A thermodynamic definition of protein domains

Lauren L. Porter[1] and George D. Rose[2]

T.C. Jenkins Department of Biophysics, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218

Edited by* S. Walter Englander, University of Pennsylvania, Philadelphia, PA, and approved April 10, 2012 (received for review February 13, 2012)

Protein domains are conspicuous structural units in globular proteins, and their identification has been a topic of intense biochemical interest dating back to the earliest crystal structures. Numerous disparate domain identification algorithms have been proposed, all involving some combination of visual intuition and/or structure-based decomposition. Instead, we present a rigorous, thermodynamically-based approach that redefines domains as cooperative chain segments. In greater detail, most small proteins fold with high cooperativity, meaning that the equilibrium population is dominated by completely folded and completely unfolded molecules, with a negligible subpopulation of partially folded intermediates. Here, we redefine structural domains in thermodynamic terms as cooperative folding units, based on *m*-values, which measure the cooperativity of a protein or its substructures. In our analysis, a domain is equated to a contiguous segment of the folded protein whose *m*-value is largely unaffected when that segment is excised from its parent structure. Defined in this way, a domain is a self-contained cooperative unit; i.e., its cooperativity depends primarily upon intrasegment interactions, not intersegment interactions. Implementing this concept computationally, the domains in a large representative set of proteins were identified; all exhibit consistency with experimental findings. Specifically, our domain divisions correspond to the experimentally determined equilibrium folding intermediates in a set of nine proteins. The approach was also proofed against a representative set of 71 additional proteins, again with confirmatory results. Our reframed interpretation of a protein domain transforms an indeterminate structural phenomenon into a quantifiable molecular property grounded in solution thermodynamics.

protein folding | protein structure | protein architecture | protein parsing

Domains are visually arresting protein substructures with an influential history in protein biochemistry (1). These familiar, self-contained structural units were first noticed in some of the earliest solved protein structures (2, 3) and soon came to be recognized as common features of protein architecture (4).

Dissecting proteins into their constituent domains provides a simple, intuitive approach to classifying protein structure, a molecular application of the time-honored principle of "carving nature at its joints" (5). Many structure-based computer algorithms have been devised to parse the ever-increasing number of solved proteins into discrete units; a highly abbreviated sample includes (6–13). Today, CATH (14) and SCOP (15) are the two most widely used domain classifications. Both are based on computational algorithms but rely ultimately on the human eye as the final arbiter of domain boundaries.

However, seeing can be deceiving. The dependence on visual intuition introduces an unavoidable element of ambiguity into procedures for domain recognition. The most enduring domain definition, "potentially independent, stable folding units" (16), conveys the fundamental concept qualitatively. Yet, visual inspection is insufficient to determine whether a protein substructure folds independently, and stability depends on folding conditions such as temperature, pressure, and solvent quality. This inherent ambiguity is reflected in conflicting domain classifications for the same protein. For example, CATH classifies human proliferating cell nuclear antigen (hPCNA) (1u7bA) as a single-domain protein, but both SCOP and those who solved its structure identify two domains (17). Again, early visual assessment found that triosephosphate isomerase "cannot be plausibly subdivided [into domains]" (16). Later experimental evidence, however, suggests that TIM barrels comprise two or even three domains (18, 19). Clearly, a rigorous, quantitative domain definition is needed.

Here, we propose a thermodynamic framework for the structural phenomenology that surrounds current domain definitions. Specifically, we redefine structural domains in thermodynamic terms as cooperative folding units, based on aqueous buffer → urea *m*-values (20, 21). These *m*-values measure the cooperativity of a protein or its substructures and also the potency of urea in denaturing them. In our analysis, a domain is equated to a contiguous segment of the folded protein whose *m*-value is largely unaffected when that segment is excised from its parent structure. In other words, domain cooperativity, as assessed by its *m*-value, depends primarily upon intrasegment interactions, not intersegment interactions.

Structure-Energy Relationship

For two-state proteins, the equilibrium constant of the folding reaction, D(enatured) ⇌ N(ative), is $K = \frac{N}{D}$, and the free energy difference between the native and unfolded states in denaturant-free buffer is given by $\Delta G^0_{[N]\rightarrow[D]} = -RT\ln K$ (R is the gas constant and T is the absolute temperature). $\Delta G^0_{[N]\rightarrow[D]}$ can be obtained experimentally using the linear extrapolation model (20), which plots the ratio of folded to unfolded species as a linear function of denaturant concentration, C. The slope of this line is the *m*-value, $m = RT\frac{\partial \ln K}{\partial C}$, with units of kcal/mol/M. The *m*-value quantifies the extent to which a given concentration of denaturant destabilizes the native structure, as measured at points in the transition zone between folding and unfolding (21). The *m*-value is akin to an intensive variable and overall stability depends upon an associated extensive variable, in this case the chain length. In general, each additional backbone unit adds another stabilizing increment under a given set of folding conditions (22, 23).

Qualifying Ratio

We propose a thermodynamically-based definition in which a domain is a segment of the folded protein for which the *m*-value in isolation and in the context of the protein is essentially the same. This definition is quantified by the qualifying ratio (QR), defined as $\frac{m_{\text{excised}}}{m_{\text{in\_situ}}}$, where m_{excised} and $m_{\text{in\_situ}}$ are the respective water → urea *m*-values of the segment in isolation and in the context of its parent structure. Segments with QRs near unity are likely domains; those with QRs less than unity are not. Implementation of this definition is made possible by the fact that the experimental *m*-value can be reproduced from the calculated *m*-value using the method of Auton and Bolen (24) and Auton

Author contributions: L.L.P. and G.D.R. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]Present address: Potomac Affinity Proteins, Institute for Bioscience and Biotechnology Research, 9600 Gudelsky Drive, Rockville, MD 20850.

