Journal Title: Proteins
ISSN: 0887-3585
Article Author(s): Holm L
Article Title: Parser for protein folding units.
Year: 1994 Jul
Volume: 19
Issue: 3
Pages: 256-68
PMID: 7937738
User's Comments: In color, if available.

PROTEINS: Structure, Function, and Genetics 19:256–268 (1994)

Parser for Protein Folding Units

Liisa Holm and Chris Sander
European Molecular Biology Laboratory, D-69012 Heidelberg, Germany

ABSTRACT **General patterns of protein structural organization have emerged from studies of hundreds of structures elucidated by X-ray crystallography and nuclear magnetic resonance. Structural units are commonly identified by visual inspection of molecular models using qualitative criteria. Here, we propose an algorithm for identification of structural units by objective, quantitative criteria based on atomic interactions. The underlying physical concept is maximal interactions within each unit and minimal interaction between units (domains). In a simple harmonic approximation, interdomain dynamics is determined by the strength of the interface and the distribution of masses. The most likely domain decomposition involves units with the most correlated motion, or largest interdomain fluctuation time. The decomposition of a convoluted 3-D structure is complicated by the possibility that the chain can cross over several times between units. Grouping the residues by solving an eigenvalue problem for the contact matrix reduces the problem to a one-dimensional search for all reasonable trial bisections. Recursive bisection yields a tree of putative folding units. Simple physical criteria are used to identify units that could exist by themselves. The units so defined closely correspond to crystallographers' notion of structural domains. The results are useful for the analysis of folding principles, for modular protein design and for protein engineering.**

Key words: unfolding, solvation, contact maps, protein design, structural domains, normal modes

INTRODUCTION

Proteins are linear polymers which fold into complicated three-dimensional shapes. From inspection of molecular models we know that in all but the smallest proteins, the polypeptide chains forms several compact, globular units, sometimes loosely connected. Such units are commonly called structural domains, although this definition based on visual inspection is intuitive and therefore rather imprecise. The goal of the present work is to provide an objective definition of structural domains, calculated unambiguously from the three-dimensional coordinates of a protein structure.

Structural domains are basic units of protein folding, function, and evolution. The increasing frequency with which apparently unrelated proteins are found to contain recurrent folding motifs suggests that the number of physically accessible folds is limited.[1] Limited proteolysis or genetic engineering can yield fragments of natural proteins which are capable of independently folding into the native structure (phosphoglycerate kinase, thermolysin, immunoglobulins, etc.). Modular architecture is an economical way to build up more complex entities. Mobile modules identified by sequence comparison are often structural domains.[2] For example, in giant structural proteins (spectrin, titin, fibronectin, etc.), internal sequence repeats reveal an underlying much simpler domain architecture. The structures of many isolated domains have been determined by NMR (fibronectin type III repeats, SH2 domains, SH3 domains, POU-specific domain, etc.). Gene duplication plus fusion is evident for example in aspartic proteinases (dimeric HIV protease vs. monomeric two-domain pepsin, chymosin, renin). Multifunctional enzymes can combine domains with different architecture, e.g. biotin repressor biotin holoenzyme synthetase (1bib in Fig. 5). Structural domains can carry complete binding functions (substrate and NAD-binding domains of alcohol and lactate dehydrogenase, etc.). Active sites are often located in clefts between domains and ligand binding can induce conformational changes where structural domains move as quasirigid bodies (hexokinase, maltose-binding protein, etc.).

A variety of techniques have been invented for locating (structural) domains in 3-D structures. These include inspection of distance maps,[3,4] clustering,[5] neighborhood correlation,[6,7] plane cutting,[8] interface area minimization,[9] specific volume minimization,[10] searching for mechanical hinge points,[11,12] maximization of compactness,[13,14] and maximization of buried surface area.[15,16] Most of these methods, in spite of their ingenuity, are not designed for detecting domains composed of more than one or two continuous pieces of chain (e.g., actin, Fig. 3E). Clustering algorithms are an exception, but they tend to give more fragmented units

Received January 4, 1994; revision accepted March 4, 1994.
Address reprint requests to Liisa Holm or Chris Sander, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany.

than the generally accepted notion of domains (ref. 5 and our unpublished results). With a rapidly growing pool of new structures, some of which represent new fold types, a new general algorithm may be useful. Here, we present a method based on the criterion of maximal interdomain fluctuation time proposed earlier by Sander.[17]

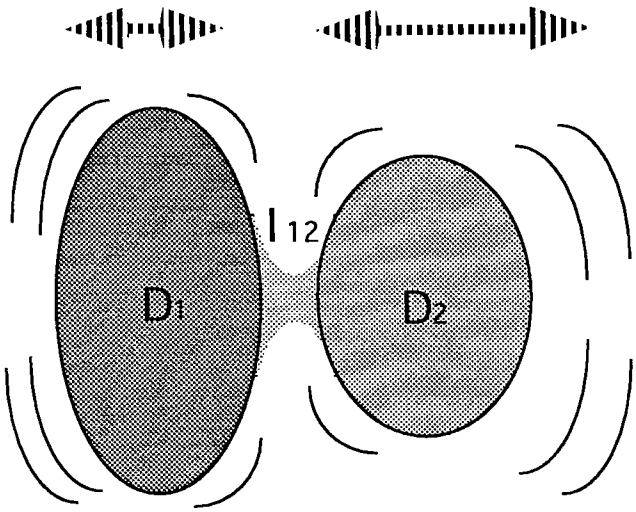
A protein may unfold in small bits and pieces (loops, ends) or in large units (structural domains). Let us focus on the second alternative and ask: What are the domains or folding units into which a globular protein separates as it unfolds? Intuitively, folding units are compact and the interactions between them weak. This intuition is made quantitative in a simple model (Fig. 1). In the underlying physical picture of the first stages of unfolding, there is a slow coherent relative motion of the units and mutual rearrangement of solvent and local protein structure near the interface between the units that results in gradual entry of solvent into the interface and finally spatial separation of independently solvated units connected by flexible hinges. For this process to occur, the relative motion of the units must be sufficiently slow to allow significant structural rearrangement: the slower, the better. As the relative motion of the units occurs on the same time scale as solvent motion, within an order of magnitude, the coupling between the two is strong and even small differences in the time scale may significantly affect the probability of unfolding. Therefore, in the present model, the main criterion for identifying folding units is the interunit fluctuation time, for which a lower limit, τ, can be calculated. For proteins of known three-dimensional structure, the model predicts the most likely decomposition into folding units.

We make the following extensions to the earlier[17] model: (1) division into units containing more than one continuous piece of chain; (2) recursive application to construct an unfolding tree; (3) distinction between nonpolar and polar interactions; (4) use of additional physical criteria to define the minimal requirements for independent structural domains. The domain dissection for a representative set of 330 proteins of known structure is reported and the physical/biological significance of the domain definition is discussed.

METHODS

General Idea of the Unfolding Model

Protein unfolding begins by the separation of two compact domains or folding units (D_1,D_2). The units interact via nonbonded atomic interaction at their interface and their relative motion is governed by the strength of the interface and the distribution of masses. The displacements of the units are assumed to be small enough for the harmonic approximation to be valid, and for solvent damping to be negligible. For an undamped harmonic oscillator the potential



$$\tau 2 \sim \frac{\mu}{I_{12}}$$

Fig. 1. Model of protein unfolding. Protein unfolding begins by the separation of two compact domains or folding units. Domains D_1 and D_2 interact via nonbonded atomic interactions at their interface I_{12} and their relative motion is governed by the strength of the interface and the distribution of masses. The most likely domain separation involves units for which the time constant of relative motion (τ) is largest. In the harmonic approximation, τ squared is proportional to the reduced mass (μ) divided by the strength of the interface.

V as a function of relative displacement x of the units is $V(x) = 1/2\ V_0x^2$, where V_0 is the force constant of the interface and $\omega^2 = V_0/\mu$ or $\tau^2 = (2\pi)^2\mu/V_0$, where μ is the reduced mass of the two units, ω is the angular frequency, and τ is the oscillation time. The dominant domain separation involves units for which the time constant of relative motion is largest.

τ as calculated is a lower limit. Displacements into the nonharmonic regime would give larger values for τ, due to the levelling off for larger x of the 6–12 potential for dispersion forces. Taking solvent contacts into account also would reduce the magnitude of the domain–domain interaction, yielding larger times τ. Qualitatively, inclusion of damping would not change the position of the best domain cut much, as both $\tau^2 \sim 1/V_o$ (undamped) and $\tau \sim 1/V_0$ (overdamped oscillator) are monotonic functions of the interface strength.

Calculation of τ

A quantitative estimate of τ is made by counting atoms and contacting atom pairs. Due to inaccuracies in the available atomic coordinates, we use a square well potential for atomic contacts rather

than evaluating a 6–12 potential (attraction $\sim r^{-6}$, repulsion $\sim r^{-12}$). Two atoms are in contact if their distance is ≤4.0 Å. A contacting atom pair is estimated to contribute $v_0 \approx 1.0$ kcal/mol/Å^2 (curvature of the Rehovot potential at the minimum) to the interface strength.[17] Backbone–backbone hydrogen bonds in β-sheets were defined using the program DSSP[18] and added to the contact matrix with a weight corresponding to 15 atom–atom contacts. From mutation experiments it is known that removing a methyl groups costs from 0.8 to 1.5 kcal/mol and a hydrogen bond stabilizes a protein by about 2.5 kcal/mol.[19] Methyl groups make maximally about 10 contacts, so the scaling between the van der Waals term and the H-bond term is reasonable. All other energy terms are ignored in the calculation of τ, and the contact map for the native conformation is also used for parts, making the assumption of no conformational changes.

The total interface strength is $V_0 = I_{12} v_0$, where I_{12} is the sum of interface contacts. The reduced mass is approximated as $\mu = m_c [N_1 N_2/(N_1 + N_2)]$, where m_C is the mass of a carbon atom (12 g/mol) and $N_{1,2}$ are the numbers of nonhydrogen atoms in domains $D_{1,2}$. Numerically,

$$\tau = \sqrt{\frac{N_1 N_2/(N_1+N_2)}{I_{12}}}\, 2\pi \sqrt{\frac{m_c}{v_0}} \approx \sqrt{\frac{N_1 N_2/(N_1 N+2)}{I_{12}}} \times 10^{-12} s\,. \qquad (1)$$

The expression under the first square root in Eq. (1) is typically of order 1 so that τ is of order 1 ps. Similar values are obtained in vacuum normal mode calculations.[20,21] If τ were small (motion fast) compared to solvent rearrangement times, interdomain motion would be averaged out before there can be any rearrangement of the solvent and domain interface structure. From known diffusion constants for water one can estimate: $\tau(\text{rotation}) \approx 10^{-13}$ s; $\tau(\text{diffusion over 3 Å}) \approx 2 \times 10^{-11}$ s. The resulting time scales suggest that typical motion of folding units is slow enough to allow solvent rearrangement. The slower the relative motion of the substructures, the more likely is such rearrangement. This argument is the basis for using $\tau(D_1,D_2)$ as the main criterion for determining the putative unfolding units.

