Potential folding-function interrelationship in proteins

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · SEPTEMBER 2004

Impact Factor: 2.63 · DOI: 10.1002/prot.20132 · Source: PubMed

4 AUTHORS, INCLUDING:

Sandeep Kumar
Pfizer Inc.
75 PUBLICATIONS **4,470** CITATIONS
SEE PROFILE

Haim J Wolfson
Tel Aviv University
209 PUBLICATIONS **13,801** CITATIONS
SEE PROFILE

Ruth Nussinov
Tel Aviv University
624 PUBLICATIONS **27,983** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Bioinformatics 56:635–649 (2004)

Potential Folding–Function Interrelationship in Proteins

Adi Barzilai,[1] Sandeep Kumar,[2] Haim Wolfson,[3] Ruth Nussinov[1,4]*

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

[2]*Laboratory of Experimental and Computational Biology, National Cancer Institute–Frederick, Frederick, Maryland*

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

[4]*Basic Research Program, SAIC–Frederick, Inc., Laboratory of Experimental and Computational Biology, National Cancer Institute–Frederick, Frederick, Maryland*

ABSTRACT **The possibility is addressed that protein folding and function may be related via regions that are critical for both folding and function. This approach is based on the building blocks folding model that describes protein folding as binding events of conformationally fluctuating building blocks. Within these, we identify building block fragments that are critical for achieving the native fold. A library of such critical building blocks (CBBs) is constructed. Then, it is asked whether the functionally important residues fall in these CBB fragments. We find that for over two-thirds of the proteins in our library with available functional information, the catalytic or binding site residues lie within the CBB regions. From the evolutionary standpoint, a folding–function relationship is advantageous, since the need to guard against mutations is limited to one region. Furthermore, conformationally similar CBBs are found in globally unrelated proteins with different functions. Hence, substituting CBBs may lead to designed proteins with altered functions. We further find that the CBBs in our library are conformationally unstable. Proteins 2004; 56:635–649.** © **2004 Wiley-Liss, Inc.**

Key words: protein folding; protein function; protein building block; protein design; folding and function; structure and function

INTRODUCTION

To function, proteins have to be in their specific three-dimensional (3D) conformations to interact with other molecules. This recognition has led to intensive investigations of the challenging relationship between protein structure (or fold) and function. Moult and Melamud[1] pointed out that functional information may be deduced from a structure, which is based on a fold relationship to an already known structure, or from an analysis of the features of the structure. Using CATH, Martin et al.[2] found some correlation between the protein class and enzyme function at the top Enzyme Commission (EC) classification level. A better correlation was observed between the protein class and ligand type (heme, carbohydrate, DNA, or nucleotide) since ligand molecules of a certain shape or polarity are recognized by certain protein folds. Comparisons of enzymes and nonenzymes in Swissprot with the domains in SCOP indicated that major SCOP fold classes tend to associate with certain functional categories.[3] Finkelstein et al.[4] rationalized that because a given structure can be obtained by different sequences, a fold should perform different functions. This phenomenon of the same fold and different functions has been widely studied. Nagano et al.[5] focused on the TIM-barrel family that shares the same fold but has diverse catalytic reactions. Dokholyan and Shakhnovich[6] and Mirny and Shakhnovich[7] proposed that function, folding kinetics, and stability are essential for proteins and nature exerted evolutionary pressure to preserve them. They developed a model to explain the hierarchical organization of proteins in fold families and predicted the pattern of amino acids conserved across protein families.

Several groups have devised methods to predict the function of a protein given its 3D structure. Some use evolutionary information while others are ab initio methods. The evolutionary trace method[8] defines binding surfaces within protein families. Aloy et al.[9] developed a structure-based algorithm to locate functional residues on a protein structure using sequence conservation. Kasuya and Thornton[10] searched for sequence patterns identified by PROSITE in known structures. Russell[11] argued that recurring 3D side-chain patterns could be used to detect protein function. Russell et al.[12] developed a method to identify structurally similar "supersites" within superfolds. Jackson and Russell[13,14] identified potential binding sites of serine protease inhibitors on the surface of known protein structures. Based on the principle of scanning a structure against a collection of structural motifs associ-

Grant sponsors: Ministry of Science (to R.N. and H.W.); Israel Science Foundation (to R.N. and H.W.); Hermann Minkowski–Minerva Center for Geometry, Tel Aviv University (to H.W.).

Grant sponsor: National Cancer Institute, National Institutes of Health; Grant number: NO1-CO-12400.

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

Received 8 September 2003; Revised 3 December 2003; Accepted 23 January 2004

Published online 11 June 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20132

ated with functions, Wallace et al.[15] scanned a structure against a library of tertiary templates to detect functional sites. Ab initio methods include functional prediction based on structural features (reviewed in Thornton et al.[16]). One example is the work of Laskowski et al.,[17] who identified ligand binding sites by scanning the protein surface for clefts. Fetrow and colleagues[18,19] suggested that functional sites should be used in models created by both ab initio and threading methods. Additional methods for deriving function from structure are described in the extensive reviews of Thornton et al.,[16] Orengo et al.,[20] Moult and Melamud,[1] Shapiro and Harris,[21] and Skolnick et al.[22]

Above we have briefly reviewed studies connecting protein structure and function. Yet, there is a third component relating to structure and to function, which is the protein folding process. Folding, structure, and function can frustrate or cooperate with one another.[23] Optimal function may frustrate folding by leading to unfavorable interactions that lower protein stability, as can be observed in directed evolution selecting for higher catalytic rates or in simulations where a building block fragment containing a functional loop region is removed. In contrast, folding and function can also cooperate with each other. Hinges, which are often crucial for function, tend to recur between elements that fold on themselves,[24] consistent with hierarchical protein folding[25–28] and with the building block folding model.[29,30] Here, we study the interrelationship between folding and function, making use of protein fragments predicted to be stand-alone elements of the protein structure.[31,32] Protein fragments may be produced experimentally[33–36] or computationally.[29,37,38] Protein fragments are frequently used in studies related to protein folding, structure, and function (e.g., refs. 39–44). Tsai et al.[29,30,45] have developed an algorithm to iteratively dissect a protein into a set of building blocks. The building block folding model[29] follows the concept of hierarchical protein folding. According to this concept, folding initiates locally, with folded elements assembling stepwise to yield the native fold. The model postulates that the driving force for protein folding is the hydrophobic effect. Through combinatorial assembly and mutual stabilization, conformationally fluctuating building blocks associate to form independently folding, compact, hydrophobic folding units. These associate to form domains and subunits. All building blocks are required for the native fold. Nevertheless, some are "more critical" than others. Critical building blocks (CBBs) tend to be buried in the protein core, frequently being spatially inserted between sequentially adjacent building blocks. We developed a computational procedure to identify a CBB in a given protein 3D structure based on the identity and number of building blocks with which it interacts, the polar and nonpolar surface areas buried by such interactions, and its location in the protein.[46] This procedure (briefly described in the Methods section) was used to identify potential CBBs in proteins. Adenylate kinase was selected for further studies in our laboratory. Molecular dynamics simulations on yeast adenylate kinase with the predicted N-terminal CBB (residues 1–36) removed lead to a shrunken, misassociated nonnative structure; however, they have native-like conformations for all remaining building blocks in the structure, consistent with the building block folding model and the cutting algorithm.[46]

Here, we ask whether such CBBs also contain residues that are important for protein function. Our initial results for dihydrofolate reductase[47] and adenylate kinase[46] have indicated that, at least for these two cases, building blocks that appear critical for folding may also be critical for function. Especially in the case of adenylate kinase, the N-terminal CBB contains the ATP-binding P-loop that is highly conserved in all mononucleotide kinases.[46]

CBBs appear to resemble the pro-region, which also fulfills a dual folding–function role. Like the pro-region, on their own they are also unstable. However, unlike the pro-region, they are not cleaved, because they are essential for the preservation of the native fold. Such a relationship appears logical from the evolutionary standpoint. CBBs are essential for folding. Hence, there has been pressure for conservation of their sequence. Function took advantage of this evolutionary pressure. Alternatively, because they are essential for function, evolution made use of it for obtaining the native fold.

In previous research we used the energy landscape to relate folding and function.[48–50] To address the folding–function question, we recently analyzed 930 sequence- and structurewise nonhomologous protein chains with known tertiary structures derived from the Protein Data Bank (PDB).[51] This led to the identification of 225 CBBs in the proteins, yielding a CBB library.[52] Our work here presents an analysis of this library. There are 67 protein chains for which the PDB files contain the SITE records describing their functional residues. Each of the 67 protein chains contains only one CBB. Large protein chains may contain more than one CBB, frequently one CBB per domain. In 46 of these 67, one or more functional (catalytic) residues fall in the CBB of the protein. Thirty-seven out of these 67 protein chains have characterized conserved regions in the PROSITE database.[53,54] Among these 37, 16 protein chains show coincidence between the PROSITE patterns and the CBB regions. Overall, our analysis appears to suggest that the CBBs may relate to both the folding and function of the protein in at least two-thirds of the proteins that we have studied, which are summarized in Figure 1 and Table I. For three cases (C-type lysozyme, bacterial lipase, and fungal lipase), to further test a potential folding–function coupling, we assemble protein families with homologous structures and carry out a detailed study of the structure and sequence conservation in all family members. We describe the results and highlight the caveats associated with this type of analysis. These largely relate to the empirical computational building block cutting algorithm, the empirical parameters defining a CBB and the reliability of the SITE records in the PDB files. The PROSITE patterns are also not fault free. These hamper a rigorous statistical analysis. Nevertheless, while bearing these caveats in mind, combined, available data suggest that folding and function may be coupled for some proteins.

