals, namely

$$\begin{aligned} R^u_{cyl}(r) = 4\pi r^2\rho_0\Bigg\{1-\frac{8a}{3\pi r}\left[1+\left(\frac{r}{2a}\right)^2\right]E\left(\theta_{\max},\frac{r}{2a}\right) \\ +\frac{8a}{3\pi r}\left[1-\left(\frac{r}{2a}\right)^2\right]F\left(\theta_{\max},\frac{r}{2a}\right)\Bigg\}, \end{aligned} \tag{A.7}$$

where

$$E(\phi,k)=\int_0^{\phi}\sqrt{1-k^2\sin^2\theta}\,d\theta \tag{A.8}$$

$$F(\phi,k)=\int_0^{\phi}\frac{d\theta}{\sqrt{1-k^2\sin^2\theta}} \tag{A.9}$$

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

For $r < 2a$, $\theta_{max} = \pi/2$ and thus $F(\pi/2, r/2a) = K(r/2a)$. For $r > 2a$, the substitution $(r/2a)\sin\theta = \sin\theta'$ allows $F(\sin^{-1}(2a/r)\,, r/2a) = K(r/2a)$ to be written as $(2a/r)K(2a/r)$. The shape factor $\alpha_{cyl}(r) = R^u_{cyl}(r)/(4\pi r^2\rho_o)$ for an infinite cylinder of radius *a* thus has the form

$$\alpha_{cyl}(r)=\begin{cases} 1-\frac{8a}{3\pi r}\left(1+\frac{r^2}{4a^2}\right)E\left(\frac{\pi}{2},\frac{r}{2a}\right)+\frac{8a}{3\pi r}\left(1-\frac{r^2}{4a^2}\right)K\left(\frac{r}{2a}\right) & r<2a \\ 1-\frac{8a}{3\pi r}\left(1+\frac{r^2}{4a^2}\right)E\left[\sin^{-1}\left(\frac{2a}{r}\right),\frac{r}{2a}\right]+\frac{16a^2}{3\pi r^2}\left(1-\frac{r^2}{4a^2}\right)K\left(\frac{2a}{r}\right) & r>2a \end{cases} \tag{A.11}$$

To the best of my knowledge, the RDF of an infinite cylinder has never been expressed in such a simplified form. The power of the autocorrelation method can be seen by comparing Eq. (A.11) to the equivalent result in Kodama *et al.* For a prolate spheroid whose three axes are *a*, *a*, and *av*, respectively, with *v* ≥ 1, the shape factor has the form (126)

$$\alpha_{prolate}(r) = \begin{cases} 1 - \frac{3r}{8av}\left(1 - \frac{r^2}{16a^2}\frac{2/3 + v^2}{v^2}\right) - \frac{3r}{8a}\left(1 + \frac{r}{4a}\right)\left(1 - \frac{r}{4a}\right)\frac{v}{\sqrt{v^2 - 1}}\tan^{-1}\sqrt{v^2 - 1} & 0 \le r \le 2a \\ 1 - \frac{3r}{8av}\left(1 - \frac{r^2}{16a^2}\frac{2/3 + v^2}{v^2}\right) - \frac{3}{8}\left(1 + \frac{r^2}{8a^2}\right)\sqrt{1 - \frac{4a^2}{r^2}}\frac{v}{\sqrt{v^2 - 1}} & \\ \quad - \frac{3r}{8a}\left(1 + \frac{r}{4a}\right)\left(1 - \frac{r}{4a}\right)\frac{v}{\sqrt{v^2 - 1}}\left(\tan^{-1}\sqrt{v^2 - 1} - \tan^{-1}\sqrt{\frac{r^2}{4R^2} - 1}\right) & 2a \le r \le 2av \end{cases}$$

(A.12)