Ordination of the Contact Matrix

The τ criterion can be used to select the most likely domain decomposition from a set of candidate bisections (D_1,D_2). Sander[17] tested all single cut points along the linear sequence. Wodak and Janin[9] extended their method to systematically search for two cut points. Here, we generalize the problem to finding a binary partition with any number of cut points in the sequence.

The maximum of $\tau(D_1,D_2)$ corresponds to a situation where rows/columns of the contact matrix have been rearranged by a permutation so that rows/columns $1, \ldots, k$ belong to D_1 and rows/columns $k+1, \ldots, L$ belong to D_2, where L is the number of residues in the intact unit, and the cut after k minimizes the number of interunit contacts I_{12} (modulo mass weighting). If one can find a permutation that makes the contact matrix block diagonal, then I_{12} equals zero and τ becomes infinite, but in general this is not possible. Clustering strongly interacting residues together (band diagonal matrix) yields a minimal number of interunit contacts for an arbitrary cut point k. A unique ordering of the residues is generated using a multivariate scaling method known as reciprocal averaging or correspondence analysis.[22,23] The analysis amounts to deriving scores for each residue so that the correlation of contacts (rows and columns) is maximized. Figure 2 illustrates a simple case.

Reciprocal averaging is a general method for the analysis of contingency tables with m columns and n rows, e.g., codon usage in differentially expressed genes.[24] Here, we present the special case of a symmetric contact matrix **A**. Let $r_i = \Sigma_j a_{ij}$ be the row totals ($a_{ij} \geq 0$). The reciprocal averaging procedure can be represented as the problem of determining a self-consistent set of residue scores (weights) x_i from

$$x'_i = \frac{\sum_j a_{ij} x_j}{r_i} \qquad (2)$$

where x' are the new scores and x are the old scores in an iterative averaging process. A self-consistent set of scores satisfies the eigenvalue problem $\rho\mathbf{x} = (\mathbf{R}^{-1}\mathbf{A})\mathbf{x}$, where **R** is a diagonal matrix of the row totals.

Some properties of the solutions of the eigenvalue problem follow.[23] There is a trivial solution (**1,1, 1,** . . .) with the maximal eigenvalue of 1, as it is not possible to exceed the limits of the original xs by the averaging procedure. Eigenvectors other than the first satisfy the relation $\Sigma_i \Sigma_j a_{ij} x_i = 0$, as the nontrivial eigenvectors are orthogonal to (**1,1,1,** . . .). The correlation of the scatter of points (weighted by contact strength) in a plot as shown in Figure 2C is

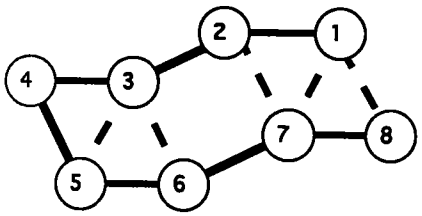
$$\rho = \frac{\sum_i \sum_j a_{ij} x_i x_j}{\sum_i \sum_j a_{ij} x_i^2} \qquad (3)$$

where we have used the constraint $\Sigma_i \Sigma_j a_{ij} x_i = 0$ to center the "points" with a weighted mean of zero. Routine differentiation using a Lagrange multiplier for the constraint shows that the stationary values of ρ are found when (ρ,x) is a nontrivial solution of the reciprocal averaging problem.

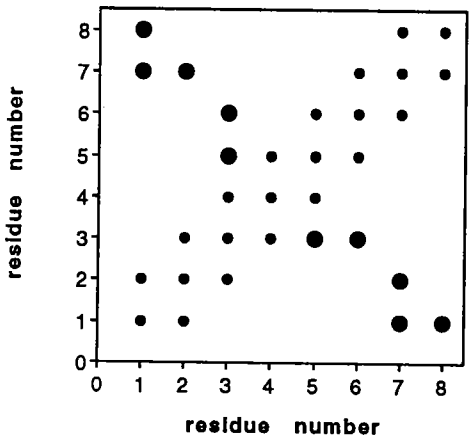
Computation of Eigenvectors

Following Hill[22] we iteratively solve for the eigenvectors of the positive-semi-definite symmetric

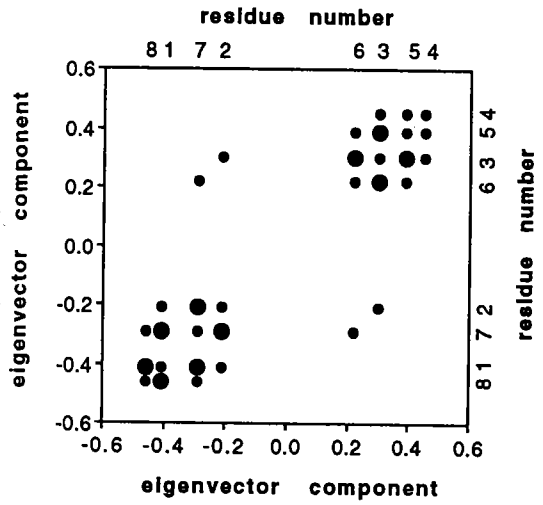
A.



B.



C.



matrix $\rho^2(\mathbf{R}^{1/2}\mathbf{x}) = (\mathbf{R}^{-1/2}\mathbf{A}\mathbf{R}^{-1/2})(\mathbf{R}^{-1/2}\mathbf{A}\mathbf{R}^{-1/2})^{\mathrm{T}}(\mathbf{R}^{1/2}\mathbf{x})$ which has a complete set of nonnegative eigenvectors. Iteration from a random start will converge to the eigenvector with the largest eigenvalue. Subsequent eigenvectors are extracted by applying Schmidt orthogonalization, i.e, subtracting the components of the previous eigenvectors from the trial vector, at each iteration step. The termination condition for iteration was that the norm of the trial vector changes by less than 0.00001 of the previous norm. The iteration usually takes 30–150 passes of the data and is very fast because the contact matrix is sparse. In the current problem, we only use the first non-trivial eigenvector.

Tree Decomposition

Sorting the residues in the contact matrix according to their component in the best nontrivial eigenvector reduces the search problem for N cuts in the linear sequence to a one-dimensional sweep along the reordered sequence. Residues before the cut point are assigned to one domain and residues after to the other domain. Figure 3 shows a worked example with the structure of actin.

The eigenvector profile plotted against the linear sequence can have sharp peaks, e.g., if the tip of a loop touches another domain across an interdomain interface (e.g., H73 in actin, Fig. 3B). To account for covalent bonds along the chain, we limit fragmentation by imposing a minimal segment length on the pieces assigned to one or the other domain. Cuts closer than 10 residues to a gap are disallowed. Gaps occur at the N- and C-terminus and where sequential C^{α} atoms have a distance larger than 5.0 Å. Short loops (<10 residues) arising from the bisection are assigned to the subdomain in which the loop starts and ends, processing the chain in the N-to-C direction.

Each trial bisection is evaluated according to Eq. (1). The bisection which gives the highest τ is remembered, and used. The bisection algorithm can be applied recursively on the subdomains until domain size reaches the lower limit (between 10 and 19 residues, see Table I). In order to identify autonomous folding units, we below define a set of termination

Fig. 2. Simple example of ordination. (**A**) Schematic structure containing residues numbered 1–8. Contacts are marked by continuous lines between sequential neighbours and broken lines for tertiary interactions. (**B**) Contact map using the discrete residue indices as axes. Small circles represent sequence neighbors, large circles represent tertiary contacts. (**C**) Contact map plotted using real-valued eigenvector components (−0.41, −0.21, 0.30, 0.45, 0.39, 0.22, −0.29, −0.46) to replace the residue indices used in **B**. Contacts between sequence neighbours (small circles) were given a weight of 1, and tertiary contacts (large circles) a weight of 2 in the reciprocal averaging procedure. One can see that the strongest contacts have moved next to the diagonal and two clusters of residues emerge (residues 8, 1, 7, 2 and 6, 3, 5, 4) with few contacts between the clusters. Mathematically, the eigenvector analysis amounts to maximizing the correlation of points in the scatterplot.

TABLE I. Tree Decomposition of Actin*

Unit	τ^2	γ	H	Size	Residues
1atnA	5.0	1.05		372	1–372
1atnA.1	2.4	0.88		183	1–146 336–372
1atnA.2	2.8	0.88		189	147–335
1atnA.1.1	1.9	0.54		145	1–32 71–146 336–372
1atnA.1.2	0.3	0.71		38	33–70
1atnA.1.1.1	1.4	0.38	1	119	1–32 97–146 336–372
1atnA.1.1.2	0.0	0.28		26	71–96
1atnA.1.1.1.1	1.0	0.33	1	80	1–32 97–109 133–146 336–356
1atnA.1.1.1.2	0.8	0.41		39	110–132 357–372
1atnA.1.1.1.1.1	0.3	0.33	3	40	1–32 97–104
1atnA.1.1.1.1.2	1.7	0.32		40	105–109 133–146 336–356
1atnA.1.1.1.1.1.1		0.18		18	1–18
1atnA.1.1.1.1.1.2	0.0	0.11		22	19–32 97–104
1atnA.1.1.1.1.2.1		0.26		19	105–109 133–146
1atnA.1.1.1.1.2.2	0.0	0.37		21	336–356
1atnA.1.1.1.2.1		0.38		18	110–127
1atnA.1.1.1.2.2	0.5	0.40		21	128–132 357–372
1atnA.1.1.1.2.2.1		0.11		11	128–132 357–362
1atnA.1.1.1.2.2.2		0.07		10	363–372
1atnA.1.2.1	0.0	0.74		20	33–52
1atnA.1.2.2		0.66		18	53–70
1atnA.2.1	1.0	0.51		97	147–179 272–335
1atnA.2.2	1.5	0.75		92	180–271
1atnA.2.1.1	0.6	0.33	2	62	147–179 272–300
1atnA.2.1.2	1.1	0.35		35	301–335
1atnA.2.1.1.1	2.4	0.30		46	147–179 272–284
1atnA.2.1.1.2		0.21		16	285–300
1atnA.2.1.1.1.1	0.4	0.26		33	147–179
1atnA.2.1.1.1.2		0.41		13	272–284
1atnA.2.1.1.1.1.1		0.10		19	147–165
1atnA.2.1.1.1.1.2		0.25		14	166–179
1atnA.2.1.2.1	0.0	0.32		21	301–321
1atnA.2.1.2.2		0.04		14	322–335
1atnA.2.2.1	1.5	0.59		47	180–215 239–249
1atnA.2.2.2	1.2	0.26		45	216–238 250–271
1atnA.2.2.1.1	1.4	0.47		36	180–215
1atnA.2.2.1.2		0.03		11	239–249
1atnA.2.2.1.1.1	0.0	0.37		22	180–201
1atnA.2.2.1.1.2		0.33		14	202–215
1atnA.2.2.2.1	0.9	0.15		30	216–238 250–256
1atnA.2.2.2.2		0.40		15	257–271
1atnA.2.2.2.1.1		0.18		18	216–233
1atnA.2.2.2.1.2		0.07		12	234–238 250–256

*For each unit, τ^2 [ps^2] for cutting in two, the globularity (γ), the number of β-sheet hydrogen bonds across the cut (H), the number of residues and the residue range are given.

criteria which are based on the ideas of weak interactions between the domains to be separated (large τ) and strong intradomain cohesion (compact shape) of each resulting domain after separation.

