A protein taxonomy based on secondary structure

Teresa Przytycka[1], Rajeev Aurora[1,2] and George D. Rose[1]

Does a protein's secondary structure determine its three-dimensional fold? This question is tested directly by analyzing proteins of known structure and constructing a taxonomy based solely on secondary structure. The taxonomy is generated automatically, and it takes the form of a tree in which proteins with similar secondary structure occupy neighboring leaves. Our tree is largely in agreement with results from the structural classification of proteins (SCOP), a multidimensional classification based on homologous sequences, full three-dimensional structure, information about chemistry and evolution, and human judgment. Our findings suggest a simple mechanism of protein evolution.

Does protein secondary structure determine the three-dimensional fold?

Historically, this question has been controversial and remains so. Both experiment[1–4] and theory[5] suggest that tertiary structure may precede secondary structure in the folding of a protein. Furthermore, a given sequence of secondary-structure elements may be able to adopt many three-dimensional arrangements[6], with the actual fold selected from this set of possibilities by sequentially long-range interactions that are particular to the sequence. Paradoxically, these conclusions are inconsistent with accumulating experimental evidence that many peptide fragments excised from proteins still adopt the native fold in the absence of longer range interactions. Again, both experiment[7,8] and theory[9] indicate that much of the native organization emerges during the earliest stages of folding, preceding contributions from nonlocal interactions.

Here we attempt to examine the question directly by analyzing proteins of known structure and assessing whether secondary structure maps into tertiary structure uniquely. If so, then fold classification and taxonomy can be based on secondary structure unambiguously. To explore this conclusion, we devise a straightforward metric that maps two patterns of secondary-structure elements into the 'distance' between them. This metric is then used to construct a tree, which is compared, in turn, with similar constructs from more sophisticated approaches (such as, structural classification of proteins, SCOP[10], based on homologous sequences, structure, evolutionary knowledge, and human judgment; and vector alignment search tool, VAST[11], based solely on three-dimensional information). Differences are discussed in detail and related controls are performed.

Protein secondary structure has been used previously in fold classification[12]. Were secondary structure alone sufficient to identify fold topology, then no further structural information would be needed because the fold would be implicit. With this in mind, several groups have developed algorithms for fold recognition based on secondary structure[13–17].

In our method, the similarity between two proteins is derived from their aligned elements of secondary structure. We adopt the simplest possible approach, using only the ordered sequence of helices, extended strands, and interconnecting loops that span the polypeptide chain. Each protein is represented by an ordered sequence of such elements, together with their lengths, termed the 'ss string'. Classification into these categories is based entirely on backbone dihedral angles; no topological or connectivity information is included. In particular, hydrogen-bonded β-sheet, which we regard as tertiary structure, is ignored deliberately.

Comparing protein structures is a challenging task[18]. In contrast, comparing protein sequences is performed routinely by aligning their sequences using a dynamic programming algorithm[19]. Usually, similar sequences have similar folds[20,21], but the converse need not be true; dissimilar sequences can also have similar folds. The reliability of such sequence-based methods diminishes as sequence similarity decreases, sinking into a region of uncertainty (called the 'twilight zone') below ~30% aligned sequence identity[22]. Both sequence families[23–27] and structural information[28,29] can be used to extend the threshold of reliable detection. To what degree can secondary structure alone enhance sensitivity? We turn now to this question.

Terminology

The terminology used in this paper follows that introduced by SCOP[10], with some additional categories. In particular, the term 'family' defines proteins having a clear evolutionary relationship, with aligned pairwise sequence identity in excess of 30%. In exceptional cases, SCOP classifies proteins into a family in the absence of high sequence identity but when similar function and structure provide definitive evidence of common descent.

