Hierarchical Protein Folding Pathways: A Computational Study of Protein Fragments

Nurit Haspel,[1] Chung-Jung Tsai,[2] Haim Wolfson,[3] and Ruth Nussinov[1,2,†]

[1]*Sackler Institute of Molecular Medicine, Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel*

[2]*Intramural Research Support Program, SAIC, Inc., Laboratory of Experimental and Computational Biology, NCI-Frederick, Frederick, Maryland*

[3]*School of Computer Science, Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel*

ABSTRACT **We have previously presented a building block folding model. The model postulates that protein folding is a hierarchical top-down process. The basic unit from which a fold is constructed, referred to as a hydrophobic folding unit, is the outcome of combinatorial assembly of a set of "building blocks." Results obtained by the computational cutting procedure yield fragments that are in agreement with those obtained experimentally by limited proteolysis. Here we show that as expected, proteins from the same family give very similar building blocks. However, different proteins can also give building blocks that are similar in structure. In such cases the building blocks differ in sequence, stability, contacts with other building blocks, and in their 3D locations in the protein structure. This result, which we have repeatedly observed in many cases, leads us to conclude that while a building block is influenced by its environment, nevertheless, it can be viewed as a standalone unit. For small-sized building blocks existing in multiple conformations, interactions with sister building blocks in the protein will increase the population time of the native conformer. With this conclusion in hand, it is possible to develop an algorithm that predicts the building block assignment of a protein sequence whose structure is unknown. Toward this goal, we have created sequentially nonredundant databases of building block sequences. A protein sequence can be aligned against these, in order to be matched to a set of potential building blocks. Proteins 2003;51:203–215.**

Key words: protein folding; building blocks; protein structure prediction; hierarchical folding; protein fragments; folding complexity

INTRODUCTION

Several models have been proposed to describe the protein folding process.[1] These include

(i) the framework model,
(ii) the nucleation and growth mechanism,
(iii) the diffusion-collision model,
(iv) the hydrophobic collapse, and
(v) the hierarchical model.

In the (i) framework model,[2–4] secondary structure formation is independent of tertiary interactions and usually precedes these. If tertiary interactions occur first, they are not necessarily native. In the (ii) nucleation and growth[5] or nucleation-condensation mechanism,[6,7] folding initiates by formation of a "nucleus," followed by its extension. The model proposes that formation of such a nucleus is dependent on contacts between key residues, which have been conserved through evolution. This model has led to searches for specific residue-conservation in families of related proteins. In the third (iii) model, largely preformed secondary structure elements assemble into complete folds through random diffusion and collision.[8] If favorable, they may lock to yield native conformations. In contrast to these, the hydrophobic collapse (iv) model highlights the hydrophobic effect.[9–11] In this scheme, folding initiates by burial of extensive nonpolar surface area. Secondary structure and specific interactions follow. In the hierarchical model (v) pioneered by G. Rose almost two decades ago,[12] folding initiates locally, with folded elements assembling step-wise to yield the native fold (reviewed in Refs. 13 and 14). These models are not necessarily exclusive of each other. The hierarchical model may include elements of hydrophobic collapse in the assembly of local folded elements. Optimization of the specific (e.g., van der Waals, electrostatic, disulfide bonds) interactions would follow. The hierarchical model may

The content of this publication does not necessarily reflect the view or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

Grant sponsor: Ministry of Science, Israel (to R.N. and H.J.W.); Grant sponsor: Israel Science Foundation (administered by the Israel Academy of Sciences) grant from the "Center of Excellence in Geometric Computing and its Applications" (to R.N. and H.J.W.); Grant sponsor: Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University (to H.J.W.); Grant sponsor: National Institutes of Health, National Cancer Institute; Grant number: NO1-CO-12400.

†Correspondence to: Ruth Nussinov, NCI-Frederick, Building 469, Room 151, Frederick, MD 21702. E-mail: ruthn@ncifcrf.gov

Received 30 June 2002; Accepted 22 August 2002

further include elements of nucleation and growth. Nucleation does not necessarily have to be restricted to specific residues. Similarly, in the framework model, we may substitute single secondary structure element formation by chain-linked local building block minima.

Since the first pioneering work,[12] numerous articles have appeared, substantiating the hierarchical folding concept. The building block folding model also considers folding as a hierarchical event.[15,16]

At the first step, local transient building block elements fold. The conformations they obtain are not necessarily stable, but they have higher population times than all other, alternate conformations. In the next step, building blocks associate, mutually stabilizing each other. The association is via selection of favorable conformers, similar to multi-molecular complex formation. Applying our procedure, we progressively dissect native protein structures[15,16] and proceed to create building block clusters separately from all-α, α+β, α/β, and all-β protein classes. The clusters represent similar conformations [low root mean squared distance (RMSD) between them]. We analyze the clusters with respect to their biological and chemical characteristics: size, buried/exposed nature, stability, and fold location.

Our goal is to relate each building block cluster to a characteristic "profile" that represents its typical sequence, super-secondary structure composition, hydrophobicity, and buried/exposed position. Because building blocks constitute local minima on the polypeptide chain, they may enable identification of building blocks on amino acid sequences, help in following the protein folding process, and be useful in modeling of local and global protein structures.

Some building blocks are "more important" for correct folding than others.[16,17] If a critical building block is cut out of the protein structure, the remaining building block fragments collapse to yield a non-native stable 3D fold. On the other hand, removal of a noncritical building block leads to a structure more similar to the native fold. Consequently, we further analyze the building blocks with respect to their "criticalness," i.e., their locations, type, and extent of interactions in the protein. We investigate whether they have a typical 3D fold and their relative stability.

Our results suggest that building blocks can be viewed as stand-alone protein fragments, with conformations repeating between different families, regardless of the overall structures and sequences.

This is reminiscent of protein folds. Nevertheless, on their own small building blocks with low stabilities may have multiple preferred conformations, with low population times. In such cases, the native conformation is stabilized by interactions with sister building blocks in the protein, similar to two-state complexes. These findings lend support to folding schemes, which are based on hierarchical concepts.

They further help in visualization of dynamic folding pathways and intermediate (mis-associated) states. With regard to critical building blocks, we find that the higher their "criticalness" score, the less stable. This makes inherent sense, because it enables large interfaces between building blocks in the protein core, while still keeping down protein size.

METHODS

The Building Block Cutting Algorithm

The algorithm[15,16] uses a scoring function that measures the relative conformational stability of a candidate building block. The stability score for a given building block is defined as:

$$SCORE^{BB}(Z,H,I) = \frac{(Z^1_{avg} - Z)}{Z^1_{Dev}} + \frac{(H - H^1_{avg})}{H^1_{Dev}} + \frac{(I^1_{avg} - I)}{I^1_{Dev}} + \frac{(Z^2_{avg} - Z)}{Z^2_{Dev}} + \frac{(H - H^2_{avg})}{H^2_{Dev}} + \frac{(I^2_{avg} - I)}{I^2_{Dev}} \quad (1)$$

where Z is compactness, H hydrophobicity, and I isolatedness. Each of the components is calculated as the deviation from the average value of known protein structures. The average and standard deviation of these quantities were calculated from a nonredundant dataset of 930 representative single-chain proteins from the PDB.[19] Terms with superscript 1 were determined with respect to fragment size and those with superscript 2 as a function of the fraction of the fragment size to the whole protein.

First we define the ASA (the solvent accessible surface area) of the fragment, which is calculated numerically.[20,21] The three components are described elsewhere.[1,15]

All candidate fragments with minimum length are tested for their stability score. Local minima are candidate building blocks. This process continues iteratively until the building blocks can no longer be dissected. The resulting tree outlines the most probable folding routes. The different levels are referred to as cutting levels.

