ELSEVIER

Journal of Molecular Graphics and Modelling 21 (2002) 195–207

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Identifying protein folding cores from the evolution of flexible regions during unfolding

Brandon M. Hespenheide[a,b], A.J. Rader[b,c], M.F. Thorpe[b,c], Leslie A. Kuhn[a,b,*]

[a] *Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA*
[b] *Center for Biological Modeling, Michigan State University, East Lansing, MI 48824, USA*
[c] *Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, USA*

Received 10 February 2002; accepted 16 July 2002

Abstract

The unfolding of a protein can be described as a transition from a predominantly rigid, folded structure to an ensemble of denatured states. During unfolding, the hydrogen bonds and salt bridges break, destabilizing the secondary and tertiary structure. Our previous work shows that the network of covalent bonds, salt bridges, hydrogen bonds, and hydrophobic interactions forms constraints that define which regions of the native protein are flexible or rigid (structurally stable). Here, we test the hypothesis that information about the folding pathway is encoded in the energetic hierarchy of non-covalent interactions in the native-state structure. The incremental thermal denaturation of protein structures is simulated by diluting the network of salt bridges and hydrogen bonds, breaking them one by one, from weakest to strongest. The structurally stable and flexible regions are identified at each step, providing information about the evolution of flexible regions during denaturation. The folding core, or center of structure formation during folding, is predicted as the region formed by two or more secondary structures having the greatest stability against denaturation. For 10 proteins with different architectures, we show that the predicted folding cores from this flexibility/stability analysis are in good agreement with those identified by native-state hydrogen–deuterium exchange experiments.

Keywords: Folding pathways; Flexibility modeling; Thermal denaturation; Graph theory; Structural stability; Hydrogen bond networks; Folding nucleus; Hydrogen-exchange NMR; Barnase; Cytochrome *c*; Apo-myoglobin; Bovine pancreatic trypsin inhibitor; Ribonuclease T1; Chymotrypsin inhibitor 2; Interleukin-1β; T4 lysozyme; Ubiquitin; α-Lactalbumin; FIRST

1. Introduction

Understanding protein folding pathways has been the subject of many recent theoretical and experimental studies [1–8]. These studies often focus on processes that occur early in folding, and models such as nucleation–condensation [9–11] and diffusion–collision [12] have been used to describe the initial step(s). Whether folding is initiated by nucleation of tertiary interactions or diffusion-controlled coalescence of already folded secondary structures is being debated, and a single model may not hold for all proteins. However, a unifying theme is that the initial steps in the folding process involve the interaction of non-local regions in the protein sequence forming a substructure that is substantially preserved in the fully folded protein. Several theoretical techniques have been designed to identify early folding substructures [13–15]. These techniques are unique in that the analysis is performed solely on the native-state conformation, instead of following the folding reaction from a denatured state to the native state. The advantage of analyzing the native state is that this conformation is largely ordered, whereas the denatured state is typically an ensemble of dissimilar, unfolded conformations.

An experimental technique that gives detailed structural information about unfolding is hydrogen–deuterium exchange NMR (H–D exchange). Under native conditions, rotation about main-chain Φ and Ψ dihedral angles leads to fluctuations in which a protein can explore its local conformational space. H–D exchange occurs when the amide and carbonyl groups involved in a hydrogen bond separate enough for deuterated water to intervene, allowing the shared proton to be replaced by a deuteron, or when a buried proton becomes solvent-accessible [16]. Because deuterium does not produce a signal in proton NMR experiments, it is

Abbreviations: H–D exchange, hydrogen–deuterium exchange NMR; CI2, chymotrypsin inhibitor 2; BPTI, bovine pancreatic trypsin inhibitor; PDB, protein data bank; 3D, three-dimensional; 1D, one-dimensional

[*] Corresponding author. Tel.: +1-517-353-8745; fax: +1-517-353-9334.
E-mail address: kuhn@agua.bch.msu.edu (L.A. Kuhn).
URL: http://www.bch.msu.edu/labs/kuhn

1093-3263/02/$ – see front matter © 2002 Elsevier Science Inc. All rights reserved.
PII: S1093-3263(02)00146-8

possible to identify which amide protons undergo hydrogen exchange by comparing the NMR spectra before and after the exchange. By allowing the experiment to run for different time steps, individual exchange rate constants can be assigned to each of the main-chain amide protons identified in the spectra. Woodward has proposed that amide protons that exchange only after long periods of exposure to deuterated water define the slow-exchange core of a protein [17]. Li and Woodward compiled the results from a number of studies on native-state H–D exchange for different proteins, tabulating the residues forming the slow-exchange core in each protein [18]. They have proposed that the secondary structures to which these residues belong define the folding core for the protein. Additionally, they have shown for barnase and chymotrypsin inhibitor 2 (CI2) that the folding core identified by H–D exchange consists of residues with high Φ-values [19], indicating that slow-exchange core residues have significant structure in the folding transition state.

For H–D exchange to occur in main-chain amides involved in hydrogen bonds, flexibility in the protein structure is required to allow access to deuterated water. Given that residues in the folding core have small exchange rates, it is reasonable to assume that the folding core protons either are not accessible to solvent or are in regions that are sufficiently rigid that the hydrogen bond donor and acceptor cannot move apart enough to allow H–D exchange. This can be probed by observing how the flexibility of a protein structure changes as it is gradually denatured.

Our hypothesis is that the folding core is stabilized by a network of particularly dense or strong non-covalent interactions, which tend to resist unfolding or denaturation. Following this hypothesis, we present a novel computational method for predicting the folding core of a protein. This approach employs the floppy inclusions and rigid substructure topography (FIRST) software, which accurately predicts flexible regions in proteins by analyzing the constraints on flexibility formed by the covalent and non-covalent bond network [20–22]. Covalent bonds, salt bridges, hydrogen bonds, and hydrophobic interactions are included in the protein representation. Because thermal denaturation or unfolding involves the breaking of hydrogen bonds and salt bridges, we compare several methods for simulating thermal denaturation, and observe how the removal of these bonds affects the stability and flexibility of the protein. As hydrogen bonds are removed, the protein structure becomes increasingly flexible, and the stable regions decrease in size. The folding core can then be predicted as the most stable region involving at least two secondary structures. The thermal denaturation model in which hydrogen bonds and salt bridges are removed from weakest to strongest predicts folding cores that correlate best with the experimentally observed folding cores. The ability to predict an early state in folding indicates that information about the folding pathway is encoded in the covalent and non-covalent bond network of the native state.

Table 1
Proteins used in this study

Protein name	PDB code	Size (residue)	Structure classification	$\langle r \rangle$ core	Number of H_2O
BPTI	1bpi	58	Few	2.38	4
Ubiquitin	1ubi	76	α–β	2.40	1
CI2	2ci2	83	α–β	2.41	0
Ribonuclease T1	1bu4	104	α–β	2.39	0
Cytochrome *c*	1hrc	104	α	2.39	4
Barnase	1a2p	110	α–β	2.39	5
α-Lactalbumin	1hml	123	α	2.38	4
Apo-myoglobin	1a6m	151	α	2.37	11
Interleukin-1β	1i1b	153	β	2.39	9
T4 lysozyme	3lzm	164	α	2.38	7

The PDB code and number of residues are listed for each protein. The fourth column gives the structure classification for each protein as defined by CATH [43]. The mean coordination, $\langle r \rangle$, of the protein, defined as the average number of covalent and non-covalent bonds per atom in the structure, is listed for each protein at the point in thermal denaturation when the largest structurally stable region is the folding core itself (see Fig. 3). Number of H_2O indicates the number of buried water molecules included in analysis of each protein.

2. Methods

2.1. Preparation of protein structures for analysis

Crystallographic structures for 10 monomeric proteins (Table 1) were selected from the protein data bank (PDB) [23] for analysis. These proteins were chosen based on their diversity of structure and the availability of native-state H–D exchange data for comparison [18]. A 3D structure was not available for apo-myoglobin (which lacks heme), though qualitative data shows its fold is very similar to that of holo-myoglobin (with heme), except for dynamic fluctuations of the *F* helix [24]. As an approximation to the apo-myoglobin structure, we analyzed the holo structure upon removal of its heme group. For this structure, FIRST analysis also found the *F* helix to be one of the two most flexible helices in the protein (data not shown). The experimental results of H–D exchange used for comparison in this study are for apo-myoglobin.