Filters for Autonomous Folding Units

Protein–protein contacts of the folded conformation are in competition with protein–solvent contacts in the unfolded conformation. Therefore, the compactness of the folded conformation is a measure of its stability (autonomy). We define globularity γ as the strength of long-sequence-range interatomic contacts per atom:

$$\gamma = \frac{1}{N}\sum_i \sum_{j<i-3} a_{ij} \qquad (4)$$

where N is the number of heavy atoms in the unit and a_{ij} is the contact strength between residues i and j. The first, second, and third sequential neighbors are excluded to enhance discrimination, as local contacts are likely to be preserved in an unfolded chain.

Five filters limit decomposition into structural domains. The filters are applied in hierarchical order, i.e., if the condition for applying a filter is true then the lower filters are not tested. (1) A lower limit on

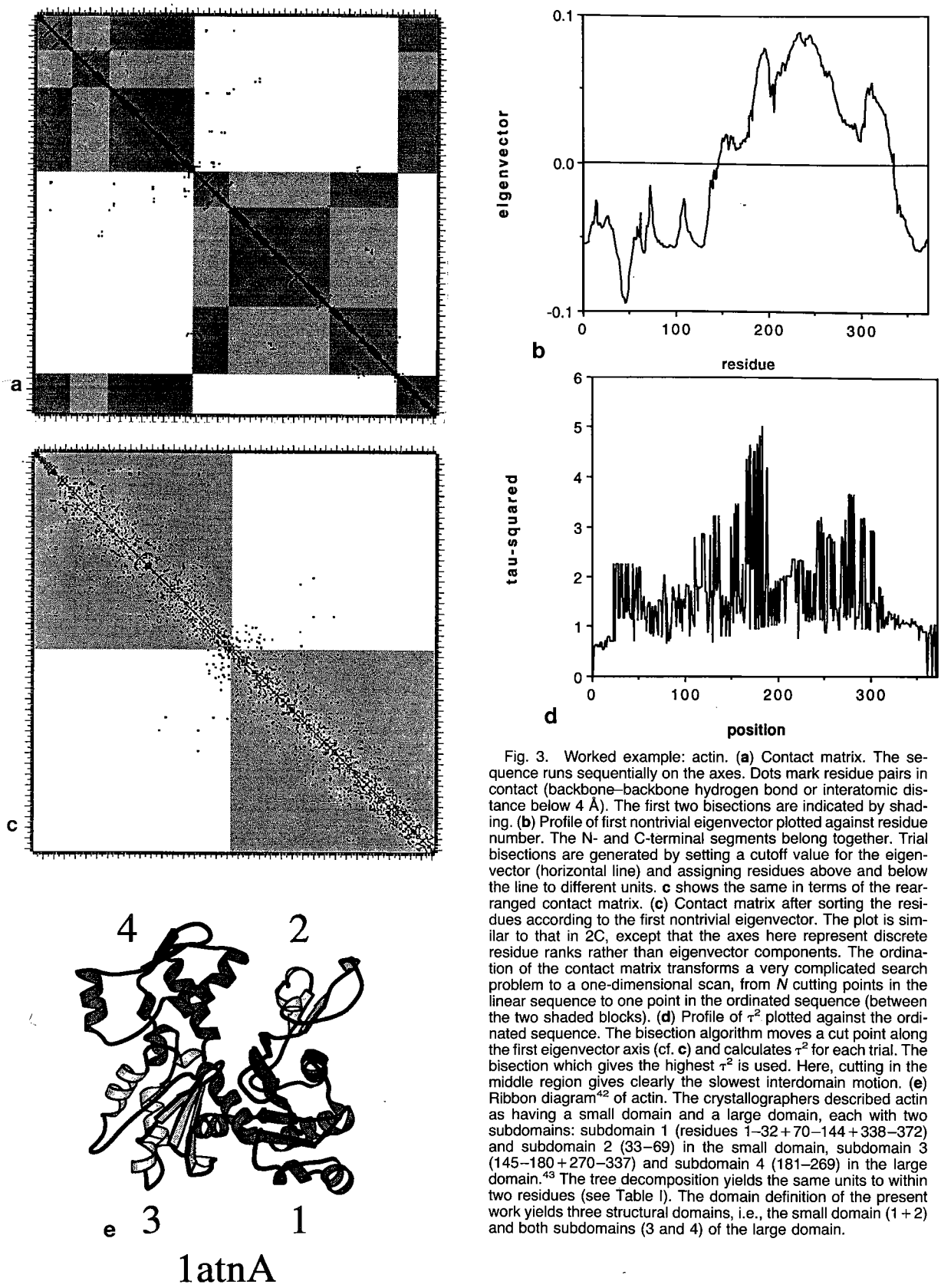


Fig. 3. Worked example: actin. (**a**) Contact matrix. The sequence runs sequentially on the axes. Dots mark residue pairs in contact (backbone–backbone hydrogen bond or interatomic distance below 4 Å). The first two bisections are indicated by shading. (**b**) Profile of first nontrivial eigenvector plotted against residue number. The N- and C-terminal segments belong together. Trial bisections are generated by setting a cutoff value for the eigenvector (horizontal line) and assigning residues above and below the line to different units. **c** shows the same in terms of the rearranged contact matrix. (**c**) Contact matrix after sorting the residues according to the first nontrivial eigenvector. The plot is similar to that in 2C, except that the axes here represent discrete residue ranks rather than eigenvector components. The ordination of the contact matrix transforms a very complicated search problem to a one-dimensional scan, from N cutting points in the linear sequence to one point in the ordinated sequence (between the two shaded blocks). (**d**) Profile of τ^2 plotted against the ordinated sequence. The bisection algorithm moves a cut point along the first eigenvector axis (cf. **c**) and calculates τ^2 for each trial. The bisection which gives the highest τ^2 is used. Here, cutting in the middle region gives clearly the slowest interdomain motion. (**e**) Ribbon diagram[42] of actin. The crystallographers described actin as having a small domain and a large domain, each with two subdomains: subdomain 1 (residues 1–32+70–144+338–372) and subdomain 2 (33–69) in the small domain, subdomain 3 (145–180+270–337) and subdomain 4 (181–269) in the large domain.[43] The tree decomposition yields the same units to within two residues (see Table I). The domain definition of the present work yields three structural domains, i.e., the small domain (1+2) and both subdomains (3 and 4) of the large domain.

domain size of 40 residues was imposed, as very few small units are appreciably globular. Thus, units smaller than 80 residues are never cut. (2) Highly flexible units ($\tau^2 > 2.6$ ps^2, see Fig. 4A) are always cut. (3) β-sheets, forming highly cooperative networks, are never cut. That is, no residue may be hydrogen bonded (backbone–backbone) to one residue in the same domain *and* another residue in the other domain. (4) The cut is accepted if both subdomains are compact ($\gamma > 0.80$, see Fig. 4B). (5) A cut which yields a small (<40 residues), nonglobular unit is accepted on condition that (recursive) application of the filters yields two domains for the larger unit.

Atomic Coordinates

The domain parser was applied to the representative set of proteins[25] as of August 1993, with a 30% sequence identity cutoff. Protein coordinates were retrieved from the Protein Data Bank.[26] If an entry contained only C^α atoms, backbone and side chain coordinates were constructed using the program MaxSprout.[27,28] Hydrogen atoms, crystal waters, ligands, and cofactors were ignored.

Computer Implementation

The algorithm was programmed in Fortran-77. The total execution time is practically linear with chain length. With unoptimized code, a protein of 200 residues is parsed in 10 s (RISC CPU) and the representative set was parsed in about 40 min. The parser for protein unfolding units, Puu (after the Finnish word for "tree"), is available on request for academic use.

RESULTS

Physical Definition of Structural Domains

Structural domains are defined using simple physical criteria involving interatomic contacts calculated from atomic coordinates. The definition for structural domains was applied to a set of 330 representative protein structures (Table II). Of these, 66 proteins were excluded from the decomposition because of their small size (≤80 residues). In the remaining set, 151 proteins contain a single structural domain and 113 multidomain proteins have a total of 286 structural domains when 55 short linker segments (≤40 residues) are ignored. Most structural domains are as globular as intact proteins (Fig. 4B,C). In contrast, below the structural domain level the putative folding units tend to become nonglobular in shape and larger interunit surfaces yield a smaller τ^2 (Fig. 4D). The size distribution of structural domains peaks around 100 residues and drops sharply after 200 residues (Fig. 4E). However, large proteins are not cut indiscriminately. Some of the largest domains (>400 residues) are found in the α/β hydrolase family (1ace, 1thg, 2had, 3sc2, 4tg1). The smallest structural domains permitted by the size threshold are the four times repeated lectin domains in wheat germ agglutinin (9wgaA).

With the present hierarchy of filters, τ^2 was the main criterion for selecting structural domains in the representative set: 118 bifurcations in the unfolding trees were accepted due to filter 2 (τ^2), compared to only 37 accepted bifurcations due to filters 4 (globularity) and 5 (short loops). Removing the β-sheet rule (filter 3) would yield 25 additional domains affecting 18 proteins, e.g. separating the unit formed by residues 80–140 in p21 ras [5p21, $\tau^2 = 1.4$ (ps)^2]. In one family of the α/β class, a bimodal mass distribution creates a weak point in the middle of a long sheet so that it is defined to consist of two structural domains by the τ^2 criterion [isocitrate dehydrogenase, 4icd in Figure 5, $\tau^2 = 4.1$ (ps)^2].

