Asymmetry in the Shapes of Folded and Denatured States of Proteins[†]

Ruxandra I. Dima and D. Thirumalai*

Institute for Physical Science and Technology and Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742

Received: October 15, 2003; In Final Form: January 29, 2004

The asymmetry in the shapes of folded and unfolded states are probed using two parameters, Δ (a measure of the sphericity) and S that describes the shape ($S > 0$ corresponds to prolate and $S < 0$ represents oblate). For the folded states, whose interiors are densely packed, the radii of gyration (R_g), Δ, and S are calculated using the coordinates of the experimentally determined structures. Although R_g scales as $N^{1/3}$, as expected for maximally compact structures, the distributions of Δ and S show that there is considerable asymmetry in the shapes of folded structures. The degree of asymmetry is greater for proteins that form oligomers. Analysis of the two- and three-body contacts in the native structures shows that the presence of near equal number of contacts between backbone and side-chains and between side-chains gives rise to dense packing. We suggest that proteins with relatively large values of Δ and S can tolerate volume mutations without greatly affecting the network of contacts or their stability. Limited analysis of the mutagenesis data on T4 lysozyme and λ-repressor support this conclusion. To probe shape characteristics of denatured states we have developed a $C_\alpha - C_\beta$ model of a WW-like domain. The shape parameters, which are calculated using Langevin simulations, change dramatically in the course of coil-to-globule transition. Comparison of the values of Δ and S between the globular state and the folded state of WW domain shows that both energetic (especially dispersion in the hydrophobic interactions) and steric effects are important in determining packing in proteins.

Introduction

The densities of the core of the compact folded states of proteins approach those observed in small-molecule organic crystals.[1] Based on the observation that packing densities of proteins are high, it can be argued that the requirement of dense packing in the core may greatly restrict the number of sequences that can fold into a unique structure.[1–3] The need for folded proteins to have a high packing density is not so stringent a requirement that volume mutations cannot be made. Indeed, extensive experiments on a number of proteins show that relatively large volume changes (up to about 10 methylene equivalents) can be made without drastically affecting protein stability.[1,4–6] Even large volume mutations can be accommodated by minor variations in the secondary structure without compromising the global fold.[7] However, it appears that maintaining the hydrophobic nature of the core may be the key factor in establishing high packing density. Current experimental evidence suggests that excluded volume effects and hydrophobic interactions are the major factors that give rise to high packing density.[1]

Recently, Liang and Dill[8] have reexamined packing in proteins using a Delaunay triangulation method of dividing space. This approach, which is used to compute the distributions of voids and occupied regions, shows that both high-density liquid-like behavior and solid-like characteristics are seen in proteins. The free volume distributions suggest that proteins have the characteristics of high-density liquids and glasses that are often accompanied by the presence of "defects". There may be a dual requirement on the overall structure of proteins: (i) Marginal stability of the folded state can be achieved by having high packing density, but not necessarily close packed structures; (ii) Plasticity of proteins, as evidenced by their ability to tolerate mutations, is better accommodated by nonspherical folds with voids. Perhaps, the two requirements are required for foldability and function of proteins.

Because proteins form dense packed structures, we expect that their overall shapes would be spherical. The purpose of this paper is two-fold: (a) We use ideas familiar in polymer physics[9–11] to describe shape parameters of proteins using only the coordinates of the native structure. Examination of the asymmetry in the overall shapes gives insights into the extent to which a protein can tolerate volume mutations. (b) The folding pathways of proteins depend on the nature of the denatured states which are difficult to characterize. To shed light on the changes in shapes of proteins that occur in the course of a folding reaction we have used a simple off-lattice $C_\alpha – C_\beta$ representation of a WW-domain to model denatured states. Comparison of the shape parameters (see below) of the folded and denatured states shows that considerable asymmetry is lost as the protein folds. These calculations show that the degree of asymmetry in the folded states of proteins arises due to dispersions in the energetic interactions involving side-chains and the peptide backbone.

Methods

Shape Parameters. To determine the shape of a given conformation we use two rotationally invariant quantities S and Δ that are determined using the inertia tensor,[10–13]

$$T_{\alpha\beta} = \frac{1}{2N^2}\sum_{i,j=1}^{N}(r_{i\alpha} - r_{j\alpha})(r_{i\beta} - r_{j\beta}) \tag{1}$$

---

where N is the number of atoms in the protein, $r_{i\alpha}$ is the αth component of the position of atom i and $\alpha, \beta = x, y, z$. The eigenvalues of T, denoted by λ_i, are the squares of the three principal radii of gyration. Thus, $R_g^2 = trT = \Sigma_{i=1}^3 \lambda_i$

Asphericity is measured using

$$\Delta = \frac{3}{2} \frac{[\sum_{i=1}^3 (\lambda_i - \bar{\lambda})^2]}{(tr\mathbf{T})^2} \qquad (2)$$

where

$$\bar{\lambda} = \frac{tr\mathbf{T}}{3}$$

is the average eigenvalue of the inertia tensor. Deviation of Δ from 0 (the value corresponding to a sphere) gives an indication of the extent of anisotropy. The overall shape of a protein is characterized using

$$S = 27 \frac{[\prod_{i=1}^3 (\lambda_i - \lambda)]}{(tr\mathbf{T})^3} \qquad (3)$$

Negative values of S correspond to oblate shapes and positive values to prolate shapes. For perfect spheres, S and Δ are $S = \Delta = 0$. The parameters Δ and S obey the inequalities $0 \le \Delta \le 1$ and $(-1/4) \le S \le 2$.

The advantage of quantifying shapes in terms of Δ and S is that they can be directly computed using *only the coordinates* of the experimentally determined structures without assumptions about overlaps between van der Waals volumes assigned to atoms[8] that are used in the Voronoi analysis.[2,3] The shape parameters are directly calculated from the inertia tensor, which uses only the structures of proteins. Because the most dense packing in the native state occurs for $\Delta = S = 0$, it follows that deviations of Δ and S from zero would suggest that packing is compromised.