[2]To whom correspondence should be addressed. E-mail: grose@jhu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1202604109/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1202604109

Downloaded at Palestinian Territory, occupied on December 9, 2021

www.manaraa.com

et al. (25), enabling the m-value of any protein fragment to be determined computationally. Most, but not all (26), domains are contiguous fragments, and the method introduced here is limited to such segments.

Using the linear extrapolation method (20), $m = \frac{\Delta G - \Delta G^0}{C}$, where ΔG is the conformational free energy (i.e., the free energy difference between the folded and unfolded states) at urea concentration, C, and ΔG^0 is the conformational free energy in buffer. Consequently, the qualifying ratio

$$QR = \frac{m_{\text{excised}}}{m_{\text{in\_situ}}} = \frac{\frac{\Delta G_{\text{excised}} - \Delta G^0_{\text{excised}}}{C}}{\frac{\Delta G_{\text{in\_situ}} - \Delta G^0_{\text{in\_situ}}}{C}} = \frac{\Delta\Delta G_{\text{excised}}}{\Delta\Delta G_{\text{in\_situ}}},$$

is a dimensionless ratio of the conformational free energy differences of the fragment in isolation and in its parent structure. At a value near unity, the destabilizing effects of urea on the excised fragment are not affected significantly upon inclusion of any *in situ* interactions between that fragment and its parent protein, an indication that the fragment unfolds as an independent cooperative unit. A fragment satisfying the condition that $QR \approx 1$ is classified as a domain.

In contrast to the typical structure-based domain definitions cited above, our thermodynamically-based definition describes domains as self-contained, cooperative folding units. With this definition, such units need not be independently stable, and in any case, stability necessarily depends upon temperature, pressure, and cosolvent conditions. Indeed, assessment of independent stability is beyond the scope of any algorithm that does not include such factors as variables. Even ostensibly disordered proteins can often be forced to fold upon addition of protecting osmolytes such as trimethylamine N-oxide (TMAO) (27). Surprisingly, suggestive earlier studies on the thermodynamic characterization of structural components (28) have not recognized the linkage between structural domains and cooperative units. To our knowledge, the definition proposed here has not been used previously.

As described next, our domain classifications are consistent with experiment in nine cases for which experimentally determined equilibrium folding intermediates are available. An additional set of 71 representative proteins was classified as well; 45 were consistent with CATH classifications, but the remaining 26 differ, usually by identifying a larger number of domains—an experimentally testable prediction. Comparison with SCOP produced similar statistics. Our algorithm's frequent agreement with CATH and SCOP demonstrates that domains are often compatible with visual intuition, but the many instances of disagreement underscore textbook wisdom that there is more to thermodynamics than meets the eye.

Results

Domain Identification Algorithm. Domains were identified in solved protein structures by using our structure-energy equivalence of domains (SEED) algorithm. The minimum size of a domain was fixed at 25 residues, approximating the size of a unit of supersecondary structure (29) and the minimum chain length needed to attain a protein-like surface/volume ratio (see figure 2 in ref. 30). No fixed limit was imposed on the maximum size. The algorithm identifies an optimal set of nonoverlapping units that maximizes both collective QRs and chain coverage. The procedure is summarized here and further described in *Methods*.

Step 1: Exhaustive search: A polypeptide chain of length N residues is subdivided into all possible contiguous n-residue segments ($30 \leq n \leq N$), and the QR of each is calculated. For each n, the three segments with the highest QRs and $\leq$50% overlap are retained for step 2.

Step 2: Likely domain selection: Overlapping segments with monotonically increasing lengths are grouped, and the segment(s) with locally maximal QRs are flagged as potential domains.

Step 3: Refinement: Potential domains are combined so as to optimize a scoring function. Close-scoring alternative divisions are also retained.

SEED Classifications Are Consistent with Experiment. A literature search turned up nine studies of experimentally determined, equilibrium folding intermediates, and in each, our domain divisions correspond to these intermediates (Fig. 1; Table S1). In contrast, the number of intermediates exceeds the number of domains identified by either CATH (14) or SCOP (15) in eight of these nine cases.

In greater detail, SEED decomposition of (*i*–*iii*) T4 phage lysozyme, α-lactalbumin, and OspA is consistent with the experimentally determined boundaries of intermediate structures (31–34). (*iv*) β-Lactamase was found to unfold into two equilibrium intermediates with a concurrent decrease in the helical CD signal of each (35), consistent with SEED decomposition into a three-domain protein with two small partially helical domains. (*v*) Notch ankyrin was split between repeats four and five, consistent with the finding that repeats 1-4 fold as a single cooperative unit (36). (*vi*) Sperm whale apomyoglobin was subdivided into two fragments, helices A-E and helices F-H; the latter fragment resembles an isolated apomyoglobin equilibrium intermediate (37) but without helix A. Experimental evidence for a G-H helix intermediate is inconclusive in sperm whale apomyoglobin, but the intermediate has been detected in the structurally similar equine myoglobin variant (38). (*vii*) HSP33 was decomposed into three domains. Although standard domain classification methods divide this protein into a C- and N-terminal unit, experimental observations have shown that the two clusters of helices in the N-terminal domain can fold independently (39), consistent with SEED division of the

