PROTEINS: Structure, Function, and Genetics 44:150–165 (2001)

Protein Flexibility Predictions Using Graph Theory

Donald J. Jacobs,[1] A.J. Rader,[1] Leslie A. Kuhn,[2*] and M.F. Thorpe[1]
[1]*Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan*
[2]*Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan*

ABSTRACT **Techniques from graph theory are applied to analyze the bond networks in proteins and identify the flexible and rigid regions. The bond network consists of distance constraints defined by the covalent and hydrogen bonds and salt bridges in the protein, identified by geometric and energetic criteria. We use an algorithm that counts the degrees of freedom within this constraint network and that identifies all the rigid and flexible substructures in the protein, including overconstrained regions (with more crosslinking bonds than are needed to rigidify the region) and underconstrained or flexible regions, in which dihedral bond rotations can occur. The number of extra constraints or remaining degrees of bond-rotational freedom within a substructure quantifies its relative rigidity/flexibility and provides a flexibility index for each bond in the structure. This novel computational procedure, first used in the analysis of glassy materials, is approximately a million times faster than molecular dynamics simulations and captures the essential conformational flexibility of the protein main and side-chains from analysis of a single, static three-dimensional structure. This approach is demonstrated by comparison with experimental measures of flexibility for three proteins in which hinge and loop motion are essential for biological function: HIV protease, adenylate kinase, and dihydrofolate reductase. Proteins 2001;44:150–165.**

Key words: conformational change; mobility and dynamics; coupled/collective motions; hydrogen-bond networks; distance constraints; dihedral angle constraints and rotations; structural stability; dihydrofolate reductase; adenylate kinase

INTRODUCTION

More than 15,000 protein structures have been determined to date, using X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy.[1] Such structures are deduced from sophisticated refinement procedures in conjunction with stereochemical modeling.[2] Proteins can be described as a collection of stable fragments,[3] ranging in size from a small number of residues to an entire domain. Different packings of the protein molecules in alternative crystal forms can trap the protein in different conformational states, providing snapshots of some of the conformations accessible to the protein.[3] NMR spectroscopy can also indicate dynamic regions of proteins by showing that several conformations are consistent with the experimental constraints, typically inter-proton distances.[4]

Other computational procedures have been developed[5–10] to characterize the intrinsic flexibility and rigidity within a protein. These procedures fall within three classes. One approach compares different conformational states, e.g., from different crystallographic or NMR conformations observed for the protein, to deduce which regions in the protein are flexible or rigid.[5–7] A second approach focuses on simulating a protein's motion, by means of molecular dynamics calculations, using a forcefield describing interatomic potentials.[11–14] The third class focuses on identifying rigid protein domains or flexible hinge joints[8–10,15,16] based on a single conformation. Each method has its limitations. In the first class, the methods are limited by the diversity of the conformational states that are available from experiment for comparison, whereas the second class is limited by the computational time involved, meaning that large motions of the protein are usually not sampled. The third class, to which our approach belongs, defines the rigid and flexible regions in the protein from a single conformation and can be used as the starting point to sample a range of motions. These methods are generally computationally fast and can provide a starting point for more efficient molecular dynamics or Monte Carlo sampling of protein conformations, as the number of degrees of freedom in the protein is reduced significantly by defining which regions are rigid. A significant question is, as always: to what extent do such methods for defining rigid/flexible regions correlate with what is observed experimentally?

The bonds within a protein can be represented as a network in which the covalent forces and strong hydrogen bonds are modeled as distance constraints between atoms. The mechanical stability of the corresponding bond net-

Grant sponsor: National Science Foundation; Grant number: DMR-96 32182; Grant number: DBI-96 00831; Grant sponsor: National Institutes of Health; Grant number: R43 GM58337-01; Grant sponsor: Michigan State University; Grant sponsor: American Heart Association; Grant number: 9940091N.

Donald J. Jacobs is currently at the Department of Physics and Astronomy, California State University Northridge, Northridge, CA 91330.

Leslie A. Kuhn and M.F. Thorpe are contributing authors.

*Correspondence to: Leslie A. Kuhn, Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI. E-mail: kuhn@agua.bch.msu.edu.

Received 23 September 2000; Accepted 15 March 2001

work of a protein can then be analyzed using new graph theoretical techniques that were originally developed for analyzing the rigidity of substructures within covalent network glasses.[17–19] The computer implementation[20] of this procedure is referred to as Floppy Inclusion and Rigid Substructure Topography (FIRST), which provides a real-time tool for evaluating the intrinsic flexibility within a protein. FIRST gives the precise mechanical properties of a protein structure under a given set of constraints. This approach defines not only the rigid regions in a protein, but also those regions that move collectively (whose motions are coupled), as well as those that move independently of other regions in the structure. Furthermore, the relative flexibility or rigidity of each region is quantified, based on the density of bonds remaining rotatable in each flexible region. This article applies this approach to defining the flexible regions and their relationship to ligand binding in three diverse proteins.

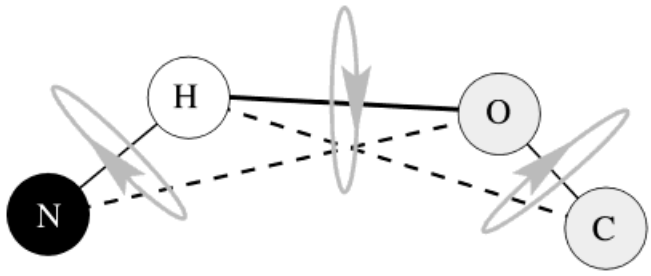

Fig. 1. Model of a hydrogen bond, involving donor and acceptor atoms, shown as nitrogen and oxygen, respectively. Covalent bonds are shown as thin black lines. The hydrogen bond is modeled as three distance constraints, consisting of a nearest-neighbor central-force constraint shown as a thick solid line (top center), and two next-nearest-neighbor bond-bending force constraints (constraining the donor–hydrogen–acceptor angle), shown as dashed lines. Each hydrogen bond is also associated with three a priori rotatable dihedral angles, indicated by the arrows.

METHODS

Both bonding and nonbonding forces play an important role in determining the structure of a protein and the dynamics about the native fold. The nonbonding forces are both short- and long-range. Although both hydrophobic interactions and van der Waals forces, as well as long-range electrostatic forces, play an important role in stabilizing a protein structure in the native state,[21] these forces are generally weak between pairs of atoms and not highly directional, and thus are not represented as interatomic distance constraints. Strong bonding forces, such as covalent and hydrogen bonds, tend to be directional and are modeled in the present study as distance and angle constraints.

Constraints

The covalent bonding within the protein resulting from bond-stretching (central), bond-bending, and torsional forces defines a natural set of interactions that can be modeled as distance constraints. It is common practice to represent the degrees of freedom accessible to a protein by fixing the covalent bond lengths and associated bond angles, while allowing the dihedral angles to rotate. The torsional forces associated with peptide bonds and the other partial-double and double bonds in proteins effectively prevent dihedral rotation about the bond and are also represented as constraints. Using the rotatable dihedral angles as a set of internal coordinates, the number of degrees of freedom to describe the flexibility of a protein is typically reduced by a factor of about 7 relative to a Cartesian representation.[22]

In addition to covalent bonding forces, hydrogen bonds (Fig. 1) have high directional dependence and act over short distances, in contrast to the hydrophobic forces, which are less specific and may be regarded as slippery. By this, we mean that the energy associated with the hydrophobic forces does not change significantly for gentle conformational shifts away from the native state. It is reasonable to expect that the buried hydrogen bonds will be substantially maintained as the protein undergoes conformational changes near its native structure. Recent results by Lu and Schulten[23] suggest that the breaking of a hydrogen bond occurs as a well-defined event, involving going over an energy barrier, as opposed to a continuous stretching until a feeble final breaking occurs. These investigators suggest that hydrogen bonds typically break one by one as the protein unfolds, while in some cases a consortium of hydrogen bonds break simultaneously, giving a much higher effective barrier.

Rigidity in Networks

Our graph-theory algorithm[20,24] for analyzing proteins is a three-dimensional (3-D) extension and implementation of results in mathematical rigidity theory that have developed over the past few years. The roots of this work go back to Lagrange's[25] introduction of constraints on the motion of mechanical systems during the late eighteenth century, which Maxwell[26] used during the late nineteenth century to determine whether structures were stable or deformable. The applications of this type of work have traditionally been to solve problems in engineering, such as the structural stability of different truss configurations in bridges. A very significant advance occurred with Laman's theorem[27] in 1970, which precisely determines the degrees of freedom within two-dimensional (2-D) networks, and allows the rigid regions and flexible joints between them to be found. The most general type of 3-D network for which results can be calculated are the so-called *bond-bending networks* (or, equivalently, *molecular frameworks*), in which vertices are connected by edges and in which every angle between edges is defined. A broader class of networks is the bar-joint framework, in which the angles are not specified; analyzing flexibility and rigidity for this case remains a significant unsolved problem in mathematics. For 3-D bond-bending networks, the flexibility in the system derives from dihedral or torsional rotations of the bonds that are not locked in by the network, and this kind of framework is used to represent the constraints within a protein.

First, we present a general algorithm using brute-force matrix diagonalization to identify all the rigid clusters in a

framework consisting of a network of atoms with interconnecting bonds. The fundamental step on which all such calculations are based is the ability to test whether a constraint (bond between atoms) is redundant or independent. A constraint is considered redundant if breaking it has no effect on the flexibility of the network, and independent if its breakage does affect the flexibility. This can be determined using the following procedure:

1. Define a network of atoms and distance constraints between them.
2. Replace each distance constraint by a spring. All spring constants are set to unity in arbitrary units, and the natural length of each spring is set equal to the distance between the associated pair of atoms.
3. Construct a dynamical matrix[28] for the spring network, and calculate all the normal mode frequencies (eigenvalues) of this network. The eigenvalues indicate whether the mode of vibration has a finite or zero frequency.
4. Count the number of zero eigenvalues, corresponding to zero-frequency vibrational modes, or floppy modes, of the system.
5. Add to the system the distance constraint being tested for redundancy or independence, and repeat the above steps.
6. If the number of zero eigenvalues remains the same, the added constraint is redundant; otherwise, it is independent.

This procedure is repeated, using steps 5 and 6 above, for each constraint in the network that is to be tested for redundancy or independence. Afterwards, in order to identify all the rigid clusters of atoms within the network, called a rigid cluster decomposition, a test constraint is placed between each pair of atoms. Only test constraints found to be redundant are added to the network. Once all pairs of atoms have been tested and all redundant constraints added, the rigid clusters are identified by selecting all atoms that are connected by a contiguous path of face-sharing tetrahedrons, which form a rigid pathway. Tetrahedra are the 3-D generalization of triangles, where a pathway of edge-sharing triangles forms a rigid path in two dimensions.