To obtain the distribution of Δ and S for folded proteins, we chose a representative set of protein structures from the Protein Databank (PDB),[14] PDBselect (ftp.embl-heidelberg.de/pub/databases/protein_extras/pdb_select/recent.pdb_select), which contains proteins with at most 25% similarity between their sequences. We further selected from this initial set those proteins that satisfy the following criteria: (i) the structures have very few missing atoms and/or amino acids; (ii) they have only a few amino acids with multiple coordinates; (iii) the secondary structure assignment must be in the PDB header, and (iv) the data set should not include membrane proteins. With these criteria our database consists of 1177 proteins, containing both single- and multichain proteins. The shape parameters Δ and S are calculated using the coordinates of the proteins in the dataset.

Number of Two- and Three-Body Contacts. To understand the origin of the dense packing in proteins, we have enumerated the number of two- and three-body contacts. The nature of these contacts is characterized by dividing the 20 amino acids into nine classes based on their van der Waals volumes and charge type (see for example Table 1.1 in ref 15): (1) Gly; (2) Pro; (3) Cys; (4) small hydrophobs (sH) Ala, Val, Ile, Leu, Met; (5) large hydrophobs (LH) Tyr, Trp, Phe; (6) small polar (sP) Ser, Thr; (7) large polar (LP) Asn, Gln, His; (8) positively charged (+) Arg, Lys; and (9) negatively charged (−) Asp, Glu.

For each amino acid, we selected all its side-chain(s) heavy atoms to determine the s−s contacts. By representing the backbone (b) by C_α atoms we also computed the number of b−b and s−b contacts. Two side-chains are in contact if at least one pair of heavy atoms (from the two residues) is separated by at most 5.2 Å. Two backbones are in contact when the distance between them is at most 6.5 Å. A side-chain and a backbone make a contact when the minimal distance between a pair of their atoms is at most 5.5 Å. We also determined the various types of three-body contacts: (b,b,b), (b,b,s), (b,s,s), and (s,s,s) using the same cutoff distances between the atoms of the three residues.

Results

Radius of Gyration of Folded State Obeys Flory Law. For a perfectly spherical molecule, volume, an extensive variable, is

$$V = \frac{4\pi}{3} R_g^3$$

From this it follows that for a spherical molecule $R_g \sim aN^{1/3}$ where N is the number of amino acid residues. The plot of R_g, for the dataset of proteins, as a function of N confirms the Flory result ($R_g \sim aN^{1/3}$ with $a = 3$ Å) with a correlation of 0.90 (Figure 1(a)).

Dobson and collaborators[16] measured, by pulse field gradient NMR techniques, the distribution of hydrodynamic radii for a set of natively folded proteins. The data fits the empirical equation: $R_h = (4.75 \pm 1.11)N^{0.29\pm0.02}$. For a spherical chain,

$$R_g = \sqrt{\frac{3}{5}} R_h$$

and therefore

$$R_g \approx (3.67 \pm 0.86)N^{0.29\pm0.02} \qquad (4)$$

which is in agreement with the best fit to our data (Figure 1(a)). Because $R_g \sim 3N^{1/3}$, as predicted by Flory for compact globules, we expect that proteins should be spherical. From the values of R_g alone it would seem that proteins are maximally compact and are likely to produce close packed structures. However, a detailed examination of the shapes shows (see below) that this is not the case.

The mean end-to-end distance, R_{end}, is often used as a measure of the size of a polypeptide chain. Fluorescence energy transfer (FRET) experiments can be used to measure R_{end}. For homopolymers in good solvents it can be shown that $R_{end} \sim R_g$ provided N is large enough i.e., both R_{end} and R_g obey the Flory scaling law. For finite sized polypeptide chains it is not clear that R_{end} provides even an approximation of R_g. The calculated R_{end} values for the dataset of proteins in their native states do not follow the Flory result, i.e., $R_{end} \sim N^{1/3}$ (Figure 1(b)). From this it follows that the implicit assumption that R_{end} provides an estimate of R_g made in several FRET experiments is erroneous. It is possible that at high temperatures or high denaturant concentrations R_{end} and R_g may be proportional to each other.