For an oblate spheroid whose three axes are a, a, and av, respectively, with $v \le 1$, the shape factor has the form (126)

$$\alpha_{oblate}(r) = \begin{cases} \left[1 - \frac{3r}{8av}\left(1 - \frac{r^2}{16a^2}\frac{2/3 + v^2}{v^2}\right) - \frac{3r}{8a}\left(1 + \frac{r}{4a}\right)\left(1 - \frac{r}{4a}\right)\frac{v}{\sqrt{1 - v^2}}\tanh^{-1}\sqrt{1 - v^2}\right] & 0 \le r \le 2av \\ \frac{v}{\sqrt{1 - v^2}}\left[\frac{3a}{4r}\left(1 + \frac{r^2}{8a^2}\right)\sqrt{1 - \frac{r^2}{4a^2}} - \frac{3r}{8a}\left(1 + \frac{r}{4a}\right)\left(1 - \frac{r}{4a}\right)\tanh^{-1}\sqrt{1 - \frac{r^2}{4a^2}}\right] & 2av \le r \le 2a \end{cases}$$

(A.13)

Lastly, for a spherical shell of radius a and thickness δ, the shape factor in the range $0 \le r \le 2a + \delta$ has the form (94, 126)

$$\begin{aligned} &\alpha_{shell}(r) = \\ &\quad \frac{\pi r \rho_0}{2(12a^2 + \delta^2)}\{r[16a^3 + 12a\delta(\delta - r) + 36a^2(2\delta - r) + 3(\delta - r)^2(2\delta + r)] + 2(\delta - r)^2[r(2\delta + r) - 12a^2]\mathrm{sg}(\delta - r) \\ &\quad - 2(2a - r)^2[r(4a + r) - 3\delta^2]\mathrm{sg}(2a - r) + r(4a - 2\delta + r)(2a - \delta - r)^2\mathrm{sg}(2a - \delta - r)\} \end{aligned}$$

(A.14)

where $sg(x) = 1$ if $x < 0$.

APPENDIX B

PROTEIN UNFOLDING UNDER FORCE:

Comparison to constant force MD simulation

This section describes unfolding results for the remaining proteins in the constraint-based unfolding study. Four of the proteins do not have accompanying flow diagrams, as the pathways of ACA and spectrin are too diverse and lack distinct recurring states to be well described by flow diagrams, while PKD and protein L lack flow diagrams because the transition of interest for each is close to the native state structure and both lack distinct states further along their pathways.

Ubiquitin

Ubiquitin is a highly conserved regulatory protein found in all eukaryotes. It is commonly used to label a protein for proteasomal degradation in which one or more ubiquitin domains are covalently attached to the protein being labeled. Its resistance to force may play an important role in proteasomal substrate unfolding (210). Ubiquitin's mechanical properties have been studied using AFM (193), AFM and steered MD (211) and MD using umbrella sampling (52). Schlierf *et al.* (193) observed an intermediate at an extension 81±7 Å beyond that of the native state and hypothesized that it was due to the unfolding of the C-terminal half of the protein. Contrary to this hypothesis, a study by Li *et al.* (52) using umbrella sampling led to the conclusion that unfolding begins from the N-terminus through the unfolding of β_A and β_B.

All 10 constraint-based pathways begin with the shearing apart of the parallel β-sheet connecting the two ends of ubiquitin (Figure B.1). In all pathways, β_A and β_B then separated from the core helix and separated from one another to form a set of mechanically robust states, as determined from the larger breaking forces observed between 110 Å and 140 Å in Figure B.2. The extension of highest force corresponding to the unfolding of the native state and the state lacking β_A and β_B were found in the

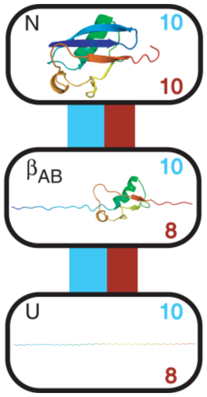


Figure B.1 Unfolding pathways of ubiquitin. Boxes serve as check points, connected by lines; colored blue (left) and red (right) for constraint-based and MD pathways respectively, and have thicknesses proportional to the number of paths that transit between the two end states. The numbers at the upper right and lower right of each box indicate the number of incoming constraint-based and MD pathways respectively.