The Critical Building Block Finding Algorithm

The algorithm[17,18] uses a scoring function (t-score) based on the contacts the building block has with other building blocks.[18] Consider a building block j that interacts with two different building blocks, k and l. The differential contacting surface area for building block j is defined as:

$$\text{Diffcontsa (j)} = \text{contsa(j,k)} + \text{contsa (j,l)} - \text{contsa (k,l)}$$

where contsa(j, k) is the surface area buried between building blocks j and k. Consider the following cases:

1. Diffcontsa(j) < 0. The interactions between k and l are stronger than the sum of their interactions with j, Diffcontsa(j) is set to zero.
2. Diffcontsa(j) > 0. The interactions between k and l are weaker than the sum of their interactions with j, Diffcontsa(j) is multiplied by different weights, so that greater weight is assigned to building blocks that mediate the interactions between building blocks that are not in direct contact.

The critical building block index [CIndex(j)] for a building block j is the sum of diffcontsa(j) computed for all combinations of k and l divided by the total surface area of building block j, totsa(j):

$$Clndex(j) = \sum diffcontsa(j) = totsa(j)$$

The total surface area of a building block has two terms: the surface area buried by the rest of the protein ([protburysa(j)] and the surface area exposed to the solvent [solvexpsa(j)]. The critical building block index is modifed to give more weight to building blocks that are largely buried:

$$Clndex(j) = Clndex(j) - protburysa(j)/solvexpsa(j)$$

At each cutting level, the average and standard deviation of the CIndex values are computed. The statistical significance of each building block is measured by its Z-score:

$$Z_{score}(j) = (CIndex(j) - \mu)/\sigma,$$

where μ is the average building block CIndex value and σ is the standard deviation. A building block is considered critical if it satisfies the following criteria:

(i) it is found at most levels below hydrophobic folding unit,
(ii) it has a consistently high CIndex at different levels, and
(iii) its CIndex is significant by at least two standard deviations in at least one hierarchical level of the protein anatomy.

Creating the Building Block Databases

The building block database was created using the data collected by Tsai et al.[15] (available at http://protein3d.ncifcrf.gov/tsai). We created 24 different databases, separately clustered from four protein classes, all-α, $\alpha+\beta$, α/β, and all-β. For each class we created a database for each cutting level (first to sixth). Not all proteins can be cut down to the sixth level, and a database of a higher cutting level contains fewer proteins than of a lower level. The coordinates of each building block were taken from the PDB, assuming that the native building block conformation is the most populated in solution.

Clustering the Building Blocks

Clustering was based on structural similarity. Each of the databases (corresponding to a different class and level) was clustered separately. The clustering algorithm is:

1. Each cluster has representative members (one or more), assuming members are similar enough so a building block that matches the representatives, matches all members.
2. For each building block: Run over all existing clusters an algorithm that finds the best rigid matching between the candidate building block and the cluster representative building block, provided that the size of each of the two building blocks is at least 70% of the other building block. The rigid matching algorithm is geometric-hashing based.[22]
3. If the two building blocks match (within RMSD of 1.8 Å at most for cutting levels 1 and 2, 1.5 Å for levels 3 and 4, and 1.3 Å for levels 5 and 6, and the match size is at least 70% the size of the smallest protein) assign the building block to this cluster.
4. If, at the end of the procedure, the building block does not match any cluster representative, open a new cluster with this building block as the representative. The cluster representative is simply the first building block that has opened this cluster. This is an approximation that can save computational time, but may also be inaccurate because building blocks in the same cluster may only match the representative, but not one another. Averaging over the cluster members' coordinates for a representative is more accurate. However, recalculating the coordinates of a representative would take time proportional to the length of the building block (because such a calculation is executed for each CA, repeated each time a new building block is added to the cluster). Additionally, not all cluster building blocks are identical in length.

Clustering has two stages:

1. Classifying the building blocks according to their original SCOP[23] protein family and clustering within the family. The motivation for this stage is that proteins from the same family almost always give very similar building blocks; therefore, clustering proteins within families reduces the number of building blocks at the next stage and saves computational time. This stage is relatively very fast, because the number of building blocks for each family is small and they cluster well, so the number of the resulting clusters is also small.
2. Merging the initial clusters, with the representative of each initial cluster now representing all cluster members. This step does not reduce the clustering accuracy, because the initial clusters contain very similar, nearly identical building blocks. If such a cluster is merged with another cluster, the initial representatives of the two clusters become the representatives of the new cluster. In such cases, a cluster may have more than one representative.

The complexity of the clustering stage is, in the worst case:

O(The number of building blocks * The number of clusters).

However, in practice, after running the first, initial stage, the complexity is closer to:

$$O(\text{The number of clusters}^2).$$

Figure 1 gives a schematic illustration of the clustering process.

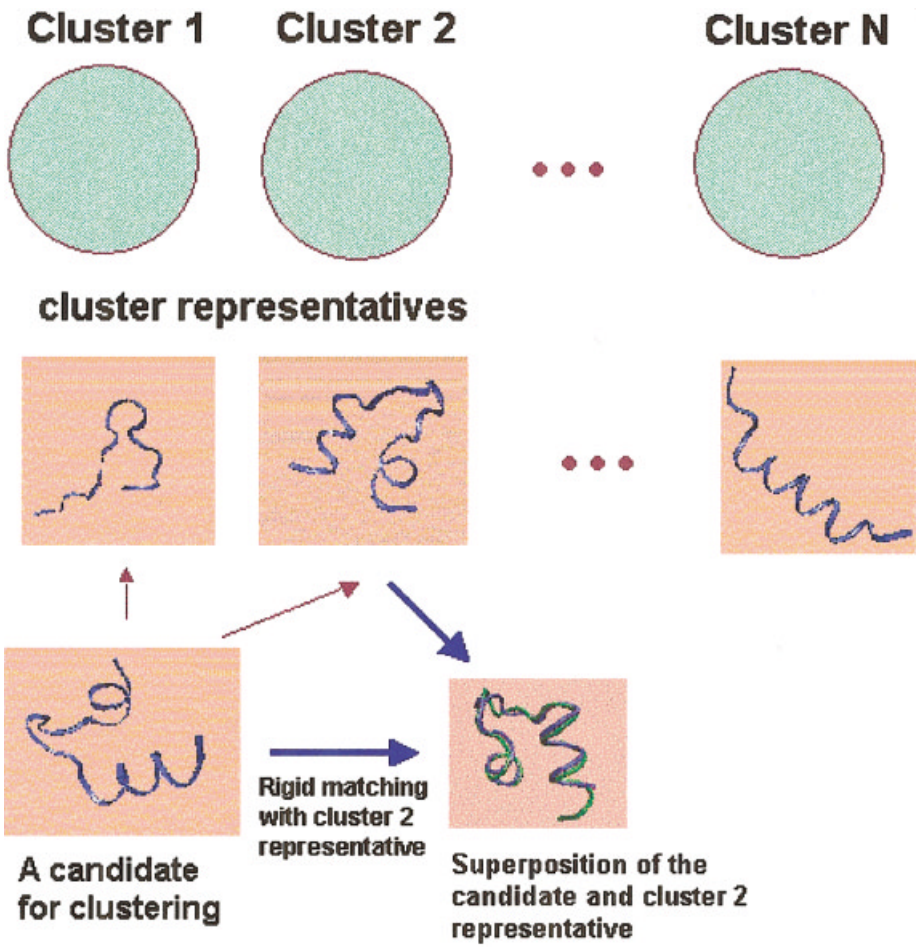

Fig. 1. A schematic illustration of the clustering process. The circles at the top represent the clusters. The building blocks below them are their representatives. In this figure there are 3 clusters. The bottom left building block is a candidate for clustering. It is trial-matched against the representatives of the clusters until a match is found (in this case, a match is obtained with the representative building block of cluster 2. The match is shown in the square at the bottom center). Consequently, the candidate building block is joined to cluster 2 and is not tested against the other cluster representatives.