Single-Chain Proteins are Approximately Aspherical. The distribution $P(\Delta)$ of Δ for the proteins in PDBselect is peaked at $\Delta \approx 0.1$ for single-chain proteins (Figure 2). Despite a relatively long tail at large values of Δ, about 78% of the 403 single-chain proteins from PDBselect have asphericities smaller than 0.2. This shows that, to a first approximation, the native-
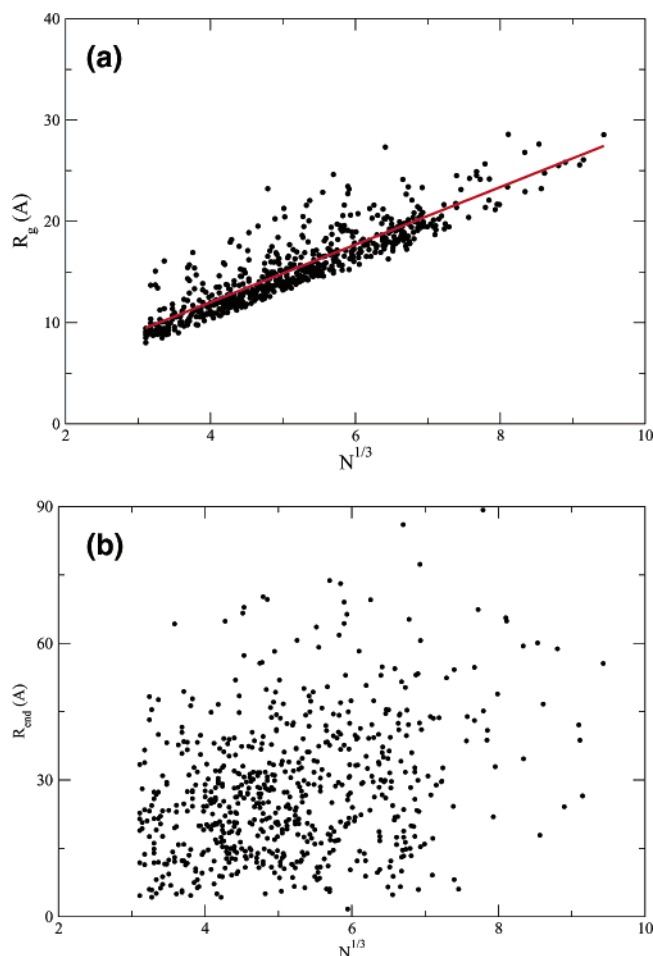
Figure 1. (a) Radius of gyration R_g, as a function of N, the number of amino acid residues for the 403 monomeric proteins in our dataset. The best linear fit to the data with a correlation of 0.9 gives $R_g = 3N^{0.333}$ Å. (b) Plot of the end-to-end distance, R_{end}, as a function of N for the dataset of proteins used in Figure 1a. Clearly, R_{end} does not obey Flory's law. Thus, for folded states of proteins, R_{end} does not provide an estimate of R_g.
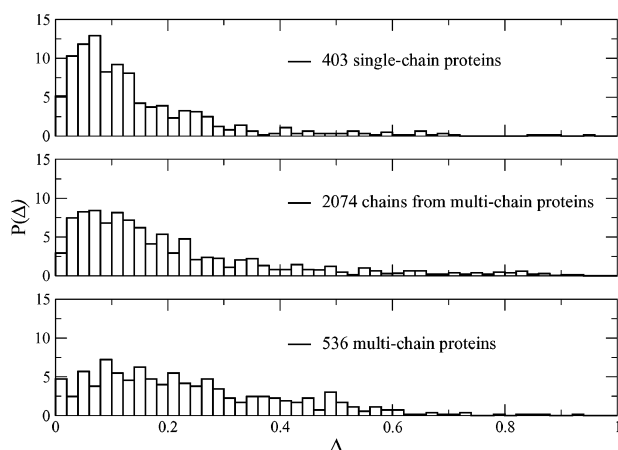

Figure 2. Distribution of the asphericity parameter Δ for the proteins in the dataset. The top panel, which corresponds to $P(\Delta)$ for the single-chain proteins, shows that a substantial fraction have $\Delta < 0.1$. The middle panel shows $P(\Delta)$ for single-chain proteins that are part of oligomers. In contrast to the top panel, $P(\Delta)$ is broad, which implies that single-chain proteins adopt aspherical shapes when they oligomerize. Large deviations from sphericity are seen in multichain proteins (bottom panel).

state conformations of globular single-chain proteins can be modeled as spheres. A similar behavior is found for chains that
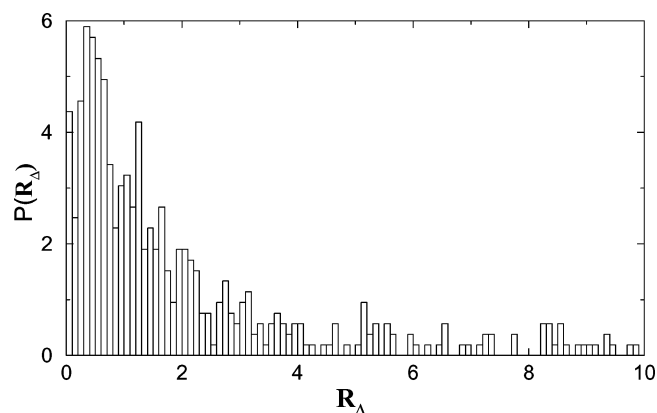

Figure 3. Distribution of R_Δ for the 536 multichain proteins. Here, $R_\Delta = (\Delta_O/\Delta_I)$, where Δ_O is the asphericity parameter for the oligomer and Δ_I is the corresponding value for the individual chains. The fraction $\int_1^\infty P(R_\Delta)\, dR_\Delta \approx 0.6$, which implies that oligomers are more asymmetric than the individual chains from which they are constructed.

are part of multichain proteins, but in this case the distribution is considerably broader (see middle panel of Figure 2).

Many proteins are organized as oligomers. For oligomeric proteins a clear deviation from sphericity appears when considering the whole oligomers. For oligomers the distribution is broad. Only about 50% of oligomeric proteins have Δ values smaller than 0.2 (bottom panel in Figure 2). The deviation from sphericity is also evident in the distribution $P(R_\Delta)$ where

$$R_\Delta = \frac{\Delta_O}{\Delta_I}$$

where Δ_O and Δ_I are the asphericity parameters for the oligomer and the individual chains in the oligomer (Figure 3). If both the oligomer and the individual chains have similar values of Δ, we expect $P(R_\Delta)$ to be peaked around 1. The distribution $P(R_\Delta)$ is peaked at R_Δ between $(0.4-0.6)$ (Figure 3). The majority (60%) of oligomers have $R_\Delta > 1$. In other words, the oligomer is less spherical than its chains. The results in Figure 2 show that, in general, the approximation that isolated proteins are spheres with $R_g \sim aN^{1/3}$, which is used for the interpretation of a wide variety of experimental techniques (like sedimentation analysis and gel filtration), is likely to produce small errors. However, such an approximation for oligomeric proteins is inaccurate.