Proteins With Clearcut Domain Structure

Even though the method allows any number of cuts in the sequence, it is striking that 75% of structural domains identified in the present analysis consist of one continuous piece of chain (excluding short loops, Fig. 4F). Of the 113 multidomain proteins 41 have only continuous domains (e.g., 1bib in Fig. 5). Many noncontinuous domains are the result of N- or C-terminal arms reaching across to another domain, e.g., in the family of bidomain binding proteins where the chain passes three times across the domain interface [e.g., 3gbp in Fig. 5, $\tau^2 = 4.3$ (ps)^2]. Even complicated aggregates are readily untangled by the algorithm (e.g., the 1pya trimer in Fig. 5).

Some domain cuts have been verified by experiment. We give only two examples: the two structural domains of thermolysin are similar, except for two helices, to two autolytic fragments which can refold independently[29] (3tln in Fig. 5, cleaved loop marked by a cross); limited proteolysis and refolding experiments confirm the existence of two structural domains in phosphoglycerate kinase (3pgk[30]).

Proteins With Somewhat Ambiguous Domain Structure

There are two principal sources of perceived ambiguity in domain structure. First, our procedure uses sharp cutoffs in τ^2 and globularity without a sharp bimodal distribution with separated peaks on either side of the cutoffs. Second, automatic domain definitions are normally compared with visual parsing, which tends to be subjective. Unfortunately there is only scanty experimental evidence about autonomous folding units, so that it may be wise to accept the perceived ambiguities for the time being. Three examples of recurrent folding motifs in different structural contexts follow.

Parallel $(\beta\alpha)_8$ barrels, also called TIM barrels, are currently described in about twenty sequence-unrelated proteins. Many TIM barrel proteins have additional domains which makes the distribution of

TABLE II. Structural Domains in a Representative Set of Proteins*

```
1021  phage T4 lysozyme
   1 F  A  111     1 to   13
                  65 to  162
   2 F A+B  52    14 to   64
1aaiB ricin
   1 T  B  138     1 to  138
   2 T  B   49   139 to  187
   3 T  B   75   188 to  262
1aaj  apoamicyanin
   1 T  B  105     1 to  105
1aak  conjugating enzyme
   1 T A+B 150     1 to  150
1aba  glutaredoxin
   1 T A/B  87     1 to   87
*1abg  SO4 binding protein
   1 T A/B 151     1 to   94
                 220 to  276
   2 T A/B 158    95 to  219
                 277 to  309
*1abh  PO4-binding protein
   1 T A/B 155     1 to   77
                 230 to  239
                 254 to  321
   2 T A/B 166    78 to  229
                 240 to  253
1abk  endonuclease III
   1 F  A   96     1 to   20
                 136 to  211
   2 F  A   83    21 to  103
   a F  A   32   104 to  135
1abmA superoxide dismutase
   1 T  A   78     1 to   78
   2 T A/B 120    79 to  198
1ace  acetylcholinesterase
   1 T A/B 432     4 to  327
                 403 to  515
   2 T  A   94   328 to  402
                 516 to  534
*1ada  adenosine deaminase
   1 T  A  349     1 to  349
1ads  aldose reductase
   1 T  - 315     1 to  315
1aps  acylphosphatase
   1 T A/B  98     1 to   98
1arb  protease
   1 T A/B 263     1 to  263
1asoA ascorbate oxidase
   1 T  B  188   123 to  142
                 170 to  337
   2 T  B  104     1 to   64
                  83 to  122
   a F  -   27   143 to  169
   3 F  -   69    65 to   82
                 371 to  392
                 524 to  552
   4 T  B  164   338 to  370
                 393 to  523
1atnA actin
   1 T A/B 183     1 to  146
                 336 to  372
   2 F A/B  97   147 to  179
                 272 to  335
   3 F  A   92   180 to  271
1avhA annexin V
   1 F  A  100   146 to  245
   2 F  A  162     3 to   89
                 246 to  320
   3 F  A   56    90 to  145
1ayh  glucanohydrolase
   1 T  B  214     1 to  214
1baa  barley endochitinase
   1 T  A  183     1 to   86
                 147 to  243
   2 F  -   60    87 to  146
1bbhA cytochrome c'
   1 F  A   71     1 to   71
   2 F  A   60    72 to  131
1bbkA dehydrogenase
   1 T  B  213    19 to   64
                 207 to  373
   2 T  B   55    65 to  119
   3 T  B   45   120 to  164
   4 T  B   42   165 to  206
1bbpA binding protein
   1 T  B  173     2 to  178
1bbt1 virus coat
   a F  -   16   193 to  208
   b T  -   36     1 to   36
   1 T  B  134    37 to  192
1bbt2 virus coat
   1 T  B  210     9 to  218
1bbt3 virus coat
   a T  -   39     1 to   39
   1 T  B  181    40 to  220
1bib  biotin repressor
   1 T  A   65     2 to   66
   2 T A/B 182    67 to  270
   3 T  B   47   271 to  317

1bmv1 virus coat
   1 T  B  185 1001 to 1185
1bmv2 virus coat
   1 T  B  194 3181 to 2192
   a F  -   18 3001 to 3018
   2 T  B  162 3019 to 3180
*1bn21 bovine neurophysin
   1 T  A   86     2 to   87
1bop  DNA-binding domain
   1 T A/B  85   326 to  410
1brd  bacteriorhodopsin
   1 F  A  170     8 to  225
1btc  beta-amylase
   1 T  A  491     5 to  495
1bw4  barwin
   1 T A/B 125     1 to  125
1c2rA cytochrome c2
   1 F  -  116     1 to  116
1caj  carbonic anhydrase
   1 T A/B 258     3 to  261
1cas  parvovirus capsid
   1 F  B  352    37 to   81
                 105 to  210
                 243 to  278
                 359 to  360
                 372 to  406
                 457 to  584
   2 F  B   71    82 to  104
                 211 to  242
                 343 to  358
   3 F  -   56   279 to  334
   4 F  -   69   335 to  342
                 361 to  371
                 407 to  456
*1cbp  cucumber protein
   1 T  B   86     1 to   86
1cbx  carboxypeptidase A
   1 T A/B 307     1 to  307
1ccr  cytochrome c
   1 T  A  111     1 to  111
1cd8  CD8
   1 T  B  114     1 to  114
1cid  CD4
   1 T  B  106     1 to  106
   2 T  B   71   107 to  177
1clm  calmodulin
   1 F  A   85     4 to   88
   2 F  A   59    89 to  147
1cmbA met aporepressor
   1 T  A  104     1 to  104
1colA colicin A
   1 T  A  197     5 to  201
1cox  cholesterol oxidase
   1 T A/B 502     5 to  506
1cpcA C-phycocyanin
   a F  A   32     1 to   32
   1 F  A  130    33 to  174
1cpcL C-phycocyanin
   a F  A   32     1 to   32
   1 F  A  140    33 to  174
1cpl  cyclophilin
   1 T  B  165     1 to  165
1dhr  reductase
   1 T A/B 236     5 to  240
*1dpi  DNA polymerase I
   1 T  -  201     1 to  201
   2 F A/B 161   202 to  222
                 274 to  333
                 467 to  546
   3 F  A   51   223 to  273
   4 F  A   81   334 to  414
   5 F  -   52   415 to  466
1dri  binding protein
   1 T A/B 123     1 to  101
                 240 to  261
   2 T A/B 148   102 to  239
                 262 to  271
1eaf  transacetylase
   a F  A   37   395 to  431
   1 T A/B 206   432 to  637
1eco  hemoglobin
   1 F  A  136     1 to  136
1egr  glutaredoxin
   1 T A/B  85     1 to   85
1end  T4 endonuclease V
   1 T  A  137     2 to  138
*1eps  synthase
   1 T A/B 206     1 to   19
                 241 to  427
   2 F A/B  74    84 to  157
   3 F A/B  75    20 to   83
                 230 to  240
   4 T A/B  72   158 to  229
1etu  elongation factor Tu
   1 T A/B 177     5 to  200

1ezm  elastase
   1 F  B  117     1 to   80
                  97 to  133
   2 T  A  181    81 to   96
                 134 to  298
1fbaA aldolase
   1 T  -  360     2 to  364
1fc1A Fc fragment
   1 T  B  100   238 to  337
   2 T  B  106   338 to  443
1fdd  ferredoxin
   1 T  -  106     1 to  106
1fha  ferritin
   1 F  A  170     6 to  183
1fnr  oxidoreductase
   1 T  B  134    19 to  152
   2 T A/B 162   153 to  314
1fxiA ferredoxin
   1 T  B   96     1 to   96
1gky  guanylate kinase
   1 F A/B 138     1 to   33
                  82 to  186
   2 T A/B  48    34 to   81
1glaG glycerol kinase
   1 T A/B 251     4 to  245
                 439 to  454
   2 T A/B 238   246 to  438
                 455 to  499
1gly  glucoamylase
   1 T  A  470     1 to  471
1gmfA growth factor
   1 T  A  119     5 to  123
1gmpA ribonuclease
   1 T  -   96     1 to   96
1gox  glycolate oxidase
   1 T  A  350     1 to  359
1gp1A peroxidase
   1 T A/B 136    10 to  111
                 159 to  193
   2 T  -   47   112 to  158
1gpb  phosphorylase
   1 F  A   66    19 to   62
                 104 to  125
   2 F A/B 424    63 to  103
                 126 to  165
                 179 to  491
                 812 to  841
   3 F  A  256   166 to  178
                 558 to  649
                 661 to  811
   4 F  A   77   492 to  557
                 650 to  660
1gpr  glucose permease
   1 T  B  158     4 to  161
1grcA transformylase
   1 T A/B 105     1 to  105
   2 T A+B  65   106 to  186
   a T  B   23   187 to  209
1grdA DNA-binding domain
   1 F  -   81    34 to  114
*1gsgP Gln-tRNA synthetase
   1 T A/B 159    95 to  253
   a F  A   17     1 to   17
   2 T  A   77    18 to   94
   3 F  -   90   254 to  320
                 452 to  474
   4 T  B  183   321 to  451
                 475 to  526
1gstA transferase
   1 F  A  217     1 to  217
1hc6  haemocyanin
   1 F  A  156     5 to  135
                 533 to  559
   2 F A+B  55   136 to  168
                 430 to  451
   3 F A/B 267   169 to  240
                 397 to  429
                 452 to  532
                 560 to  653
   4 T  -  156   241 to  396
1hgeA hemagglutinin
   1 T  B   48     1 to   38
                 319 to  328
   2 F  B   65    39 to   56
                 272 to  318
   3 T  B  215    57 to  271
1hgeB hemagglutinin
   1 F A/B 118     1 to   50
                 108 to  175
   2 F  A   57    51 to  107
*1higA interferon-gamma
   1 F  A   94     1 to   94
   a F  A   29    95 to  123
1hilA IgG2a Fab fragment
   1 T  B  114     1 to  108
   2 T  B  103   109 to  211

1hsbA HLA
   1 T A/B 180     1 to  180
   2 T  B   90   181 to  270
*1hsc  heat shock protein
   1 F A/B 125     1 to   38
                 115 to  180
                 362 to  382
   2 T  -   76    39 to  114
   3 T A+B  99   181 to  226
                 309 to  361
   4 T A/B  82   227 to  308
*1hsdA dehydrogenase
   a F  -   14   242 to  255
   1 T  -  198     1 to   36
                  80 to  241
   2 T  -   43    37 to   79
*1ifa  interferon-beta
   1 T  A  158     1 to  158
1ifc  binding protein
   1 T  B  110     1 to   12
                  34 to  131
   a T  A   21    13 to   33
1ipd  dehydrogenase
   1 T A/B 196     1 to  103
                 253 to  345
   2 T A/B 149   104 to  252
1lap  Leu aminopeptidase
   1 T A/B 159     1 to  162
   2 T A/B 322   163 to  484
1lig  binding domain
   1 T  A  149    25 to  180
1lmbA lambda repressor
   1 F  A   87     6 to   92
1lpe  apolipoprotein-E3
   1 F  A  144    23 to  166
1lte  lectin
   1 T  B  239     1 to  239
1ltsA enterotoxin (LT)
   1 T A/B 185     4 to  188
1ltsD enterotoxin (LT)
   a F  A   12     1 to   12
   1 T  B   72    13 to   51
                  71 to  103
   b T  A   19    52 to   70
1lz3  lysozyme
   1 T  A  129     1 to  129
1mamH Fab fragment
   1 T  B  122     1 to  122
   2 T  B   95   123 to  217
1mba  myoglobin
   1 F  A  146     1 to  146
1mbd  myoglobin
   1 T  A  153     1 to  153
1mdc  binding protein
   1 T  B  106     1 to   40
                  67 to  131
   a T  B   25    41 to   66
1minA nitrogenase
   1 T  A  127    62 to  188
   a F  A   22     5 to   26
   2 T A/B 158    27 to   61
                 189 to  191
                 342 to  447
                 459 to  481
   3 T  A  161   192 to  341
                 448 to  458
1minB nitrogenase
   1 T  A   42     2 to   43
   2 F  A  226    70 to  230
                 250 to  314
   3 F A/B 140    44 to   69
                 363 to  476
   4 F  A  114   231 to  249
                 315 to  362
                 477 to  523
*1mli  isomerase
   1 T A/B  85     1 to   85
   a F  -   11    86 to   96
*1mon  monellin
   1 T  -  188     1 to  189
   2 T  A   94   190 to  283
   3 T  A   94   284 to  377
1mrm  mandelate racemase
   1 T A/B 145     3 to  129
                 320 to  329
                 340 to  359
   2 T A/B 200   130 to  319
                 330 to  339
1ms2A MS2 virus
   a F  A   37    93 to  129
   1 T  B   75     1 to   65
                  83 to   92
   b T  B   17    66 to   82
1mup  urinary protein
   1 T  B  157     5 to  161
1nipB nitrogenase
   1 T  A  256     1 to  256
   a F  -   31   257 to  287

*1nrcA riboprotein U1-SNRP
   1 T A/B  85     1 to   85
*1nrd  nitrite reductase
   a F  -   17   317 to  333
   1 T  -  154     1 to  154
   2 T  B  162   155 to  316
1nsbA sialidase
   1 T  B  390    76 to  465
1ofv  flavodoxin
   1 T A/B 169     1 to  169
1omf  matrix porin (ompF)
   1 T  B  340     1 to  340
1omp  binding protein
   1 T A/B 161     1 to  110
                 261 to  311
   2 F A/B 209   111 to  260
                 312 to  370
1ovaC ovalbumin
   1 T A/B 204    24 to   55
                  99 to  114
                 188 to  295
                 336 to  391
   2 T A/B 169    56 to   98
                 115 to  187
                 296 to  335
1ovb  ovotransferrin
   1 T A/B 159    94 to  248
1pafA antiviral protein
   1 T A/B 262     1 to  262
1pba  procarboxypeptidase
   1 T  -   81     1 to   81
1pbxA hemoglobin
   1 F  A  142     1 to  142
*1pcdA dioxygenase
   1 T  -  201     1 to  201
1pda  pdeaminase
   1 T A/B 109     3 to   99
                 200 to  220
   2 T A/B 100   100 to  199
   3 T A/B  87   221 to  307
*1pec  pectate lyase
   1 T  -  324     1 to   23
                  52 to  352
   a T  -   28    24 to   51
1pfkA phosphofructokinase
   1 T A/B 193     0 to  141
                 253 to  303
   2 T A/B 111   142 to  252
   a F  A   16   304 to  319
1pgd  dehydrogenase
   1 T A/B 183     1 to  183
   2 F  A  249   184 to  432
   a F  -   37   433 to  469
1phg  cytochrome P450CAM
   1 T  A  405    10 to  414
1phh  hydroxylase
   1 T A/B 244     1 to   72
                  96 to  180
                 269 to  343
                 383 to  394
   2 F  B  111    73 to   95
                 181 to  268
   a F  A   39   344 to  382
*1phs  phaseolin
   1 T A/B 189     1 to   11
                  27 to   42
                 203 to  364
   2 F  B  142    12 to   26
                  43 to  144
                 178 to  202
   a F  A   33   145 to  177
*1phy  photoactive protein
   1 T  -  123     1 to  123
1pii  isomerase:synthase
   1 T A/B 255     1 to  255
   2 T A/B 197   256 to  452
1plc  plastocyanin
   1 T  B   99     1 to   99
1ppfE elastase
   1 T  B  218    16 to  243
1ppl  penicillopepsin
   1 T  B  212     1 to  192
                 304 to  323
   2 T  B  111   193 to  303
1ppn  papain
   1 T  B  120     1 to   18
                 111 to  212
   2 T  A   92    19 to  110
1prcC reaction center
   1 F  - 196     1 to   23
                 143 to  315
   2 F  A  136    24 to  142
                 316 to  332
1prcM reaction center
   1 F A/B  51     1 to   51
   2 F  A  140    52 to  191
   3 F  A  132   192 to  323
```