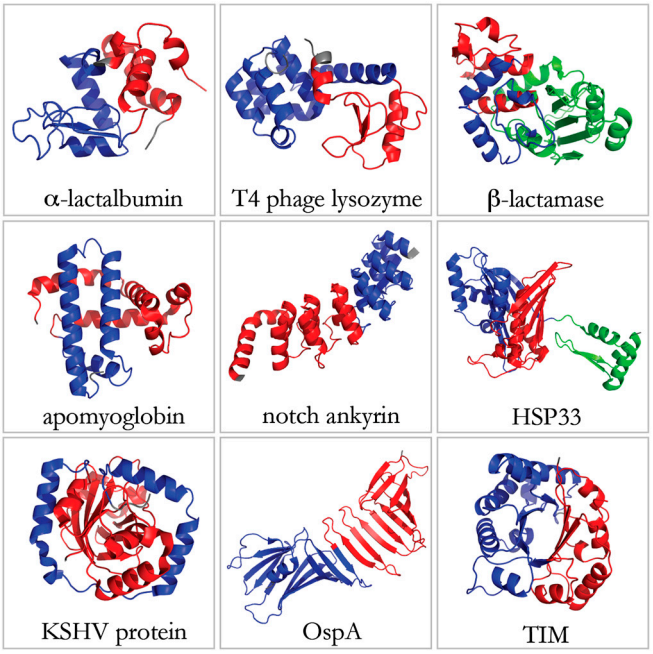

Fig. 1. SEED classifications of nine proteins with experimentally determined equilibrium folding intermediates. All are consistent with experimental observations. Distinct SEED domains are color-coded; colors progress sequentially from N- to C-terminal in order red, blue, and green; unclassified segments are shown in gray. Protein Data Bank (PDB) and chain identifiers are: row 1 (1alc_, 206l_, 3blm_); row 2 (101m_, 1ot8A, 1vzyA); row 3 (2pbkB, 2g8c_, 1wyiA).

Downloaded at Palestinian Territory, occupied on December 9, 2021

www.manaraa.com

N-terminal unit into two domains, each containing one helical cluster. (*viii*) KSHV protease was divided into two domains: residues 4–161, the protein core, and residues 162–226, most of which is known to be unfolded in the monomeric form but folded upon dimerization (40). (*ix*) Decomposition of human TIM resulted in two close-scoring classifications—one with two domains and the other with three. Experiments have demonstrated that rabbit muscle TIM, structurally similar to human TIM, contains two domains (19), but others have shown *Trypanosoma brucei* TIM contains three domains (18), raising the possibility that human TIM also has an alternative three-domain structure. Neither CATH nor SCOP detected the intermediates detailed above in eight of these nine cases (Table S1), a vivid demonstration that visually based methods are blind to thermodynamics.

Domain Classification. A representative, structurally diverse set of 71 additional proteins was selected for analysis and comparison with CATH. These 71 structures have between one and three CATH-defined domains; proteins with more than three domains are rare and were excluded. Specifically, there are 31,286 CATH structures with resolution ≤2.0 Å and more than 98% have three domains or fewer (22,104, 7,525, and 1,184 have one, two, or three domains, respectively). To assure completeness, every CATH architecture was represented in at least one single-, double- and triple-domain protein if available.

SEED and CATH classifications differ significantly: The two methods assigned dissimilar boundaries to over one-third (26/71) of the structures analyzed (Fig. 2, Table 1, and Table S2). Among those that differ, seven of the SEED classifications are supported by experimental data, but similar evidence is thus far unavailable for the remaining 19 (Table S2). An almost equal proportion (22/71) differ in the number of discovered domains; on average, SEED divides proteins into a larger number of smaller substructures than CATH. Explicitly, when these 22 structures are categor-