Creation of the Sequence Databases

Next, we created a sequentially nonredundant sequence database that represents each clustered database the following way:

1. For each cluster, the sequences of the cluster members were extracted to a FASTA format file.[24]
2. Within each cluster, the sequences are clustered using the utility blastclust (in the BLAST Package[25]) with default parameters. Thus, each structural cluster can be associated with a nonredundant group of sequences that give a local structural pattern.
3. All the nonredundant sequence groups of all structural clusters are gathered and reclustered using blastclust. The goal of this stage is to eliminate all redundancies among clusters, caused by similar sequences that fall in different structural clusters. Ideally, this should not occur, because similar sequences almost always give similar structures and should be in the same cluster. However, because of our clustering method, if there are similar structural clusters, a building block that can match both clusters will be assigned to the first one it encounters. The goal of this stage is to compensate for such cases.

The result of this procedure is a sequentially nonredundant database that represents the whole structural database by means of sequences. Each item in that database is associated with a specific structural cluster, such that a structural cluster can be represented by more than one sequence.

The cutting and clustering programs were written as a library of C-shell, perl and C++ programs and were run on a two processor Red Hat 6.1 Linux machine. The statistical analysis was performed using Matlab, version 6.

RESULTS

Building Block Database Creation and Clustering

Figure 2 presents an example of the distribution (here for $\alpha+\beta$) of the number of clusters in the different cutting

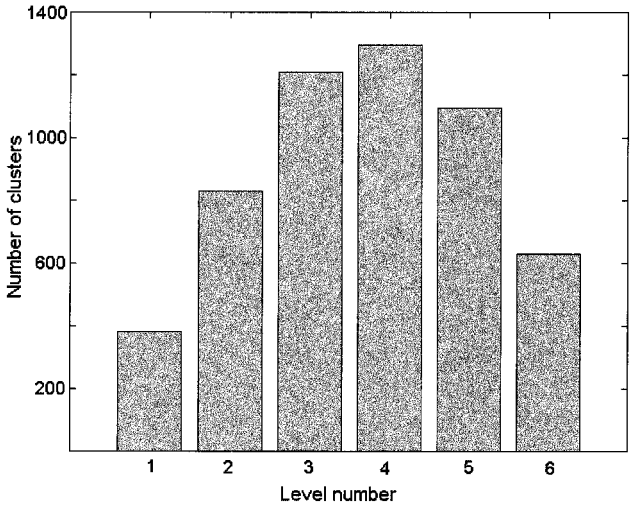

Fig. 2. The number of clusters at each cutting level (1–6) for $\alpha+\beta$ proteins. Other protein classes exhibit a similar behavior.

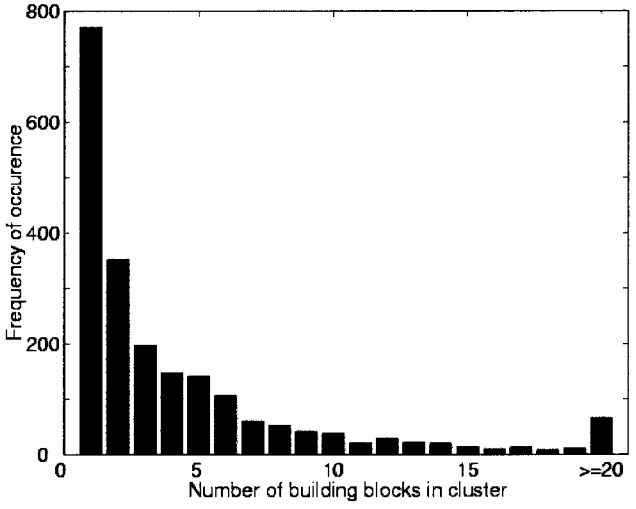

Fig. 3. The distribution of cluster sizes in level 5 of α/β proteins. Other protein classes behave in a similar manner.

levels. The average building block size hardly changes beyond level 3. This is because at lower levels (5–6), the proteins that were cut to short fragments and lowered the average building block size can no longer be cut and therefore are not included. This keeps the average size quite constant at levels 4–6. Figure 3 shows the distribution of cluster sizes among level 5 of α/β proteins as an example. If our clustering parameters were more liberal, there would be more large clusters, but their members would be less similar. We preferred small clusters with more similar members. Table I shows that the average cluster size usually ranges between 3 and 6 building blocks, and depends on the protein class.

General Cluster Analysis by Means of Sequence and Structure

Clusters with nearly identical building blocks usually derive from the same family. However, sequence conservation changes considerably among clusters. Figure 4(a,b) illustrates examples of clusters. The superpositioning has been carried out using MUSTA, our multiple structure comparison algorithm.[26] In these examples, the cluster members do not all belong to the same sequence family. For clarity, building blocks that are nearly identical, sequentially and structurally, were removed, with only one (or two) from each sequence family shown here. Thus, even though the sequence similarity is low, the structural similarity is high. This phenomenon resembles whole proteins, where the structure is better conserved than the sequence. As expected, the structurally least conserved are the building block ends, turns and loops, regardless of their sequence conservation. The building blocks vary from a nearly single secondary structure element (e.g., an α-helix with a short loop attached to it) to an almost independent protein domain.

The most frequent cases are of super-secondary structure elements, like two helices connected by a loop or two strands and a loop.

Building block stability score analysis was calculated from the distribution of the stability scores for nonredundant building blocks in every cluster with over 15 building blocks. The compactness component of the stability score may differ within the cluster, because of loops and dangling ends. The hydrophobicity depends both on the amino acid composition of the building block and on its structure, and the "isolatedness" depends on the composition and the environment the building block is in, that is, on the rest of the protein. Therefore, we expect building blocks from different protein families (thus, different sequences and environments) to have different stability scores. Figure 5 shows an example of a cluster with a broad stability distribution, with the building blocks derived from proteins that do not all belong to the same structural and sequential families. Although some building blocks are very similar to each other sequentially, others differ substantially. This broad distribution was observed in most clusters that contained building blocks from different protein families. On the other hand, homogeneous clusters exhibit a narrow stability distribution. Greater similarity in the stability scores reflects higher similarity of the original proteins and not only the structural similarity between building blocks.

Do Different Proteins Contribute Building Blocks to Different Clusters?

We expect the answer to be generally no. Proteins from different families may share structural elements and some of them may be local stable elements. On the other hand, the probability that two dissimilar proteins will share more than one such local substructure is low. For each of our 24 building block cluster databases we created a nonredundant subset of building blocks from proteins of different families.

For each we calculated how many building blocks it shares with all others. The results are presented in the form of a matrix M, where $M(i,j)$ is the number of building blocks shared by proteins i and j. We expect most of the

TABLE I. Details for Each Database†

Protein class	Cutting level	No. of protein chains	No. of building blocks	No. of clusters	Average building block size	Average cluster size
α/β	1	1514	2121	515	138.80	4.12
α/β	2	1514	5257	1286	58.83	4.09
α/β	3	1514	8013	1899	36.84	4.22
α/β	4	1495	7813	1727	29.25	4.52
α/β	5	1174	9944	2124	27.41	4.68
α/β	6	595	5663	1402	27.16	4.04
α + β	1	1619	2491	379	101.95	6.57
α + β	2	1616	5445	828	47.47	6.58
α + β	3	1558	7454	1199	31.82	6.22
α + β	4	1339	7702	1285	26.65	5.99
α + β	5	520	4361	1089	26.20	4.00
α + β	6	153	1606	628	26.75	2.56
all-α	1	870	1419	405	81.93	3.50
all-α	2	870	3181	751	42.47	4.23
all-α	3	794	3858	830	31.14	4.65
all-α	4	594	2971	727	27.09	4.09
all-α	5	263	2662	707	26.52	3.77
all-α	6	94	1299	466	27.73	2.88
all-β	1	1313	1757	428	109.90	4.10
all-β	2	1311	3814	1049	51.16	3.64
all-β	3	1299	5815	1566	34.87	3.71
all-β	4	1132	4683	1456	29.80	3.22
all-β	5	523	4068	1283	28.22	3.17
all-β	6	167	1914	729	28.75	2.63

†The data here were calculated only from the representatives of the clusters, to avoid bias toward large families that contain nearly identical building blocks.

matrix to consist of 0 and 1, and only a few indices with more than 2. Figure 6 shows matrices constructed from nonredundant building block databases representing the four folds, all-α, all-β, α/β, and α+β, at level 3. α/β, α+β, and all-β matrices mostly consist of 0's with small "isles" of 1's and few isles of larger numbers (see color bar). This leads us to conclude that the building blocks in proteins from different families are largely independent and that two unrelated proteins usually do not share more than one building block of similar structure. This conclusion is more profound than the simple fact that two unrelated proteins usually do not share many structural elements. The building blocks are different from other structural elements in the sense that they are structurally stable and can exist independently in solution. This conclusion further supports our assumption that the building blocks are independent units. The all-α matrix illustrates that these proteins have a higher likelihood of sharing more than one building block, probably because the number of ways helices can combine to yield local stable elements is smaller.