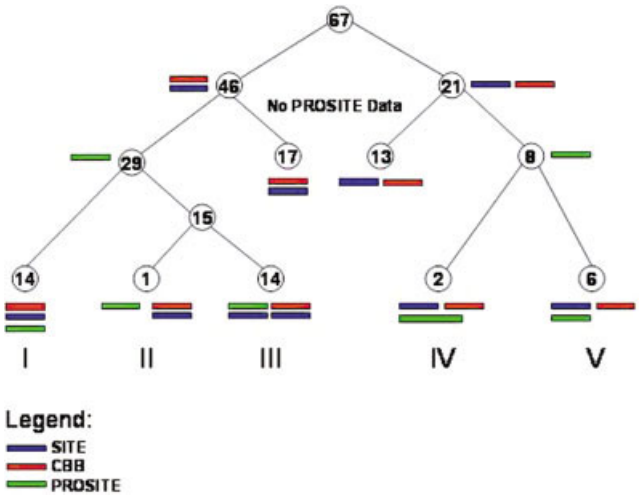

Fig. 1. A schematic representation of the data set categories. Starting with the 67 CBBs having SITE records in their PDB files, in 46 protein chains the CBBs contain at least one functional residue (left branch) and 21 do not (right branch). Next, we split the data according to the availability of PROSITE annotation. Among the 46 and 21 chains, 29 and 13 chains have available data, respectively; 14 protein chains out of the 29 (left leaf) show a correlation between the CBBs, SITE functional residue(s), and the PROSITE conserved region. These cases are classified as category I. The remaining categories are described in Table II.

METHODS

Building Blocks and Their Empirical Stability Scores

We have used an algorithm, Anatomy, developed by Tsai et al.[29,30,45] to iteratively dissect a protein into a set of stand-alone building blocks. The algorithm and the computational cutting procedure have been described over the last few years, and the details of the algorithm and its rationale are given in ref. 29. It has been used to analyze and predict the folding complexity of sequential versus nonsequential folding proteins.[55] Tsai et al.[30] have further reviewed the literature and showed its consistency with the independently folding building block concept and definition. Most recently, Haspel et al.[31,32] have clustered all the building blocks, studied the characteristic properties of the clustered building block library[31] and applied these clusters in a first step toward protein folding.[32] The comparison of the computational cutting with experimental limited proteolysis experimental cutting[56] and with hydrogen exchange data[30,56] has been undertaken to inspect its agreement with the experiment. There is no preset length to the building blocks. The only stipulation is that the minimal length be at least 15 residues to allow for a hydrophobic core. The algorithm can perform a cutting independently of window sizes because the scoring function is independent of the fragment length.[29] The most serious drawback in the identification of a building block is the fact that the algorithm is based solely on the native state of the protein. In addition, in the scoring function we do not account for electrostatic interactions. Briefly, for a given protein 3D structure, all fragment candidates are assigned a stability score calculated by the building block scoring function. To collect a basket of building blocks, we locate all local minima on the fragment map. We define a local minimum in the simplest way, that is, a local minimum should have the highest value in a defined local region. Thus, all candidate fragments with minimum length (15 residues) are tested for their stability score and the local minima are candidate building blocks. This process continues iteratively until the building blocks can no longer be dissected. The different levels are referred to as cutting levels. The empirical stability score of a building block is evaluated by the following scoring function, which is independent of the length of the building block:

$$\begin{aligned}\text{score}^{B,B}(Z,H,I) = {} & (Z^1_{\text{Avg}} - Z)/Z^1_{\text{Dev}} + (H^1_{\text{Avg}} - H)/H^1_{\text{Dev}} \\ & + (I^1_{\text{Avg}} - I)/I^1_{\text{Dev}} + (Z^2_{\text{Avg}} - Z)/Z^2_{\text{Dev}} + (H^2_{\text{Avg}} - H)/H^2_{\text{Dev}} \\ & + (I^2_{\text{Avg}} - I)/I^2_{\text{Dev}}, \quad (1)\end{aligned}$$

where Z is the compactness, I is the degree of isolation, and H is the hydrophobicity of the building block. The Z value is the solvent accessible surface area (ASA) divided by its minimum possible surface area (the area of a sphere with an equal volume) of the fragment. The I value is the ratio of the fragment's nonpolar ASA that was originally buried but is exposed to the solvent after cutting to the ASA of the isolated fragment. The H value is the fraction of the buried nonpolar area out of the total nonpolar area of the fragment. The average and standard deviation were calculated from a data set of 930 dissimilar protein chains. Terms with a superscript 1 were determined with respect to fragment size. Those with a superscript 2 are a function of the fraction of the fragment size to the whole protein. Note that the building block score does not refer to the thermodynamic stability of the building block in solution.

Identification of CBBs, Criticalness Score, and CBB Library

Kumar et al.[46] have developed an algorithm to identify CBB(s) in a protein using the set of building blocks generated by the Anatomy program at different levels of cutting. The program examines whether a protein contains a potential CBB(s). At each level of building block cutting, a building block is assigned a "criticalness" score (CIndex) based on the identity and number of building blocks with which it interacts, the polar and nonpolar surface areas buried by such interactions, and its location in the protein. Consider a building block a such that it interacts with building blocks b and c. Its CIndex(a) is given by

$$\text{CIndex}(a) = \sum \text{diffcontsa}(a)/\text{totsa}(a)*\text{protburyasa}(a)/\text{solyexpsa}(a), \quad (2)$$

where

$$\text{diffcontsa}(a) = \text{contsa}(a,b) + \text{contsa}(a,c) - \text{contsa}(b,c), \quad (3)$$

in which contsa(a, b) is the surface area buried between building blocks a and b. Similarly for contsa(a, c) and contsa(b, c), diffcontsa(a) is the differential contacting sur-

TABLE IA. The Data Set of 67 Protein Chains Containing SITE Record in PDB

PDB entry	Protein name	Functional class	Protein size	CBB	CBB-SITE	PROSITE entry and pattern position	
1aa6	Formate dehydrogenase H	Enzyme	696	Ile103-Asp134	+	PS00490	Thr433-Glu450
						PS00551	Thr6-Asn23
						PS00932	Ala615-Ser642
1abrB	Abrin-A	Binding	267	Leu110-Trp136	+		—
1ac5	Kex1p serine carboxypeptidase	Enzyme	483	Leu365-Val391	+	PS00131	Leu172-Gly179
						PS00560	Leu438-Ser455
1aco	Aconitase	Enzyme	753	Ile146-Ala186	+	PS00450	Ile350-Met366
						PS01244	Gly413-Gly426
1ad3A	Aldehyde dehydrogenase	Enzyme	446	Val106-Met145	−	PS00070	Phe236-Asp247
						PS00687	Leu208-Pro215
1aihA	Hp1Integrase	Enzyme	170	Asp193-Thr213	+		—
1aorA	Aldehyde ferredoxin oxidoreductase	Enzyme	605	Ala432-Lys450	−		—
1ast	Astacin	Enzyme	200	His181-Leu200	−	PS00142	Thr89-Ile98
1asu	Integrase	Enzyme	162	Gln62-Thr97	+		—
1atlA	Atrolysin C	Enzyme	200	Ser130-Gly149	+	PS00142	Thr139-Leu148
1bco	Bacteriophage Mu transposase	Enzyme	295	Trp290-Ser308	−		—
1bp1	Bactericidal/Permeability-increasing protein	Binding	456	Val1-Tyr16	+	PS00400	Pro3-Pro35
1btkB	Bruton's tyrosine Kinase	Enzyme	169	Tyr100-Asn130	−		—
1cex	Cutinase	Enzyme	197	Met98-Ser129	+	PS00155	Pro110-Gly122
						PS00931	Cys171-His188
1chd	Cheb methylesterase	Enzyme	198	Leu158-His190	+		—
1chkA	Chitosanase	Enzyme	238	Asn115-Val148	−		—
1cmvB	Human cytomegalovirus protease	Enzyme	206	Tyr128-Arg175	+		—
1cvl	Triacylglycerol hydrolase	Enzyme	316	Lys80-Thr112	+	PS00120	Val81-Gly90
1daaA	D-Amino acid aminotransferase	Enzyme	277	Ile17-Thr43	+	PS00770	Glu177-Arg206
1def	Peptide deformylase	Enzyme	147	Leu125-Ser147	+		—
1dim	Sialidase	Enzyme	381	Gly26-Ser61	+		—
1doi	2Fe-2s Ferredoxin	Transport	128	Phe61-Asp79	+	PS00197	CYS63-CYS71
1dtp	Diphtheria toxin catalytic domain	Enzyme	190	Thr89-Thr111	−		—
1ecfA	Glutamine phosphoribosylpyrophosphate	Enzyme	492	Ala157-Ile173	−	PS00103	Val362-Thr374
						PS00443	Cys1-Gly5
1esl	E-selectin	Binding	157	His25-Ile51	−	PS00022	Cys142-Cys153
						PS00615	Cys90-Cys117
1fieB	Coagulation factor Xiii	Enzyme	715	Ala332-Leu348	−	PS00547	Gly312-Gly329
1han	2,3-Dihydroxybiphenyl 1,2-dioxygenase	Enzyme	288	Gly239-Arg266	+	PS00082	Gly239-Glu260
1iba	Glucose permease	Enzyme	78	Asp50-Ile72	−	PS01035	Asn28-Asp45
1imbA	Inositol monophosphatase	Enzyme	273	Asn199-Gly232	+	PS00629	Trp87-His100
						PS00630	Trp219-Gly233
1ivd	Sialidase	Enzyme	388	Ser101-Gly135	+		—
1kcw	Ceruloplasmin	Enzyme	1017	Met601-Gly631	+	PS00079	Gly313-Phe333
							Gly674-Tyr694
						PS00080	Gly1015-Tyr1035
							His1020-Met1031
						PS00205	Tyr92-Gly101
							Tyr93-Gly101
							Tyr435-Ser444
1lcf	Lactoferrin	Transport	691	Tyr528-Thr549	+	PS00206	Tyr192-Phe208
							Tyr528-Phe544
						PS00207	Glu226-Val256
							Asp570-Val600
1mkaA	β-Hydroxydecanoyl thiol ester dehydrase	Enzyme	171	Gly82-Tyr126	+		—
1mla	Malonyl-coenzyme	Enzyme	305	Ser92-Arg117	+		—
1mtyD	Methane	Enzyme	512	Leu348-Leu363	−		—
1nipB	Nitrogenase	Enzyme	287	Ser16-Gln54	+		—
1oacB	Copper amine oxidase	Enzyme	722	Glu454-Gly488	+	PS01165	Thr684-Pro697
1opc	Ompr	Binding	99	Leu161-Pro179	−		—