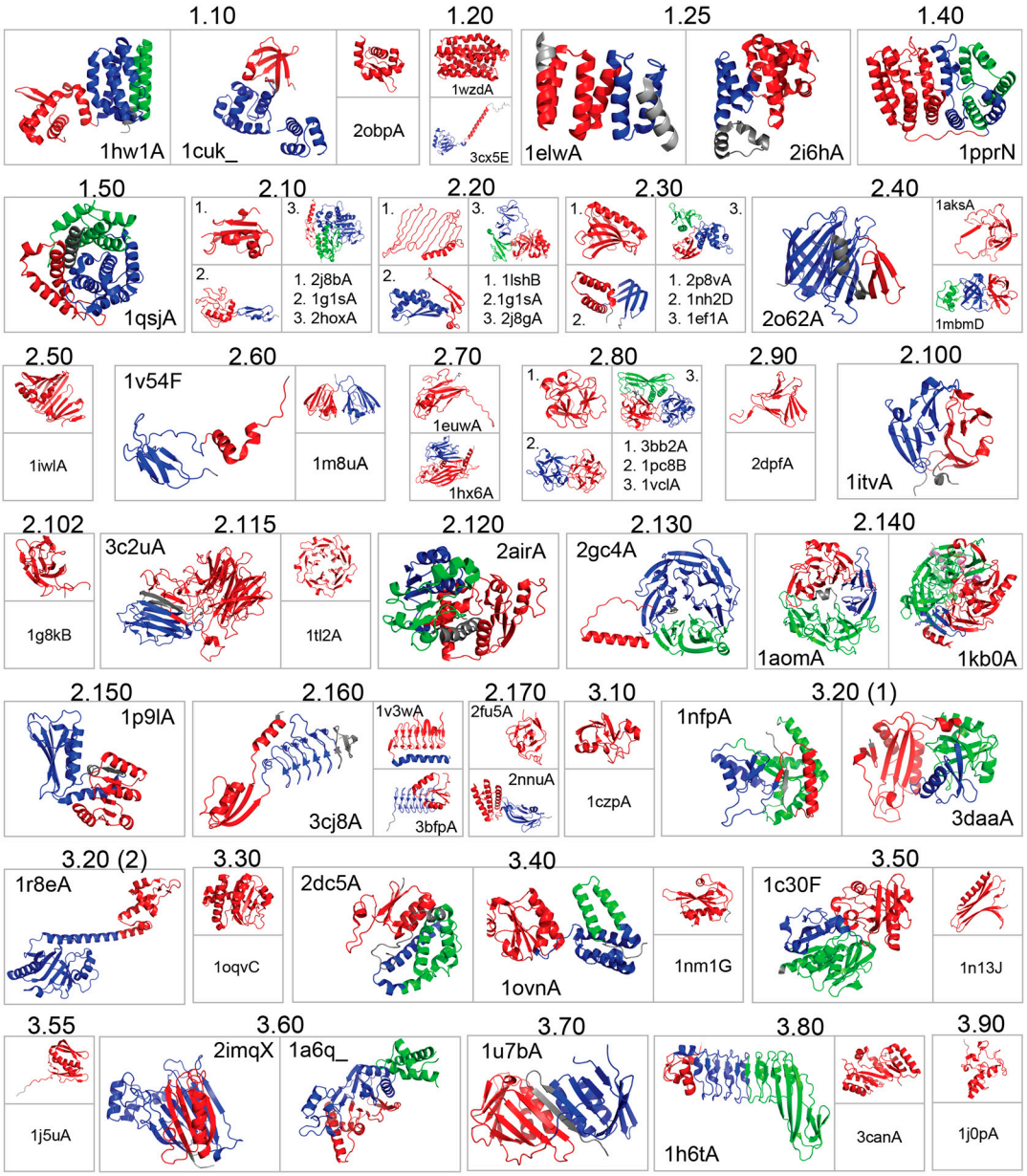

Fig. 2. SEED classifications of a structurally diverse protein set. Large boxes indicate disagreement with CATH; small boxes indicate agreement. Numbers above boxes denote CATH architecture. Distinct SEED domains are color-coded as in Fig. 1, in order red, blue, green and purple; unclassified segments in gray. PDB ID codes and chain identifiers are shown. The figure includes 63 of the 71 proteins. Table S2 lists all domain boundaries, including the remaining eight (1h6uA, 2rikA, 1eflA, 1tdqA, 3i7w_, 5i1b_, and 1stn_, and 1ubq_).

Downloaded at Palestinian Territory, occupied on December 9, 2021
www.manaraa.com

Table 1. Comparison of SEED domain classifications with CATH and SCOP

PDB ID*,†	SEED‡	CATH‡,§	SCOP‡,¶	PDB ID*,†	SEED‡	CATH‡,§	SCOP‡,¶
2obpA	1	1	1	1aomA	3	1	1
1hw1A	3	2	2	1kb0A	4	1	2
1cuk_	2	3	3	1p9lA	2	2	2
1wzdA	1	1	1	1v3wA	1	1	1
3cx5E	2	2	2	3bfpA	2	2	1
1elwA	2	1	1	3cj8A	2	3	X
2i6hA	2	~~2~~	1	2fu5A	1	1	1
1pprN	3	2	2	2nnuA	2	2	1
1qsjA	3	1	1	1czpA	1	1	1
2j8bA	1	1	1	1ewfA	2	~~2~~	~~2~~
1g1sA	2	2	2	1nfp_	3	1	1
2hoxA	3	~~3~~	1	3daaA	3	2	1
1lshB	1	1	1	1r8eA	2	3	~~2~~
1zb9A	2	2	X	1oqvC	1	1	1
2j8gA	3	3	2	1nm1G	1	1	1
2p8vA	1	1	X	2dc5A	3	2	X
1nh2D	2	2	2	1n13J	1	1	1
1ef1A	3	3	3	1c30F	3	2	2
1aksA	1	1	1	1j5uA	1	1	1
2o62A	2	2	2	2imqX	2	1	1
1mbmD	3	3	1	1a6q_	3	2	2
1iwlA	1	1	1	1u7bA	2	1	2
1v54F	2	1	1	1h70A	1	1	1
1m8uA	2	2	2	3canA	1	1	X
1euwA	1	1	1	1h6tA	3	2	2
1hx6A	2	2	2	1h6uA	3	~~3~~	2
3bb2A	1	1	X	1j0pA	1	1	1
3ah2A	2	2	X	1ovnA	3	2	2
1vclA	3	3	3	2rikA	3	3	X
2dpfA	1	1	X	1eflA	3	2	2
1itvA	2	1	1	1tdqA	3	3	3
1g8kB	1	1	1	3i7w_	1	1	X
1tl2A	1	1	1	5i1b_	1	1	1
3c2uA	2	2	X	1stn_	1	1	1
2airA	3	2	2	1ubq_	1	1	1
2gc4A	3	1	1				

*Color-coded comparisons: SEED, CATH, and SCOP agree (black); SEED and CATH differ (green); SEED and SCOP differ (purple); SEED and both CATH and SCOP differ (red).

†Last character indicates PDB chain; single chains indicated by underscore symbol (_).

‡Domain boundaries are listed in Table S2.

§Numbers with strikethrough indicate equal domain numbers but different domain boundaries in comparison to SEED.

¶X indicates unclassified proteins; comparison with SEED is not possible and color code is not applied.

ized by the number of discovered domains, SEED vs. CATH identified 25 vs. 36 single-domain proteins, 22 vs. 24 with two domains, 22 vs. 11 with three domains, and one vs. zero with four domains (Table S2). Among the similarly classified proteins, over half (24/45) are single-domain structures with fewer than 200 residues.

