FLEXIBLE AND RIGID REGIONS IN PROTEINS

Donald J. Jacobs[1], Leslie A. Kuhn[2] and Michael F. Thorpe[1]

[1]Department of Physics and Astronomy
[2]Department of Biochemistry
Michigan State University
East Lansing, MI 48824, U.S.A.

INTRODUCTION

We represent the microstructure of a protein as a generic bar-joint truss framework, where the hard covalent forces and strong hydrogen bonds are modeled as distance constraints. The mechanical stability of the corresponding bar-joint network is then analyzed using graph theoretical techniques. The computer program for analyzing the rigidity of substructures within macromolecules[1] is referred to as FIRST (Floppy Inclusion and Rigid Substructure Topography). This program provides a real-time tool for evaluating the intrinsic flexibility within a protein by applying a new combinatorial constraint counting algorithm. Unlike many methods for parsing protein folds, this new approach gives exact mechanical properties of a protein structure (or other macromolecules) under a given set of distance constraints. These properties include; counting the number of independent degrees of freedom, locating overconstrained regions where internal strain arises, partitioning the protein structure into rigid clusters that are separated by flexible joints and identifying underconstrained regions where continuous deformations can take place.

This article proceeds as follows. We begin by discussing some basic notions about intrinsic flexibility in proteins. We then highlight the underlying physical assumptions of our approach, introducing concepts from graph rigidity and sketching the computational algorithm. Some example applications are given, including the FIRST analysis of a lysine-binding protein.

Intrinsic Flexibility of Proteins

Over seven thousand protein structures have been determined to date using X-ray crystallography[2]. Such crystallographic structures are deduced from sophisticated

Rigidity Theory and Applications
Edited by Thorpe and Duxbury, Kluwer Academic/Plenum Publishers, 1999

refinement analysis in conjunction with stereochemical modeling[3], and they represent the best average structure over many realizations of individual protein molecules making up the solid crystal. Frequently in crystallographic structural studies, independent crystal forms trap the protein in different conformational states, suggesting an intrinsic flexibility within the protein[4]. The study of protein motion in solution using NMR spectroscopy shows that many different conformational states can be explored by a protein in its natural environment[5]. The intrinsic flexibility of a protein is often manifested in conformational changes[6] that are responsible for the large-scale rearrangements of domains or the relative motion of smaller fragments within individual domains.

A collection of X-ray crystal structures for a protein in different conformations allows a limited number of structural comparisons, from which some information about the intrinsic flexibility can be deduced. Prediction of the flexibility of a protein structure that is known only in one conformation can sometimes be inferred from other, similar protein structures. Molecular dynamics (MD) simulations can also be used to observe the dynamical evolution of a given protein as it explores various conformational states having energies near that of its native state. In practice, simulations cannot be performed for long enough times to explore the full range of available conformations.

Some computational procedures have been developed[7–12] to characterize the intrinsic flexibility and rigidity within a protein. These procedures fall into two classes. One makes use of comparing different conformational states[7–9] while the other class deals with identifying protein domains or characterizing protein structure[10–12] for any given conformation. These methods are generally fast (ranging from hours to minutes of CPU time), and give considerable insight into the mechanical stability of a protein. Nevertheless, each method has its limitations. In a broad sense, the methods in the first class are limited to the diversity (and accuracy) of the conformational states that are available from experiment for comparison, whereas the methods in the second class are limited to the correlation between the selected empirical criteria such as packing density or structural protrusions. FIRST falls into the second class, but it has the advantage that it runs in a fraction of a second, and it is based on a systematic distance constraint approach where rigidity can be directly calculated.

Characterizing Mechanical Stability

What would the *ideal* scenario be for an automated computational procedure to determine flexibility and rigidity, or more specifically the *mechanical stability* of a protein? Besides being fast and efficient, the method should only require knowing the protein structure. The output should consist of a universal scheme for characterizing mechanical stability that gives a robust quantified description of the degree of stability throughout the protein.

Perhaps the ideal universal way to characterize mechanical stability of a protein is via a hierarchical scheme of substructuring, reflecting in a self-consistent way the interaction strengths between the identified substructures. Within this characterization, gross features of mechanical stability between different protein conformations should be the same with some local differences. Of course, when comparing experimentally observed protein structures in different conformations, differences in mechanical stability due to binding different substrates or to different packing within the crystal lattice can occur.

Developing such an ideal automated procedure is our goal. A distance constraint approach is used to characterize mechanical stability of a protein, where only knowing the underlying *connectivity* of the structure is required. Overconstrained regions, rigid substructures and underconstrained regions can be determined within a hierarchical scheme. Distance constraints are used to model the covalent bond forces (bond-stretching and bond-bending) as well as any selected set of hydrogen bonds. Selection of the imposed distance constraints for the hydrogen bonding is determined by introducing a cut-off criterion on the interaction strength. A hierarchical characterization of the mechanical stability can be constructed by changing the chosen cut-off.

Applications to Drug Design and Protein Engineering

Rigidity and structural stability in proteins is commonly associated with close packing of amino acids into modules or domains[6,10,11,13]. Multifunctional proteins are often built from several closely packed domains linked by flexible regions which can allow large scale, hinge-type conformational changes[6,14]. There can also be shear-type conformational changes within densely packed regions[6], which are more subtle to identify. Flexible regions in proteins are often associated with loop structures, which are thought to be functionally important. For example, as a means of regulating interactions between proteins, one region in the CksHs1 protein switches between loop and beta-strand conformations, resulting in its dimerization with a protein kinase and regulation of the cell cycle in humans[15].

Conformational changes within a protein are often observed during and after ligand binding. In addition to the flexible regions, rigid motifs are important in several areas of protein structure and function. For many proteins, active or ligand-binding sites are described as rigid templates, which do not undergo significant conformational change upon ligand binding.

MICROSCOPIC FORCES AS DISTANCE CONSTRAINTS

Imposing distance constraints to replace important bonding forces between atoms reduces the total number of degrees of freedom available to the protein. If we imagine atoms interacting pairwise within a protein, the effect of a distance constraint is to fix the interaction between a pair of atoms. Applying pairwise distance constraints is the *input*, and the *output* is the set of rigid clusters, separated by flexible joints.

Obviously, in reality no set of atoms within a protein will ever define a *perfectly* rigid cluster. By imposing distance constraints on the strongest interactions, to quench mainly higher frequency motions, it is expected that defining a rigid cluster is meaningful only on sufficiently long time scales. Floppy regions consist of atoms that continue to have relative motion on these long time scales. The inherent assumption that we make in our distance constraint approach is to neglect vibrational modes within a rigid cluster and the coupling between inter cluster vibrational modes. The lowest frequency from the intra cluster vibrational modes will set the relevant time scale for which the rigid and floppy regions become meaningful. The cleanest situation occurs when the separation between strong and weak interactions has a significant gap.

Separation Between Strong and Weak Forces

The hard covalent bonding within the protein, consisting of the bond-stretching and bond-bending forces, defines a natural set of distance constraints. The energies associated with central bond-bending and torsional forces are of the order 25 (Kcal/mol)/Å^2, 4 (Kcal/mol)/rad^2 and 0.06 Kcal/mol respectively[16]. Note that although the energy units are different, a direct comparison in magnitudes can roughly be made upon realizing that the covalent bond lengths are about 1.5Å. It is common practice to fix the covalent bond lengths and bond angles while allowing the dihedral angles to be free to rotate. Using the dihedral angles as a set of internal coordinates, the number of degrees of freedom to describe the motion of a protein is typically reduced by a factor of about seven[17].

In addition to the central and bond-bending forces, some torsional forces associated with the peptide or resonant bonds lock rotational motion about the bond axis. In this case, the dihedral angle can be fixed by using a third neighbor distance constraint. Modeling only the strongest set of forces associated with covalent bonding as distance constraints will not define any large rigid region. Instead, the protein flexibility will be described by a floppy set of connected small rigid clusters that can be identified by inspection. This happens because in all proteins, the covalent bonds form a tree-like structure (i.e no loops) with a few exceptions consisting of rings found in some residues (proline, histidine, phenylalanine, tyrosine and tryptophan) or perhaps more interestingly, crosslinking from a few disulfide bonds.

Groups of atoms tend to cluster into well defined substructures such as alpha helices and beta sheets (common secondary structures[18]). Furthermore, many proteins are made up of a collection of stable fragments[4], that range in size from a small part of a domain to an entire domain. Protein domains can be defined in various ways[11,19], either by function or some particular structural characteristic, such as packing density. In general a protein domain will consist of many atoms (typically 1000 atoms or more) and will contain common secondary structures. From empirical evidence, it is believed that conserved substructures exist within domains during the course of conformational changes. These are the rigid clusters we seek to identify a priori.

We must apply distance constraints associated with other (weaker) forces in order to determine larger rigid regions. We could of course (erroneously) produce one entire rigid region by placing *all* third neighbor distance constraints associated with torsional forces, thereby fixing all dihedral angles. The same trivial result would occur by modeling van der Waals interactions by distance constraints. Keep in mind, however, that a reasonably clean separation between strong and weak forces is desired to make distinctions between rigid and floppy regions. Therefore we need to apply distance constraints to bonding forces that are stronger than torsional forces (not associated with resonant bonds), but these forces will still be weaker than the covalent bond-bending forces.

Hydrogen Bonding: A Hierarchical Approach

It is natural to consider hydrogen bonds as the next set of (weaker) forces, after the covalent forces, and to model them also as distance constraints. The strength of a hydrogen bond varies from nearly as strong as the covalent bonds to as weak as the van der Waals interactions[20,21]. This broad range of variation in strength indicates that the hydrogen bond, unlike covalent bonding, is very sensitive to its local environment. We know hydrogen bonding is very important in proteins because secondary structures, such as alpha helices, beta sheets and hairpin turns, can be readily identified in terms of hydrogen bond patterns that form along the mainchain of the protein[18]. These substructures frequently occur in proteins (evident from the Brookhaven Protein Data Bank[2], which is an on line computer-based archive for bio-macromolecular structures), and their relative arrangements further define interesting structural motifs.