As described in the Methods section, proteins analyzed in this paper are from PDB_select (August 1996)[30]. As such, any pair of molecules is guaranteed to have an aligned sequence identity <30%, and, accordingly, none satisfies the usual definition of a SCOP family. Thus, the classification by SCOP of two proteins from our test set into the same family must be due to either similar structure/function or a transitive series of high sequence associations (that is, A ~B ~ C ~..., where X ~ Y means that the aligned sequence identity between X and Y exceeds 30%), in which only the first and last members belong to our test set. We

[1]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, 725 N. Wolfe Street, Baltimore, Maryland 21205, USA. [2]Current address: Monsanto, Mail Zone AA3G, 700 Chesterfield Parkway, St. Louis, Missouri 63198, USA.

Correspondence should be addressed to G.D.R. *E-mail: rose@grserv.med.jhmi.edu*

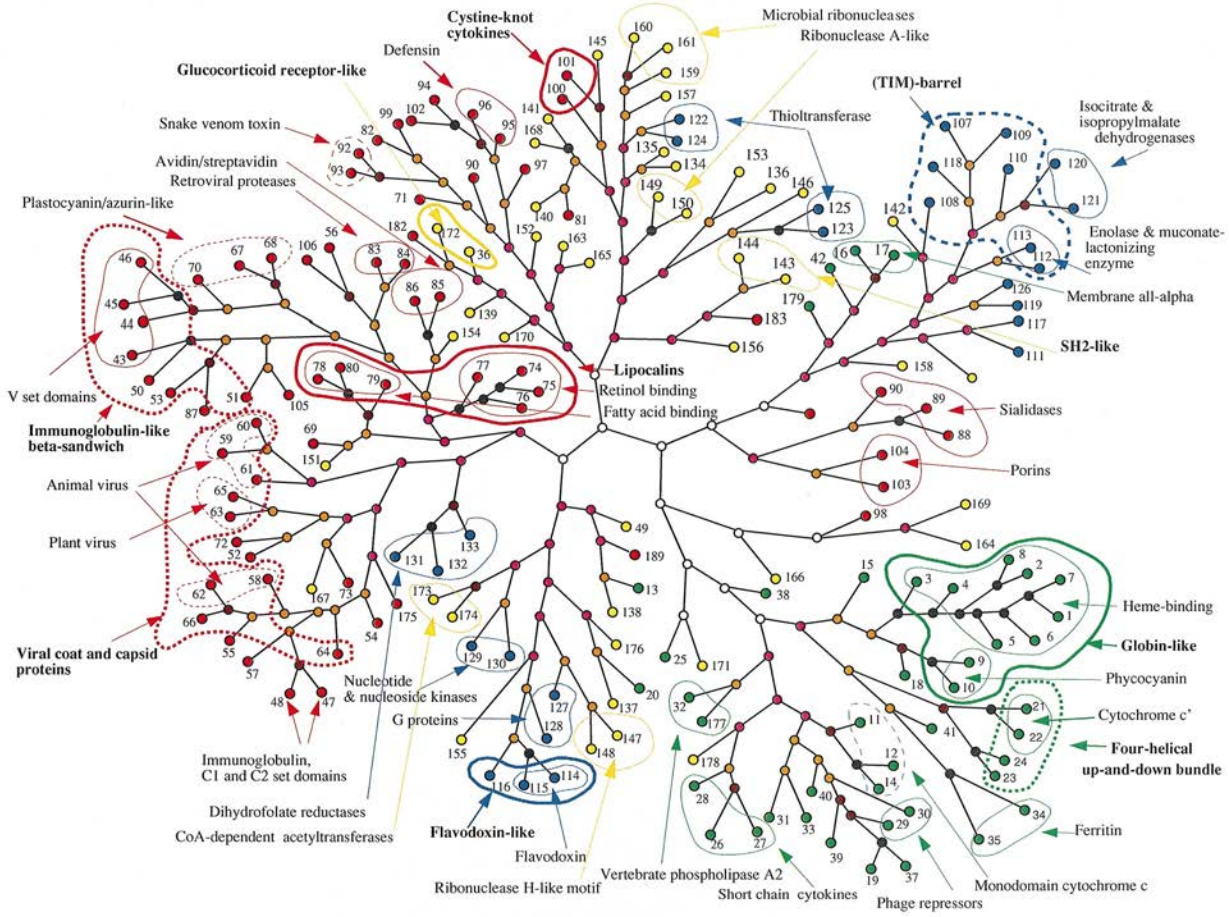

Fig. 1 Similarity tree for the 183 proteins in our data set, generated automatically as described in the Methods. Proteins are color-coded by class: α, green; β, red; α+β, yellow; α/β, blue. Proteins belonging to the same SCOP[10] fold are circumscribed, either by solid (complete family) or dashed (incomplete family) lines. Thin lines indicate folds that include a single SCOP extended family; thick lines indicate folds that contain members of at least two different extended families. Encircled groupings are intended only as an approximate guide. For clarity, only categories with cluster scores ≥0.3 are shown, categories that did not cluster well are not encircled, and thin lines that clutter the tree to the point of obstruction have also been omitted. The tree was generated[44] by the weighted-pair group (WPG) method[43]. The branch lengths on the figure are unrelated to distance. To provide distance information, each node is color-coded according to the average distance, τ, between pairs of leaves in its cluster: black, $\tau \le 0.15$; brown, $0.15 < \tau \le 0.20$; orange $0.20 < \tau \le 0.30$; pink, $0.30 < \tau \le 0.40$; white, $\tau > 0.4$.

introduce the term 'extended family' to denote proteins that belong to the same SCOP family, but with sequence identity <30% for any two members.