Turning now to the seven proteins for which corroborative experimental evidence is available, four of them—(*i*) the TPR1 domain of Hop, (*ii*) the methylamine dehydrogenase heavy chain, (*iii*) nitrite reductase, and (*iv*) glutathione *S*-transferase mu7—either have one or more detectable folding intermediates or are structurally similar to other proteins that have such intermediates. In each case, the number of intermediates is equal to the number of SEED domains, although the experimental data are insufficient to define specific domain boundaries. The TPR1 domain of Hop (1elwA) spans several tetratricopeptide (TPR) repeats; in another protein, isolated TPR repeats were shown to undergo conformational transition, but multiple TPR repeats are mutually stabilizing, as expected for a cooperative domain (41). Both the methylamine dehydrogenase heavy chain (2gc4A) and nitrite reductase (1aomA) are β-propeller proteins that subdivide into the same three domains (Fig. 2 and Table 1), and similar folding intermediates have been observed in other β-propellers (42). Furthermore, the SEED-defined N-terminal domain glutathione *S*-transferase mu 7 (2dc5A) has an independently stable structural counterpart (43). The remaining three proteins—(*v*) the PEX domain of MMP9, (*vi*) BmrR, and (*vii*) Internalin B—exhibit modular independence, as expected for a domain. SEED divides the PEX domain of MMP9 (1itvA) into two domains, one of which shifts upon dimerization (44). BmrR (1r8eA) is also divided into two domains. The protein contains two small globular regions interconnected by a long intervening helix (Fig. 2 and Table 1); CATH dissects these two globules and the helix into three separate domains. However, the 33-residue N-terminal fragment of this helix folds with the N-terminal globule (45), indicating that the entire helix need not fold as a single unit, consistent with the SEED classification that groups a portion of this helix with the N-terminal domain. Internalin B (1h6tA), a protein from Listeria, contains a leucine rich repeat (LRR) protein flanked by a truncated EF-hand-like cap and an immunoglobulin-like fold (46). SEED classifies the EF-hand-like protein as one domain, but then divides the LRR into its first four repeats, with the last three repeats plus the Ig-like domain grouped together. In a different LRR protein, YopM, it was found that the first four and one half repeats plus N-terminal hairpin can fold independently (47), and the four N-terminal LRR units of Internalin B may fold similarly.

Finally, we note that 19 of the 79 proteins analyzed here had two or more high-scoring domain decompositions. For example, the top-scoring division of the notch-ankyrin domain (1o8tA) divides the protein between repeats four and five, but a close-scoring alternative divides the protein between repeats five and six. In fact, both divisions are consistent with experimental data: Repeats one through four are known to fold as a single cooperative unit (36), whereas repeats one through five are known to fold stably in solution (48). Although the highest scoring division of human BPI (1ewfA) differs from the corresponding CATH classification, a close runner-up does have similar boundaries. We also note that a moderately close-scoring alternative ubiquitin decomposition separates the two N-terminal strands from the rest of the protein, consistent with NMR experiments showing that these two strands can fold independently under nonphysiological solution conditions (60% methanol; 40% water; pH < 2) (49). On reflection, there is no inherent reason why domain divisions should be unique. Alternative decompositions are reported in Tables S1 and S2.

Comparison with Surface Area. *QR*s are ratios of *m*-values, quantities proportional to the summed groupwise transfer free energies scaled by changes in the solvent-accessible surface area (SASA) of each group. Other algorithms identify domains using SASA alone (9), prompting us to question whether inclusion of transfer free energies makes a meaningful difference. To answer this question, the *QR* was redefined from a ratio of *m*-values to a corresponding ratio of solvent accessibilities, $\frac{\text{SASA}_{\text{in\_situ}}}{\text{SASA}_{\text{excised}}}$, and all proteins were reanalyzed.

Unlike *m*-value based decomposition, SASA-based decomposition detected only one equilibrium intermediate in the nine-protein test set, a clear indication that experimentally compatible domain classification is improved by inclusion of scaled energies (Table S1). Differences with the 71 protein test set were detected as well (Table S2). To probe the basis of this disparity, *m*-value and SASA ratios of backbone-only, side chain-only, and backbone + side chain were determined for the all domains. Backbone vs. backbone + side chain *m*-value ratios are strongly correlated ($\rho = 0.88$), whereas side chain vs. backbone + side chain *m*-value ratios are essentially uncorrelated ($\rho = 0.14$), indicating that the backbone is the major determinant of *m*-value ratios. In contradistinction, for SASA ratios, side chain vs. backbone + side chain SASA ratios are strongly correlated ($\rho = 1.00$), indicating that side chains are the major determinant of SASA ratios; the corresponding comparison for backbone (i.e., backbone + side chain SASA) is less well-correlated ($\rho = 0.83$).

Downloaded at Palestinian Territory, occupied on December 9, 2021
www.manaraa.com

These correlations are in agreement with experiment, where it is known that the peptide backbone plays the determinative role in urea denaturation (24, 25, 50). Upon unfolding, the exposure of backbone units is the predominant energy term although it accounts for only approximately 25% of the total newly exposed SASA.

Discussion

From molecules to skyscrapers, persisting structure is ultimately a consequence of the unseen stabilizing energetics. Accordingly, we have shifted the domain recognition problem from conventional structure-based methods to one that is thermodynamically-based, implemented here via *m*-value ratios. In essence, structural domains have been equated to cooperative units, for which rigorous identification is possible. This interpretation of a domain agrees with experimentally observed equilibrium intermediates (Fig. 2 and Table 1), and it also agrees with visual intuition often enough to be plausible (Fig. 2 and Table 1).

