doi:10.1016/j.jmb.2004.09.061

J. Mol. Biol. (2004) **344**, 797–811

JMB

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER

The Optimal Fraction of Hydrophobic Residues Required to Ensure Protein Collapse

Jiangbo Miao[1], Judith Klein-Seetharaman[1,2] and Hagai Meirovitch[3]*

[1]*Carnegie Mellon University*
School of Computer Science
Language Technologies Institute
Pittsburgh, PA 15213, USA

[2]*University of Pittsburgh*
School of Medicine, Department
of Pharmacology, Biomedical
Science Tower E1058
Pittsburgh, PA 15261, USA

[3]*University of Pittsburgh*
School of Medicine, Center for
Computational Biology and
Bioinformatics and Department
of Molecular Genetics and
Biochemistry, Biomedical
Science Tower W1058
Pittsburgh, PA 15261, USA

**Corresponding author*

The hydrophobic interaction is the main driving force for protein folding. Here, we address the question of what is the optimal fraction, f of hydrophobic (H) residues required to ensure protein collapse. For very small f (say $f<0.1$), the protein chain is expected to behave as a random coil, where the H residues are "wrapped" locally by polar (P) residues. However, for large enough f this local coverage cannot be achieved and the thermodynamic alternative to avoid contact with water is burying the H residues in the interior of a compact chain structure. The interior also contains P residues that are known to be clustered to optimize their electrostatic interactions. This means that the H residues are clustered as well, i.e. they effectively attract each other like the H-monomers in Dill's HP lattice model. Previously, we asked the question: assuming that the H monomers in the HP model are distributed randomly along the chain, what fraction of them is required to ensure a compact ground state? We claimed there that $f \approx p_c$, where p_c is the site percolation threshold of the lattice (in a percolation experiment, each site of an initially empty lattice is visited and a particle is placed there with a probability p. The interest is in the critical (minimal) value, p_c, for which percolation occurs, i.e. a cluster connecting the opposite sides of the lattice is created). Due to the above correspondence between the HP model and real proteins (and assuming that the H residues are distributed at random) we suggest that the experimental f should lead to percolating clusters of H residues over the highly dense protein core, i.e. clusters of the core size. To check this theory, we treat a simplified model consisting of H and P residues represented by their α-carbon atoms only. The structure is defined by the C^{α}–C^{α} virtual bond lengths, angles and dihedral angles, and the X-ray structure is best-fitted onto a face-centered cubic lattice. Percolation experiments are carried out for 103 single-chain proteins using six different hydrophobic sets of residues. Indeed, on average, percolating clusters are generated, which supports our theory; however, some sets lead to a better core coverage than others. We also calculate the largest actual hydrophobic cluster of each protein and show that, on average, these clusters span the core, again in accord with our theory. We discuss the effect of protein size, deviations from the average picture, and implications of this study for defining reliable simplified models of proteins.

Keywords: hydrophobic clusters; percolation theory; protein collapse; HP model

Introduction

The hydrophobic interaction is the main driving force for protein folding.[1] Therefore, a great deal of work has been done toward understanding the thermodynamic basis of hydrophobicity,[2,3] as well as the effect of this phenomenon on the details of protein structures.[4–7] The ability of a protein chain to organize itself in a stable compact structure is expected to depend strongly on the fraction, f, of the hydrophobic (H) residues. For small f (say $f<0.1$) of randomly distributed H residues, an H residue could become "wrapped" locally by several

Abbreviations used: SYS, systematic search; MCS, Monte Carlo search; fcc, face-centered cubic.

E-mail address of the corresponding author: hagaim@pitt.edu hydrophilic (P) residues to form a "blob". This would lead to an effectively shorter random coil chain of blobs connected by flexible segments, which gains further stability from its high entropy (see Figure 1). However, when f is large enough, the local coverage of the H residues cannot be achieved any more and the only thermodynamic alternative to avoid contact with water is burying them in the interior of a compact chain structure. Obviously, if f is too large the molecule will precipitate and therefore the optimal value observed in real proteins will be a balance between these effects and others.

It has been shown that most of the H residues are located in the interior of protein structures, while the exterior is populated mostly by P residues, which interact favorably with the surrounding water.[8–16] More specifically, in an inner sphere of radius R around the center of mass, where R is the radius of gyration, the concentration of the H residues is larger than their fraction, f, in the entire sequence; this concentration decreases significantly in concentric spherical layers of increasing radii, i.e. in going from the core towards the surface, whereas an opposite trend is observed for the P residues.[12–14] The interior contains P residues that are clustered in groups to optimize their electrostatic interactions. Therefore, the H residues are clustered as well,[15] and even though the H residues only seek to avoid the contact with water, effectively they can be viewed as attracting each other. This picture is the basis for the HP model proposed by Dill.[17,18] Moreover, at least one H cluster should span most of the core, because if all the H clusters were localized (i.e. each of them surrounded by P residues), it would mean that a chain of blobs would provide the most stable solution rather than a compact structure. One objective of this work is to examine whether core-size H clusters exist in the interior of proteins.

However, our main objective is to explain the experimental fraction f of H residues in terms of percolation theory that, as argued later, provides a relation between f, the clustering of H residues, and hence the compactness of protein structures. In its basic form, this theory is developed for a simple experiment carried out on the sites of a large empty lattice (say, a square lattice) as follows:[19,20] each site is visited and a particle is placed there with probability p or the site remains vacant with probability $1-p$ (using a random number). After completing this experiment for the entire lattice, one asks whether percolation has occurred, i.e. whether a cluster of occupied sites connecting (bridging) the opposite sides of the lattice has been created. It has been shown that for each lattice a critical probability, p_c, exists (called the site percolation threshold), where p_c is the minimal probability such that for $p \geq p_c$, percolation will always occur.[19] For a square lattice $p_c \approx 0.59$, and p_c deceases as the coordination number of the lattice increases; thus, $p_c \approx 0.31$, 0.25 and 0.18 for simple cubic, body-centered cubic, and face-centered cubic (fcc) lattices, respectively (see Figure 2).[19] We seek to establish a connection between f and p_c.

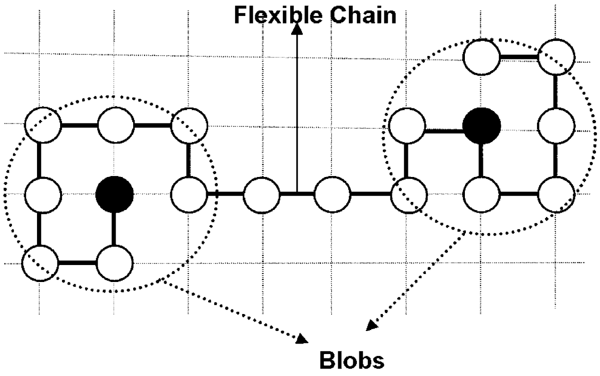

Figure 1. A schematic lattice illustration of a typical arrangement expected for a protein with a low fraction f ($f \ll 1$) of hydrophobic residues (filled circles). To avoid the contact with water these residues are "wrapped" locally by hydrophilic residues (open circle) to form "blobs". This chain of flexible blobs gains further stability from its high entropy.

Previously, we took the first step in this direction,[21] by applying percolation theory to the simplified HP model.[17,18] In the HP model, a protein is described by a self-avoiding chain on a lattice consisting of N monomers (i.e. N–1 bonds) of two kinds, H and P. Two non-bonded H monomers that are nearest neighbors on the lattice interact with an attractive energy ε ($\varepsilon = -|\varepsilon|$), where the interaction of PP and HP contacts is zero. Thus, for a given distribution of the H monomers, the ground state (which might be degenerate) is a chain configuration with the lowest possible energy, $E = n_{\max}\varepsilon$, where $n_{\max}$ is the maximal number of HH contacts. Assuming that the H-monomers are distributed at random along the chain, we have asked the following question: what should be their minimal fraction, f that would lead to a compact collapsed ground state?

To answer this question, we first considered[21] a self-avoiding chain consisting only of P monomers and arranged it in a perfect compact structure (a square shape in Figure 3); then, each of the P monomers was visited and changed to an H monomer with probability f. This process is exactly a percolation experiment that for $f \geq p_c$ would lead to a percolating cluster of the H monomers over the compact (square) chain structure, where in this case p_c is the percolation threshold of both the perfect compact structure and the square lattice. However, the cluster is not symmetric, in the sense that only contacts of H monomers that reside on parallel (or anti-parallel) segments of the chain (horizontal HH contacts in Figure 3) contribute to the energy (hence to the stability of the structure), while HH contacts along the chain (perpendicular in Figure 3) do not contribute to the energy (see further discussion in Methods). Elsewhere, we have argued that such a percolating cluster is of low energy, "holding" the perfect compact structure together;[21] thus, we have suggested that $f \approx p_c$ is approximately the minimal