In SCOP, a 'fold' contains proteins with substantial structural similarity. In our analysis, comparison is focused on the family and/or fold, and we introduce the term 'category' to describe proteins in our test set that belong to either the same SCOP extended family or SCOP fold.

A set of proteins, P, is said to form a 'cluster' with respect to some similarity tree, T, when there exists an internal node in T such that all members of P are descendents of that node. Given a set of proteins, 'cluster evaluation' is a quantitative estimate of the relatedness among its elements in that similarity tree. Cluster evaluation is performed with respect to a scoring function introduced in this paper. We adopt informal terms such as 'a set clustered well' or 'a set did not cluster' to describe the results of cluster evaluation.

Results

A representative set of 183 proteins (Table 1) with pairwise aligned sequence identity <30% was chosen previously from the Protein Data Bank[31]. Using SCOP[10], each protein was placed in one of four classes: α (mostly helical), β (mostly sheet), α/β (mixed regions of helix and sheet), and α+β (separate regions of helix and sheet)[32]. Figure 1 is a taxonomic tree of this data set.

To evaluate our taxonomic tree (Fig. 1), we compare it with classifications of the same proteins by SCOP and VAST, two established, information-rich approaches. The SCOP program of Murzin *et al.*[10] uses sequence homology, three dimensional information, chemical knowledge and informed human judgment. The VAST program of Bryant and co-workers[11] uses three-dimensional coordinates in a completely automated comparison. All SCOP categories that do not form a cluster with respect to our tree are examined individually. In addition, we introduce a quantitative cluster-scoring function and use it to evaluate SCOP categories in our tree and in the tree that would have been obtained using scores from VAST. This latter evaluation can reveal potential fold dissimilarities within a SCOP category.

In the ensuing analysis, most comparisons are between our classification and the corresponding SCOP family, that is,