constraint-based model to be separated by an average of 78±11 Å, in agreement with the value of 81±7 Å observed experimentally by Schlierf *et al.* Interestingly, despite the extension being indistinguishable from that of Schlierf *et al.*, the structure of the partially unfolded state differs from their hypothesized intermediate and instead lends support to the conclusions of Li *et al.* The results from constant force MD trajectories were less clear than those from the constraint-based method, as ubiquitin is seen to unfold completely as a single event without a stable intermediate, although β_A and β_B do separate from the core shortly before the C-terminal strands during the sudden unfolding events and therefore represents the same pathway.

Ribonuclease H

Ribonuclease H, or simply RNase H, is a non-specific endonuclease that cleaves RNA by a hydrolytic mechanism. In DNA replication, RNase H also removes the RNA primer to allow DNA synthesis to be completed. Extensive work has been done on RNase H, including bulk studies that have found that the most stable region and the first to fold consists of α_{1-4} and β_{4-5}. In a study by Cecconi *et al.* (160) using optical tweezers, an intermediate was observed and it was deduced that the region that remains folded in the

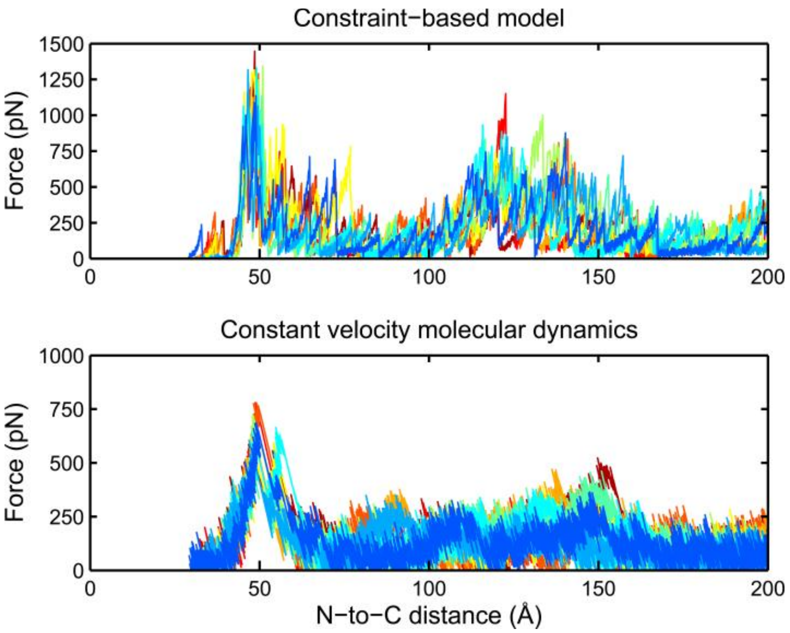


Figure B.2 Force profiles for the mechanical unfolding of ubiquitin obtained from crack propagation of the constraint network of 10 starting structures compared with those from constant velocity MD simulations.

intermediate is the same as the stable core observed in folding experiments (194, 195). The unfolding and folding of RNase H were studied by Schmitt *et al.* (212) and Clementi *et al.* (213) respectively using Gö-like models.