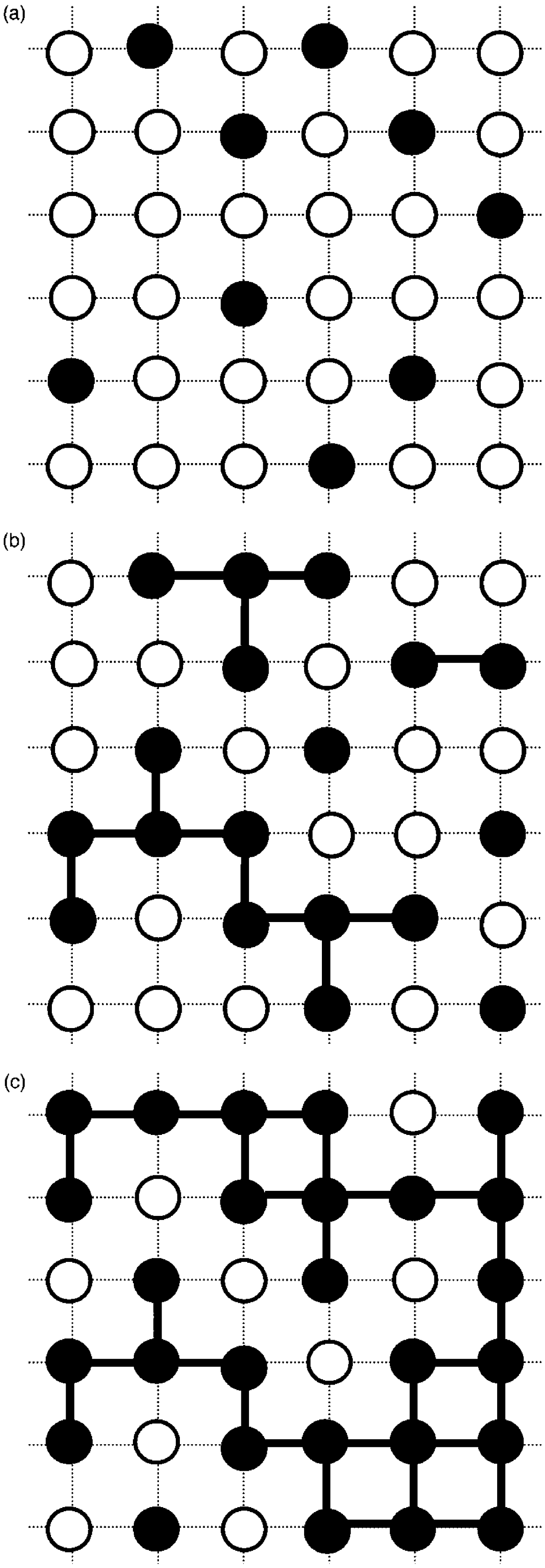

Figure 2. Percolation experiments on a square lattice populated initially by empty circles. Each lattice site is visited and a full circle replaces the empty one with probability p using a random number. Two nearest-neighbor full circles on the lattice are defined to be connected, which is illustrated by a bold-faced line. (a) For $p=0.25$, all the full circles are isolated, i.e. no cluster has been created. (b) For $p=0.50$, the number of full circles increases and three local clusters are observed. (c) For $p=0.75$, which is larger than the percolation threshold, 0.59, a percolation cluster is created, connecting the opposite sides of the lattice.

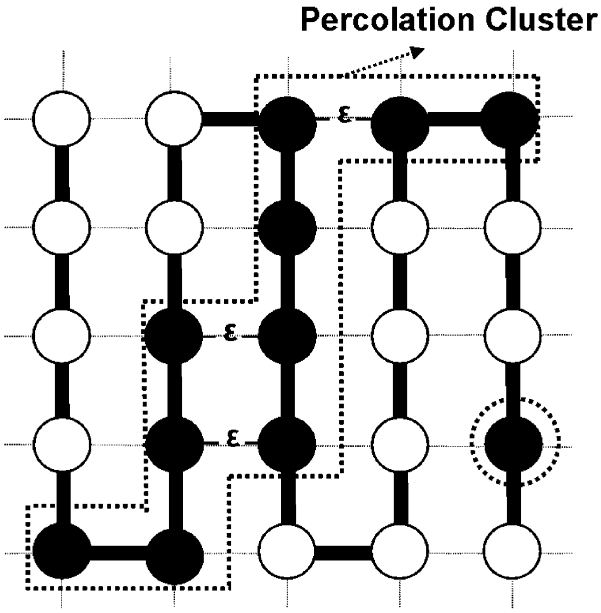

Figure 3. A perfect compact structure (square) of a self-avoiding HP chain on a square lattice. Filled and open circles stand for hydrophobic (H) and polar (P) monomers, respectively. Non-bonded, nearest-neighbor H monomers interact with negative energy $\varepsilon=-|\varepsilon|$. This perfect compact chain structure has total energy, $E=3\varepsilon$. Two nearest-neighbor H monomers (bonded or non-bonded) are defined to be connected. The Figure displays an isolated H monomer and a percolating cluster of H monomers that bridges the opposite sides of the chain structure.

fraction of H monomers required to guarantee a collapsed ground state. Indeed, simulations of the HP model on square and simple cubic lattices at low temperatures have supported this idea. Elsewhere, we have given heuristic arguments that this picture also applies to globular proteins,[21] and here the connection between the HP model and real proteins is established further.

While for $f=p_c$ the perfect compact structure introduced above is of low energy, in most cases this energy would not be the lowest possible for the given sequence, i.e. this structure is not the ground state with the maximal number of HH contacts (see Figure 4). Typically, the ground state is characterized by a higher concentration of H monomers in the interior than in the periphery (surface) and an opposite distribution of the P monomers. Correspondingly, the chain loses its perfect compact

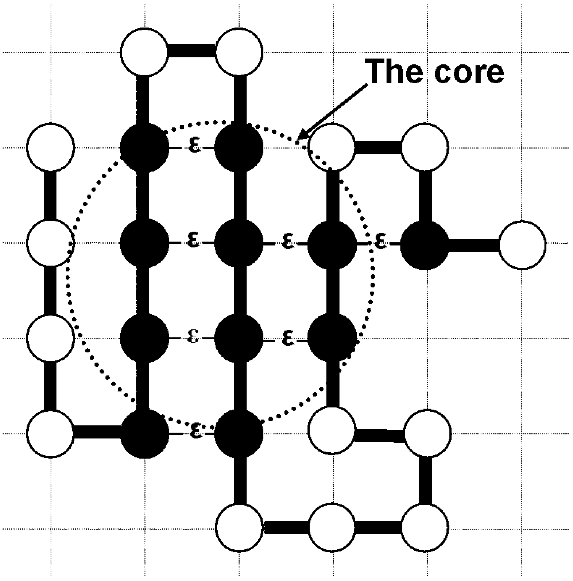

Figure 4. The ground state of the chain depicted in Figure 3. The energy decreases significantly to $E=7\varepsilon$. This minimal energy is achieved by increasing the concentration of the H monomers in the interior, maximizing thereby the number of HH contacts; as a result all the P monomers move to the surface. However, the chain loses its perfect compact shape acquiring a ramified periphery, while its interior core still has the maximal density, which is the characteristic of the perfect compact structure. The dotted circle represents the spherical core, and its size (radius) is probably not optimal. This circle does not include the left segment of four P monomers, which does not contribute to the energy, and is thus expected to be fully flexible maximizing its contribution to the entropy of the chain.

shape (square in Figure 3), which is manifested by a ramified periphery, while the maximal density, characteristic of the perfect compact structure, remains only in the interior. We call this region of maximal density, "the core" (circled in Figure 4). It should be pointed out that a percolation experiment can be carried out over compact chain configurations with a ramified surface such as the ground state in Figure 4; in this case, the site percolation threshold (p_c) will increase because of the decrease in the effective coordination number of the ramified part. However, the percolation threshold for the core alone remains the same as for the perfect compact structure. The distinction between a percolation experiment over the core and over the entire structure will become important in what follows.

We have argued earlier that, due to their clustering, the H residues in a real protein effectively attract each other; that is, they behave basically as in the HP model. Therefore, the main conclusions drawn for the HP model should apply to real proteins. Thus, a folded protein structure (i.e. an X-ray structure from the Protein Data Bank (PDB)) corresponds to the ground state of the HP model (for $f=p_c$). Like the latter, the folded structure has a ramified surface and a core defined approximately as the spherical region of highest density around the center of mass. On the other hand, the perfect compact structure of a protein is unknown. One can assume, however, that the density of such a structure would be approximately the same as that of the core. Thus, to test our hypothesis that the experimental fraction, f of randomly distributed H residues is approximately equal to the percolation threshold for the perfect compact structure, we can carry out percolation experiments based on the experimental f on the protein core.

To apply our analysis to real proteins, a simplified model of a protein is used, where an amino acid residue is represented by its α-carbon atoms and the structure is thus defined by the C^{α}–C^{α} virtual bond lengths, virtual bond angles, and virtual dihedral angles.[22,23] To keep the lattice picture alive (we shall argue that this is not mandatory), the PDB structure is best-fitted onto an fcc lattice (as described in Methods),[23] the core region is defined, and percolation experiments are performed. The size of the largest percolation cluster is compared with the core size to determine whether percolation has occurred. We identify the largest H cluster to compare its size with respect to the core size. Note the distinction between the largest H cluster, which is based on the actual distribution of the (relatively concentrated) H residues in the core, and the clusters generated by the percolation procedure based on the smaller experimental fraction f of H residues in the entire sequence. Percolation experiments are carried out for 103 single-chain proteins using six different hydrophobic sets of residues. Indeed, on average, percolating clusters are generated, which supports our theory; however, some sets lead to a better core coverage than others. We calculate the largest actual hydrophobic cluster of each protein and show that, on average, these clusters span the core, again in accord with our theory. We discuss the effect of protein size, deviations from the average picture, and implications of this study for defining reliable simplified models of proteins.

Our analysis is based on the assumption that the H residues are distributed at random over the sequence. Indeed, an early study of protein sequences by White and Jacobs supports this assumption,[24] while in a later study these authors have found a slight bias toward the creation of shorter consecutive blocks of H residues than would be anticipated from a random distribution.[25] Similar conclusions were found by Schwartz *et al.*,[26] who studied sequences of proteins that are known to fold in aqueous solutions, and by others.[27,28] Therefore, our assumption of random distribution is an approximation, which is justified within the accuracy of our approach. It should also be pointed out that Stauffer[29,30] and de Gennes[31] have applied percolation theory to describe the cluster creation in sol–gel transition in polymers.[32]

Results

Fitting a protein structure to a lattice

The first step in our calculations requires fitting the PDB X-ray structures to an fcc lattice (see Methods). Structures of 103 single-chain proteins are used in the calculations presented. In Table 1 results are shown for the root-mean-square deviation (RMSD) between the best-fitted structures of eight of the longer proteins and the corresponding X-ray structures using two different methods, systematic search (SYS) and Monte Carlo search (MCS). Table 1 reveals that the fittings obtained by MCS are slightly (but not significantly) better (i.e. with smaller RMSD) than those performed by the SYS. On the other hand, for the smaller proteins, the advantage of the systematic search increases because of the dramatic increase in the number of trials and SYS leads to slightly better results than MCM. Overall, for the 103 proteins studied, the RMSD values are less than 1.4 Å, where the smallest values were obtained for the shorter proteins.

Sets of hydrophobic amino acid residues studied

In our calculations, amino acids are defined as either hydrophobic (H) or polar (P). However, there are significant differences in the literature regarding the identity of H and P residues. Hydrophobicity has been defined either experimentally by studying the free energy of transfer of amino acids from water to organic solvents or empirically by examination of X-ray structures of proteins. The empirical approach has been carried out in various ways and is itself different in principle from the experimental approach, which does not reflect the influence of chain connectivity and other interactions.[13] In recent studies, the sets of amino acids classified as H include nine amino acids, and we shall adhere to this convention. Of the nine residues, most of the hydrophobicity sets proposed in the literature include the six residues, Val, Leu, Ile, Phe, Trp and Met but differ by the additional three residues.

