ELSEVIER

Protein flexibility and dynamics using constraint theory

M.F. Thorpe,* Ming Lei,* A.J. Rader,* Donald J. Jacobs,† and Leslie A. Kuhn‡

**Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, USA*
†Department of Physics and Astronomy, California State University, Northridge, CA 91330, USA
‡Department of Biochemistry and Molecular Biology and Center for Biological Modeling, Michigan State University, East Lansing, MI 48824, USA

A new approach is presented for determining the rigid regions in proteins and the flexible joints between them. The short-range forces in proteins are modeled as constraints and we use a recently developed formalism from graph theory to analyze flexibility in the bond network. Forces included in the analysis are the covalent bond-stretching and bond-bending forces, salt bridges, and hydrogen bonds. We use a local function to associate an energy with individual hydrogen bonds, which then can be included or excluded depending on the bond strength. Colored maps of the rigid and flexible regions provide a direct visualization of where the motion of the protein can take place, consistent with these distance constraints. We also define a flexibility index that quantifies the local density of flexible or floppy modes, in terms of the dihedral angles that remain free to rotate in each flexible region. A negative flexibility index provides a measure of the density of redundant bonds in rigid regions.

A new application of this approach is to simulate the maximal range of possible motions of the flexible regions by introducing Monte Carlo changes in the free dihedral angles, subject to the distance constraints. This is done using a method that maintains closure of the rings formed by covalent and hydrogen bonds in the flexible parts of the protein, and van der Waals overlaps between atoms are avoided. We use the locus of the possible motions of HIV protease as an example; movies of its motion can be seen at http://www.pa.msu.edu/~lei.

INTRODUCTION

In this paper, we develop methods to probe the flexibility of proteins. Within a given force constant model, it is possible to simplify the forces to give a first approximation of the flexibility of the protein, subject to the strongest constraints within the structure. In this approach, the strong local forces [salt bridges, covalent and hydrogen bonds] are taken to be infinitely strong and treated as constraints. All other, weaker forces are not included. This simplification allows the use of modern constraint theory to find the rigid clusters within the protein and the flexible joints between them. This is reviewed in the first half of this study in the section entitled Statics. We then use these static results to explore possible motions of the protein, using an unbiased Monte Carlo sampling of the free dihedral angles in the structure. Such motion involves the flexible joints while maintaining the constraints. In the Discussion section that follows, we examine the possibility of going further and using more realistic potentials in the Monte Carlo simulations, and also using the static results to speed up molecular dynamics simulations by using two time scales.

Corresponding author: Dr. M.F. Thorpe, Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, USA.

E-mail address: thorpe@pa.msu.edu

STATICS

Modeling the Forces Within Proteins

Bonding forces within proteins impose distance constraints between atoms and reduce the total number of degrees of freedom available to the protein. These distance constraints can be viewed as a network of interactions in which flexibility and rigidity can be computed by graph theoretic approaches implemented in a computer program called *FIRST*,[1] which stands for *F*loppy *I*nclusion and *R*igid *S*ubstructure *T*opography. This approach, originated by Thorpe et al., has been applied previously to problems in material science, such as percolation through rigid networks and the prediction of phase transitions.[2]

To utilize *FIRST*, it is important to include those forces that carry the most biological significance. An objective way to select the correct forces is to consider the spectrum of forces, from strongest to weakest, as shown in Figure 1. The covalent bonding within the protein resulting from bond-bending, torsional, and bond-stretching (central) forces defines a natural set of distance constraints. A central force bond is a bond that has a force acting along the line connecting the two atoms and

Journal of Molecular Graphics and Modelling 19, 60–69, 2001

655 Avenue of the Americas, New York, NY 10010

1093-3263/01/$–see front matter
PII S1093-3263(00)00122-4

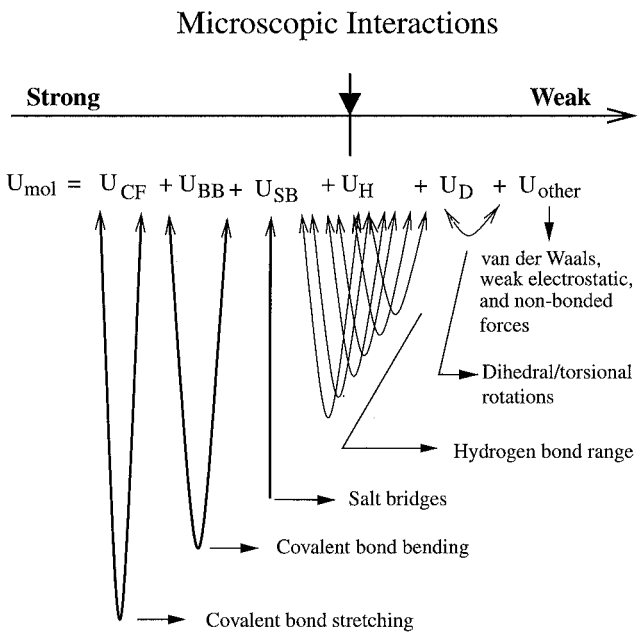

Figure 1. A schematic representation of the ordering of microscopic forces from the strongest to the weakest. Distance constraints are used in FIRST *to model strong bonding forces to the left of a sliding pointer. This approach defines a network of covalent bonds, salt bridges, and hydrogen bonds in the protein.* FIRST *analyzes the resulting constraint network to locate the rigid and flexible regions. Note that the peptide bond and other bonds with partial-double or double-bond character are always locked, and therefore to the left of the pointer, but are not shown explicitly in the diagram.*

bringing it back to an equilibrium length if it gets too long or too short. The torsional forces associated with peptide bonds and the other partial-double and double bonds in proteins effectively prevent dihedral rotation about that bond. In considering the effects of thermal energy on the kinetics of a protein structure, it is useful to note that bonds with energies of approximately 0.6 kcal/mol (the energy RT associated with room temperature, where R is the gas constant) are susceptible to breaking and reforming at room temperature.

Hydrogen bonds have high directional dependence and act over short ranges in contrast to the hydrophobic force. Therefore it is reasonable to expect that the buried hydrogen bonds will be substantially maintained as the protein undergoes conformational changes near its native structure. Modeling hydrogen bonds as distance constraints is a simple way to incorporate these directional and short-range properties of hydrogen bonding. Some recent molecular dynamics results by Lu and Schulten[3] suggest that the breaking of hydrogen bonds occurs as a well-defined event involving going over an energy barrier, as opposed to a continuous stretching until a feeble final breaking occurs.

