FLEXIBILITY AND CRITICAL HYDROGEN BONDS IN CYTOCHROME *C*

M. F. THORPE
Physics & Astronomy Department, Michigan State University
East Lansing, MI 48824, USA
e-mail: thorpe@pa.msu.edu

BRANDON M. HESPENHEIDE, YI YANG, and LESLIE A. KUHN
Department of Biochemistry, Michigan State University
East Lansing, MI 48824, USA
e-mail: brandon@sol.bch.msu.edu, yangyi@pilot.msu.edu, kuhn@agua.bch.msu.edu

We show that protein flexibility can be characterized using graph theory, from a single protein conformation. Covalent and hydrogen bonds are modeled by distance and angular constraints, and a map is constructed of the regions in this network that are flexible or rigid, based on whether their dihedral bonds remain rotatable or are locked by other interactions in the network. This analysis takes only a second on a typical PC, and interatomic potentials; the most time-consuming aspect of molecular dynamics calculations, are not required. Our preliminary work has shown that this approach identifies the experimentally observed, biologically important flexible regions in HIV protease and lysine-arginine-ornithine binding protein. Here we analyze three evolutionarily distant cytochromes *c*, and find strong similarity between their flexible regions, despite having only 39% sequence identity. Furthermore, we show how the structural flexibility increases as the weaker hydrogen bonds are removed, as would happen under thermal denaturation of the protein. This approach identifies the critical hydrogen bonds that cross-link the tertiary structure.

1 Introduction

Predicting flexibility in proteins has proven elusive. For example, molecular dynamics simulations with realistic potentials cannot be run for long enough times to extract the large-scale motions that correspond to low frequency motions in normal modes analysis [1]. Thus, the development of techniques for studying protein flexibility remains an important current problem. Proteins are held together by several kinds of forces, of which the most important are the covalent forces that determine many bond lengths and angles, including the dihedral angles associated with peptide bonds. During protein folding, once hydrophobic collapse of the polypeptide chain has occurred, hydrogen bonds are responsible for holding together the secondary structures, principally α-helices and β-sheets, and stabilizing the higher-order structures in which helices and sheets fold together into the complex units responsible for the diverse biological

activity of proteins.

2 Representing the Covalent and Hydrogen-Bond Structure as a Network

A useful insight into modeling protein flexibility is that a protein structure can be reduced to its essentials, and viewed as a mechanical system of points (atoms) whose motion is limited by distance and angle constraints [2,3] representing the interatomic bonds. These constraints are sometimes sufficient to render a region immobile (i.e. rigid) and sometimes some freedom remains which allows motion to occur. When applied to the covalent and hydrogen-bond network of a single, static protein structure, this approach [4,5,6] can predict the important rigid and flexible features in proteins, including HIV protease, dihydrofolate reductase, adenylate kinase and lysine-arginine-ornithine binding protein [6,7].

Proteins are held together by several kinds of forces, of which the most important are the covalent forces that fix many bond lengths and angles, including the dihedral angles associated with peptide bonds. Many other forces are also important, including hydrophobic, van der Waals, hydrogen-bond, and electrostatic interactions. Our hypothesis is that once the protein is in the folded state, the strong short-ranged forces mainly determine the flexibility; thus, we focus our attention on the covalent and hydrogen bonds and salt bridges, considered as a special case of hydrogen bonds. The idea is that when parts of the folded protein move, this motion must be consistent with maintaining the covalent bonds and a majority of the hydrogen bonds, as their breakage would involve too high an energy penalty. On the other hand, local motion is not so costly against the other forces in the protein, as they are more diffuse and do not involve individual bonds, but rather consist of *regions* interacting with each other. Even hydrophobic interactions do not seriously inhibit local flexibility, as they tend to be fairly slippery and nonspecific.