The hydrogen bond patterns within the mainchain, associated with the nitrogen (donor) and oxygen (acceptor) atoms provide crosslinking between residues. The crosslinking of hydrogen bonds is responsible for the high degree of mechanical stability found within secondary structures. Similarly, we expect that the additional crosslinking provided by hydrogen bonds outside of the mainchain will support larger mechanically stable substructures, which may become as large as an entire domain. Modeling these hydrogen bonds as distance constraints gives a simple way to characterize the degree of stability within the protein. Since placing a distance constraint between two atoms does not make a distinction between a strong or weak interaction, a weak hydrogen bond becomes just as important as a strong hydrogen bond or a covalent bond in this regard. Therefore, at our discretion, we must choose whether or not to model certain bonding forces with distance constraints.

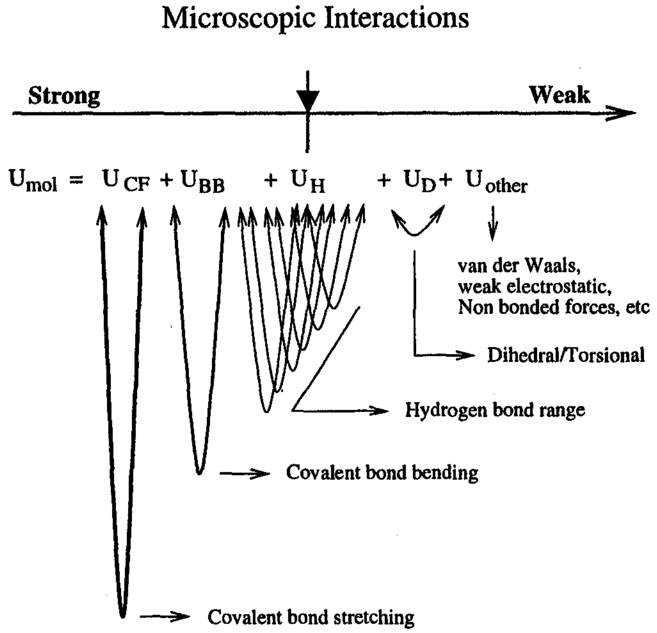


Figure 1. A schematic representation of ordering the microscopic forces from strongest to weakest. Distance constraints are used to model strong bonding forces to the left of a sliding pointer. This approach defines a system of interacting rigid clusters.

Consider modeling only those hydrogen bonds having an interaction strength *greater* than some predefined cut-off, as distance constraints. This will divide the hydrogen bonds into two groups, but a gap between strong and weak forces will *not* be found in this case. Unlike covalent network glasses[22], where there are no hydrogen bonds, obtaining a clean separation in bond strength is not possible in proteins, because the hydrogen bonds cover a broad range of interaction strengths. As a result, no justification can be given for selecting a single set of hydrogen bonds to be modeled as distance constraints.

The schematic diagram shown in Figure 1 represents a hierarchical approach that we have adopted for selecting which interactions to model as distance constraints. Naturally, the covalent bond-stretching and angular or bond-bending forces will be modeled as distance constraints. The nearest neighbor bond-stretching force defines a nearest neighbor distance, and the angular or bond-bending force defines a second nearest neighbor constraint. In addition, we introduce a *pointer* on an interaction scale such that any hydrogen bonds interacting more strongly are modeled as distance constraints, while those that have a weaker interaction strength are ignored. The pointer is allowed to slide down the scale starting from an interaction strength just below that of covalent bonding. As the pointer slides down, more and more hydrogen bonds are included. This approach allows us to uncover various rigid substructures, that will merge together as more crosslinks are added, albeit via weaker hydrogen bonds.

Intuitively, one expects that a large rigid cluster, consisting of many weak hydrogen bonds, will not be as *rigid* as some internal parts that form rigid substructures involving strong hydrogen bonds. Of course, the strongest rigid substructures within any large rigid region will be the set of small rigid clusters defined by the covalent bonding. The hierarchical approach of gradually selecting weaker and weaker hydrogen bonds allows us to access the relative degree of stability (as a continuum measure) between different regions in the protein. However, before we accomplish this task, we first construct a quantitative measure for the degree of stability applied locally, rather than globally.

Van der Waals and Hydrophobic Forces

In proteins there are both short- and long-ranged non-bonding forces. Individually these forces are generally weak, but collectively they are important in governing the dynamics of a protein. For example, each van der Waals interaction is too weak to model as a distance constraint, yet collectively the van der Waals interactions play an important role in determining steric conformational constraints. The hydrophobic force is a thermodynamic force that is generally regarded as a dominant contribution that drives a protein to fold[23], and it plays an important role in stablizing a protein structure in the native state. It is not possible to apply distance constraints to model the hydrophobic force, because it is an entropic force, depending on the ensemble of configurations available to all atoms of the protein and solvent molecules.

During conformational changes, certain regions of the protein are essentially preserved as a rigid body. Our viewpoint is that the rigid clusters, defined by the covalent bonding and the crosslinking hydrogen bonds, will interact with one another via the weaker non-bonding forces as a system of coupled rigid bodies. Therefore, rather than modeling the protein stucture as a system of interacting particles, we focus on the long time dynamics where conformational changes are controlled by weaker, but collectively important, forces acting between coupled rigid bodies.

The van der Waals and hydrophobic forces can be regarded as the most important for driving and stablizing protein folds, because they will essentially determine the most probable conformations having the lowest thermodynamic free energy. As a result, it should not be surprising to experimentally observe thermodynamically stable regions within the protein that are identified as floppy regions using a simple distance constraint approach. These floppy regions locate low energy pathways that enable a protein to explore different conformations. These pathways are important to facilitate domain rearrangement or allow some local flexibility for the protein to successfully bind with a ligand. Thus, we are focusing on the mechanical stability of a network of rigid clusters, rather than the thermal stability of the protein as a whole.

GENERIC RIGIDITY

A brute force method to *identify all rigid clusters* in a system with N atoms would involve diagonalizing a dynamical matrix N^2 times to check for redundant distance constraints between all pairs of atoms[24]. This leads to a polynomial time algorithm that scales as $O(N^5)$. Other methods to check for redundant distance constraints, involving the rigidity matrix for example, can be used instead of diagonalizing the dynamical matrix, but the final result will still lead to a polynomial time algorithm that also scales as $O(N^5)$. Finding stressed regions requires an additional numerical calculation involving a stress-strain relaxation of the network, using for example a conjugate gradient method, that scales as $O(N^2)$.

The distance constraint approach will only be useful in as much as the calculation for identifying floppy and rigid regions is fast and scales linearly or nearly linearly with protein size. By considering the protein as a generic structure, one can deduce the rigidity properties using concepts from graph rigidity[25]. This is the contribution of this paper.

A generic structure is one that has no special symmetries, such as parallel bonds or bond angles of 180 degrees, that could create geometrical singularities[26]. Under these conditions, the study of rigidity becomes much easier to deal with because the properties of network rigidity depend only on the *connectivity* of the network, as determined by an underlying graph. Although removing concerns about the *particular* coordinates of atoms gives a great simplification conceptually, the calculational problem at hand is still far from trivial.

There is a theorem by Laman[27] that states how generic rigidity of any *two* dimensional bar-joint network can be completely characterized by applying constraint counting to *all* subgraphs. Applying Laman's theorem directly leads to a combinatorial calculation that scales as the exponential in the number of atoms within the network. However, by applying Laman's theorem recursively, a very fast and efficient algorithm (the so called *pebble game*) has been constructed[28,31,32] for identifying rigid clusters and stressed regions. The *pebble game* uses computer memory that scales linearly with the size of the network, and has a performance that scales in the worst case as $O(N^2)$ for pathological networks, but in practice performs nearly linearly with N.

In three dimensional generic bar-joint networks, constraint counting over all subgraphs is known to *fail* in general. However, for the special class of truss frameworks, the theorem of Laman can be generalized[29] to three dimensions. Again constraint counting over all subgraphs is enough to completely characterize generic rigidity. We will model the microstructure of a protein using distance constraints that define a truss framework, which is also called a bond-bending network.

Three Dimensional Bond-Bending Networks

Covalent bonding naturally defines bond distances and angles. As shown in Figure 2, the covalent bonding defines a graph, where each vertex represents an atom, nearest neighbor distances represent central forces and next nearest neighbor distances represent bond-bending forces. This type of graph is called a *squared graph*[30], a truss framework or a bond-bending network. The network connectivity is completely described by the nearest neighbor central-force constraints, which are viewed as inducing the bond-bending next nearest neighbor distance constraints. The only elementary floppy element that exists within a bond-bending network is a *hinge joint.* Rotations

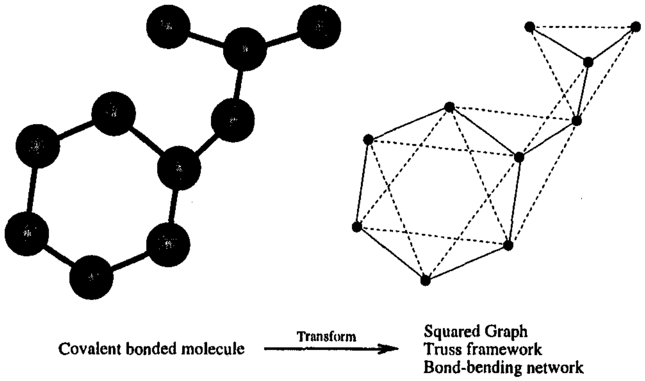


Figure 2. Once a covalent bonded molecule or network is regarded as generic, the rigidity of the network is completely determined by the connectivity of the underlying structure. A transformation is made from a physical molecule to a mathematical graph, where vertices represent atoms, solid lines represent distance constraints between nearest neighbor atoms associated with central forces and dashed lines represent distance constraints between next nearest neighbor atoms associated with bond-bending forces. The graph is of a special kind, called a squared graph. This structure is also referred to as a truss framework or bond-bending network.

through a dihedral angle about the axis of a central-force constraint is a priori possible, but may be locked because of the network properties.