Critical Building Block Analysis

Relative position of the critical building blocks in the protein

Because critical building blocks tend to be buried, being able to identify critical blocks in protein sequences may provide a clue about the 3D location of the fragment. Consequently, we are interested in features that make a building block critical.

Critical building blocks are frequently unstable. Nevertheless, the native conformation is likely to prevail, similar to the situation in other building blocks. Their instability implies low population times; however, through their binding to other building blocks, the native conformations are greatly stabilized, leading to population shifts toward the native conformations. Kumar et al.[18] have suggested that the critical building blocks tend to appear at the N-terminal part of the protein sequence. This was supported by both the locations of the proregion and by analysis of dihydrofolate reductase[17] and adenylate kinase.[18] Critical building blocks can be viewed as intramolecular chaperon-like fragments, except that they are not cleaved and remain an integral part of the protein. Here we have carried out an analysis of all critical building blocks, and the results are summarized in Table II, with a typical distribution shown in Figure 7. Critical building blocks can appear anywhere along the sequence, but they have a tendency toward the termini, with a slight preference toward the C-terminus. Figure 8 shows an example of a building block that has been identified as critical by the algorithm. The critical building block (at the C-terminus, 8th in level 3) is darkened with the rest of the protein in a lighter shade.

Characteristic size and shape of a critical building block

Our analysis did not detect characteristic sizes and shapes of critical building blocks. Sequence alignments of nonredundant sets of critical building blocks from differ-

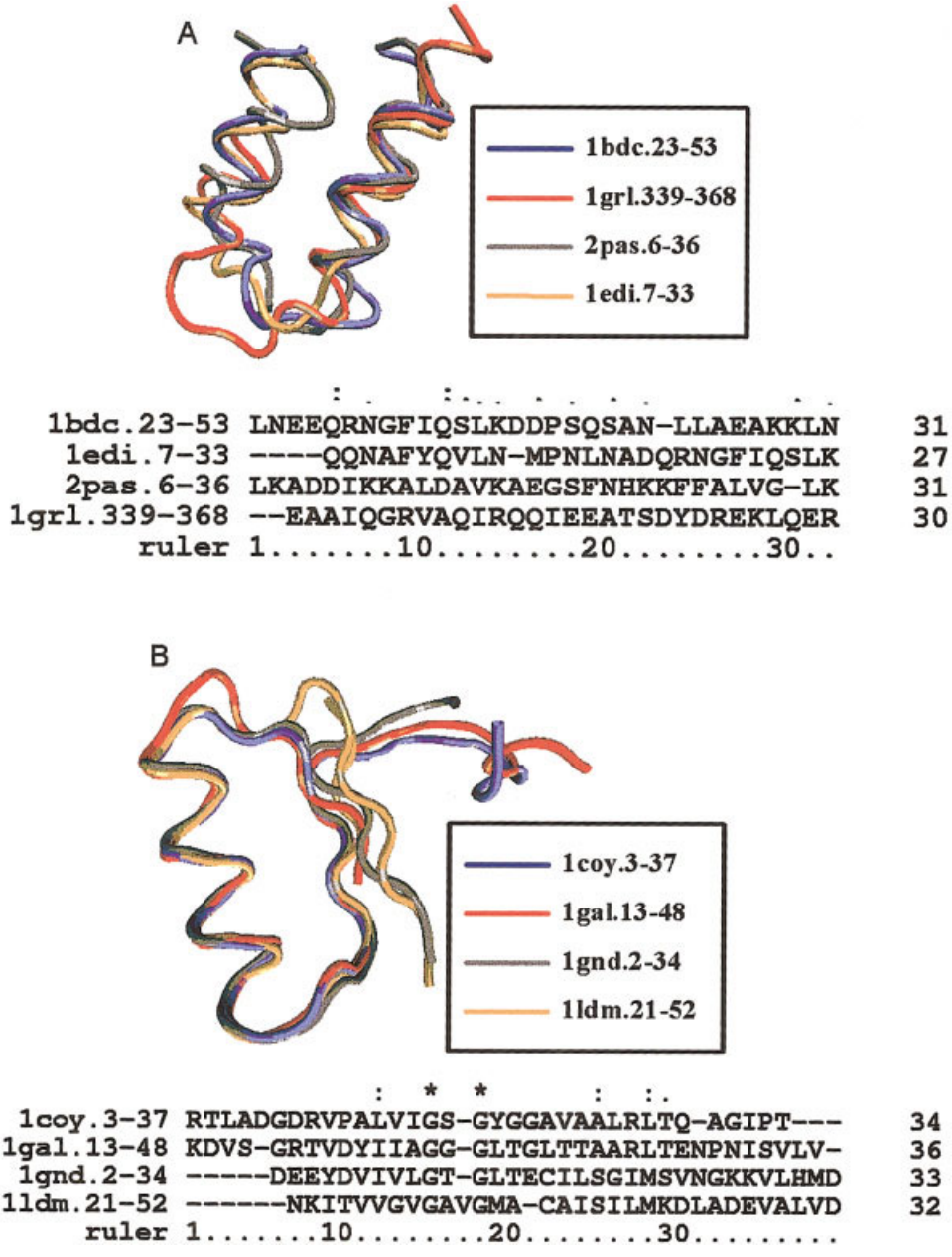

Fig. 4. Two examples of clusters. The figures show multiple structural alignments of the cluster members and the legends depict the pdb18 codes of each building block member, and its range of residues. (A) 6ldh, apo-lactate dehydrogenase from dogfish; 1avo, apo duck ovotransferrin; 1gnd, guanine nucleotide dissociation inhibitor, alpha-isoform from cow. (B) 1bbhA, cytochrome c from *Chromatium vinosum*, chain A; 1cgn, cytochrome c from *Alcaligenes denitrificans*; 1cpr, cytochrome c from the purple phototropic bacterium, *Rhodobacter capsulatus*; 1jsw, native l-aspartate ammonia lyase from *Escherichia coli*, chain B; 1rcp, cytochrome c from *R. capsulatus*; 2fuoA, fumarase c with bound citrate from *E. coli*, chain A.

ent folds and at different levels did not reveal any significant sequence similarity. Critical building blocks have different sizes and secondary structure composition and folds and cannot be characterized by their sequences. This was expected, because the "criticalness" does not depend on the details of the shape, but on the position inside the protein. However, the frequency of hydrophobic residues is considerably higher than in other parts of the protein. The difference gets larger at lower levels. A χ^2 test of the frequency differences between critical and noncritical building blocks is very significant, in the majority of the cases with a p-value close to 0.