In modeling a protein as a bond network or graph, the covalent bond between two adjacent atoms, say C_α–C in an amino acid, is considered to fix the distance between these two points, so that all motions must be consistent with this constraint. An angular constraint, reflecting the bond angle, is represented by a constraint between second-neighbor atoms, e.g., between the N and C neighbors tetrahedrally coordinated to a main-chain C_α atom. An additional constraint is introduced between third-neighbor O and H atoms in the O–C–N–H group, to prohibit any rotation around the C–N peptide bond linking amino acids along the protein main chain. Such constraints restrict the possible motions of the main and side chains, and have also been used in

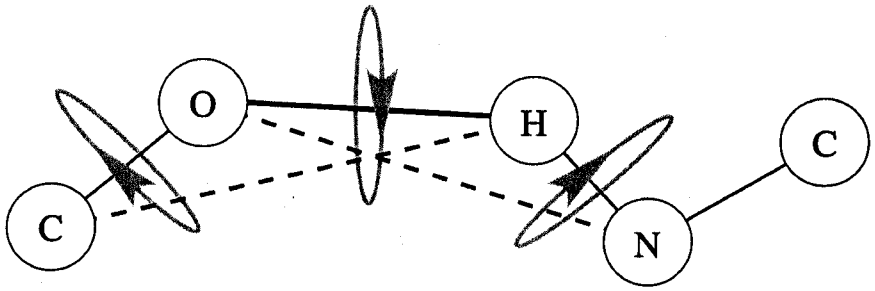

Figure 1: Model of a hydrogen bond, involving donor and acceptor atoms, shown as nitrogen and oxygen, respectively. Covalent bonds are shown as thin black lines. The hydrogen bond is modeled as three generic distance constraints, consisting of a nearest-neighbor central-force constraint shown as a thick solid line (top center), and two next-nearest-neighbor bond-bending force constraints (maintaining the donor-hydrogen-acceptor angle at the experimentally observed value), shown as dashed lines. Each hydrogen bond is also associated with three *a priori* rotatable dihedral angles, indicated by the arrows. (In this example, no rotation is allowed about the C–O bond when it is a double bond – for instance, in the carbonyl group of the protein main chain.)

molecular dynamics calculations[8,9] to reduce the number of dynamical degrees of freedom. A wide range of hydrogen-bond strengths are found in proteins, and we use the following criteria. A hydrogen bond is said to form[6] if the donor–acceptor distance is less than 3.6 Å and the donor–hydrogen–acceptor angle is greater than 90°, and the energy is less than −0.1 kcal/mol, using a hydrogen bond energy function[10]. These hydrogen bonds are included in the network used to represent the protein for graph theoretical analysis. The acceptance criteria can be made more stringent by decreasing the energy cut-off (remembering that the hydrogen-bond energy is negative), resulting in only the strongest hydrogen bonds being included. Thus the stronger terms in the potential are included as constraints and the weaker forces are not. The scheme for incorporating hydrogen bonds into the network is shown in Figure 1.

Once these constraints are identified, the search for rigid regions and the intervening flexible joints can begin. This is referred to as *rigid region decomposition*[6]. For small molecules, the rigid regions and flexible joints are found as follows. An unconstrained point (atom) has 3 degrees of freedom in 3-dimensional space. For the r atoms in an r-fold ring (e.g., $r = 5$ for a proline side chain), there are $3r$ degrees of freedom, which are reduced by the r covalent bond-length constraints in the ring and by another r covalent bond-angle constraints. This leaves a residual r degrees of freedom, of which 6 are the macroscopic rigid body motions involving the whole molecule. Therefore,

there are $r-6$ internal bond-rotational degrees of freedom, or floppy modes [12] associated with a ring. These are zero-frequency continuous deformations of the molecule consistent with the constraints and therefore do not cost any energy. When $r = 6$, the number of floppy modes is zero, and the region is said to be isostatically rigid [4,5,6]. For $r > 6$, the ring is underconstrained, or flexible. Thus a seven-fold ring has a single floppy mode that can be visualized as a rolling motion around the perimeter. A five-fold ring has one constraint more than is needed for rigidity; thus, it is overconstrained with one redundant constraint, creating stress within the ring. The relative flexibility of each bond in the ring can be quantified by a flexibility index, $f = (r-6)/r$, giving the number of floppy modes (residual independently rotatable dihedral angles) within the ring, divided by the number of bonds in the ring. For an overconstrained ring, the flexibility index is negative, giving the number of redundant constraints per bond. Thus, for flexible regions the flexibility index represents the *density of floppy modes*, and for rigid regions it represents the *density of redundant constraints* within the region.