Given the absence of hydrogen atom positions in most X-ray crystal structures, positions for polar hydrogen atoms (including those in bound water molecules) that optimize hydrogen bonding were assigned using the software WHATIF [25]. Only buried water molecules, identified using the PRO_ACT software [26], were included in the subsequent analysis. Each potential hydrogen bond was identified using the following modification of the Mayo potential [27], which evaluates the favorability of the observed hydrogen-bond length relative to the optimal, equilibrium length for that pair of atoms based on their chemistry, as well as the favorability of the angles between the donor and acceptor groups. Our modification strengthens the angular dependence to avoid the inclusion of non-physical H-bonds with angles near 90° (e.g. between C=O(i) and NH($i + 3$), rather than the important C=O(i)–NH($i + 4$) interactions, in the middle of α-helices). Salt bridges were identified between the negatively charged groups of aspartate, glutamate, or the carboxy-terminus of the protein, and the positively charged groups of histidine, lysine, arginine, or the amino-terminus. The energies of hydrogen bonds, E_{HB}, and salt bridges, E_{SB}, were calculated using the following equations:

$$\begin{aligned}
&E_{\mathrm{HB}} = V_0 \left\{ 5\left(\frac{R_0}{R}\right)^{12} - 6\left(\frac{R_0}{R}\right)^{10} \right\} F(\theta,\phi,\gamma),\\
&\mathrm{sp}^3 \text{ donor pairing with } \mathrm{sp}^3 \text{ acceptor} : F(\theta,\phi,\gamma) = \cos^2\theta\, \mathrm{e}^{-(\pi-\theta)^6} \cos^2(\phi - 109.5^\circ),\\
&\mathrm{sp}^3 \text{ donor pairing with } \mathrm{sp}^3 \text{ acceptor} : F(\theta,\phi,\gamma) = \cos^2\theta\, \mathrm{e}^{-(\pi-\theta)^6} \cos^2\phi,\\
&\mathrm{sp}^3 \text{ donor pairing with } \mathrm{sp}^3 \text{ acceptor} : F(\theta,\phi,\gamma) = \cos^4\theta (\mathrm{e}^{-(\pi-\theta)^6})^2,\\
&\mathrm{sp}^2 \text{ donor pairing with } \mathrm{sp}^2 \text{ acceptor} : F(\theta,\phi,\gamma) = \cos^2\theta\, \mathrm{e}^{-(\pi-\theta)^6} \cos^2(\max[\phi,\gamma]),\\
&\text{where } V_0 = 8\,\text{kcal/mol} \quad \text{and} \quad R_0 = 2.8\,\text{Å};\\
&E_{\mathrm{SB}} = V_{\mathrm{S}} \left\{ 5\left(\frac{R_{\mathrm{S}}}{R+a}\right)^{12} - 6\left(\frac{R_{\mathrm{S}}}{R+a}\right)^{10} \right\},\\
&\text{where } V_{\mathrm{S}} = 10\,\text{kcal/mol}, \quad R_{\mathrm{S}} = 3.2\,\text{Å}, \quad \text{and} \quad a = 0.375\,\text{Å}.
\end{aligned} \tag{1}$$

In each equation, R is the distance between the donor and acceptor atoms. The θ angle is the donor-hydrogen-acceptor angle, and ϕ is the hydrogen-acceptor-base atom angle, where the base atom is the atom bonded to the acceptor (e.g. carbonyl carbon for a carbonyl oxygen acceptor atom). The angle γ is an out-of-plane angle that arises when both the donor and acceptor have sp^2 hybridization. For the salt-bridge energy function, also a modification of the Mayo hydrogen-bond potential, the values of V_{S}, R_{S}, and a were selected such that the computed energies matched those of experimental results on salt bridges [28]. No angular term between donor and acceptor is included for salt bridges because of their significant Coulombic component, which is only dependent on distance.

Hydrogen bonds and salt bridges were included in the flexibility analysis if their energies were less than (more favorable than) −0.1 kcal/mol. Because salt bridges are essentially a special case of hydrogen bonds in which the donor and acceptor are charged, for simplicity, we will refer to both hydrogen bonds and salt bridges as hydrogen bonds.

Pairs of carbon and/or sulfur atoms in the protein and ligands were considered to make hydrophobic contacts if their van der Waals surfaces were within 0.25 Å, using van der Waals radii of 1.7 and 1.8 Å for carbon and sulfur atoms, respectively [29]. Because the FIRST program models interactions as inter-atomic constraints, hydrophobic interactions [30] were modeled as flexible tethers [31] in order to constrain the protein structure less than hydrogen bonds do. This representation of hydrophobic interactions allows the two atoms forming a hydrophobic interaction to slip relative to one another, while remaining close enough (with their van der Waals surfaces constrained to stay within 0.25 Å) that water molecules cannot intervene. Bonds between the protein and any ligands, including metals and other ions, were treated as covalent bonds if so specified in the PDB file, or if they were within covalent bonding distance of the protein; otherwise, they were subject to the above rules for identifying hydrogen bonds and hydrophobic interactions.

2.2. Flexibility analysis

The structural flexibility of a protein structure is a property that depends upon how the motion of each atom is restricted by bond forces. In the absence of any bond forces, each atom has 3 degrees of freedom associated with motion in three dimensions. To calculate the flexible regions in a protein, it is necessary to accurately identify which bond forces remove degrees of freedom from the system by restricting the motion between atoms. The strongest of these bond forces are the covalent bonds. In the absence of non-covalent forces, the single bonds in a protein could rotate about any dihedral angle that did not result in steric overlap. The protein would be free to adopt a large number of conformations with comparable energies. Thus, the non-covalent forces largely define the secondary, tertiary, and quaternary structure observed in proteins. The non-covalent interactions, such as hydrogen bonds and hydrophobic interactions, impose constraints on bond rotation that can be observed by identifying the stable and flexible regions in a protein structure. We use the software FIRST to represent the covalent and non-covalent constraints present in a protein and to compute the resulting flexibility of the main chain and side chains [20,22].

Because we are interested in macroscopically significant flexibility, rather than the high-frequency fluctuations associated with thermal motion, bond lengths and angles are assigned their equilibrium values as observed in the protein crystal structure. These fixed bonds lengths and angles give rise to distance constraints between pairs of atoms in the protein, either explicitly from chemical bonds or implicitly

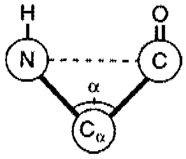


Fig. 1. Example of a distance constraint that arises due to a fixed bond angle, α. The positions of the N, C_α, and C atoms are crystallographically defined, and the sp^3 hybridization of the C_α atom defines the bond angle α. Because this angle is held constant, the distance between the N and the C atoms, shown as a gray dashed line, is also fixed.

from other local bond lengths and angles. For example, each of the covalent bonds between adjacent N, C_α, and C atoms in the backbone has a constant bond length and forms a constant bond angle, α (Fig. 1). This fixes the distance between the second nearest neighbor N and C atoms, shown as a dashed gray line in Fig. 1. All such fixed bond angles can be represented by the associated distance constraints. In this manner, we identify all the distance constraints that arise due to covalent bonds and angles, and add constraints for non-rotatable peptide and other double or partial double bonds, as well as those arising from salt bridges, hydrogen bonds, and hydrophobic interactions [31], as described earlier.

FIRST uses 3D constraint counting [22] on this network of distance constraints to identify the flexible and rigid (structurally stable) regions within a protein. This graph-theory algorithm [20,32] for analyzing proteins is a 3D 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 the introduction of constraints on the motion of mechanical systems during the late 18th century by Lagrange [33]. Maxwell [34] used this approach during the late 19th century to determine whether structures were stable or deformable. Traditional applications have been to problems in engineering, such as determining the structural stability of different truss configurations in bridges. A very significant advance occurred with Laman's theorem [35] in 1970, which precisely determines the degrees of freedom within two-dimensional networks, and allows the rigid regions and flexible joints between them to be found. The details of the extension of this approach to 3D systems, including the FIRST software developed to analyze proteins, are presented in [20]. The results of FIRST native-state flexibility analysis have been shown to compare well with experimental definitions of flexible regions in a series of proteins including lysine–arginine–ornithine binding protein [20], cytochrome *c* [21], HIV protease, adenylate kinase, and dihydrofolate reductase [22]. (The FIRST software is available to interested academic and commercial researchers; see http://first.pa.msu.edu.)