In all constraint-based and MD unfolding simulations, unfolding began at the C-terminus through the detachment of the terminal strand from the core, followed by the unraveling of the terminal α_5. At this point the pathways diversify, with half of the constraint-based pathways predicting the β-sheet to break between strands β_4 and β_5 while the other half separate between strands β_3 and β_4. This proved to be a critical step, as strands β_4 and β_5 were observed to serve as a clamp protecting the core helices from the external force. In most cases where strands β_4 and β_5 separated prior to β_3 and β_4, the core helices proceeded to unfold completely, leaving $\beta_{(1\text{-}3)}$ (the bracketed numbers signify that they are intact) as the last secondary structure to disappear. When strands β_3 and β_4 separate before β_4 and β_5, in 4 out of the 5 cases the N-terminal β-sheet proceeded to

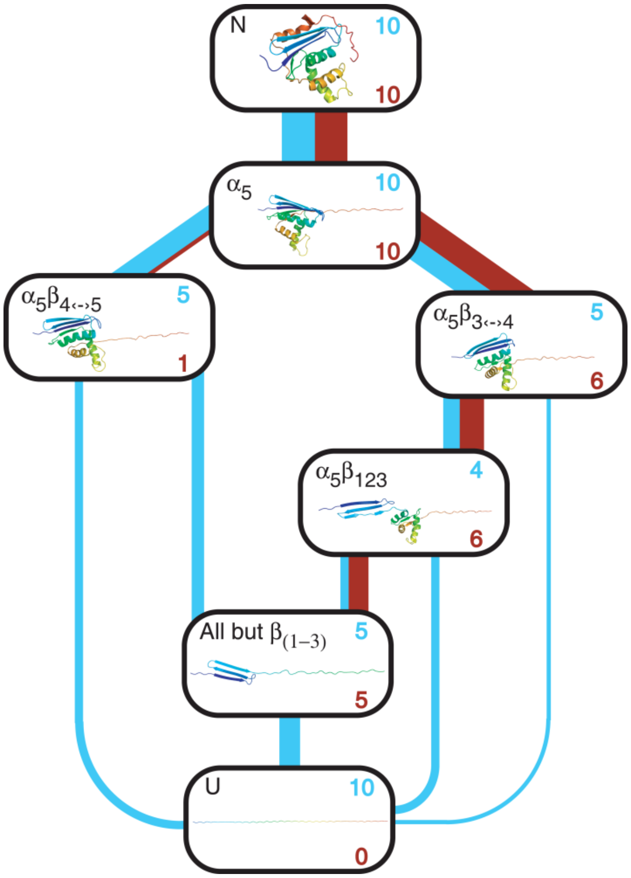


Figure B.3 (Color online) Unfolding pathways of ribonuclease H. Boxes serve as check points, connected by lines; colored blue (left) and red (right) for constraint-based and MD pathways respectively, and have thicknesses proportional to the number of paths that transit between the two end states. The numbers at the upper right and lower right of each box indicate the number of incoming constraint-based and MD pathways respectively.

separate (either together or individually), leaving the stable core observed in bulk studies (194, 195). At the level of detail described in Figure B.3, the set of pathways found by the constraint-based method included the MD trajectories as a subset. All but one MD trajectory followed the same unfolding pathway, leading to a state in which α_5 is unfolded and $\beta_{(1\text{-}3)}$ is detached, leaving the core observed in bulk studies, but in none of the pathways did $\beta_{(1\text{-}3)}$ unfold prior to the core due to the high stability of $\beta_{(1\text{-}3)}$ on the timescales simulated. Due to the ability of multiple constraint-based pathways to reproduce the order of detachment observed in the majority of MD simulations, as well as the observation in both models of an alternative pathway in which strands β_4 and β_5

separated prior to β_3 and β_4 leading most often to the unfolding of the α-helical core, we consider the constraint-based model to have adequately captured the unfolding behavior from MD simulation.

Acyl-CoA binding protein

Acyl-CoA binding protein, herein called ACA, is a small helical protein that binds acyl-CoA esters with high affinity. To the authors' knowledge, the mechanical unfolding of ACA has not been experimentally investigated, but has been studied using MD by Paci *et al.* (185).