As shown in Figure 3 the hydrogen bond is modeled in a similar way to a covalent bond with generic geometry, in which the donor, hydrogen and acceptor atoms are *not co-linear.* Typically, each hydrogen bond will introduce three distance constraints, corresponding to one central force between the hydrogen and acceptor atoms and two bond-bending forces associated with the hydrogen and acceptor atoms. If an acceptor atom is involved in n_h hydrogen bonds, there will be $(n_h - 1)$ additional bond-bending force constraints between hydrogen-acceptor-hydrogen atoms. This particular model for a hydrogen bond is mathematically convenient because it allows the protein structure to be described as a bond-bending network. Physically, the model is also reasonable because hydrogen bonds are almost never linear and the three dihedral angle degrees of

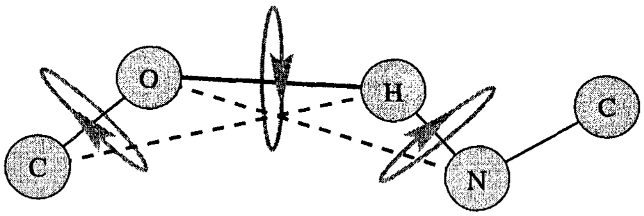


Figure 3. A diagram showing a hydrogen bond involving a donor and acceptor atom taken here as nitrogen and oxygen respectively. It is modeled as three generic distance constraints, consisting of a nearest neighbor central-force constraint shown as a thick solid line, and two next nearest neighbor bond-bending force constraints shown as dashed lines. Each constrained hydrogen bond is also associated with three a priori rotatable dihedral angles indicated by the arrows.

freedom associated with the hydrogen bond allows it to be relatively flexible. Although modeling the hydrogen bond to be more or less constrained is easy to do, we consider the model in Figure 3 to strike a good balance between neither over- nor under-representing the effectiveness of a hydrogen bond.

The 3D Pebble Game

At the heart of the FIRST computer program is the *3D pebble game* algorithm that is constructed in a very similar way as the two dimensional *pebble game*[28,29,31,32]. Here, three pebbles, representing three translational degrees of freedom, are assigned to each vertex in the graph. For each independent distance constraint between two vertices, a pebble from either one of the incident vertices must be used to *cover* the constraint. Pebbles associated with vertices are called *free* pebbles, and they represent the independent degrees of freedom remaining within the network. Each independent distance constraint uses up one independent degree of freedom. Then the following covering rule is applied: once an independent constraint is covered by a pebble, it must always remain covered by any of the pebbles associated with either of its incident sites. Rearrangement of pebbles throughout the network is possible provided this covering rule is not violated.

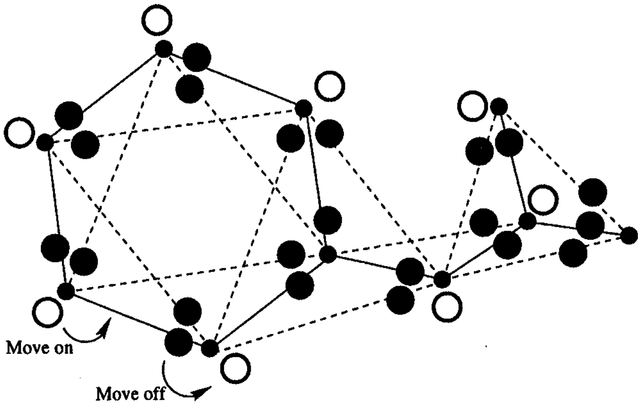


Figure 4. A diagram showing the final pebble covering of a simple network. The open big circles represent free pebbles that are placed directly on vertices and denote free degrees of freedom available to the network. The filled big circles represent pebbles that are covering a distance constraint, and are placed directly on the edges of a graph. The pebble covering is not unique, because pebbles can be rearranged according to a few simple rules as explained in the text. Here, we show an example of how an elementary pebble exchange works, as indicated by the two arrows. A free pebble can be moved on to a covered edge (off from a vertex) provided a corresponding pebble, which is presently covering the edge and is associated with the neighboring vertex, is moved off (on to a vertex).

In Figure 4 an example of a pebble covering is given for the graph shown in Figure 2 (rotated by 45 degrees). The pebble covering is a convenient way to represent a dynamically changing directed graph[28] facilitated by the rearrangement of pebbles through a series of elementary pebble exchanges. Note that if another constraint is added between a pair of vertices within the six fold ring, it would be found to be redundant, because there are not enough free pebbles in this region to cover the constraint. This physically corresponds to the fact that the six fold ring is generically rigid.

The *3D pebble game* is a recursive algorithm like its 2D counterpart. The network is built up by placing one distance constraint at a time. An essential feature of the *3D pebble game* is that each central-force distance constraint associated with vertices v_1 and v_2 must have associated with it angular (i.e. second nearest neighbor) constraints around both vertices v_1 and v_2. For each new independent distance constraint that is introduced, pebbles are rearranged in a way to test if the new distance constraint is independent or not. If the new distance constraint is found to be independent, it is then covered, otherwise it is not covered. This process continues until all distance constraints within the network have been completely placed in the network. The algorithm can be sketched in the following way:

1. Place a central-force distance constraint between vertices v_1 and v_2 as appropriate and as described in the previous sections.
2. Rearrange the pebble covering to collect three pebbles on vertex v_1.
3. Rearrange the pebble covering to collect the maximum number of pebbles on vertex v_2 while holding the three pebbles at vertex v_1.
4. If the number of pebbles on vertex v_2 is two, the distance constraint is redundant. Otherwise, three pebbles reside at vertices (v_1 and v_2).
   Continue to rearrange the pebble covering:
   (a) Hold the three pebbles on both vertices v_1 and v_2.
   (b) For each neighbor of vertex v_2: Attempt to collect a pebble.
   (c) If for any neighbor of vertex v_2 a pebble cannot be obtained, then that distance constraint is redundant.
5. If the distance constraint is not redundant, cover it with a pebble from vertex v_2.

Unlike the 2D *pebble game*, the distance constraints cannot be placed in *any* random order. There is an additional rule about the placement of distance constraints. The first distance constraint that is introduced must correspond to a central-force constraint. After each central-force distance constraint is placed, all of its associated induced angular or bond-bending constraints (next nearest neighbor distance constraints) must be placed before another central-force constraint can be placed. Within this restriction, the order of placing either central force or the induced bond-bending constraints is completely arbitrary. The restriction on the order of placing distance constraints *nearly* maintains the form of the network to be that of a truss framework throughout the building process. The entire network is a truss framework, except in the local region where the additional central-force bond is initially placed. The network is restored to a truss framework *everywhere*, after all the induced bond-bending constraints are added. This local deviation from a truss framework does not create any pivot points or implied hinge joints. Therefore this restriction on recursively placing constraints is sufficient[29] for combinatorial constraint counting to remain valid in characterizing generic network rigidity.

Finally, torsional constraints for the peptide and resonant bonds are fixed by *third* nearest neighbor distance constraints. These are most conveniently placed after the central and bond-bending distance constraints have been placed. No induced bond-bending constraints are associated with the torsional constraints, because these are auxiliary distance constraints for locking in certain dihedral angles.

After all these distance constraints have been placed, the number of free pebbles remaining on the vertices gives the total number of degrees of freedom required to describe the motion of the network. This includes the six trivial rigid body translational and rotational degrees of freedom of the whole network. The free pebbles can be rearranged, but are restricted to certain regions because of the rule for pebble covering. For example, no more than six free pebbles can be found within a rigid cluster. Based on the location and number of free pebbles throughout the network, one can identify overconstrained regions, rigid clusters and underconstrained regions.

Identifying Overconstrained Regions

A redundant constraint is identified when a failed pebble search occurs. A failed pebble search consists of a set of vertices that have no extra free pebbles to give up. This physically corresponds to the region of vertices and distance constraints that predefines the length between a pair of vertices. As we are considering generic networks, placing a distance constraint between this pair of vertices will cause a length mismatch. Physically, this means that the bond lengths and angles within this region of the failed pebble search will become distorted as this region will be internally stressed. Thus, by recording the failed pebble searches, which we call Laman subgraphs as in two dimensions, we automatically identify the overconstrained regions.

Overconstrained regions will always consist of closed loops. As distance constraints are added to the network, more overconstrained regions will be found, and generally these regions will overlap. Overlapping overconstrained regions merge together into a single overconstrained region. As these networks are generic, stress will propagate throughout such a merged set of overconstrained regions.

A subtle point that does not occur in two dimensions is that stress can propagate from one floppy region to another in three dimensions. In bond-bending networks, the effect is a trivial propagation of stress between neighboring atoms that are both four (or more) coordinated. This occurs because a four coordinated atom has six angular constraints, but only five are independent. As a result, every four (or more) coordinated atom is part of at least a locally stressed region consisting of itself and its neighbors. The loops are formed by central-force and bond-bending constraints.

As an example, consider a long floppy chain constructed such that along the mainchain, there are carbon atoms, each connected to two other carbon atoms along the mainchain, and two hydrogen atoms forming dangling ends, so that each carbon atom has four neighbors. Although the chain is floppy, having a rotatable dihedral angle between each pair of carbon atoms in the mainchain, it will carry stress from one end to the other! Again the reason for this effect is in modeling the local chemistry of a four coordinated covalently bonded atom with six angular constraints. By explicitly monitoring this type of overconstrained region, we can include or exclude this effect. This effect can be excluded by removing *any* one of the six angular constraints at each four fold coordinated atom in the model. For the remainder of this article we will not be concerned with this type of locally induced stress. Instead, we are interested in network induced overconstrained regions caused by loops formed by central-force constraints.