Our method is deliberately focused on equilibrium structures. However, many of the intermediates cited here were detected using denaturants other than urea. Why should a water → urea *m*-value identify an intermediate induced by an alternative method of denaturation? For a highly cooperative process, like the denatured ⇌ native folding reaction, a population of natively folded molecules persists even under predominantly denaturing conditions (51). Whereas different cosolvents may affect the concentration of denaturant needed to achieve a given degree of destabilization, they do not result in alternative folds (52). However, it is possible that differences in stability may give rise to differences in domain divisions, particularly in those proteins having multiple, close-scoring alternative decompositions (*Results*). Even visually imperceptible changes in boundary selection can impact domain stability significantly. For example, the third fibronectin type III domain from human tenascin was initially defined to be 90 residues long, but it was later found that a two-residue C-terminal extension stabilizes the structure by approximately an additional 3 kcal/mol (53).

Previous algorithms to determine domain boundaries have focused primarily on side chain interactions. Instead, our method is based largely on exposure of backbone surface area, the predominant energy term by far when forcing either folding or unfolding using natural osmolyte cosolvents (24, 25, 27, 50).

It has been assumed that there is no upper limit on the size of a domain because large (>300 residues) domains are present in nature (16). To our knowledge, all other methods of domain recognition include single-domain proteins with 300 residues or more, such as the TIM barrel. Thus far, our results indicate that large proteins are typically composites of smaller—and often less obvious—domains. Of our two largest identified domains in 3c2u and 1kb0A, each with more than 300 residues, the latter one allows for an alternative decomposition in which this large domain is divided in half (1kb0A in Table S2). These results suggest that domains may in fact be subject to a fundamental size limit. If so, arguments about the limited number of protein folds would be directly applicable to protein domains (54). Systematic enumeration of all possible domains would provide a basis set for protein architecture (55) and chart a way for pursuits in protein design and engineering.

Methods

Surface Area Calculations. Solvent-accessible surface areas were calculated based on the method of Lee and Richards (56) using a 1.4 Å water probe and the following atomic radii: tetrahedral carbons, 2.00 Å; carboxamide and carboxylic acid carbons, 1.70 Å; aromatic carbons, 1.85 Å; ammonium nitrogens, 2.00 Å; amide and aromatic nitrogens, 1.70 Å; guanidino nitrogens, 1.80 Å; imide nitrogens, 1.50 Å; carbonyl oxygens, 1.40 Å; phenol and alcohol oxygens, 1.60 Å; all other oxygens, 1.50 Å; thiol sulfurs, 2.00 Å; thioether sulfers, 1.85 Å; and backbone amide hydrogens, 1.0 Å.

Blocking Groups. Excised fragments were terminated by added N-terminal acetyl and C-terminal *N*-methyl amide blocking groups.

QR Distribution. Both *QR*s and SASAs were calibrated against values observed for CATH domains (14). Explicitly, the frequencies of *QR*s and SASA ratios were calculated for all CATH domains with resolution ≤2.0 Å. Distributions peaked in the range 0.9–1.05 for *QR*s (Fig. S1*A*) and 0.94–1.06 for SASAs (Fig. S1*B*). Accordingly, putative domains with *QR*s/SASAs within these respective ranges were considered more likely to be authentic domains. CATH version 3.4 and SCOP version 1.75 were used in all comparisons.

Scoring. Each cluster of putative domains was scored as

$$\text{Score} = g * \left[\left(\sum_{i=1}^{n} QR_i \cdot r_i \right) - \left(\frac{\sum_{i=1}^{n} QR_i}{n} \right) \cdot \left(R - \sum_{i=1}^{n} r_i \right) \right],$$

where QR_i is the *QR* of putative domain *i*, r_i is the number of residues in putative domain *i*, *n* is the number of putative domains in the cluster, and *R* is the total number of residues in the protein. The first parenthesized term is an overall *QR* weight for the cluster, which is reduced by the average *QR* (second parenthesized term) scaled by the extent to which the clusters cover the protein (third parenthesized term). The weighting factor, *g*, is introduced to bias the score against overweighting a larger number of smaller fragments or, conversely, by a protein-sized large fragment. Specifically, let n_j = the number of putative domains in the range 0.9–1.05 (or 0.94–1.06 for *SASA*s) with $r_i \geq 60$ and n_k = the number in this range with $r_i < 60$, excluding chain termini; $n_j + n_k = n$. Then $g = n_j - n_k$ unless $n_k = 1$, in which case $g = n_j + 1 = n$. In words, *g* is incremented (rewarded) for each putative domain in the cluster ≥60 but decremented (penalized) for each putative domain <60, biasing the score in favor of domains ≥60 residues and against small domains between 25–59 residues (25 residues is the minimum allowed size for any domain). However, a singlèton domain <60 residues is rewarded, not penalized. If the final value of $g \leq 0$, this factor is set to unity. Fragments with *QR*s ≥ 1.2 were not considered.

Grouping. Overlapping segments of monotonically increasing length were accumulated into groups, with two conditions:

Condition 1: $(\forall i \in \text{group})\ \text{length}(\text{segment}_{i+1}) - \text{length}(\text{segment}_i) < 20 \text{residues}$, and

Condition 2: $(\forall i \in \text{group})\ (\text{segment}_i \cup \text{segment}_{i+1}) - (\text{segment}_i \cap \text{segment}_{i+1}) \leq 15$ residues.

Condition 1 restricts the increase in length between successive segments and condition 2 restricts the length of the nonoverlapping region between successive segments. Members of each group were sorted by length from shortest to longest.

Locally Maximal QRs. The *QR* of segment$_i$ was defined to be locally maximal if $QR(\text{segment}_{i-1}) < QR(\text{segment}_i) > QR(\text{segment}_{i+1})$. However, the first/last segment in the group was defined to be locally maximal if its *QR* was greater than that of its immediate successor/predecessor.