Thus, according to the hydrophobicity scales suggested by Meirovitch *et al.*[13] and Wertz & Scheraga,[33] the additional three residues are Cys, His and Tyr, whereas Cys, Tyr and Ala is the triplet defined by Manavalan & Ponnuswamy,[34] and Krigbaum & Komoriya.[12] Eisenberg & McLachlan[35] have suggested Ala, Pro and Tyr, a triplet that has been used recently by Schwartz *et al.*[26] Mandel-Guetfreund & Gregoret[4] used, Ala, Gly and Pro, while according to Chothia's scale,[36] Rose *et al.*,[37] Kyte & Doolittle,[38] Janin,[39] and Wolfenden *et al.*,[40] Cys, Ala and Gly are the three additional residues. Finally, Tyr, Cys and Pro were defined by Levitt,[22] Zhou & Zhou,[41] and Sharp *et al.*;[42] this last set was also derived experimentally by Fauchere & Pliska.[43] We examine all these sets, besides the experimentally based set reported by Nozaki & Tanford,[44] because their triplet includes Lys (and Tyr and Pro). These sets are listed in Table 2, identified by the three differing residues.

Table 1. RMSD between eight protein structures fitted onto an fcc lattice and the corresponding crystal structures

PDB ID	No. residues	RMSD (Å)	
		SYS	MCS
5cpa	307	1.32	1.31
2apr	325	1.34	1.34
1pmi	440	1.40	1.38
6taa	478	1.39	1.38
3cox	507	1.41	1.38
1gal	583	1.39	1.39
7acn	754	1.42	1.39
1yge	839	1.41	1.39

The fitting was carried out by a systematic search (SYS) and a Monte Carlo search (MCS). The total number of fitting trials (rotations) is limited to 36,000. SYS runs three rounds, with $n_1=30$, $n_2=20$, $n_3=10$ ($30^3+20^3+10^3=36000$, for details, see the text). MC runs until the limit is reached.

Dataset

All of our calculations are applied to 103 single-chain proteins, most of them chosen from the set used by Tobi *et al.*,[45] which is a subset of the database accumulated by Hinds & Levitt.[46] These are proteins ranging in size from 104 to 839 residues. Their X-ray structures have been retrieved from the PDB.

Average results for the entire group of proteins

For each protein i 100 percolation experiments over the core are performed using its fraction f_i of H residues. Table 2 shows results averaged over the entire group of 103 proteins for the six hydrophobicity sets of amino acid residues identified only by their three differing residues. These sets are arranged in an increasing population of the three amino acid types in proteins. For example, in set 3 the frequently occurring Ala replaces the less frequent Pro of set 2, and correspondingly the average fraction of the nine members of set 2 in sample of 103 proteins, $\langle f_{\text{prot}}\rangle_{103}=0.37$ increases to 0.42 for set 3. Thus, the values of $\langle f_{\text{prot}}\rangle_{103}$ in Table 2 range from 0.35 to 0.49 and the average fractions of H residues in the core, $\langle f_{\text{core}}\rangle_{103}$, as expected, are larger, ranging from 0.45 to 0.57, where the corresponding differences lie between 0.07 and 0.1.

Table 2 presents results for the average size of the longest hydrophobic and percolation clusters, $\langle L_{\text{H-clust}}\rangle$ and $\langle L_{\text{perco}}\rangle$, respectively, where L is calculated with respect to the core radius, R_c, $L=$ (length of largest cluster$/2R_c$), as defined in Methods. The monotonic increase of $\langle f_{\text{prot}}\rangle_{103}$ and $\langle f_{\text{core}}\rangle_{103}$ in going from set 1–6 (discussed above) suggests that the average size of the largest hydrophobic and percolation clusters $\langle L_{\text{H-clust}}\rangle_{103}$ and $\langle L_{\text{perco}}\rangle_{103}$ will increase monotonically, as well;

Table 2. Results for the different hydrophobic sets averaged over the entire group of 103 proteins

Hydrophobic set	$\langle f_{\mathrm{prot}}\rangle_{103}$	$\langle f_{\mathrm{core}}\rangle_{103}$	$\langle L_{\text{H-clust}}\rangle_{103}$	$\langle L_{\mathrm{perco}}\rangle_{103}$
1:{Cys, His, Tyr}[13,33]	0.35	0.45	0.76 (2)	0.77 (1)
2:{Cys, Pro,Tyr}[22,41–43]	0.37	0.46	0.82 (2)	0.83 (1)
3:{Cys, Ala, Tyr}[12,34]	0.42	0.52	0.90 (2)	0.93 (1)
4:{Pro, Ala, Tyr}[26,35]	0.44	0.51	0.98 (2)	1.01 (2)
5:{Cys, Ala, Gly}[36–40]	0.47	0.56	1.08 (3)	1.05 (2)
6:{Pro, Ala, Gly}[4]	0.49	0.57	1.16 (3)	1.11 (2)

$\langle f_{\mathrm{prot}}\rangle_{103}$ and $\langle f_{\mathrm{core}}\rangle_{103}$ are the fractions of H residues in the protein sequence and in the spherical cores, respectively averaged over the group of 103 proteins. $L_{\text{H-clust}}$ = (length of largest actual H cluster)/2R_c, where R_c is the core radius; L_{perco} = (average length of the largest clusters obtained in 100 percolation experiments)/2R_c. All hydrophobicity sets share the same six residues; Ile, Leu, Val, Phe, Trp, and Met. Therefore, the sets are defined by the three differing amino acid residues. The statistical errors are (one standard deviation/($103^{1/2}$ = 10.1)). The errors of the last digit are denoted by parentheses, e.g. 0.98 (2) = 0.98 ± 0.02. The errors of $\langle f_{\mathrm{prot}}\rangle_{103}$ and $\langle f_{\mathrm{core}}\rangle_{103}$ are not larger than ±0.006.

this indeed is shown to occur, where the corresponding ranges are 0.76–1.16 and 0.77–1.11. Thus, for all of the hydrophobicity sets studied, the average of the largest hydrophobic and percolation clusters span most of the core region in accord with our theoretical expectations. It is of interest to point out that the corresponding values of $\langle L_{\text{H-clust}}\rangle_{103}$ and $\langle L_{\mathrm{perco}}\rangle_{103}$ are the same within the statistical error (one standard deviation/$103^{1/2}$), even though the actual H clusters are based on f_{core}, which is larger than f_{prot} used in the percolation experiments. This demonstrates that, on average, the H residues in the actual hydrophobic clusters are packed somewhat tighter than in the percolation clusters, as illustrated in Figures 3 and 4 for the HP model.

However, a more detailed picture about these clusters is given in Tables 3 and 4, where results are presented for sets of proteins grouped according to size, and in Table 5, where results for the 103 individual proteins are provided. Finally, if one seeks to define an optimal hydrophobicity set based on our criteria, the choice would be set 4, where $\langle L_{\text{H-clust}}\rangle_{103} \approx \langle L_{\mathrm{perco}}\rangle_{103} \approx 1$; therefore, the results in Table 5 were calculated with set 4.

Average results for proteins grouped according to size

As pointed out above, to examine the effect of protein size on cluster size the sample of 103 proteins was divided into five groups according to chain length, from 100–200, 200–300, 300–400, 400–500 and greater than 500. Table 3 provides the number of proteins in each group (at least 13) and information about the core, which will be used for a later analysis. The density of C^{α} atoms within the spherical core, R_c, and within a sphere of the radius of gyration, R_G, is larger for the shorter proteins (groups 100–200 and 200–300) than for the longer ones, a fact that has been pointed out by others.[47] These results do not stem from changes in the average core radius, $\langle R_c/R_G\rangle$, which is shown to remain approximately the same for the different groups.

Table 4 displays results for the averages, $\langle f_{\mathrm{prot}}\rangle$, $\langle f_{\mathrm{core}}\rangle$, $\langle L_{\text{H-clust}}\rangle$, and $\langle L_{\mathrm{perco}}\rangle$ for the six hydrophobicity sets, as well as for the five groups (denoted i, i = 1,5) of proteins of increasing size. The Table reveals that for each hydrophobicity set, the five values of $\langle f_{\mathrm{prot}}\rangle_i$ basically remain unchanged. In contrast, $\langle f_{\mathrm{core}}\rangle_i$ (except for sets 4 and 6) shows a tendency to decrease as i increases, i.e. the fraction of H residues in the core increases with increasing protein length. This probably is a result of the difficulty to protect the hydrophobic side-chains from the contact with water as the protein size decreases, which leads for the smaller proteins to a higher density of C^{α} atoms of H residues in the protein core; this picture is in accord with the higher core densities shown in Table 3 (and discussed above) for the smaller proteins (100–200 and 200–300). The somewhat different behavior of $\langle f_{\mathrm{core}}\rangle_i$ for sets 4–6 stems, in particular, from the ambivalent character of Ala and Gly in smaller and larger proteins. Thus, for the smaller proteins these residues have been found to exhibit a hydrophilic character, i.e. their distance from the center of mass is, on average, relatively large, comparable to the distance of typical hydrophilic residues; on the other hand, for the larger proteins, Ala and Gly are

Table 3. Average core size and core density for the different groups of proteins

Protein groups by no. residues	Proteins in group (n)	Average R_G (Å)	Average R_c (Å)	$\langle R_c/R_G\rangle$	Core density ($\times 10^{-4}$)	R_G density ($\times 10^{-4}$)
100–200	41	14.6 (3)	12.8 (3)	0.88 (1)	67 (2)	60 (2)
200–300	14	16.6 (3)	14.3 (5)	0.86 (3)	73 (1)	67 (2)
300–400	15	20.1 (3)	16.8 (2)	0.84 (1)	61 (1)	53 (2)
400–500	20	22.7 (3)	19.1 (3)	0.84 (2)	58 (2)	51 (2)
>500	13	24.7 (6)	21.3 (7)	0.87 (2)	60 (3)	54 (3)

R_c and R_G are the core radius and the radius of gyration, respectively. The density is the number of C^{α} atoms divided by the spherical volume (in Å^3). The statistical errors are (one standard deviation/$n^{1/2}$); see the legend to Table 2.