Rigid Cluster Decomposition

The vertices within each rigid cluster are divided into two types. A vertex is classified as a *bulk vertex* when all of its neighboring vertices also belong to the same rigid cluster, otherwise it is classified as a *surface vertex.* Thus a surface vertex has at least one neighbor belonging to a different rigid cluster than itself. Isolated vertices and the vertices within dimers are regarded as bulk vertices. This is a very useful classification scheme because in general a vertex can belong to more than one rigid cluster simultaneously, yet a bulk vertex can be assigned to one and only one rigid cluster. A unique rigid cluster labeling scheme can be constructed by labeling only the bulk vertices. This labeling scheme for rigid clusters is appropriate to three dimensional bond-bending networks[1,28], where it is possible to take advantage of some special properties. In general this labeling scheme will not work. For example, this labeling scheme cannot be used at all in two dimensional networks[28,31,32].

Each rigid cluster consists of a set of bulk vertices, all having the same *cluster label*, and a set of surface vertices all having a different cluster label than the bulk vertices. Any non-bulk vertex that is nearest neighbor to a bulk vertex for a particular rigid cluster is actually a surface vertex for that rigid cluster. There are two additional Properties about the rigid clusters within a bond-bending network[29] that will be needed here.

P1 For each vertex, v_1, all its neighbors are automatically mutually rigid with respect to one another as well as with vertex v_1 itself because of the bond-bending constraints.

P2 All vertices within a rigid cluster are connected via a path of central-force constraints.

Using the above properties, an algorithm to obtain the rigid cluster decomposition within bond-bending networks is given as:

1. Initialize rigid cluster counter. Label all vertices null.
2. For each vertex, v_1, not already associated with a rigid cluster:
   - (a) If an isolated vertex: Increment counter and label vertex.
   - (b) If the vertex belongs to a dimer: Increment counter, label both vertices.
   - (c) If vertex is one fold coordinated; return to (2).
   - (d) Rearrange pebbles: Collect 3 pebbles on vertex v_1, 2 pebbles on its 1st neighbor, v_2, and 1 pebble on its 2nd neighbor, v_3.
   - (e) By property **P1**, use vertices $\{v_1, v_2, v_3\}$ as a rigid base, and place them in a stack defining the rigid cluster.
   - (f) By property **P2**, use a breadth first search[33] via nearest neighbors, $\{v_{new}\}$, to grow the rigid cluster stack to completion.
     - i. For each vertex v_{new}: Rearrange pebbles attempting to collect a pebble while holding the six pebbles on the rigid base.
     - ii. If attempt fails: Include new vertex in the rigid cluster stack.
   - (g) Increment cluster label: Label all bulk vertices within the rigid cluster.

After the above algorithm is finished, every bulk vertex will have a cluster label assigned ranging from 1 to the number of rigid clusters in the network. One fold coordinated vertices are regarded as bulk vertices.

Identifying Underconstrained Regions

After the rigid cluster decomposition is made, locating hinge joints is an easy task. A hinge joint can never occur about a bond-bending distance constraint[29]. To find all the hinge joints within the network, one needs to check the two incident vertices associated with each central-force distance constraint. If the two vertices have different cluster labels, then a dihedral rotation is possible and the central-force constraint is a hinge joint, otherwise the dihedral angle motion is locked, as it is part of a rigid cluster.

The number of hinge joints will generally be more than the number of residual internal degrees of freedom in the network. This means that not all the rotatable dihedral angles associated with the hinge joints are *independent.* We call a hinge joint independent if its dihedral angle can be changed without affecting other dihedral angles within the network. Collective motions will take place in underconstrained regions within the protein. It is convenient to partition hinge joints into distinct underconstrained regions, which may include only one independent hinge joint.

Let the set of dihedral angles associated with the hinge joints define a reference set of internal coordinates, not all of which are independent. We will partition the independent degrees of freedom of the network into subsets, which correspond to underconstrained regions. In a similar way that there are redundant constraints within overconstrained regions, there are floppy modes within underconstrained regions. Physically, collective motions can occur within a particular underconstrained region without affecting internal coordinates outside of the region. The underconstrained regions are identified by attempting to specify a value for each dihedral angle. Specifying a dihedral angle is equivalent to placing an external torsional constraint to lock in this choice of angle. Independent externally imposed torsional constraints represent independent degrees of freedom available to the system, while redundant externally imposed constraints indicate the angle is predetermined as part of a collective motion. Therefore, the algorithm for identifying the underconstrained regions is given by:

1. Initialization: Record null Laman subgraphs by regarding all previously identified redundant constraints within the network as not present.
2. Continue to play the *pebble game* by placing a third nearest neighbor distance constraint to lock the dihedral angle for each hinge joint in the network.
3. After completion: The set of newly formed Laman subgraphs identifies the underconstrained regions.

After all torsional constraints are placed on each hinge joint, a completely connected network will become totally rigid having six degrees of freedom that are represented by six remaining free pebbles. The Laman subgraphs within this bond network now define the underconstrained regions, which consist of loops involving central-force bonds. The number of independent dihedral angles within an underconstrained region is given by the number of torsional constraints covered by a pebble. Independent hinge joints are associated with a covered torsional constraint not belonging to a Laman subgraph.

FLOPPY INCLUSION AND RIGID SUBSTRUCTURE TOPOGRAPHY

A large amount of detailed information about the mechanical stability of a protein under a fixed set of distance constraints becomes available from a FIRST analysis. All overconstrained regions, rigid clusters and underconstrained regions are determined. We discuss in detail the output of such a FIRST analysis for a hand made 32 atom example shown in Figure 5, which ties together the discussion of the previous section. In addition to partitioning a structure into different types of regions, a continuous measure for the degree of stability is introduced and is worked out for the network shown in Figure 5. The FIRST analysis is then applied to protein structure, where the floppy inclusion and rigid substructure topography is described by a one dimensional representation in terms of what we define as a *stability index.*

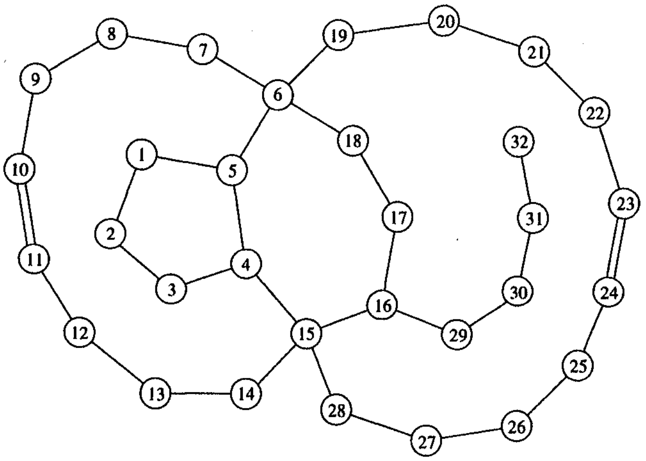


Figure 5. A simple hand made 32 atom network. Each atom is labeled from 1 to 32. Single lines between labeled atoms represent central-force distance constraints that have a priori rotatable dihedral angles, while the double lines represent double bonds where the dihedral angle is locked.

3D Graphical Display

Constructing a three dimensional graphical display for the different types of mechanical regions amounts to devising a systematic labeling scheme. We take our basic set of reference labels to be the atom numbers. For the network shown in Figure 5, it is possible to obtain the complete solution for the floppy inclusion and rigid substructure topography by inspection. Each part of the solution will be discussed in turn.

Recall that overconstrained regions are identified within the *3D Pebble Game* as a result of failed pebble searches. Each time a failed pebble search is encountered a Laman subgraph is identified, which contains more distance constraints than needed to make that region rigid. A convenient labeling system for identifying Laman subgraphs is to initially assign each atom a Laman subgraph label. The atom numbers are used as the *initial* set of Laman subgraph labels.

When a Laman subgraph is uncovered from a failed pebble search, all atoms found within are re-assigned the lowest Laman subgraph label encountered. This allows Laman subgraphs to naturally merge together. Since Laman subgraphs (representing overconstrained regions) consist of loops, stress will reside in the central-force bonds connecting pairs of atoms having the same Laman subgraph label. Loops of atoms with

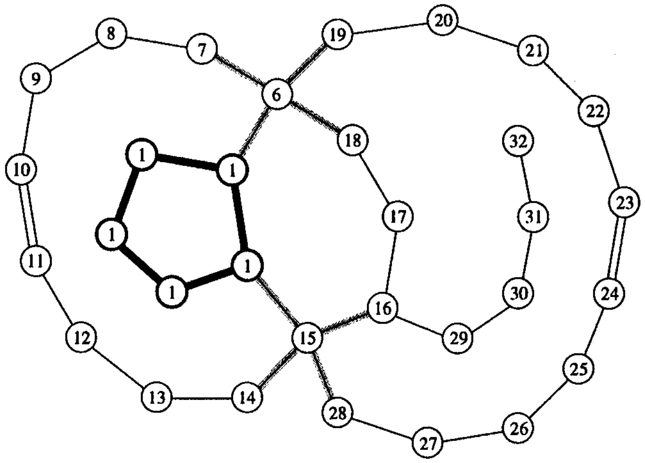


Figure 6. The FIRST analysis for internally stressed regions. Each atom is assigned a Laman subgraph label. The thick solid black lines show a network induced overconstrained region where there will be internal stress. In addition, there are two localized regions that are internally stressed (highlighted using thick grey lines) each associated with the presence of a four-coordinated atom.

the same Laman subgraph label define a particular overconstrained region. As shown in Figure 6 only one network induced overconstrained region is present, consisting of a five fold ring with one redundant constraint. Since the incident atoms about all other central-force bonds are assigned different Laman subgraph labels, no network induced stress is present anywhere else.

Induced stressed regions, localized around four (or higher) coordinated atoms, can also be monitored. Four coordinated atoms, such as atoms 6 and 15, produce a local Laman subgraph with one redundant constraint within a set of edge sharing loops consisting of the central and bond-bending constraints associated with the four coordinated atom and its nearest neighbor atoms. The Laman subgraph labeling scheme works for both kinds of induced stressed regions, where the former is identified by checking the atom coordination number. We eliminate this effect by removing one of the six angular constraints, chosen arbitrarily, at each four fold coordinated site.