The results of FIRST indicate for each bond in the protein whether it is flexible (free to rotate) or rigid (not rotatable) due to the covalent and non-covalent constraints within the structure. Groups of atoms coupled to each other via rigid bonds form a rigid cluster. One or more independent rigid clusters with intervening flexible regions may occur in a protein structure. The distribution of rigid clusters and flexible bonds identified by FIRST can be viewed graphically by color-mapping this information onto the 3D structure of the protein, as shown in the top part of Fig. 2 (for clarity, the side chains are not shown). Flexible (rotatable) bonds are shown as thin black lines, while rigid bonds are depicted by thick, colored tubes, with each independently rigid cluster distinguished by a different color.

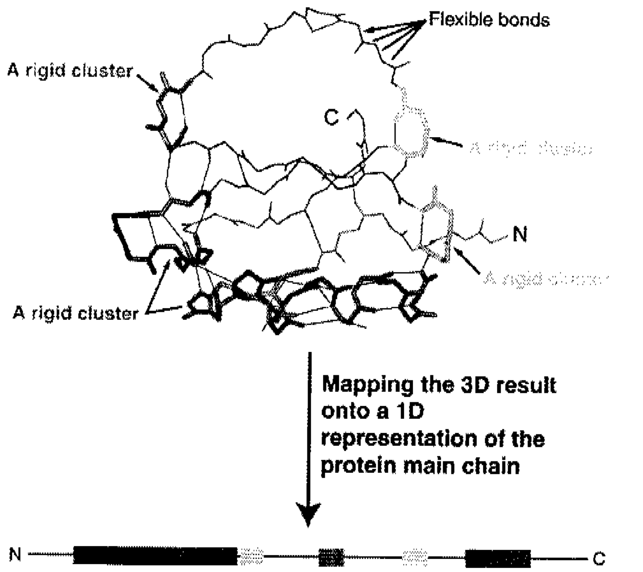


Fig. 2. Displaying the results of FIRST flexibility analysis by mapping the rigid and flexible regions in the 3D structure onto a 1D representation, from N- to C-terminus. The structure of chymotrypsin inhibitor 2 is shown above, with FIRST-defined flexible bonds shown as thin black lines and independently rigid regions shown as differently colored thick tubes. For simplicity, only the rigid and flexible regions of the main chain are shown, but the side chains were also included in the analysis. To create a one-dimensional summary of the results, the same coloring of lines/tubes is then mapped onto a line representing the main chain from N- to C-terminus, shown below.

2.3. Simulating denaturation

As a protein is gradually denatured, the covalent bonds remain intact, whereas hydrogen bonds begin to break. The flexibility in the protein will increase as the number of hydrogen bonds in the protein decreases. Our hypothesis is that the folding core is the region that will remain structurally stable the longest under denaturing conditions. This hypothesis was tested by incrementally removing hydrogen bonds from a protein structure to simulate thermal denaturation, then using FIRST to observe the evolution of flexible regions in the structure. The results depend upon the order in which hydrogen bonds are removed. Because hydrophobic interactions actually become somewhat stronger with moderate temperature increases [30], these interactions are maintained throughout the simulation. Three methods for diluting the hydrogen bond network of a protein are presented, each designed to test the importance of the strength and/or density of the hydrogen bonds when selecting which bond to remove next.

2.3.1. Thermal denaturation

As the temperature of a protein is gradually increased, the hydrogen bonds are expected to break in an energy-dependent manner. We mimic this process by using the following procedure. Initially, the flexibility of the native protein structure is analyzed with all its covalent and non-covalent interactions (hydrogen bonds and hydrophobic interactions). The weakest hydrogen bond in the structure is then broken by removing any constraints created by that bond [31]. The effect of removing this bond is then observed by applying FIRST to identify the flexible regions in the protein. We continue this process of breaking the weakest hydrogen bond remaining in the structure and updating the identification of flexible regions until all the hydrogen bonds have removed.

2.3.2. Random removal of non-covalent bonds over a small energy window

The thermal denaturation scheme above removes hydrogen bonds strictly in order of energy. To introduce some noise into the method, reflecting the stochastic nature of thermal denaturation and testing the effect of inaccuracies in the hydrogen-bond energy function, the next hydrogen bond to be removed is randomly selected from the 10 weakest bonds remaining in the protein. This method was developed to test whether the small fluctuations expected to occur during thermal denaturation will influence the flexibility or folding core predictions.

2.3.3. Completely random removal of non-covalent bonds

To check whether the relative energies of hydrogen bonds, and not just their density in the structure, are indeed important in thermal denaturation, we have also performed completely random dilutions of the hydrogen bonds in the network, without respect to their energies. In this case, the next hydrogen bond to be removed from the protein is selected randomly from all remaining hydrogen bonds.