It is common practice to represent the degrees of freedom accessible to a protein by fixing the covalent bond lengths and covalent bond angles, while allowing the dihedral angles to rotate. 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 seven relative to a Cartesian representation.[4] Even more degrees of freedom are eliminated when the bonds from salt bridges and hydrogen bonds are incorporated, resulting in larger rigid regions.

Beyond covalent bonds, salt bridges and hydrogen bonds form the next strongest interactions within proteins (see Figure 1). Periodic hydrogen-bonding patterns between main-chain amide and carbonyl groups form the secondary structures of α-helices, β-sheets, and reverse turns. Hydrogen bonds also stabilize the tertiary structure of proteins through side-chain interactions that interlock nonadjacent parts of the protein chain. Hydrogen bonds vary in strength from nearly as strong as covalent bonds to almost as weak as van der Waals interactions.[5,6]

Whether a hydrogen bond is included in the bond network used in the graph-theoretic analysis of *FIRST* depends upon its geometry and energy. For the majority of crystallographic structures of proteins determined by X-ray diffraction, the hydrogen atom positions are not defined; therefore, we have used the *WhatIf* software package to assign polar hydrogen atoms in positions optimal for hydrogen bonding.[7] Initially a superset of possible hydrogen bonds is assigned, based on the following geometric criteria[8]: the donor–acceptor distance, d, is less than 3.6Å; the hydrogen–acceptor distance, r, is less than 2.6Å; and the donor–hydrogen–acceptor angle θ is between 90° and 180°. These parameters are illustrated in Figure 2.

We use a local energy function that includes a central force part and an angular term $F(\theta, \phi, \gamma)$ depending on the chemistry of the donor and acceptor sites. This is similar to the potential used by Dahiyat et al.,[9] except that we have modified the angular part, $F(\theta, \phi, \gamma)$, to contain an additional exponential term so that the function becomes exponentially small for q angles less than ~120° rather than falling to zero at 90°. This modified angular function $F(\theta, \phi, \gamma)$ generally removes hydrogen bonds with $\theta \leq 120°$ when coupled with a cutoff energy of −0.01kcal/mol. This avoids the inclusion of main-chain hydrogen bonds between the i and $i+3$ residues in the middle of α-helices and other unphysical hydrogen bonds. In this model, helices are properly held together by hydrogen bonds between the residues i and $i+4$. Using this function, we apply an energy threshold of −0.01 kcal/mol (only including those hydrogen bonds with this or more favorable energy) to eliminate the large number of extremely weak hydrogen bonds. Whereas the ab-

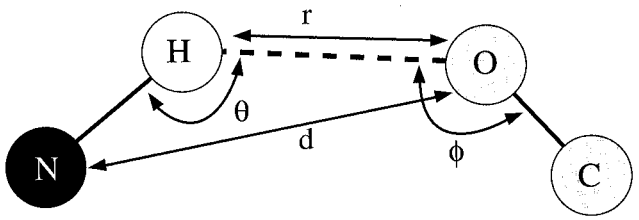

Figure 2. The geometry used in the hydrogen-bond energy potential, E_{HB}, *given in Equation 1. Here* θ *is the donor – hydrogen – acceptor angle,* ϕ *is the hydrogen – acceptor – base angle,* d *is the donor – acceptor distance,* r *is the hydrogen – acceptor distance, and* γ *(not shown) is the angle between the normals to the planes defined by the bonds from the donor and acceptor.*

solute energy values from this function may not be exact, it provides a useful way to rank the hydrogen bonds in a protein according to their relative energy. The hydrogen bond energy, E_{HB}, is a function of the equilibrium hydrogen bond distance, d_0, and well depth (optimal energy), V_0, as well as the angular term $F(\theta, \phi, \gamma)$:

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

where, for an

$$\text{sp}^3 \text{ donor} - \text{sp}^3 \text{ acceptor}, \quad F = \cos^2\theta \exp(-[\pi - \theta]^6) \cos^2(\phi - 109.5) \tag{1}$$

$$\text{sp}^3 \text{ donor} - \text{sp}^2 \text{ acceptor}, \quad F = \cos^2\theta \exp(-[\pi - \theta]^6) \cos^2\phi$$

$$\text{sp}^2 \text{ donor} - \text{sp}^3 \text{ acceptor}, \quad F = \{\cos^2\theta \exp(-[\pi - \theta]^6)\}^2$$

$$\text{sp}^2 \text{ donor} - \text{sp}^2 \text{ acceptor}, \quad F = \cos^2\theta \exp(-[\pi - \theta]^6) \cos^2(\max[\phi,\gamma])$$

and V_0 = 8 kcal/mol and d_0 = 2.8Å. Figure 2 illustrates how d_0, θ, and ϕ relate to the donor (N), hydrogen (H), acceptor (O), and base atom (C) geometry for the case of an amide-carbonyl main-chain hydrogen bond. In this case, γ is the out-of-plane angle of C—O relative to N—H.

Salt bridges can be viewed as strong hydrogen bonds[5] with average energies of −6 ±4 kcal/mol.[10] Salt bridges have broader distance and angular distributions than are found for nonionic hydrogen bonds, likely due to their Coulombic component. These observed distributions are not well reflected by hydrogen-bond energy functions as those found in (1). Salt bridges within these geometric ranges are generally stronger interactions than hydrogen bonds. Our identification of such salt bridges follows previous studies[11–13] by extending the maximum distance between donor and acceptor to 4.6Å and including all such salt bridges as constraints.

For hydrogen bonds, we can tune the energy threshold used to define which hydrogen bonds are included in the bond network (the sliding pointer in Figure 1). Moving the pointer to the right includes weaker hydrogen bonds, which are common in proteins and can contribute significantly to stability. The ability to select or exclude hydrogen bonds based on strength allows investigation of how the flexibility in each region of the protein varies as hydrogen bonds are added or subtracted from the network. Individual hydrogen bonds or small sets of hydrogen bonds that form critical crosslinks can also be identified in this way.[14]

Van der Waals and hydrophobic forces are examples of nonbonding forces found in proteins. Each individual 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. Hydrophobic interactions are likely dominant in driving a protein to fold,[15] and play an important role in stabilizing the native state. However, the hydrophobic and van der Waals forces are nonspecific (slippery) and therefore are not modeled by distance constraints between pairs of atoms. Our goal here is to test the extent to which salt bridges and covalent and hydrogen bonds can accurately define the degrees of freedom accessible to a protein.