Although the bulk of globular proteins have relatively low values of Δ (<0.2), there are a number of proteins in the tail of the distribution. Two proteins for which both Δ_O and Δ_I have very large values (>0.80) are spectrin (PDB code 2spc) and a signaling protein (PDB code 1qu7). Both the individual chains and the whole protein are very elongated, and the chains appear intertwined in the oligomer. Large deviations from a spherical shape for these helical proteins might be required for functional reasons.

Majority of the Proteins are Prolate. The distribution of the shape parameter for protein chains (see Figure 4) reveals that only about 20% of single-chains have $S \approx 0$. An additional 20% have $S < 0$, that is they have an oblate shape, while the remaining 60% are prolate. For multichain proteins, only about 10% have $S \approx 0$. Even if Δ seems to indicate that usually proteins are approximately spherical, their actual shape tends to be prolate. Thus, there is considerable asymmetry in the shapes of globular proteins.

To put these findings into context we note that the values of the shape parameter for real proteins are smaller than for
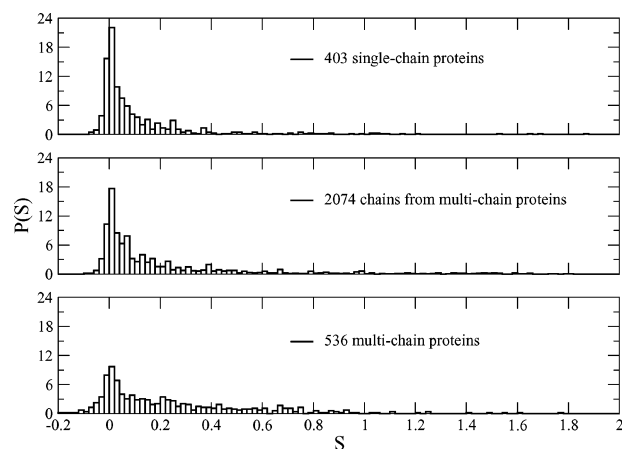
Asymmetry in Folded and Denatured Proteins

J. Phys. Chem. B, Vol. 108, No. 21, 2004 **6567**

Figure 4. Distribution of $P(S)$ of S, the shape parameter from (eq 3). The distributions of $P(S)$ in the top and middle panels are similar, which show that the shapes of single-chain proteins (either in isolation or as part of a multichain) are roughly the same. The distribution $P(S)$ for the oligomers is broader with a larger fraction of $S < 0$ corresponding to an overall oblate shape.
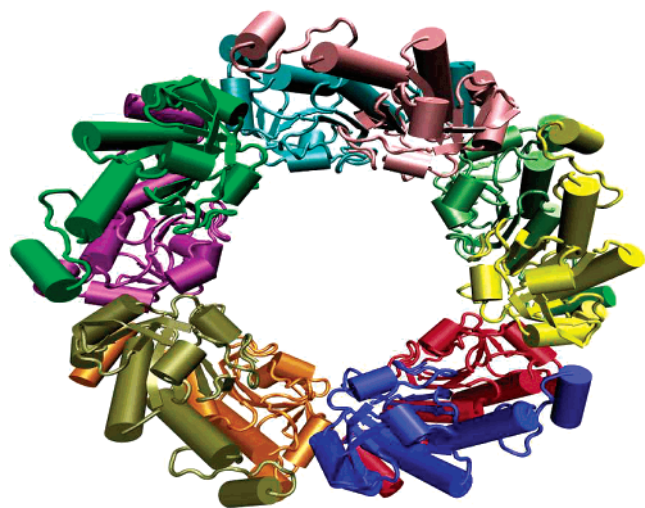

Figure 5. Structure of the human thioredoxin peroxidase-B (1qmv), which has one of the most negative values of S ($S = -0.15$). In the denatured states it is likely that the overall shape is prolate. Thus, during folding and oligomerization there is a qualitative and large change in the overall shapes of such proteins. The figure was produced with the packages VMD[32] and PovRay.

Gaussian chains.[11] In the latter case $S > 0.63$ which shows that Gaussian chains are always prolate. Because Gaussian chains are useful models of denatured states, it is reasonable to assume that upon folding the chain becomes compact but retains the prolate shape. It is therefore surprising that about 20% of proteins in the native state have oblate shapes, i.e., $S < 0$. Two such proteins with the largest (in absolute value) negative S, and therefore having the most pronounced oblate shapes, are 1qb3 ($S = -0.164$) and 1qmv ($S = -0.153$). Both are multichain proteins in which the identical individual chains are disposed in an almost perfect circular registry as in human peroxidase-B (1qmv) (Figure 5).

For proteins with $S < 0$, there have to be substantial changes in the shape during the transition from denatured states to the folded state. The transition from $S > 0$, which is typical of unfolded states, to values of $S < 0$ in all likelihood involves substantial variations in the accessible surface area. Therefore, such a transition might involve a larger energetic barrier than the typical case when S does not change sign. The helical
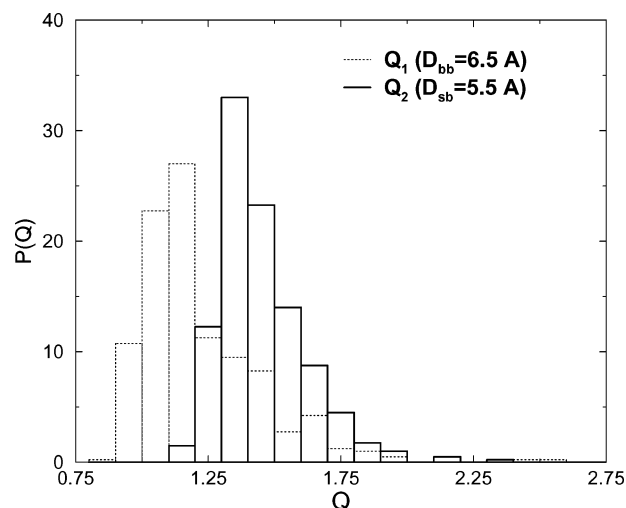