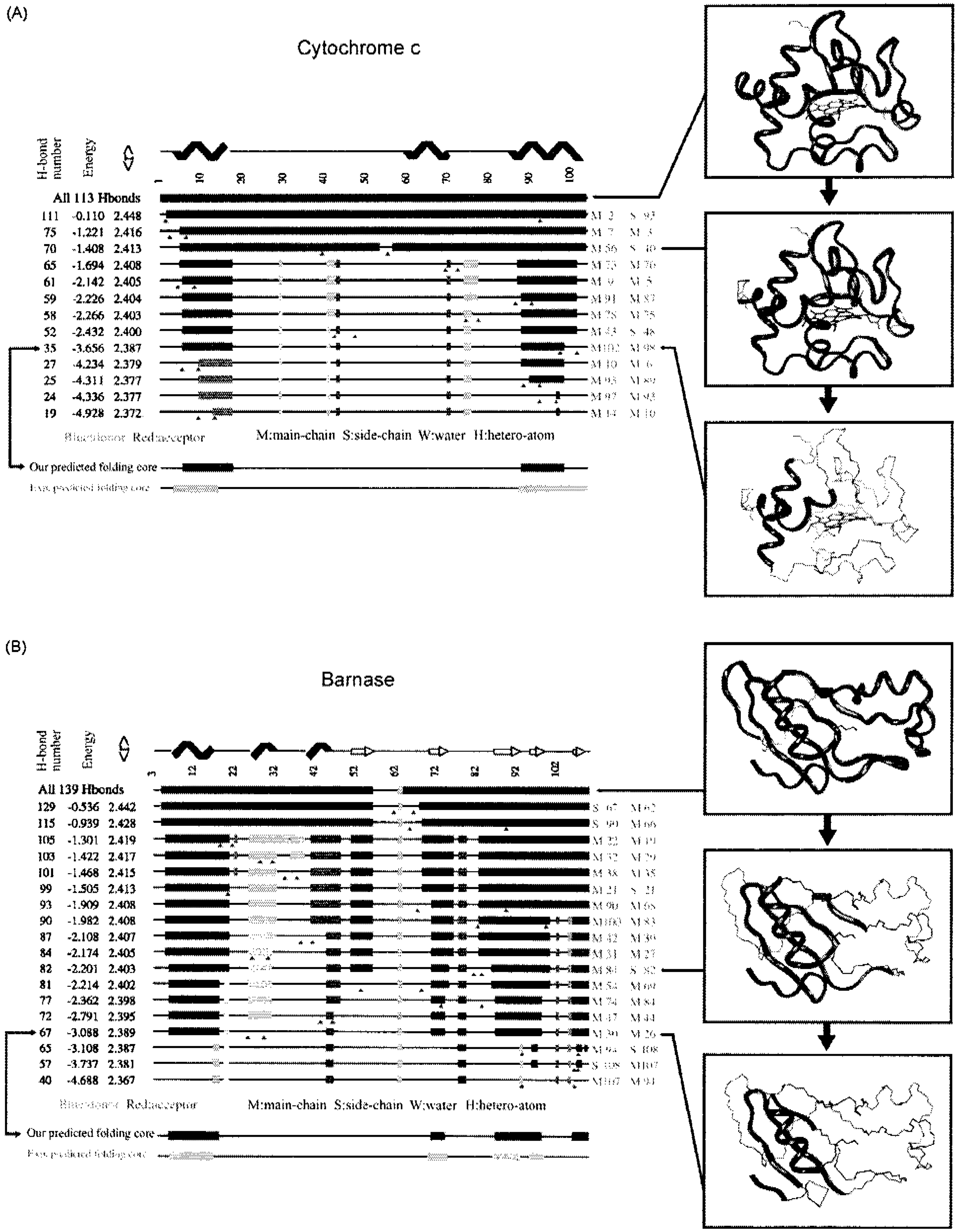
2.4. Visualizing results

Due to the difficulty in comparing flexibility results mapped onto 3D structures for a large number of steps in the denaturation of a protein, we employ the reduced one-dimensional (1D) representation shown at the bottom of Fig. 2. The calculation used for this result includes side-chain and ligand atoms, as well as main-chain atoms. As in the 3D figure, each backbone bond is represented as a thin black line if it is flexible (rotatable), or as a colored tube if it is rigid. A single rigid region (represented by a single color) may consist of non-contiguous regions of the sequence, as demonstrated by the red rigid region in Fig. 2. The complete denaturation can now be viewed as a series of horizontal lines (Fig. 3), ordered from native state (top) to a substantially flexible, or denatured state (bottom). Each line shows the current regions of structural stability and flexibility for the backbone atoms after a step in the denaturation process. Frequently, several successive lines are identical because the flexibility of the backbone has not been affected by the changes in the non-covalent bond network. These redundant lines are omitted, and only those steps that result in a change in backbone flexibility are displayed. Fig. 3A provides an example of a complete thermal denaturation simulation for cytochrome *c*. The three columns on the left-hand side describe: the number of remaining hydrogen bonds in the protein at each step; the energy of the just-broken bond (in kcal/mol), according to the modified Mayo potential; and the mean coordination, $\langle r \rangle$, of the atoms in the network at that step, counted as the number of covalent bonds, hydrogen bonds, salt bridges, and hydrophobic interactions per atom, averaged over all atoms in the protein [31]. The mean coordination decreases along the unfolding pathway and is a structure-based variable that can be usefully regarded as a folding/unfolding reaction coordinate. Regular secondary structure content is shown at the top, as determined by DSSP [36]. The right-hand columns, together with the solid triangles beneath each line, show the residue locations of the donor (blue) and acceptor (red) atoms of the hydrogen bond or salt bridge broken to generate this step. For instance, "M2" indicates the main chain of residue 2, "S93" indicates the side chain of residue 93, and "W120" indicates water molecule 120 in the PDB structure. "H" indicates other heteroatoms, belonging to non-protein functional groups such as bound heme.

2.5. Identifying the folding core

Generally, in the native state, most of the residues belonging to an α-helix or β-strand are rigid, and the secondary structures are mutually rigid, or approximately rigid. As the hydrogen bonds are removed from the protein, parts of the secondary structures may become flexible, particularly the ends of helices and strands. Also, the secondary structures tend to become independently rigid at intermediate steps in denaturation, due to loss of inter- and intra-secondary structure bonds.

The protein folding core is defined in this study as the set of secondary structures that remain mutually rigid the longest in the simulated denaturation. The secondary structures for the native states of each of the 10 proteins were identified by using DSSP and tracked during the unfolding simulation. Not all residues in the secondary structure are required to be rigid when identifying the folding core. An α-helix is considered to be rigid if at least five consecutive residues, corresponding to one complete turn of an α-helix, belong to the rigid cluster. If a helix is defined by DSSP to contain less than five residues, as can occur with 3_{10} helices, all its residues must be mutually rigid to be considered a


(A)
Cytochrome c
H-bond number
Energy
⟨r⟩
1
10
20
30
40
50
60
70
80
90
100
All 113 Hbonds
111 -0.110 2.448 M 2 S 93
75 -1.221 2.416 M 7 M 3
70 -1.408 2.413 M 56 S 40
65 -1.694 2.408 M 73 M 76
61 -2.142 2.405 M 9 M 5
59 -2.226 2.404 M 91 M 87
58 -2.266 2.403 M 78 M 75
52 -2.432 2.400 M 43 S 48
35 -3.656 2.387 M 102 M 98
27 -4.234 2.379 M 10 M 6
25 -4.311 2.377 M 93 M 89
24 -4.336 2.377 M 97 M 93
19 -4.928 2.372 M 14 M 10
Blue:donor Red:acceptor
M:main-chain S:side-chain W:water H:hetero-atom
Our predicted folding core
Exp. predicted folding core
(B)
Barnase
H-bond number
Energy
⟨r⟩
3
12
22
32
42
52
62
72
82
92
102
All 139 Hbonds
129 -0.536 2.442 S 67 M 62
115 -0.939 2.428 S 90 M 66
105 -1.301 2.419 M 22 M 19
103 -1.422 2.417 M 32 M 29
101 -1.468 2.415 M 38 M 35
99 -1.505 2.413 M 21 S 21
93 -1.909 2.408 M 90 M 68
90 -1.982 2.408 M 100 M 83
87 -2.108 2.407 M 42 M 39
84 -2.174 2.405 M 31 M 27
82 -2.201 2.403 M 84 S 82
81 -2.214 2.402 M 54 M 69
77 -2.362 2.398 M 74 M 84
72 -2.791 2.395 M 47 M 44
67 -3.088 2.389 M 30 M 26
65 -3.108 2.387 M 94 S 108
57 -3.737 2.381 S 108 M 107
40 -4.688 2.367 M 107 M 93
Blue:donor Red:acceptor
M:main-chain S:side-chain W:water H:hetero-atom
Our predicted folding core
Exp. predicted folding core

rigid secondary structure. The β-strands are required to have at least three consecutive residues rigid to be considered as part of the folding core. This criterion of three consecutive rigid residues allows for at least two hydrogen bonds to an adjacent strand. If a strand defined by DSSP consists of less than three residues, the entire strand is required to be rigid to be counted as part of the folding core.

3. Results

3.1. Thermal denaturation to probe unfolding pathways and folding cores

For cytochrome *c*, the native state is composed of a single, structurally stable region represented by the top line in Fig. 3A, and the 3D structure is shown at the right. When hydrogen bonds 113 through 65 (the weakest 49) were removed, the large rigid cluster (colored red) significantly decreased in size (at the fifth line in panel A), resulting in new flexibility in those residues between the N- and C-terminal helices. These helices formed the only significantly rigid region in the protein. The folding core was predicted as the last point in the denaturation when at least two secondary structures formed a single rigid region. This point in cytochrome *c* occurred in the fifth-to-last line, where the N- and C-terminal helices remained mutually rigid. On the next line, no single rigid cluster contained more than one secondary structure. The predicted folding core is shown structurally at bottom right in panel A, and summarized in a 1D representation just below the denaturation results, along with the folding core determined by H–D exchange [18,37]. The predicted and observed folding cores correspond well, both indicating that the N- and C-terminal helices together form the stable folding core.

The detailed unfolding pathway and folding core predictions upon thermal denaturation are shown for barnase in Fig. 3B. There was a significant change in the flexibility of the protein observed after 35 hydrogen bonds had been removed (line 4), resulting in several small rigid regions that could move independently of one another (as indicated by their different colors in the plot), and one large rigid region (shown in red). Our study of folding transition states [31] has shown that the rigid core of proteins disintegrates into several independent rigid regions when the mean atomic coordination decreases below ~2.415. This is seen for both barnase and cytochrome *c* in Fig. 3, yielding a transition between rigid and flexible states that is also found for network glasses near the same mean coordination [38]. An intermediate structural state in barnase is formed by the packing of an α-helix against the β-sheet (second structural panel at right in Fig. 3B). The β-sheet in this super-secondary structure partially denatures to form the folding core itself, consisting of the α-helix packed against part of the β-sheet (fourth line from bottom in Fig. 3B, with 3D structure shown in the last panel at right). The H–D exchange folding core, shown at bottom (orange), matches the predicted folding core (red) well, with the exception of the short, C-terminal β-sheet.

Fig. 4 shows the unfolding pathway for interleukin-1β, a protein consisting of only β-strands. The structure shows little breakup during the initial steps of the unfolding simulation. A significant event occurred when hydrogen bond 106 was broken, resulting in flexibility for a large portion of the structure. The β-strands formed by residues between 50 and 135 remain rigid, and form the folding core on the fourth line from the bottom. A comparison to the experimental folding core, shown at the bottom in orange, shows significant similarity.

The hydrogen bond dilution results for BPTI are shown in Fig. 5. BPTI is a member of the DSSP secondary structure class "few" due to its small size and few secondary structures, and its disulfide bonds were included as part of the covalent bond network. The steps in the unfolding pathway represented in Fig. 5 show a gradual breakup of the structure into small rigid regions linked by flexible hinges. The N-terminal helix becomes flexible when hydrogen bond 29 is broken, followed by the C-terminal helix when hydrogen bond 15 is broken. The remaining two secondary structures (β-strands between residues 15 and 35) remain mutually rigid, along with residues 45 and 51, to form the predicted folding core of BPTI. Again, the predicted and experimentally determined folding cores correspond closely.