Three-Dimensional Bond-Bending Networks

The covalent and hydrogen-bond connectivity of a protein can be completely described by the nearest neighbor central-force constraints, which include the associated bond-bending next nearest neighbor distance constraints. The latter define the angular geometry around an atom, e.g., that an sp^3-hybridized atom maintains tetrahedral coordination. The only elementary flexible unit that exists within a bond-bending network is a hinge joint, which corresponds to a rotatable dihedral angle. As with rigid clusters, flexible regions can be separated into underconstrained regions based on their collective motions. Hinge joints can only occur about axes defined by central-force distance constraints, never about a bond-bending distance constraint, which defines the bond angles according to an atom's chemistry (e.g., sp^3 hybridization).[16] Hinge joints separate rigid clusters. The number of hinge joints will generally be considerably more than the number of residual internal degrees of freedom in the network. This means that many of the dihedral angles associated with the hinge joints are interdependent, and all hinge joints sharing the same degree(s) of freedom are grouped together into collective motions. Rotations through a dihedral angle about the axis of a central-force constraint corresponding to a single bond are possible, but may be locked because of the surrounding network, leading to conformational constraints.[17,18] Furthermore, for double and partial-double bonds, the dihedral angle is fixed, represented here by incorporating a third neighbor distance constraint. Along the main chain there are two dihedral angles about which rotations are possible for each residue (conventionally called Φ and Ψ), and the dihedral angle associated with the peptide bond (Ω) is locked.

Intuitively, one expects that a large rigid cluster, consisting of many weak hydrogen bonds, will not be as stable as a similar rigid cluster involving stronger 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 assess 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 local flexibility and rigidity (stability).

As part of the analysis of *FIRST*, a protein is decomposed into rigid and flexible regions.[1,2] These rigid regions are further classified as overconstrained or isostatic. An overconstrained region (rigid cluster) is one that has more constraints than are necessary to lock all dihedral angles (hinge joints). Thus in an overconstrained region, there are no independent degrees of freedom, but in fact a number of extra constraints. This leads to the idea of redundant constraints, which then can be removed, leaving the cluster still rigid. Isostatic clusters are just rigid, because the number of independent internal degrees of freedom is exactly balanced by the number of constraints. If a constraint is removed from an isostatic cluster, that cluster becomes flexible and decomposes into at least two parts.

It is worth mentioning that although the worst-case computational complexity of *FIRST* is $O(N^2)$, it runs in $O(N)$ in practice, where N is the number of atoms.[1] The same results as those obtained by *FIRST* can be obtained using a brute force matrix method, but such a method is infeasible for large molecular systems because its computational complexity is $O(N^5)$.

This can be reduced to $O(N^3)$ by special methods. We have checked the rigid region decomposition using the brute force method against *FIRST* for many networks up to $N = 450$, and there has been exact agreement between the two methods in each case.

Quantitative Flexibility and Rigidity

A flexible region consisting of many interconnected rigid clusters may define a collective motion having only a few independent degrees of freedom. This flexible region, although underconstrained, could be nearly rigid and thus mechanically stable. However, some rigid regions have more constraints than are needed to maintain structural stability. Due to this continuum between underconstrained and overconstrained regions in the network, a continuous flexibility index is useful.

The total number of floppy modes in a protein, F, corresponds to the number of internal independent degrees of freedom. To obtain F, the six trivial rigid body degrees of freedom (three rotational and three translational degrees of freedom for the molecule as a whole) must be 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 flexibility. However, *FIRST* also locates each underconstrained region and the number of floppy modes within each such region. A quantitative measure of flexibility can be obtained by tracking how the floppy modes are spatially distributed throughout the protein. Similarly, a global count for the number of redundant constraints gives a sense of the overall stability (rigidity) of a protein. *FIRST* identifies where the overconstrained regions are located and how many redundant constraints are present in each region. In a similar way to that done with the floppy modes, a better measure for the degree of rigidity can be obtained by tracking how the redundant constraints are distributed throughout the protein.

The flexibility index, f_i, characterizes the degree of flexibility for the i^{th} central-force bond in the protein. Let H_k and F_k respectively denote the number of hinge joints (rotatable dihedrals) 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 these, a quantitative measure for both the degree of flexibility associated with floppy modes, and the degree of rigidity associated with redundant constraints, is obtained. This flexibility index is given by:

$$f_i = \begin{cases} \dfrac{F_k}{H_k} & \text{for flexible region } k \\ 0 & \text{for an isostatic region} \\ \dfrac{-R_j}{C_j} & \text{for rigid region } j \end{cases} \qquad (2)$$

When the i^{th} central-force bond is a hinge joint, the flexibility index is defined by the number of floppy modes divided by the total number of hinge joints within that underconstrained region. The number of floppy modes corresponds to the number of independent dihedral angle rotations that can be made within the underconstrained region. When the i^{th} central-force bond is *not* a hinge joint, it must be part of a rigid cluster, which may or may not be overconstrained. 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.

FIRST Results for HIV Protease

Figure 3 shows the analysis *FIRST* can provide for any individual protein structure. In this case, the protein structure is the open conformation of HIV protease. Crystallographically, this is a structure of moderate resolution, 2.7 Å, with a crystallographic residual error (R-factor) of 0.19. Two β-hairpin flaps in HIV protease structures are essential for allowing access of substrates and inhibitors to the active site, and are shown at the top of Figures 3A and 3B. These flaps close over such ligands to isolate them in the active site away from solvent[19]. Without an inhibitor in the structure, these flaps are free to move[20]. However, attaching an inhibitor to the protein restricts the motion of these flaps, which are then linked to each other or the ligand through hydrogen bonds. In Figure 3A, we show the rigid region decomposition of the complete protein, including the side groups. The protein's hydrogen bonds, with an energy cut-off of -0.01 kcal/mol, are shown as thin black lines. In Figure 3B, we show a ribbon diagram of flexibility, with the scale indicated to the right. The most flexible regions are red, the isostatic (just rigid) regions are gray, and the most rigid (overconstrained) regions are dark blue. Because only the main chain is represented (though the side groups were included in the calculation), this view provides an indication of regions in which significant conformational change is possible. The two flaps appear at the top, and along with the small loops at the sides (two β-turns appearing in red and yellow in Figure 3B), they are the most flexible regions in the protein. The majority of HIV protease forms a rigid region (blue) forming a cavity (center) for interaction with ligands.