The implementation described above can be used on any type of 3-D framework, but the resulting computational performance scales as $O(N^2 \times N^3)$, where N is the number of vertices (atoms) in the framework, rendering it useless for application to protein structures with thousands of atoms. Also of interest are the regions that have more distance constraints than are required for rigidity, that is, that have redundant constraints within the rigid clusters. When redundant constraints are present, this creates regions of internal strain. In order to measure the strain, an additional separate numerical algorithm must be used to relax the network and calculate the stress in the springs. These methods, such as conjugate gradient, typically scale as $O(N^2)$.

Clearly, the distance constraint approach will only be useful in as much as the calculation for identifying rigid clusters and flexible regions is fast and scales linearly or nearly linearly with system size. The two main contributions of the present article are to show how to use concepts from graph rigidity[29] to make this type of calculation feasible and then to present applications of this approach to three diverse proteins, showing that their predicted conformational flexibility correlates well with known, biologically relevant flexible regions in these proteins.

Generic Rigidity

A generic structure is one that has no special symmetries, such as parallel bonds or bond angles of 180°, that could create geometric singularities.[30] Under these conditions, the study of rigidity becomes much easier because the properties of network rigidity depend only on the connectivity as determined by an underlying graph. Laman's theorem[27] states how generic rigidity within any 2-D generic framework can be characterized completely by applying constraint counting to all the subgraphs within the framework. Applying Laman's theorem directly leads to a combinatorial calculation that scales exponentially in the number of atoms within the framework. However, by applying Laman's theorem recursively, a very fast and efficient integer algorithm (the so-called "pebble game") has been constructed[17,31,32] for identifying rigid clusters and stressed regions. The computational complexity of the pebble game scales in the worst case as $O(N^2)$ for pathological networks, where N is the number of vertices or atoms in the network, and scales linearly in practice. The algorithm also scales linearly with N in computer memory.

Unfortunately, it has been a longstanding problem that in 3-D generic bar-joint frameworks, constraint counting over the subgraphs is known to fail in general; that is, Laman's theorem does not simply generalize to 3-D bar-joint frameworks. However, for the special class of truss frameworks, in which the bars and joints have extent and the bars are subject to bond-bending angle constraints, Laman's theorem can be generalized[18,24] to three dimensions. In addition, the molecular framework conjecture proposed by Tay and Whiteley[33,34] indicates that Laman constraint counting extends to nongeneric molecular models in which all bonds (hinges) of an atom pass through a single central point of the atom. Although the molecular framework conjecture requires a rigorous proof, there are no known exceptions after years of exact testing. Thus, we model the microstructure of a protein using distance constraints that define a bond-bending network and apply the 3-D pebble game to determine precisely the rigid clusters, stressed regions, and internal degrees of freedom in terms of dihedral angles.

Pebble Game

At the heart of the FIRST computer program designed to analyze protein structures is the 3-D pebble game algorithm, constructed in a very similar way to its 2-D counterpart.[17,18,31,32] The pebble game is an implementation of the counting implicit in Laman's theorem in two dimensions, as well as its 3-D generalization. Here, three pebbles are assigned to each vertex or atom, representing the 3

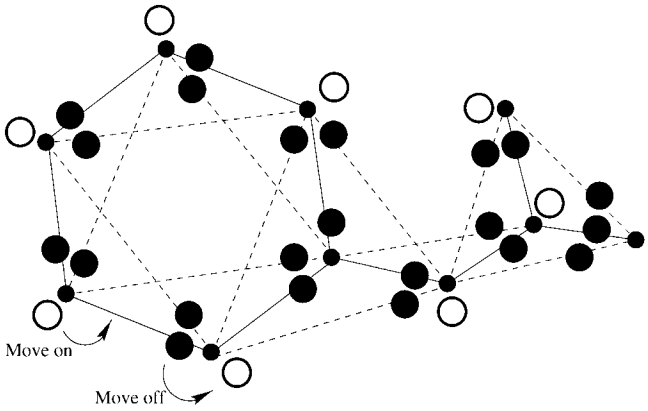

Fig. 2. Diagrammatic representation of the final pebble covering of a simple network. Free pebbles are placed directly on vertices and denote degrees of freedom (○). Pebbles covering a bond are placed directly on the edges of the graph and represent distance constraints (●). The pebble covering is not unique, because pebbles can be rearranged according to a few simple rules as explained in the text. An example is shown of how an elementary pebble exchange works (two arrows). A free pebble can be moved onto a covered edge (adjacent to a vertex), provided that a corresponding pebble, presently covering the edge and associated with the neighboring vertex, is moved off.

degrees of freedom of the atom. Each bond is represented by a distance constraint, or edge. For each independent edge between two vertices, a pebble from one of the two vertices must be used to cover the edge. Pebbles that remain associated with vertices are called free pebbles, and they represent the independent degrees of freedom remaining to this vertex within the framework. Each independent edge thus uses up one independent degree of freedom from a vertex. Then the following covering rule is applied: once an edge is covered by a pebble (forming an independent constraint), it must always remain covered by one of the pebbles associated with its incident vertices. Rearrangement of pebbles throughout the graph is possible, provided that this covering rule is maintained and that three pebbles remain associated with each vertex, as free pebbles or as pebbles associated with edges coming from the vertex. Figure 2 shows a pebble covering for a network. The pebble covering is a convenient way to indicate the degrees of freedom accessible to atoms, and to mark which constraints are independent and which are redundant in the bond network. Essentially the redistribution of pebbles from vertices to bonds, by a series of stepwise pebble exchanges, is an intuitive way of representing a dynamically changing directed graph.[24,31]

The 3-D pebble game is a recursive algorithm. The framework is built up by adding one distance constraint at a time, until the final framework is complete. To maintain a bond-bending network, each bond-stretching distance constraint having incident vertices v_1 and v_2 has associated with it angular (i.e., second-nearest-neighbor) constraints about both vertices. In the case of proteins, these angular constraints correspond to the angles of bond coordination about an atom, defined by its chemistry. For each new independent distance constraint introduced, pebbles are rearranged to test whether the new constraint is independent; this is indicated by whether a pebble can be moved to that constraint from one of its vertices. If the new distance constraint is independent, it is then covered by a pebble; otherwise, it is not covered. This process continues until all distance constraints within the network have been placed and tested for independence. The algorithm can be sketched as follows:

1. Begin with a set of unconstrained vertices, each with three pebbles, representing the three degrees of freedom for each vertex in 3-D space.
2. Place a central-force distance constraint between vertices v_1 and v_2, thereby building the framework up one distance constraint at a time.
3. Rearrange the pebble covering, if necessary, to collect three pebbles on vertex v_1.
4. Rearrange the pebble covering to collect as many pebbles as possible (where three is the maximum) on vertex v_2, while holding the three pebbles at vertex v_1.
5. If the number of pebbles on vertex v_2 is two, the edge is redundant. Otherwise, three pebbles reside at vertices (v_1 and v_2) and both remain independently mobile. Continue to rearrange the pebble coverings:
   a. Hold the three pebbles on vertex v_1 and the three pebbles on vertex v_2.
   b. For each neighbor of vertex v_2, attempt to collect a pebble from a neighboring bond, restoring a pebble to that bond from its other vertex.
   c. If for any neighbor of vertex v_2 a pebble cannot be obtained, the edge joining the two vertices is redundant.
   d. If the edge is not redundant, cover it with a pebble from vertex v_2.

Unlike the 2-D pebble game, the distance constraints cannot be placed in random order. The first distance constraint that is introduced must correspond to a central-force constraint.

After each central-force distance constraint is placed, all its associated angular or bond-bending constraints (next-nearest-neighbor distance constraints) must be placed before another central-force constraint can be introduced. Within this restriction, the order of placing either central force or the associated bond-bending constraints is completely arbitrary, and the resulting decomposition of the network into rigid clusters is unique. This restriction on recursively placing constraints is sufficient[18,24] for constraint counting to remain valid in characterizing the rigidity of 3-D bond-bending networks within proteins or other structures. Finally, torsional constraints for the peptide and resonant bonds are fixed by third-nearest-neighbor distance constraints (e.g., fixing the distance between the amide H and carbonyl O in peptide bonds). These are most conveniently placed after the central and bond-bending distance constraints.

After all 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 framework. 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 pebble-covering rule. 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 framework, one can identify overconstrained regions, rigid clusters, and underconstrained regions, as described below.

Strained 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 placing an additional distance constraint between a pair of atoms that have a predefined fixed distance. Placing a distance constraint between this pair of atoms generally causes a length mismatch and leads this region to become internally strained. Thus, a failed pebble search identifies overconstrained regions. Overconstrained regions always consist of closed loops. Because distance constraints are added to the framework, more overconstrained regions will be found, and generally these regions will overlap. Overlapping, overconstrained regions merge together into a single overconstrained region. As these frameworks are generic, stress will propagate and redistribute throughout the merged overconstrained regions. Redundant bonds reside within overconstrained regions. Therefore, the more redundant bonds that are present within a given rigid region, the more stable that region will be against removal of constraints, e.g., crosslinking salt bridges or hydrogen bonds.

Rigid Cluster Decomposition

The method used to identify the rigid clusters, including overconstrained regions, is very simple once all edges in the graph are in place and the pebble game is finished. All rigid clusters can at most have six free pebbles distributed over the vertices within the cluster. Therefore, to identify these clusters, select a vertex and two of its bonded nearest-neighbor vertices. Collect three pebbles on the selected vertex and two pebbles and one pebble, respectively, on its two neighboring vertices. Because we are considering bond-bending networks, it is always possible to collect these six pebbles, and never any more. Mark these three vertices. Then, iteratively in a breadth-first search, check all bonded, unmarked nearest neighbors to the current set of marked vertices, to see whether a free pebble can be obtained. If a free pebble cannot be obtained, mark the new vertex, and note that it is part of the same rigid cluster. This method works because all rigid clusters in bond-bending networks are contiguous through bonded nearest neighbors[17,20]; this point is essential and implicit in the generalization of Laman's theorem to 3-D bond-bending networks.

The rigid cluster decomposition using the 3-D pebble game has been compared with the numerical brute force method described earlier. For a variety of generic bond-bending networks containing as many as 450 atoms, exact agreement has always been found.[35] It is worth mentioning that in contrast to the numerical method, the 3-D pebble game can be used on networks with more than 10 million atoms with the security of knowing the results are exact, because the pebble game is an integer counting algorithm, eliminating the possibility of any numerical round-off errors. Therefore, even the largest proteins and protein complexes can be analyzed both precisely and rapidly.