Thermal denaturation simulations were performed to predict the folding core for each protein in our dataset. Fig. 6 summarizes the folding core predictions from these simulations, comparing the predicted folding core to that observed experimentally. For a majority of the proteins (8 out of 10), the folding core predictions agree well with

Fig. 3. Results of thermal denaturation, in order of hydrogen-bond energy, for cytochrome *c* and barnase. (A) Cytochrome *c*: this figure shows how the structure fragments into smaller rigid regions, with intervening flexible linkers, as the hydrogen bond network denatures with increased thermal energy. Alpha helices within the native structure are indicated as red zigzags at the top. Shown at right is the 3D representation of the largest stable region (colored red) in the protein for the native state (top), an intermediate, state (middle), and the folding core (bottom), defined as the last point in denaturation at which the largest rigid region consists of more than one secondary structure. The summary of the folding core prediction at the bottom of panel A indicates that there is close correspondence between the prediction of the folding core as the most stable supersecondary region and the folding core as defined by protection from H–D exchange [37]. (B) Barnase: the native-state secondary structure for barnase is shown at the top (red zigzags indicate α-helical structure and yellow arrows represent β-strands). The three figures on the right show the location of the largest rigid cluster in the protein at that step. The folding core is predicted at the fourth-to-last line, and includes the N-terminal helix and the four C-terminal strands. This predicted folding core overlaps well with the observed folding core from H–D exchange experiments [44], shown in orange at the bottom of the figure.

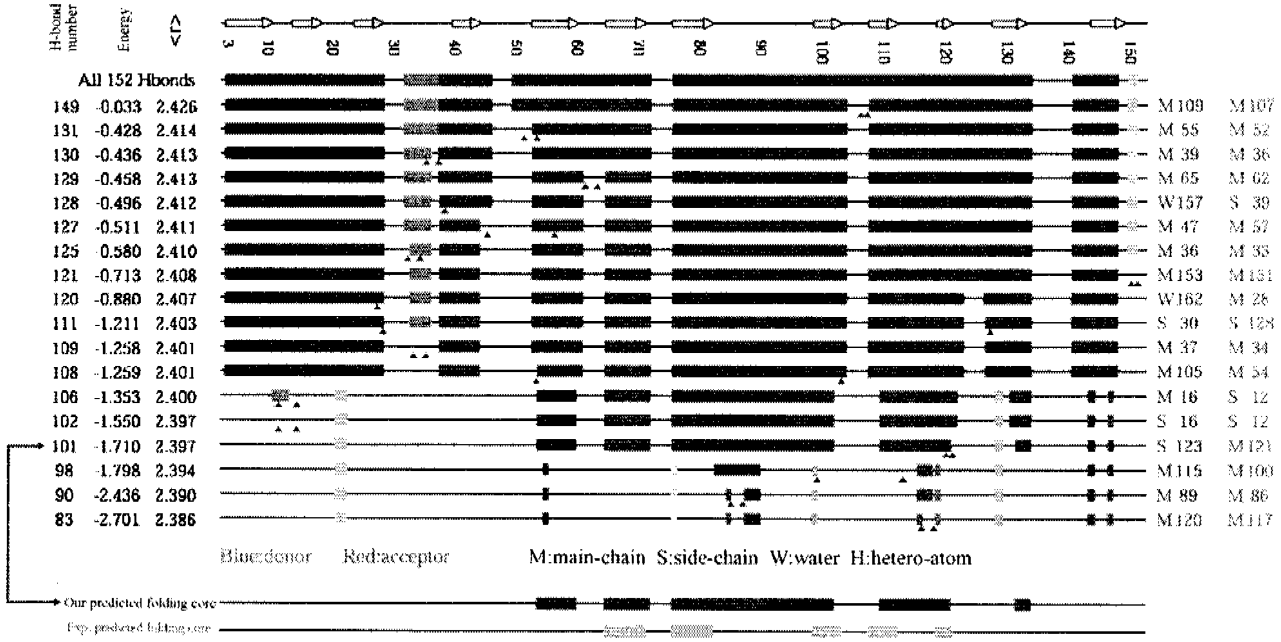


Fig. 4. Thermal denaturation results for interleukin-1β. The native-state secondary structure for interleukin-1β is shown at the top (yellow arrows represent β-strands). The experimental folding core is composed of β-sheet formed by strands 6–10. The main-chain rigidity of the predicted folding core is shown on the fourth line from the bottom. The predicted folding core (summarized in red at bottom) includes strands 6–10, and also portions of strands 5 and 11, and is compared with H–D exchange folding core [49] (shown in orange).

folding cores predicted by regions of slow H–D exchange, and often involve tertiary interactions between sequence-distant secondary structures. For α-lactalbumin, half of the folding core region is in agreement, and for T4 lysozyme, the folding core identified by experiment is much larger than that identified by flexibility analysis. Given that different experimental conditions can also produce different results, we are consulting a broader range of experimental probes of T4 lysozyme folding, as well as doing further structural analysis.

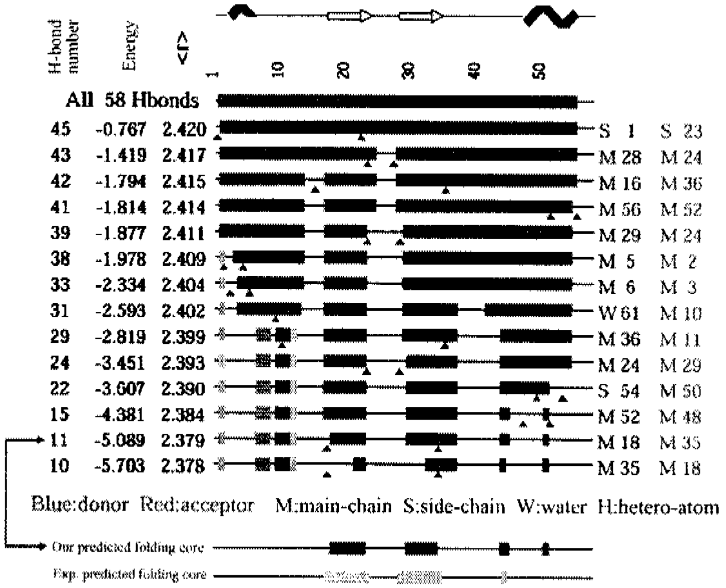


Fig. 5. Thermal denaturation results for BPTI. The native-state secondary structure for BPTI is shown at the top (red zigzags indicate α-helical structure and yellow arrows represent β-strands). The secondary structure of BPTI consists of a 3_{10}-helix near the N-terminus, two β-strands, and an α-helix near the C-terminus. The bottom of the figure compares the predicted folding core, in red, to the experimentally identified folding core [46], in orange. The comparison shows significant overlap between the two results, both of which identify the two β-strands and residue 45 as belonging to the folding core.