Table 4. Results for the different hydrophobicity sets and protein groups

Protein groups by number of residues	$\langle f_{\text{prot}}\rangle$	$\langle f_{\text{core}}\rangle$	$\langle L_{\text{H-clust}}\rangle$	$\langle L_{\text{perco}}\rangle$
Set 1: {Cys, His, Tyr}				
100–200	0.34 (1)	0.48 (1)	0.78 (4)	0.76 (2)
200–300	0.35 (1)	0.46 (2)	0.79 (5)	0.88 (4)
300–400	0.35 (1)	0.44 (2)	0.84 (5)	0.77 (4)
400–500	0.36 (1)	0.42 (1)	0.70 (5)	0.74 (2)
>500	0.35 (1)	0.40 (1)	0.63 (4)	0.69 (3)
Set 2: {Cys, Pro, Tyr}				
100–200	0.36 (1)	0.48 (1)	0.84 (4)	0.85 (3)
200–300	0.37 (1)	0.48 (2)	0.86 (7)	0.85 (3)
300–400	0.38 (1)	0.46 (1)	0.87 (8)	0.82 (3)
400–500	0.39 **(1)**	0.44 (1)	0.75 (5)	0.77 (2)
>500	0.38 (1)	0.42 (1)	0.75 (5)	0.74 (3)
Set 3: {Cys, Ala, Tyr}				
100–200	0.41 (1)	0.55 (1)	0.91 (4)	0.93 (3)
200–300	0.42 **(1)**	0.54 (1)	1.03 (6)	1.07 (4)
300–400	0.42 (1)	0.50 (1)	0.93 (7)	0.93 (3)
400–500	0.43 **(1)**	0.48 (1)	0.87 (5)	0.89 (2)
>500	0.41 (1)	0.46 (1)	0.77 (6)	0.81 (3)
Set 4: {Pro, Ala, Tyr}				
100–200	0.43 (1)	0.51 (1)	0.98 (5)	0.95 (4)
200–300	0.44 (1)	0.52 (1)	1.09 (6)	1.11 (5)
300–400	0.45 (1)	0.51 (1)	1.01 (7)	1.00 (3)
400–500	0.46 (1)	0.51 (1)	0.94 (5)	1.05 (2)
>500	0.46 (1)	0.49 (1)	0.86 (7)	0.99 (4)
Set 5: {Cys, Ala, Gly}				
100–200	0.46 (1)	0.58 (1)	1.07 (5)	1.04 (4)
200–300	0.48 (1)	0.61 (2)	1.22 (5)	1.22 (4)
300–400	0.48 (1)	0.55 (2)	1.12 (5)	1.07 (3)
400–500	0.47 (1)	0.53 (1)	1.04 (4)	0.99 (2)
>500	0.46 (1)	0.52 (1)	0.95 (6)	0.94 (4)
Set 6: {Pro, Ala, Gly}				
100–200	0.47 (1)	0.57 (1)	1.13 (5)	1.07 (4)
200–300	0.49 (1)	0.61 (2)	1.29 (7)	1.26 (4)
300–400	0.51 (1)	0.58 (1)	1.24 (9)	1.15 (4)
400–500	0.50 (1)	0.56 (1)	1.17 (6)	1.09 (3)
>500	0.50 (1)	0.56 (1)	1.03 (9)	1.06 (5)

f_{prot}, f_{core}, $L_{\text{H-clust,}}$ and L_{perco} are defined in the legend to Table 2. All the hydrophobicity sets share the same six residues; Ile, Leu, Val, Phe, Trp and Met. Therefore, the sets are defined by the three differing amino acid residues. The statistical errors are one standard deviation/$n^{1/2}$; see the legend to Table 2. Boldfaced errors, **(1)** are smaller than ±0.01.

distributed in the interior as typical H residues.[13,14] This tendency is reflected in sets 5 and 6, where $\langle f_{\text{core}}\rangle_2$ (i.e. for proteins of size 200–100) is slightly larger than $\langle f_{\text{core}}\rangle_1$. Note that a behavior similar to that of Ala and Gly (even though less drastic) was found for Pro, and an opposite behavior for Tyr, while the average distance of Cys from the center of mass is the same for smaller and larger proteins. These tendencies determine the almost constant $\langle f_{\text{core}}\rangle_i$ values obtained for sets 4 and 6.

The highest core density and the largest fraction of H residues in the core found for the shorter proteins explains the tendency of both $\langle L_{\text{H-clust}}\rangle$ and $\langle L_{\text{perco}}\rangle$ to decrease with increasing protein size (see also Figures 5 and 6, for $\langle L_{\text{H-clust}}\rangle$ and $\langle L_{\text{perco}}\rangle$, respectively). In most cases these values are maximal for the proteins of size 200–300, which have the largest core density (Table 3), and the highest $\langle f_{\text{core}}\rangle$ for sets 5 and 6. Table 4 shows that, in general, the corresponding results for $\langle L_{\text{H-clust}}\rangle$ and $\langle L_{\text{perco}}\rangle$ are close to each other, while one would expect that the H clusters that are based on $\langle f_{\text{core}}\rangle$ would be larger than the percolation clusters that were generated with the smaller $\langle f_{\text{prot}}\rangle$. On the other hand, the actual H clusters tend to be more packed than the percolation clusters, as illustrated in Figures 3 and 4. In any case, the cluster sizes are characterized by relatively large statistical errors that are larger for $\langle L_{\text{H-clust}}\rangle$ than for $\langle L_{\text{perco}}\rangle$ because the former is based on a single cluster per protein, where the latter is an average over 100 percolation experiments for each protein. These errors mean that strong deviations from the average picture are expected for individual proteins (see below). The results reported in Table 4 for $\langle L_{\text{H-clust}}\rangle$ and $\langle L_{\text{perco}}\rangle$ for all the hydrophobicity sets averaged over ten groups of proteins of increasing size are displayed in Figures 5 and 6, respectively. The visualization of Table 4 in these Figures emphasizes that the results for $\langle L_{\text{perco}}\rangle$ are fluctuating less than those for $\langle L_{\text{H-clust}}\rangle$.

It should be pointed out that we could have adopted a different type of analysis in which the percolation threshold, $p_c(i)$ is determined for each protein i and the averages and fluctuations of $p_c(i)$ over the five groups of proteins are calculated and compared to the corresponding $\langle f_{\text{prot}}\rangle$ and $\langle f_{\text{core}}\rangle$ values. This kind of analysis is expected to lead to

Table 5. Results based on hydrophobicity set 4 for the individual proteins arranged in five groups of increasing size