TABLE IA. (Continued)

PDB entry	Protein name	Functional class	Protein size	CBB	CBB-SITE	PROSITE entry and pattern position	
1osa	Calmodulin	Binding	148	Ser81-Ile100	+	PS00018	Asp20-Leu32 Asp56-Phe68 Asp93-Leu105 Asp129-Phe141
1pbn	Purine nucleoside phosphorylase	Enzyme	289	Thr221-Ser239	+	PS01240	Val79-Leu120
1pii	*N*(5′phosphoribosyl) anthranilate isomerase	Enzyme	452	Ile111-Cys134	+	PS00614	Phe50-Ile64
1pkp	Ribosomal protein S5	Ribosomal	145	Lys126-Leu147	−	PS00585	Gly28-Ile60
1pth	Prostaglandin H2 synthase-1	Enzyme	551	Leu295-Asn310	−		—
1put	Putidaredoxin	Transport	106	Cys85-Trp106	+	PS00814	Cys39-His49
1pyaB	Pyruvoyl-dependent histidine decarboxylase	Enzyme	228	Thr84-Lys103	+		—
1que	Ferredoxin-Nadp^{+} reductase	Enzyme	303	Leu142-Tyr165	−		—
1rcy	Rusticyanin	Transport	151	Thr71-Phe87	+	PS00196	Ala131-Met148 Gly132-Met148
1smnA	Extracellular endonuclease	Enzyme	241	Ser22-Ala43	−	PS01070	Asp86-Ala94
1tca	Triacylglycerol lipase	Enzyme	317	Asp75-Tyr135	+		—
1thtA	Thioesterase	Enzyme	294	Gly109-Val138	+		—
1utg	Uteroglobin	Binding	70	Ile2-Pro18	+	PS00404	Gln50-Cys69
1vhh	Sonic hedgehog	Binding	157	Ala169-Ser185	+		—
1whi	Ribosomal Protein L14	Ribosomal	122	Ile2-Val21	+	PS00049	Ala60-Ile86
1xikA	Protein R2 Of Ribonucleotide Reductase	Enzyme	340	Lys191-Val210	+	PS00368	Ser114-Val130
1zia	Pseudoazurin	Transport	124	Thr30-Tyr82	+	PS00196	Gly72-Met84 Gly72-Met86
2abk	Endonuclease iii	Enzyme	211	Pro168-Cys187	+	PS00764 PS01155	Cys187-Cys203 Gly102-Gly131
2azaA	Azurin	Transport	129	Thr96-Leu127	+	PS00196	Gly105-Met120 Gly105-Met121 Ala107-Met120 Ala107-Met121
2bbkL	Methylamine dehydrogenase	Enzyme	125	Trp13-Leu43	−		—
2ebn	Endo-*N*-acetylglucos-aminidase F1	Enzyme	285	Lys11-Ser46	−		—
2mcm	Macromomycin	Binding	112	Val54-Val76	+		—
2mtaC	Cytochrome C55li	Transport	147	Gly82-Met117	+	PS00190	Cys57-Gly62
2vik	Villin 14t	Binding	126	Ile18-Ser49	+		—
3dni	Deoxyribonuclease I	Enzyme	258	Phe128-Asn145	+	PS00918 PS00919	Gly167-Ser174 Ile130-Val150
3pte	D-Alanyl-*d*-alanine carboxypeptidase/transpeptidase	Enzyme	347	Arg317-Phe343	−		—
4blmA	β-Lactamase	Enzyme	256	Ser126-Gly144	+	PS00146	Phe66-Leu81
4pgaA	Glutaminase-asparaginase	Enzyme	330	Gly93-Ser125	+	PS00144 PS00917	Ile14-Ala22 Gly93-Leu103
8rucA	Ribulose-1,5-bisphosphate carboxylase/oxygenase	Enzyme	466	Leu375-Gly405	−		—

The proteins are in alphabetical order. Protein names are taken from the header record of the PDB files, and protein sizes are the number of residues having coordinates. The location of the PROSITE pattern on the structure is taken from PDBsum database. +, the CBB and the SITE record overlap; −, when they do not.

face area of building block a buried by building blocks b and c, subtracting the contacting surface area between b and c; and totsa(a) is the total surface area of building block a. It has two components, the surface area buried by the rest of the protein [protburyasa(a)] and the surface area exposed to the solvent [solvexpasa(a)]. The surface area calculation involved both the polar and nonpolar surfaces.[46]

The significance of the CIndex value for building block a is measured by its Z score. At each level of cutting, the average (μ) and standard deviation (σ) for all building block CIndex values are calculated. The Z score measures the difference of the CIndex value of building block a at level k from the average CIndex value for that level.

$$Z\ \text{score}_{(ak)} = (\text{CIndex}_{(ak)} - \text{AvCIndex}_{(k)})/S_{(k)}. \qquad (4)$$

The Z score of a building block with the greatest CIndex at the lowest level of cutting increases with the protein size

and hence with the number of building blocks.[52] The t score for building block a at level k is defined as

$$t\ \text{score}_{(ak)} = Z\ \text{score}_{(ak)} \times \sqrt{\text{NumBB}_{(k)}}, \quad (5)$$

where NumBB(k) is the number of building blocks at the kth cutting level. A building block is considered critical if it is found at most levels of the protein cutting and has consistently high CIndex values, Z scores, and t scores.

The algorithm was applied to a nonredundant data set of 930 protein chains. Of these, 756 chains contain building blocks with a t score of at least >1.0 at the lowest level. At the lowest level of building block cutting, the average and standard deviation in the maximum Z score and maximum t score was calculated for a given number of building blocks. A protein with N building blocks at the lowest level of cutting K is selected in the CBB library if its maximum t score is greater than the average plus 1σ of the maximum t score expected for the N building blocks.

$$\max\{t\text{score}_{(1K)}, t\ \text{score}_{(2K)}, \ldots, t\ \text{score}_{(NK)}\} > \text{AvMax}t\ \text{score}(N) + 1\sigma. \quad (6)$$

Using these criteria, a library of 225 CBBs was constructed and reported elsewhere.[52] Many of the proteins in our CBB library have been well studied and a large number of high resolution crystal structures are available for many of these proteins. The examples include hen egg white lysozyme, flavodoxin, sialidase, β-lactamase, tyrosine kinase, α-lytic protease, calmodulin, α-amylase, c-H Ras p21 protein, ferredoxin, diphtheria toxin, rubisco, reverse transcriptase, lipase, class I major histocompatibility complex, porin, alkaline phosphatases, and so forth. Hence, a variety of proteins may contain CBBs. Most of the proteins in the library fold in a complex nonsequential manner, particularly the midsized and large ones. Many of the CBBs lie at or near the N- or the C-termini of the polypeptide chains. Larger proteins may contain more than one CBB with different CBBs in different protein domains. In some cases, a single domain contains more than one CBB. Although all the CBBs are located in the protein core, a mere presence in the core is an insufficient condition for a building block to be critical. Because the minimal size of a building block is 15 residues, no building block is completely buried, that is, the ASA never approaches zero.

We are currently studying the proteins whose CBBs are in our library. The data set used in this study consists of 67 CBBs in as many protein chains for which information on functional residues is available from the respective PDB entry SITE records.

Database for Protein Folding and Function: Comparison Among Locations of PDB SITE Record Residues, PROSITE Patterns, and CBBs

Our database contains protein chains in the CBB library if their PDB files contain a SITE record. The SITE record describes important functional groups and may include heteroatoms. In 67 protein chains in the library, the PDB files contain the SITE records. These include 1aa6, 1abrB, 1ac5, 1aco, 1ad3A, 1aihA, 1aorA, 1ast, 1asu, 1atlA, 1bco, 1bp1, 1btkB, 1cex, 1chd, 1chkA, 1cmvB, 1cvl, 1daaA, 1def, 1dim, 1doi, 1dtp, 1ecfA, 1esl, 1fieB, 1han, 1iba, 1imbA, 1ivd, 1kcw, 1lcf, 1mkaA, 1mla, 1mtyD, 1nipB, 1oacB, 1opc, 1osa, 1pbn, 1pii, 1pkp, 1pth, 1put, 1pyaB, 1que, 1rcy, 1smnA, 1tca, 1thtA, 1utg, 1vhh, 1whi, 1xikA, 1zia, 2abk, 2azaA, 2bbkL, 2ebn, 2mcm, 2mtaC, 2vik, 3dni, 3pte, 4blmA, 4pgaA, and 8rucA. In two cases, 2mtaC and 1pyaB, the SITE records include small heteroatoms, which are the heme group and pyruvolyl group cofactor, respectively. In these cases, the information on the functional residues was extracted from the original crystallographic articles.[57,58] Each entry in the database contains information on the protein type; SCOP fold classification[59]; folding complexity[55]; conserved residues identified as signature by the PROSITE database[52,59]; functional residues reported by the SITE record of the PDB file; building blocks at the lowest level; and the CBBs with their respective CIndex score, Z score, and t score. The PROSITE patterns were extracted using the PDBsum database.[17,60] To determine whether there is coupling between protein folding and function, we search for cases where one or more functional residues falls within the CBB region. We consider a PROSITE pattern to be near a CBB if it lies within seven residues of either of the CBB termini. This is because our building blocks assignment is uncertain by seven residues at each terminus.