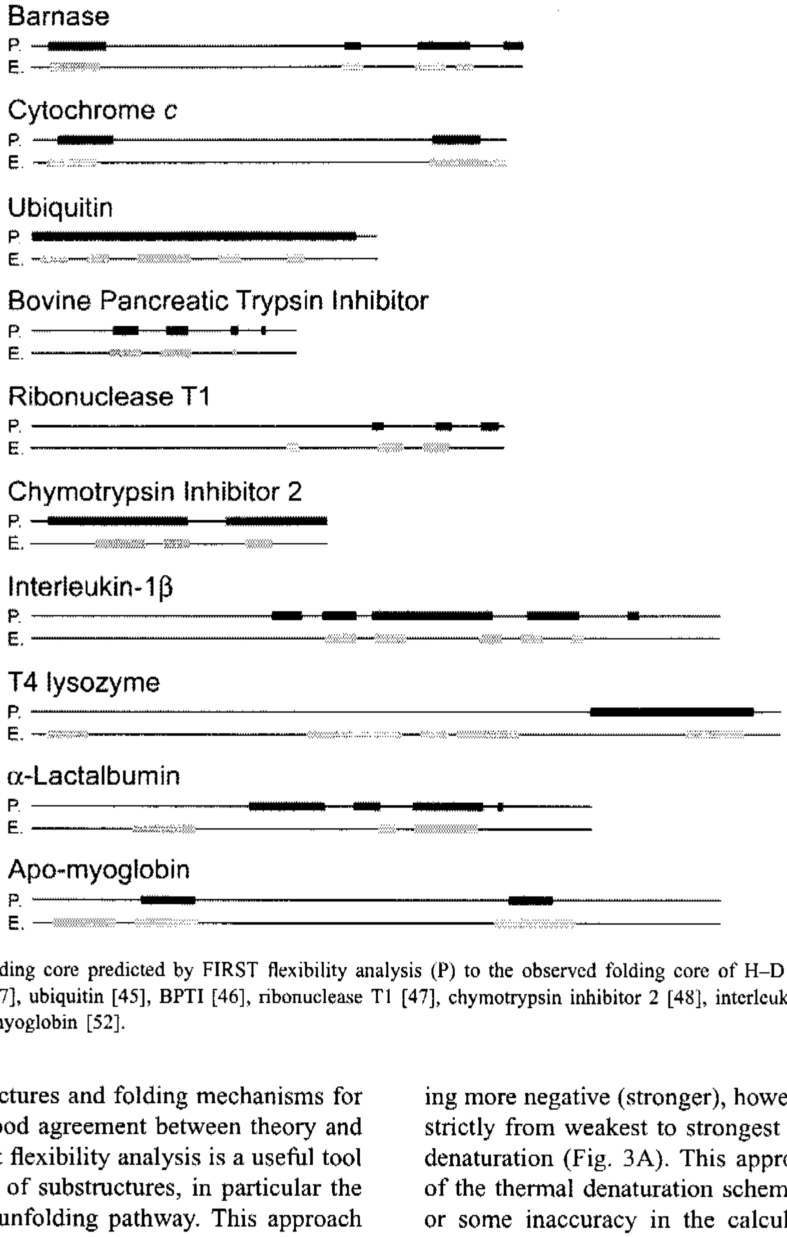


Fig. 6. Comparison of the folding core predicted by FIRST flexibility analysis (P) to the observed folding core of H–D exchange experiments (E) for barnase [44], cytochrome *c* [37], ubiquitin [45], BPTI [46], ribonuclease T1 [47], chymotrypsin inhibitor 2 [48], interleukin-1β [49], T4 lysozyme [50], α-lactalbumin [51], and apo-myoglobin [52].

Given the diverse structures and folding mechanisms for these 10 proteins, the good agreement between theory and experiment indicates that flexibility analysis is a useful tool for probing the stability of substructures, in particular the folding core, along the unfolding pathway. This approach provides explicit 3D structural maps of the stable regions predicted in the protein at each step during denaturation, as well as providing a model for the interactions important in stabilizing folding cores: a dense network of hydrogen-bond interactions that augment the ubiquitous, but less specific, hydrophobic interactions.

3.2. Evaluating other models of denaturation

3.2.1. Random removal of non-covalent bonds over a small energy window

Fig. 7 shows the result of simulating cytochrome *c* denaturation by removing a hydrogen bond randomly from the 10 lowest-energy bonds remaining in the protein at each step. It can be seen in the second column on the left that the energies of the bonds being removed are generally becoming more negative (stronger), however they are not removed strictly from weakest to strongest energy as in the thermal denaturation (Fig. 3A). This approach tests the robustness of the thermal denaturation scheme to thermal fluctuations or some inaccuracy in the calculation of hydrogen-bond energies. Comparing Fig. 3A and Fig. 7 show that introducing some randomness into the thermal denaturation has little effect on accurate prediction of the folding core for cytochrome *c*, and mainly predicts a more rigid unfolding intermediate state between −1 and −2.3 kcal/mol. Twenty separate runs were performed with different random selection of the hydrogen bonds to be removed from the 10 lowest-energy hydrogen bonds (data not shown), and all runs predicted the same folding core.

3.2.2. Completely random removal of non-covalent bonds

As an extreme example of a random dilution, we simulated denaturation in which the hydrogen bond energies were not taken into account. Each hydrogen bond was weighted equally, and the next bond to be removed was chosen randomly from all hydrogen bonds remaining in the protein.

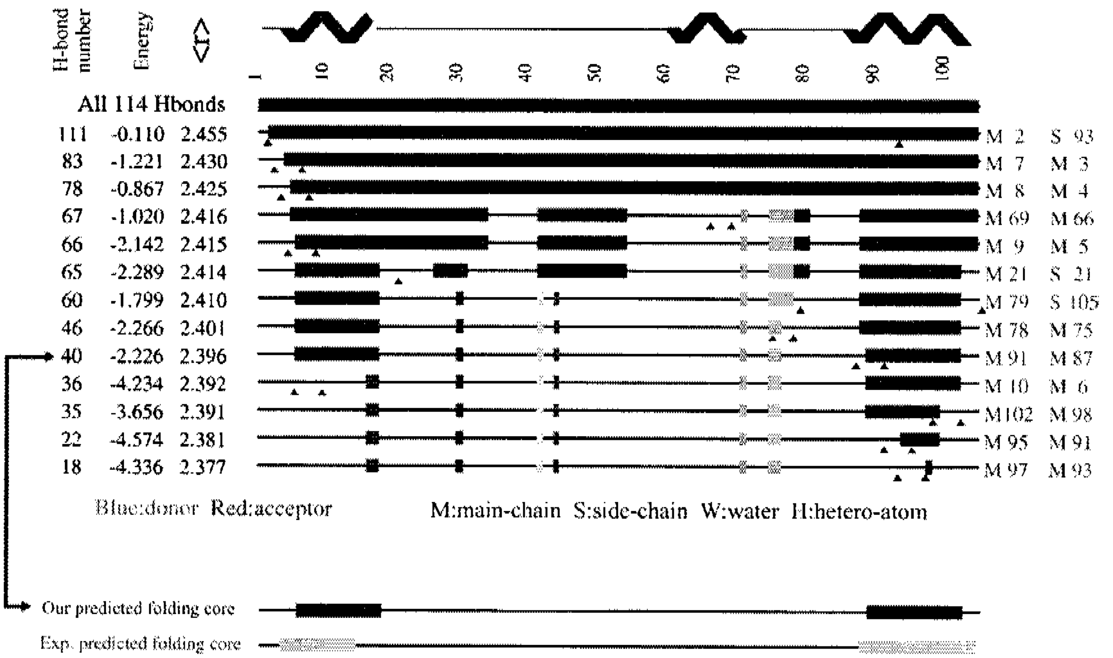


Fig. 7. Results of random hydrogen bond dilution within a window of hydrogen-bond energies for cytochrome *c*. Denaturation is simulated by removing hydrogen bonds as in the thermal denaturation scheme. However, instead of always removing the weakest hydrogen bond in the protein, a hydrogen bond is randomly selected from the 10 weakest hydrogen bonds remaining in the protein. Beneath the figure, the predicted folding core (red) is compared to the observed folding core (orange). The similarity in folding core prediction with that of the thermal denaturation simulation in Fig. 3A indicates that the results of thermal denaturation are robust, in that they are insensitive to small thermal fluctuations or inaccuracies in the hydrogen-bond energy function.

If the folding core of a protein could be identified solely by having the highest density of covalent bonds, hydrogen bonds and hydrophobic interactions, regardless of their energies, then the results of this approach would be accurate. Four separate, random denaturation simulations for cytochrome *c* are shown in Fig. 8. Below each panel, a comparison between the folding core predicted from this simulation and the experimentally observed folding core is shown. Panel C in Fig. 8 shows that a completely random simulation can, by chance, produce a correct folding core prediction and have similar intermediate features to thermal denaturation according to hydrogen-bond energy (compare with Fig. 3A). However, the other panels in Fig. 8 indicate that a random hydrogen bond removal scheme most commonly mispredicts the folding core. Thus, the energy of hydrogen bonds is a significant factor in simulating the denaturation and unfolding of proteins, as validated by folding core prediction.

4. Discussion

Several theoretical techniques have been developed to probe protein folding pathways through an analysis of the native state [13,15,39–41]. Galzitskaya and Finkelstein have developed a technique to computationally analyze the energetics of all possible substructures in the native-state conformation and define a subset of these structures as the transition state ensemble. Computed Φ-values, which measure the similarity between transition-state structure and native-state structure for a given residue, from their ensemble show good correlation to experimentally determined values [13]. Hilser et al. partition the protein into blocks along the sequence, then generate alternative partitions by shifting these blocks [15]. The blocks are then kept folded or unfolded in all possible combinations to generate an ensemble of states. Folding cooperativity between one residue and all other residues in the protein is assessed by performing an energy-perturbing mutation of the residue, in all its occurrences within folded states, then observing the effects on all other residues. An alternative approach is that of Tsai et al. [40], in which the native-state structure is also partitioned, first into domains (visually), then into potential hydrophobic folding units based upon a scoring function measuring compactness, degree of isolation, and hydrophobicity. A combinatorial approach is then used to reassemble possible folded states from these folding units. Similarly, Wallqvist et al. partition the structure by using a sequence mask, and assess pairwise and higher-order interactions in a unified-atom representation of the protein by using a knowledge-based folding potential [41]. Essentially, all these approaches exhaustively partition the structure into substructures, and use a potential or scoring function to assess the interactions between substructures as potential intermediate states in folding.