DYNAMICS

We can now use the static results from *FIRST* as input and introduce dynamics to study the range of motions accessible to the protein. The simplest way is to let the flexible regions of the protein move in an unbiased way using Monte Carlo moves. In this approach, small atomic motions are made in the flexible regions, then care is taken to ensure that all the original bond constraints are obeyed. This involves paying particular attention to ring closures, which we do by introducing a fictitious energy and subsequently adjusting the location of the other atoms. In all cases we have to avoid van der Waals overlaps between atoms.

RING DYNAMICS AND CLOSURE

An isolated n-fold single ring, with no double bonds, has $3n$ total degrees of freedom. The number of bond length constraints is n, the number of bond angle constraints is n, and the number of trivial rigid-body degrees of freedom is 6. Thus, the total number of floppy modes is $3n - n - n - 6 = n - 6$. If $n < 6$, the ring is overconstrained. If $n = 6$, the ring is isostatic or just rigid. A six-fold ring has two conformations: the chair and the boat, and there is a potential barrier between these two conformations. A ring is floppy only if it is larger than 6-fold, since there are $n - 6$ floppy modes in an n-fold ring.

An n-fold single ring has $3n$ Cartesian coordinates but only

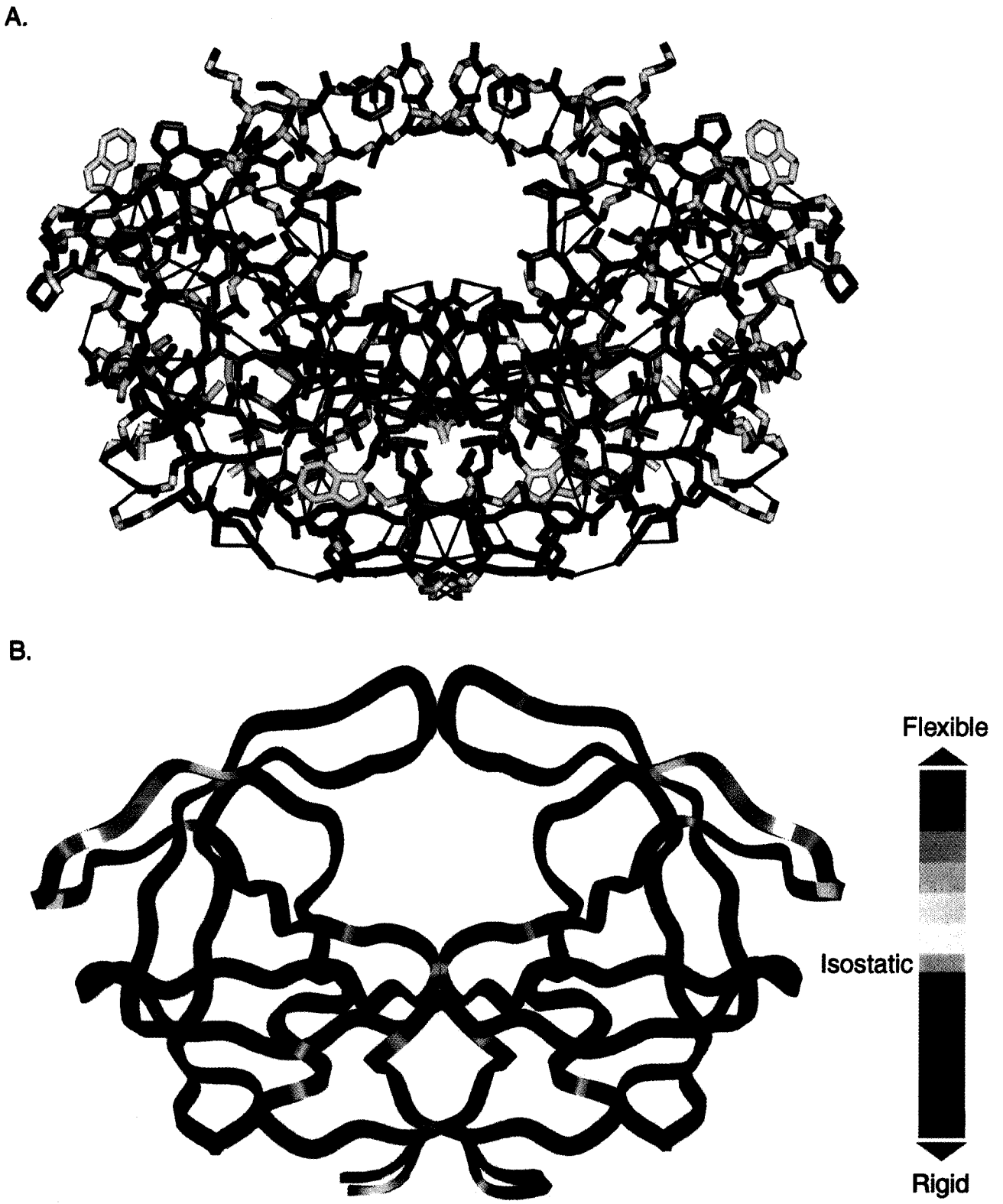

Figure 3. The open conformation of HIV protease (PDB code: 1hhp) with a hydrogen-bond energy threshold of $E_{cut} \leq -0.01$ kcal/mol. (A) The rigid cluster decomposition. Any region of continuous color, blue for example, signifies that these atoms belong to the same rigid cluster. Bonds that split into two colors (e.g. yellow and red) indicate flexible joints. At each of these color interfaces is a rotatable dihedral bond (hinge joint). Hydrogen bonds and salt bridges are shown as black lines. (B) A ribbon diagram of the flexibility index. This image shows the flexibility index, f_i, mapped onto the C_α atoms on the main chain of the protein. The spectrum used to color the ribbon goes from red (flexible and underconstrained) through gray (isostatic) to blue (rigid and overconstrained).