Analysis of Protein Families for Folding and Function Coupling

To further test the potential relationship between folding and function, we have carried out an in-depth protein family analysis for three protein families. Therefore, we have selected protein families containing several homologous structures and asked how well the CBBs are conserved structure- and sequencewise within members of the family. We have cut all family members into building blocks and computed the CIndex on all building blocks of each structure. In addition, we have looked for consistency in the assignment of the highest criticalness. Next, we have superimposed the building blocks of each family member on the corresponding building blocks of the representative of the family and compared the respective root mean square deviations (RMSDs). We have further carried out multiple sequence alignments of the family members using SEQLAB and the conserved regions were marked using PLOTSIMILARITY. This analysis has been carried out for three families: 253 C-type lysozyme structures and 27 additional sequences from the SRS database, 15 bacterial lipase structures supplemented by 25 bacterial lipase sequences retrieved from the SRS, and 18 fungal lipase structures supplemented by 10 additional sequences detected through a database search. The structures in a family are of the same protein, with or without the ligand, at different temperatures in the presence or absence of one or more substrates and in different crystallographic space groups. We have used an in-house developed clustering program to create groups of similar conformation building blocks.

RESULTS

Description of Data Set

Sixty-seven protein chains in the CBB library[52] have available data on functionally important residues in the SITE records of their PDB files. The data set (Table I) contains proteins with diverse functional classes: 49 enzymes, nine binding proteins, seven transport proteins, and two ribosomal proteins. In binding proteins, the SITE records consist of residues involved in binding carbohydrates (PDB entry 1abr), actin (2vik), chromophores (2mcm), metal ions (zinc-1vhh, calcium-1osa, 1esl), steroids (1utg), DNA (1opc), and lipopolysaccharides (1bp1). Electron transport protein (1doi, 1put, 2aza, 1zia, 1rcy, and 2mtaC) SITE records contain residues involved in metal ion binding. Two ribosomal proteins (1whi and 1pkp) contain large binding sites, constituting 12% of the protein size. The SITE records in the PDB entries are based on the data supplied by the crystallographers. Frequently, these records do not contain all the residues that form an active site around the protein substrate. These records are often a subset of the active site residues that are directly involved in protein function. For example, for enzymes, the SITE records usually refer to the catalytic residues that number only a few (usually 3–5). Although the SITE records tend to be inconsistent across the PDB entries, they are still valuable because they present the annotations of the authors working on these proteins.

The proteins in our data set belong to various structural classes. According to SCOP, most of our proteins have mixed α and β structures: there are 21 α + β chains and 20 α/β protein chains. There are 8 all-α and 13 all-β chains. Four structures are multidomain and one can be considered as a small protein.

Folding and Function Coupling

The average size of the protein chains in our database is 302 ± 194 residues. In comparison, the average size of the CBBs is 27 ± 9 residues. Figure 2 plots the size of the CBBs versus the size of the proteins. Most CBBs are shorter than 40 residues, regardless of the protein size. In 46 out of the 67 (69%) protein chains, the CBBs also contain one or more functionally important (i.e., catalytic or binding site) residues. We further verified our results by using the pages of ligplots of interactions wherever available at the PDBSum database (www.biochem.ucl.ac.uk/bsm/pdbsum/). Ligplots are automatically generated plots of protein–ligand interactions that are computed using HBplus.[61] In most cases, the CBB contains several active or binding site residues. Among the 46 proteins, 31 (67%) are enzymes, 7 are binding proteins, 7 are transport proteins, and 1 is a ribosomal protein.

In 21 out of the 67 (31%) proteins in our data set, the CBB regions do not contain a catalytic or binding site residue. All the residues in the SITE records of their PDB files fall outside the CBB regions. Out of these 21 proteins, 18 (86%) are enzymes, 2 are binding proteins, and 1 is a ribosomal protein. They are distributed in all fold classes. Absence of the PDB SITE record residues in the CBBs does not necessarily imply nonexistence of a coupling between protein folding and function for these proteins. The CBB regions in these proteins still contain other residues in the active or binding sites and in other functionally important regions such as flexible loops, residues involved in hinge-bending motions, or allostery. However, these have not been investigated in the present study.

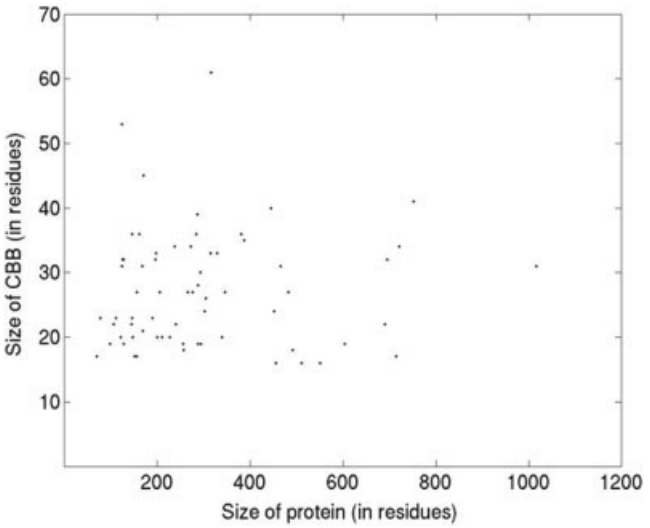

Fig. 2. The size of the critical building blocks versus the protein size for the 67 protein chains in our data set.

Fold Type and Folding–Function Coupling

Does the occurrence of protein folding–function coupling show a preference for certain protein fold type(s)? SCOP[59] uses a protein domain as the unit for structural classification. Among the 46 protein chains that show coincidence between folding and function, 4 (9%) are all-α fold, 12 (26%) are all-β fold, 12 (26%) are α + β, 15 (33%) are α/β, and 3 (6%) are multidomain proteins. Based on our sample size, folding–function coupling is not a distinct characteristic of a specific fold type.

Evolutionary Conservation of Protein Folding and Function Coupling

A coupling between folding and function via a common protein fragment may be evolutionarily advantageous for proteins. This is because proteins need to avoid mutations in a single region to preserve both the native protein fold and its function. To look into the possibility that protein folding and function coupling might be conserved, we first use PROSITE (release 17.18)[53,54,62] on the 67 proteins for which there are both CBB and SITE records. PROSITE is a collection of patterns and profiles that show high sequence conservation in the protein families through evolution. The patterns (and profiles) are derived from multiple alignments of homologous sequences. These patterns frequently contain functionally important residues. PROSITE patterns are available for 37 out of the 67 (55%) protein chains in our data set. PDBsum[17,60] identifies the location of these patterns in the protein structure.

Table I compares the locations of the CBBs and the PROSITE patterns in the 37 proteins. For 16 (out of 37,

TABLE II. Detected Interrelationships between PROSITE pattern, functional SITE Residues; and Critical Building Blocks (CBBs)

	Coincidence between					
Category	CBB SITE	CBB PROSITE	PROSITE SITE	CBB, SITE, PROSITE	No. cases	PDB entries
I	+	+	+	+	14	1cvl, 1cex, 1atlA, 2azaA, 1zia, 1osa, 1lcf, 1doi, 1bp1[a]
II	+	−	−	−	1	1aco
III	+	−	+	−	14	1ac5, 1aa6, 4blmA, 1xikA, 1pbn, 1pii, 1daaA, 1oacB, 1whi, 1utg, 1rcy, 1put, 1kcw, 2mtaC
IV	−	+	+	−	2	1iba, 1fieB
V	−	−	+	−	6	1ad3A, 1pkp, 1esl, 1ast, 1ecfA, 1smnA

+, the respective sites in the header of the column overlap; −, when they do not. The categories are schematically illustrated in Figure 1 and an example for each is given in Figure 3. They are described in detail in the text.
[a]1bp1 and 1ad3A do not show coincidence between SITE and PROSITE residues. Their corresponding categories are defined by the relationships of CBB SITE and CBB PROSITE.

43%) proteins, the PROSITE patterns are located either within or near the CBB regions. We consider a PROSITE pattern to be near a CBB if it lies within seven residues of either of the CBB termini. This is because our building blocks assignment is uncertain by seven residues at each terminus. In the remaining 21 proteins (57% of 37) the CBBs and the PROSITE patterns do not coincide. This observation is rather surprising. CBBs are potentially important elements of the protein structure. Hence, we expect them to be conserved within protein families. However, noncoincidence of the PROSITE patterns does not imply that the CBB regions in these 21 proteins are not conserved.

We present a few interesting examples of colocalization of the PROSITE pattern with the CBBs in proteins. Atrolysin C (PDB entry 1atlA) is a 200 amino acid Zn^{2+}-dependent metallopeptidase that is found in snake venom.[63] Its PROSITE pattern (PROSITE entry PS00142) is involved in Zn^{2+} binding and it is positioned at Thr139-Leu148. The CBB for this protein spans Ser130 to Gly149, including this pattern. Triacylglycerol hydrolase (PDB 1cvl) contains a 30-residue long (Lys80-Thr112) CBB in the middle of the protein. The CBB is made up of an α–β–α unit containing the known nucleophilic elbow, which is identified by the consensus pentapeptide G-X-S-X-G with the nucleophilic serine in the center.[64,65] This pentapeptide and the adjacent residues form the lipase pattern (PROSITE PS00120). In the electron transfer protein 2Fe-2S ferredoxin (PDB 1doi), four cysteine residues bind a single 2Fe-2S iron–sulfur cluster.[66] The PROSITE pattern (PROSITE PS00197) spans this iron–sulfur binding region located at Cys63-Cys71. The CBB (Phe61-Asp79) flanks the pattern region. Additional examples include cutinase, estradiol dioxygenase, lactoferrin, pseudoazurin, azurin, and glutaminase–asparginase. Significant overlaps between the CBBs and the PROSITE patterns were observed in the bacteriocidal protein, calmodulin, and deoxyribonuclease I. For glucose permease, inositol monophosphatase, and endonuclease III, the PROSITE patterns are located within seven residues from the CBB termini. The building block termini, as assigned by the building block cutting program,[29] are also uncertain by seven residues, suggesting possible overlaps there as well. These examples appear to indicate coincidence between regions being conserved through evolution and regions considered critical for folding. A comparison between PROSITE patterns and functional residues in the PDB SITE records (data not shown) indicates that, in 34 out of 37 (92%) protein chains, at least one of the patterns contains a SITE functional residue. The three remaining protein chains (PDB entries 1bp1, 1ad3A, and 1aco) do not show coincidence between the PROSITE pattern and the SITE functional residues.