This simple counting by inspection cannot be extended much beyond single rings. However new concepts from graph rigidity theory [6,13] can be applied to macromolecules of the size and complexity of proteins, where it is difficult to determine which constraints are independent, and it is no longer possible to count constraints manually. The formal way to tackle such problems is via a branch of graph theory that was first formalized by Laman [14]. For two-dimensional systems, Laman's theorem states that, beyond a global count of degrees of freedom, [3] every subgraph in the network must be evaluated in a similar way. Here, these ideas are extended to 3-dimensional systems.

Laman's theory cannot be generalized to 3-dimensions in general, but we have been able to make a generalization if there are angle bending forces for *every* pair of central forces at an atom, meaning that the bond angle defined by the molecular orbitals is held fixed. Fortunately this is always so with covalent bonds, and hydrogen bonds can also be modeled this way, with the hydrogen atom being placed explicitly between the acceptor and donor. All such bond angles, which are represented by second-neighbor forces, are then constrained [5,6]. In addition the dihedral angle rotation associated with the peptide bond is locked, by including a central force constraint between third-neighbor atoms located on either side of the peptide bond. There are no additional angular constraints associated with this third neighbor constraint, which is added after all the other kinds of constraints are in place.

A new combinatorial algorithm has been applied that balances local constraints with degrees of freedom. We find that the computational performance is linear in the number of atoms in the protein. This general approach to

constraint counting is referred to as the *pebble game*[15,16,17,18]. Three virtual pebbles (representing three degrees of freedom) are associated with each atom in the network, and subsequently moved onto bonds to identify the independent constraints. The *pebble game* has been used in physics problems to study rigidity percolation in 2-dimensional triangular networks[15,16] and also in 3-dimensional glass networks[18]. The *pebble game* has now been embedded in the *FIRST* algorithm, which incorporates protein stereochemistry, particularly the dependence of covalent bond rotation on bond order (single, partial double, or double), and the dependence of hydrogen bonds and their strength on the bond length and angle. Here we apply *FIRST* to analyze and compare flexibility for cytochromes *c* from three species.

3 Calculating Flexible and Rigid Regions in Cytochrome *c*

The bond network representation for horse cytochrome *c* is shown in Figure 2, based on the crystallographic structure (Protein Data Bank, or PDB, code 1hrc). Covalent (thick tubes) and hydrogen bonds (thin black lines) define the network; side-chain covalent and hydrogen bonds are also included in the calculation but are omitted from the figure for clarity. Because hydrogen atoms are not visible at typical crystallographic resolutions, they are added to the structures using the *WhatIf* software package[19]. In this analysis, crystallographic waters with $\geq$ 50% solvent accessibility are discarded from the structure. We regard these waters as mobile and therfore not part of the static structure. In some cases, the hydrogen atom position is unambiguous due to the bond type (e.g., for amide nitrogens). In cases where the position is ambiguous (e.g., for hydrogen atoms in hydroxyl groups), *WhatIf* positions them so as to optimize hydrogen-bond opportunities. Hydrogen bonds are then defined using a rough geometric screen (donor–acceptor distance of $\leq$ 3.6 Å and donor-hydrogen-acceptor angle of $\geq$ 90°), followed by identifying the subset of these bonds having favorable energy ($\leq$ -0.1 kcal/mol), considering the distance and angles between donor and acceptor groups[10]. Salt bridges are considered to be a special case of hydrogen bonds, with lengths up to 4.6 Å and angles of $\geq$ 80° allowed in the initial screen.