n dihedral angles. Thus there is an advantage to using the n dihedral angles as variables. Since the number of degrees of freedom of the ring is $n - 6$, there must be 6 independent equations to fix the ring conformation. In previous work, Go and Scheraga [21] developed nine equations as the requirements for the ring to close, which we give in a modified and compact form below:

$$\mathbf{d}_0 + T_0R_1\mathbf{d}_1 + \cdots + T_0R_1T_1R_2 \cdots T_{n-2}R_{n-1}\mathbf{d}_{n-1} = 0$$

$$T_0R_1T_1R_2 \cdots T_{n-2}R_{n-1}T_{n-1}R_n = I \quad (3)$$

in which

$$\mathbf{d}_i = \begin{pmatrix} d_i \\ 0 \\ 0 \end{pmatrix} \qquad T_i = \begin{pmatrix} \cos\theta_i & -\sin\theta_i & 0 \\ \sin\theta_i & \cos\theta_i & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$R_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\omega_i & -\sin\omega_i \\ 0 & \sin\omega_i & \cos\omega_i \end{pmatrix} \quad (4)$$

and d_i is the distance from the atom $i-1$ to the atom i, θ_i is the supplement of the angle within a ring at atom i, and ω_i is the dihedral angle associated with the bond between atom $i-1$ and atom i. The first vector equation in (3) expresses the fact that atom 0 and atom n in an n-fold ring are at the same position in space, and the second matrix equation in (3) expresses the fact that the angle at the ring closure has its correct prescribed value. The quantity I is the unit matrix. The first equation in (3) introduces 3 independent constraints. Of the 9 additional constraints introduced in the second matrix equation in (3), an additional 3 are independent, giving a total of 6 independent constraints associated with ring closure.

These equations are called ring closure equations. Every atom is positioned obeying the bond length and bond angle constraints, and rotated about the previous bond with the dihedral angle specified by ω_{i-1}. Atom n will be at the origin if the ring closes itself. This is accomplished by the first vector equation (3). The direction of atom n to atom $n+1$should be the same as that of first atom to the second atom, which is the second matrix equation in (3). Once the $n-6$ dihedral angles are set, the remaining 6 unknowns can be solved from the ring closure equations. All these equations are nonlinear. Go and Scheraga[21,22] showed that if the 6 unknown dihedral angles are in sequence, these nonlinear equations can be solved analytically and the number of independent solutions or ring conformations can be as high as 4. This solution was used to calculate the structure of long peptide chains such as *cyclo*-hexaglycyl with C_n, I, or S_{2n} symmetries.[22] Numerical methods have to be applied to deal with the most common cases where the unknown dihedral bonds are not in sequence, where there is no symmetry, or where nonpeptide bonds such as hydrogen bonds are present in the rings.

Intercorrelated Rings

When rings are interconnected, additional angle constraints arise between the rings. An example is shown in Figure 4, in which two rings share the same bond *AB*. In addition to the bond angle and bond distance constraints already counted in each ring, two more angle constraints, $\angle CAD$ and $\angle EBF$, appear. These extra angle constraints introduce a correlation between the dihedral angle of bond *AB* in the left and right rings. In the coordinates where atoms *E*, *B*, and *A* are fixed, a change of the dihedral angle of bond *AB* in the left ring alters the position of atom *C*. Because of the additional inter-ring angle constraints of $\angle CAD$, the atom *B* has to move to keep this angle fixed. Because atom *F* is fixed by the angle constraints of $\angle EBF$ while atom *D* changes its position, the dihedral angle of bond *AB* in the right ring changes as well. In other words, the dihedral angles of the same bond in different rings that share this bond are not independent. The changes of dihedral angles of the same bond are the same in all rings sharing this bond. Thus the interconnection between rings reduces the number of independent variables, and hence reduces the degrees of freedom. For example, two separate seven-fold rings have two floppy modes, but only one when they share a common bond, as shown in Figure 4.

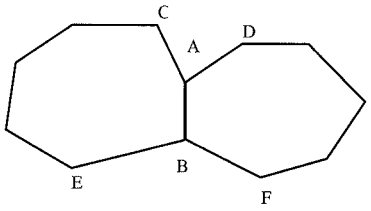

Figure 4. Two seven-fold rings share the same bond AB, *which introduces two extra angle constraints,* $\angle$CAD *and* $\angle$EBF.

Proteins often contain some four-, five-, and six-fold rings, as well as larger rings. The large rings, with size greater than six, are flexible if they are not connected with any other rings and do not contain any peptide bonds. Not only can the inter-ring connections hold the rings together in one rigid cluster, but they also reduce the number of floppy modes in flexible rings. The majority of such inter-ring connections involve hydrogen bonds that link the rings together sufficiently to form complicated correlated motions in the protein.

Network and Branch Atoms

We can partition the atoms in the protein into two types according to the topological properties of the atoms. We define *network* atoms as atoms that belong to a ring—excluding the internal rings associated with the side groups—proline, phenylalanine, histidine, tyrosine, and tryptophan. The atoms in these internal rings may become network atoms if they are members of other rings, such as those formed by hydrogen bonds. Therefore, each network atom has at least two other network atoms as neighbors. All other atoms are called *branch* atoms. In Figure 5, atom *A* is a network atom, while atom *B* is

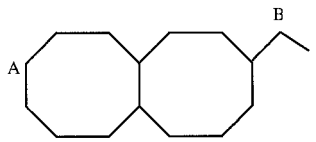

Figure 5. This figure shows the difference between a network atom (A) and a branch atom (B). Network atoms are atoms that belong to a ring—excluding the internal rings associated with the side groups. All other atoms are branch atoms.

a branch atom.

Once the rigidity or flexibility of every atom is determined by *FIRST*, the floppy regions are decomposed into irreducible rings. An irreducible ring is such that there is no shorter path between any two atoms within the ring than the path within the irreducible ring itself. Every irreducible ring has its own ring closure equations to satisfy. The total number of equations is $9M$, where M is the number of irreducible rings. The number of variables is the total number of bonds that belong to these irreducible rings. This means that if the value of the dihedral angle of one shared bond is adjusted in one set of ring closure equations (3), its dihedral angles in all the rings that share this bond should be adjusted accordingly. The dihedral angles of peptide bonds are always fixed, of course.