Locations of PROSITE Patterns, CBBs, and Functional Residues in SITE Records of PDB Files

We have classified the results of the analysis into five categories (Fig. 1). Table II presents a summary of the results and Figure 3 gives selected examples.

Category I Analysis Results

The CBB contains both the PROSITE pattern and the functional residue(s) in the PDB file SITE records [Fig. 3(a)]. This occurs in 14 out of 37 (38%) protein chains. In these protein regions the folding and function coupling is evolutionarily conserved. For example, deoxyribonuclease I (PDB entry 3dni, 258 residues) has five residues forming its active site: Glu39, Gly78, His134, Asp212, and His 252. His134 is located in the middle of both the CBB (Phe128-Asn145) and the PROSITE pattern (Ile130-Val150).

Category II Analysis Results

The CBB contains the functional residue(s) in the PDB file SITE record; however, the PROSITE pattern is outside this region [Fig. 3(b)]. This indicates that folding and function are coupled but may not contain a known evolutionarily conserved pattern. This appears in only one case, mitochondrial aconitase (1aco). Aconitase is a large protein, which has 753 residues with 22 building blocks at the lowest level of cutting. It catalyzes the reversible isomerization of citrate and isocitrate in the Krebs cycle and contains an 4Fe-4S iron–sulfur cluster as a prosthetic group bound by three cysteine residues. However, none of this Cys residues is included in the SITE record of the PDB

file. Rather, the SITE record describes residues participating in protonation events as donors or acceptors.[67] The CBB (Ile146-Ala186) contains active site residues His147, Asp165, Ser166, His167, and Asn170 out of 17 residues identified by the SITE record. Three additional active site residues fall in a building block with the second highest criticalness score (CIndex: see Methods section and Kumar et al.[46]). By contrast, they do not fall within the PROSITE pattern. The pattern for the aconitase family includes two conserved regions that contain the three cysteine ligands of the 4Fe-4S cluster. None fall within the CBB.

Category III Analysis Results

The PROSITE pattern includes functional residues different from those falling in the CBB [Fig. 3(c)]. This may occur when the active site involves spatially close but sequentially distant residues. Here, residues of the same active site can overlap both with the region of the CBB and with the region of the PROSITE pattern. Such cases may indicate that a given part of the protein structure may contain residues important for protein folding and function. It may also be conserved by evolution. This occurred in 14 protein chains. For example, Kexlp serine carboxypeptidase (1ac5, 483 residues) has a catalytic triad formed by serine-aspartic acid-histidine.[68] Out of 16 building blocks at the lowest level of cutting, these three residues fall in three separate regions, one in the CBB and the other two in the PROSITE pattern.

Category IV Analysis Results

The CBB and the functional residue(s) noted in the PDB SITE records are in adjacent sequence positions [Fig. 3(d)] and the PROSITE pattern overlaps both the CBB and the SITE residues. Two protein chains fall into this category. One is transglutaminase coagulation factor XIII (1fieB). This large protein (715 residues, 24 building blocks) has three catalytic residues: Cys314, His373, and Asp396. None fall in the CBB (Ala332-Leu348). Instead, the neighboring PROSITE pattern Gly312-Gly329 (PROSITE entry PS00473) contains the catalytic cysteine. Here, the N-terminus of the CBB borders the pattern within the 7-residue threshold.[29]

Category V Analysis Results

The CBB contains neither the functional residue(s) in the PDB SITE records nor the PROSITE pattern [Fig. 3(e)]. Instead, the PROSITE pattern overlaps some functional residues. Six protein chains fall into this category.

Control Study

It is difficult to rigorously assess the statistical significance of these trends because of the empirical nature of the computational cutting, the criticalness score, and the lack of accuracy of the SITE and PROSITE records in the PDB. Nevertheless, in an attempt to obtain some measure of an assessment, we took building blocks with the lowest criticalness scores (CIndex score) in the 67 protein chains at the lowest level of the protein Anatomy algorithm cutting. These building blocks are least important for protein folding because they do not mediate the interactions among other building blocks and they lie mostly outside the protein core. However, low CIndex scores do not mean lower stability or size for these building blocks. For example, the average size of these 21 least important building blocks, foldingwise, is similar to that of the 46 CBBs (average length for 46 CBBs = 27 $\pm$ 10 residues; average length for the 21 least important building blocks = 29 $\pm$ 12 residues). We checked how many of these contain functional SITE residues or match with the PROSITE pattern. For the 67 protein chains, 21 building blocks with the lowest CIndex scores were found to contain functional residues in comparison to 46 CBBs. We detected 7 cases of coincidence between building blocks with the lowest criticalness score and the PROSITE pattern out of the total 37 with available PROSITE data (compared to 16 cases for the CBBs). Our results suggest that CBBs are roughly twice as likely to contain either a functional residue or a PROSITE pattern when compared with a building block with the lowest CIndex value.

Protein Families

A second way to test a potential protein folding and function coupling is to carry out comprehensive structure and sequence database analyses on the conservation of the entire protein versus the CBB regions. If a protein building block is critical for folding, it should be highly conserved within the family. Changes in its conformation may lead to either unstable or nonnative global folds. Hence, we may expect lower mutation rates for CBBs. We have therefore picked protein families containing homologous structures and asked how well the CBBs are conserved structure- and sequencewise within members of the family.

This analysis was carried out for three families: C-type lysozymes, bacterial lipase, and fungal lipase. The multiple sequence alignments of the lysozyme sequences show two relatively conserved regions. The conserved regions are around catalytic residues Glu35 and Asp52. The first region falls within the N-terminal CBB region (residues 2–39, CIndex = 1σ). Bacterial and fungal lipase structures share the same α/β hydrolase fold and their corresponding CBBs were identified with a high level of confidence (CIndex = 2σ). In both lipase families, the CBB consists of a common $\beta/\alpha/\beta$ unit containing the nucleophilic serine. Stronger conservation of the CBB region was observed here in both the sequence and structure as suggested by multiple sequence and structural alignments (data not shown). Our results suggest that CBBs tend to be conserved through evolution more than other building blocks.

Interesting Cases of Similar CBBs in Structurally Unrelated Proteins with Similar and Different Functions

By their definition, building blocks are stand-alone elements of the protein structure. We previously consistently observed that proteins with dissimilar global structures and dissimilar sequences can share building blocks with similar conformations.[31] Here, we compared the

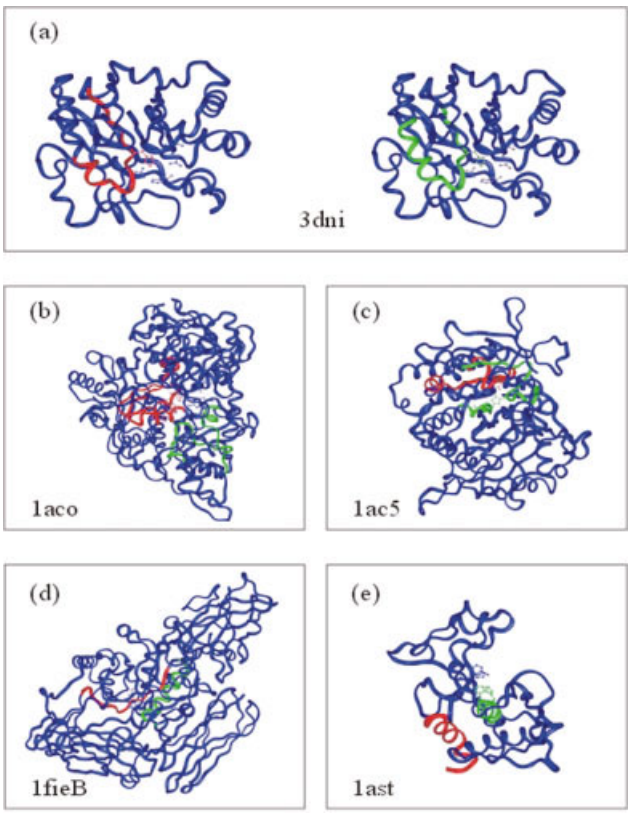

Fig. 3. Examples of the various categories of Figure 1, relating to whether the critical building block, PROSITE pattern, and functional residues in the SITE record overlap. The structures are displayed in ribbon representation: CBBs are red, PROSITE patterns are green, and active site residues are shown as ball and sticks. (**a**) Category I: the correspondence between the critical building block region, PROSITE pattern, and functional residues in deoxyribonuclease I (PDB entry 3dni). The critical building block, Phe128-Asn145 (left side), overlaps the PROSITE pattern Ile130-Val150 (right side). His134 is a well-conserved catalytic residue. (**b**) Category II: the correspondence between the critical building block region and the functional residues, although not with the PROSITE pattern as seen in mitochondrial aconitase (PDB entry 1aco). (**c**) Category III: the correspondence between the critical building block region and functional residues, where different functional residues correspond with PROSITE patterns. It is shown here for Kexlp serine carboxypeptidase (PDB entry 1ac5). (**d**) Category IV: no correspondence between the critical building block and the functional residues. The PROSITE pattern is located near the critical building block and contains the functional residues, as shown for transglutaminase coagulation factor XIII (PDB entry 1fieB). (**e**) Category VP: no correspondence between the critical building block and the functional residues; however, the functional residues are within the PROSITE pattern, which is shown here for astacin (PDB entry 1ast). The functional residues (His92, His96, His102, and Tyr149) do not fall within the critical building block His181-Leu200. However, two histidine residues fall in the PROSITE pattern Thr89-Ile98.