Protein name	PDB	$N_{\mathrm{AA}}{}^{\mathrm{a}}$	$R_{\mathrm{c}}/R_{\mathrm{G}}$	f_{prot}	f_{core}	$L_{\mathrm{H\text{-}clust}}$	L_{perco}
Ribonuclease T_1 (EC 3.1.27.3)	9rnt	104	0.86	0.37	0.55	1.01	1.01
Cytochrome *c*	5cyt	104	1.02	0.37	0.47	0.87	0.71
Fk506 binding protein (Fkbp)	1fkb	107	0.86	0.42	0.52	1.46	1.15 (3)
Actinoxanthin	1acx	108	0.80	0.44	0.51	1.40	1.39
Rat oncomodulin	1rro	108	0.99	0.37	0.46	0.52	0.71
Cytochrome *c*	1ccr	112	0.93	0.41	0.46	0.74	0.85
Cytochrome c_2 (reduced)	3c2c	112	0.89	0.42	0.53	1.00	0.84
Neocarzinostatin	1noa	113	0.84	0.43	0.47	1.22	1.20
Phospholipase A2 (EC 3.1.1.4)	1ppa	121	0.80	0.32	0.35	0.55	0.72
α-Lactalbumin	1alc	123	0.80	0.39	0.47	0.79	0.87
Pseudoazurin	1paz	123	1.05	0.51	0.66	1.12	1.08
Ribonuclease A	1rat	124	0.99	0.35	0.42	0.94	0.77
CheY	3chy	128	1.05	0.51	0.60	1.09	0.80
Heroin esterase	1lz1	130	0.81	0.41	0.57	0.98	0.91
Prophospholipase A2	4bp2	130	0.80	0.32	0.39	0.57	0.69
Hemoglobin (erythrocruorin, aquo met)	1eca	136	0.84	0.49	0.55	1.08	0.96
Endonuclease V (EC 3.1.25.1)	2end	138	0.89	0.47	0.52	0.51	0.85
Flavodoxin	2fox	138	1.07	0.42	0.59	0.75	0.80
Apolipoprotein-E3 (LDL receptor binding domain)	1lpe	144	0.80	0.42	0.47	0.60	0.78
Basic fibroblast growth factor (hbFGF)	4fgf	146	0.85	0.40	0.53	0.87	1.10
Myoglobin (met) (pH 7.0)	1mba	147	0.94	0.55	0.57	1.23	1.02
Hemoglobin (deoxy)	2hbg	147	0.83	0.52	0.64	1.13	1.17
Hemoglobin V (cyano, met)	2lhb	149	0.86	0.51	0.62	0.95	0.97
Staphylococcal nuclease (EC 3.1.33.1)	2sns	149	0.87	0.42	0.49	1.11	0.99
Sindbis virus capsid protein	2snv	151	0.81	0.40	0.52	1.21	1.05
Ubiquitin conjugating enzyme	2aak	152	0.80	0.47	0.59	1.40	1.09
Leghemoglobin (aquo, met)	1lh2	153	0.83	0.52	0.59	1.05	1.08
Selenomethionyl ribonuclease H (EC 3.1.26.4)	1rnh	155	0.80	0.39	0.46	1.13	1.01
Dihydrofolate reductase (EC 1.5.1.3)	3dfr	162	0.87	0.47	0.56	0.64	1.10
Troponin-C	5tnc	162	0.86	0.39	0.45	0.42	0.49
Lysozyme (EC 3.2.1.17) (high salt)	4lzm	164	0.80	0.44	0.49	0.49	0.80
Flavodoxin (oxidized form)	1ofv	169	0.80	0.41	0.51	1.35	1.11
Erythrina trypsin inhibitor (Kunitz) De-3	1tie	172	0.80	0.43	0.51	1.47	1.28
Flavodoxin	2fcr	173	0.88	0.45	0.59	1.27	1.11
γ-B Crystallin (previously –II crystallin)	4gcr	174	0.80	0.41	0.47	0.60	0.97
Proteinase A (SGPA)	2sga	181	0.96	0.38	0.43	0.82	1.00
Retinol binding protein	1rbp	182	0.90	0.42	0.48	1.34	1.10
Guanylate kinase (EC 2.7.4.8)	1gky	187	1.02	0.42	0.49	1.01	0.71
Dihydrofolate reductase (EC 1.5.1.3)	8dfr	189	0.82	0.47	0.56	1.07	1.14
Adenylate kinase (EC 2.7.4.3)	3adk	195	0.92	0.39	0.52	0.99	0.82
α-Lytic protease (EC 3.4.21.12)	2alp	198	0.80	0.40	0.43	1.32	1.19
Papain Cys25 with bound atom	1ppn	212	0.89	0.44	0.53	0.94	1.05
Actinidin (sulfhydryl proteinase)	2act	220	0.85	0.43	0.54	1.35	1.06
β-Trypsin	5ptp	223	0.83	0.39	0.45	0.95	1.04
Trypsin (SGT) (E.C. 3.4.21.4)	1sgt	223	0.81	0.45	0.52	1.05	1.28
Trypsin (orthorhombic, 2.4 M ammonium sulfate)	2ptn	223	0.80	0.39	0.46	0.95	1.07
Tonin	1ton	235	0.84	0.44	0.52	1.05	1.22
Native elastase (EC 3.4.21.11)	3est	240	0.80	0.43	0.49	1.28	1.23
γ-Chymotrypsin (EC 3.4.21.1)	8gch	244	0.90	0.43	0.54	1.16	1.17
Carbonic anhydrase II (carbonate dehydratase)	2ca2	259	1.08	0.43	0.58	0.91	0.87
Triacylglycerol acylhydrolase (EC 3.1.1.3)	4tgl	269	0.80	0.45	0.51	0.90	1.22
d-Ribose-binding protein complex with β-d-ribose	2dri	271	0.80	0.47	0.49	1.71	1.14
Thermitase (EC 3.4.21.66)	1thm	279	1.08	0.45	0.56	0.86	0.88
Proteinase K (EC 3.4.21.14)	2prk	279	0.82	0.42	0.53	1.04	1.15
Rhodanese (EC 2.8.1.1)	1rhd	293	0.80	0.47	0.53	1.12	1.09
Elastase (EC 3.4.24.26) (zinc metalloprotease)	1ezm	301	0.89	0.43	0.50	0.76	0.84
l-Arabinose-binding protein	1abe	306	0.80	0.47	0.49	1.06	1.07
Carboxypeptidase Aα (Cox) (EC 3.4.17.1)	5cpa	307	1.06	0.44	0.54	0.83	0.86
d-Galactose d-glucose-binding protein	2gbp	309	0.80	0.47	0.53	0.88	1.02
NADP^+ oxidoreductase (ferredoxin reductase)	1fnb	314	0.82	0.44	0.51	1.00	1.17
Bira bifunctional protein (EC 6.3.4.15)	1bia	321	0.80	0.51	0.56	1.21	0.99
Annexin V	1ala	321	0.80	0.43	0.44	0.68	0.75
Acid proteinase (penicillopepsin) (EC 3.4.23.20)	3app	323	0.83	0.40	0.47	0.80	0.98
Acid proteinase (rhizopuspepsin) (EC 3.4.23.6)	2apr	325	0.83	0.42	0.48	0.82	0.98
Pepsin (EC 3.4.23.1)	4pep	326	0.84	0.44	0.56	1.68	1.06
Renin (EC 3.4.23.15)	2ren	340	0.80	0.45	0.53	1.39	1.10
Leucine/isoleucine/valine-binding protein (LIVBP)	2liv	344	0.81	0.46	0.48	1.06	0.95
Reca protein (EC 3.4.99.37)	2reb	352	0.82	0.46	0.49	0.97	0.95
Pepsin (renin) (EC 3.4.23.23)	1mpp	361	0.84	0.43	0.54	0.92	1.02
Elongation factor Tu (domain I)	1etu	379	0.80	0.48	0.56	1.09	1.32
Cytochrome P450Cam (camphor monooxygenase)	2cpp	405	0.85	0.48	0.51	0.97	1.04
Glycinamide ribonucleotide synthetase	1gso	411	0.95	0.50	0.54	0.67	1.10
Cytochrome P450-Cam (EC 1.14.15.1)	1phd	414	0.84	0.48	0.54	0.89	1.09

Table 5 (*continued*)

Protein name	PDB	N_{AA}[a]	R_c/R_G	f_{prot}	f_{core}	$L_{H\text{-}clust}$	L_{perco}
3-Phosphoglycerate kinase	16pk	415	0.80	0.45	0.45	0.89	0.89
Phosphoglycerate kinase (EC 2.7.2.3)	3pgk	416	0.80	0.47	0.51	1.06	0.89
Phosphomannose isomerase	1pmi	440	0.82	0.44	0.52	0.78	1.12
Pancreatic lipase-related protein 2	1bu8	446	0.80	0.42	0.43	0.78	0.91
NADH peroxidase (EC 1.11.1.1)	1npx	447	0.84	0.48	0.52	1.51	0.98
N-(5′ Phosphoribosyl)anthranilate isomerase	1pii	452	0.87	0.50	0.54	1.01	0.88
Tetanus neurotoxin	1a8d	452	0.80	0.45	0.57	0.95	1.48 (4)
Sulfite reductase hemoprotein	1aop	452	0.80	0.45	0.43	0.67	1.01
Glucoamylase-471	1gai	453	0.94	0.45	0.53	0.63	1.06
para-Nitrobenzyl esterase	1qe3	467	0.86	0.50	0.62	1.28	1.46
α-Amylase	1vjs	469	0.84	0.43	0.52	1.32	1.15
Glucoamylase-471 (1,4-α-d-glucan glucohydrolase)	3gly	470	0.96	0.44	0.51	0.38	0.78
Photolyase (DNA cyclobutane dipyrimidine photoly.)	1qnf	475	0.82	0.51	0.48	0.75	0.89
α-Amylase (Taka-amylase) (EC 3.2.1.1)	6taa	478	0.81	0.45	0.54	1.03	0.93
Chondroitinase B	1dbg	481	0.80	0.45	0.42	1.20	1.03
Amylase	1smd	495	0.80	0.42	0.49	0.93	1.07
Cholesterol oxidase	1b4v	498	0.88	0.47	0.52	1.12	1.24
Cholesterol oxidase (E.C. 1.1.3.6)	3cox	507	0.92	0.45	0.51	0.99	0.98
5′-Nucleotidase (UDP-sugar hydrolase)	1ush	515	0.85	0.45	0.46	0.92	1.00
Phosphoenolpyruvate carboxykinase	1ayl	532	0.91	0.45	0.50	1.04	0.96
Acetylcholinesterase	2ace	537	0.94	0.46	0.53	1.15	0.92
Flavocytochrome *c*3	1qjd	568	0.85	0.43	0.44	1.18	0.92
Vanadium chloroperoxidase	1vns	574	0.80	0.50	0.55	0.92	1.38
DNA polymerase I	1xwl	580	0.83	0.49	0.54	0.81	1.05
Glucose oxidase (EC 1.1.3.4)	1gal	583	0.90	0.46	0.49	0.79	1.03
Soluble lytic transglycosylase Slt70	1qsa	618	0.98	0.47	0.50	0.40	0.58
Galactose oxidase (EC 1.1.3.9) (pH 4.5)	1gof	639	0.82	0.41	0.46	0.93	1.06
Prolyl oligopeptidase (prolyl endopeptidase)	1qfm	705	0.85	0.45	0.47	0.71	1.07
Aconitase (EC 4.2.1.3)	7acn	754	0.80	0.43	0.46	0.72	0.96
Lipoxygenase-1	1yge	839	0.80	0.47	0.52	0.67	0.99

f_{prot}, f_{core}, $L_{H\text{-}clust}$, and L_{perco} are defined in the legend to Table 2. For each protein, the statistical error of L_{perco} calculated from 100 percolation experiments is one standard deviation/10. For most proteins this error is not larger than ±0.02; for several proteins, the error is larger and it appears in the Table according to the convention defined in the legend to Table 2.

[a] The number of amino acid residues.

the same conclusion drawn above from Tables 3 and 4, that hydrophobicity set 4 is the optimal set. The advantage of the present analysis is that $\langle L_{perco}\rangle$ can be compared with $\langle L_{H\text{-}clust}\rangle$.

Results for the individual proteins

As discussed above, some of the individual proteins are expected to show significant deviations from the average picture presented in Tables 3 and 4. One reason for these deviations is the fact that our cluster definition is tailored for a globular protein with approximately constant density in the interior, conditions that are never satisfied completely, especially by the shorter proteins. Therefore, in Table 5 we present results for R_c, f_{prot}, f_{core}, $L_{H\text{-}clust}$, and L_{perco} for the 103 individual proteins, based on hydrophobicity set 4, which has been found to best satisfy our criterion, $\langle L_{H\text{-}clust}\rangle_{103} \approx \langle L_{perco}\rangle_{103} \approx 1$. Our main interest is to examine the strongest deviations, in particular the smallest values of $L_{H\text{-}clust}$, and L_{perco} that are related to clusters that span only part of the core in contrast to our theoretical considerations.

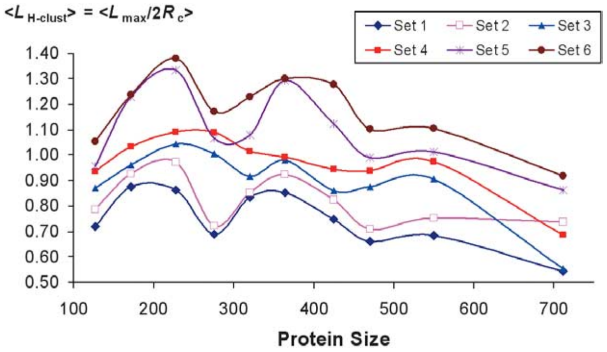

Figure 5. Results for $\langle L_{H\text{-}clust}\rangle$ averaged over ten groups of proteins of increasing size, calculated for the six hydrophobicity sets.