Monte Carlo Constrained Dynamics

The Monte Carlo dynamics proceeds as follows. We randomly select a few bonds in each flexible region and randomly make small changes in their dihedral angles. These bonds are called *selected* bonds. The dihedral angles of the selected bonds are not allowed to change during one iteration of searching for a new conformation. The more selected bonds there are, the less likely it will be to find a solution for the dihedral angles of the remaining bonds to satisfy the ring closure equations (3) in every irreducible ring. An extreme case will illustrate this point. Assume the dihedral angles of six out of seven bonds in a seven-fold ring are arbitrarily changed from their initial values. In that case, it is highly unlikely that there is a suitable value for the remaining dihedral angle that will close the ring. But selecting too few bonds limits the range of conformational change that is sampled. We have found that a good ratio of the fraction of the number of selected bonds to the total number of bonds in a flexible region lies between 0.1 and 0.2. As for the dihedral angles of selected bonds, they should not be changed too much at each iteration, otherwise the system may go over a barrier to get from one allowed conformation to another. We are only searching for continuous deformations in this work, and introduce a fictitious energy E,

$$E = (\mathbf{d}_0 + T_0R_1\mathbf{d}_1 + \cdots\cdots + T_0R_1T_1R_2\cdots T_{n-2}R_{n-1}\mathbf{d}_{n-1})^2 + \sum_{ij}(T_0R_1T_1R_2\cdots\cdots T_{n-2}R_{n-1}T_{n-1}R_n - I)_{ij}^2 \quad (5)$$

where the elements of the 3-by-3 matrix are given by the subscripts ij.

We solve for the dihedral angles of the remaining flexible bonds. To solve this general case, it is necessary to go beyond previous work, which has only addressed special cases (as described previously). The nonlinear equations for all the rings must be solved simultaneously. To do this, we use the fictitious energy function as given in (5), and search for a minimum. This is not a real energy, but is constructed such that if a minimum energy of zero can be found, this means that *all* the equations (3) are all satisfied simultaneously. Any numerical minimization method is suitable for this, and we use the conjugate gradient code from Numerical Recipes.[23] This procedure is essentially the same as putting in central forces between atoms 0 and n and atoms 1 and $n-1$ to close each irreducible ring at the correct angle.

If the above energy (5) can be minimized to zero, then there exist solutions to the ring closure equations (3). Otherwise, that step is aborted and any changes are discarded, and a new iteration is begun. If a solution to the ring closure equations (3) exists, we then use a simplex method to calculate the coordinates of the associated branch atoms. If there are no van der Waals collisions between atoms, we accept the new conformation and proceed to another iteration until a large number of conformations is sampled. The simplex method[24] is outlined in the next section.

Simplex Method

Once a new conformation of the bond network is available, the coordinates of the branch atoms have to be adjusted, subject to the constraints and making sure to avoid all collisions due to the van der Waals radii overlapping. The bond angle and bond length constraints form equations that are equalities, while the van der Waals constraints, which are lower bounds for the distances between pairs of atom centers (based on the sums of their van der Waals radii), lead to inequalities. All these equations (equalities and inequalities) are nonlinear.

The simplex method requires a linear function to minimize. The new coordinates of the branch atoms should not be far from their original values; otherwise, they are likely to collide with other atoms or move away from the protein. This requirement produces a nonlinear function to be minimized. The nonlinear equations can then be solved by successive linear iterations. So now we have a linear function to minimize, under the condition that the new conformation must satisfy the linear equations, which can be written as a constraint matrix. Both equality and inequality constraints are present in the matrix.

The basic idea of the simplex method is that in a linear system the extreme point of the function is located at one of the extremities of the allowed domain. Imagine an n-dimensional space, where every equality constraint is an $n-1$ dimensional hyperplane in the space. Every inequality equation cuts the space into two. When all the equality and inequality constraints are applied, the space is reduced into a lower dimensional subspace in which all points satisfy all the constraints. Because the function is linear, the minimum point of the function must be at one of the extremities of the subspace, which limits the search to a rather small number of isolated points. A detailed explanation of the simplex method can be found in many function optimization books.[25,26] This algorithm is very fast in handling the minimization of linear functions. Because we use linear constraints and a linear function to approximate the nonlinear expression, the procedure must be repeated several times to achieve an acceptable accuracy. Even so, this procedure is very fast compared with other nonlinear function minimization methods.

Calculation Results

By repeating the procedure outlined above, we are able to create 300 possible conformations of the HIV protease in a few hours on a DEC Alpha 433 MHz workstation. A superposition of these 300 conformations is shown in Figure 6. Figure 6A shows the motion accessible to the side groups, based on salt-bridge and covalent and hydrogen-bond constraints; Figure 6B shows the accessible range of main-chain motion, using the same 300 conformations. We emphasize that there is no temperature in this calculation, though it is a Monte Carlo procedure. A new conformation is either accepted or rejected based