(Continued on overleaf)

TABLE II. Structural Domains in a Representative Set of Proteins* (Continued)

*1pte carboxypeptidase
1 F - 197 1 to 23
77 to 123
135 to 261
2 F - 56 124 to 134
262 to 271
314 to 348
3 F - 72 24 to 76
272 to 290
a T - 23 291 to 313
1pyaA His decarboxylase
1 F A/B 41 1 to 41
a F A/B 40 42 to 81
1pyaB His decarboxylase
1 T A/B 228 83 to 310
1pya* His decarboxylase
1 T A/B 309 A1 to B310
2 T A/B 307 C1 to D308
3 T A/B 311 D309 to F310
1pyp pyrophosphatase
1 T - 280 1 to 280
1rla2 rhinovirus coat
1 T B 253 11 to 263
*1rleE ECO RI endonuclease
1 T A/B 198 1 to 97
126 to 153
189 to 261
2 F A/B 63 98 to 125
154 to 188
1rbp binding protein
1 T B 174 1 to 174
1rcb interleukin-4
1 T A 129 1 to 129
*1rea recA protein
a T A 31 1 to 31
1 T A/B 213 32 to 244
2 F A/B 60 245 to 304
1rhd rhodanese
1 T - 156 1 to 156
2 T - 137 157 to 293
1rnbA barnase
1 T - 79 2 to 21
52 to 110
a T A 30 22 to 51
1rnd ribonuclease A
1 T A+B 73 1 to 49
80 to 103
2 T A/B 51 50 to 79
104 to 124
1rveA ECO RV endonuclease
1 T A/B 49 2 to 33
145 to 161
2 F A/B 107 34 to 101
117 to 144
162 to 172
3 T A/B 88 102 to 116
173 to 245
1s01 subtilisin BPN'
1 T A/B 275 1 to 275
1sas Ca-binding protein
1 F A 185 1 to 185
1sdhA hemoglobin
1 F A 146 1 to 146
1sgt trypsin
1 T B 139 16 to 28
69 to 80
121 to 234
2 T B 73 29 to 68
81 to 120
a F A 11 235 to 245
1shaA SH2 domain
1 T A/B 103 2 to 104
1snc nuclease
1 T A/B 135 7 to 141
1spa Asp aminotransferase
1 T A/B 222 71 to 299
a F - 28 5 to 32
2 F A 102 33 to 48
323 to 409
3 F A 44 49 to 70
300 to 322
1ten fibronectin type III
a T B 28 803 to 830
1 T B 61 831 to 891
1tfg growth factor
1 T A/B 112 1 to 112
1thg lipase
1 T A/B 544 1 to 544
1tho thioredoxin
1 T A/B 109 1 to 108
1tie trypsin inhibitor
1 T B 166 1 to 170
1tlk telokin
1 T B 103 33 to 135