The coloring of rigid clusters is associated with the bulk atoms, each assigned a particular color according to its cluster label. The assignment of colors is restricted by the requirement that two nearest neighbor bulk atoms having different cluster labels must have different colors. As a general property of bond-bending networks, each central-force bond belongs to either two rigid clusters when it is a hinge joint[29], or one rigid cluster otherwise. Therefore, the first half of each central-force bond stemming from a bulk atom is half colored. A central-force bond is completely colored only when both incident atoms have the same cluster label. As a result, only hinge joints appear as half colored bonds.

Empirically we find that a complete color assignment is possible using as little as four colors, although we generally use more colors by having a dark and bright set of colors to denote rigid and floppy regions respectively. The relationship between uniquely coloring the rigid cluster decomposition and the chromatic coloring of a graph, such as the famous four color problem on planar graphs[30,34] has not been investigated.

In Figure 7, we show the rigid cluster decomposition of our example 32 atom

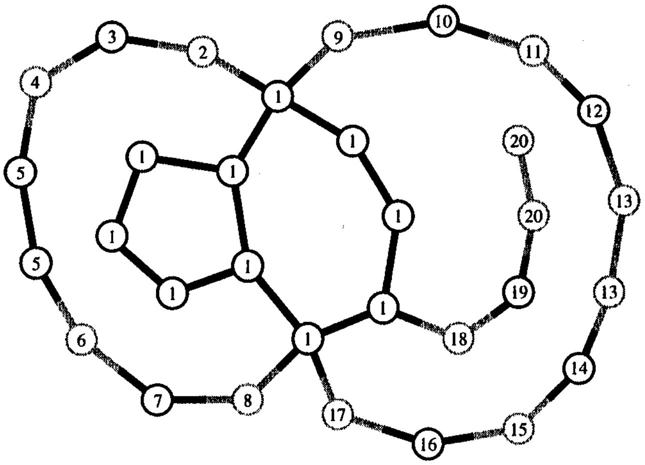


Figure 7. The FIRST analysis for the rigid cluster decomposition. Each atom is assigned a cluster label. Each rigid cluster is assigned a color, with the constraint that neighboring clusters must be assigned a different color. Half colored bonds represent hinge joints.

network. Here we are able to completely color the rigid cluster decomposition using only two colors (ie. black and grey). Comparing Figure 5 with Figure 7 indicates that rigid cluster #1 consist of 14 atoms (bulk atoms $\{1, 2, 3, 4, 5, 6, 15, 16, 17, 18\}$ and surface atoms $\{7, 14, 19, 28\}$), rigid cluster #2 consist of 3 atoms (bulk atom 7 and surface atoms $\{6, 8\}$), rigid cluster #5 consist of 4 atoms (bulk atoms $\{10, 11\}$ and surface atoms $\{9, 12\}$), etc. Within rigid cluster #1, only the five fold ring is overconstrained as shown in Figure 6 while the seven fold ring is isostatically rigid. Each half colored bond identifies a hinge joint where its dihedral angle can be rotated, but this does not indicate the inter dependence of rotating a set of dihedral angles in the floppy regions.

Each underconstrained region is displayed with a unique color to indicate an interdependent set of hinge joints. The labeling scheme is applied directly to the hinge joints, where each hinge joint is *initially* assigned a distinct label representing the underconstrained regions. We denote the hinge joint between atoms i and j as $H_{i,j}$. As external torsional constraints are placed on each hinge joint, all H_{ij} found within a failed pebble search are re-assigned the lowest common label encountered. Therefore, underconstrained regions merge together whenever failed pebble searches have overlapping H_{ij}. Only central-force bonds that were originally hinge joints (labeled by the set $\{H_{ij}\}$ and actually are hinge joints within the protein) are colored. Therefore, underconstrained regions are generally not contiguous.

In Figure 8 it can be seen that the 32 atom example network has two large underconstrained regions and three other underconstrained regions, each consisting of one independent hinge joint. The left underconstrained region consist of 8 hinge joints, each colored black, while it has only two independent dihedral angles indicated by the circular arrows. Comparing with Figure 7 it is seen how intermediate rigid clusters #1 and #5 cause the underconstrained region to be non-contiguous. Similarly, the underconstrained region on the right is also non-contiguous. It consist of 10 hinge joints, shown in grey, while it has four independent dihedral angle rotations. The three independent hinge joints are indicated only by circular arrows.

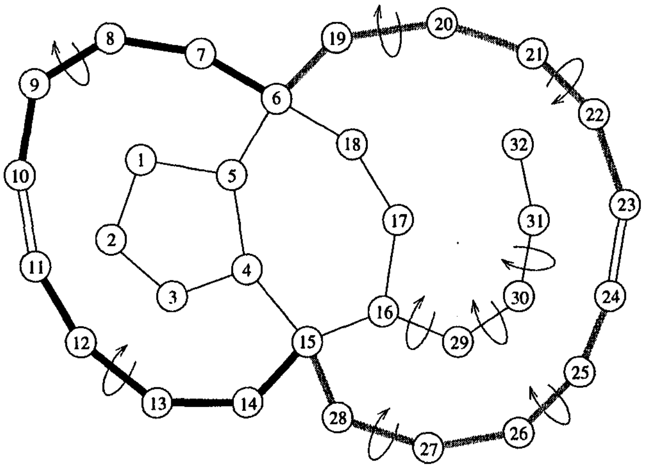


Figure 8. The FIRST analysis for the underconstrained regions. Each atom is numbered from 1 to 32. Two underconstrained regions are found. The black thick lines on the left represent one underconstrained region consisting of two floppy modes extending over eight hinge joints. The grey thick lines on the right represent another underconstrained region consisting of four floppy modes extending over ten hinge joints. The nine arrows represent internal independent dihedral angles. All central-force bonds that are not part of the two identified underconstrained regions or have a circular arrow have locked dihedral angles. Three independent hinge joints are also shown in the section from 16 to 32.

The independent dihedral angles within an underconstrained region correspond to free pebbles in the the *3D pebble game*. The choice of which hinges are taken as independent is arbitrary, and is analogous to choosing which bonds are redundant within an overconstrained region.

In general, rigid clusters will subdivide a floppy region into multiple underconstrained regions, where independent collective motions can take place. Figure 8 shows an example of how rigid cluster #1, shown in Figure 7, partitions free degrees of freedom into two localized regions. It is seen that once the two independent dihedral angles are specified in the underconstrained region on the left, the relative position of atoms $\{8, 9, 10, 11, 12, 13\}$ are all fixed relative to rigid cluster #1. The specification of these two angles, however, has no affect on the relative position of atoms $\{20, 21, 22, 23, 24, 25, 26, 27\}$ within the underconstrained region on the right. Notice that if atoms $\{1, 2, 3\}$ were removed from the network, rigid cluster #1 would fall apart. Moreover, the two underconstrained regions would merge together with the seven fold ring to form a single underconstrained region with 25 hinge joints and 11 independent dihedral angles. This illustrates how a local structural change can have important consequences to the overall mechanical stability of the network.

Quantifying the Degree of Flexibility

With the above system of coloring, we are able to make a 3D multi-color map for overconstrained regions, the rigid cluster decomposition and the underconstrained regions. Although it is insightful to view the three dimensional multi-color rendering of these regions, one must exercise caution not to be mislead about the stability within a protein.

Modeling only a few additional hydrogen bonds as distance constraints will often make some floppy regions (or parts within) rigid. Likewise, the removal of distance constraints associated with just a few hydrogen bonds will sometimes make some rigid regions break apart like a house of cards. Modeling just a *few* more or less hydrogen bonds as distance constraints can sometimes dramatically change the mechanical stability of a protein. However these observed changes in the 3D graphical displays do not necessarily reflect the actual change in mechanical stability that is occurring physically.

A floppy region consisting of many interconnected rigid clusters may define a collective motion having only a few independent degrees of freedom. This floppy region, although underconstrained, would be quite stable mechanically as it is nearly rigid. An isostaticly rigid region is not expected to be as stable as an overconstrained region. For this reason a continuous index of stability is useful. As the hydrogen bond selection criteria is relaxed using a hierarchical protocol, the observed changes in the mechanical stability of the protein can be tracked in a continuous fashion.

The total number of floppy modes in a protein, denoted by F, corresponds to the number of *internal* independent degrees of freedom. To obtain F, the six trivial rigid body degrees of freedom must be subtracted out from the total number of independent degrees of freedom. For purposes of simplifying the discussion, it has been assumed that all atoms in the protein are connected via covalent bonding. The global count of the number of floppy modes gives a good sense of intrinsic flexibility. However, a better measure for the degree of floppyness can be obtained by tracking how the total number of floppy modes are spatially *distributed* throughout the protein. In particular we are interested in locating underconstrained regions and the number of floppy modes contained within each of these regions.

Regions containing more constraints are regarded as being more stable than regions with less constraints. Overconstrained regions have more constraints than necessary to be rigid, and therefore are considered to be more stable. A global count for the number of redundant constraints, denoted by R, gives a sense of the overall stability of a protein. However, a better measure for the degree of stability can be obtained by tracking how the total number of redundant constraints are distributed throughout the protein. In a similar way as done with the floppy modes, we will be interested in where the overconstrained regions are located and count the number of redundant constraints present in each region.

We will define a quantity s_i, as a *stability index* characterizing the i-th central-force bond in the protein. Let H_k and F_k respectively denote the number of hinge joints and the number of floppy modes (internal independent degrees of freedom) within the k-th underconstrained region. Let C_j and R_j respectively denote the number of central-force bonds and the number of redundant constraints within the j-th overconstrained region. Combining a quantitative measure for both the degree of floppyness and stability associated with the distribution of floppy modes and redundant constraints respectively, the definition for the stability index is given by:

$$s_i \equiv \begin{cases} \frac{-F_k}{H_k} & \text{in an underconstrained region} \\ 0 & \text{in an isostatically rigid region} \\ \frac{R_j}{C_j} & \text{in an overconstrained region.} \end{cases} \qquad (1)$$

When the i-th central-force bond is a hinge joint, the stability index is defined to be a negative quantity with magnitude given by the the number of floppy modes divided by the total number of hinges within the underconstrained region. The number of floppy modes correspond to the number of independent dihedral angle rotations that can be made within the underconstrained region. Since the number of independent dihedral angle rotations must be less than or equal to the number of hinge joints, the stability index can never be less than -1. For an independent hinge joint the stability index works out to be -1. Furthermore, the stability index for a hinge joint must always be less than zero because there will always be at least one floppy mode within an underconstrained region.