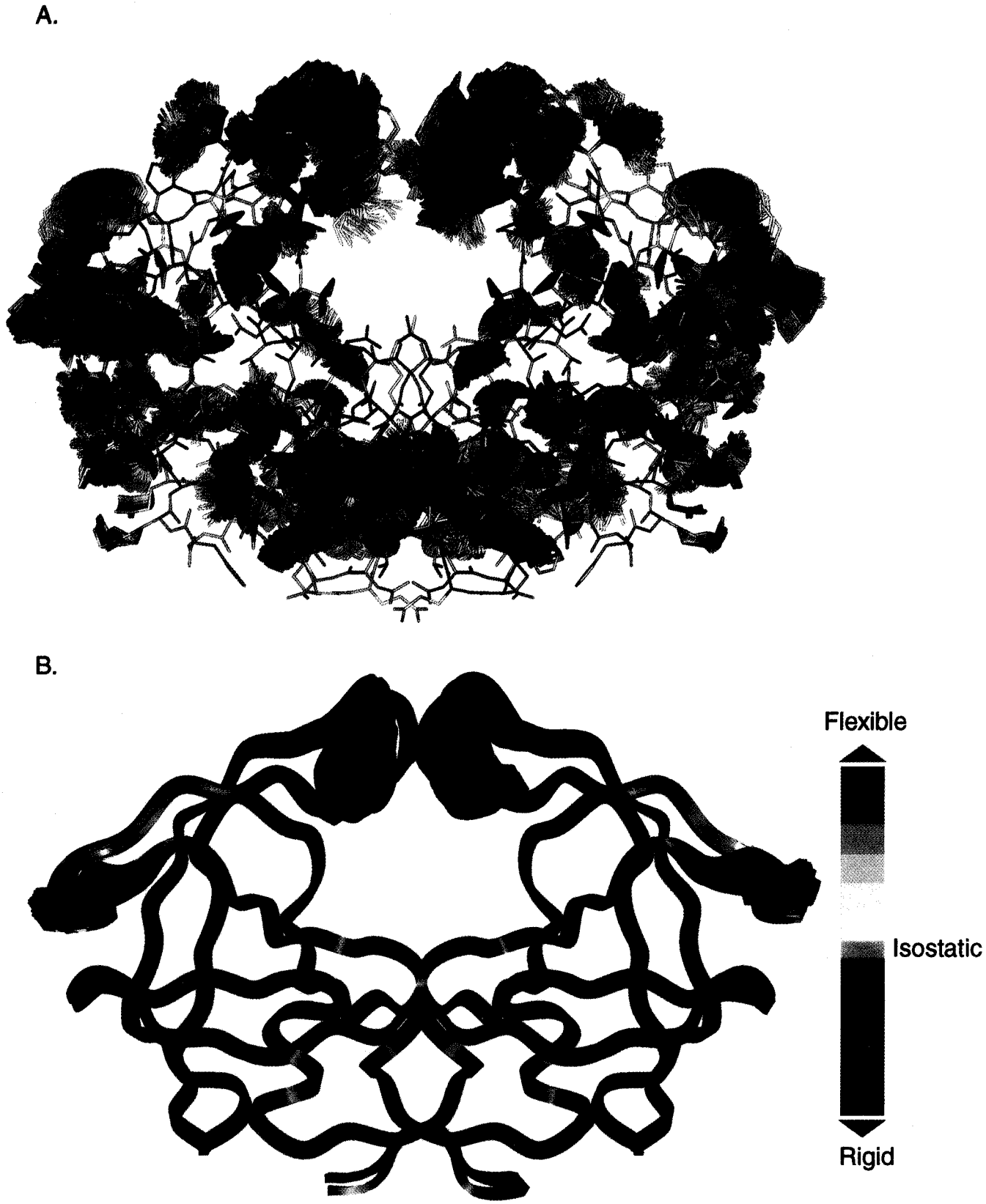

Figure 6. This figure shows the superimposition of 300 allowed conformations associated with the open conformation of HIV protease (PDBcode:1hhp) with a hydrogen bond energy cutoff of $E_{cut} \leq -0.01$ kcal/mol. An allowed conformation obeys all the covalent and hydrogen-bond length and bond angle constraints and does not violate any van der Waals constraints. The rigid core is indicated by the color blue. (A). The rigid region decomposition (as in Figure 3A), where motions of the main chain and side chains can be seen. (B) The ribbon graph of the flexibility index as in Figure 3B. The motion of the flexible red regions is apparent.

on whether every constraint is satisfied or not, and whether there are van der Waals overlaps or not.

A movie has been made to show the path of the conformational evolution of HIV protease. This is not a MD method, so the movie composed of the 300 conformations does not show a time evolution of HIV protease, but merely samples the unbiased, allowed conformations of HIV protease. These movies can be found at the web site http://www.pa.msu.edu/~lei/. These results should not be compared with molecular dynamics simulations, which are on short time scales of up to typically a few nanoseconds, but rather with the motion made up by interpolating between different experimentally observed conformations, as observed by Gerstein et al.,[27] available at http://bioinfo.mbb.yale.edu/perl/motreport.pl?ID=hivprot. We will refer to these as "interpolations" between crystallographically observed conformations. Comparing our results, visualized here in Figure 6, with the interpolations, we find that the red regions in Figure 6B correspond to the regions of the largest motion in the animation, with the main flaps at the top moving by ~7Å in both cases. The motion of the other red regions in Figure 6B is in qualitative, but not quantitative, agreement with the animations. This is not unexpected, as the interpolations involve conformations observed upon HIV protease binding to different ligands, whereas we are analyzing a single, ligand-free structure. Also, some conformations found by *FIRST* conformational sampling may be feasible but not observed in the crystallographic structures, which only include those conformations that will crystallize in a regular lattice.

DISCUSSION

We have shown how the rigid regions and the flexible joints between them can be identified for a protein using a static approach involving constraints via the program *FIRST*. This knowledge is extremely useful when moving on to study the dynamics of the protein, in particular, diffusive, large-scale motion. We have taken initial steps in the study of the dynamics by just exploring the space available to the flexible part of the protein using unbiased Monte Carlo moves and avoiding any collisions due to overlapping van der Waals radii. We refer to such moves as legal moves. This could be improved by using a potential function for the protein and accepting or rejecting legal moves based on a Metropolis[28] criterion. This would involve using a Boltzmann factor that depends on the difference in energy between the initial and new conformation, and the temperature.

A more ambitious approach would be to use the static input provided by *FIRST* in a molecular dynamics[29] program that separates into two distinct time scales. Most of the time the molecular dynamics would proceed using only the diffusive motion associated with the flexible regions, as defined by *FIRST*, and this would involve a much longer time step than can usually be used. This would use a potential function for the protein, projected into the subspace defined by the floppy modes. Occasionally a full molecular dynamics simulation would be done, which would allow the whole protein to execute higher frequency motions for a short time. While this remains untested, it opens up the possibility of using *FIRST* in a much more powerful way to probe the dynamics of proteins using a realistic potential, going beyond the constraint model.

CONCLUSIONS

We have shown how the strong local forces in a protein can be used to find the rigid regions and the flexible joints between them. This procedure takes only a few seconds of computer time and provides a useful way to get a good initial idea of the rigid and flexible regions of a protein. The results can be visualized in a number of ways, some of which are shown here. The results of this approach correlate favorably with experimental measures of protein flexibility.[30] We have used these static results to investigate the possible motions of the flexible regions of the protein, such that the bond constraints are maintained and collisions between atoms, due to their being closer than the sum of their van der Waals radii, are avoided. To date we have explored the simplest dynamics using unbiased Monte Carlo moves, utilizing a fictitious energy function to ensure closure of rings of atoms in the bond network, with a simplex method used to assure that the motion of branch atoms does not induce van der Waals collisions. Our results have been illustrated using HIV protease, and these techniques can be applied to any protein whose three-dimensional structure is known.