*1tmd dehydrogenase
1 T A/B 381 1 to 381
2 F A 169 382 to 489
645 to 705
3 T A/B 155 490 to 644
a F - 24 706 to 729
1tnfA necrosis factor
1 T B 152 6 to 157
*1tpt phosphorylase
1 T A 100 1 to 69
156 to 186
2 F A 244 70 to 155
187 to 334
431 to 440
3 T A/B 96 335 to 430
1trb thioredoxin reductase
1 T A/B 125 118 to 242
2 T A/B 155 1 to 41
77 to 117
243 to 315
a T A 35 42 to 76
1troA Trp repressor
a F A 37 5 to 41
1 F A 67 42 to 108
1ttbA transthyretin
1 T B 127 1 to 127
1ula phosphorylase
1 T A/B 289 1 to 289
1vaaB MHC class I
1 T B 99 1 to 99
1vsgA glycoprotein
1 T A/B 205 1 to 34
85 to 255
2 F A 157 35 to 84
256 to 362
1wsyA tryptophan synthase
1 T A 248 1 to 265
1wsyB tryptophan synthase
1 T A/B 294 9 to 96
188 to 393
2 F A/B 91 97 to 187
1yat binding protein
1 T B 113 -5 to 107
256bA cytochrome b562
1 F A 106 1 to 106
2aaa acid alpha-amylase
1 T A 374 1 to 374
2 T B 102 375 to 476
*2at2C transcarbamoylase
1 T A 169 1 to 144
271 to 295
2 T - 126 145 to 270
2aviA avidin
1 T B 121 3 to 123
2azaA azurin
1 T A/B 129 1 to 129
2bpa1 phi-174 capsid
1 T A/B 320 1 to 166
214 to 297
357 to 426
2 F A/B 73 167 to 180
298 to 356
a F A 33 181 to 213
2bpa2 phi-174 capsid
1 T B 175 1 to 175
2ccyA cytochrome c'
1 F A 127 2 to 128
2cdv cytochrome c3
1 F A/B 44 1 to 44
2 F - 63 45 to 107
2cmd malate dehydrogenase
1 T A/B 312 1 to 312
2cts citrate synthase
a F A 17 421 to 437
b F A 31 19 to 49
1 F A 292 1 to 18
50 to 280
378 to 420
2 F A 97 281 to 377
2cyp cyt c peroxidase
1 T A 172 2 to 144
266 to 294
2 T A 121 145 to 265
2dnjA deoxyribonuclease I
1 T A/B 118 1 to 88
231 to 260
2 T A/B 135 89 to 230
2glsA glutamine synthetase
a F A 11 458 to 468
1 T - 209 113 to 130
267 to 457
2 T A/B 112 1 to 112
3 F A/B 136 131 to 266
2had dehalogenase
1 T A 310 1 to 310
2hhmA monophosphatase
1 T A+B 142 5 to 150
2 T A 130 151 to 272

*2hhrC growth hormone
1 T B 91 1 to 91
2 T B 104 92 to 195
*2hvp HIV-1 protease
1 T - 94 1 to 94
*2ila interleukin-1alpha
1 T B 145 1 to 145
2lbp Leu-binding protein
1 T A/B 151 120 to 251
328 to 346
a F A 27 252 to 278
2 T A/B 121 1 to 119
279 to 280
3 T A/B 47 281 to 327
2lh7 leghemoglobin
1 F A 153 1 to 153
2madL dehydrogenase
1 T - 124 7 to 130
2mcm macromomycin
1 T B 112 1 to 112
2mev1 virus
a F - 31 1 to 31
1 T B 218 32 to 249
b F - 19 250 to 268
2mhr myohemerythrin
1 T A 118 1 to 118
2msbA lectin domain
1 T A/B 111 110 to 220
2pf2 prothrombin
1 T A 62 1 to 62
2 T - 83 63 to 145
2pia dioxygenase reductase
1 T - 96 226 to 321
2 T B 104 1 to 104
3 T A/B 121 105 to 225
2plv1 virus coat
1 T - 61 6 to 75
a F - 19 284 to 302
2 F B 138 76 to 115
131 to 197
235 to 265
3 F - 70 116 to 130
198 to 234
266 to 283
2plv3 virus coat
1 T - 45 1 to 45
2 T B 190 46 to 235
2pmgA phosphoglucomutase
1 T A/B 188 1 to 188
2 T B 141 421 to 561
3 T - 115 189 to 303
4 F A 117 304 to 420
2por porin
1 T B 301 1 to 301
2ren renin
1 T B 108 202 to 318
2 F B 85 4 to 21
152 to 201
319 to 340
3 T B 127 22 to 151
2rn2 ribonuclease H
1 T A/B 155 1 to 155
2rspB virus protease
1 T B 113 1 to 124
2scpA Ca binding protein
1 T A 174 1 to 174
2sga proteinase A
1 T B 181 16 to 242
2sicI subtilisin inhibitor
1 T B 107 7 to 113
2snv virus capsid protein
1 T B 65 114 to 178
2 T B 86 179 to 264
2stv virus coat
a F A 15 12 to 26
1 T B 169 27 to 195
2tbvA virus coat
1 T B 169 102 to 270
2 T B 117 271 to 306
2tmvP tobacco mosaic virus
1 T A 154 1 to 154
2yhx hexokinase
1 T A/B 160 56 to 187
431 to 458
2 F A 96 2 to 18
284 to 362
3 F A/B 201 19 to 55
188 to 283
363 to 430
3adk adenylate kinase
1 F A/B 144 1 to 37
88 to 194
2 T A 50 38 to 87
3b5c cytochrome b5
1 T A+B 85 3 to 87
*3cbh cellobiohydrolase II
1 T - 365 1 to 365

3cd4A CD4
1 T B 98 1 to 98
2 T B 80 99 to 178
3chy cheY
1 T A/B 128 2 to 129
3cla acetyltransferase
1 T A/B 213 6 to 219
3dfr reductase
1 T A/B 162 1 to 162
*3dpa papD
1 T B 132 1 to 132
2 T B 86 133 to 218
3gapA activator protein
1 T A/B 137 1 to 137
2 F - 71 138 to 208
3gbp binding protein
1 T A/B 139 3 to 108
259 to 291
2 T A/B 166 109 to 258
292 to 307
3grs glutathione reductas
1 F A/B 173 18 to 60
105 to 162
290 to 361
a F A 40 61 to 88
404 to 415
2 F A/B 105 362 to 403
416 to 478
b F A 16 89 to 104
3 T A/B 127 163 to 289
3inkC interleukin-2
1 T A 121 06 to 133
3pgk kinase
1 T - 199 1 to 188
405 to 415
2 T A 216 189 to 404
3rubL RUBISCO
1 T A/B 153 22 to 148
301 to 316
353 to 367
2 F A 288 149 to 300
317 to 352
368 to 467
3rubS RUBISCO
a T - 38 1 to 38
1 T A/B 85 39 to 123
3sc2 carboxypeptidase
1 T A/B 406 -4A to 422
3sod0 superoxide dismutase
1 T B 151 1 to 151
3tgl acylhydrolase
1 T A/B 265 5 to 269
3tln thermolysin
1 T A/B 135 1 to 135
2 T A 181 136 to 316
451c cytochrome c551
1 F A 82 1 to 82
4blmA beta-lactamase
1 T A/B 211 31 to 86
132 to 291
2 T A 45 87 to 131
4bp2 prophospholipase A2
1 T A 117 1 to 123
4dfrA reductase
1 T A/B 159 1 to 159
4enl enolase
1 T A+B 126 1 to 126
2 T A/B 310 127 to 436
4fgf growth factor
1 T B 124 20 to 143
4fxn flavodoxin
1 T A/B 138 1 to 138
4gcr gamma-crystallin
1 T B 82 1 to 82
2 T B 92 83 to 174
4gpd1 dehydrogenase
1 T - 167 1 to 146
313 to 333
2 T A/B 166 147 to 312
4icd dehydrogenase
a F - 37 162 to 198
1 T A/B 220 3 to 124
319 to 416
2 F A/B 157 125 to 161
199 to 318
4rcrH reaction center
1 T - 73 12 to 84
2 F - 41 85 to 115
239 to 248
3 T - 123 116 to 238
4sbvA virus coat protein
1 T A/B 199 62 to 260
4tms thymidylate synthase
1 F A/B 223 1 to 52
146 to 316
2 F A 93 53 to 145

4ts1A Tyr-tRNA synthetase
1 T A/B 219 1 to 221
2 F A 98 222 to 319
5fbpA bisphosphatase
1 T A/B 209 6 to 201
274 to 291
302 to 313
2 F A/B 104 202 to 273
292 to 301
314 to 335
5nn9 neuraminidase
1 T B 222 82 to 102
268 to 468
2 T B 76 103 to 177
3 T B 90 178 to 267
5p21 ras p21 protein
1 T A/B 166 1 to 166
7timA isomerase
1 T A/B 247 2 to 248
7xia xylose isomerase
1 T A 315 1 to 315
2 F A 72 316 to 387
8abp binding protein
1 T A/B 140 2 to 109
254 to 285
2 T A/B 165 110 to 253
286 to 306
8acn aconitase
1 T A/B 224 531 to 754
2 T A/B 306 2 to 101
122 to 211
230 to 317
503 to 530
a T A 20 102 to 121
3 T A/B 203 212 to 229
318 to 502
8adh dehydrogenase
1 T A/B 234 1 to 177
318 to 374
2 T A/B 140 178 to 317
8atcA transferase
1 T A/B 147 1 to 129
293 to 310
2 T - 163 130 to 292
8atcB transferase
1 T A/B 90 8 to 97
2 T B 56 98 to 153
8catA catalase
1 F - 65 3 to 67
a F - 39 381 to 419
2 F A/B 269 68 to 152
200 to 380
420 to 422
3 F A 111 153 to 199
437 to 500
b F - 14 423 to 436
8ilb interleukin 1-beta
1 T B 146 5 to 151
9ldtA dehydrogenase
a F - 20 1 to 22
1 T A/B 311 23 to 331
9rnt ribonuclease T1
1 T A/B 104 1 to 104
9rubB RUBISCO
1 T A/B 192 2 to 139
289 to 316
337 to 362
a F - 11 326 to 336
2 T A/B 176 140 to 288
317 to 325
363 to 370
384 to 393
3 F A 79 371 to 383
394 to 459
9wgaA agglutinin
1 T - 43 1 to 43
2 T A/B 44 44 to 87
3 T A/B 41 88 to 128
4 T A/B 43 129 to 171

*The table is based on the PDB_select.aug_1993 list of representative protein structures.[25] Chains shorter than 80 residues were not parsed and are excluded from the table. For each protein, the first line lists the Protein Data Bank (PDB) code, chain identifier, and protein name. C^{α}-only entries are marked by an asterisk before the PDB code. The subsequent lines list for each structural domain its number (1,2,3, . . . ; letters a,b,c,.. for nonglobular short linkers), whether or not the unit is compact (true or false), the structural class, the number of residues in the domain, and sequence ranges. Structural classes are defined on the basis of secondary structure[18] content: class A has >40% of the residues in helix and <15% of the residues in β-strands; class B has <15% helix and >30% strand; class A + B has either the N- or C-terminal half classifying as A *and* the other half classifying as B; class A/B has >15% helix and >15% strand; otherwise the class is "—."