In the constraint-based pathways, the first two events are always the detachment of α_A and α_D from the core, occurring simultaneously as well as in an ordered fashion, with no significant unfolding of the helices themselves. The remaining core was found to be resistant to force, causing varying amounts of unfolding of α_A and α_D prior to the unfolding of the core. These features agree with those from the MD simulations, for which α_A and α_D detach first, typically doing so simultaneously. There was usually, but not always, little loss of secondary structure of the terminal helices before detachment. A core resistant to unfolding formed for several of the simulations, composed of approximately the same residues as those from the constraint-based pathways and the study of Paci *et al.* (185).

Filamin

Filamin is part of the cytoskeleton and is subject to force as part of its physiological role. It consists of long chains of actin-binding modules separated by varying numbers of immunoglobulin rod domains. This study uses the 4th filamin domain of *Dictostelium descoideum* (ddFLN4). In an AFM study, Schwaiger *et al.* (192) inserted segments of glycine residues multiple loop regions to probe which strands unfold to form

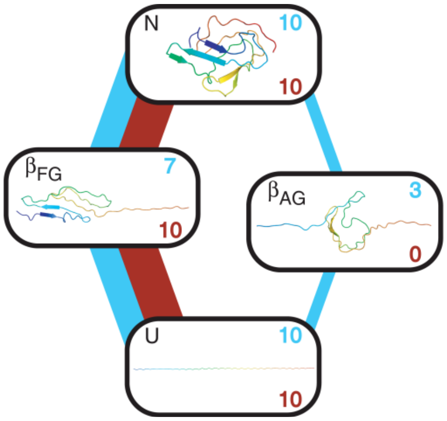


Figure B.4 Unfolding pathways of filamin. Boxes serve as check points, connected by lines; colored blue (left) and red (right) for constraint-based and MD pathways respectively, and have thicknesses proportional to the number of paths that transit between the two end states. The numbers at the upper right and lower right of each box indicate the number of incoming constraint-based and MD pathways respectively.

the observed intermediate, concluding that unfolding is restricted to β_A and β_B. Theoretical modelling studies have recently been performed using MD simulation (214) as well as coarse-grained simulation (215, 216).

The constraint-based pathways unfolded through two routes, as displayed in Figure B.4. In the majority of the pathways, unfolding began entirely at the C-terminus as β_F and β_G unfolded, leaving a mechanically stable state. In the other pathways, unfolding occurred at both termini through the detachment of β_A and β_G prior to the unfolding of the remaining core. All 10 MD trajectories have β_F and β_G detaching first, consistent with the dominant constraint-based unfolding pathway as well as that observed in a previous study (215). Unlike all other proteins in this study, neither the constraint-based model nor the MD simulations capture the unfolding pathway leading to the stable intermediate observed by Schwaiger *et al.* (192). Interestingly, the study of Li *et al.* (216) suggests that the pulling velocity determines the end at which unfolding begins, switching to the N-terminus and passing through the experimentally observed intermediate at very low pulling rates.

Spectrin

Spectrin domains are triple-helical coiled-coil units located within many protein filaments that frequently bear a mechanical load and may function as elastic elements. This study pertains to the 16th repeat of α-spectrin. The combined AFM and MD study by Altmann *et al.* (217) found that the spectrin unfolding pathways, unlike those of proteins such as titin I27 and fibronectin, possess a broad range of extensions in which it maintains its native fold. Despite the MD simulations of Altmann *et al.* (217) resulting in a diverse set of unfolding pathways, mutation analysis allowed them to conclude that the experimentally probed pathways involve the kinking of the central helix.

The constraint-based pathways display great variability in the amount in which α_A and α_C unfold prior to the loss of the native state packing of the helices, consistent with the findings of Altmann *et al.* In the majority of the pathways, the central helix α_B kinks in the middle prior to the detachment of both α_A and α_C. These features agree with those of the MD simulations, for which the initial effect of force was to cause the terminal helices to unravel by amounts that varied greatly among the simulations. As with the constraint-based model, α_B sometimes developed a kink in it before, and often during, the simultaneous detachment of the terminal helices.