Figure 6. Distributions of $Q_1 = N_{bb}/N_{ss}$ and $Q_2 = N_{sb}/N_{ss}$, where N_{bb} is the number of b−b contacts, N_{sb} is the number of s−b contacts, and N_{ss} is the number of s−s contacts in the 403 single-chain proteins. Both distributions are peaked at values larger than 1. This shows that in the native states of globular proteins, contacts involving backbone atoms and especially those between backbones and side-chains occur frequently.

proteins, 2spc and 1qu7, have the largest preference for prolate shape (with $S = 1.5$ and 1.8 respectively). More generally, the examination of the distribution of S and Δ for the 403 monomeric proteins shows that usually proteins with small asphericities have also small values of S. This rule breaks down somewhat unexpectedly for proteins with substantially oblate shapes (i.e., large negative S) that have relatively small asphericities, especially when compared to the proteins with pronounced prolate shapes (i.e., large positive S).

High Degree of Packing in Proteins Correlates with Substantial Contacts Involving Backbone Atoms. To probe the origin of the small values of Δ, which is linked to high packing density in nearly 80% of globular proteins, we analyzed the number of two-body and three-body contacts in our dataset of proteins. We have previously reported that there are a large number of two-body contacts involving the protein backbone.[17,35] Using the nine classes of amino acids (see the Methods section) with the cutoffs $D_{bb} = 6.5$ Å, $D_{sb} = 5.5$ Å, and $D_{ss} = 5.2$ Å, we determined the number and nature of two- and three-body contact between backbones, between backbones and side-chains, and between side-chains. The distributions $P(Q_1)$ of $Q_1 = (N_{bb}/N_{ss})$ and $P(Q_2)$ of $Q_2 = (N_{sb}/N_{ss})$, where N_{bb} is the number of b−b contacts, N_{sb} is the number of s−b contacts, and N_{ss} is the number of s−s contacts in the 403 single-chain proteins (Figure 6), have several interesting and somewhat unexpected features. The distributions are peaked at values of Q_1 and Q_2 larger than 1, the shift being most evident for the s−b contacts. This implies that, in the native states of globular proteins, contacts between backbone atoms and especially those between backbones and side-chains occur at least as frequently as contacts between the side-chains. Second, both distributions have relatively long tails that extend well into the region corresponding to Q_1 and Q_2 greater than 1.5. Upon closer inspection we find that in general the b−b and s−b contacts occur in a correlated manner with the s−s contacts. In other words, if two residues are found to have their side-chains in contact, then their backbones are also very likely to be in contact or at least the backbone of one residue is in contact to the side-chain of the other residue. In addition, the backbones of the pairs of their adjacent residues along the chain are also usually

in contact with the other or/and with the side-chains. These distributions are similar for chains in the multichain proteins from PDBselect (data not shown). The only difference is that in this case $P(Q_1)$ is peaked at somewhat lower values while $P(Q_2)$ is peaked at slightly larger values than in the case of single-chain proteins.

Analysis of the nature of amino acids that participate in the s−s contacts shows that, as expected, the largest percentages come from contacts between hydrophobic residues. On an average 18% are contacts between small hydrophobs (Ala, Val, Ile, Leu, Met) and 12% are between small hydrophobs and large hydrophobs (Tyr, Trp, Phe). Contacts among large hydrophobs are scarce as they represent only approximately 1.5% of all s−s contacts.

Enumeration of the number of three-body contacts shows a pronounced shift with respect to the distributions of two-body contacts. The number of contacts between side-chains, which is on the same order as the number of contacts between 2 side-chains and a backbone, is typically larger than contacts between backbones and between 2 backbones and a side-chain. The distribution of the types of s−s−s contacts shows that the two strongest contributions are due to interactions among small hydrophobs or between 2 small hydrophobs and a large hydrophob. The contacts between three large hydrophobs become even less favored; in about 80% of proteins their contribution adds up to less than 1% of the total number of s−s−s contacts. Other types of three-body contacts that appear as highly unfavored in the proteins from our dataset are contacts between three Gly residues, three Cys residues, three Pro residues, or between three charged amino acids of the same sign. Due to Coulomb repulsion, contacts between any other combination of three charged residues are infrequent, usually contributing with about 0.4% to the total number of s−s−s contacts, while the contribution from contacts between three charges of the same sign is at best 0.06%. The presence of near equal number of two- and three-body contacts involving backbone and side-chains confers tight packing, which in turn explains the low Δ values in proteins.

Shape Parameters for "Denatured" States of Proteins. A complete understanding of how proteins fold requires characterizing the structures of the ensemble of unfolded states. Several recent studies have argued that unfolded states can have considerable residual structure even at elevated denaturant concentrations.[18−22] To a first approximation, unfolded states can be modeled as random coils. Therefore, the radius of gyration for denatured states should be given by $R_g \approx l_u N^{\nu_F}$, where the Flory exponent $\nu_F \approx 0.6$ and l_u is a "persistence" length. The Gaussian chains (corresponding to Θ-solvent condition) are the simplest approximation for random coils. The values of Δ and S for Gaussian chains can serve as estimates of the shape parameters for unfolded states of proteins. Explicit calculation of the asphericity parameter for long Gaussian chains, averaged over an ensemble of denatured states, shows that[23]

$$\langle \Delta \rangle = \frac{1}{8}(26 - 9K_3) \quad (5)$$

where $K_3 = \int_0^\infty dx x^{5/2}(\sinh x)^{-3/2}$. Thus, numerical calculations give $\Delta = 0.52$ and $S = 0.87$. If excluded volume interactions are taken into account we find that $\Delta = 0.55$ and $S = 0.92$. These calculations show that, if denatured states are modeled as random coils without side-chains, then the average asphericity should show considerable deviation from globularity. It also

follows that large changes in Δ and S are accompanying the folding process.