Segments with a locally maximal *QR* were further refined by determining whether a minor extension (<10 residues) improved the *QR*. Explicitly,

for a given segment from residue$_i$ to residue$_j$ (length $j - i + 1$), and for window extensions, $\mathscr{W}$, $\mathscr{W} \in [0, 9]$ find $\max\{QR^{w}_{x=0}(\text{residue}_{i-x} \ldots \text{residue}_{j+(w-x)})\}$.
All variables are positive integers.

This procedure extends the segment by as many as nine residues and finds the maximum *QR* of the augmented set.

A further optimization test was applied for instances of two or more disjoint putative domains that constitute a larger domain (but not as large as the entire protein). In such cases, the larger domain was excised from the protein and divided into the previously identified smaller domains. *QR*s of these smaller domains were then recalculated, and if they increased, the cluster was rescored.

Boundary Refinement. Minor changes in boundaries between adjoining domains within a domain cluster can change the score, and in turn, affect the rank of that cluster in comparison with alternative clusters. To account for this possibility, adjoining domains within clusters having scores ≥99% of

Downloaded at Palestinian Territory, occupied on December 9, 2021
www.manaraa.com

the highest scoring cluster were jointly excised from the parent protein and subjected to boundary refinement. Explicitly, the dividing point between adjacent domains was shifted one residue at a time, exhaustively, and the site of maximal *QR* for both domains was selected as the boundary. Upon completion, the cluster score was recalculated. In a few exceptional cases, the resulting candidates were rejected if one or more of the domains was below the minimum size (≤25 residues), in which case the threshold was lowered to 90% and, if necessary, further decremented in steps of 10% until these refinement criteria were satisfied.

Experimentally Characterized Protein Set. The nine experimentally characterized proteins (Fig. 1) were analyzed as described above. However, protein termini are often crisscrossed (57), and in this event, the protein was circularly permuted computationally so as to attach the N-terminal segment to the C terminus or vice versa, followed by manual recalculation of the *QR*s, using boundaries from the original SEED calculation. Boundaries were adjusted up to 10 residues such that all *QR*s ≥ 0.9. However, if this condition could not be satisfied, boundaries from the next SEED runner-up were used instead.

ACKNOWLEDGMENTS. We thank Matthew Auton for valuable discussion and Buzz Baldwin and Pat Fleming for useful suggestions. Support from the Mathers Foundation and the National Science Foundation is gratefully acknowledged.