When the i-th central-force bond is not a hinge joint, it must be part of a rigid cluster, although it may or may not be part of an overconstrained region. If the central-force bond is within an overconstrained region, the stability index is assigned a positive value given by the number of redundant constraints divided by the total number of central-force bonds within the region. Although there is no useful upper bound for the stability index it rarely goes above unity and generally stays below $1/2$.

A stability index of zero is assigned to central-force bonds within isostatic rigid regions, because this is the limiting value for an underconstrained region that is nearly rigid and for an overconstrained region that is nearly isostatic. The intrinsic flexibility of a protein can be studied using the stability index within a one-dimensional representation. Here, the stability index for central-force bonds along the mainchain of a protein, where a plot is made against residue number along the protein sequence, gives a good sense for the mechanical stability.

The number of independent degrees of freedom and redundant constraints within the protein vary gradually as the number of hydrogen bond constraints is varied. Likewise, the stability index also varies gradually when hydrogen bond constraints are added or removed. It is worth mentioning that two independent sum rules result directly from the definition of the stability index. It follows that adding the stability index over all hinge joints sums to the negative of the total number of floppy modes, while summing over all non-hinge joints results in the total number of redundant constraints.

As a simple example, consider a single n-fold ring of atoms that are connected by covalent bonds. From simple constraint counting, the number of degrees of freedom less the number of constraints is given by $F = n - 6$. The number of hinge joints (and central-force bonds) is simply given by n. Therefore the stability index for a n-fold ring is given by

$$s_i = \frac{6-n}{n} \quad \text{for each central-force bond in a } n\text{-fold ring.} \tag{2}$$

Notice that as the ring becomes very large, the stability index goes to the limit of -1, as each dihedral angle is nearly independent and is almost as flexible as a linear chain. For a six fold ring, the stability index is zero, indicating that the generic ring structure is isostatically rigid. For a three fold ring, the stability index works out to be $+1$ and is highly overconstrained. Note that for the case of a n-fold ring, the stability index is bounded between -1 and 1.

As another example, consider once again the hand made 32 atom network. From Figure 8 the assigned value for the stability index to each hinge joint within the black and grey underconstrained regions are $-1/4$ and $-2/5$ respectively, indicating that the underconstrained region on the right (grey) is less stable, or more floppy. Each independent hinge joint has a stability index of -1. From Figure 6 the stability index for each central-force constraint within the network induced overconstrained region is $1/5$. All other central-force bonds, including the resonant bonds, have a stability index of zero.

The Alpha Helix

The alpha helix is a secondary structure[18] that is frequently found in proteins, and they are generally regarded as rigid substructures. The structure of an alpha helix is that of a helix with mainchain hydrogen bonds forming between every fourth consecutive residue as shown in Figure 9. Each turn, through 2π along the alpha helix consists on average of 3.6 residues.

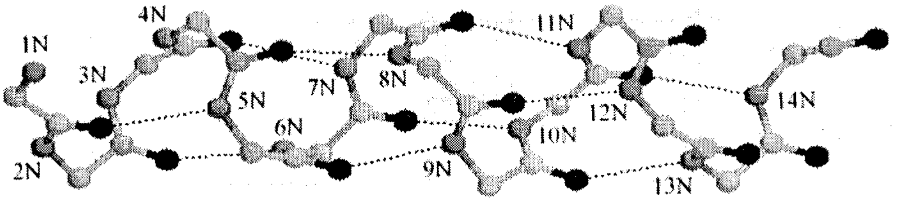


Figure 9. An example of an alpha helix made from 14 glycine residues. The hydrogen bonds are shown as thin black dashed lines.

We will analyze the intrinsic mechanical stability of an alpha helix using our distance constraint approach. Underconstrained, isostaticly rigid and overconstrained regions can easily be calculated because of the linear repeating pattern of the mainchain hydrogen bonds. The stability index will be worked out for an alpha helix made up of entirely glycine residues because we are interested in understanding the effects of the crosslinking hydrogen bonds along the mainchain. The same effects will be present for any alpha helix substructure within a protein regardless of its constituent residue types, although other rigiditifying affects may result from (additional) sidechain or end cap hydrogen bonding.

The internal independent degrees of freedom (floppy modes) and redundant constraints will be counted explicitly by hand. The distance constraints associated with the covalent bonding between atoms along the mainchain allow two free dihedral angles within each residue as shown in Figure 10. Recall that the dihedral angles associated with peptide bonds are locked by third nearest neighbor distance constraints. The two free dihedral angles are conventionally referred to as the $\{\phi_i, \psi_i\}$ angles within the i-th residue. The number of floppy modes for the alpha helix is given by

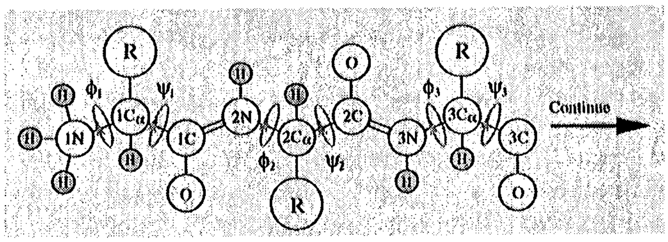


Figure 10. A simple illustration for the mainchain of a protein. Double lines represent the peptide bond for which we do not allow any dihedral rotation. There are two a priori free dihedral angles denoted by ϕ and ψ respectively per residue. In our simple alpha helix, all residues (denoted by R) are glysine, which is equivalent to replacing R by a hydrogen atom.

$$F(n) = (6 + 2n) - 3N_h - 6 \qquad (3)$$

where $(6 + 2n)$ is the number of degrees of freedom for n residues, $3N_h$ is the number of crosslinking distance constraints (three per hydrogen bond), and the 6 trivial rigid body motion degrees of freedom are subtracted out.

The regular repeating pattern of the alpha helix gives $N_h = n - 4$ hydrogen bonds for $n \geq 4$ residues. Since we want the alpha helix to have at least one turn, we will only consider an alpha helix consisting of four or more residues. The hydrogen bond pattern does not crosslink the very ends of the alpha helix. As a result, two independent dihedral angle motions are allowed at both ends of the alpha helix regardless of its length. The region that is of interest to us is along the main body of the helix. Depending on the number of residues in the alpha helix, the region of interest will form one underconstrained, isostatically rigid, or overconstrained region.

The two independent hinge joints at each end of the helix do not belong to the region of interest and must be accounted for separately. To prove that the main body of the alpha helix does not break into multiple regions (assuming all hydrogen bonds are present) requires one to perform an all subgraph constraint counting exercise either manually or with the aid of the FIRST program. Since the structure is quite regular, one can prove this by combining inspection with mathematical induction as an alpha helix is built up one residue at a time.

A global count in the number of floppy modes within a n-residue alpha helix is given by $F(n) = 12 - n$ where Eq. 3 has been used. However, there are five distinct regions within the alpha helix, consisting of the two hinge joints at each end with independent dihedral angles and the single region of interest within the main body. Therefore, the number of floppy modes within the main body of a n-residue alpha helix, $F_\alpha(n)$, is given by:

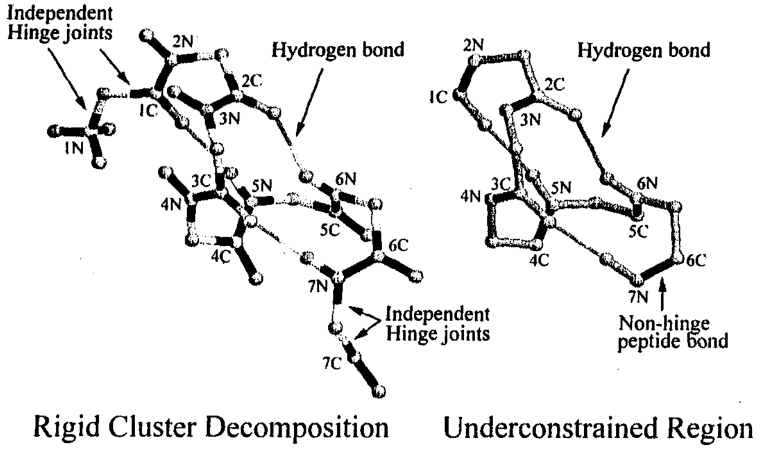
$$F_\alpha(n) = 8 - n \quad \begin{cases} > 0 \text{ for } n < 8 & \text{an underconstrained region} \\ = 0 \text{ for } n = 8 & \text{an isostaticly rigid region} \\ < 0 \text{ for } n > 8 & \text{an overconstrained region} \end{cases} \qquad (4)$$

The rigid cluster decomposition is shown in Figure 11 for a seven residue alpha helix. From Eq. 4 the main body of the alpha helix is floppy with one floppy mode, describing a twisting motion along the axis. In this floppy region there are 21 interconnected rigid clusters that form an underconstrained region consisting of nineteen hinge joints, which is also shown in Figure 11. For a n-residue alpha helix, the number of hinge joints in the region of interest can be counted as

$$H_\alpha = 2n - 4 + 3N_h \;\; = \;\; 5n - 16 \qquad (5)$$

where there are $2n$ hinge joints consisting of the ϕ and ψ rotatable dihedral angles in total. Four hinge joints correspond to the two independent hinge joints at each end of the helix, and recall there are 3 dihedral angles associated with each hydrogen bond as shown in Figure 3.