Because interactions between side-chains and backbone play a crucial role in determining the packing in proteins, realistic estimates of Δ and S even for denatured states must include side-chain effects. To obtain reliable values of Δ and S for models of denatured states we used Langevin simulations of an off-lattice model of the 34-residue WW-domain. These calculations provide a picture of the states sampled by the denatured state ensemble (DSE) in a good and a poor solvent conditions. Solvent quality is mimicked by tuning the interaction potentials between side-chains (see below).

Each residue in a chain is represented by two beads: one is the position of its C_α carbon atom and the other one is the center of mass of its side-chain. The conformation of the chain is specified by the set of C_α positions $(\vec{r}_{b,i})$ and the set of side-chain centers of mass positions $(\vec{r}_{s,i})$. The potential energy of each conformation is given by

$$E_{pot}(\{\vec{r}_{b,i}\},\{\vec{r}_{s,i}\}) = V_{BL} + V_{SBC} + V_{BA} + V_{DIH} + V_{NB} \quad (6)$$

where V_{BL}, V_{SBC}, V_{BA}, V_{DIH}, and V_{NB} are bond-length potential, side-chain to backbone connectivity potential, bond-angle potential, dihedral angle potential, and nonbonded long-range potential. The bond length potential is

$$V_{BL} = \sum_{i=1}^{N-1} \frac{k_r}{2}(|\vec{r}_{b,i+1} - \vec{r}_{b,i}| - r_{i,i+1}^{WT})^2 \quad (7)$$

where $k_r = 200\ \epsilon_h/a^2$, $a = 3.8$ Å, ϵ_h is the average strength of the hydrophobic interaction (i.e., it is the unit of energy in our model), and $r_{i,i+1}^{WT}$ is the distance between the C_α at positions i and $i+1$ in the wild-type sequence from the PDB (N-terminal of 1pin)

$$V_{SBC} = \sum_{i=1}^{N} \frac{k_r}{2}(|\vec{r}_{s,i} - \vec{r}_{b,i}| - r_{sb,i}^{WT})^2 \quad (8)$$

where $r_{sb,i}^{WT}$ is the distance between the C_α and the center of mass of the side-chain at position i in the wild-type sequence from the PDB, and V_{BA} is given by

$$V_{BA} = \sum_{i=1}^{N-2} \frac{k_\theta}{2}(\theta_i - \theta_i^{WT})^2 \quad (9)$$

where $k_\theta = 20\ \epsilon_h/(rad)^2$ and θ_i^{WT} is the bond angle at positions i in the wild-type sequence from the PDB. The dihedral angle potential is given by

$$V_{DIH} = \sum_{i=1}^{N-3}[A_i(1 - \cos d\phi) + B_i(1 + \cos 3d\phi) + C_i(1 - \sin d\phi)] \quad (10)$$

where $A_i = 1.0\ \epsilon_h$, $B_i = 1.6\ \epsilon_h$, $C_i = 2.0\ \epsilon_h$, $d\phi = \phi - (\phi^{WT} - 60.5)$ when $\phi^{WT} < 2\pi/3$, $A_i = 1.2\ \epsilon_h$, $B_i = 1.2\ \epsilon_h$, $C_i = 0.0\ \epsilon_h$, $d\phi = \phi - (\phi^{WT} - 180)$ when $2\pi/3 < \phi^{WT} < 4\pi/3$ and $A_i = 1.0\ \epsilon_h$, $B_i = 1.6\ \epsilon_h$, $C_i = 2.0\ \epsilon_h$, $d\phi = \phi - (\phi^{WT} - 299.5)$ when $4\pi/3 < \phi^{WT} < 2\pi$. The nonbonded potential that gives rise to globularity is given by

$$V_{NB} = \sum_{i=1}^{N-3}\sum_{j=i+3}^{N} V_{ij}(r_{ij}) \quad (11)$$

Asymmetry in Folded and Denatured Proteins

J. Phys. Chem. B, Vol. 108, No. 21, 2004 **6569**

where $V_{ij}(r_{ij}) = 4\epsilon_h(R_{ij}/r_{ij})^{12}$ in good solvents and $V_{ij}(r_{ij}) = 4\epsilon_h[(R_{ij}/r_{ij})^{12} - (R_{ij}/r_{ij})^6]$ for bad solvents. The values for R_{ij} were taken from Table 1 in ref 24.

It has been argued that the two major factors that contribute to packing in proteins are steric effects and energetic interactions between the largely hydrophobic core in the protein interior.[1,25] Modeling energetic effects is more difficult than excluded volume interactions. To dissect these two contributions separately, we first neglect attractive interactions between the side chain C_β spheres. This corresponds to good solvent conditions. In this case the unfolded state is expected to have the characteristics of Flory random coil. In our model the coil-to-globule transition is induced by assigning attractive interactions that mimic poor solvent conditions between the C_β spheres. Assigning a uniform value for ϵ_h, which masks the sequence dependence entirely, allows us to examine shapes of maximally compact globular structures. Comparison of the Δ and S values for the globules, which are favored under poor solvent conditions, and the native state gives an indication of the dispersion in interactions that stabilizes a given fold.

We performed Langevin simulations in the underdamped limit with a low friction coefficient, $\mu = 0.05\ \tau_L^{-1}$, and a time integration step $h = 0.005\tau_L$, where $\tau_L = (ma^2/\epsilon_h)^{1/2}$ (ref 26). We generated 15 trajectories, each starting from different initial conditions. All trajectories start at a high temperature ($T = 3.0$ in units of ϵ_h/k_B) and after a quench to $T = 2.5$ the temperature is reduced sequentially in steps of 0.050 degrees until the final temperature, $T = 1.0$, is reached. At each temperature the shape parameters are determined as averages over 0.5×10^6 steps, but only after an initial lag of 10^6 steps which are needed for equilibration.