PKD

The mechanical properties associated to two PKD domain structures have been predominantly used in previous MD studies, that of the 1st PKD domain of human polycystin-1 and the archaeal PKD domain from *Methanosarcina Mazei*. In anticipation of possible future mutation studies, the structure of the archaeal PKD domain was used because the human structure has been found experimentally to be only marginally thermodynamically stable (218); which is a strategy that has been adopted previously (219). MD simulations have suggested that PKD's remarkable strength may be due to non-native contacts between the A-B loop and the G-strand that forms when the domain is subjected to force.

In the constraint-based model, the application of force causes the A-A' loop to be drawn towards the G strand, but fails to approach close enough to the G strand to form the non-native hydrogen bonds observed in previous studies (218, 219). The clamp between the $\beta_{A'}$ and β_G therefore does not grow prior to the shearing apart of the two halves of the domain. Upon shearing, unfolding was observed to proceed from either terminus. The results of the MD simulations are not dissimilar to those reported by Forman *et al.* (219), with the A-A' loop being pulled towards the G-strand resulting in the formation of non-native contacts and a structure referred to previously as S2 (219).

Tenascin

The third fibronectin type III domain of human tenascin, abbreviated TNfn3, has a β-sandwich fold and forms part of the extracellular matrix. The mechanical unfolding of TNfn3 has been studied both with AFM (220, 221) and computationally using coarse-grained models (165). A comprehensive study using protein engineering, AFM and MD simulation was also conducted by Ng *et al.* (54). Intermediates have not been observed experimentally, but ϕ-value analyses of mutants suggest that a significant amount of the protein is in a non-native conformation at the transition state (54), the greatest rearrangement taking place in β_A and β_G.

In all constraint-based pathways, tenascin formed intermediate I_1 characterized by the loss of several hydrogen bonds between β_A and β_G near the N-terminus that allows β_A to extend further without significant rotation and rearrangement of the core (Figure B.5). Many hydrophobic contacts in the core proceed to break and non-native contacts form as the two β-sheets rotate until the strands of each are roughly parallel, forming I_2. From here, strands β_A and β_G unfold in 9 of the pathways, followed by the complete

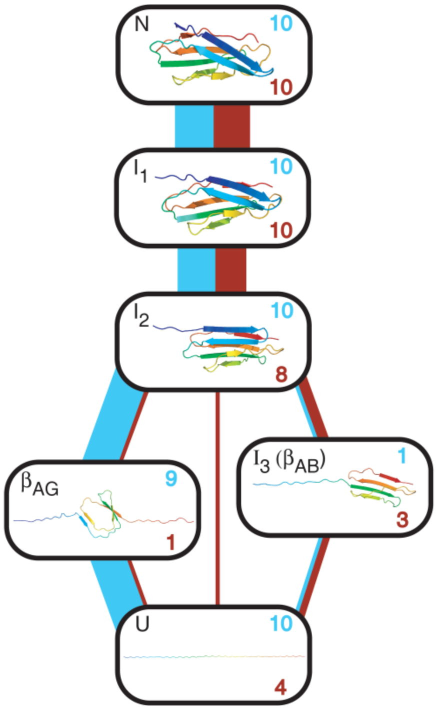


Figure B.5 Unfolding pathways of tenascin. Boxes serve as check points, connected by lines; colored blue (left) and red (right) for constraint-based and MD pathways respectively, and have thicknesses proportional to the number of paths that transit between the two end states. The numbers at the upper right and lower right of each box indicate the number of incoming constraint-based and MD pathways respectively. Check points possessing the additional label "I" have been identified as intermediates in previous studies.

unfolding of tenascin. In the remaining path, β_A and β_B unfold to form the intermediate I_3. The constraint-based unfolding pathways correspond well with those from MD, with the major distinction being the frequency at which I_3 is formed. It should be stressed that no experiment has ever detected I_3, in contrast to fibronectin (190). It is interesting that this asymmetry in the experimental detection of I_3 in the structurally homologous proteins tenascin and fibronectin is captured in the constraint-based model. We performed both constraint-based and constant force MD simulations on a conservative Ile8 → Ala8 mutant which has been experimentally found to weaken the molecule (54). This mutation did not significantly alter the constraint-based or MD pathways.