Once the covalent and hydrogen-bond networks are defined, this information is provided as input to *FIRST* for identifying dihedral angles that remain rotatable in the system, versus those fixed by the bond and angular constraints. *FIRST* also identifies rigid clusters and intervening flexible joints based on this information, and calculates the flexibility index. The flexibility index is calculated by identifying a flexible region and counting the number of remaining degrees of freedom (i.e. floppy modes) in that region. The flexibility index is

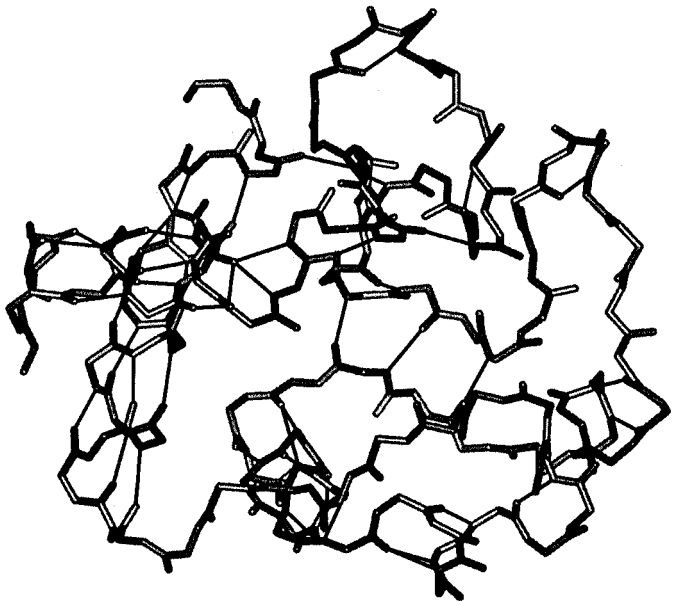

Figure 2: The network of main-chain covalent (grey tubes) and hydrogen bonds (thin black lines) in horse cytochrome c (PDB code 1hrc). The different shades of grey for the covalent bonds are shadows due to the 3-dimensional rendering

then given by the number of floppy modes per bond within that region. The flexibility index for bonds in a region that is rigid but contains no redundant bonds (i.e. an isostaticly rigid region) is zero. The flexibility index is negative for the bonds in an overconstrained region, and is its magnitude is defined as the number of redundant constraints per bond within that region where the bonds are locked. As a result, the rigid and flexible regions within the protein are defined; when horse cytochrome c is analyzed this way, with all hydrogen bonds and salt bridges included, the main chain and covalently-bound heme (latter not shown) comprise a single rigid cluster in Figure 2.

4 Decoding the Hierarchy of Flexibility in Cytochrome c

Considering the structure of cytochrome c as one rigid cluster is overly simplistic, as not all parts of the structure are equally rigid, depending on how many redundant bonds are in the vicinity. We are especially interested in

Figure 3: Multiple sequence alignment of cytochromes *c* from horse (PDB code 1hrc), yeast (1ycc), and bacteria (1co6_A). Identical residues are shaded in black, and those with structural or chemical similarity are shaded gray; - indicates a gap in alignment, due to no corresponding residue in the sequence.

those regions where flexibility will be induced first upon breakage of hydrogen bonds, say by interactions with a molecular partner such as cytochrome *c* oxidase, or by thermal or chemical denaturation. Therefore, we consider the result of diluting the hydrogen-bond network by individually removing the weakest hydrogen bonds (those most susceptible to thermal breakage), then observing the effect on the flexibility of the protein. We are also interested in the extent to which the flexible and rigid regions are similar for different species, and therefore have analyzed cytochromes *c* from three distant relatives: horse, yeast, and *Rhodopseudomonas.* In Figure 3, the ClustalW (http://www2.ebi.ac.uk/clustalw) sequence alignment is shown for the three species, which exhibit 39% identity over the aligned region.

The ease in which the bond network description of a protein may be modified in *FIRST* – for example, to reflect thermal denaturation or a side-chain mutation – allows considerable ability to probe protein rigidity and flexibility. Another potential application is the prediction of protease-sensitive sites in proteins. Thornton and colleagues [20] found that sites susceptible to proteolytic cleavage or nicking must adopt a canonical main-chain conformation to fit the protease active site, and this usually involves changes in the conformation of at least 12 residues. Thus, *FIRST* prediction of main-chain flexibility for 12 or more adjacent residues is expected to correlate with sensitivity to proteolysis.