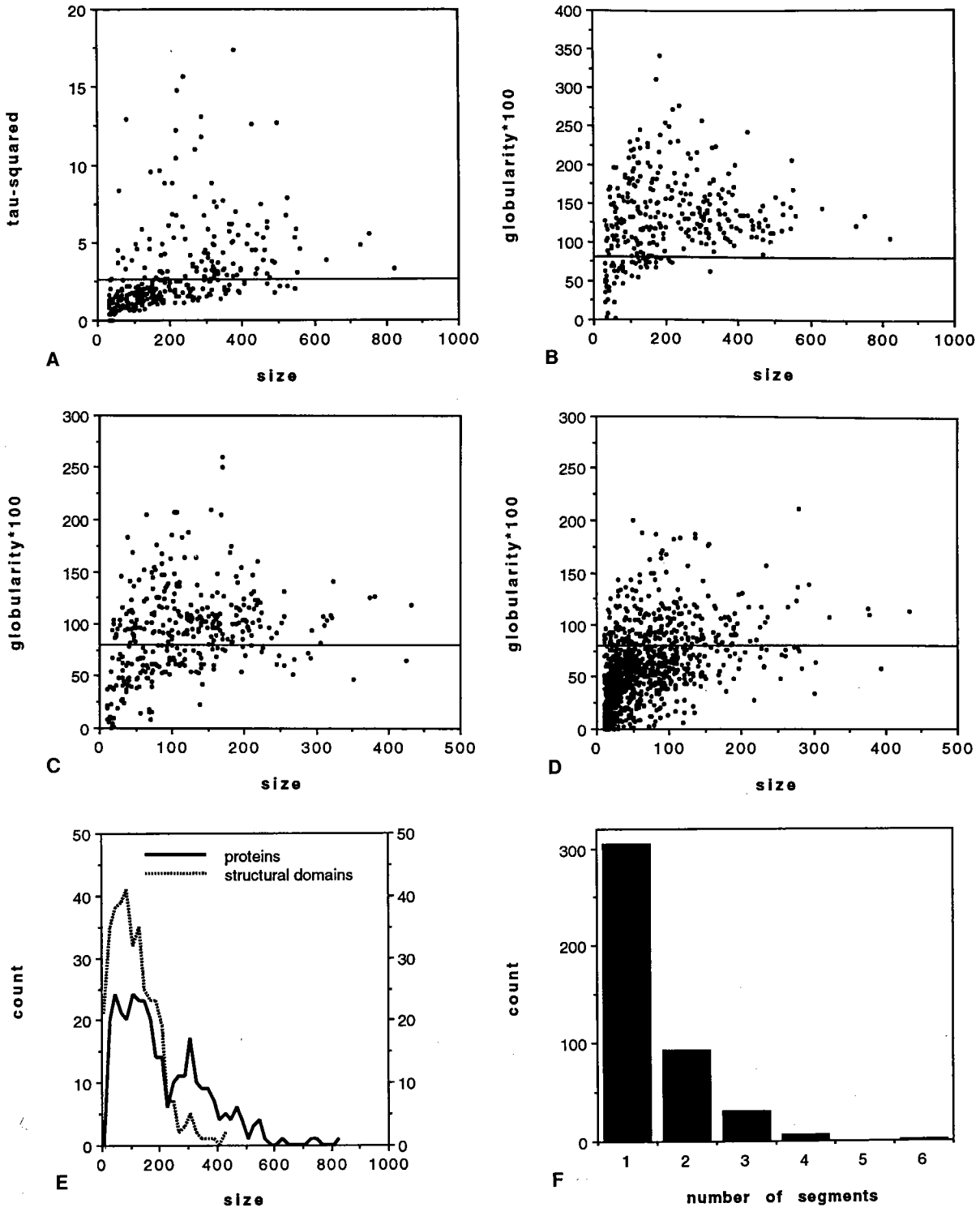


Fig. 4. Physical characteristics of intact proteins, structural domains, and sub-domains. (**a**) τ^2 and (**b**) globularity are plotted against the number of residues for 330 protein chains in the representative set. Cutoffs used by the structural domain definition are shown by horizontal lines. The τ^2 values have a rising trend with larger proteins, as larger mass corresponds to slower frequencies. There is a relatively sharp lower limit of globularity for intact proteins (excluding very short chains). (**c**) Globularity is plotted against the number of residues in structural domains of multidomain proteins identified by the present method. Single-domain proteins (no cuts) are excluded. Structural domains are nearly as compact as intact proteins (plot b). (**d**) Globularity is plotted against the number of residues in folding units one level lower than the structural domains identified by the present method. There is a notable increase in nonglobular units compared to **b** or **c**. The globular subdomains are paired with nonglobular ones, because the criteria limiting decomposition are for pairs of subdomains. (**e**) Size distributions of intact proteins and structural domains. The hump around 300 residues in intact proteins is reduced significantly in structural domains which are mostly smaller than 200 residues. Short proteins (<80 residues) were excluded from the domain decomposition. (**f**) Histogram of the number of segments in structural domains, excluding short linkers (units smaller than 40 residues) from the statistics. Continuous domains (1 segment) include the 151 single-domain proteins and 153 domains from multidomain proteins. The tail falls off with about a factor of three between bins.

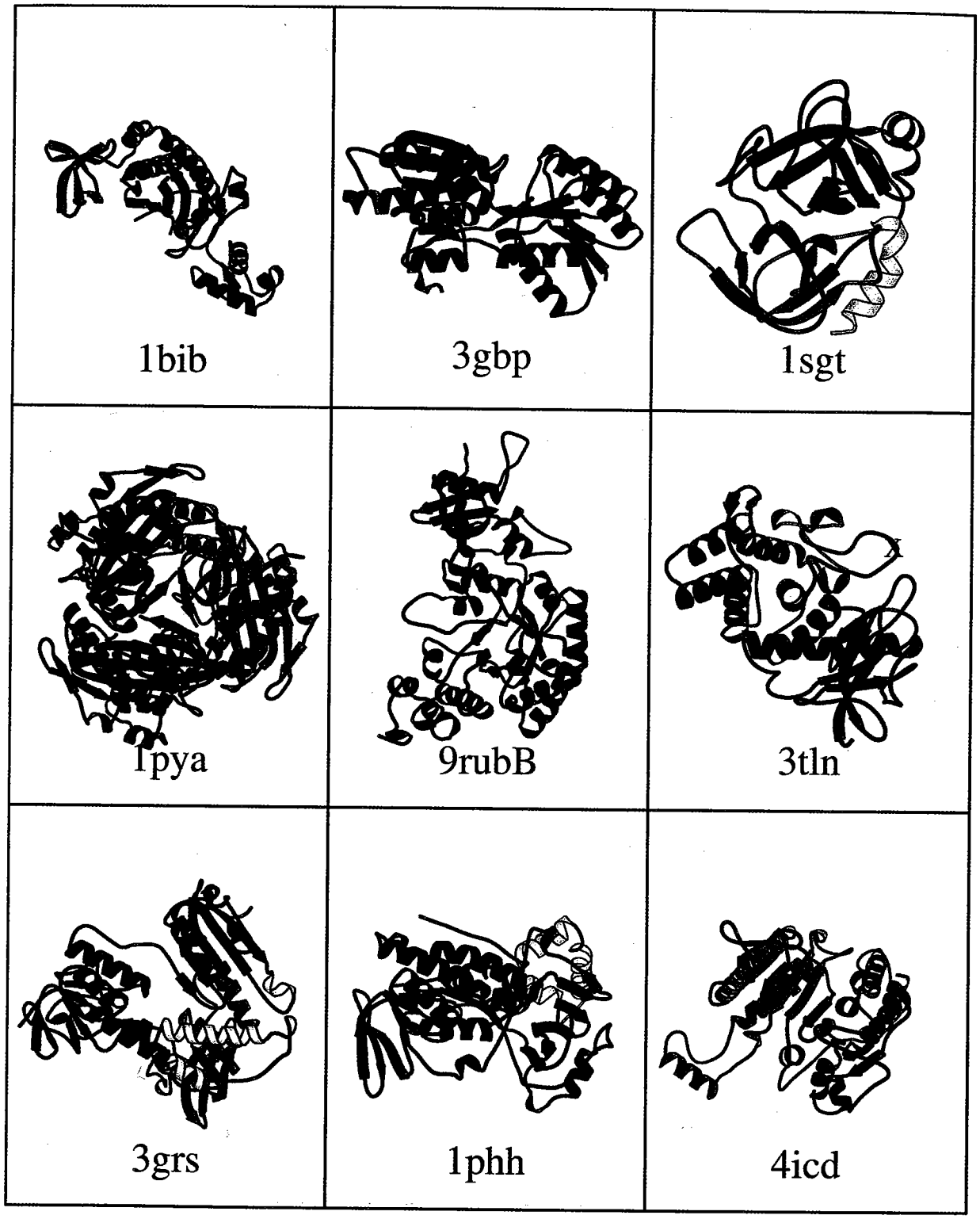


Fig. 5. Examples of structural domains. Structural domains identified by the present method are shown[42] in different color. Short linkers have lighter hue. The PDB code is given below each structure. Consult Table II for protein names and residue ranges.

mass around the barrel quite non-uniform. In Rubisco, the N-terminal domain has a very tight interface with the barrel domain. Two helices, which by analogy with other TIM barrels "belong" to the barrel domain, are in this case assigned to the N-terminal domain and another helix pair to the C-terminal domain [9rubB in Figure 5, τ^2 values for the cuts 5.3 and 4.7 $(ps)^2$].

A common FAD-binding domain is one of three well separated structural domains in glutathione reductase (red domain of 3grs in Fig. 5). The domains are identified in agreement with the original description (FAD-domain: 1–139+273–344; NADP-domain: 139–272; interface domain: 345–461; linker: 45–104).[31] The main difference in the present automatic parsing is that the long helices of the linker region are split between the three major domains (see Table II). The two first domains of glutathione reductase reappear at the rear of a TIM barrel domain in thymidylate synthase (1tmd domains 2 and 3, see Table II). The FAD-binding domain is also found in parahydroxybenzoate hydroxylase (red domain of 1phh in Fig. 5). The crystallographers saw three domains (domain I: 1–64+84–172; domain II: 65–83+173–261; domain III: 289–391).[32] The present method essentially identifies domains I and II while half of domain III is assigned to domain I (see Table II). Cholesterol oxidase (1cox) has a similar topography as 1phh. The cores of two domains seen by the crystallographers, i.e., residues 5–44+226–316+426–506 in the FAD-binding domain and residues 45–225+317–461 in the steroid-binding domain,[33] are identified in the tree decomposition [(5–159+191–323 + 383–405 + 443–506) (160–190+324–381+ 406–442)] but fall just below the τ^2-cutoff because of a tighter interface [τ^2 = 2.2 $(ps)^2$].