Protein L

The B1 domain of Protein L is expressed in *P. magnus* as a tandem domain located in its cell walls. Protein L has no known mechanical function, but its parallel β-sheet structure suggests a high mechanical resistance. Its response to force was studied both by AFM and through MD simulation by Brockwell *et al.* (222), as well as through a Gö-like model by West *et al.* (223).

In all 10 constraint-based pathways were consistent with previous findings (222), with the major force peaks correspond to the shearing apart of β_A and β_D followed by the sequential detachment firstly of β_C and β_D and secondly of β_A and β_B. In the MD simulations, unfolding was a single sudden event; however β_C and β_D tended to detach slightly before β_A and β_B, consistent with that found by Brockwell *et al.*

Comparison of unfolding forces

The choice of parameters for the constraint-based model was made to maximize agreement between the model's unfolding pathways and those from constant force MD. The model can nevertheless distinguish mechanically strong conformations from those that are less effective at supporting a load, as shown by comparisons of the force profiles for von Willebrand factor, fibronectin, and ubiquitin. To test the model's ability to capture differences in mechanical strength *between* proteins, the maximum forces applied to the termini along the entire unfolding pathways of the constraint-based and constant velocity MD simulations are compared in Figure B.6. Each force value represents an average of the maximum force along each of the 10 pathways. The maximum forces from the two techniques correlate strongly (correlation coefficient of 0.82) despite the unfolding force not being a consideration during parameter optimization. Experimentally measured unfolding forces are incorporated into the figure by assigning point sizes according to experimental forces, as summarized in Table B.1. It should be noted that the

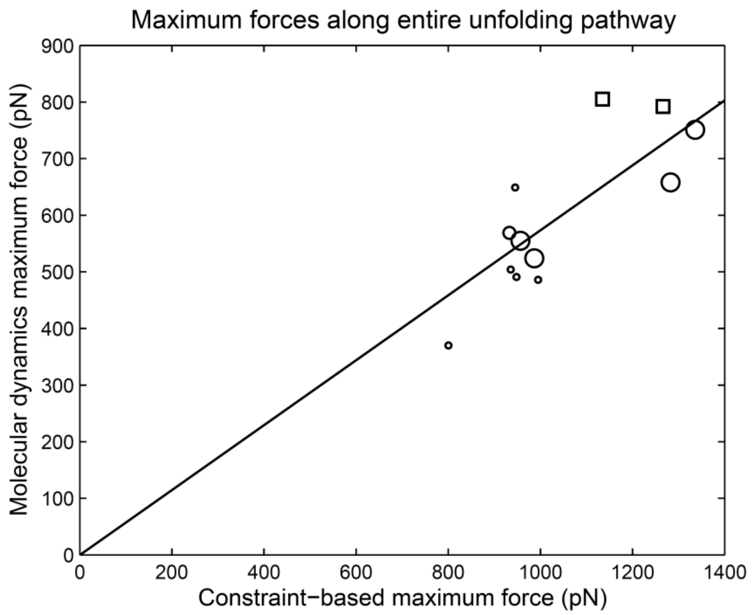


Figure B.6 Comparison of the maximum applied force along the unfolding pathways for the constraint-based model and constant velocity MD simulation. The size of the circular points reflects the unfolding forces observed experimentally by AFM, with small, medium, and large circles representing forces F < 75 pN, 75 pN < F < 150 pN, and F > 150 pN respectively. Experimental forces measured by optical tweezers are represented by squares of fixed size.