The decomposition of the protein into rigid and flexible regions can be compared before and after the modification in the hydrogen-bond network, so that any change in flexibility can be attributed directly to that set of hydrogen bonds modified. Here, this approach is used to study the onset of flexibility in cytochrome *c* upon thermal dilution of the hydrogen-bond network. Each hydrogen bond in the protein is assigned an energy, based on the distance and angle-dependent function of Mayo *et al.* [10]. Then, the hydrogen bonds are broken individually, from weakest to strongest, and any resulting change in

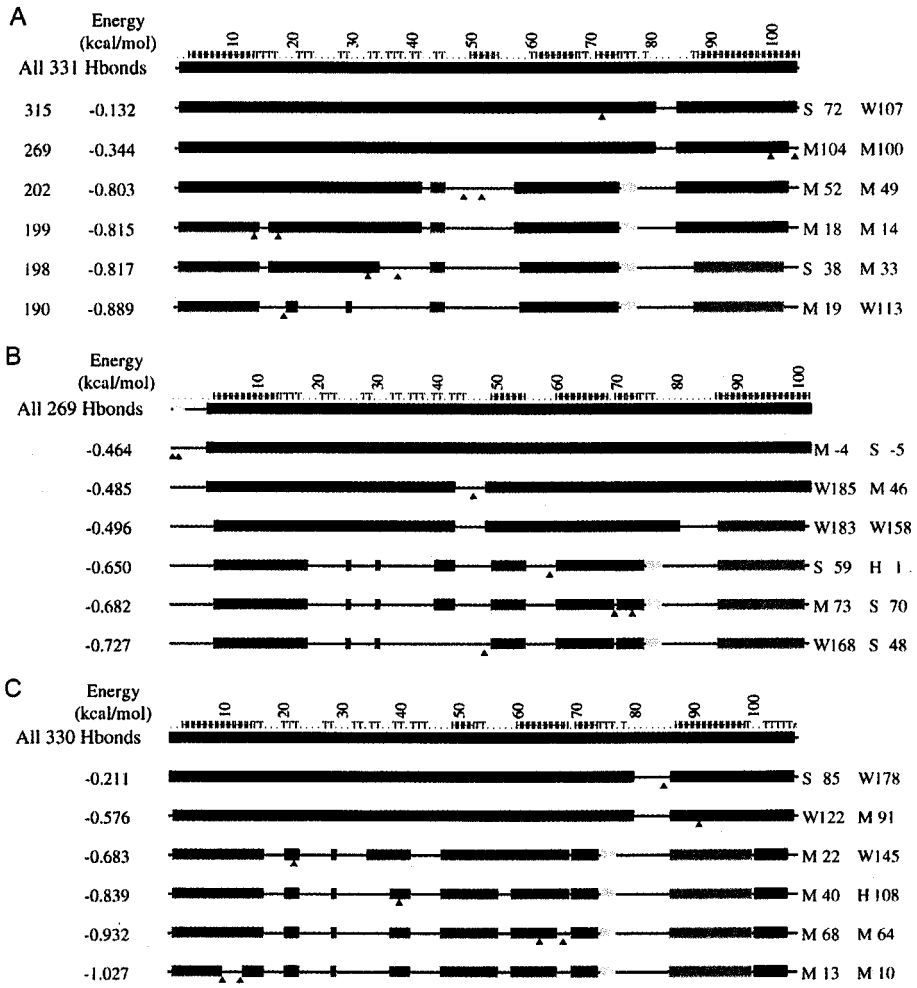

Figure 4: The structural decomposition of cytochrome *c* into rigid clusters, showing the increase in flexible regions as hydrogen bonds are removed from weakest to strongest. Panel (A) is for the horse structure, (B) for yeast, and (C) for *Rhodopseudomonas*. With all hydrogen bonds included (top line in each panel), the protein and its heme form a single rigid cluster (red bar). New rigid cluster decompositions, with a separate colored bar for each cluster, are shown for each hydrogen-bond deletion that changes the clusters; intervening black lines represent flexible regions. The first column indicates the strength of the hydrogen bond that was removed, and carets beneath the clusters indicate the hydrogen bond whose breakage resulted in that set of rigid clusters. The code at the end of each line gives the donor identity in blue, with PDB residue number and S for side-chain, M for main-chain, W for water molecule, or H for heme, with the same information in red for the acceptor. Secondary structure assignments (H for helix, T for turn) from *DSSP* appear at the top.