The stability index for each of the rotatable dihedral angles within the main body of the alpha helix (ie. not those associated with the peptide bond) is given by $s_\alpha = -F_\alpha/H_\alpha = -1/19$. The stability index associated with each of the central-force bonds in the peptide bond is zero because the third neighbor distance constraint



Rigid Cluster Decomposition Underconstrained Region

Figure 11. The rigid cluster decomposition (right panel) and the underconstrained region (left panel) along the main body of an alpha helix of 7 residues is shown.

used to lock the dihedral angle makes a small isostatic rigid cluster. A one dimensional representation of the degree of flexibility along the mainchain of the seven residue alpha helix is shown in Figure 12. A topological representation of the corresponding rigid cluster decomposition and the main body underconstrained region is also shown. Unlike the 32 atom example network worked out earlier, a protein structure can be unfolded into a linear covalent polymer chain and sequenced according to residue in a one dimensional fashion, otherwise known as the primary structure. Therefore, in Figure 12 we plot the degree of flexibility for the central-force bonds along the mainchain of the protein structure against residue number.

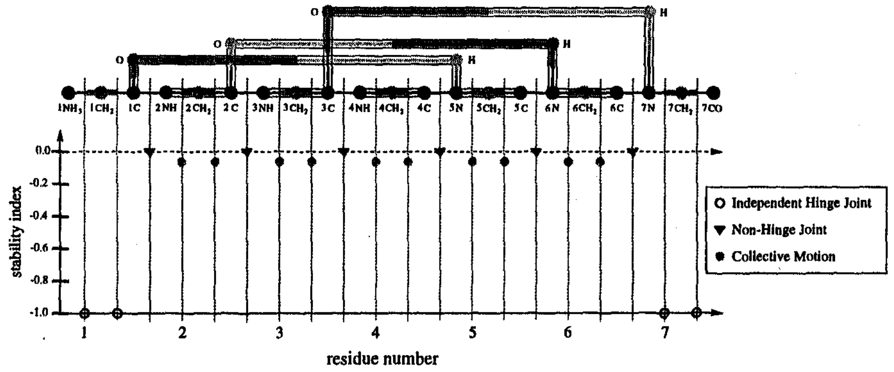


Figure 12. The stability index is shown along the mainchain of the 7 residue alpha helix. To show the correspondence clearly, a schematic diagram for the rigid cluster decomposition (black, grey and light grey regions) and the single underconstrained region (outlined by solid lines) are also shown.

From Eq. 4 it is seen that the main body of the alpha helix will be rigid with

$(n-8)$ redundant constraints for an alpha helix having eight or more residues. In this region, the total number of central-force constraints is given by

$$C_\alpha = H_\alpha + (n-1) \;\; = \;\; 6n - 17 \tag{6}$$

where H_α (from Eq. 5) is used because it represents the number of central-force constraints associated with a priori rotatable dihedral angles, and the addition of $(n-1)$ corresponds to the number of (non-rotatable) peptide bonds between residues. Unlike within underconstrained regions where hinge joints are counted, we count all central-force bonds within a rigid region. Using Eq. 1, the stability index for each central-force bond associated with the $\{\phi, \psi\}$ angles along the mainchain and within the main body of a n-residue alpha helix works out to be

$$s_\alpha(n) = \begin{cases} \frac{n-8}{5n-16} & \text{for } 4 \leq n \leq 8 \\ \frac{n-8}{6n-17} & \text{for } n > 8. \end{cases} \tag{7}$$

It is worth noting that for long alpha helices (ie. in the limit $n \to \infty$), the stability index approaches 1/6 corresponding to one redundant constraint per residue.

Lysine/Arginine/Ornithine-Binding Protein

The lysine/arginine/ornithine binding protein is a bacterial periplasmic protein found in *Salmonella typhimurium*, and is responsible for the transport of various substrates (or ligands) such as lysine from the periplasm to the cytoplasm[35]. The X-ray crystal structures of the lysine binding protein in the closed conformation with lysine bound and in its open conformation without a ligand are known[35] to within 1.9 Å. This protein has two globular domains connected by a double flexible linkage, which allows a large-scale hinge-type ("clam shell") movement of its two domains[6,14]. Each globular domain essentially define the two halfs of the "clam shell" separated by loop hinges as defined by the crystallographers[35]. Domain 1 contains residues between 1-87 and 195-237, consisting of parts of the N- and C-termini. Domain 2 contains all residues between 94-181.

By comparing the ϕ and ψ mainchain dihedral angles in the protein structure in its closed and open conformations, given in the *1lst* and the *2lao* PDB files, (from the Brookhaven Protein Data Bank[2] of atomic resolution macromolecular structures) it is ascertained that the large-scale domain motion is localized to a few dihedral angles within a flexible linkage consisting of two loops[35]. From kinetic studies[36], the time scale for the opening and closing of the structure is found to be 10 nanoseconds without the substrate and 10 miliseconds with the substrate – a difference of six orders of magnitude. We selected this protein as an initial test case in applying FIRST to predict intrinsic flexibility.

Hydrogen atoms are generally not resolved in a protein structure derived from crystallography, because their low electronic densities do not contribute much to the scattering of X-rays. As we consider generic networks, the exact position of the hydrogen atoms is not important. However, the placement of hydrogen atoms is useful for identifying hydrogen bonds with good geometry[37] within the protein structure. Therefore, we use the WhatIf software[38] for geometrically correct optimal placement of hydrogen atoms for hydrogen bonding.

Once the hydrogen atom positions are explicitly provided, we model hydrogen bonds as distance constraints within the bond network analyzed by FIRST, provided they satisfy the following criteria:

1. The hydrogen bond is intra-molecular.
2. The donor-hydrogen-acceptor angle must be greater than 120 degrees.
3. Either the donor-acceptor distance is less than 3.5 Å or the hydrogen-acceptor distance is less than 2.5 Å.

It is possible to consider additional hydrogen bonds that form between the protein and solvent (water) molecules, and between the protein and other ligands. However, here we will focus on the effect of intra-protein hydrogen bonds on rigidity.

In Figures 13 and 14, we show the one dimensional representation for the intrinsic flexibility of the lysine-binding protein in the closed conformation, with lysine bound, and in the open conformation without lysine. In both figures, the stability index associated with the $\{\phi_i, \psi_i\}$ angles in each residue along the mainchain is plotted against residue number. The one dimensional representation alone gives a clear overview of the rigid and flexible regions, and can be compared directly with the three-dimensional molecular graphics of those regions.

It can be seen that the gross features in mechanical stability of each conformation are the same. Most of the rigid substructures and underconstrained regions identified in the two conformations correspond to one another.

Domain 1, consisting of both the N- and C-terminus of the lysine binding protein, is seen to be mechanically more stable than domain 2 consisting of residues between 94 to 181 (between the inner two vertical dashed lines). Both conformations show that domain 1 is almost completely rigid and made up almost entirely by rigid cluster R1. Domain 2 contains a large underconstrained region (labeled as U2) with several rigid substructures (labeled as R2, R3, ... R8) that are mainly alpha helices.

For the same hydrogen bond criteria, it is seen that the intrinsic flexibility is greater for domain 2 than that for domain 1. The large underconstrained region (labeled as U2) consists of the known double flexible linkage, residues 88-93 and 182-194, in addition to a beta sheet in domain 2. Thus, domain 2 apparently does not have enough sidechain hydrogen bonding to rigidify the beta-sheet together with other rigid substructures to form a large rigid domain. Instead, the various rigid substructures move collectively. The underconstrained region U2 decreases in size as the hydrogen bond selection criteria are relaxed and more hydrogen bonds are modeled as constraints. Eventually, the beta-sheet region within domain 2 rigiditifies in both the closed and open conformations, making the floppy motion in the double flexible linkage more pronounced. These results taken together suggest that during conformational changes of the lysine binding protein, some motion will take place within domain 2.

Differences found in the stability index between the closed and open conformations are caused predominately by differences in hydrogen bonding. This must be the case since the covalent bonding is identical in each conformation with the sole exception of the presence of the lysine substrate in the closed conformation, which itself is non-covalently bound. It is important to discriminate between real physical differences as a result of the presence of the lysine substrate, in contrast to the sensitivity of our approach in selecting a particular set of hydrogen bonds that are modeled as distance constraints. We consider differences as physical when the results are robust against small perturbations in the hydrogen bond selection criteria.

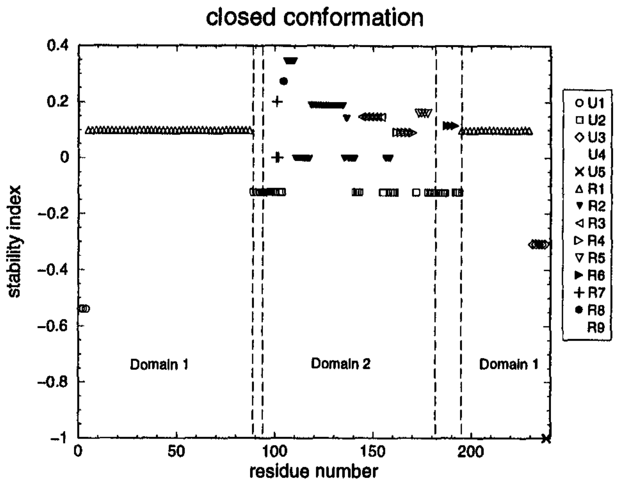


Figure 13. The stability index associated with the ϕ and ψ dihedral angles (hinge joints) along the mainchain is shown for the lysine binding protein in its closed conformation (PDB code 1lst). In the legend on the right, U1-U5 represent underconstrained regions and R1-R9 represent rigid substructures that FIRST identified. The labeling scheme and choice of symbols used for the underconstrained regions and rigid substructures are consistent with the open conformation shown below to aid in direct comparisons. Note that a single rigid region (eg. R1) can be discontinuous in sequence number along the mainchain. The two sets of dashed lines locate the double flexible linkage between residues 88-93 and 182-194

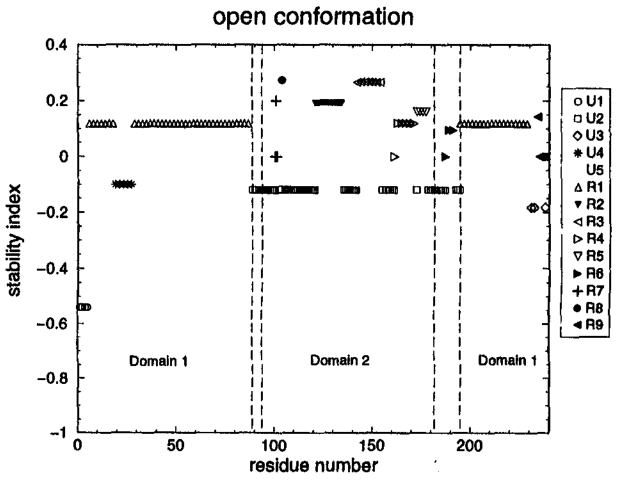


Figure 14. The stability index associated with the ϕ and ψ dihedral angles (hinge joints) along the mainchain is shown for the lysine binding protein in its open conformation (PDB code 2lao). The legend on the right and the dashed lines are the same as used above for the closed conformation.