CBBs in our library, clustering them structurally. Most of the clusters that we obtained consisted of CBBs derived from globally similar proteins. However, some clusters were particularly interesting. These consisted of CBBs derived from proteins with dissimilar structures and different functions. Interesting cases are those of cutinase, bacterial lipase, fungal lipase, and thioesterase (PDB entries 1cex, 1cvl, 1tca, and 1thtA, respectively). The CBBs of these proteins are structurally similar, being characterized by a common β/α motif (Fig. 4). Three

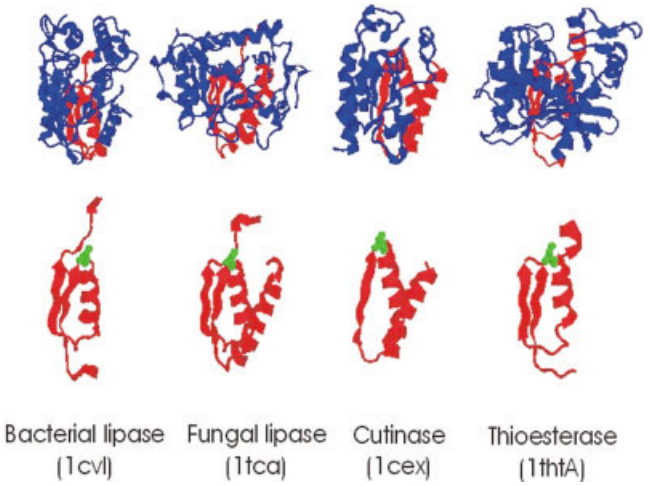

Fig. 4. Structurally similar critical building blocks (red) derived from globally unrelated, dissimilar protein structures: bacterial lipase, fungal lipase cutinase, and thioesterase (PDB entries 1cvl, 1tca, 1cex, and 1thtA, respectively). The critical building blocks are characterized by common β/α subunits.

(bacterial lipase, fungal lipase, and thioesterase) are classified as α/β hydrolase folds whereas cutinase is a flavodoxin-like fold. All enzymes belong to the functional hydrolase family. Both lipase and cutinase are carboxylic ester hydrolases (EC.3.1.1.-) whereas thioesterase is a thioester hydrolase (EC.3.1.2.-). The catalytic machinery of all enzymes consists of the triad serine-aspartic acid-histidine.[65,69–71] The essential catalytic serine residue is located at a distinct 3D location, called the "nucleophilic elbow," that is positioned between an α helix and a β strand in all CBBs. The CBBs are derived from structures having similar folds, although they are distant. Superposition of the complete chains of these proteins could detect only a small contiguous fragment shared by all structures. The common fragment consists of an α/β/α unit found to correspond to the CBB region. Such a situation suggests that the CBBs can be exchanged between otherwise globally dissimilar structures, altering protein function.

CBBs May Be Unstable

We have estimated the relative conformational stability of building blocks using a building block stability score based on the hydrophobicity, compactness, and isolatedness of the individual building blocks (see Methods section). Although this empirical score does not accurately assess the thermodynamic stability of the building blocks, it accounts for parameters known to make dominant contributions to the protein stability, that is, the hydrophobic effect and the compactness.

We have already shown[31] that for all building blocks from the entire nonredundant data set derived from the PDB, the higher the criticalness score (see Methods section and Kumar et al.[46]) the lower the empirical stability score.[29] Figure 5 plots the stability scores versus the criticalness scores for all building blocks derived from all the proteins in our CBB library at the lowest cutting level from which the CBBs were taken. A similar trend is

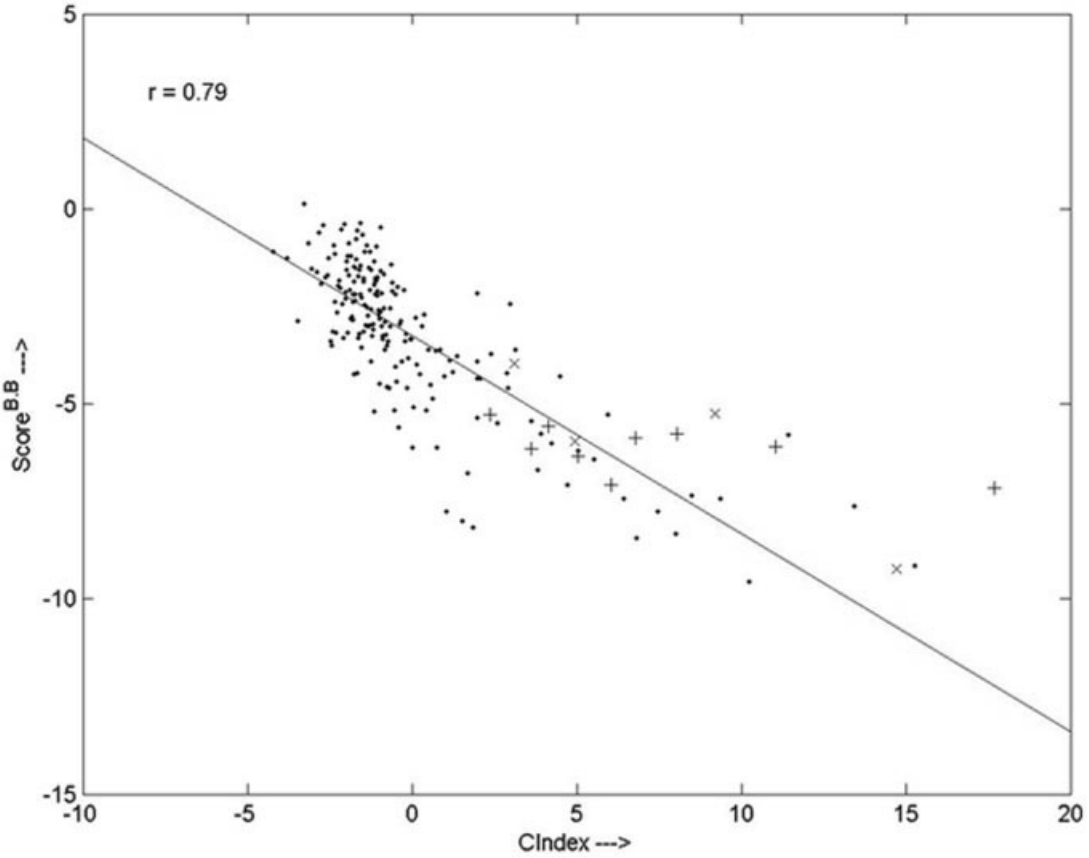

Fig. 5. The criticalness score (CIndex) versus the empirical building block stability score ($Score^{B.B}$) for 1880 building blocks belonging to all the protein chains in the CBB library. (· · ·) The averages of all building blocks; (+) the averages of the 46 CBBs, showing a folding–function correspondence; and (×) the 21 CBBs that do not. The 46 and 21 CBBs belong to the 67 protein chains that constitute the data for this study. The values for the 1880, 46, and 21 building blocks are averaged over windows of 10, 5, and 5, respectively. The regression line in the figure is obtained by least squares fitting of the 1880 building blocks. The correlation coefficient (r) is −0.79. The scoring functions are described in the Methods section. Additional details are given by Tsai et al.[29] and Kumar et al.[46]

observed here: the more critical the building blocks the lower the empirical stability scores. Thus, the 67 CBBs are expected to be conformationally unstable. No difference is observed between the 46 CBBs that contain a functional residue (from the SITE records in the PDB) and the 21 that do not. Figure 2 illustrates that, regardless of the protein size, the CBBs are small, being usually less than 40 residues. With extensive interfaces with their neighboring building blocks in the protein interior, they largely adopt extended shapes, lacking a sufficiently large hydrophobic core to keep them stable.[72]

DISCUSSION

Study Limitations

There are several limitations to our present study. The building blocks cutting program is based on empirical parameters derived from statistical analyses of protein structures available from the PDB. Although it is encouraging that building blocks detected computationally agree fairly well with protein fragments obtained from limited proteolysis experiments,[56] there is no certainty that they are the "real" building blocks of the proteins. Moreover, our assignment of building blocks termini is uncertain by about seven residues on each side. The assignment of CBBs is also empirical because it is based on calculations of the polar and nonpolar surfaces buried among the building blocks and on a weighting scheme for the relative importance of the building blocks of the protein structure. Although the relative weights in this scheme were tested extensively and tuned on a large number of families with many members and at different building block cutting levels, nevertheless, despite the observed consistency, they are still arbitrarily parameterized. In the case of adenylate kinase we have extensively studied the N-terminal building block. This building block emerged as critical using molecular dynamics simulation.[46] However, the CBBs library with 225 CBBs is too large for such an individual testing. The SITE records in the PDB files also have limitations, and they may not be consistent with automated detection of the ligand binding residues.

To roughly estimate the probability (P) of finding functionally important residues in CBBs by random chance, we use the average number of residues in the SITE records (NS), the average number of residues in the CBBs (NCBB), the average number of building blocks in the proteins

(NBB), and the average protein size (NP) in the following formula:

$$P = (\mathrm{NS} \times \mathrm{NCBB})/(\mathrm{NBB} \times \mathrm{NP} \times \mathrm{NP}).$$

In our data set of 46 protein chains that contain functional residues in their CBBs, NS = 7, NBB = 5, NCBB = 27, and NP = 302. To account for the uncertainty in building block termini, we add 14 (7 residues for each BB terminus) to the NCBB to make it 41. Hence, the probability of finding a functional residue in a CBB by random chance is 6.29×10^{-4}.

Study Implications

One of the paradigms in biology states that proteins need to fold into their native (functional) states in order to function. Currently, understanding and ultimately predicting protein function is perhaps the most challenging problem in protein science. Consequently, an increasing number of studies have addressed the relationship between structure and function (e.g., see refs. 1, 2, and 5). Here, we probe the relationship between the protein folding process and protein function, complementing studies of structure and function. Folding and structure are related: the conformation (topology) of the final folded state dictates the folding pathway. Proteins with similar topologies fold via similar pathways. For small two-state proteins (or domains) topology is a dominant factor in folding kinetics.[73–81] Proteins with similar topologies have similar protein anatomies[30] and their building block cutting patterns are similar, regardless of the details of the sequence or the structural stability. At the same time, proteins with similar topologies have hinges at similar locations and similar patterns of slow and fast motions,[24] even though their functions may differ.

The fact that the number of fold types is limited and that some topologies are observed repeatedly for different functions (as in the TIM barrels[5]) suggests that function took advantage of "good" folds. On the other hand, function can lead to more complex protein folds and to unfavorable local interactions. Here, we focus on folding and function. Several observations have been made that bear on the question of a potential relationship between the protein folding process and protein function.