In good solvents, Δ and S for the WW-like domain are 0.5 and 0.7, respectively. The S values are less than those obtained by neglecting side-chains. A typical conformation of the Flory random coil is shown in Figure 7(a). When the solvent is poor, so that the coil-to-globule transition occurs, Δ and S reduce drastically. For our model $\Delta = 0.04$ and $S = 0.01$, which implies that the presence of side-chains drives the structures to acquire maximally compact spherical shapes (Figure 7). The spherical shape formed under poor solvent conditions with $\Delta \approx S \approx 0$ is different from the most probable values of Δ and S for proteins (Figures 2 and 4). In our model the radii of side-chains coincide with the van der Waals sizes of the residues in the WW-domain. Thus, steric effects which represent one factor in packing are properly taken into account. However, hydrophobic interactions in our model are uniform (ϵ_h is the same for interactions between all side-chains). This is not the case for the WW-domain in which dispersion in hydrophobic interactions along with energetics associated with hydrogen bond formation are necessary to produce the three-stranded β-sheet protein. In the native state (Figure 7(c)) Δ and S for the WW-domain are 0.15 and 0.07, respectively. These values, which are different from the ones for the maximally compact structures (nearly close packed) (Figure 7(b)), show that both excluded volume and energetic stabilization are important in providing the observed packing in proteins. These conclusions are consistent with the findings of Kussell et al.[25] who showed that in addition to excluded volume interactions, energetic effects (hydrophobic, electrostatic, and polar) must play a role in determining the fold of a protein.

Conclusions

A hallmark of native states of proteins is that their cores are densely packed.[1] In this paper we have shown that, although the radii of gyration of folded proteins follow the Flory law R_g
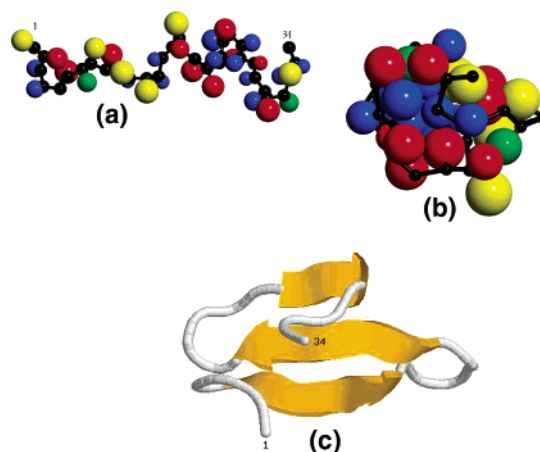
Figure 7. (a) Typical conformation of the off-lattice model $C_\alpha - C_\beta$ of the WW-like domain in a good solvent is a coil. The shape parameters for the Flory coil Δ and S are 0.5 and 0.7, respectively. These are much larger than the values for the native states of proteins seen in Figures 2 and 4. (b) Typical globular conformation of the off-lattice model of the WW-like domain in a bad solvent. In this case $\Delta = 0.04$ and $S = 0.01$, which implies that the presence of the attractive interactions between side-chains drives the structures to acquire a spherical shape. These figures have been produced with the packages RasMol[33] and Povray. (c) Conformation of the WT WW-domain (1pin). The Δ and S values are 0.15 and 0.07, respectively. Comparison with (b) shows that the acquisition of the three stranded β-sheet occurs at the expense of deviation from maximally compact structures. It is likely that the WT WW-domain can tolerate volume mutations without compromising its stability. This figure has been produced with the package MolScript.[34]

$\sim 3N^{1/3}$, there is considerable asymmetry in the shapes of proteins. Oligomeric proteins in general are found to deviate substantially from spherical shapes. The majority of single chain proteins are prolate ($S > 0$) with small Δ values (<0.2). Large deviations from sphericity show that proteins are not close packed, which is in accord with the conclusions of ref 8. This observation indicates that a vast number of proteins can tolerate substantial volume mutations. We hypothesize that in proteins with larger Δ and S values mutations that change the van der Waals volume of side-chains in the core can be made without compromising their stability or the network of contacts.

Mutagenesis experiments on T4 lysozyme and λ-repressor lend support to our hypothesis. The Δ (0.20) and S (0.19) values for the wild-type T4 lysozyme, a 164 residue single-chain protein (3lzm), are larger than those for a typical single-chain protein (Figure 2). Various volume mutations in the hydrophobic core of T4 lysozyme (32 of the approximately 350 structures of T4 lysozyme variants deposited in the PDB) lead to only minor increases in the Δ and S values (data not shown). The average Δ and S values for these mutants are 0.22 and 0.20, respectively. Irrespective of the sign of the volume change, the overwhelming majority of these mutations lead to increased stability of the structure (see Table 3 in ref 1). In addition, comparison of the number of 3-body s−s−s contacts in the WT T4 lysozyme and in one of the mutants (140l), shows that only 10% of the contacts are different. This suggests that in proteins with large Δ and S nonpolar core mutations induce only minimal rearrangements in the network of contacts.

A corollary of our hypothesis is that in fairly spherical proteins volume mutations would lead to a larger degree of rearrangements in the network of contacts. Indeed, upon analysis of the WT N-term of the λ-repressor (1lbm) and a mutant (1lli), both structures are spherical ($\Delta \approx 0.08$ and $S \approx 0.01$) but there is a 18% difference between the number of 3-body s−s−s

contacts. These two examples (mutational experiments on T4 lysozyme and λ-repressor) provide anecdotal evidence for our hypothesis that volume mutations in proteins with large Δ and S can be readily made without drastically altering the number of contacts. On the other hand, in highly spherical proteins substantial rearrangements of contacts accompanied perhaps by shifts in secondary structures occur upon mutations that change the volume of the side-chains. This anecdotal evidence establishes the possibility of our hypothesis relating Δ and S values to tolerance to volume mutations. A larger body of experimental data has to be analyzed to further confirm the proposed hypothesis.