protein flexibility is noted. In Figure 4, those hydrogen bonds whose removal caused a change in the number or size of the rigid clusters in the main chain are shown for structures from the three cytochromes *c*. The first line in each panel identifies the rigid clusters when all of the hydrogen bonds are present. The total number of hydrogen bonds in each protein is listed on the upper left-hand corner of each panel. Subsequent lines depict changes in rigid clusters due to the incremental removal of hydrogen bonds. Each line is interpreted as follows (from left to right): the energy of the hydrogen bond removed, the rigid clusters present in the protein after removal of that hydrogen bond, and the residues that formed the broken hydrogen bond. Non-adjacent regions of the main chain may belong to the same rigid cluster; see, for instance, the second line of Figure 4A, in which a short, flexible loop around residue 83 is attached to one rigid cluster (red bars) formed by two discontiguous regions of the sequence.

Removal of very weak hydrogen bonds (top bars in each panel) has little or no effect on the rigidity of the protein; in all three cases, the structures retain a single large cluster, with small regions becoming floppy (Figure 4). However, once hydrogen bonds in the region of -0.65 to -0.80 kcal/mol begin to be removed (corresponding to the thermal fluctuation of hydrogen bonds, in k_BT, just above room temperature), each of the proteins fragment into smaller, independent clusters that correspond to the major secondary structural features (the N- and C- terminal helices and "60's" helix). In fact, these regions are also found to be the most stable units in hydrogen-exchange NMR experiments on horse cytochrome *c* [21]. The study of Ptitsyn [22] indicates that only seven residues are conserved across all known subfamilies of cytochromes c. He notes that the four of these residues that are not heme ligands form a "folding cluster", the core of the packing interface between the N- and C-terminal helical pair, which is the most stable region in cytochrome *c*. A main-chain hydrogen bond between two of the absolutely conserved heme ligand residues, Cys 14 and His 18, likely is important in orienting their side chains for heme ligation, and proves to be one of the most critical hydrogen bonds for the protein's stability identified by *FIRST* (line 5 in Figure 4A). This approach of removing hydrogen bonds, then observing the resulting stable and flexible regions of the protein structure, provides a means of identifying structurally critical hydrogen bonds whose removal fragments the tertiary structure. They are prime candidates for mutagenesis to probe or increase protein stability.

The effects of the hydrogen-bond dilution on rigidity/flexibility, when viewed in the context of the three-dimensional structures, are also quite similar for the three species (Figure 5). The left-hand panels in this figure show the rigid clusters (ribbons) in horse, yeast, and bacterial cytochromes *c* at the point