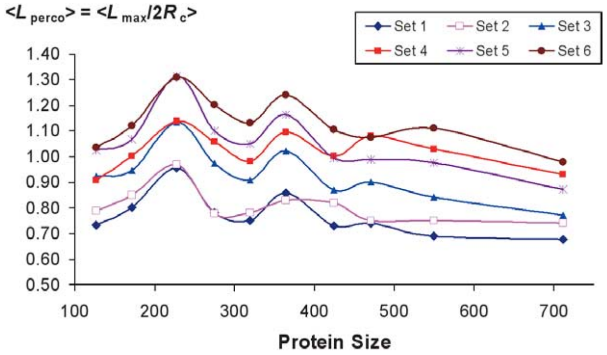

Figure 6. Results for $\langle L_{perco}\rangle$ averaged over ten groups of proteins of increasing size, calculated for the six hydrophobicity sets.

Low values of $L_{\text{H-clust}}$, (smaller than 0.61) are found in the nine proteins, 1rro (108), 1ppa (121), 4bp2 (130), 2end (138), 1lpe (144), 5tnc (162), 4lzm (164), 4gcr (174), and 3gly (470), where, as expected most of them (eight) belong to the group of smallest size, $N=100$–200. However, $L_{\text{perco}}<0.61$ is found in only two of these proteins, which is in accord with the typical smaller fluctuations in $\langle L_{\text{perco}}\rangle$ than in $\langle L_{\text{H-clust}}\rangle$, discussed above. In an attempt to understand these small L values, we inspected graphical visualizations of these structures provided by the PDB, and indeed have found that most of them (seven) deviate from a globular shape and/or lack homogeneous core density. Thus, 5tnc ($N=162$) (Figure 7(a)) is extremely elongated, while 4bp2 ($N=130$), 2end ($N=138$) (Figure 7(b)), and 1lpe ($N=144$) are moderately elongated, having elliptical shapes. It should be pointed out that the percolation threshold for elongated structures increases (becoming 1 for a rod); indeed, for all these proteins L_{perco} is also small, even though only for 5tnc and 1qsa it is smaller than 0.61. The structures of 4lzm ($N=164$) and 1aop ($N=452$) (Figure 7(c)) consist of two parts, while that of 1qsa ($N=618$) (Figure 7(d)) seems to have holes in its interior. Evaluating the remaining two structures would require a more detailed analysis.

For 16 proteins, L_{perco} or $L_{\text{H-clust}}$ are larger than 1.29. Nine of these proteins are small ($N<300$) and in none of them is $L_{\text{perco}}>L_{\text{H-clust}}$, while this relation is found in four of the seven larger proteins. Graphical visualization did not reveal drastic deviations from globularity. It should be pointed out that these relatively large clusters are defined because we allow for clusters started in the core to "grow" outward (see Methods); in some cases the structure

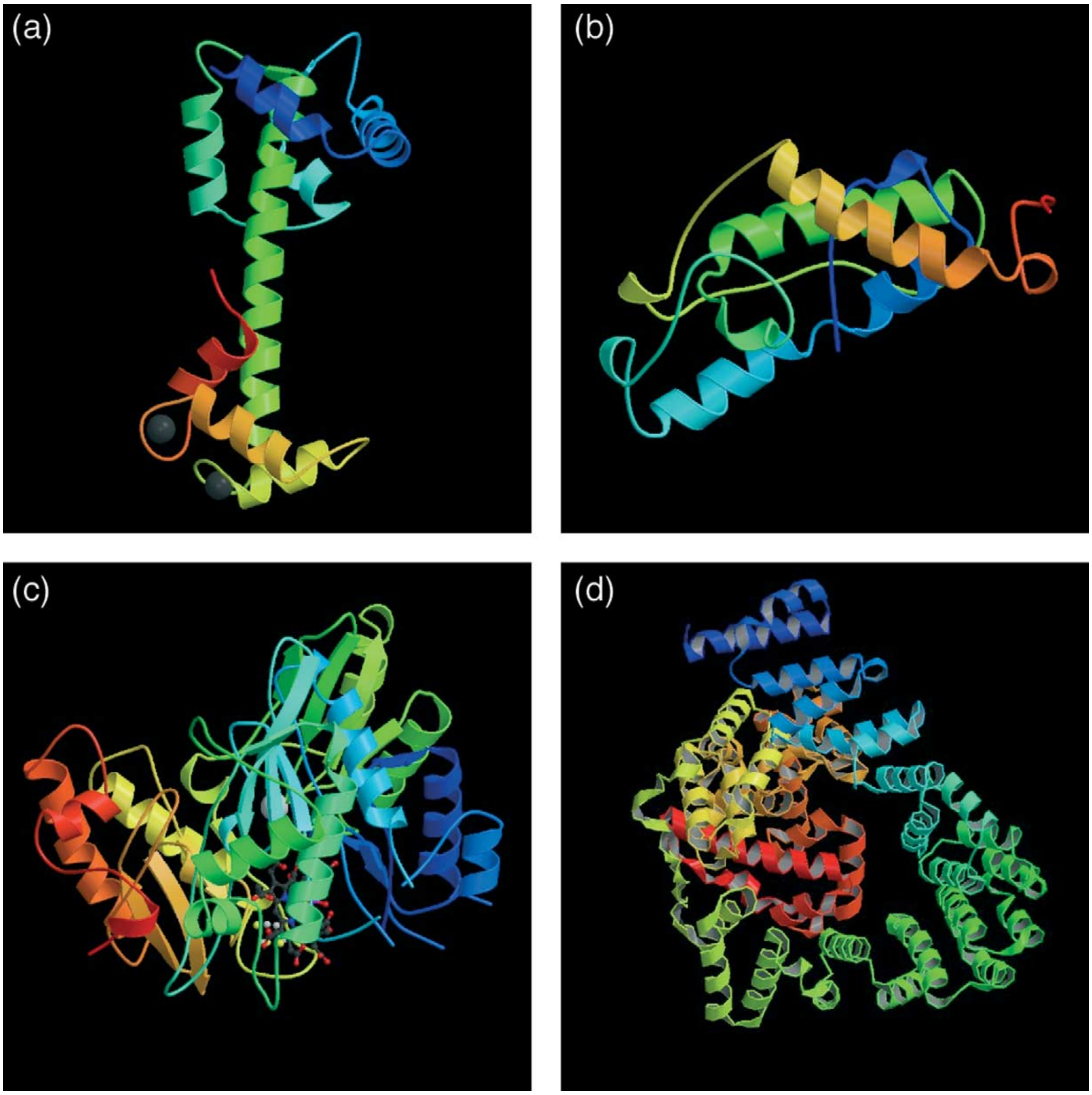

Figure 7. Graphical representation of protein structures with relatively small $L_{\text{H-clust}}$ values (see Table 5). The structures were taken from the PDB. (a) Highly elongated protein (5tnc, $N=162$). (b) Elliptical protein (2end, $N=138$). (c) Two-domain protein (1aop, $N=452$). (d) Protein with a "hole" (1qsa, $N=618$).

will return immediately to the core, whereas in others it will expand towards the periphery.

Summary and Discussion

In this work we have asked the question: what is the optimal fraction f of hydrophobic (H) residues required to ensure protein collapse? We have argued that an f that is too small is expected to lead to a random coil chain of blobs, while for f that is large enough the only thermodynamic alternative to avoid the contact of the H residues with water is burying them in the interior of a compact chain structure; if f is too large, the molecule will precipitate. Indeed, the fraction of H residues in the interior of proteins is known to be larger than their fraction (f) in the sequence. However, the interior contains hydrophilic residues that cluster in groups to optimize their electrostatic interactions. This means that the H residues are clustered as well, and therefore effectively can be viewed as attracting each other, as in Dill's HP model. Moreover, at least one hydrophobic cluster should span most of the core, because if all of the H clusters were localized, it would mean that a chain of blobs would provide the most stable solution rather than a compact structure.

Previously, we have argued that to ensure a collapsed ground state for randomly distributed H-monomers in the HP model, the minimal fraction f of H-monomers should be approximately equal to the percolation threshold of the lattice.[21] Because the H monomers in the HP model and the H residues in real proteins behave similarly, we argue here that the actual value, f, of a protein is approximately equal to the percolation threshold of the protein core. Therefore, percolation experiments based on f should lead to percolating clusters over the core, i.e. clusters of the core size.

To test our hypothesis, a simplified model of a protein was used, where an amino acid residue is represented by its α-carbon atoms and the structure is thus defined by the C^{α}–C^{α} virtual bond lengths, virtual bond angles, and virtual dihedral angles. These models of the PDB structures were best-fitted onto an fcc lattice. It should be pointed out that fitting of structures to an fcc lattice has been applied to conform with the lattice picture, the HP model and the usual percolation theory; however, this is not necessary, since by defining a nearest-neighbor distance one could define clusters directly over the X-ray structure of C^{α} atoms and carry out percolation experiments as well. Therefore, embedding protein structures in a lattice with a large effective coordination number, such as 90 or 210 (used by Skolnick, Kolinski, and collaborators[48,49]) might improve the fittings somewhat, but the results of our analysis are not expected to change significantly. Following the best-fit process, the core region was defined, and percolation experiments were performed. We also identified the largest actual hydrophobic cluster of a protein to compare its size with the core size.