Relative stability of critical building block

Is there a correlation between the criticalness score and building block stability? It is reasonable to assume a negative correlation, i.e., the more critical the building block, the less stable it will be. The overall contact area

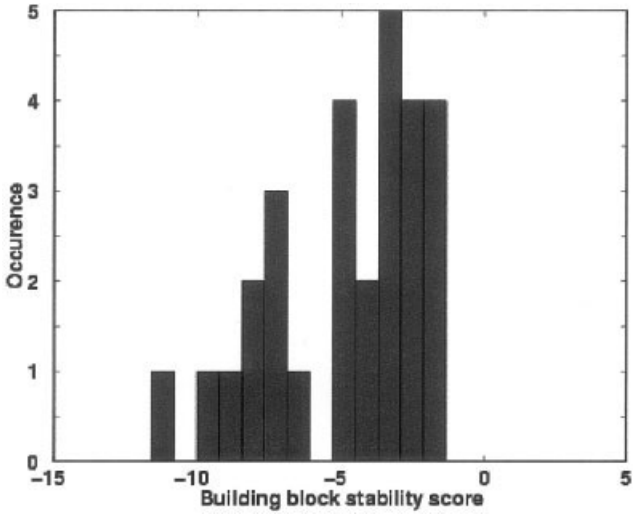

Fig. 5. An example of a stability score distribution in a cluster of α+β proteins, level 3, where the building blocks belong to various protein families and the stability score distribution is broad (mean: −4.843; SD: 2.63). The sequence alignment (not shown) illustrates weak similarity.

between a critical building block and all other building blocks is large. On the other hand, critical building blocks are relatively small. Small building blocks usually have lower stability, because they do not have a large hydrophobic core. Combined, large interfaces and small size argue for low stability. In solution, on their own, critical building blocks are likely to persist in a disordered state. To be stable, with such an extended surface area necessitates considerably larger fragment sizes. However, that would be counterproductive to the cell, as it would inflate protein sizes.[27] This explanation is further supported by the fact that the criticalness score gets higher with the level of the cutting. Thus, although the critical building blocks tend to be relatively unstable, they are still considered building blocks, with higher population time than other conformations. Figure 9 plots the stability against the criticalness score (t-score), for all building blocks from α/β, α+β, all-α, and all-β proteins, at level 3. The linear fit of the two data series is also plotted. The correlation coefficients are given in the Figure 9 legend. Note that the main stability score is less than zero. This results from the way the hydrophobicity score is calculated, as a linear extrapolation instead of the number of standard deviations. The correlation coefficient is very significant statistically (p-value $\approx$ 0 using t-test).

Next we examined whether members of the same clusters tend to have similar criticalness scores. We expected the criticalness score to be less dependent on the building block itself and more on its environmental context, consistent with our finding that there is no typical critical building block structure. Because members of the same cluster have similar 3D structures but may be located in different areas of their proteins, they are likely to have different criticalness scores. In contrast, building blocks derived from similar proteins have similar criticalness scores. This is likely to occur not because these building blocks are similar, but because proteins from the same family tend to be cut at similar locations. Therefore, building blocks from proteins of the same family would not only be very similar in sequence and structure, but also be in similar 3D environments within their proteins. Figure 10 shows an example of the distributions of the criticalness scores within large building block clusters of proteins from different families. The criticalness scores distribute broadly within the clusters. Within families, they distribute narrowly.

Previously we have compared in detail sequences and structures of building blocks within the same family. We found[17,18] that both are more conserved in critical building blocks than in other building blocks. This may be understood when we consider the instability of the critical building blocks. Even small changes in sequence can shift the energy landscape, leading to misfolded proteins. It is intriguing that at least for the cases we have examined, building blocks critical for folding are also critical for function. Currently, we are exploring further this potential folding-function inter-relationship.

DISCUSSION

Hierarchical protein folding schemes and the building block folding model

There are a number of strategies that have been adopted in protein folding schemes.[28] These range from "real" ab initio folding, to threading and homology modeling. The first is currently impractical for chains sizes over 40–50 residues. On the other hand, success in modeling is a function of the similarity between the target and sequences whose structures are available. In-between are hierarchy-based schemes.[29–31]

These pre-pick chain pieces and fold them through modeling of numerous short, overlapping, fixed-sized fragments. A well-known example is that of Baker et al.,[32,33] who studied short sequences that folded into known 3D local substructures by shifting an eight-residue window along the sequence. Oliva et al.[34] classified loop structures. Other methods search the conformational space on the energy landscape.[35–38] Combined with multiple sequence alignments and ensemble-clustering schemes, these significantly improve structural prediction.[39–42] Alternatively, cutting the target sequence into fragments and sampling space through parallel tempering (replica exchange) enables overcoming barriers between local minima, frequently associated with rough energy landscapes. Here, multiple (MC, MD) trajectories are run at given temperature intervals, switching conformations at specified steps. Hence, "folding by parts" with subsequent combinatorial docking, appears a practical promising approach.

Here, we provide validation of hierarchical schemes. In contrast to previous algorithms,[29–31] our size independent scoring function enables generating longer fragments, whose size is neither arbitrary nor fixed. Through analysis of the protein fragments, we show that chain segments may fold independently of their surroundings. Our heuristic energy function for the calculation of local minima has obvious deficiencies; for example, it does not include

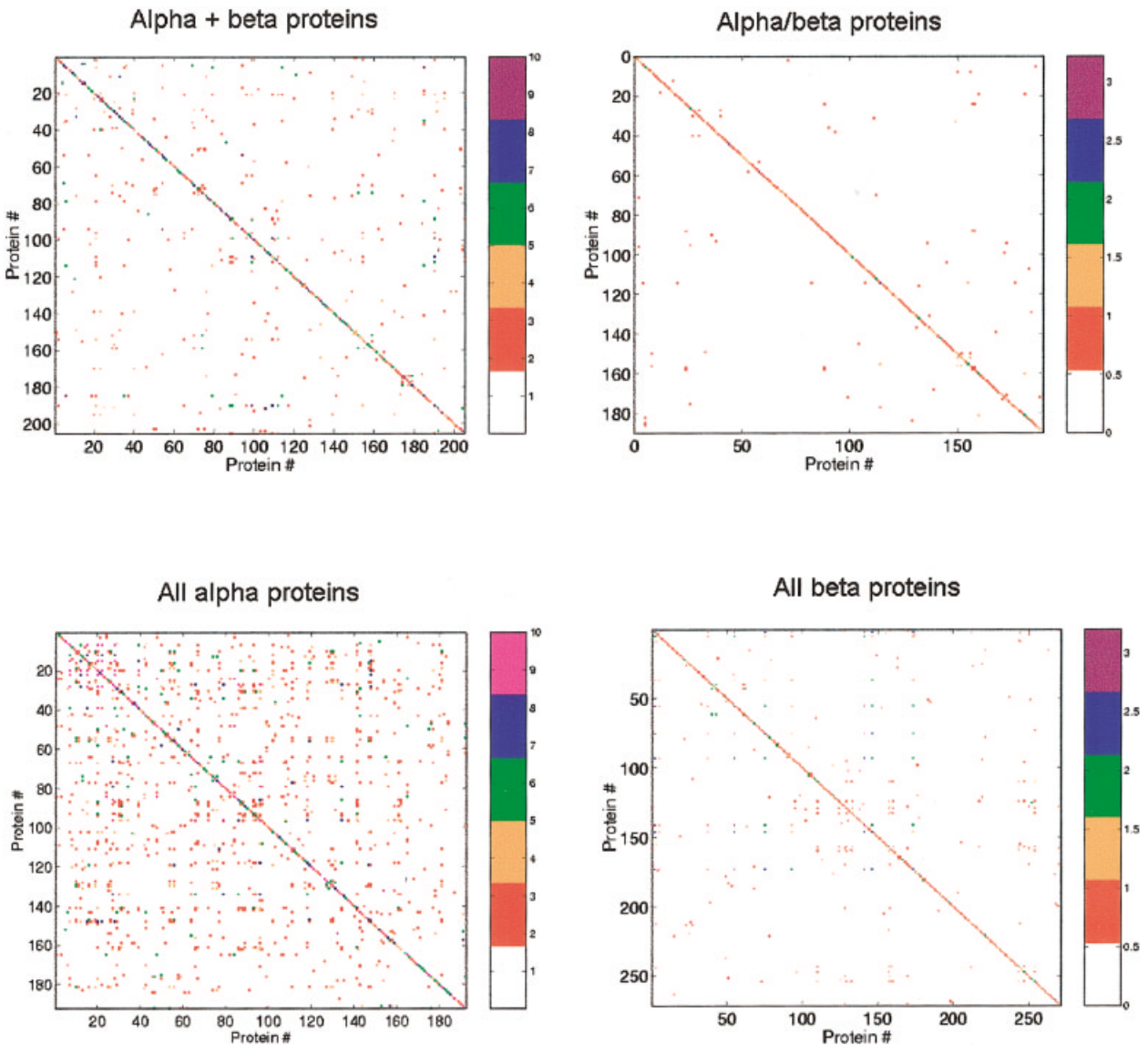

Fig. 6. Matrices representing the number of building blocks shared by nonrelated proteins of a nonredundant building block database representing the proteins classes, level 3. The color bars on the right hand-side represent the coloring scheme. White areas represent proteins that do not share any building blocks, and red dots represent proteins that share one building block, etc.

electrostatics; it is based solely on nonpolar buried surface area, compactness, and fragment "isolatedness"; and it accounts only for the native state.