ACKNOWLEDGMENTS

We acknowledge support from NSF grant DMR 0078361 and a grant from Michigan State University. We thank Brandon Hespenheide and Yuquing Xiao for useful comments and A. Roy Day for a critical reading of the manuscript. We also thank S.D. Guest for introducing us to the simplex method.[31,32]

REFERENCES

1 Jacobs, D.J., Kuhn, L.A., and Thorpe, M.F. Flexible and rigid regions in proteins. In: *Rigidity Theory and Applications*, Thorpe, M.F., and Duxbury, P.M., Eds., Kluwer Academic/Plenum Publishers, New York, 1999, pp. 357–384

2 Jacobs, D.J., and Thorpe, M.F. Generic rigidity percolation in two dimensions. *Phys. Rev. Lett.* 1996, **75**, 4051–4054

3 Lu, H., and Schulten, K. Steered molecular dynamics simulation of conformational changes in immunoglobulin domain 127 interpret atomic force microsocopy observations. *Chemical Physics* 1999, **247**, 141–153

4 Abagyan, R., Totrov, M., and Kuznetsov, D. ICM – A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comp. Chem.* 1994, **15**, 488–506

5 Jeffrey, G.A. *An Introduction to Hydrogen Bonding*, Oxford University Press, New York, 1997

6 Fersht, A.R. The hydrogen bond in molecular recognition. *Trends in Biochem. Sci.* 1987, **87**, 301–304

7 Vriend, G. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* 1990, **8**, 52–56

8 Habermann, S.M., and Murphy, K.P. Energetics of hydrogen bonding in proteins: A model compound study. *Protein Sci.* 1996, **5**, 1229–1239

9 Dahiyat, B.I., Gordon, D.B., and Mayo, S.L. Automated design of the surface positions of protein helices. *Protein Sci.* 1997, **6**, 1333–1337

10 Kumar, S., and Nussinov, R. Salt bridge stability in

monomeric proteins. *J. Molec. Biol.* 1999, **293**, 1241–1255
11 Barlow, D.J., and Thornton, J.M. Ion-pairs in proteins. *J. Molec. Biol.* 1983, **168**, 867–885
12 Gandini, D., Gogioso, L., Bolognesi, M., and Bordo, D. Patterns in ionizable side chain interactions in protein structures. *Proteins: Structure, Function, and Genetics* 1996, **24**, 439–449
13 Xu, D., Tsai, C.-J., and Nussinov, R. Hydrogen bonds and salt bridges across protein–protein interfaces. *Protein Engineering* 1997, **10**, 999–1012
14 Thorpe, M.F., Hespenheide, B.M., Yang, Y., and Kuhn, L.A. Flexibility and critical hydrogen bonds in cytochrome c. In: *Pacific Symposium on Biocomputing*, Altman, R.B., Dunker, A.K., Hunter, L., Lauderdale, K., and Klein, T., Eds., World Scientific, Singapore, 2000, pp. 191–205.
15 Dill, K.A. Dominant forces in protein folding. *Biochemistry* 1990, **29**, 7133–7155
16 Jacobs, D.J. Generic rigidity in three-dimensional bond-bending networks. *J. Phys. A: Math. Gen.* 1998, **31**, 6653–6668
17 Lesk, A.M. and Chothia, C. Mechanisms of domain closure in proteins. *J. Molec. Biol.* 1984, **174**, 175–192
18 Gerstein, M., Schultz, G. and Chothia, C. Domain closure in adenylate kinase: joints on either side of two helices close like neighboring fingers. *J. Molec. Biol.* 1993, **229,** 494–501
19 Rose, R.B., Craik, C.S., and Stroud, R.M. Domain flexibility in retroviral proteases: Structural implications for drug resistant mutations. *Biochemistry.* 1998, **37**,2607–2621
20 Nicholson, L.K., Yamazaki, T., Torchia, D.A., Grzesiek, S., Bax, A., Stahl, S.J., Kaufman, J.D., Wingfield, P.T., Yam, P.Y.S., Jadhav, P.K., Hodge, C.N., Domaille, P.J., and Chang, C.-H. Flexibility and function in HIV-1 protease. *Nature Structural Biology* 1995, **2**, 274–280
21 Go, N., and Scheraga, H.A. Ring closure and local conformational deformations of chain molecules. *Macromolecules* 1970, **3**, 178–184
22 Go, N., and Scheraga, H.A. Calculation of the conformation of cyclo-hexaglycyl. *Macromolecules*, 1973, **6**, 525–535
23 Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T, *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1986
24 Kuenzi, H.P., Tzschach, H.G., and Zehnder, C.A. *Numerical Methods of Mathematical Optimization*, Academic Press, New York, 1971
25 Gill, P. E., Murray, W., and Wright, M. H. *Practical Optimization*, Academic Press, New York, 1981
26 Bradley, S. P., Hax, A. C., and Magnanti, T. L. *Applied Mathematical Programming*, Addison-Wesley Publishing Company, Massachusetts, 1977
27 Gerstein, M., and Jansen, R. Studying the macromolecular motions in a database framework: From structure to sequence. In: *Rigidity Theory and Applications*, Thorpe, M.F. and Duxbury, P.M., Eds., Kluwer Academic/Plenum Publishers, New York, 1999, pp. 401–419
28 Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 1953, **21**, 1087–1092
29 Case, D. Molecular dynamics and normal mode analysis of biomolecular rigidity. In: *Rigidity Theory and Applications*, Thorpe, M.F., and Duxbury, P.M., Eds., Kluwer Academic/Plenum Publishers, New York, 1999, pp. 329–344
30 Jacobs, D.J., Rader, A.J., Kuhn, L.A., and Thorpe, M.F., Graph theory prediction of protein flexibility. *Proteins: Structure Function and Genetics 2001, in press*
31 Kangwai, R.D., and Guest, S.D. Symmetry-adapted equilibrium matrices. *Int. J. Solids Struct.*, 2000, **37,** 1525–1548
32 Kangwai, R.D., Guest, S.D., and Pellegrino S. Introduction to the analysis of symmetric structures. *Computers & Structures* 1999, **71**, 671–688