Underconstrained Regions

Locating rotatable bond dihedral angles in the protein, which correspond to hinge joints in the network, is an easy task after the rigid cluster decomposition is made. Note that a hinge joint can never occur about a next-nearest-neighbor distance constraint, which corresponds to a bond-bending angle.[17] If the two incident vertices of a bond-stretching constraint belong to different rigid clusters, a dihedral angle rotation is possible, and the bond is recorded as a hinge joint; otherwise, the dihedral angle motion is locked, as it is part of a rigid cluster. The number of rotable dihedral angles will generally be considerably more than the number of residual internal degrees of freedom in the network. Not all the rotatable dihedral angles associated with hinge joints are independent, as they are part of a ring of bonds (e.g., within loops formed by crosslinking hydrogen bonds).

Collective motions consist of coupled dihedral angles within the protein and take place in underconstrained regions. Distinct underconstrained regions are partitioned such that collective motions can occur within one underconstrained region without directly affecting internal coordinates within all the other underconstrained regions. The underconstrained regions are identified by attempting to specify a value for each dihedral angle and determining whether it can be satisfied. 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 finding these distinct underconstrained regions is the same as that for finding the overconstrained regions, except that now the only constraints placed in the network are the external torsional constraints.

Proteins as Bond-Bending Networks

The connectivity of a bond-bending network is completely defined by the nearest-neighbor bond-stretching (central-force) constraints. Next-nearest-neighbor distance constraints represent the angular bond-bending constraints due to the atom's chemistry. Dihedral rotation about the central-force bonds is the elementary flexible element in this type of network. Rotation through a dihedral angle is possible a priori but may be locked because of crosslinking bonds in the network. Furthermore, dihedral angles associated with double or partial-double bonds are represented as fixed by using a third-nearest-neighbor distance constraint. For instance, to lock

the peptide bonds within a protein structure, the distance between the carbonyl oxygen and amide hydrogen is fixed. Therefore, along the backbone of a protein the Φ, Ψ dihedral angles are a priori allowed to rotate, but the peptide bonds and other double-bonded groups are kept planar in this way.

As shown in Figure 1, the hydrogen bond is modeled similarly to a covalent bond, in which the donor, hydrogen, and acceptor atoms are typically generic and hence noncolinear. 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. This model for a hydrogen bond allows the protein structure to be described as a bond-bending network and is physically reasonable because hydrogen bonds are almost never precisely linear; the three dihedral angle degrees of freedom associated with this representation of the hydrogen bond also allow it to have some flexibility. Modeling the hydrogen bond to be more or less constrained than this has been tested, and the model in Figure 1 provides a good balance between neither over- nor underrepresenting the flexibility of a hydrogen bond. This model results in ideal rigid α-helices, and β-sheets ranging from rigid to somewhat flexible, depending on size and the regularity of their hydrogen-bonding patterns. Moreover, protein structures typically show a substantial proportion of rigid regions, while having regions that remain flexible. Determination of constrained and rotatable dihedrals, as implemented in the FIRST software's implementation of the pebble game, has been tested against exact counting and shown to agree for all the elementary structures: α-helices, parallel and antiparallel β-sheets, and reverse turns.

APPLICATION TO PROTEINS

Detailed information about the mechanical stability of a protein under a fixed set of distance constraints is provided by FIRST analysis. All overconstrained regions, rigid clusters, and underconstrained regions are determined and can be colored and viewed using standard molecular rendering packages, including Rasmol and InsightII. To analyze a protein, it must first be decided which hydrogen bonds to include and model as distance constraints.

Hydrogen Bonding

Beyond covalent bonds, salt bridges and then hydrogen bonds form the next strongest interactions within proteins (Fig. 3). Hydrogen bonds vary in strength from nearly as strong as the covalent bonds to as weak as the van der Waals interactions.[36,37] Hydrogen bonds form directional crosslinks in the bond-bending network that lead to large-scale rigid regions. In proteins, the regular hydrogen-bonding patterns between main-chain amide and carbonyl groups form the regular secondary structures: α-helices, β-sheets, and reverse turns. Hydrogen bonds also stabilize the tertiary structure of proteins through side-chain interactions that interlock parts of the protein chain distant in sequence.

For protein structures in which the hydrogen atom positions are not experimentally defined (the case for most

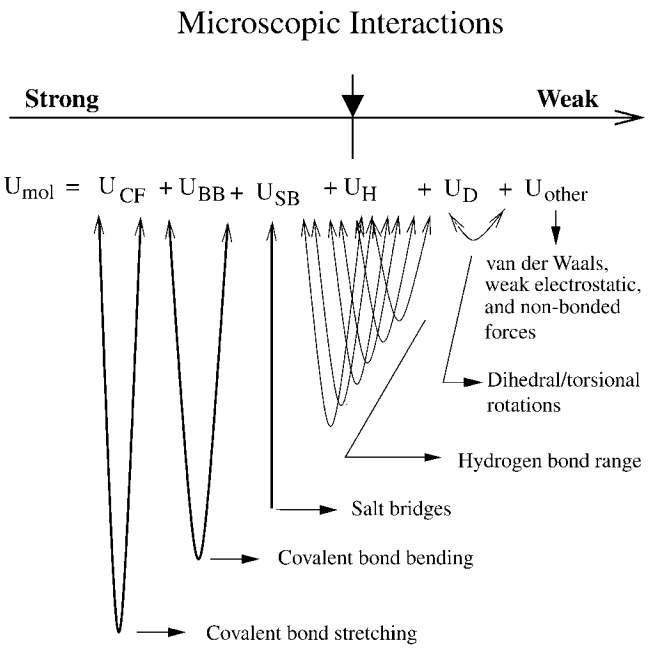

Fig. 3. Schematic representation of the ordering of microscopic forces, from strongest to weakest. Distance constraints are used in FIRST to model strong bonding forces to the left of a sliding pointer. This approach defines a network of covalent and hydrogen bonds and salt bridges in the protein.

crystallographic structures determined by X-ray rather than neutron diffraction), the WhatIf software package is used to assign polar hydrogen atoms positioned such that their hydrogen-bonding opportunities are optimized.[38] Because FIRST results will depend on the accuracy of placement of polar hydrogen atoms (because of their influence on hydrogen bonds being assigned, or not), we have tested how well WhatIf-defined hydrogen positions compare with the experimentally determined positions in five neutron diffraction structures from the Protein Data Bank (PDB; www.rcsb.org/pdb).[1] Five neutron structures available in the PDB are lysozyme (PDB entry 1lzn), trypsin (1ntp), insulin (3ins), myoglobin (2mb5), and ribonuclease A (5rsa). For each structure, we processed the file through FIRST using two different hydrogen-bond energy threshold or cutoff values, as discussed below. We then created a modified version of the neutron structure by stripping the hydrogen atoms from it and adding new hydrogens with WhatIf. This new structure was then processed through FIRST for comparison with the results obtained using neutron diffraction-defined hydrogen atom positions.

Table I contains results for these five neutron structures at two different energy cutoff values: −0.1 kcal/mol and −0.6 kcal/mol (the latter corresponding to thermal fluctuation at room temperature). The comparative results show only slight differences due to a few hydrogens placed differently in the two structures. The percentages shown are calculated by dividing twice the number of hydrogen bonds in common by the total number of hydrogen bonds for both versions of the protein structure. While the energy threshold of −0.6 kcal/mol is more restrictive and includes

TABLE I. Effect of Hydrogen Atom Placement on the Number of Hydrogen Bonds Identified*

H-Bond Energy		PDB Code 1lzn	1ntp	2mb5	3ins	5rsa	Total
	No. of residues	129	223	153	102	124	—
	No. of protein atoms	1762	1790	1836	1305	1556	—
	Resolution (Å)	1.70	1.80	1.80	1.50	2.00	—
No. of H-bonds with $E \leq -0.1$ kcal/mol	No. common to both	184	216	249	108	140	897
	No. unique to neutron[a]	3	9	13	3	4	76
	No. unique to modified[b]	11	17	6	6	4	
	% in common	96.3	94.3	96.3	96.0	97.2	95.9
No. of H-bonds with $E \leq -0.6$ kcal/mol	Common to both	130	168	182	80	116	676
	Unique to neutron[a]	7	16	22	3	3	84
	Unique to modified[b]	6	9	8	6	4	
	% in common	95.2	93.1	92.4	94.7	97.1	94.2

*Experimentally defined hydrogen positions and resulting hydrogen bonds of five neutron structures are compared with the hydrogen bonds resulting from assignment of hydrogen positions by WhatIf in the same five structures.
[a]Rows labeled unique to neutron include the number of hydrogen bonds from experimentally determined hydrogen atom positions.
[b]Rows labeled unique to modified include the number of hydrogen bonds from WhatIf calculated hydrogen atom positions in neutron diffraction structures from which the experimentally determined hydrogen atom positions had been removed.

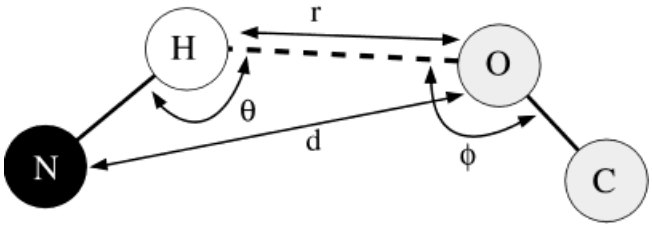

Fig. 4. Geometry used in the hydrogen bond energy potential. θ is the donor–hydrogen–acceptor angle; ϕ is the hydrogen–acceptor–base angle, where base is the atom (C, in this case) covalently bonded to the acceptor; d is the donor–acceptor distance; r is the hydrogen–acceptor distance; and φ (not shown) is the angle between the normals of the planes defined by the covalent bonds of the donor and base atoms (e.g., the planes defined by the two sp^2 centers, N and C, in this case).

only the strongest hydrogen bonds, there was slightly less overall agreement (94%) in the hydrogen bonds at this energy threshold in the WhatIf and neutron versions of the structure than was found at the chosen threshold of −0.1 kcal/mol (96%). Thus, on average, only 4% of the hydrogen bonds were assigned differently in the two types of structures. Not surprisingly, many of the hydrogen-bond differences resulted from different placement of hydrogens on histidine residues. Histidine has two side-chain nitrogen atoms that can bond to 0, 1, or 2 hydrogens, depending on the local environment. Overall, we conclude that the WhatIf software package positions hydrogen atoms sufficiently accurately to permit analysis of the resulting hydrogen-bond network.

To define the hydrogen bonds and salt bridges for inclusion in the analysis of rigid and flexible regions, the geometry and energy of these interactions is assessed. A superset of possible hydrogen bonds is assigned based on meeting the following geometric[39,40] criteria: the donor–acceptor distance, d ≤ 3.6 Å, the hydrogen-acceptor distance, r ≤ 2.6 Å, and the donor–hydrogen-acceptor angle, θ, falls between 90° and 180° (Fig. 4). Salt-bridge (ion-pair) interactions are considered as a special case of hydrogen bonds. Salt bridges are similar to hydrogen bonds, with a more significant ionic or Coulombic component, which is less geometrically sensitive. Our identification of such salt bridges follows previous studies[41–43] by extending the maximum distance between donor and acceptor to 4.6 Å and softening the angular dependence such that θ falls between 80° and 180°.