Nevertheless, the computational cutting procedure has been shown to be consistent with experimental limited proteolysis for a number of cases for which experimental data are available (α-lactalbumin, thermolysin, hen egg white lysozyme, cytochrome c, ribonuclease A apomyoglobin).[43] Combined with the finding that structurally similar building blocks are obtained from overall different protein structures with different sequences, it suggests that we can adopt hierarchical strategies for protein folding.

Furthermore, this general agreement with experiment and the results from our cluster analysis suggest that rather than try all fixed-window sized fragments along the chain, a strategy to consider is focusing on local minima.

We note, however, that to determine if a building block is really stable, it is insufficient to obtain a score of a single conformation as we have done here. Its native conformation needs to be compared to alternative ones. The only way to achieve this goal this is through molecular simulations. Such simulations, using the parallel tempering method, are currently underway.

Building Blocks Often Can Be Viewed as Stand-alone Units

Can then building blocks be referred to as stand-alone units, independent of their environmental context? Does a building block exhibit similar properties—tertiary and secondary structure and stability—if its sequence is placed within another protein and therefore in a different environment?

TABLE II. Relative positions of critical building blocks inside the protein structures†

Protein class	Cutting level	Five most common positions (percent, in descending order)
α/β	1	(5–10),(15–20),(65–70),(85–90),(90–95)
α/β	2	(90–95),(85–90),(5–10),(70–75),(20–25)
α/β	3	(75–80),(5–10),(70–75),(90–95),(85–90)
α/β	4	(75–80),(30–35),(70–75),(5–10),(10–15)
α/β	5	(30–35),(75–80),(5–10),(60–65),(70–75)
α/β	6	(75–80),(30–35),(65–70),(55–60),(5–10)
α + β	1	(20–25),(15–20),(40–45),(50–55),(5–10)
α + β	2	(15–20),(80–85),(70–75),(85–90),(40–45)
α + β	3	(80–85),(85–90),(15–20),(25–30),(50–55)
α + β	4	(80–85),(85–90),(15–20),(70–75),(65–70)
α + β	5	(70–75),(60–65),(15–20),(65–70),(80–85)
α + β	6	(70–75),(5–10),(65–70),(85–90),(60–65)
all-α	1	(85–90),(80–85),(10–15),(20–25),(15–20)
all-α	2	(70–75),(85–90),(80–85),(60–65),(10–15)
all-α	3	(80–85),(10–15),(85–90),(70–75),(35–40)
all-α	4	(10–15),(80–85),(85–90),(70–75),(90–95)
all-α	5	(60–65),(80–85),(85–90),(50–55),(10–15)
all-α	6	(65–70),(15–20),(55–60),(85–90),(35–40)
all-β	1	(10–15),(95–100),(65–70),(50–65),(70–75)
all-β	2	(85–90),(80–85),(5–10),(15–20),(65–70)
all-β	3	(80–85),(10–15),(85–90),(30–35),(15–20)
all-β	4	(80–85),(10–15),(35–40),(60–65),(65–70)
all-β	5	(30–35),(10–15),(35–40),(60–65),(5–10)
all-β	6	(30–35),(10–15),(35–40),(60–65),(5–10)

†The position inside the protein was calculated as:

$$\frac{\text{Building block midpoint residue number}}{\text{Total protein length}} \times 100$$

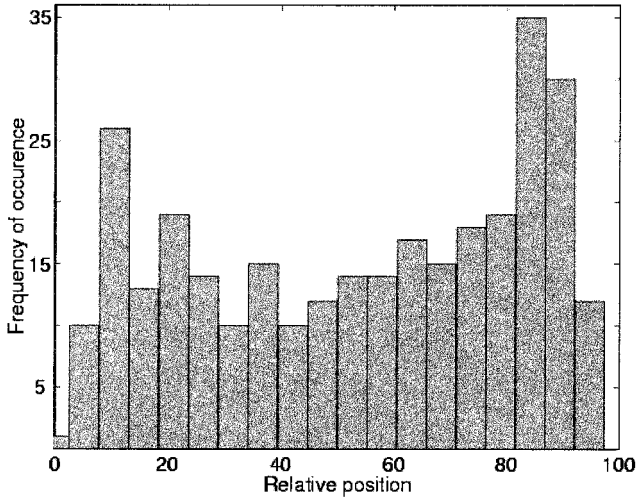

Fig. 7. Distribution of the critical building blocks with respect to their relative positions in α-proteins, level 3 (mean: 53.896; SD: 28.65).

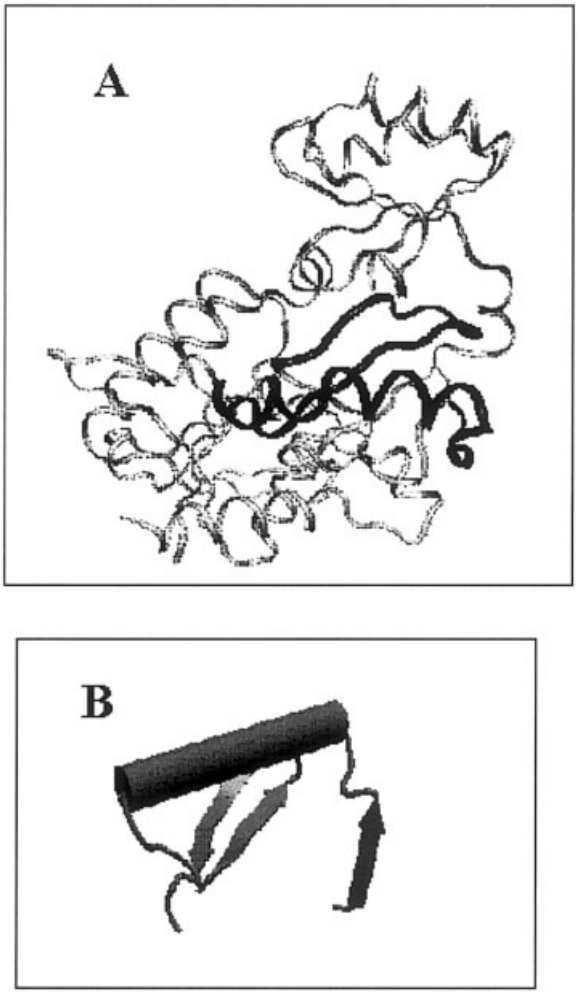

Fig. 8. An example of a critical building block for prokaryotic glutathione synthetase (pdb: 1gsh). The critical building block is located at the C-terminus of the protein. (A) The critical building block (dark) inside the protein. (B) The critical building block in secondary structure drawing.