More recently, Vendruscolo et al. [8] and Dokholyan et al. [42] probed the transition-state ensembles of small

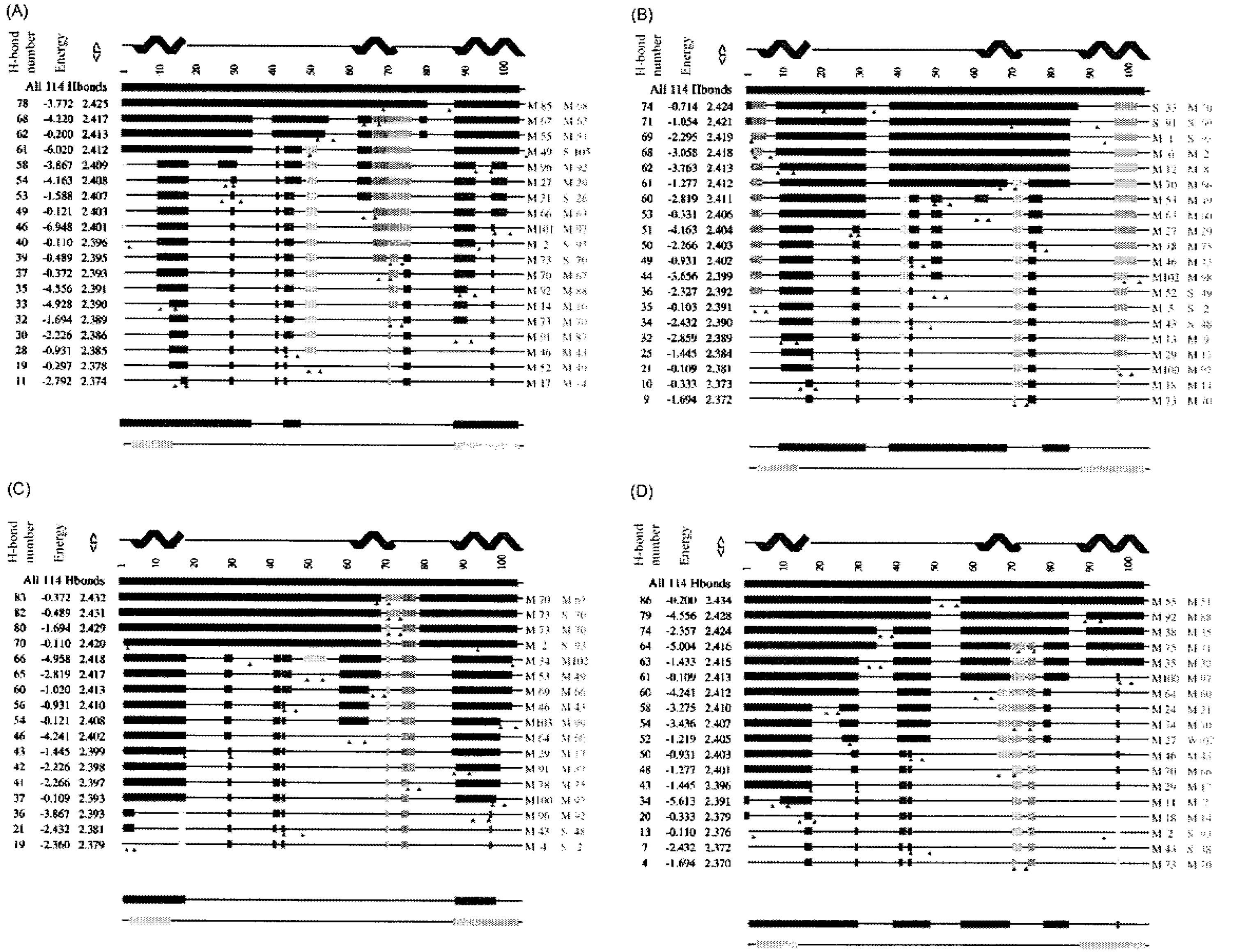


Fig. 8. Four random dilutions of the hydrogen bonds and salt bridges in cytochrome *c*. Each panel represents a single unfolding simulation in which the hydrogen bonds were removed in random order. The DSSP [36] assigned secondary structures are shown at the top of each panel (the red zigzags represent α-helical structures, and the straight lines represent irregular secondary structures). The predicted folding core is shown at the bottom of each panel (in red) and compared to the folding core observed by NMR (shown in orange). The panel at the lower left shows that an accurate folding core prediction can by chance be obtained completely random hydrogen bond removal. However, the results in the other three panels are in poor agreement with the observed folding core. Thus, dilution of the hydrogen-bond network in order of hydrogen-bond strength (Fig. 3A) is more appropriate than random breakage of hydrogen bonds in representing the thermal denaturation of proteins.

proteins for residues important in forming the transition state, and represented the results in terms of networks of interactions between residues. In particular, Dokholyan et al. identify three residues, A16, L49, and I57 that have experimentally been shown to be important for forming the folding nucleus in CI2 [42]. This agrees with our results on CI2, as residues A16, L49 and I57 are predicted to be part of the folding core. A difference between these methods is that the FIRST approach directly predicts from the native state which residues contribute to the folding core, and does not require an ensemble of near-transition-state conformers for the analysis. FIRST also predicts which residues are mutually rigid or flexible from the complete network of interactions, rather than focusing on the number of interactions with neighboring residues.

The FIRST approach, coupled with thermal dilution of the hydrogen-bond network, also has the goal of identifying structurally stable states along the unfolding/folding pathway, and does so by decoding the hierarchy of stable substructures within the native state. The FIRST program treats a protein structure as a network of atoms and bonds, and the analysis decomposes the structure into rigid regions and flexible regions. Given that the experimentally identified folding core represents a region of structure that resists unfolding, we have used FIRST to identify the region of structure that resists becoming flexible as we simulate unfolding. The good correlation between the predicted and experimental folding cores shown in Figs. 3–6 supports the hypothesis that the native-state structure of a protein, specifically the distribution and strength of the non-covalent forces, encodes information about the folding pathway.

The power of FIRST flexibility analysis lies in its simplicity and computational speed; all steps in the thermal denaturation of a large protein can be calculated in a minute on a personal computer. FIRST, combined with thermal denaturation of the non-covalent bond network, also provides an explicit structural description of which regions of the protein are flexible or structurally stable at each step along the unfolding pathway. Using this approach, the phase transition from the folded state to unfolded [31] can be tracked structurally as rigidity in the protein is lost, and, as shown here, the folding cores can be identified and prove to be in good agreement with experimental results.

Acknowledgements

We would like to thank Clare Woodward for valuable discussions during the early stages of this work. We also thank the MSU Center for Protein Structure, Function, and Design, the MSU Center for Biological Modeling, and the Barnett C. Rosenberg Fellowship for support to B.M.H. and A.J.R., and the NSF for support under Grants DBI-9600831 and DMR-0078361.

References

[1] J.N. Onuchic, Z. Luthey-Schulten, P.G. Wolynes, Theory of protein folding: the energy landscape perspective, Ann. Rev. Phys. Chem. 48 (1997) 545–600.

[2] M. Gruebele, The fast protein folding problem, Ann. Rev. Phys. Chem. 50 (1999) 485–516.

[3] J.-E. Shea, C.L. Brooks III, From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding, Ann. Rev. Phys. Chem. 52 (2001) 499–535.

[4] L. Mirny, E. Shakhnovich, Protein folding theory: from lattice to all-atom models, Ann. Rev. Biophys. Biomol. Struct. 30 (2001) 361–396.

[5] S.E. Jackson, How do small single-domain proteins fold? Fold. Des. 3 (1998) R81–R91.

[6] S.W. Englander, Protein folding intermediates and pathways studied by hydrogen exchange, Ann. Rev. Biophys. Biomol. Struct. 29 (2000) 213–238.