Because the strength of hydrogen bonds and salt bridges depends on the chemistry of the particular donor and acceptor atoms as well as their orientation, an energy function[44] is then used to rank hydrogen bonds:

$$E_{HB} = V_0\left\{5\left(\frac{d_0}{d}\right)^{12} - 6\left(\frac{d_0}{d}\right)^{10}\right\} F(\theta, \phi, \varphi) \qquad (1)$$

where

sp^3 donor–sp^3 acceptor	$F = \cos^2\theta\cos^2(\phi - 109.5)$
sp^3 donor–sp^2 acceptor	$F = \cos^2\theta\cos^2\phi$
sp^2 donor–sp^3 acceptor	$F = \cos^4\theta$
sp^2 donor–sp^2 acceptor	$F = \cos^2\theta\cos^2(\max[\phi,\varphi])$
V_0 = 8kcal/mol and d_0 = 2.8 Å	

The hydrogen bond energy (E_{HB}) is a function of the equilibrium hydrogen bond distance, d_0, and well depth, V_0. The donor to acceptor distance is d. The angular dependence of the function, F, is dependent on the hybridization of the donor and acceptor atoms. The four possible cases shown rely on the angular terms θ, ϕ, and φ. Angle θ defines the donor atom–hydrogen–acceptor atom angle, and ϕ is the angle between the hydrogen atom, the acceptor, and the base atom bonded to the acceptor. The φ angle is between the normals of the two planes defined by the sp^2 centers. If ϕ is less than 90°, the supplement of the angle is used. Figure 4 illustrates how these parameters are defined.

Salt bridges can be viewed as strong hydrogen bonds[36] with average energies of −6 (±4) kcal/mol.[45] Salt bridges have broader distance and angular distributions than are

found for nonionic hydrogen bonds, and these observed distributions are not well reflected by the hydrogen-bond energy functions we have tested. Salt bridges within the above specified geometric ranges generally have stronger interactions than those exhibited by hydrogen bonds. Therefore, salt bridges meeting the above geometric criteria are always included in the bond network analyzed for flexibility.

For hydrogen bonds, we can tune the energy threshold (the sliding pointer in Fig. 3) used to define which hydrogen bonds are included in the network. Setting the threshold at less negative (less favorable) energy values includes weaker hydrogen bonds, which tend to be common in proteins and have a significant influence on structural stabilization. The ability to select hydrogen bonds based on strength allows investigation of how the stability in each region of the protein varies as the hydrogen-bond network is strengthened or weakened. By changing the criteria for modeling a hydrogen bond as a constraint, a protein can be substructured from containing a few, large rigid clusters down to being completely floppy, with many small rigid clusters involving single atoms with their covalent bonds acting as rotatable dihedral angle hinges. Individual hydrogen bonds or small sets of hydrogen bonds that form critical crosslinks can therefore be identified by shifting the energy threshold and observing which hydrogen bonds, when included or omitted, have a large effect on the rigidity of the network.

Ideally, we would like to be able to set this energy cutoff value, perform a single FIRST analysis, and know that the output describes the physically relevant flexibility of a protein. One way of setting the energy threshold objectively is to choose it such that maximum agreement in the hydrogen-bond network is obtained for pairs of independently determined structures for a protein (e.g., by different researchers or in different crystallographic packings) in which the main-chain conformations are the same. This ensures that the results of FIRST analysis are not sensitive to the sorts of fluctuations known to occur within protein structures. Since the energy cutoff is the tunable parameter in using FIRST, we tested which setting gave the most similar results between pairs of such structures. Such a cutoff value should naturally be below when all hydrogen bonds are present ($E_{\mathrm{cut}} \leq 0.0$ kcal/mol) and above when the protein substructures into many tiny rigid clusters (E_{splinter}). A very natural place to look at the behavior of these proteins is near room temperature which corresponds to $E_{\mathrm{rt}} = -0.6$ kcal/mol. However, because the energy function is approximate (does not take into account the effect of more distant neighboring atoms on hydrogen-bond strength), it is important not to take these energies too literally, and to consider them as relative rather than absolute.

To determine a reasonable default energy threshold for hydrogen bonds, we evaluated which threshold best conserves the hydrogen bonds within a family of protein structures. Multiple structures within four different protein families were studied to find such a threshold. The PDB codes used for each family are as follows: trypsin (1tpo, 2ptn, 3ptn), trypsin inhibitor (4pti, 5pti, 6pti, 9pti), adenylate kinase (1zin, 1zio, 1zip), and human immunodeficiency virus protease (HIVP) (1dif, 1hhp, 1htg). Figure 5 shows the hydrogen-bond energy distribution for one of these families, namely the three HIVP structures. A large spike in the distribution of possible bonds located between -0.1 kcal/mol and 0.0 kcal/mol for the number of hydrogen bonds in the three structures appears in Figure 5. This spike is largely due to the fact that quite generous definitions of hydrogen bonds are allowed initially (donor–hydrogen–acceptor angle, $\theta \geq 90°$ and donor–acceptor distance, $d \leq 3.6$ Å, as shown in Fig. 4). The inset in Figure 5 expands the region near 0.0 kcal/mol, demonstrating how a large number of very weak hydrogen bonds, often with θ angles near 90°, can be removed by setting $E_{\mathrm{cut}} \leq -0.1$ kcal/mol. Thus, the generous hydrogen bond distance and angle screening criteria can be effectively filtered by setting E_{cut}. When these geometric criteria and an energy threshold of -0.1 kcal/mol are applied to analyze the hydrogen bonds and salt bridges in five neutron diffraction structures, a Gaussian distribution is observed for the number of hydrogen bonds as a function of donor–acceptor distance, with virtually all hydrogen bonds and salt bridges having distances between 2.6 and 3.6 Å. The distribution in donor–hydrogen–acceptor angles is bimodal, with a strong, Gaussian peak between 130° and 180° and a weaker peak between 90° and 130°.

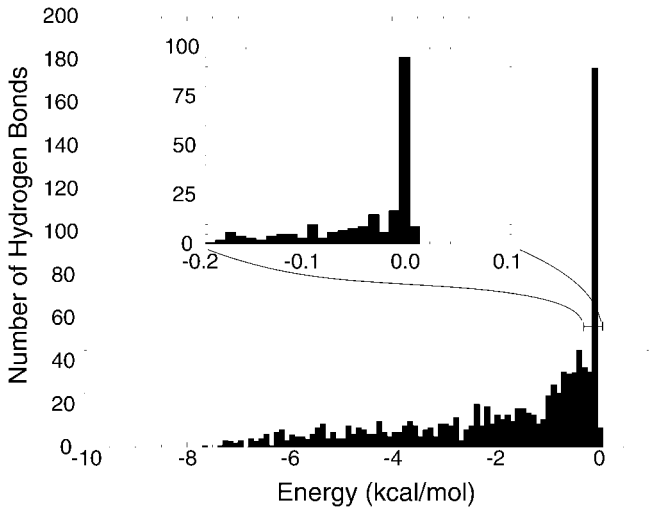

Fig. 5. Histogram of energies for hydrogen bonds. Distribution of hydrogen bond energy for three structures of human immunodeficiency virus protease (HIVP) (PDB codes 1dif, 1hhp, 1htg). Hydrogen positions were established by WhatIf. The inset expands the low-energy (weak hydrogen bond) region between -0.2 and 0 kcal/mol. An energy threshold of -0.1 kcal/mol is used to eliminate the large number of very weak hydrogen bonds in the spike near 0 kcal/mol.

In the choice of protein structures to analyze, the stereochemical quality of the structure can have a significant influence on the definition of its network of hydrogen bonds, due to their angular dependence. The result is that FIRST analysis on a structure with poor stereochemistry is likely to indicate the protein as being more flexible than it actually is, due to missing hydrogen bonds. It is advisable to assess the main-chain stereochemistry through a

Φ, Ψ plot, as well as focus on high-resolution, well-refined structures for FIRST analysis, to avoid this possibility of missing hydrogen bonds due to the misorientation of main-chain hydrogen-bonding groups.

Flexibility Index

A flexible region consisting of many interconnected rigid clusters within a protein may define a collective motion having only a few independent degrees of freedom. Although underconstrained, this region could be nearly rigid and thus mechanically stable. An isostatically rigid region, however, which contains no redundant constraints and is only rigid, is not expected to be as stable as an overconstrained region. Overconstrained regions have more constraints than necessary to be rigid, and therefore are considered more stable. Because of this continuum between rigidity and flexibility, a continuous index is useful.

The total number of floppy modes in a protein, denoted by F, corresponds to the number of independent, internal degrees of freedom. To obtain F, the six trivial rigid body degrees of freedom are subtracted from the total number of independent degrees of freedom. The global count of the number of floppy modes gives a good sense of overall intrinsic flexibility. However, a more useful measure is to track how the degrees of freedom are spatially distributed throughout the protein. In particular, we are interested in locating the underconstrained or flexible regions.

The quantity, f_i, is defined as a flexibility index that characterizes the degree of flexibility of the ith central-force bond in the protein. Let H_k and F_k, respectively, denote the number of hinge joints (rotatable bonds) and the number of floppy modes within the kth underconstrained region. Let C_j and R_j, denote the number of central-force bonds and the number of redundant constraints within the jth overconstrained region, respectively. The flexibility index provides a quantitative range from most to least constrained and is given by

$$f_i \equiv \begin{cases} \dfrac{F_k}{H_k} & \text{in an underconstrained region} \\ 0 & \text{in an isostatically rigid region} \\ \dfrac{-R_j}{C_j} & \text{in an overconstrained region} \end{cases} \qquad (2)$$

When the ith central-force bond is a hinge joint, the flexibility index is defined to be given by the the number of floppy modes divided by the total number of hinges within the underconstrained region. Because the number of independent dihedral rotations must be less than or equal to the number of hinge joints, the flexibility index is always ≤ 1. When the ith central-force bond is not a hinge joint, it is part of a rigid cluster. If the central-force bond is within an overconstrained region, the flexibility index is assigned a negative value with magnitude given by the number of redundant constraints divided by the total number of central-force bonds within the region. This number becomes more negative as the region becomes more overconstrained.