The main tool used to address these questions is the building block clusters that contain large amounts of data. These vary considerably from one another. Although some clusters contain a single building block and others contain only building blocks from the same protein family, still other clusters contain building blocks derived from different families, varying in their sequences, stability scores, and criticalness. Such clusters are the most interesting ones. These clusters show that there are structural motifs that recur across different protein families and are relatively stable in each one. The fact that we have many such clusters shows that this is probably not an incident. If it is generally true that building blocks are independent units, then it should be possible to at least partially assign a set of building blocks with known structures to a target sequence, even if the sequence does not belong to any known structural family. This assignment is based on local sequence resemblance, relative stability, and possibly other properties such as secondary structure. If the assignment is correct, it would considerably reduce the complexity of structure prediction, because structures of building blocks are at least partially known and are preserved for this sequence, independent of the global sequence environment. We have already developed a prototype of an assignment algorithm that, given a protein sequence, finds the "optimal" building block

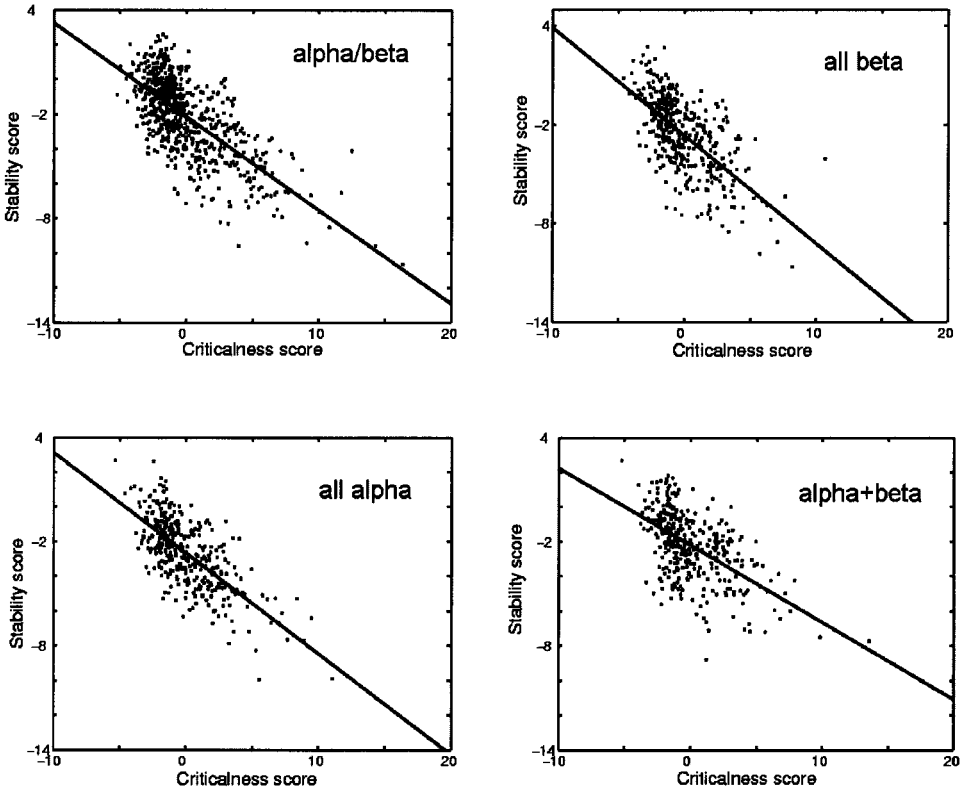

Fig. 9. Plots of the criticalness score vs. stability score for level 3 of all protein classes. Each graph point is an average of five data points. The straight line is a linear fitting of the data. Top left: α/β proteins, correlation coefficient -0.692. Top right: all-β proteins, correlation coefficient -0.654. Bottom left: all-α proteins, correlation coefficient -0.702. Bottom right: $\alpha+\beta$ proteins, correlation coefficient -0.561.

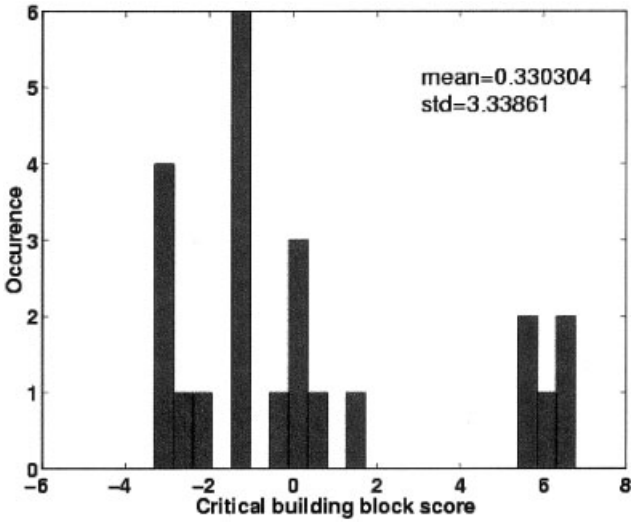

Fig. 10. An example of a broad distribution of criticalness scores among members of one cluster of α/β proteins. The sequence similarity (not shown) is weak (mean: 0.330; SD: 3.334).

assignment. The assignment is carried out by an alignment of that sequence against the building block database, followed by finding a set of matching building blocks, using a graph theoretic algorithm. Description of the algorithm will be presented elsewhere. The next, combinatorial assembly problem, involving putting these together in an optimal way, is the computationally most complex part of a hierarchical folding scheme.

Critical Building Blocks

Critical building blocks are largely buried. They play a critical role in correct folding and in maintaining the native structure. Because mutations may critically affect the structure, Ma et al.[17] and Kumar et al.[18] have proposed that critical building blocks tend to be more conserved, sequence and structure-wise, than other building blocks in the protein.

We find that critical building blocks may appear anywhere along the protein sequence, but they have a propensity toward the N and C termini. This is understandable given their role in protein folding. Remarkably, for all fold types, we find that critical building blocks are considerably less stable than other building blocks. This finding is not surprising. Though quite small, critical building blocks have extensive interfaces with other building blocks. Hence, on its own, a critical building block may be expected to be in a disordered protein state. This situation resembles that observed in a functional dimer. The two monomers are intertwined, with extensive interface. Yet, each monomer is relatively small. Pulling apart two-state dimers leads to disordered states. To have extensive interfaces and be

stable, a substantial increase in size would be required. This is counterproductive to the cell.

Additionally, their lower stability may reflect more hydrophobic residues on their surfaces. In the limited number of cases we have studied in detail, we have observed that building blocks that are critical for folding are also critical for function. Consistently, Luque and Friere[44] have observed that binding sites contain both stable and unstable areas. Elcock[45] has recently noted that charged residues in protein binding sites are mostly destabilizing. Instability at binding sites may be advantageous. The larger range of conformations at the site may be complementary to different enzymes, receptors, or ligands. This may explain both the preference of cleavage by a variety of proteases at certain sites[46] and the different, diverse ligands binding at the same site.[47] We have now undertaken a large-scale study of the potential relationship between folding and function. Our current results support this conservation for folding and for function paradigm. In proteins where critical building blocks have been identified, residues that are critical for protein function appear to be frequently located in critical building blocks.

Although critical building blocks may be more conserved than other building blocks within a protein family, there is no universal sequential or structural pattern for critical building blocks. Different protein families have critical building blocks with different secondary structure composition, conformations, and sequence patterns. However, because they are buried, they share a propensity for hydrophobic amino acids. Criticalness does not depend on the building block conformation, because building blocks from the same cluster (thus with similar conformations) have different criticalness scores. Finally, identification of a critical building block in a target sequence can reduce the complexity of the combinatorial assembly process in structure prediction. Given that critical building blocks are largely buried, if the assignment algorithm predicts that a certain fragment along the protein chain is a critical building block, we may have a clue to its 3D location.

CONCLUSIONS

Recently we have shown that results obtained by our computational building block cutting algorithm yield fragments that are in agreement with those obtained experimentally by limited proteolysis.[43]

Here our analysis indicates that building blocks are often stand-alone fragments, with folds repeating among different families, regardless of the overall structures, and with different protein sequences.

This leads us to conclude that the building block unit is frequently independent of the surrounding protein environment. Nevertheless, although a small building block also has a preferred conformation, we cannot rule out the possibility of multiple such conformations. Under such circumstances, the protein environment stabilizes the native conformer, increasing its population time. On the other hand, for larger, independently folding units with hydrophobic cores, the influence of the protein environment is limited.