[7] W.A. Eaton, V. Muñoz, S.J. Hagen, G.S. Jas, L.J. Lapidus, E.R. Henry, J. Hofrichter, Fast kinetics and mechanisms in protein folding, Ann. Rev. Biophys. Biomol. Struct. 29 (2000) 327–359.

[8] M. Vendruscolo, M. Paci, E. Dobson, M. Karplus, Three key residues form a critical contact network in a protein folding transition state, Nature 409 (2001) 641–645.

[9] A.R. Fersht, A. Matouschek, L. Serrano, The folding of an enzyme. Part I. Theory of protein engineering analysis of stability and pathway of protein folding, J. Mol. Biol. 224 (1992) 771–782.

[10] L.S. Itzhaki, D.E. Otzen, A.R. Fersht, The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation condensation mechanism for protein folding, J. Mol. Biol. 254 (1995) 260–288.

[11] A.R. Fersht, Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism, Proc. Natl. Acad. Sci. 97 (2000) 1525–1529.

[12] M. Karplus, D.L. Weaver, Protein folding dynamics: the diffusion–collision model and experimental data, Protein Sci. 3 (1994) 650–668.

[13] O.V. Galzitskaya, A.V. Finkelstein, A theoretical search for folding/unfolding nuclei in three-dimensional protein structures, Proc. Natl. Acad. Sci. 96 (1999) 11299–11304.

[14] I.Y. Torshin, R.W. Harrison, Charge centers and formation of the protein folding core, Proteins 43 (2001) 353–364.

[15] V.J. Hilser, D. Dowdy, T.G. Oas, E. Freire, The structural distribution of cooperative interactions in proteins: analysis of the native state ensemble, Proc. Natl. Acad. Sci. 95 (1998) 9903–9908.

[16] S.W. Englander, L. Mayne, Y. Bai, T.R. Sosnick, Hydrogen exchange: the modern legacy of Linderstrøm-Lang, Protein Sci. 6 (1997) 1101–1109.

[17] C. Woodward, Is the slow-exchange core the protein folding core? TIBS 18 (1993) 359–360.

[18] R. Li, C. Woodward, The hydrogen exchange core and protein folding, Protein Sci. 8 (1999) 1571–1591.

[19] M. Oliveberg, A.R. Fersht, Thermodynamics of transient conformations in the folding pathway of barnase: reorganization of the folding intermediate at low pH, Biochemistry 35 (1996) 2738–2749.

[20] D.J. Jacobs, L.A. Kuhn, M.F. Thorpe, Flexible and rigid regions in proteins, in: M.F. Thorpe, P.M. Duxbury (Eds.), Rigidity Theory and Applications, Kluwer Academic Publishers/Plenum Press, Dordrecht/New York, 1999, pp. 357–384.

[21] M.F. Thorpe, B.M. Hespenheide, Y. Yang, L.A. Kuhn, Flexibility and critical hydrogen bonds in cytochrome *c*, in: R.B. Altman, A.K. Dunker, L. Hunter, K. Lauderdale, T.E. Klein (Eds.), Pacific Symposium on Biocomputing, World Scientific, New Jersey, 2000, pp. 191–202.

[22] D.J. Jacobs, A.J. Rader, L.A. Kuhn, M.F. Thorpe, Protein flexibility predictions using graph theory, Proteins: Struct. Func. Genet. 44 (2001) 150–165.

[23] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.

[24] A. Fontana, M. Zambonin, P. Polverino de Laureto, V. de Filippis, A. Clementi, E. Scaramella, Probing the conformational state of apomyoglobin by limited proteolysis, J. Mol. Biol. 266 (1997) 223–230.

[25] G. Vriend, What If: a molecular modeling and drug design program, J. Mol. Graph 8 (1990) 52–56.

[26] M.A. Williams, J.M. Goodfellow, J.M. Thornton, Buried waters and internal cavities in monomeric proteins, Protein Sci. 3 (1994) 1224–1235.

[27] B.I. Dahiyat, D.B. Gordon, S.L. Mayo, Automated design of the surface positions of protein helices, Protein Sci. 6 (1997) 1333–1337.

[28] X. Dong, C.-J. Tsai, R. Nussinov, Hydrogen bonds and salt bridges across protein–protein interfaces, Protein Eng. 10 (1997) 999–1012.

[29] A. Bondi, van der Waals volumes and radii, J. Phys. Chem. 68 (1964) 441–451.

[30] C. Tanford. The hydrophobic effect, 2nd ed., Wiley/Interscience, New York, 1980.

[31] A.J. Rader, B.M. Hespenheide, L.A. Kuhn, M.F. Thorpe, Protein unfolding: rigidity lost, Proc. Natl. Acad. Sci. 99 (2001) 3540–3545.

[32] M. F. Thorpe, D. J. Jacobs, Computer-implemented system for analyzing rigidity of substructures within a macromolecule, US Patent number 6,104,449, 1998.

[33] J. Lagrange, Méchanique analytique, Paris, p. 1788.

[34] J.C. Maxwell, On the calculation of the equilibrium and stiffness of frames, Philos. Mag. 27 (1864) 294–299.

[35] G. Laman, On graphs and rigidity of plane skeletal structures, J. Eng. Math. 4 (1970) 331–340.

[36] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.

[37] M.-F. Jeng, S.W. Englander, G.A. Eløve, A.J. Wand, H. Roder, Structural description of acid-denatured cytochrome *c* by hydrogen exchange, Biochemistry 38 (1990) 10433–10437.

[38] M.F. Thorpe, D.J. Jacobs, D.J. Djordjevic. The structure and rigidity of glass networks, in: P. Boolchand (Ed.), Insulating and Semiconducting Glasses, World Scientific, New Jersey, 2000, pp. 95–145.

[39] M. Levitt, M. Gerstein, E. Huang, S. Subbiah, J. Tsai, Protein folding: the endgame, Ann. Rev. Biochem. 66 (1997) 549–579.

[40] C.-J. Tsia, J.V. Maizel Jr., R. Nussinov, Anatomy of protein structures: visualizing how a one-dimensional protein chain folds into a three-dimensional shape, Proc. Natl. Acad. Sci. 97 (2000) 12038–12043.

[41] A. Wallqvist, G.W. Smythers, D.G. Covell, Identification of cooperative folding units in a set of native proteins, Protein Sci. 28 (1997) 1627–1642.

[42] N.V. Dokholyan, L. Li, F. Ding, E.I. Shakhnovich, Topological determinants of protein folding, Proc. Natl. Acad. Sci. 99 (2002) 8637–8641.

[43] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, CATH: a hierarchic classification of protein domain structures, Structure 5 (1997) 1093–1108.

[44] S. Perrett, J. Clarke, A.M. Hounslow, A.R. Fersht, Relationship between equilibrium amide proton exchange behavior and the folding pathway of barnase, Biochemistry 34 (1995) 9288–9298.

[45] Y. Pan, M.S. Briggs, Hydrogen exchange in native and alcohol forms of ubiquitin, Biochemistry 31 (1992) 11405–11412.

[46] C.K. Woodward, B.D. Hilton, Hydrogen isotope exchange kinetics of single protons in bovine pancreatic trypsin inhibitor, Biophys. J. 32 (1980) 561–575.

[47] L.S. Mullins, C.N. Pace, F.M. Raushel, Conformational stability of ribonuclease T1 determined by hydrogen–deuterium exchange, Protein Sci. 6 (1997) 1387–1395.

[48] J.L. Neira, L.S. Itzhaki, D.E. Otzen, B. Davis, A.R. Fersht, Hydrogen exchange in chymotrypsin inhibitor 2 probed by mutagenesis, J. Mol. Biol. 270 (1997) 99–110.

[49] P.C. Driscoll, A.M. Wingfield, G.M. Clore, Determination of the secondary structure and molecular topology of interleukin-1β by use of two- and three-dimensional heteronuclear 15N-1H NMR spectroscopy, Biochemistry 29 (1990) 4668–4682.

[50] D.E. Anderson, J. Lu, L. McIntosh, F.W. Dahlquist, NMR of Proteins, CRC Press, Boca Raton, 1993, pp. 258–304.

[51] B.A. Schulman, C. Redfield, Z.-Y. Peng, C.M. Dobson, P.S. Kim, Different subdomains are most protected from hydrogen exchange in the molten globule and native state of human α-lactalbumin, J. Mol. Biol. 253 (1995) 651–657.

[52] F.M. Hughson, P.E. Wright, R.L. Baldwin, Structural characterization of a partially folded apo-myoglobin intermediate, Science 249 (1990) 1544–1548.