Trypsin-like serine proteinases are another well known protein family where the strength of the domain interface varies considerably. The fold contains an internal duplication of an antiparallel beta-barrel motif. Trypsin [1sgt in Fig. 5, τ^2 = 1.6 $(ps)^2$] and a remote viral homolog [2snv, τ^2 = 2.5 $(ps)^2$] are parsed into two compact structural domains. The duplication is identified in the tree decomposition of proteinase A [2sga, (16–124+234–242)(125–233), τ^2 = 1.1 $(ps)^2$], but two hydrogen bonds gluing loops to the other domain prevent cutting. *Achromobacter* protease I (1arb), defined here to be a single structural domain, has such a tight interface that the tree decomposition splits through the second β-barrel rather than between the barrels [(1–22+74–97+144–161+180–193+212–225)(23–73+ 98–143 +162–179+194–211+226–263), τ^2 = 1.3 $(ps)^2$].

Comparison of Unfolding Trees With Experiment

Early folding intermediates of a number of small proteins have been probed by NMR methods. (1) The tree decomposition of barnase is [((22–33)(34–51))((57–69)((2–21,52)((96–110) ((83–95)(53–56,70–92)))))]. Loops at the extremities are removed first while the compact sheet-helix motif (at the right in the unfolding tree) is resistant to unfolding with the present algorithm. Experiment shows that all the regions that fold early interact extensively with the β-sheet.[34] (2) The first cut in the tree decomposition of apomyoglobin [(1–19+71–153)(20–70))] splits open the haem pocket. The first unit contains an early folding intermediate (helices A-G-H).[35] (3) The first cut in the tree decomposition of apocytochrome *c* [1–31)(32–79))] also splits open the haem pocket. The N-terminal unit contains three helices, which are protected from amide exchange in a folding intermediate.[36] (4) The tree decomposition of pancreatic trypsin inhibitor [(1–18)(((19–32)(42–58))(33–44))] identifies the flexible N-terminal arm.[37] (5) Ubiquitin folds very fast in a single step.[38] The strong cohesion is reflected in a small τ^2 = 0.8 $(ps)^2$ for the first cut compared to the cutoff for autonomous units $\tau^2 \geq 2.6$ $(ps)^2$. (6) H40–H71 and M180–H200 are two early folding segments of elastase, a relative of trypsin.[39] The segments map to each of the two domains and survive 4 and 6 cuts in the unfolding tree, which has a largest depth of 6 cuts.

The qualitative features of the unfolding trees remain similar without constraining the minimal segment length, i.e., setting this parameter to 1 residue instead of 10 residues (as above and elsewhere in this work).

DISCUSSION

We have presented a general and computationally efficient method for the elucidation of the borders between structural domains in proteins of known three-dimensional coordinates. The novel aspects are the physical criteria used and the eigenvalue analysis of contact maps. Although the harmonic oscillator model is a simplified approximation, its qualitative features are more general and our results agree well with visual intuition. The method comes close to the goal of a fully objective definition of domains and can be a useful tool in the automatic classification of recurrent folding motifs[40] in the flood of newly solved structures.

There is sufficient experimental evidence for independently folding domains in many larger natural proteins, reflecting a partitioning of the folding problem into units of simple structure and intermediate size. The key question is then whether the present method is a valid extrapolation to identify physically independent folding units in unknown cases. There are a number of open technical questions regarding the calibration of the method. One area of potential improvement is that of correct energetics, as we only used a simple force field. The choice of thresholds for autonomous units is a first approximation and may be adjusted as more data becomes available. Reassuringly, the domain defini-

tion was reasonably robust with respect to different cutoff values or rule sets, i.e., the number of domains of only a small number of proteins was affected as parameters were explored.

The experimental verification of the predictions implied, e.g., by Table II is not straightforward and touches on the question of protein stability vs. specificity of the folded conformation.[41] Separation of domains with a large hydrophobic interface is likely to produce unspecific aggregates. One way to perform experiments is to make the interdomain contact surface more polar, taking care not to affect the core of the domain, and then produce the isolated fragment and test for folding into the native conformation. The goal of such experiments would be to ascertain whether the units defined here can have an autonomous existence as building blocks, either in evolution, or in protein folding, or in protein design.

REFERENCES

1. Holm, L., Sander, C. Searching protein structure databases has come of age. Proteins 19:165–173, 1994.
2. Bork, P. Mobile modules and motifs. Curr. Opin. Struct. Biol. 2:413–421, 1992.
3. Nemethy, G., Scheraga, H.A. A possible folding pathway of bovine pancreatic RNase. Proc. Natl. Acad. Sci. U.S.A. 76:6050–6054, 1979.
4. Murphy, K., Bhakuni, V., Xie, D., Freire, E. Molecular basis of cooperativity in protein folding. III. Structural identification of cooperative folding units and folding intermediates. J. Mol. Biol. 227:293–306, 1992.
5. Crippen, G. The tree structural organization of proteins. J. Mol. Biol. 126:315–332, 1978.
6. Schulz, G.E., Schirmer, H. In "Principles of Protein Structure," Chapter 5. New York: Springer-Verlag, 1979.
7. Noguti, T., Sasakibara, H., Gō, M. Localization of hydrogen-bonds within modules in barnase. Proteins 16:357–363, 1993.
8. Rose, G.D. Hierarchic organization of domains in globular proteins. J. Mol. Biol. 134:447–470, 1979.
9. Wodak, S.J., Janin, J. Location of structural domains in proteins. Biochemistry 20:6544–6552, 1981.
10. Lesk, A.M., Rose, G.D. Folding units in globular proteins. Proc. Natl. Acad. Sci. U.S.A. 78:4304–4308, 1981.
11. Rashin, A. The hinge-bending mode in lysozyme. Nature (London) 262:325–326, 1976.
12. Segawa, S., Richards, F.M. Identification of regions of potential flexibility in protein structures: Folding units and correlations with intron positions. Biopolymers 27:23–40, 1988.
13. Zehfus, M.H., Rose, G.D. Compact units in proteins. Biochemistry 25:5759–5765, 1986.
14. Zehfus, M.H. Improved calculations of compactness and a reevaluation of continuous compact units. Proteins 16: 293–300, 1993.
15. Rashin, A.A. Location of domains in globular proteins. Nature (London) 291:85–87, 1981.
16. Moult, J., Unger, R. An analysis of protein folding pathways. Biochemistry 30:3816–3824, 1991.
17. Sander, C. Physical criteria for folding units of globular proteins. In "Structural Aspects of Recognition and Assembly in Biological Macromolecules." Balaban, M. et al., eds. Rehovot: Balaban ISS, 1981: 183–195.
18. Kabsch, W., Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymer 22:2577–2637, 1983.
19. Serrano, L., Kellis Jr., J.T., Cann, P., Matouschek, A., Fersht, A.R. The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. J. Mol. Biol. 224:783–804, 1992.
20. Go, N., Noguti, T., Nishikawa, K. Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc. Natl. Acad. Sci. U.S.A. 80:3696–3700, 1983.
21. Levitt, M., Sander, C., Stern, P.S. Protein normal mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. J. Mol. Biol. 181:423–447, 1985.
22. Hill, M.O. Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61:237–251, 1973.
23. Hill, M.O. Correspondence analysis: a neglected multivariate method. Appl. Statist. 23:340–354, 1974.
24. Holm, L. Codon usage and gene expression. Nucl. Acids Res. 14:3075–3087, 1986.
25. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Protein Sci. 1:409–417, 1992.
26. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
27. Holm, L., Sander, C. Database algorithm for generating protein backbone and side-chain co-ordinates from a C^{α} trace. Application to model building and detection of co-ordinate errors. J. Mol. Biol. 218:183–194, 1991.
28. Holm, L., Sander, C. Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology. Proteins 14:213–223, 1992.
29. Corbett, R.J., Roche, R.S. Independent folding of autolytic fragments and their domain-like properties. Int. J. Pept. Res. 28:549–559, 1986.
30. Missiakas, D., Betton, J.M., Minard, P., Yon, J.M. Unfolding of the domains in yeast phosphoglycerate kinase: Comparison with the isolated engineered domains. Biochemistry 29:8683–8689, 1990.
31. Schulz, G.E., Schirmer, R.H., Sachsenheimer, W., Pai, E.F. The structure of the flavoenzyme glutathione reductase. Nature (London) 273:120–124, 1978.
32. Wierenga, R.K., Jong, R.J.d., Kalk, K.H., Hol, W.G.J., Drenth, J. Crystal structure of p-hydroxybenzoate hydroxylase. J. Mol. Biol. 131:55–73, 1979.
33. Vrielink, A., Lloyd, L.F., Blow, D.M. Crystal structure of cholesterol oxidase from *Brevibacterium sterolicum* refined at 1.8 Å resolution. J. Mol. Biol. 219:533–554, 1991.
34. Serrano, L., Matouschek, A., Fersht, A.R. The folding of an enzyme. VI. The folding pathway of barnase: Comparison with theoretical models. J. Mol. Biol. 224:847–859, 1992.
35. Barrick, D., Baldwin, R.L. Stein and Moore Award address. The molten globule intermediate of apomyoglobin and the process of protein folding. Protein Sci. 2:869–876, 1993.
36. Jeng, M.F., Englander, S.W., Elove, G.A., Wand, A.J., Roder, H. Structural description of acid-denatured cytochrome c by hydrogen exchange and 2D NMR. Biochemistry 29:10433–10437, 1990; erratum Biochemistry 30:2988, 1991.
37. van Mierlo, C.P., Darby, N.J., Creighton, T.E. The partially folded conformation of the Cys-30 Cys-51 intermediate in the disulfide folding pathway of bovine pancreatic trypsin inhibitor. Proc. Natl. Acad. Sci. U.S.A. 89:6775–6779, 1992.
38. Briggs, M.S., Roder, H. Early hydrogen-bonding events in the folding reaction of ubiquitin. Proc. Natl. Acad. Sci. U.S.A. 89:2017–2021, 1992.
39. Ghelis, C. Transient conformational states in proteins followed by differential labeling. Biophys. J. 32:503–514, 1980.
40. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233:123–138, 1993.
41. Lattman, E.E., Rose, G.D. Protein folding—what's the question? Proc. Natl. Acad. Sci. U.S.A. 90:439–441, 1993.
42. Kraulis, P. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24:946–950, 1991.
43. Kabsch, Q., Mannherz, H.G., Suck, D., Pai, E.F., Holmes, K.C. Atomic structure of the actin:DNase I complex. Nature (London) 347:37–44, 1990.