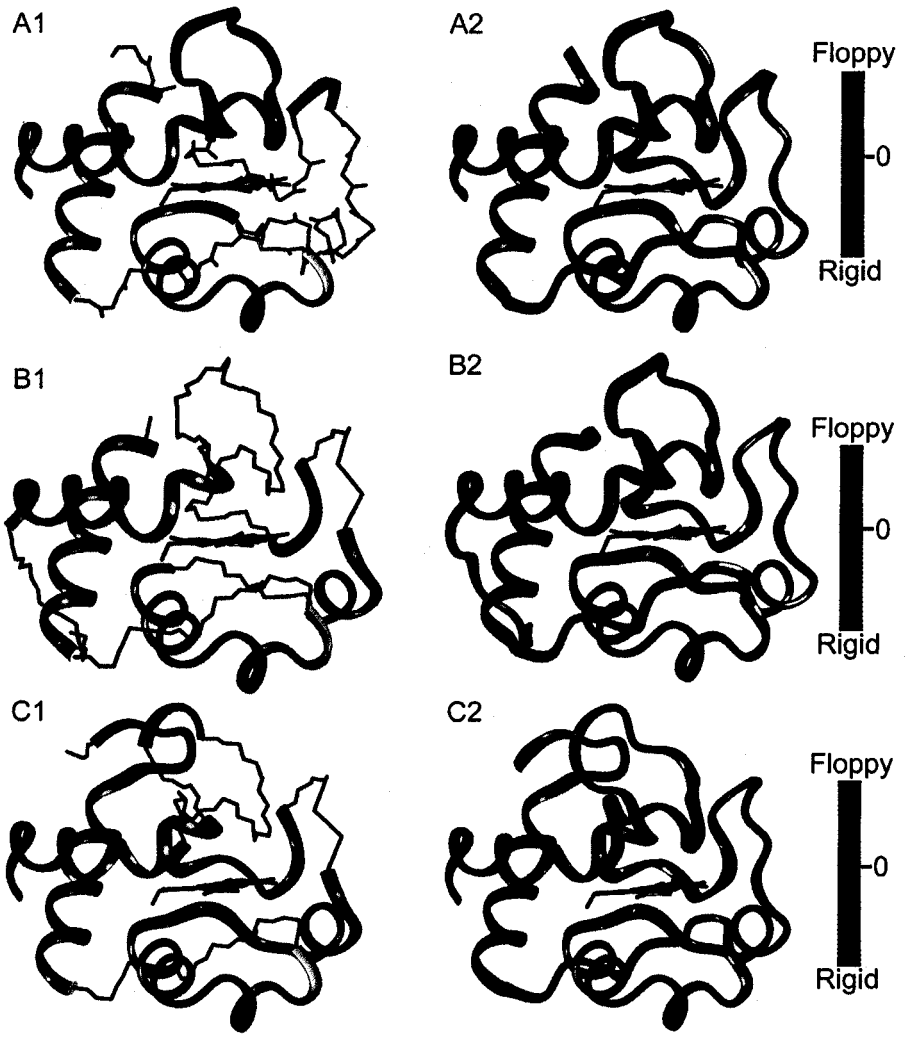

Figure 5: Effect of hydrogen-bond dilution on the flexibility of cytochromes *c* from horse (panel A; PDB code 1hrc), yeast (B, PDB 1ycc), and ***Rhodopseudomonas*** (C, PDB 1co6). The structures in the left column (A1, B1, C1) show the rigid regions (colored ribbons) that result when hydrogen bonds are removed one by one, starting with the weakest, until the tertiary structure fragments into secondary and partial tertiary structures. The flexible regions of the main-chain are colored magenta and shown as thin tubes; the heme appears edge-on at center, in gold tubes. The right column (panels A2, B2 and C2) displays the flexibility index mapped onto the main-chain ribbon at the same point in the hydrogen-bond dilution as the left panels. Note the region of stable (rigid) structure nucleated by the N-terminal and C-terminal helices, at left.

in hydrogen-bond dilution corresponding to fragmentation into substructures (line 6 in Figure 4 for horse, line 5 for yeast, and line 4 for *Rhodopseudomonas*). The left side of each structure is a rigid domain formed by the N-terminal, C-terminal, and 60's helices in contact with the heme (seen edge-on, at center), whereas the right edge of the structure, corresponding to the region in cytochrome *c* that docks with its electron transfer partners [23], is flexible (thin magenta tubes). The right-hand panels in this figure show the relative flexibility index (at the same point in hydrogen-bond dilution) mapped onto a main-chain ribbon for each of the structures. Again, the left edge of each structure is highly rigid (blue), and the flexible region lies at the exposed edge of the heme, with some shift in the region of maximal flexibility in *Rhodopseudomonas* (towards the top edge of panel C2) relative to the other two species.

There are some intriguing differences in the three cytochromes shown in Figure 5. For example the loop at top center is rigid in A2, but flexible with a small rigid inclusion in B2 and C2. The biochemical significance of this, if any, is unclear.