First, the pro-region at the N-termini of some extracellular proteases has been shown to be critical for correct folding of the protein. It serves both as an intramolecular chaperone and as an inhibitor. Self-cleavage of the fragment by the folded protein leads to an active protease. An a priori absence of the pro-region yields an altered nonfunctional protein conformation. Its noncovalent subsequent addition leads to the functional, native conformation. Hence, the pro-region catalyzes protein folding. Its self-cleavage places the protease under kinetic (rather than thermodynamic) control, which is functionally advantageous in an extracellular protease-infested environment.[82] The dual role of the pro-region in folding and function suggests cooperativity.

Second, it has been estimated that a relatively large fraction of proteins are in a natively disordered state. In this state, the native conformation of the protein is unstable, with a low population time. We have recently suggested that natively disordered proteins have large interfaces with their bound sister molecules and have a large interface to size (in terms of the number of residues in the chain) ratio. To have functionally dictated large interfaces and be stable, the protein size needs to be large to accommodate a significant hydrophobic core. Yet, larger proteins are disadvantageous. Larger proteins aggravate the problems that cells already fight to overcome, such as cellular crowding. They further increase genomic sizes and lead to a larger expenditure of energy.[72] From the functional standpoint, even a low population time presents no problem. The favorable conformer will bind, and equilibrium will be shifted in its direction.

Third, as Figure 5 indicates, CBBs may be conformationally less stable than other building blocks. A similar observation has been made for the pro-region of subtilisin.[83] Here, too, the instability derives from large surface areas and small hydrophobic cores, largely the outcome of the small CBB sizes and large interfaces. Smaller CBB sizes lead to smaller protein sizes. Furthermore, binding sites are characterized by regions of instability.[84] Unstable binding sites are advantageous for catalysis[85,86] and for binding a range of ligands.[87] Hence, the dual function–folding roles for CBBs suggest cooperativity. A further indication of preferred lower stability at binding sites comes from an analysis of citrate synthase. There, overall, the number of salt bridges and their strengths are similar in thermophiles versus psychrophiles. However, whereas in the thermophile a large number of salt bridges cluster in the enzyme active site, in the psychrophile they are practically absent in the active site region.[88]

Fourth, it has been suggested that faster folding rates are advantageous to a protein, reducing the chances of misfolding. In contrast, functional constraints may necessitate more complex protein folds with native interactions that are distant on the polypeptide chain, leading to longer folding time scales and frustration.

Fifth, natively folded proteins are marginally stable at the source organism's living temperature. Marginal stability is dictated by function. Rigid proteins are nonfunctional. In a recent analysis of experimental data, we found that the thermodynamic stability of a protein at the organism's optimum living temperature is invariably less than the maximal protein stability.[89]

Function can cooperate with folding or frustrate it. At the same time, folding may also frustrate function. In vitro protein evolution may select for enzymes with faster catalytic rates. Yet, these enzymes may be less stable. In the end, what we observe is an optimal viable interplay between folding and function and between structure and function. Folding, structure, and function are interrelated. They can enhance or frustrate each other. The majority of the studies to date have focused on the relationship between structure and function. Previous work on folding–function involved studies of energetic frustration and ϕ value analysis.[90,91] In this work we question whether folding and function can be coupled for some proteins.

There have been indications that such a coupling can take place.[24,92] Our recent studies on adenylate kinase and dihydrofolate reductase[46,47] have further suggested such a potential coupling. In these cases important folding elements are essential for function. Using building blocks that are critical for folding, we ask if they are also functionally important. Our analysis indicates that, although such coupling is not observed in all proteins, in over two-thirds of the proteins the residues that are important functionally (largely catalytic residues) fall in the CBBs. We did not probe a broader range of functional residues, such as those participating in binding or in domain/loop flexibility. These are likely to increase the percentage of folding–function coupling.

Mirny and Shakhnovich[7] studied evolutionary conservation in nine protein families containing homologous sequences. The residues involved in the folding nuclei of these proteins were identified using protein engineering techniques. We recently observed that the PDB SITE records contain information on functional residues in five of these proteins. Four out of the five cases contain functional residues that are either identical or lie adjacent in sequence to the residues in the folding nuclei. In one case, acyl-coenzyme A binding protein (ACBP),[93] Phe5, Ala9, and Tyr73 are part of both the folding nucleus and acyl-coenzyme A binding site. These residues are also important for the structural stability of ACBP (N. A. Petrova, S. Kumar, unpublished data). These observations provide further indications for coupling between protein folding and function.

CONCLUSIONS

In this study we probed a potential relationship between folding and function. We started with building block fragments predicted to be critical for folding. Next, we asked whether functionally important residues appear to fall within them. We found that for over two-thirds of the CBBs for which functional information is available the functionally important residues fall in these fragments. Because we used the PDB SITE record to obtain the functional residues, in most proteins (74%) these are the catalytic residues. However, these numbers are approximate, bearing in mind the computational algorithms, empirical parameters, and relative inaccuracy of the PDB SITE records.

Although we mapped functionally important (mostly catalytic) isolated residues, function actually requires a region. Because a building block that is critical for folding is in contact with many other fragments in the protein, it has higher chances of fulfilling a functional role as well. From the evolutionary standpoint such a folding–function relationship appears logical. Given the critical importance of a fragment for folding, there is evolutionary pressure to preserve its sequence. Slight changes may lead to a misfolded protein. Function took advantage of the evolutionary pressure for conservation. To explore the conservation of this region, we performed a detailed analysis of the protein families consisting of the sequences and structures of lysozyme and of bacterial and fungal lipases, in addition to the previous analyses of dihydrofolate reductase and adenylate kinase. This observed conservation may assist in the prediction of protein function of an unknown sequence.

Interestingly, clustering the CBBs by their structures reveals that CBBs with similar conformations may occur in globally dissimilar proteins with different functions. This is reasonable, because their sequences are different and thus may have different catalytic residues. This situation resembles the well-known observation for global protein folds: proteins with similar structures may fulfill different functions. We further note that all CBBs in our library are conformationally unstable.

ACKNOWLEDGMENTS

We thank Dr. Chung-Jung Tsai and Natalia V. Petrova for many useful discussions and comments, Nurit Haspel and John Owens for their help with manuscript preparation, and the members of the Nussinov–Wolfson Structural Bioinformatics group. The research of the second (R.N.) and fourth (H.W.) authors in Israel was partially supported by a Ministry of Science grant and by the Center of Excellence in Geometric Computing and Its Applications funded by the Israel Science Foundation (administered by the Israel Academy of Sciences). The research of the fourth author (H.W.) is partially supported by the Hermann Minkowski–Minerva Center for Geometry at Tel Aviv University. This project was funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health. 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 organizations imply endorsement by the U.S. Government.

REFERENCES

1. Moult J, Melamud E. From fold to function. Curr Opin Struct Biol 2000;10:384–389.
2. Martin ACR, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JBO, Taroni C, Thornton JM. Protein folds and functions. Structure 1998;6:875–884.
3. Hegyi H, Gerstein M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol 1999;288:147–164.
4. Finkelstein AV, Gutun AM, Badretdinov AYa. Why are the same protein folds used to perform different functions? FEBS Lett 1993;325:23–28.
5. Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 2002;321:741–765.
6. Dokholyan NV, Shakhnovich EI. Understanding hierarchical protein evolution from first principles. J Mol Biol 2001;312:289–307.
7. Mirny LA, Shakhnovich I. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999;291:177–196.
8. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
9. Aloy P, Querol E, Aviles FX, Sternberg MJ. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 2001;311:395–408.