While our analysis depends on the set of hydrophobic residues used, no consensus exists about this issue and we therefore studied six sets of nine residues each, where six of these residues are shared by all sets (Val, Ile, Leu, Met, Phe and Trp) that thus differ by the additional three residues. For all sets, the percolation clusters and the actual H clusters calculated for 103 proteins were found to span, on average, most of the core in accord with our theoretical expectations. However, these results differ from set to set, where set 4 (Pro, Ala, Tyr) provides the best agreement with respect to our criteria ((core size)/(cluster size) ∼1), and sets 3 (Cys, Ala and Tyr) and 5 (Cys, Ala and Gly) give the second-best results. The effect of protein size on the average results has been discussed. In accord with expectations, we have found that the largest H clusters in proteins span the core and their average size is comparable to the average size of the percolation clusters. The contribution of these H clusters to the stability of proteins has been demonstrated in NMR experiments, where partially unfolded proteins have been shown to have clusters of residual structure that are arranged along the protein sequence and are correlated strongly with hydrophobic residues. These clusters stabilize each other, and disruption of the largest cluster that ties the core of the native protein results in loss of residual structure in all of the clusters.[50]

Our theory assumes that the proteins are globular with a homogenous dense core, which is expected to resemble the density of the unknown perfect compact structure. In reality, many proteins deviate from this ideal picture, which is one reason for the observed deviations from the expected behavior ((core size)/(cluster size)∼1), discussed here. The assumption of randomly distributed H residues inherent in this analysis is an approximation that enables us to recruit the theory of percolation and establish the connection between the fraction of H residues and collapse. The results described here suggest that for simplified lattice models of proteins to be realistic, the fraction of hydrophobic residues (with any degree of specificity) should be close to the percolation threshold of a perfect compact structure of the chain model.[51,52] An interesting question is whether the distribution of H residues and the size of percolation clusters found for single proteins change for individual chains of multi-chain proteins. We carried out calculations for 47 of the latter proteins and obtained results for $\langle R_c/R_G\rangle_{47}$, $\langle f_{\text{prot}}\rangle_{47}$, $\langle f_{\text{core}}\rangle_{47}$, $\langle L_{\text{H-clust}}\rangle_{47}$, and $\langle L_{\text{perco}}\rangle_{47}$ that are exactly the same as those obtained in Tables 2 and 3 for the 103 single-chain proteins, suggesting that most of the 47 protein chains studied fold before association. This problem will be studied in detail in future work. One would suggest also that comparing the values of $L_{\text{H-clust}}$, and L_{perco} of individual protein structures to $\langle L_{\text{H-clust}}\rangle_{103}$, and $\langle L_{\text{perco}}\rangle_{103}$, respectively, might serve as a criterion to identify misfolded protein structures. We applied this criterion to several misfolded structures of the CASP4 competition but found it to be inconclusive because of the relatively large fluctuations in the values of $L_{\text{H-clust}}$, and L_{perco}. This subject will be studied further.

Methods

Fitting protein structures onto a lattice

To conform to the lattice picture of the HP model, protein structures are fitted to a lattice. As in the work of Covell & Jernigan,[23] we use a simplified protein model where the amino acid residues are represented by their backbone α-carbon atoms connected successively by virtual bonds;[22] the structure is thus defined by the virtual bond lengths, virtual bond angles, and virtual torsion angles, where the correlations between these parameters as reflected in known protein structures have been studied by Levitt.[22] A short summary of Levitt's results is given by Covell & Jernigan, who also discuss the fitting of protein structures to lattices and show, as expected, that the fitting quality improves as the coordination number of the lattice increases, i.e. in going from a simple cubic to a body-centered cubic, and to face-centered cubic (fcc) lattice. Therefore, to obtain the best fitting, we use the fcc lattice in this work.

An fcc lattice with a cubic edge a, is characterized by a coordination number 12, i.e. each cubic vertex and center face point has 12 neighbor points of distance $a/\sqrt{2}$. However, like Covell & Jernigan, we seek to improve the fitting by increasing the coordination number of the lattice, and for that, we also consider the six points of distance a, and points of the larger distance, $a\sqrt{3/2}$; thus, a center face point has eight neighbors of the latter distance (i.e. altogether 26 neighbors) and a cubic vertex has 24 such neighbors (i.e. altogether 42 neighbors). Taking $a=3.8$ Å, the two other options for fitting a virtual C^{α}–C^{α} bond are $3.8/\sqrt{2}=2.687$ A, and $3.8\sqrt{3/2}=4.654$ A.

The fitting process begins with an X-ray structure of a protein taken from the PDB, where only the C^{α} coordinates are considered. Starting from the N terminus the α-carbon atoms are fitted successively, where a candidate lattice point for placing the $(i+1)$th C^{α} must be a vacant neighbor lattice point of the ith C^{α}, meaning that double occupancy of a lattice point is forbidden and the excluded volume interaction is thus satisfied. The chosen (best-fitted) $(i+1)$th lattice point is that with the minimum distance (squared), d^2_{i+1} from the corresponding C^{α} coordinates of the PDB structure. During the fitting process, the d^2_i values are accumulated and the root-mean-square deviation (RMSD) of the fitted structure from the actual structure is calculated. To minimize the RMSD, it is calculated for many rotations of the lattice with respect to the PDB structure defined by the Eulerian angles φ, θ, and ψ, using SYS and MCS procedures. With the systematic search, we start by dividing evenly the range [0, 2π] into n_1 values for each of the three angles, and perform the corresponding n_1^3 rotations that lead to a minimum RMSD value for φ_1, θ_1, and ψ_1. Then, $n_2(n_2<n_1)$ values are chosen evenly within a small range around each of the angles, φ_1, θ_1, and ψ_1 and a lower RMSD is obtained from the corresponding n_2^3 rotations, and so on; i.e. this fitting procedure is based on a total of $n_1^3+n_2^3+\ldots$ trial rotations. In practice, for small proteins (with less than 100 amino acid residues), n_1 usually ranges from 100 to 300, where for the larger proteins, $n_1\leq 50$ due to the high cost of computation.

In an attempt to improve the optimization for the larger proteins, we have applied a Monte Carlo procedure combined with Powell's local minimization method.[53] Thus, at step k of the process, a set of rotation angles is generated at random and the corresponding RMSD_k is calculated and compared to the minimal value RMSD_m achieved in previous steps. If RMSD_k is smaller than RMSD_m, Powell's method is used to minimize RMSD_k further. Otherwise, such a minimization is carried out with probability:

$$p=\exp[-(\text{RMSD}_k-\text{RMSD}_m)/K]$$

where $K=0.01$ is a parameter; RMSD_m is upgraded and the process continues.

Clustering procedure

After a protein is mapped onto a lattice, it can be viewed as a self-avoiding chain with a specific sequence of "beads" of two kinds, hydrophobic (H) and polar (P). Neighbor H residues (beads) on the lattice can be clustered using the following procedure: First, any un-clustered H residue is chosen and its neighboring H residues (i.e. the first shell) are identified and added to the same cluster; then the neighbor H residues of the first shell are identified and added to the cluster, etc., i.e. using a breath-first-search approach. This procedure is repeated over and over again, each time starting from any of the remaining un-clustered H residues, until each H belongs to one of the several clusters thus created.

Figure 8 provides a 2-D illustration of such clustering applied to a chain of 22 beads, where the H residues are grouped into four clusters, C_j $(j=1, 4)$. We define the concept of bond distance, $D_{j,k}$ between clusters j and k as the smallest number of bonds required to walk from any residue of cluster j to any residue of cluster k. In Figure 8,

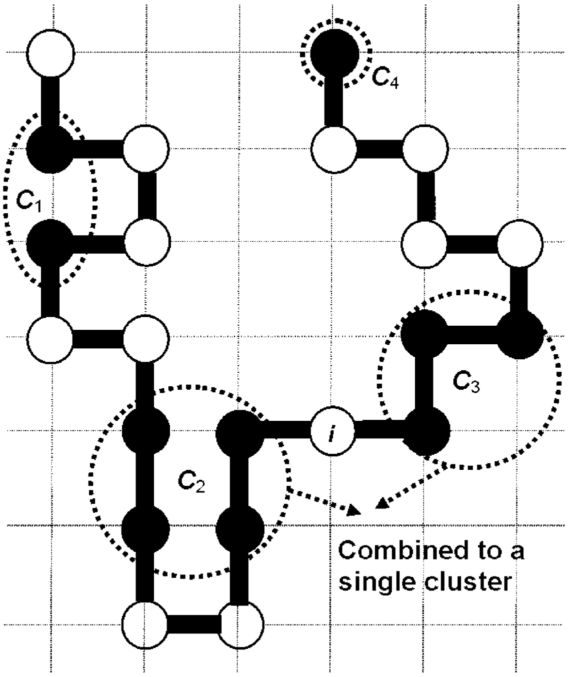

Figure 8. Hydrophobic clusters (C_j, $j=1$, 4) of a 22 residue protein on a square lattice. A filled (●) and an open (○) circle represent a hydrophobic and a polar residue, respectively. The distance between clusters C_2 and C_3 is two bonds, which is the minimal possible distance. These clusters are "weakly" connected by the polar residue i and we combine them, creating a single cluster.

$D_{1,2}=3$, $D_{2,3}=2$, $D_{3,4}=5$, $D_{1,3}=10$, etc. It is obvious that the minimum bond distance between two clusters is 2. As pointed out in Introduction, a percolation experiment over an HP chain model is not symmetric, i.e. bonded H contacts do not contribute to the energy, which is determined only by the non-bonded contacts. In this respect, the connectivity of nearest neighbors along the chain is less important for structure stability than non-bonded contacts and, therefore, we change slightly the definition of clusters for protein lattice model. Thus, if two clusters have a bond distance of only two, we consider them to be connected to each other by this bond. An example is illustrated in Figure 8, where clusters 2 and 3 that are not "strongly" connected to each other by an HH bond, become "weakly" connected by the polar residue *i* to form a single cluster. This definition is adopted in this work for a percolation cluster as well as for an actual H cluster. This definition increases the average size of the clusters by 15%.

The spherical core

After fitting the structures onto fcc lattices, their centers of mass and radii of gyration, R_G, are calculated. We calculate also the density profile around the center of mass and for $R \geq 0.8\ R_G$ determine the radius R_c as the point where the density profile starts decreasing monotonically; R_c defines the spherical core of each protein. This definition of the core is suitable for most protein structures, as demonstrated in Figure 9 for PDB structure 1rat ($N=124$) that its density fluctuates around 0.6 for small R_c and starts decreasing monotonically at $R_c/R_G \approx 1$. However, this definition of the core radius is not suitable for a (almost) monotonically decreasing profile such as that of 1gof ($N=639$), which probably stems from a structure consisting of three substructures; Still, in all cases we use the above prescription based on $0.8\ R_G$.

The cluster length, *L*, is defined as the maximal distance between any two of the cluster's beads, where the length of a single-residue cluster is considered to be zero. We measure the cluster's size with respect to the core size by the ratio, $L_{\text{cluster}}=L_{\max}/2R_c$, where $L_{\max}$ is the length of the largest cluster; this applies to both hydrophobic and percolation clusters, which are denoted $L_{\text{H-clust}}$ and L_{perco}, respectively. It should be pointed out that during the clustering process, which always starts from a core residue, the generated cluster might "overflow" beyond the core limits; we allow this to occur and thus consider clusters that are arranged with respect to the core. Unlike the usual case where a percolating cluster should connect the opposite sides of a lattice, we define such a cluster when $L_{\text{perco}}/2R_c \approx 1$ (or larger than 1).