The dense packing in most proteins arises because there is a substantial number of contacts between side-chains and backbone atoms. A number of recent studies[25,28-27,30,31] have also concluded that orientations of side-chains and their interactions with the protein backbone are important in determining the shapes of proteins. In this paper we have shown that the total number of b−b and s−b contacts is of the same order of magnitude as the number of s−s contacts. This is also true at the level of three-body contacts. Our results also suggest that the knowledge-based interaction potentials should include terms that account for the large number of b−b and s−b contacts.[27] Statistical potentials that focus only on interaction parameters between side-chains cannot successfully account for native state stability.

A model of denatured states of proteins using realistic representations ($C_\alpha - C_\beta$ model) shows that unfolded states are more spherical than polymer models that neglect side-chains. Even when the solvent is good we find that Δ and S values are less than the theoretically predicted values for Gaussian chains or chains with excluded volume. The simulations of collapsed structures for a model of WW-like domain that neglects the dispersion in the hydrophobic interactions shows that $\Delta \approx S \approx 0$. These values, which are substantially different for the WT structure, show that the variations in the sequence-dependent energies and excluded volume interactions are both responsible for the high-density packing observed in proteins.

References and Notes

(1) Richards, F. M.; Lim, W. A. *Q. Rev. Biophys.* **1994**, *26*, 423−498.
(2) Richards, F. M. *J. Mol. Biol.* **1974**, *82*, 1−14.
(3) Finney, J. L. *J. Mol. Biol.* **1975**, *96*, 721−732.
(4) Lim, W. A.; Farruggio, D. C.; Sauer, R. T. *Biochemistry* **1992**, *31*, 4324−4333.
(5) Karpusas, M.; Baase, W. A.; Matsumura, M.; Matthews, B. W. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 8237−8241.
(6) Hurley, J. H.; Baase, W. A.; Matthews, B. W. *J. Mol. Biol.* **1992**, *224*, 1143−1159.
(7) Lesk, A. M.; Chothia, C. *J. Mol. Biol.* **1980**, *136*, 225−270.
(8) Liang, J.; Dill, K. A. *Biophys. J.* **2001**, *81*, 751−766.
(9) Solc, K.; Stockmayer, W. *J. Chem. Phys.* **1971**, *54*, 2756−2766.
(10) Aronovitz, J. A.; Nelson, D. R. *J. Phys (Paris)* **1986**, *47*, 1445−1460.
(11) Honeycutt, J. D.; Thirumalai, D. *J. Chem. Phys.* **1989**, *90*, 4542−4559.
(12) Cannon, J. W.; Aronovitz, J. A.; Goldbart, P. *J. Physique I* **1991**, *1*, 629−645.
(13) Jagodzinski, O.; Eisenriegler, E.; Kremer, K. *J. Physique I* **1992**, *2*, 2243−2279.
(14) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Sindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(15) Creighton, T. E. *Proteins: Structures and molecular properties*; W. H. Freeman and Company: New York, 1993.
(16) Wilkins, D. K.; Grimshaw, S. B.; Receveur, V.; Dobson, C. M.; Jones, J. A.; Smith, L. J. *Biochemistry* **1999**, *38*, 16424−16431.
(17) Betancourt, M. R.; Thirumalai, D. *J. Phys. Chem. B* **2002**, *106*, 599−609.
(18) Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*, 487−489.
(19) Klein-Seetharaman, J.; Oikawa, M.; Grimshaw, S. B.; Wirmer, J.; Duchardt, E.; Ueda, T.; Imoto, T.; Smith, L. J.; Dobson, C. M.; Schwalbe, H. *Science* **2002**, *295*, 1719−1722.
(20) Mok, Y. K.; Kay, C. M.; Kay, L. E.; Forman-Kay, J. *J. Mol. Biol.* **1999**, *289*, 619−638.
(21) Wong, K. B.; Freund, S. M. V.; Fersht, A. R. *J. Mol. Biol.* **1996**, *259*, 805−818.
(22) SaabRincon, G.; Gualfetti, P. J.; Matthews, C. R. *Biochemistry* **1996**, *35*, 1988−1994.
(23) Diehl, H. W.; Eisenriegler, E. *J. Phys. A* **1989**, *22*, L87−L91.
(24) Kolinski, A.; Godzik, A.; Skolnick, J. *J. Chem. Phys.* **1993**, *98*, 7420−7433.
(25) Kussell, E.; Shimada, J.; Shakhnovich, E. I. *J. Mol. Biol.* **2001**, *311*, 183−193.
(26) Guo, Z.; Thirumalai, D. *Biopolymers* **1995**, *36*, 83−102.
(27) Buchete, N. V.; Straub, J. E.; Thirumalai, D. *J. Chem. Phys.* **2003**, *118*, 7658−7671.
(28) Bahar, I.; Jernigan, R. L. *Fold. Des.* **1996**, *1*, 357−370.
(29) Maritan, A.; Micheletti, C.; Trovato, A.; Banavar, J. R. *Nature* **2000**, *406*, 287−290.
(30) Bagci, Z.; Jernigan, R. L.; Bahar, I. *J. Chem. Phys.* **2002**, *116*, 2269−2276.
(31) Bagci, Z.; Kloczkowski, A.; Jernigan, R. L.; Bahar, I. *Proteins* **2003**, *52*, 56−67.
(32) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33−38.
(33) Sayle, R. A.; Milner-White, E. J. *Trends Biochem. Sci.* **1995**, *20*, 374−376.
(34) Kraulis, P. J. *J. Appl. Cryst.* **1991**, *24*, 946−950.
(35) Buchete, N. V.; Straub, J. E.; Thirumalai, D., submitted to *Protein Sci.*