10. Kasuya A, Thornton JM. Three-dimensional structure analysis of PROSITE patterns. J Mol Biol 1999;286:1673–1691.
11. Russell RB. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 1998;279:1211–1127.
12. Russell RB, Sasieni PD, Sternberg MJ. Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol 1998;282:903–918.
13. Jackson RM, Russell RB. The serine protease inhibitor canonical loop conformation: examples found in extracellular hydrolases, toxins, cytokines and viral proteins. J Mol Biol 2000;296:325–334.
14. Jackson RM, Russell RB. Predicting function from structure: examples of the serine protease inhibitor canonical loop conformation found in extracellular proteins. J Comput Chem 2001;26:31–39.
15. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997;6:2308–2323.
16. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. From structure to function: approaches and limitations. Nature Struct Biol 2000;7:991–994.
17. Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM. PDBsum: a Web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci 1997;22: 488–490.
18. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 1998;281:949–968.
19. Zhang B, Rychlewski L, Pawlowski K, Fetrow JS, Skolnick J, Godzik A. From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions. Protein Sci 1999;8:1104–1115.
20. Orengo CA, Todd AE, Thornton JM. From protein structure to function. Curr Opin Struct Biol 1999;9:374–382.
21. Shapiro L, Harris T. Finding function through structural genomics. Curr Opin Biotechnol 2000;11:31–35.
22. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. Nature Biotechnol 2000;18: 283–287.
23. Gruebele M. Protein folding: the free energy surface. Curr Opin Struct Biol 2002;12:161–168.
24. Sinha N, Tsai CJ, Nussinov, R. Building blocks, hinge-bending motions and protein topology. J Biomol Struct Dyn 2001;19:369–380.
25. Lesk AM, Rose GD. Folding units in globular proteins. Proc Natl Acad Sci USA 1981;78:4304–4308.
26. Rose GD. Hierarchic organization of domains in proteins. J Mol Biol 1979;134:447–470.
27. Baldwin RL, Rose GD. Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biochem Sci 1999;24:26–33.
28. Baldwin RL, Rose GD. Is protein folding hierarchic? II. Folding intermediates and transition states. Trends Biochem Sci 1999;24: 77–83.
29. Tsai CJ, Maizel J, Nussinov R. Anatomy of protein structure: visualizing how a one-dimensional protein chain folds into a three-dimensional shape. Proc Natl Acad Sci USA 2000;97:12038–12043.
30. Tsai CJ, Ma B, Kumar S, Wolfson H, Nussinov R. Protein folding: binding of conformationally fluctuating building blocks via population selection. Crit Rev Biochem Mol Biol 2001;36:399–433.
31. Haspel N, Tsai CJ, Wolfson H, Nussinov R. Hierarchical protein folding pathways: a computational study of protein fragments. Proteins 2003;51:203–215.
32. Haspel N, Tsai CJ, Wolfson H, Nussinov R. Reducing the computational complexity of protein folding via fragment folding and assembly. Protein Sci 2003;12:1177–1187.
33. Eliezer D, Wright PE. Is apomyoglobin a molten globule? Structural characterization by NMR. J Mol Biol 1996;263:531–538.
34. Eliezer D, Yao J, Dyson HJ, Wright PE. Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding. Nature Struct Biol 1998;5:148–155.
35. Fontana A, Polverino de Laureto P, De Filippis V, Scaramell E, Zambonin M. Probing the partly folded states of proteins by limited proteolysis. Fold Des 1997;2:R17–R26.
36. Peng Z-Y, Wu LC. Autonomous protein folding units. Adv Protein Chem 2000;53:1–47.
37. Dobrodumov A, Gronenborn AM. Filtering and selection of structural models: combining docking and NMR. Proteins 2003;53:18–32.
38. Jiang S, Tovchigrechko A, Vakser IA. The role of geometric complementarity in secondary structure packing: a systematic docking study. Protein Sci 2003;12:1646–1651.
39. Chavez LG Jr, Scheraga HA. Intrinsic stabilities of portions of the ribonuclease molecule. Biochemistry 1980;19:1005–1012.
40. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. J Mol Biol 2002;323:297–307.
41. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 1989;5:355–373.
42. Prestrelski SJ, Williams AL Jr, Liebman MN. Generation of a substructure library for the description and classification of protein secondary structure. I. Overview of the methods and results. Proteins 1992;14:430–439.
43. Zehfus MH. Continuous compact protein domains. Proteins 1987; 2:90–110.
44. Michel SLJ, Berg JM. Building a metal binding domain, one half at a time. Chem Biol 2002;9:667–668.
45. Tsai CJ, Nussinov R. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. Protein Sci 1997;6:24–42.
46. Kumar S, Sham YY, Tsai CJ, Nussinov R. Protein folding and function: the N-terminal fragment in adenylate kinase. Biophys J 2001;80:2439–2454.
47. Ma B, Tsai CJ, Nussinov R. Binding and folding: in search of intramolecular chaperone-like building block fragments. Protein Eng 2001;9:617–627.
48. Ma B, Kumar S, Tsai CJ, Nussinov R. Folding funnels and binding mechanisms. Protein Eng 1999;12:713–720.
49. Tsai CJ, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels and protein function. Protein Sci 1999;8:1181–1190.
50. Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R. Folding and binding cascades: dynamic landscapes and population shifts. Protein Sci 2000;9:10–19.
51. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucl Acids Res 2000;28:235–242.
52. Kumar S, Barzilai A, Haspel N, Sham YY, Tsai CJ, Wolfson H, Nussinov R. Critical building blocks in proteins: a common theme for folding and function. In: Recent research developments in protein folding, stability and design. Gromiha MM, Selvaraj S, editors. Trivandrum, India: Research Signpost; 2002. p 207–217.
53. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A. The PROSITE database, its status in 2002. Nucleic Acids Res 2002;30:235–238.
54. Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 2002;3:265–274.
55. Tsai CJ, Maizel J, Nussinov R. Distinguising between sequential and nonsequentially folded proteins: implication for folding and misfolding. Protein Sci 1999;8:1591–1604.
56. Tsai CJ, Polverino de Laureto P, Fontana A, Nussinov R. Comparison of protein fragments identified by limited proteolysis and by computational cutting of proteins. Protein Sci 2002;11:1753–1770.
57. Chen L, Durley R, Mathews FS, Davidson VL. Structure of an electron transfer complex: methylamine dehydrogenase, amicyanin and cytochrome c551i. Science 1994;264:86–89.
58. Gallagher T, Rozwarski DA, Ernst SR, Hackert ML. Refined structure of the pyruvoyl-dependent histidine decarboxylase from lactobacillus 30a. J Mol Biol 1993;230:516–528.
59. 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.
60. Laskowski RA. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res 2001;29:221–222.
61. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. Protein Eng 1995;8:127–134.
62. Bucher P, Bairoch A. A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation. In: Altman R, Brutlag D, Karp P, Lathrop R, Searls D,

editors. Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology. Menlo Park, CA: AAAI Press; 1994. p 53–61.
63. Zhang D, Botos I, Gomis-Ruth FX, Doll R, Blood C, Njoroge FG, Fox JW, Bode W, Meyer EF. Structural interaction of natural and synthetic inhibitors with the venom metalloproteinase, atrolysin C (form d). Proc Natl Acad Sci USA 1994;91:8447–8451.
64. Derewenda ZS, Derewenda U. Relationships among serine hydrolases: evidence for a common structural motif in triacylglyceride lipases and esterases. Biochem Cell Biol 1991;69:842–851.
65. Lang DA, Hofmann B, Haalck L, Hecht HJ, Spener F, Schmid RD, Schomburg D. Crystal structure of a bacterial lipase from *Chromobacterium viscosum* ATCC 6918 refined at 1.5 Å resolution. J Mol Biol 1996;295:704–717.
66. Frolow F, Harel M, Sussman JL, Mevarech M, Shoham M. Insights into protein adaptation to a saturated salt environment from the crystal structure of a halophilic 2Fe-2S ferredoxin. Nature Struct Biol 1996;3:452–458.
67. Lauble H, Kennedy MC, Beinert H, Stout CD. Crystal structures of aconitase with *trans*-aconitate and nitrocitrate bound. J Mol Biol 1994;237:437–451.
68. Shilton BH, Li Y, Tessier D, Thomas DY, Cygler M. Crystallization of a soluble form of the Kex1p serine carboxypeptidase from *Saccharomyces cerevisiae*. Protein Sci 1996;5:395–397.
69. Uppenberg J, Hansen MT, Patkar S, Jones TA. The sequence, crystal structures determination and refinement of two crystal forms of lipase B from *Candida antarctica*. Structure 1994;2:293–308.
70. Lawson DM, Derewenda U, Serre L, Ferri S, Szittner R, Wei Y, Meighen EA, Derewenda ZS. Structure of a myristoyl-ACP-specific thioesterase from *Vibrio harveyi*. Biophys J 1994;33:9382–9388.
71. Longhi S, Czjzek M, Lamzin V, Nicolas A, Cambillau C. Atomic resolution (1.0 Å) crystal structure of *Fusarium solani* cutinase: stereochemical analysis. J Mol Biol 1997;268:779–799.
72. Gunasekaran K, Tsai CJ, Kumar S, Zanuy D, Nussinov R. Extended disordered proteins: targeting function, with less scaffold. Trends Biochem Sci 2003;28:81–85.
73. Alm E, Baker D. Matching theory and experiment in protein folding. Curr Opin Struct Biol 1999;9:189–196.
74. Alm E, Baker D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. Proc Natl Acad Sci USA 1999;96:11305–11310.
75. Munioz V, Eaton W. A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc Natl Acad Sci USA 1999;96:1131–1136.
76. Galzitskaya OV, Finkelstein AV. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. Proc Natl Acad Sci USA 1999;96:11299–11304.
77. Kim DE, Gu H, Baker D. The sequences of small proteins are not extensively optimized for rapid folding by natural selection. Proc Natl Acad Sci USA 1998;95:4982–4986.
78. Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi O, Baker D. Functional rapidly folding proteins from simplified amino acid sequences. Nature Struct Biol 1997;4:805–809.
79. Perl D, Welker C, Schindler T, Schroder K, Marahiel MA, Jaenicke R, Schmid FX. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. Nature Struct Biol 1998;5:229–235.
80. Martinez JC, Pisabarro MT, Serrano L. Obligatory steps in protein folding and the conformational diversity of the transition state. Nature Struct Biol 1998;5:721–729.
81. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
82. Sauter NK, Mau T, Rader SD, Agard DA. Structure of α-lytic protease complexed with its proregion. Nature Struct Biol 1998;5:945–950.
83. Wang L, Ruan B, Ruvinov S, Bryan PN. Engineering the independent folding of subtilisin BPN′ prodomain: correlation of prodomain stability with the rate of subtilisin folding. Biochemistry 1998;37:3165–3171.
84. Luque I, Freire E. Structural stability of binding sites: consequences for binding affinity and allosteric effects. Proteins 2000;4:63–71.
85. Piana S, Parrinello M, Carloni P. Role of conformational fluctuations in the enzymatic reaction of HIV-1 protease. J Mol Biol 2002;319:567–583.
86. Piana S, Carloni P, Rothlisberger U. Drug resistance in HIV-1 protease: flexibility-assisted mechanism of compensatory mutations. Protein Sci 2002;11:2393–2402.
87. Ma B, Shatsky M, Wolfson H, Nussinov R. Multiple ligands binding at a single site: a matter of pre-existing conformations. Protein Sci 2002;11:184–197.
88. Kumar S, Nussinov R. Different roles of electrostatics in heat and in cold adaptation by citrate synthase. ChemBioChem 2004;V:280–290.
89. Kumar S, Tsai CJ, Nussinov R. Thermodynamic differences among homologous thermophilic and mesophilic proteins. Biochemistry 2001;40:14152–14165.
90. Plotkin SS, Onuchic JN. Investigation of routes and funnels in protein folding by free energy functional methods. Proc Natl Acad Sci USA 2001;97:6509–6514.
91. Jager M, Nguyen H, Crane J, Kelly J, Grubele, M. The folding mechanisms of a β-sheet: the WW domain. J Mol Biol 2001;311:373–393.
92. Cunningham EL, Jaswal SJ, Sohl JL, Agard DA. Kinetic stability as a mechanism for protease longevity. Proc Natl Acad Sci USA 1999;96:1108–11014.
93. Kragelund BB, Knudsen J, Poulsen FM. Acyl-coenzyme A binding protein (ACBP). Biochem Biophys Acta 1999;1441:150–161.