Protein	Constraint-based Maximum Force (pN)	Constant Velocity MD Maximum Force (pN)	Experimental Unfolding Force (pN)	Pulling Speed (nm/s)	Reference
ACA	800	370	None known		
Barnase	948	491	70	100-500	(186)
Fibronectin	945	649	74 +/- 20	600	(190, 224)
Filamin	995	486	56.5 +/- 1.4	250-350	(192)
PKD	987	524	181, 183	300, 600	(218)
Protein L	1336	751	152 +/- 5	700	(222)
RNase H	1135	805	19*	10-1000	(160)
Spectrin	936	504	25-35	300	(225)
Tenascin	933	569	137 +/- 12	200–600	(221)
Titin I27	957	555	204 +/- 26	400–600	(226)
Ubiquitin	1283	658	203 +/- 35	400	(211)
vWF	1266	792	7-14*	0.35-350 pN/s(227)	

*Obtained by optical tweezers.

Table B.1 Unfolding forces from models and experiment.

experimentally measured strength of a protein is a different quantity than the mechanical strength defined as the gradient of an energy Hamiltonian, as the experimental strength is sensitive to the free energy of the unfolding transition state, the displacement of the transition state along the vector of applied force, and the rate at which force is applied.

Number of non-native constraints

Added constraints in the form of hydrogen bonds, salt bridges, and hydrophobic interactions are considered to be non-native if the constraint occurs between different pairs of atoms than those in the starting structure. Constraints that break and reform are not double counted. The average number of non-native constraints along each pathway are compared with the number of constraints in the starting structure in Table B.2.

Protein	Av. No. of Native Constraints	Av. No. of Non-native Constraints
ACA	89.8	91.5
Barnase	113.1	114.4
Fibronectin	83.8	79.4
Filamin	79.3	79.2
PKD	80.0	94.6
Protein L	71.3	59.8
RNase H	169.8	170.6
Spectrin	109.9	114.5
Tenascin	87.1	84.4
Titin I27	90	73.5
Ubiquitin	81.9	74.5
vWF	196.9	235.6

Table B.2 Number of non-native constraints compared with the number in the native state.

Computational Cost

The CPU time needed for a single constraint-based and MD pathway are compared in Table B.3. The CPU time for the constraint-based model scales quadratically with the number of residues for all protein in this study, as both the number of steps as well as the computational cost per step scales linearly with protein size.

Protein	CPU Time, MD (h)	CPU Time, Constraint Model (h)	No. of Steps, Constraint Model	No. of Residues
ACA	10.433	0.58	3200	86
Barnase	12.555	0.92	4100	110
10FNIII	13.81	0.7	3400	94
ddFLN4	9.453	0.68	3600	100
PKD	9.094	0.72	2900	83
Protein L	9.858	0.25	2200	64
RNase H	20.967	1.87	5800	155
Spectrin	15.254	0.74	3400	98
Tenascin	17.62	0.56	3200	90
Titin I27	11.686	0.56	3100	89
Ubiquitin	8.763	0.42	2700	76
vWF	24.976	2.78	6900	177

Table B.3 CPU time and number of steps for a constraint-based pathway compared to the CPU time for a constant force MD pathway.

Parameter sensitivity

Of the model parameters not fixed by experimental values, namely k_{hb}, k_{ph}, k_{sh}, k_{st}, k_{rmsd}, Ω, and x_{max}, only the dependence of the unfolding pathways on the values of k_{hb}, k_{ph}, and Ω were varied during optimization against MD results (summarized in Table B.4), as these three control the relative load that each bond can support before breaking. The choice of values of k_{sh}, k_{st}, x_{max}, and k_{rmsd} is described in Chapter 7. The value $\Omega = 0.5$ was chosen to ensure finite load bearing capacity of all hydrogen bonds and salt bridges.

		Relative side-chain hydrogen bond to hydrophic bond strength (k_{hb}/k_{ph})	
		15/5	30/5
Ω	1	7	7.5
	0.5	5.5	9
	0	5.5	9

Table B.4 Dependence of the number of proteins whose pathways agree with those from constant force MD as a function of the free parameters.