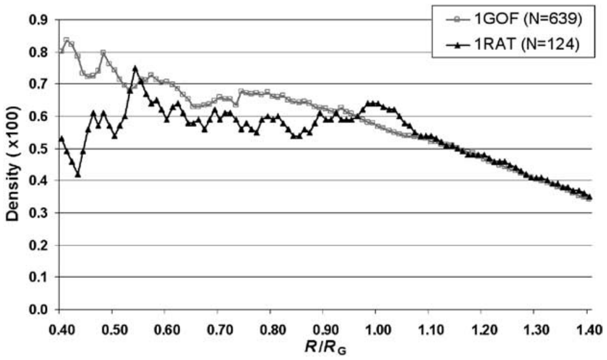

Figure 9. Good and bad density profiles. The density is defined as the number of α-carbon atoms within a sphere divided by the volume of the sphere. The density is calculated in concentric spheres around the center of mass, as a function of R/R_G, where *R* is the radius of the sphere, and R_G is the radius of gyration. The density profile of 1rat is "good", since it fluctuates around 0.6 from small *R* and starts decreasing monotonically at $R/R_G \approx 1$, while the profile of 1gof (bad) decreases almost monotonically from small *R*. Still, in all cases R_c is determined as $R \geq 0.8\ R_G$, for which the density starts decreasing.

Percolation experiments

For each of the proteins studied, the percolation experiments are performed with the corresponding fraction *f* of hydrophobic residues. Thus, all the lattice sites occupied by α-carbon atoms are considered initially to contain only P residues (beads); then, each of these sites is visited and P is replaced by H with probability *f*. After completing this procedure, the clusters of H are identified in the same way (i.e. shell-by-shell) as described above for the actual hydrophobic clusters. For each protein, 100 such percolation experiments based on different sets of random numbers are performed and the average size (and fluctuation) of the largest clusters generated in the core is calculated. We emphasize again the difference between a typical percolation experiment applied to a lattice and our percolation experiments, which are carried out over a chain structure embedded in a lattice. Because the effective coordination number (i.e. the average number of nearest-neighbor residues) of the chain is much smaller than that of the lattice, the percolation threshold of the former is significantly larger than that of the latter.

Acknowledgements

The authors gratefully acknowledge financial support from National Science Foundation Information Technology Research grants NSF 0225656 and NSF 0225636, and NIH grant GM61916.

References

1. Kauzmann, W. (1959). Some factors in the interpretations of protein denaturation. *Advan. Protein Chem.* **14**, 1–63.
2. Silverstein, K. A. T., Haymet, A. D. J. & Dill, K. A. (1998). A simple model of water and the hydrophobic effect. *J. Am. Chem. Soc.* **120**, 3166–3175.
3. Silverstein, K. A. T., Haymet, A. D. J. & Dill, K. A. (1999). Molecular model of hydrophobic solvation. *J. Chem. Phys.* **111**, 8000–8009.
4. Madel-Gutfreund, Y. & Gregoret, L. M. (2002). On the significance of alternating patterns of polar and non-polar residues in beta-strands. *J. Mol. Biol.* **323**, 453–461.
5. Broome, B. M. & Hecht, M. H. (2000). Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. *J. Mol. Biol.* **296**, 961–968.
6. Hennetin, J., Le tuan, K., Canard, L., Colloc'h, N., Mormon, J.-P. & Callebaut, I. (2003). Non-intertwined

binary patterns of hydrophobic/nonhydrophobic amino acids are considerably better markers of regular secondary structures than nonconstrained patterns. *Proteins: Struct. Funct. Genet.* **51**, 236–244.

7. Némethy, G. & Scheraga, H. A. (1962). The structure of water and hydrophobic bonding in proteins:III. The thermodynamic properties of hydrophobic bonds in proteins. *J. Phys. Chem.* **66**, 1773–1789.
8. Klotz, I. M. (1970). Comparison of molecular structures of proteins: helix content; distribution of apolar residues. *Arch. Biochem. Biophys.* **138**, 704–706.
9. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
10. Kuntz, I. D. (1972). Tertiary structure in carboxypeptidase. *J. Am. Chem. Soc.* **120**, 3166–3175.
11. Chothia, C. (1975). Structural invariants in protein folding. *Nature*, **254**, 304–308.
12. Krigbaum, W. R. & Komoriya, A. (1979). Local interactions as a structure determinant for protein molecules:II. *Biochim. Biophys. Acta*, **576**, 204–228.
13. Meirovitch, H., Rackovsky, S. & Scheraga, H. A. (1980). Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids. *Macromolecules*, **13**, 1398–1405.
14. Meirovitch, H. & Scheraga, H. A. (1980). Empirical studies of hydrophobicity. 2. Distribution of the hydrophobic, hydrophilic, neutral, and ambivalent amino acids in the interior and exterior layers of native proteins. *Macromolecules*, **13**, 1406–1414.
15. Meirovitch, H. & Scheraga, H. A. (1981). Empirical studies of hydrophobicity. 3. Radial distribution of clusters of hydrophobic and hydrophilic amino acids. *Macromolecules*, **14**, 340–345.
16. Rose, G. D. & Roy, S. (1980). Hydrophobic basis of packing in globular proteins. *Proc. Natl Acad. Sci. USA*, **77**, 4643–4647.
17. Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry*, **24**, 1501–1509.
18. Lau, K. F. & Dill, K. A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, **22**, 3986–3997.
19. Stauffer, D. & Aharony, A. (1992). *Introduction to Percolation Theory*. Taylor & Francis, London.
20. Alexandrowicz, Z. (1980). Critically branched chains and percolation clusters. *Phys. Letters A*, **80**, 284–286.
21. Meirovitch, H. (2002). Polymer collapse, protein folding, and the percolation threshold. *J. Comput. Chem.* **23**, 166–171.
22. Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.
23. Covell, D. G. & Jernigan, R. L. (1990). Conformations of folded Proteins in restricted spaces. *Biochemistry*, **29**, 3287–3294.
24. White, S. H. & Jacobs, R. E. (1990). Statistical distribution of hydrophobic residues along the length of protein chains: implications for protein folding and evolution. *Biophys. J.* **57**, 911–921.
25. White, S. H. & Jacobs, R. E. (1993). The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. *J. Mol. Evol.* **36**, 79–95.
26. Schwartz, R., Istrail, S. & King, J. (2001). Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. *Protein Sci.* **10**, 1023–1031.
27. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1994). Nonrandomness in protein sequences: evidence for physically driven stage of evolution. *Proc. Natl Acad. Sci. USA*, **91**, 12972–12975.
28. Irbäck, A., Peterson, C. & Potthast, F. (1996). Evidence from nonrandom hydrophobicity in protein chains. *Proc. Natl Acad. Sci. USA*, **93**, 9533–9538.
29. Stauffer, D. (1976). Gelation in concentrated critically branched polymer solutions. *J. Chem. Soc. (London) Faraday Trans. II*, **72**, 1354–1364.
30. Stauffer, D., Coniglio, A. & Adam, M. (1982). Gelation and critical phenomena. *Advan. Polym. Sci.* **44**, 103–158.
31. de Gennes, P. G. (1976). On a relation between percolation theory and the elasticity of gels. *J. Phys. (Paris) Letters*, **37**, 1–2.
32. de Gennes, P. G. (1985). *Scaling Concepts in Polymer Physics, chapt. 5*. Cornell University Press, Ithaca, NY.
33. Wertz, D. H. & Scheraga, H. A. (1978). Influence of water on protein structure. An analysis of the preference of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules*, **11**, 9–15.
34. Manavalan, P. & Ponnuswamy, P. K. (1978). Hydrophobic character of amino acid residues in globular proteins. *Nature (London)*, **275**, 673–674.
35. Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.
36. Chothia, C. J. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1–12.
37. Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985). *Science*, **229**, 834–838.
38. Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.
39. Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature*, **277**, 491–492.
40. Wolfenden, R., Andersson, L., Cullis, P. M. & Southgate, C. C. B. (1981). Affinities of amino acid chains for solvent water. *Biochemistry*, **20**, 849–855.
41. Zhou, H. & Zhou, Y. (2002). Stability scales and atomic solvation parameters extracted from 1023 mutations experiments. *Proteins: Struct. Funct. Genet.* **49**, 483–492.
42. Sharp, K. A., Nicholls, A., Friedmann, R. & Honig, B. (1991). Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models. *Biochemistry*, **30**, 9686–9697.
43. Fauchere, J.-L. & Pliska, V. (1983). Hydrophobic parameters p of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides Eur. *J. Med. Chem. (Chim. Ther.)*, **18**, 369–375.
44. Nozaki, Y. & Tanford, C. J. (1971). The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J. Biol. Chem.* **246**, 2211–2217.
45. Tobi, D., Shafran, G., Linial, N. & Elber, R. (2000). On the design and analysis of protein folding potentials. *Proteins: Struct. Funct. Genet.* **40**, 71–85.
46. Hinds, D. A. & Levitt, M. (1994). Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* **243**, 668–682.
47. Liang, J. & Dill, K. A. (2001). Are proteins well-packed. *Biophys. J.* **81**, 751–766.
48. Godzik, A., Skolnick, J. & Kolinski, A. (1992). Simulations of the folding pathway of triose phosphate isomerase-type α/β barrel proteins. *Proc. Natl Acad. Sci. USA*, **89**, 2629–2633.

49. Kolinski, A., Milik, M., Rycombel, J. & Skolnick, J. (1995). A reduced model of short range interactions in polypeptide chains. *J. Chem. Phys.* **103**, 4312–4323.
50. Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T. *et al*. (2002). Long-range interactions within a nonnative protein. *Science*, **295**, 1719–1722.
51. Dill, K. A. (1999). Polymer principles and protein folding. *Protein Sci.* **8**, 1166–1180.
52. Shakhnovich, E. I. (1998). Protein design: a perspective from simple tractable models. *Fold. Des.* **3**, R45–R58.
53. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1994). *Numerical Recipes in Fortran*. Cambridge University Press, New York.

Edited by M. Levitt

(Received 30 July 2004; received in revised form 14 September 2004; accepted 21 September 2004)