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

$$f_i = \frac{n-6}{n} \quad \text{for each central-force bond in a } n\text{-fold ring} \qquad (3)$$

Note that as the ring becomes larger, the flexibility index goes to the limit of $+1$; in this case, each dihedral angle is nearly independent, and the ring is almost as flexible as a linear chain. For a six-fold ring, the flexibility index is zero, indicating an isostatically rigid structure. For a three-fold ring, the ring is overconstrained, and the flexibility index is -0.2. The flexibility index of a protein can be plotted as a function of residue number, and regions within the plot (corresponding to segments within the sequence) can be colored according to whether they are coupled in motion (Fig. 6). A nice property of the flexibility index is that it varies gradually as hydrogen bond constraints are added or removed.

RESULTS AND DISCUSSION

An important feature of FIRST is that it can predict the intrinsic flexibility of a protein given a single 3-D structure. However, it is typical that upon ligand binding, the hydrogen-bond pattern will change. Therefore, the predicted conformational flexibility using FIRST will depend on whether the structure being analyzed is an open (ligand-free) form, or a closed (ligand-bound) form. Crystal contacts can also influence the flexibility of a protein, and their influence can be assessed in two ways by FIRST: (1) by analyzing the flexibility of the protein independent of its crystal lattice neighbors (in which case the effects of intermolecular hydrogen bonds are removed from analysis); and (2) by comparing the flexible regions found for the same protein crystallized in different lattice packings. However, the general features of flexible and rigid regions found by FIRST are remarkably consistent among different 3-D structures (in the same ligand-binding state) for a protein, as will be shown for HIVP, dihydrofolate reductase, and adenylate kinase.

HIV Protease

An initial application is to HIVP, a major inhibitory drug target for current acquired immunodeficiency syndrome

Fig. 6. **A:** Rigid cluster decomposition of the open conformation of human immunodeficiency virus protease (HIVP) (PDB code 1hhp). **B:** Flexibility index, color-mapped onto the same HIVP structure. Four regions of interest, α, β, γ, and δ, are identified for one of the dimers. **C:** Flexibility index plotted versus residue number. Of the four regions, α, β, and γ are most flexible (shown red in **B**), while δ is rigid (blue) in this open conformation of HIVP. Parts of the sequence that are coupled in motion are plotted in the same color; the same regions in **D** and **E** are then colored accordingly. **D:** Mobility plotted versus residue for this conformation. Mobility is determined as the average crystallographic temperature factor (B-value, or Debye–Waller factor) divided by the average atomic occupancy, for the main-chain atoms in each residue. **E:** Dihedral angle changes between the main chains of the above open conformation and the closed conformation (Fig. 7, PDB code 1htg). The three flexible regions identified by FIRST are also those with the greatest experimentally defined mobility values and dihedral angle changes.

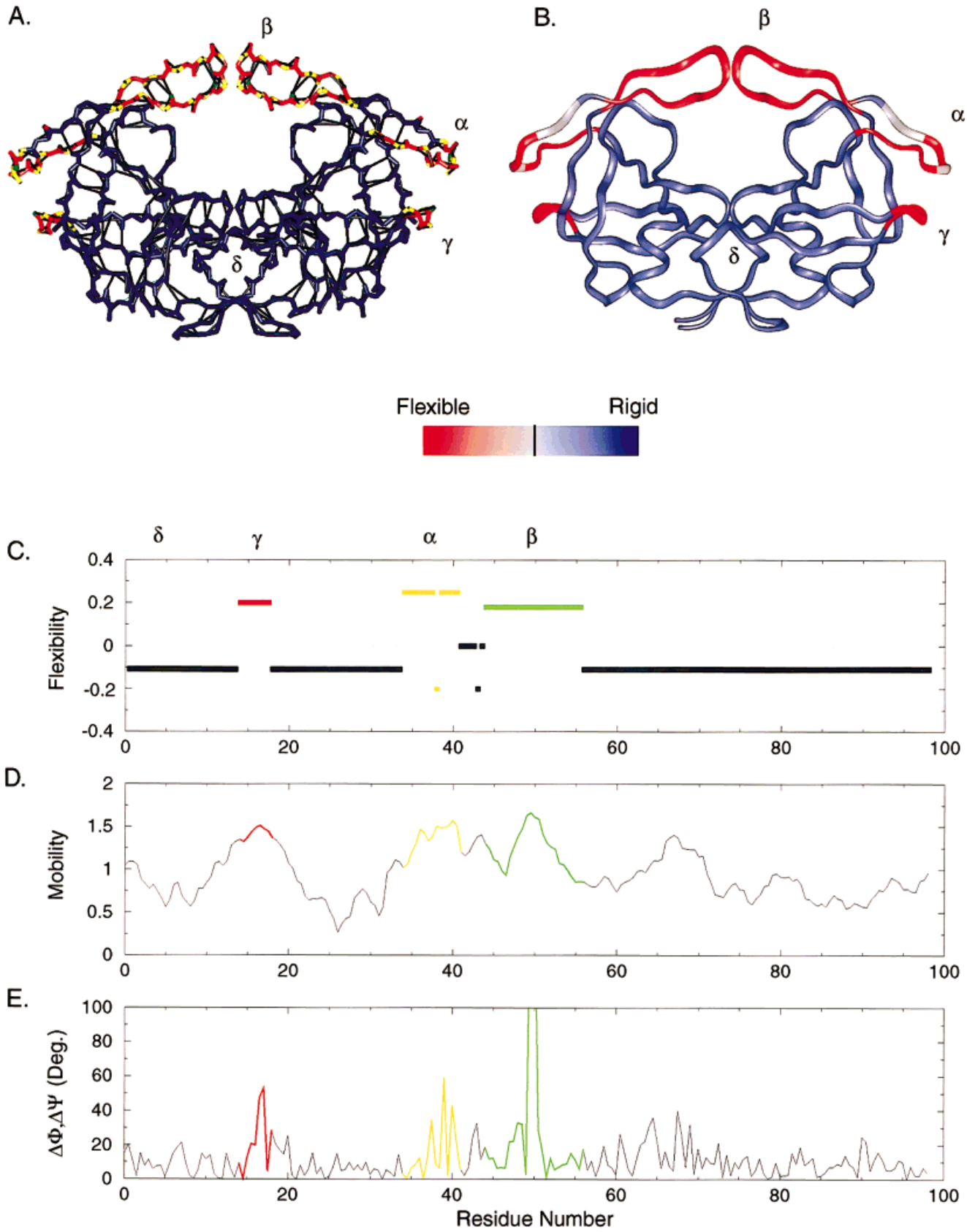
A.
β
α
γ
δ
B.
β
α
γ
δ
Flexible
Rigid
C.
δ
γ
α
β
Flexibility
0.4
0.2
0
-0.2
-0.4
0
20
40
60
80
100
D.
Mobility
2
1.5
1
0.5
0
0
20
40
60
80
100
E.
ΔΦ,ΔΨ (Deg.)
100
80
60
40
20
0
0
20
40
60
80
100
Residue Number

Figure 6.

(AIDS) therapy. Two ligand-free X-ray crystal structures available for HIV protease, PDB entries 1hhp and 3phv, are superficially very similar in structure and have similar resolution and crystallographic residual error (2.7-Å resolution for both, and an *R*-factor of 0.190 for 1hhp and 0.191 for 3phv). However, PROCHECK[46] showed that 3phv had significantly fewer residues with stereochemically favored Φ, Ψ values, which results in distorted main-chain hydrogen-bond geometries; therefore we chose 1hhp to represent the open HIVP conformation. The open form of the protein (PDB code 1hhp) is dominated by a single rigid cluster shown in blue (Fig. 6A), including the base and walls of the substrate and inhibitor binding site (cavity at center), and three flexible regions shown as alternating-colored bonds (each color indicating a rigid microcluster within the flexible region). The ends of the flaps (β region labeled in Figure 6A, residues 45–56) are known from crystallographic and NMR structures to be important for closing over and binding inhibitors[47] and appear as the most flexible (red) regions when the structure is characterized by the flexibility index (Fig. 6B,C). Other flexible regions include the base of each flap (region α, residues 39–42), which may act as a cantilever, and the γ region.

The FIRST results for HIVP have been compared with experimental measures of protein flexibility. The major peaks in main-chain thermal mobility (B-value), measured crystallographically and shown in Figure 6D, correlate directly with the α, β, and γ flexible regions predicted by FIRST. The region labeled δ is the dimer interface, formed by the N- and C-termini of the two, identical protein chains. It should be noted that for proteins with mobile domains or other moving rigid bodies, such as α-helices, the crystallographic mobility and FIRST results will not necessarily compare well with B-values. Crystallographically, they appear as mobile regions, whereas in FIRST they appear as rigid regions flanked by flexible loops (allowing the motion). This confusion can be avoided when NMR order parameters are available for comparison, since they also indicate moving rigid bodies as rigid regions flanked by flexible loops. The Indiana Dynamical Database[48] contains such data for a number of proteins, including HIVP. These data are provided for PDB entry 1bvg, a ligand-bound form; as in the FIRST results for a different ligand-bound structure described in Figure 7, the base of the flaps are the most flexible region.

HIVP has been crystallized with various inhibitors bound, resulting in a closed conformation with the flaps lowered. The main-chain dihedral angle (Φ, Ψ) changes (similar to the analysis reported by Korn and Rose[49]) observed for crystal structures of the open (entry 1hhp) and closed (entry 1htp) conformations are shown in Figure 6E. The FIRST-predicted flexible regions also directly correspond with the regions of greatest dihedral angle change. In the three flexible regions (α, β, and γ), the flexibility is associated with a flip in at least one dihedral angle (defined as a change of more than 60°) within a rigid β-turn in the center of each flexible region (Fig. 6A,E). The results here are consistent with the motion observed by interpolation between different HIVP crystal structures[50] and an earlier dihedral analysis for a different pair of HIVP structures[49] indicating that large changes at residues 40, 50, and 51 in the α and β regions result in a large, concerted movement of the flaps. Flexibility of the γ region has not been emphasized in other studies of HIVP; however, it is known that drug-resistant mutants of the protease include two residues that pack against the γ region, 63 and 71, with residue 63 proposed to induce a conformational perturbation.[51,52] Thus, conformational coupling between the γ region and the flaps, through the γ–α loop interactions, may explain why mutations in the γ region, which are distal from the active site, cause resistance to drug binding.

Ligand binding restricts the motion of the flaps through new hydrogen bonds linking the two flaps to each other and to the ligand. Some of these hydrogen bonds between the flaps and ligand are mediated by an conserved water molecule found in retroviral but not mammalian homologs of HIVP,[53] providing a useful basis for designing more HIV-specific drugs. To compare the influence of ligands on HIVP flexibility, there were a number of ligand-bound structures of good stereochemistry from which to choose. For brevity, here we show the results from PDB entry 1htg, with GR137615 bound, to represent the closed form of HIVP. (We have also analyzed two other ligand-bound structures, 1hiv and 1dif, and found the influence of these ligands on protein flexibility to be substantially similar.) Unlike the open form, the closed structures were resolved crystallographically as dimers; thus, independent structural information is available for the two subunits of the dimer. This means it is possible to assess the influence of different side-chain conformations in the two halves (due to ligand interactions, thermal fluctuations and environmental differences) in terms of their effects on the hydrogen-bonding network and flexibility. The left and right sides of HIVP in Figure 7 indicate that the only substantial difference in their flexibility is caused by the asymmetry of the ligand bound (at center).