5 Conclusions

Significant insights into protein flexibility can be gained from the analysis of a single protein conformation using *FIRST*. Only a few seconds of computational time are needed and and interatomic potentials are not required. The increase in speed means that flexibility calculations for large systems, which were previously infeasible, can now be accomplished in real time. This method can be used to instantaneously assess changes in protein flexibility due to natural or designed side-chain mutations, substrate or inhibitor binding, and interactions with other molecules, including crystal lattice neighbors and solvent molecules. Using this approach, information about protein (or other molecular) flexibility can be extracted from a single snapshot of the protein structure, which can aid drug design, protein engineering, and decoding protein folding pathways. The application of this technique to cytochromes *c* has allowed the identification of critical hydrogen bonds that stabilize the tertiary structure, as well as aspects of flexibility and stability shared by different species, and differences in flexibility that may account for their different specificities and thermostabilities.

Acknowledgments

We would like to thank Shelagh Ferguson-Miller (Michigan State University) and Duncan McRee (The Scripps Research Institute) for helpful discussions on cytochrome *c* structure and flexibility, and Don Jacobs (Michigan State

University) for discussions on algorithmic aspects of this work. We thank the NSF, NIH and the Center for Protein Structure, Function, and Design at Michigan State University for supporting this interdisciplinary research.

References

1. A. Amadei, A. B. M. Linssen and H. J. C. Berendsen, Proteins: Science, Function, and Genetics, **17**, 412 (1993).
2. J. L. Lagrange, Mécanique Analytique, Paris (1788).
3. J. C. Maxwell, Philos. Mag. **27**, 294 (1864).
4. W. Whiteley, Structural Topology **1**, 46 (1979).
5. D. J. Jacobs, J. Phys. A: Math. Gen. **31**, 6655 (1998).
6. D. J. Jacobs, L. A. Kuhn, M. F. Thorpe, in Rigidity Theory and Applications, M. F. Thorpe and P. M. Duxbury, Eds. (Kluwer/Plenum, New York, 1999), p. 357.
7. D. J. Jacobs, M. F. Thorpe, and L. A. Kuhn, submitted to Physical Review Letters (1999).
8. R. Abagyan, M. Totrov, D. Kuznetsov, J. Comp. Chem. **15**, 488 (1994).
9. T. R. Forrester and W. Smith, J. Comp. Chem., **19**, 102 (1998).
10. B. I. Dahiyat, D. Gordon and S. L. Mayo, Prot. Sci. **6**, 1333 (1997).
11. S. M. Habermann, K. P. Murphy, Protein Science **5**, 1229 (1996).
12. M. F. Thorpe, J. Non-Cryst. Solids, **57**, 355 (1983).
13. Rigidity Theory and Applications, M. F. Thorpe and P. M. Duxbury, Eds. (Kluwer/Plenum, New York, 1999).
14. G. Laman, J. Engrg. Math., **4**, 331 (1970).
15. D. J. Jacobs and M. F. Thorpe, Phys. Rev. Letts. **75**, 4051 (1995).
16. D. J. Jacobs and M. F. Thorpe, Phys. Rev. **E 53**, 3682 (1996).
17. D. J. Jacobs and B. Hendrickson, J. Comp. Phys. **137**, 346 (1997).
18. M.F. Thorpe, D. J. Jacobs, N. V. Chubynsky, A.J. Rader, in Rigidity Theory and Applications, M. F. Thorpe and P. M. Duxbury, Eds. (Kluwer/Plenum, New York, 1999), p. 239.
19. R. W. W. Hooft, C. Sander, G. Vriend, Proteins **26**, 363 (1996).
20. S. J. Hubbard, F. Eisenmenger, J. M. Thornton, Protein Sci. **3**, 757 (1994).
21. Y. Bai, R. R. Sosnick, L. Mayne, S. W. Englander, Science **269**, 192 (1995).
22. O. B. Ptitsyn, in Pacific Symposium on Biocomputing, R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein, K. Lauderdale, Eds. (World Scientific, Singapore, 1999), p. 494.
23. V. A. Roberts, M. E. Pique, J. Biol. Chem., in press (1999).