1. Doolittle RF (1995) The multiplicity of domains in proteins. *Annu Rev Biochem* 64:287–314.
2. Phillips DC (1966) The three-dimensional structure of an enzyme molecule. *Sci Am* 215:78–90.
3. Drenth J, Jansonius JN, Koekoek R, Swen HM, Wolthers BG (1968) Structure of papain. *Nature* 218:929–932.
4. Wetlaufer DB (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci USA* 70:697–701.
5. Jowett B (1937) *The Dialogues of Plato; Phaedrus [section 266]* (Random House, New York).
6. Crippen GM (1978) The tree structural organization of proteins. *J Mol Biol* 126:315–332.
7. Rose GD (1979) Hierarchic organization of domains in globular proteins. *J Mol Biol* 134:447–470.
8. Lesk AM, Rose GD (1981) Folding units in globular proteins. *Proc Natl Acad Sci USA* 78:4304–4308.
9. Wodak SJ, Janin J (1981) Location of structural domains in protein. *Biochemistry* 20:6544–6552.
10. Holm L, Sander C (1994) Parser for protein folding units. *Proteins* 19:256–268.
11. Zehfus MH (1994) Binary discontinuous compact protein domains. *Protein Eng* 7:335–340.
12. Tsai CJ, Maizel JV, Jr, Nussinov R (2000) Anatomy of protein structures: Visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc Natl Acad Sci USA* 97:12038–12043.
13. Tai CH, et al. (2011) Protein domain assignment from the recurrence of locally similar structures. *Proteins* 79:853–866.
14. Orengo CA, et al. (1997) CATH—a hierarchic classification of protein domain structures. *Structure* 5:1093–1108.
15. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540.
16. Richardson JS (1981) The anatomy and taxonomy of protein structure. *Adv Protein Chem* 34:167–339.
17. Bruning JB, Shamoo Y (2004) Structural and thermodynamic analysis of human PCNA with peptides derived from DNA polymerase-delta p66 subunit and flap endonuclease-1. *Structure* 12:2209–2219.
18. Chanez-Cardenas ME, et al. (2002) Unfolding of triosephosphate isomerase from Trypanosoma brucei: Identification of intermediates and insight into the denaturation pathway using tryptophan mutants. *Arch Biochem Biophys* 399:117–129.
19. Pan H, Raza AS, Smith DL (2004) Equilibrium and kinetic folding of rabbit muscle triosephosphate isomerase by hydrogen exchange mass spectrometry. *J Mol Biol* 336:1251–1263.
20. Greene RF, Jr, Pace CN (1974) Urea and guanidine hydrochloride denaturation of ribonuclease, lysozyme, alpha-chymotrypsin, and beta-lactoglobulin. *J Biol Chem* 249:5388–5393.
21. Pace CN (1986) Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol* 131:266–280.
22. Robertson AD, Murphy KP (1997) Protein structure and the energetics of protein stability. *Chem Rev* 97:1251–1268.
23. Dill KA, Ghosh K, Schmit JD (2011) Physical limits of cells and proteomes. *Proc Natl Acad Sci USA* 108:17876–17882.
24. Auton M, Bolen DW (2005) Predicting the energetics of osmolyte-induced protein folding/unfolding. *Proc Natl Acad Sci USA* 102:15065–15068.
25. Auton M, Holthauzen LM, Bolen DW (2007) Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc Natl Acad Sci USA* 104:15317–15322.
26. Zehfus MH (1997) Identification of compact, hydrophobically stabilized domains and modules containing multiple peptide chains. *Protein Sci* 6:1210–1219.
27. Bolen DW, Rose GD (2008) Structure and energetics of the hydrogen-bonded backbone in protein folding. *Annu Rev Biochem* 77:339–362.
28. Murphy KP, Bhakuni V, Xie D, Freire E (1992) Molecular basis of cooperativity in protein folding. III. Structural identification of cooperative folding units and folding intermediates. *J Mol Biol* 227:293–306.
29. Levitt M, Chothia C (1976) Structural patterns in globular proteins. *Nature* 261:552–558.
30. Rose GD, Wetlaufer DB (1977) The number of turns in globular proteins. *Nature* 268:769–770.
31. Chen L, Hodgson KO, Doniach S (1996) A lysozyme folding intermediate revealed by solution X-ray scattering. *J Mol Biol* 261:658–671.
32. Kuwajima K (1996) The molten globule state of α-lactalbumin. *FASEB J* 10:103–109.
33. Wu LC, Peng ZY, Kim PS (1995) Bipartite structure of the α-lactalbumin molten globule. *Nat Struct Biol* 2:281–286.
34. Koide S, et al. (1999) Multistep denaturation of Borrelia burgdorferi OspA, a protein containing a single-layer β-sheet. *Biochemistry* 38:4757–4767.
35. Uversky VN, Ptitsyn OB (1994) Partly folded state, a new equilibrium state of protein molecules: Four-state guanidinium chloride-induced unfolding of β-lactamase at low temperature. *Biochemistry* 33:2782–2791.
36. Mello CC, Barrick D (2004) An experimentally determined protein folding energy landscape. *Proc Natl Acad Sci USA* 101:14102–14107.
37. Hughson FM, Wright PE, Baldwin RL (1990) Structural characterization of a partly folded apomyoglobin intermediate. *Science* 249:1544–1548.
38. Chi Z, Asher SA (1998) UV resonance Raman determination of protein acid denaturation: Selective unfolding of helical segments of horse myoglobin. *Biochemistry* 37:2865–2872.
39. Graf PC, et al. (2004) Activation of the redox-regulated chaperone Hsp33 by domain unfolding. *J Biol Chem* 279:20529–20538.
40. Nomura AM, Marnett AB, Shimba N, Dotsch V, Craik CS (2005) Induced structure of a helical switch as a mechanism to regulate enzymatic activity. *Nat Struct Mol Biol* 12:1019–1020.
41. Taylor P, et al. (2001) Two structures of cyclophilin 40: Folding and fidelity in the TPR domains. *Structure* 9:431–438.
42. Juhasz T, Szeltner Z, Fulop V, Polgar L (2005) Unclosed β-propellers display stable structures: Implications for substrate access to the active site of prolyl oligopeptidase. *J Mol Biol* 346:907–917.
43. Bhattacharyya S, et al. (2002) Identification of a novel archaebacterial thioredoxin: Determination of function through structure. *Biochemistry* 41:4760–4770.
44. Cha H, Kopetzki E, Huber R, Lanzendorfer M, Brandstetter H (2002) Structural basis of the adaptive molecular recognition by MMP9. *J Mol Biol* 320:1065–1079.
45. Newberry KJ, Brennan RG (2004) The structural mechanism for transcription activation by MerR family member multidrug transporter activation, N terminus. *J Biol Chem* 279:20356–20362.
46. Schubert WD, et al. (2001) Internalins from the human pathogen *Listeria monocytogenes* combine three distinct folds into a contiguous internalin domain. *J Mol Biol* 312:783–794.
47. Kloss E, Barrick D (2009) C-terminal deletion of leucine-rich repeats from YopM reveals a heterogeneous distribution of stability in a cooperatively folded protein. *Protein Sci* 18:1948–1960.
48. Zweifel ME, Barrick D (2001) Studies of the ankyrin repeats of the Drosophila melanogaster Notch receptor. 2. Solution stability and cooperativity of unfolding. *Biochemistry* 40:14357–14367.
49. Stockman BJ, Euvrard A, Scahill TA (1993) Heteronuclear three-dimensional NMR spectroscopy of a partially denatured protein: The A-state of human ubiquitin. *J Biomol NMR* 3:285–296.
50. Cannon JG, Anderson CF, Record MT, Jr. (2007) Urea-amide preferential interactions in water: Quantitative comparison of model compound data with biopolymer results using water accessible surface areas. *J Phys Chem B* 111:9675–9685.
51. Lattman EE, Rose GD (1993) Protein folding—what's the question? *Proc Natl Acad Sci USA* 90:439–441.
52. Rose GD, Fleming PJ, Banavar JR, Maritan A (2006) A backbone-based theory of protein folding. *Proc Natl Acad Sci USA* 103:16623–16633.
53. Hamill SJ, Meekhof AE, Clarke J (1998) The effect of boundary selection on the stability and folding of the third fibronectin type III domain from human tenascin. *Biochemistry* 37:8071–8079.
54. Przytycka T, Aurora R, Rose GD (1999) A protein taxonomy based on secondary structure. *Nat Struct Biol* 6:672–682.
55. Perskie LL, Street TO, Rose GD (2008) Structures, basins, and energies: A deconstruction of the Protein Coil Library. *Protein Sci* 17:1151–1161.
56. Lee B, Richards FM (1971) The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol* 55:379–400.
57. Krishna MM, Englander SW (2005) The N-terminal to C-terminal motif in protein folding and function. *Proc Natl Acad Sci USA* 102:1053–1058.

Downloaded at Palestinian Territory, occupied on December 9, 2021

www.manaraa.com