Comparison of this ligand-bound structure with the open HIVP also demonstrates how a ligand can rigidify part of the protein through new hydrogen bonds even though the ligand itself is not rigid (note black bonds indicating H-bonds between the protease flaps in Figure 7A, and that the flaps are now rigidified), while making other parts of the protein more flexible. In particular, note the dimer interface, where inter-subunit rotation occurs upon ligand binding, breaking some of the interfacial stabilizing hydrogen bonds, and the loop to the right of the binding cavity, shown as a flexible (orange) region of the main-chain ribbon in Figure 7B. This loop flexibility is not reflected in the other HIVP subunit, due to ligand asymmetry. Flexibility of the dimer interface in a ligand-bound structure is also a prominent feature found by NMR[54] and MD analyses[55]; MD also identifies flap flexibility in the ligand-free conformation.

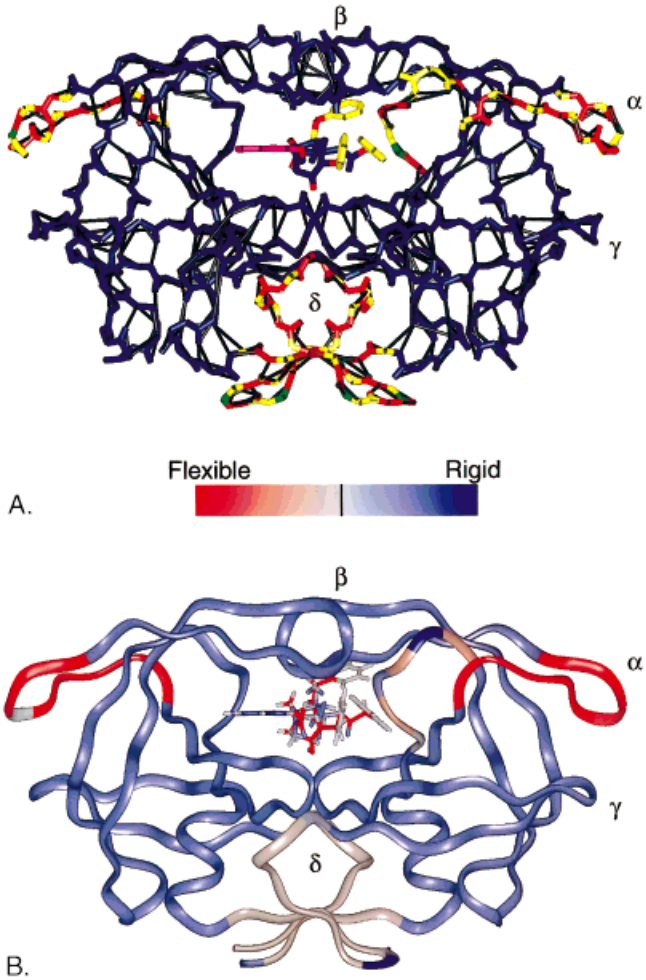

Fig. 7. **A:** Rigid cluster decomposition of the closed conformation of human immunodeficiency virus protease (HIVP) (PDB code 1htg). **B:** Flexibility index plot of the same, closed conformation of HIVP (PDB code 1htg). Four regions of interest, α, β, γ, and δ, are identified for one of the monomers. Contrast the change in rigidity for these regions between this ligand-bound structure and the ligand-free case (Fig. 6).

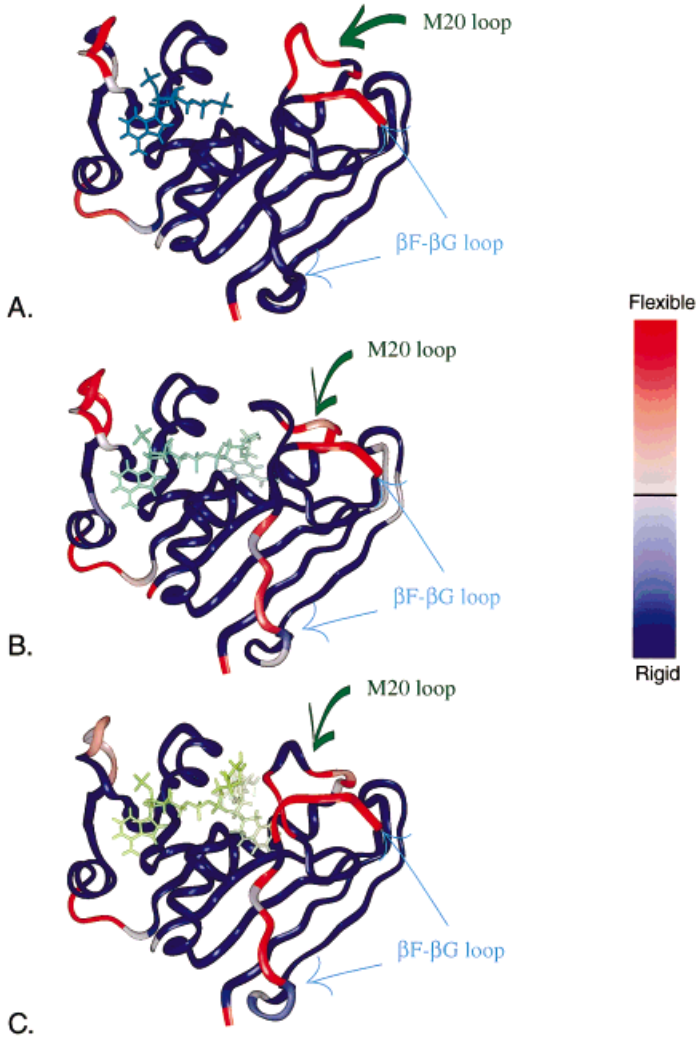

Fig. 8. Flexibility map of dihydrofolate reductase for the open conformation (PDB code 1ra1) (**A**) and for the closed conformation (PDB code 1rx1) (**B**). **C:** Occluded conformation (PDB code 1rx6). Shown in green are the ligands bound in these reaction pathway intermediates. Two loops experimentally determined to be flexible, M20 and βF–βG, are also noted. The motion of the M20 loop is essential to accommodate a variety of ligands during catalysis. The flexible βF–βG loop participates in ligand-induced conformational changes. At the top of the graph is the scale for the flexibility index used to map color onto the Cα trace. The scale runs from red (flexible) through gray (isostatic) to blue (rigid).

The influence of water is easily seen in a comparison of the liganded cases. A specific water molecule (WAT301) positioned between the β flaps and the ligand is conserved in most HIVP ligand-bound structures.[53] In this rigid region analysis, we found the inclusion of this water essential to rigidify the β flaps. This particular water molecule serves as a hydrogen acceptor from residue 50 of each protein chain and a hydrogen donor to the ligands. Adding one water molecule introduces four hydrogen bonds to the structure, with energies within the range of −1.5 to −7.5 kcal/mol. For both 1htg and 1dif, the β flaps become flexible without the intermolecular hydrogen bonds created by this water molecule. We have only included buried water molecules making intermolecular hydrogen bonds in HIVP, as these tend to be reliably assigned between structures, whereas surface water molecules are unevenly assigned in many of the crystal structures of HIVP, as well as being variable in crystallographic temperature factor.

Dihydrofolate Reductase

By trapping different ligand-bound states crystallographically, Sawaya and Kraut[56] noted five conformational states for *Escherichia coli* dihydrofolate reductase (DHFR) during its catalytic cycle. We analyzed three of these crystallographic structures (PDB codes: 1ra1, 1rx1, and 1rx6) with the FIRST algorithm. These three structures correspond to the open, occluded, and closed conformations of DHFR, as shown in Figure 8. The C_α traces are colored according to each residue's flexibility index, f_i, with the most flexible regions colored red and the most rigid colored blue.

According to the previous study,[56] the regions of greatest interest are the rotations of two major subdo-