The recurrence of building block conformations is reminiscent of protein folds. Combined with the agreement with limited proteolysis, this suggests that these fragments are able to fold independently, in principle enabling visualization of dynamic folding pathways and intermediate states. These findings validate protein folding schemes that are based on hierarchical folding, as first proposed by Rose.[12]

They suggest that it should be possible to develop an algorithm that predicts the building block assignment of a protein sequence whose structure is unknown. Here we have created sequentially nonredundant databases of building block sequences, against which a target protein sequence can be aligned in order to be matched to a set of potential building blocks.

We further probe building blocks that are critical for "correct" folding and for maintaining native protein structures. We find that although they can be located anywhere along the protein sequence, they are most likely to appear at chain termini. Although the critical building block sequence and structure may be conserved within a protein family, they do not have typical secondary structure composition, sequence or tertiary structure that tend to recur across families. The sole property shared among critical building blocks from different protein families is their larger proportion of hydrophobic amino acids compared to other building blocks, likely the outcome of their being buried in the protein core.

Criticalness scores of building blocks belonging to the same structural clusters distribute broadly. Hence, criticalness does not depend on the 3D structure. Remarkably, we further find that the higher the criticalness score, i.e., the more critical they are, the less stable. This makes inherent sense, as it enables large interfaces between building blocks in the protein core while still keeping "reasonable" protein sizes.

REFERENCES

1. Tsai CJ, Ma B, Kumar S, Wolfson H, Nussinov R. Protein folding: binding of building blocks via population selection. Crit Rev Biochem Mol Biol 2001;36:399–433.
2. Kim PS, Baldwin RL. Specifc intermediates in the folding reactions of small proteins, and the mechanism of protein folding. Annu Rev Biochem 1982;51:459–489.
3. Kim PS, Baldwin RL. Intermediates in the folding reactions of small proteins. Annu Rev Biochem 1990;59:631–660.
4. Udgaonkar JB, Baldwin RL. NMR evidence for an early framework intermediates on the folding pathway of ribonuclease A. Nature 1988;335:694–699.
5. Wetlaufer DB. Nucleation, rapid folding and globular intrachain regions in proteins. Proc Natl Acad Sci USA 1973;70:697–701.
6. Shakhnovitch E, Abkevitch V, Ptitsyn O. Conserved residues and the mechanism of protein folding. Nature 1996;379:96–98.
7. Fersht AR. Nucleation mechanism in protein folding. Curr Opin Struct Biol 1997;7:3–9.
8. Karplus M, Weaver DL. Protein folding dynamics: the diffusion-collision model and experimental data. Prot Sci 1994;3:650–668.
9. Rackovsky S, Scheraga HA. Hydrophobicity, hydrophilicity and the radial and orientational distributions of residues in native proteins. Proc Natl Acad Sci USA 1977;74:5248–5251.
10. Dill KA.Theory for the folding and stability of globular proteins. Biochemistry 1985;24:1501–1509.

11. Dill KA. Dominant forces in protein folding. Biochemistry 1990;9: 7135–7155.
12. Lesk AM, Rose GD. Folding units in globular proteins. Proc Natl Acad Sci USA 1981;78:4304–4308.
13. Baldwin RL, Rose GD. Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biol Sci 1999;24:26–33.
14. Baldwin RL, Rose GD. Is protein folding hierarchic? II. Folding intermediates and transition states. Trends Biol Sci 1999;24:77–84.
15. Tsai CJ, Maizel JV, Nussinov R. Anatomy of protein structures: visualizing how a 1D protein chain folds into a 3D shape. Proc Natl Acad Sci USA 2000;97:12038–12043.
16. Tsai CJ, Ma B, Sham Y, Kumar S, Wolfson H, Nussinov R. A hierarchical, building-blocks based computational scheme for protein structure prediction. IBM J Res Dev, issue on Life Sci 2001;45:513–523.
17. Ma B, Tsai CJ, Nussinov R. Binding and folding: in search of intramolecular chaperone-like building block fragments. Protein Eng 2000;13:617–627.
18. Kumar S, Sham YY, Tsai CJ, Nussinov R. Protein folding and function: the N-terminal fragmentin adenylate kinase. Biophys J 2001;80:2439–2454.
19. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein databank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
20. Lee B, Richards F. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
21. Shrake A, Rupley J. Environment and exposure to solvent of protein atoms. lysozyme and insulin. J Mol Biol 1973;79:351–371.
22. Fischer D, Nussinov R, Wolfson HJ. 3-d substructure matching in protein molecules. In: Third Symposium on Combinatorial Pattern Matching, Tucson, Arizona. Springer-Verlag. Lecture Notes in Computer Science 644.1992. p 136–150.
23. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540. http://scop.mrc-lmb.cam.ac.uk/scop/.
24. Pearson WR, Lipman DJ. Improved tools for biological sequence analysis. Proc Natl Acad Sci USA 1988;85:2444–2448. Fasta: University of Virginia, Fasta world wide web URL, http://fasta.bioch.virginia.edu.
25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410. Blast: NCBI, Blast world wide web URL, http://www.ncbi.nlm.nih.gov/BLAST/.
26. Leibowitz N, Fligelman Z, Nussinov R, Wolfson H. Automated multiple structure alignment and detection of a common structural motif. Proteins 2001;43:235–245.
27. Gunasekaran K, Tsai CJ, Kumar S, Zanuy D, Nussinov R. Extended disordered protein: function with less scaffold. 2002. Submitted for publication.
28. Hardin C, Pogorelov TV, Luthey-Schulten Z. Ab initio protein structure prediction. Curr Opin Struct Biol 2002;12:176–181.
29. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins 2001;5(Suppl):119–126.
30. Koretke KK, Luthey-Schulten Z, Wolynes PG. Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. Prot Sci 1996;5:1043–1059.
31. Sasai M, Wolynes PG. United theory of collapse, folding and glass transitions in associative memory. Hamiltonian models of proteins. Phys Rev A 1992;46:7979–7997.
32. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
33. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301:173–190.
34. Oliva B, Bates P, Quero E, Aviles F, Sternberg MJ. An automated classification of the structure of protein loops. J Mol Biol 1997;266: 814–830.
35. Osguthorpe DJ. Ab initio protein folding. Curr Opin Struct Biol 2000;10:146–152.
36. Hardin C, Eastwood MP, Prentiss M, Luthey-Schulten Z, Wolynes PG. Folding funnels: the key to robust protein structure prediction. J Comput Chem 2002;23:138–146.
37. Yue K, Dill K. Constraint-based assembly of tertiary protein structures from secondary structure elements. Prot Sci 2000;9: 1935–1946.
38. Eastwood MP, Hardin C, Luthey-Schulten Z,Wolynes PG. Evaluating protein structure-prediction schemes using energy landscape theory. IBM J Res Dev 2001;45:475–497.
39. Betancourt MR, Skolnick J. Finding the needle in a haystack: deducing native folds from ambiguous ab initio protein structure predictions. J Comput Chem 2001;22:339–353.
40. Skolnick J, Kolinski A, Kihara D, Betancourt MR, Rotkiewicz P, Boniecki M. Ab initio protein structure prediction via a combination of threading, lattice folding, clustering and structure refinement. Proteins 2001;5(Suppl):149–156.
41. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 2001;44:133–149.
42. Reva BA, Skolnick J, Finkelstein AV. Averaging interaction energies over homologues improves protein fold recognition in gapless threading. Proteins 1999;45:353–359.
43. Tsai CJ, Polverino de Laureto P, Fontana A, Nussinov R. Comparison of protein fragments identified by limited proteolysis and computational cutting of proteins. Prot Sci 2002;11:1753–1770.
44. Luque I, Friere E. Structural stability of binding sites: consequences for binding affinity and allosteric effects. Proteins 2000; Suppl 4:63–71.
45. Elcock AH. Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 2001;312, 885–896.
46. Fontana A, Polverino de Laureto P, De Filippis V, Scaramella E, Zambonin M. Probing the partly folded states of proteins by limited proteolysis. Fold Des 1997;2:R17–R26.
47. Ma B, Shatsky M, Wolfson H, Nussinov R. Multiple ligands binding at single site: a matter of pre-existing populations. Prot Sci 2002;11:184–197.