There is one main difference in the intrinsic flexibility between the closed and open conformations that is worth mentioning here. This difference is associated with the underconstrained region involving the double flexible linkage (labeled U2) that extends into domain 2. There is also a corresponding difference in a rigid substructure (labeled as R2). In both conformations rigid substructure R2 consists of a relatively stable region that is overconstrained. In the closed conformation it has some additional regions that are (mostly) isostaticly rigid to the stable base, while in the open conformation these regions remain floppy and are part of underconstrained region U2. Thus in the open conformation, underconstrained region U2 is larger in size, consisting of 300 rotatable dihedral angles from which only 36 are independent. In the closed conformation underconstrained region U2 has 219 rotatable dihedral angles from which only 27 are independent. Note that the count in dihedral angles include the sidechain covalent bonding as well as hydrogen bonding.

As the hydrogen bond selection criteria is relaxed or tightened slightly, it is observed that underconstrained region U2 in the closed conformation is always smaller than that in the open conformation. We therefore suspect that this result is a robust signature for the presence of the lysine substrate, and that domain 2 is mechanically more stable when a substrate is present. This rigiditifying effect from the substrate could be a contributing factor in explaining why the time scale is six orders of magnitude slower for the hinge motion when the substrate is present. This is an exciting prospect as this suggests the possibility of identifying allosteric effects directly from characterizing the mechanical properties of a structure.

In our preliminary investigation on the lysine-binding protein, the FIRST analysis already has provided new insights as well as correlated with known motions within the protein. The next step is to implement the hierarchical protocol in modeling hydrogen bonds as distance constraints from strongest to weakest using a hydrogen bonding energy term.

SUMMARY

In this article, we have introduced a novel distance constraint approach for characterizing the intrinsic flexibility of a protein. The conceptual framework is to develop a hierarchical protocol for adding more distance constraints that model successively weaker bonding forces. The underlying physical and mathematical assumptions behind this idea were layed out. We then focused on the first step toward our goal of developing a robust automated program, FIRST, that can characterize the mechanical stability of a protein. Namely, we have shown the application of concepts from graph rigidity to identify a series of mechanical properties of a protein structure under a given set of bonding forces modeled as distance constraints.

The initial implementation of FIRST determines the floppy inclusion and rigid substructure topography of a given protein structure based on a single set of distance constraints as determined by covalent and hydrogen bonding. The *stability index* is introduced as a *continuous* measure for quantifying the local mechanical stability of each dihedral angle. When the measure falls between $(-1, 0)$, this physically corresponds to a dihedral angle within an underconstrained region that is anywhere from being completely independent to nearly isostatically rigid. A zero value indicates that the dihedral angle is within an isostaticly rigid region. A stability index that is greater than zero physically corresponds to a dihedral angle within an overconstrained region, where stability is enhanced as the number of redundant constraints within the region increases.

There are several advantages of FIRST relative to previous methods for analyzing protein flexibility. The FIRST analysis requires calculations that can be done in real-time (fraction of a second). For a given set of distance constraints the rigidity is calculated exactly, which includes identifying overconstrained and underconstrained regions in addition to obtaining a rigid cluster decomposition. The ability to determine underconstrained regions gives the FIRST analysis a distinctive advantage over other methods, because these regions localize important collective motions, whereby changing one dihedral angle will generally change other dihedral angles within the same underconstrained region. The stability index is introduced as a continuous measure to give meaningful one-dimensional representations for the mechanical stability of a protein structure, and this has already provided new insights on the lysine-binding protein.

Acknowledgements

The work by DJ and MT was supported in part by NSF grant DMR-96 32182, the work by LK and MT was supported in part by The Center for Protein Structure, Function, & Design at Michigan State University. DJ and MT would also like to thank John Grinstead for help in working out the stability index for regular secondary structures.

REFERENCES

1. D.J. Jacobs and M.F. Thorpe. US Patent pending: *Computer-implemented System for Analyzing Rigidity of Substructures Within a Macromolecule* (1998)
2. www.pdb.bnl.gov is the electronic data base. Also see, E.E. Abola, F.C. Bernstein, S.H. Bryant, T.F. Koetzle and J. Weng, *Protein Data Bank* 107-132, in *Crystallographic Databases - Information Content, Software Systems, Scientific Applications*, Editors: F.H. Allen, G. Bergerhoff and R. Sievers, (Data Commission of the International Union of Crystallography) (1987)
3. D. Ringe and G.A. Petsko, *A Consumer's Guide to Protein Crystallography* 210-229, in *Protein Engineering and Design*, Editor: P.R. Carey. (Academic Press) San Diego (1996)
4. W.S. Bennett and R. Huber, *Crit. Rev. Biochem.* **15**, 291 (1984)
5. K. Wuthrich and G. Wagner, *Trends Biochem. Sci.* **3**, 227 (1978)
6. M. Gerstein, A.M. Lesk and C. Chothia, *Biochem.* **33**, 6739 (1994)
7. W.L. Nichols, G.D. Rose, L.F. Ten Eyck and B.H. Zimm, *Proteins* **23**, 38 (1995)
8. A.S. Siddiqui and G.J. Barton, *Protein Sci.* **4**, 872 (1995)
9. N.S. Boutonnet, M.J. Rooman and S.J. Wodak, *J. Mol. Biol.* **253**, 633 (1995)
10. L. Holm and C. Sander, *Proteins* **19**, 256 (1994)
11. M.H Zehfus and R.D. Rose, *Biochem.* **25**, 5759 (1986)
12. P.A. and G.E. Schulz, *Naturwissenschaften* **72**, 212 (1985)
13. B.W. Matthews, *Ann. Rev. Biochem.* **62**, 139 (1993)
14. V. Maiorov and R. Abagyan, *Proteins* **27** 410 (1997)
15. Y. Bourne, M.H. Watson, M.J. Hickey, W. Holmes, W. Rocque, S.I. Reed and J.A. Tainer, *Cell* **84**, 863 (1996)
16. A. Askar, B. Space and H. Rabitz, *J. Phys. Chem.* **99**, 7330 (1995)
17. R. Abagyan, M. Totrov and D. Kuznetsov, *J. Comp. Chem.* **15**, 488 (1994)
18. C. Branden and J. Tooze. *Introduction to Protein Structure* Garland Publishing, New York and London (1991)
19. J. Janin and S.J. Wodak, *Prog. Biophys. Molec. Biol.* **42**, 21 (1983)
20. G.A. Jeffrey and W. Saenger. *Hydrogen Bonding in Biological Structures* Springer-Verlag, Germany (1991)

21. A.R. Fersht, *Trends Biochem. Sci.* **87**, 301 (1987)
22. M.F. Thorpe, B.R. Djordjevic and D.J. Jacobs, *Amorphous Insulators and Semiconductors* M.F. Thorpe and M.I. Mitkova, ed. NATO ASI Series 3: High Technology **23**, Kluwer Academic (1997): M.F. Thorpe, D.J. Jacobs and B.R. Djordjevic, *Insulating and Semiconducting Glasses* P. Boolchand, ed, World Scientific (1998): Also see article within this book. M.F. Thorpe, *et al.*
23. K.A. Dill, *Biochem.* **29**, 7133 (1990)
24. D. Franzblau, Private communication (1996)
25. J. Graver, B. Servatius and H. Servatius *Combinatorial Rigidity (Graduate Studies in Mathematics)* American Mathematical Society, Providence RI (1993)
26. E. Guyon, S. Roux, A. Hansen, D. Bibeau, J.P. Troadec and H. Crapo, *Rep. Prog. Phys.* **53** 373 (1990)
27. G. Laman, *J. Eng. Math.* **4** 331 (1970)
28. D.J. Jacobs and B. Hendrickson, *J. Comp. Phys.* **137** 346 (1997)
29. D.J. Jacobs, *J. Phys. A: Math. Gen.* **31** 6653 (1998)
30. F. Harary. *Graph Theory* Addison-Wesley, Reading MA (1969)
31. D.J. Jacobs and M.F. Thorpe, *Phys. Rev. Lett.* **75** 4051 (1995)
32. D.J. Jacobs and M.F. Thorpe, *Phys. Rev. E* **53** 3683 (1996)
33. D. C. Kozen. *The Design and Analysis of Algorithms* Chapter 4, Springer-Verlag, New York (1992)
34. J. Saks, American Mathematical Society, #441, 91 (1991)
35. B. Oh, J. Pandit, C. Kang, K. Nikaido, S. Gokcen, G.F. Ames and S. Kim, *J. Biol. Chem.* **268** 11348 (1993)
36. D.M. Miller III, J.S. Olson, J.W. Pflugrath and F.A. Quiocho, *J. Biol. Chem.* **258** 13665 (1983)
37. S.M. Habermann and K.P. Murphy, *Protein Science* **5** 1229 (1996)
38. http://swift.embl-heidelberg.de/whatif/