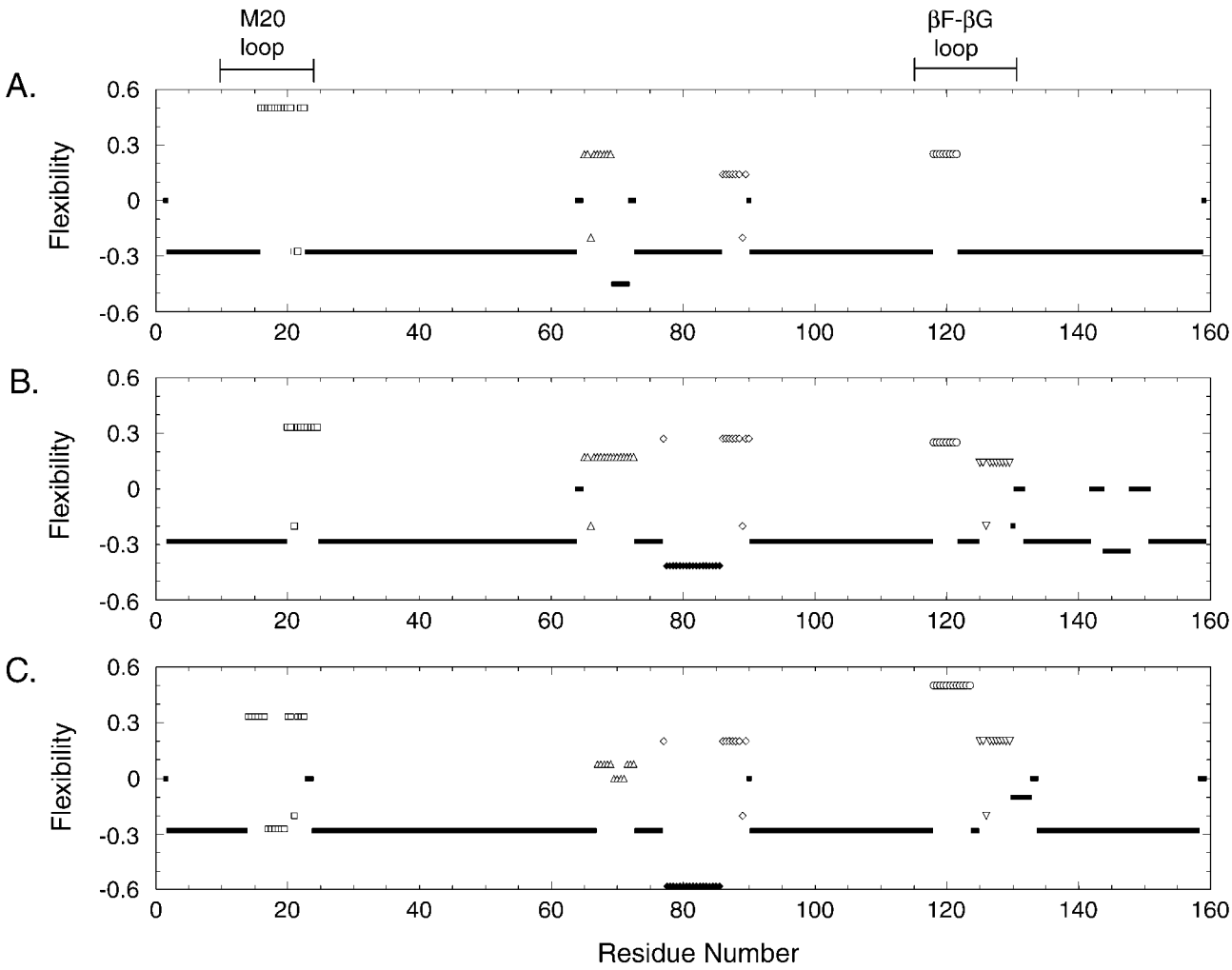

Fig. 9. Flexibility index plot for the three conformations of dihydrofolate reductase (DHFR) shown in Figure 8. The value of the flexibility index as defined in eq. (2), plotted versus residue number (a monochrome version of the type of plot shown for human immunodeficiency virus protease (HIVP) in Fig. 6C). The two experimentally determined loops, M20 and βF–βG, are shown at the top of the plot, and correspond to highly flexible regions. For each collective motion within the structure, a unique symbol is plotted.

mains: the adenoside binding subdomain and the loop subdomain. The M20 and βF–βG loops of the loop subdomain are denoted in Figures 8 and 9. Studies by Miller and Benkovic[57] concluded that the flexibility of these loops is interrelated, such that the flexibility of the outer βF–βG loops guides the conformation of the M20 loop. It is this correlated flexibility that gives DHFR ligand specificity. We would expect the regions that move the most, and that are important for binding, to appear flexible in the FIRST analysis, at least in the open conformation. Figure 8A shows that the M20 loop is detected by FIRST to be fully flexible, but in the closed form (Fig. 8B), this loop has moved and become partially locked into place. Comparing all three conformations quantitatively in Figure 9, the residues within these two mobile loops tend to be most flexible, and this flexibility is fairly independent of conformation being analyzed. This relates to a functional requirement for these loops to remain flexible during the catalytic cycle. The flexible region around residue 88 in all three conformations of Figure 9 corresponds to a hinge between the two subdomains. Similarly, the flexible region found by FIRST around residue 70 (open triangles in the flexibility index plot, Fig. 9) is within the adenoside binding subdomain and has also been identified as flexible by NMR techniques.[58] Plots of crystallographic mobility and regions of greatest main-chain dihedral angle change (data not shown) for the three DHFR structures in Figure 8 are remarkably consistent between the structures and also agree with FIRST results that the most flexible regions are in the M20 loop and the βF–βG loop. We have analyzed several other DHFR structures by FIRST (PDB entries 1rx2, 3, 4, and 5) and have found substantial agreement with the above conclusions. In general, changes in ligands between these structures can influence the flexibility of their neighborhood in the protein, but the major features of flexibility remain consistent.

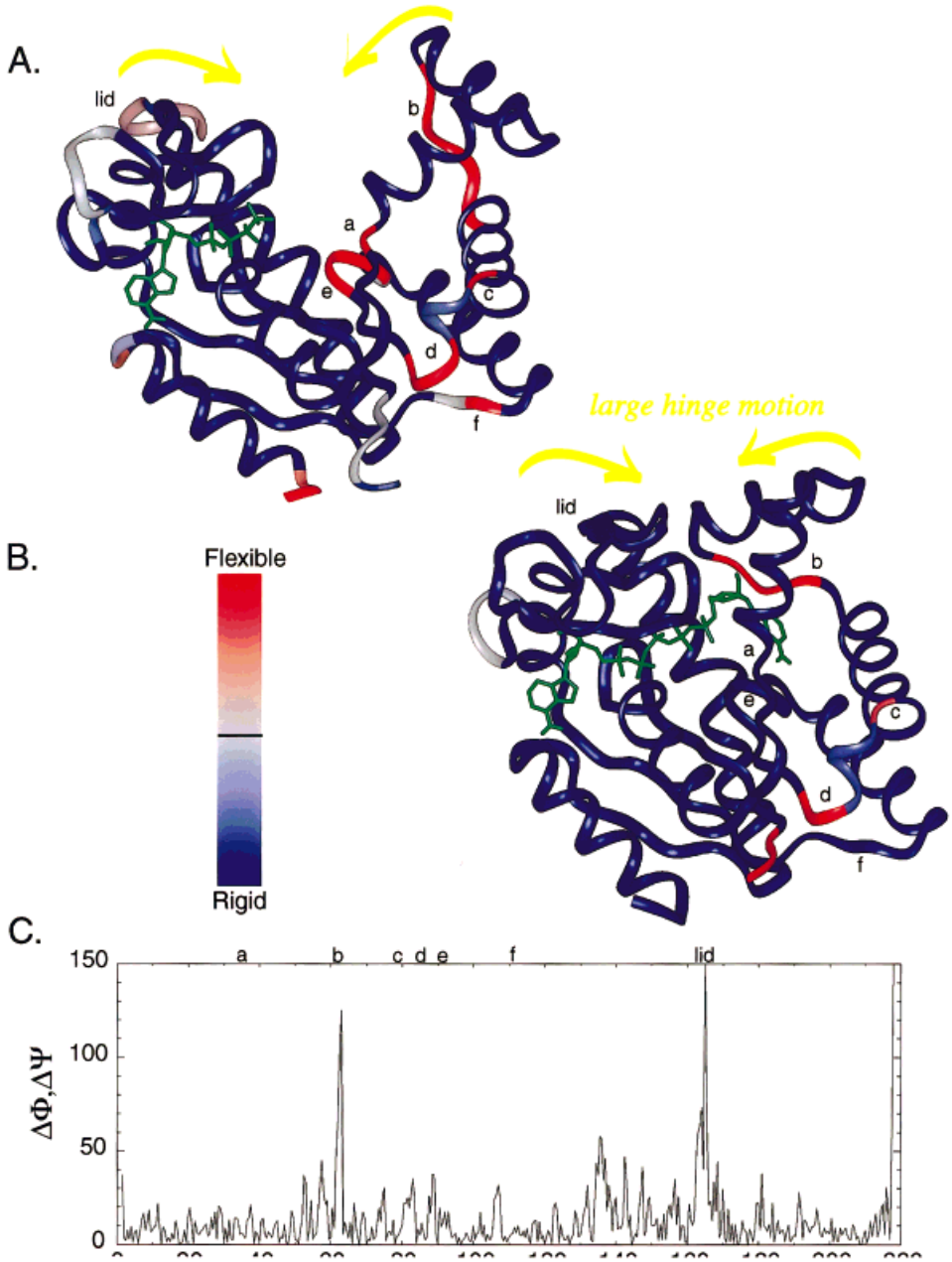

Fig. 10. Flexibility plot of adenylate kinase for the open conformation (PDB code 1dvr) (**A**) for the closed conformation (PDB code 1aky) (**B**). **C:** Difference in main-chain dihedral angles between these two conformations, indicating the locations of large, localized conformational changes. Ligands bound to these structures are shown in green tubes. The open state (**A**) has only adenosine triphosphate (ATP) bound. In the closed state (**B**), the ligand, P^1,P^5-bis(adenosine-5′-)pentaphosphate (AP_5A) mimics the roles of AMP and ATP binding concurrently. Dark blue corresponds to highly overconstrained and rigid, with a flexibility index; bright red corresponds to highly flexible. All labeled regions are discussed in the text.

Adenylate Kinase

Another protein whose motion has been studied experimentally is adenylate kinase.[59–61] Previous work indicates that adenylate kinase uses hinges rather than shear motions for conformational change. This intrinsic, large-scale hinge motion upon ligand binding is easily identifiable in both the open and closed conformations analyzed by FIRST, as indicated in Figure 10 by gold arrows. Several helices at the extreme right of Figure 10 move in like fingers via hinges defined as the most flexible (red) portions of the protein by FIRST.

Adenylate kinase binds two ligands, ATP and AMP, in a two-step mechanism. Unfortunately, structures of adenylate kinase from the same species are not available for all three steps (ligand-free, followed by two binding/conformational change events). By comparing enzymes from differ-

ent species, four hinges were previously defined to contribute to the conformational change between open and ligand-bound forms.[59] These hinges move in a concerted way to account for the large conformational change closing the lid domain (residues 131–165) over the ligand. Figure 10A,B shows structures with ligands bound to this domain and not this initial ligand-binding step.[59] However, the peak in main-chain dihedral change shown around residue 167 in Figure 10C corresponds to a conformational switch made by the red flexible loop (part of the lid) in Figure 10A between structures. Crystallographic mobility analysis for the open structure (10A) and the closed structure (10B) are generally in good agreement with FIRST results, the only difference being that regions between residues 135 and 160 appear crystallographically mobile in the closed structure.

Closing of the lid domain is associated with the binding of ATP (green tubes at center in Fig. 10A). Binding of the ligand AP_5A (green tubes, Fig. 10B) produces the fully closed conformation of adenylate kinase and locks many of the domain linking hinges (a–f), as seen by the transition between flexible (red) and rigid (blue) regions in Figure 10A,B. The NMP_{bind} site is where the part of AP_5A that is nonoverlapping with ATP binds, and this site is formed by the interface between the two domains as they clamp down on the inhibitor.[61]

Comparing these structures, FIRST shows that the flexibility of the NMP_{bind} domain (especially hinges a, e, and f) decreases upon AP_5A binding to this domain. The flexible linkage (b) between helices $\alpha3$ and $\alpha4$ around residue 62 (identified in the Fig. 10C plot of change in Φ, Ψ angles between the open and closed structures in Fig. 10A,B) seems to account for a large part of the motion transforming the open to the closed conformation. This region (b in Fig. 10A,B) is found to decrease in size but remains flexible in the closed conformation. The persistent flexibility in the closed conformation hints at the reversibility of the motion required for catalytic turnover. Even in the ATP-bound closed lid conformation exhibited by both of these structures, certain key hinges[59,60] remain flexible. For example, hinges c and d between helices $\alpha4$ and $\alpha5$ remain flexible in both states. Thus, FIRST results correlate well with the crystallographically observed conformational changes upon ligand binding for the complex motions within adenylate kinase.

SUMMARY

A novel distance constraint approach is introduced for characterizing the intrinsic flexibility of a protein. The underlying physical and mathematical assumptions are outlined and implemented computationally in the FIRST software. FIRST determines the floppy inclusion and rigid substructure topography of a given protein structure, based on a set of distance constraints determined by the covalent and hydrogen bonding network within a single conformation of the protein. A flexibility index is introduced as a continuous measure for quantifying the flexibility or stability of each bond within the protein, based on the FIRST identification of the density of floppy modes as well as the density of redundant bonds within the protein.

There are several advantages of FIRST relative to previous methods for analyzing protein flexibility. FIRST calculations can be done virtually in real time (a few seconds of CPU time) once the bond network has been defined. Analysis of a single protein structure is able to indicate regions likely to undergo conformational change as part of the protein's function. For a given set of distance constraints, the rigid regions and the flexible joints between them are determined precisely. The ability to determine coupled motions very quickly among the dihedral angles of a flexible region gives FIRST an advantage over other methods. Collective motions, in which changing one dihedral angle will influence the other dihedral angles within the region, are localized within the flexible regions of the protein. Analysis of the relative flexibility within HIV protease, dihydrofolate reductase, and adenylate kinase, even when performed on a single structure, captures much of the functionally important conformational flexibility observed experimentally between different ligand-bound states. A distribution version of the FIRST software is in preparation and will be available through the authors.

ACKNOWLEDGMENTS

The authors thank Yuquing Xiao for help with programming and Vishal Thakkar and Brendan Hespenheide for their contribution to molecular graphics.

REFERENCES

1. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr., Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
2. Ringe D, Petsko GA. A consumer's guide to protein crystallography. In: Protein engineering and design. San Diego: Academic Press, 1996; p 210–229.
3. Bennett W, Huber R. Structural and functional aspects of domain motions in proteins. Crit Rev Biochem 1984;15:291–384.
4. Wuthrich K, Wagner G. Internal motion in globular proteins. Trends Biochem Sci 1978;3:227–230.
5. Nichols WL, Rose GD, Ten Eyck LF, Zimm BH. Rigid domains in proteins: an algorithmic approach to their identification. Proteins 1995;23:38–48.
6. Siddiqui AS, Barton GJ. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definition. Protein Sci 1995;4:872–884.
7. Boutonnet N, Rooman M, Wodak S. Automatic analysis of protein conformational changes by multiple linkage clustering. J Mol Biol 1995;253:633–647.
8. Holm L, Sander C. Parser for protein folding units. Proteins 1994;19:256–268.
9. Zehfus MH, Rose GD. Compact units in proteins. Biochemistry 1986;25:5759–5765.
10. Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. Naturwissenschaften 1985;72:212–213.
11. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
12. Ma J, Karplus M. The allosteric mechanism of the chaperonin GroEL: a dynamic analysis. Proc Natl Acad Sci USA 1998;95:8502–8507.
13. Case DA. Molecular dynamics and normal mode analysis of biomolecular rigidity. In: Thorpe M, Duxbury P, editors. Rigidity theory and applications. Kluwer Academic/Plenum, 1999.
14. Keskin O, Jernigan RL, Bahar I. Proteins with similar architecture exhibit similar large-scale dynamic behavior. Biophys J 2000;78:2093–2106.
15. Janin J, Wodak S. Structural domains in proteins and their role in

the dynamics of protein function. Prog Biophys Mol Biol 1983;42: 21–78.
16. Maiorov V, Abagyan R. A new method for modeling large-scale rearrangements of protein domains. Proteins 1997;27:410–424.
17. Jacobs DJ, Thorpe MF. Generic rigidity percolation: The pebble game. Phys Rev Letters 1995;75:4051–4054.
18. Jacobs DJ. Generic rigidity in three-dimensional bond-bending networks. J Phys A Math Gen 1998;31:6653–6668.
19. Thorpe MF, Jacobs D, Chubynsky N, Rader A. Generic rigidity of network glasses. In: Thorpe MF, Duxbury P, editors. Rigidity theory and applications. New York: Kluwer Academic/Plenum; 1999.
20. Jacobs D, Thorpe MF. Computer-implemented system for analyzing rigidity of substructures within a macromolecule. US Patent number 1998:6,014,449.
21. Dill KA. Dominant forces in protein folding. Biochemistry 1990;29: 7133–7155.
22. Abagyan R, Totrov M, Kuznetsov D. ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem 1994;15:488–506.
23. Lu H, Schulten K. Steered molecular dynamics simulation of conformational changes in immunoglobulin domain 127 interpret atomic force microscopy observations. Chem Phys 1999;247:141–153.
24. Jacobs D, Kuhn LA, Thorpe MF. Flexible and rigid regions in proteins. In: Thorpe MF, Duxbury P, editors. Rigidity theory and applications. New York: Kluwer Academic/Plenum; 1999.
25. Lagrange J. Mećanique analytique. Paris, 1788.
26. Maxwell JC. On the calculation of the equilibrium and stiffness of frames. Philos Mag 1864;27:294–299.
27. Laman G. On graphs and rigidity of plane skeletal structures. J Eng Math 1970;4:331–340.
28. Ashcroft N, Mermin N. Solid state physics. Fort Worth, TX: Saunders College Publishing; 1976.
29. Graver J, Servatius B, Servatius H. Combinatorial rigidity. Graduate studies in mathematics. Providence, RI: American Mathematical Society; 1993.
30. Guyon E, Roux S, Hansen A, Bibeau D, Troadec J, Crapo H. Non-local and non-linear problems in the mechanics of disordered systems: application to granular media and rigidity problems. Rep Prog Phys 1990;53:373–419.
31. Jacobs DJ, Hendrickson B. An algorithm for two-dimensional rigidity percolation: the pebble game. J Comput Phys 1997;137: 346–365.
32. Jacobs DJ, Thorpe MF. Generic rigidity percolation in two dimensions. Phys Rev E 1996;53:3683–3693.
33. Tay T-S, Whiteley W. Recent advances in generic rigidity of structures. Struct Topol 1985;9:31–38.
34. Whiteley W. Rigidity of molecular structures: generic and geometric analysis. In: Thorpe M, Duxbury P, editors. Rigidity theory and application. New York: Kluwer Academic/Plenum; 1999.
35. Xiao Y, Jacobs D, Thorpe M. Unpublished results; 1997.
36. Jeffrey GA. An introduction to hydrogen bonding. New York: Oxford University Press; 1997.
37. Fersht AR. The hydrogen bond in molecular recognition. Trends Biochem Sci 1987;12:301–304.
38. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph 1990;8:52–56. http://swift.embl-heidelberg.de/whatif/
39. Stickle D, Presta L, Dill K, Rose G. Hydrogen bonding in globular proteins. J Mol Biol 1992;226:1143–1159.
40. McDonald I, Thorton J. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1991;238:777–793.
41. Barlow DJ, Thorton JM. Ion-pairs in proteins. J Mol Biol 1983;168: 867–885.
42. Gandini D, Gogioso L, Bolognesi M, Bordo D. Patterns in ionizable side chain interactions in protein structures. Proteins 1996;24:439–449.
43. Xu D, Tsai C-J, Nussinov R. Hydrogen bonds and salt bridges across protein–protein interfaces. Protein Eng 1997;10:999–1012.
44. Dahiyat BI, Gordon DB, Mayo SL. Automated design of the surface positions of protein helices. Protein Sci 1997;6:1333–1337.
45. Kumar S, Nussinov R. Salt bridge stability in monomeric proteins. J Mol Biol 1999;293:1241–1255.
46. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291. ftp.biochem.ucl.ac.uk/pub/procheck/tar3_4/procheck.tar.Z
47. Nicholson LK, Yamazaki T, Torchia DA, Grzesiek S, Bax A, Stahl SJ, Kaufman JD, Wingfield PT, Yam PYS, Jadhav PK, Hodge CN, Domaille PJ, Chang C-H. Flexibility and function in HIV-1 protease. Nature Struct Biol 1995;2:274–280.
48. Goodman J, Pagel M, Stone M. Relationships between protein structure and dynamics from a database of NMR-derived backbone order parameters. J Mol Biol 2000;295:963–978. http://pooh.chem.indiana.edu/IDD.html
49. Korn AP, Rose DR. Torsion angle differences as a means of pinpointing local polypeptide chain trajectory changes for identical proteins in different conformational states. Protein Eng 1994; 7:961–967.
50. Gerstein M, Krebs W. A database of molecular motions. Nucleic Acids Res 1998;26:4280–4290.
51. Chen Z, Li Y, Schock HB, Hall D, Chen E, Kuo LC. Three dimensional structure of a mutant HIV-1 protease displaying cross-resistance to all protease inhibitors in clinical trials. J Biol Chem 1995;270:21433–21436.
52. Patrick A, Rose R, Greytok J, Bechtold C, Hermsmeier M, Chen P, Barrish J, Zahler R, Colonno R, Lin P. Characterization of a human immunodeficiency virus type 1 variant with reduced sensitivity to an aminodiol protease inhibitor. J Virol 1995;69: 2148–2152.
53. Wlodawer A, Erickson J. Structure-based inhibitors of HIV-1 protease. Annu Rev Biochem 1993;62:543–585.
54. Ishima R, Freedber D, Wang Y-X, Louis J, Torchia D. Flap opening and dimer–interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. Struct Fold Design 1999;7:1047–1055.
55. Scott W, Schiffer C. Curling of flap tips in HIV-1 protease as a mechanism for substrate entry and tolerance of drug resistance. Struct Fold Design 2000;9:1259–1265.
56. Sawaya M, Kraut J. Loop and subdomain movements in the mechanism of *Escherichia coli* dihydrofolate reductase: crystallographic evidence. Biochemistry 1997;36:586–603.
57. Miller G, Benkovic S. Stretching exercises—flexibility in dihydrofolate reductase catalysis. Chem Biol 1998;5:R105–R113.
58. Epstein D, Benkovic S, Wright P. Dynamics of the dihydrofolate reductase–folate complex: catalytic sites and regions known to undergo conformational change exhibit diverse dynamical features. Biochemistry 1995;34:11037–11048.
59. Gerstein M, Schulz G, Chothia C. Domain closure in adenylate kinase: joints on either side of two helices close like neighboring fingers. J Mol Biol 1993;229:494–501.
60. Schlauderer GJ, Proba K, Schulz GE. Structure of a mutant adenylate kinase ligated with an ATP-analogue showing domain closure over ATP. J Mol Biol 1996;256:223–227.
61. Zhang H-J, Sheng X-R, Pan X-M, Zhou J-M. Activation of adenylate kinase by denaturants is due to the increasing conformational flexibility at its active sites. Biochem Biophys Res Commun 1997;238:382–386.