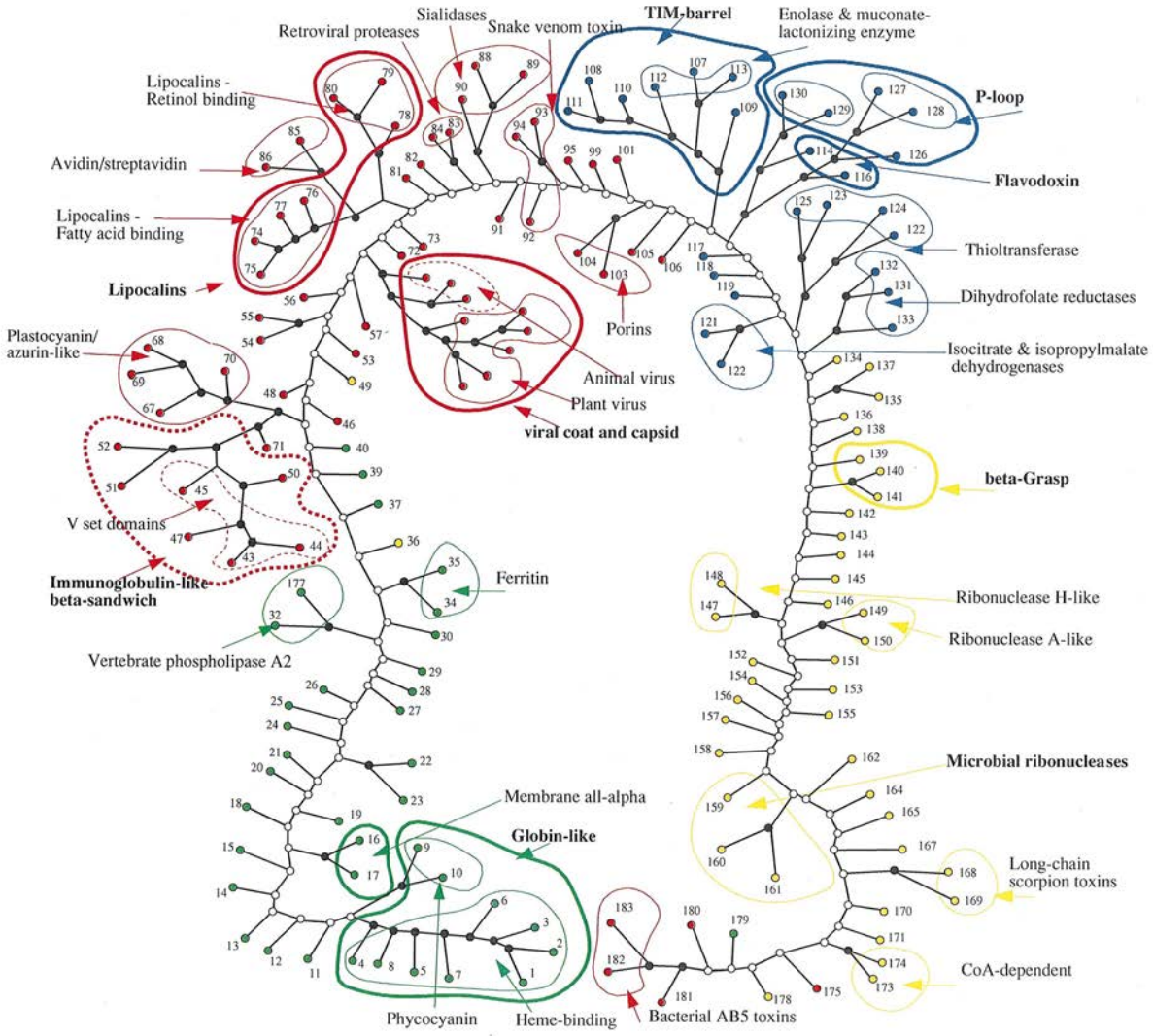

Fig. 2 The tree obtained using VAST[11]. All distances are binary, that is, folds that are the same have a distance of 0, folds that differ have a distance of 1. The color coding of leaves and nodes is the same as in Fig. 1.

between a structure-based method and a largely sequence-based method. Given that every pair of proteins in our test set has an aligned sequence identity <30%, there is insufficient information for sequence-based classification into SCOP families. Nevertheless, the clusters in our tree (Fig. 1) do recapitulate their corresponding SCOP families with considerable success. In sequence-based comparisons, such family relationships usually evade detection at the low level of sequence similarity present in our test set (although such methods have gained sensitivity in recent approaches). Because our taxonomy uses only ss strings, the relevant family information must be embededed there. Indeed, this is the principal conclusion to emerge in comparisons against SCOP. We emphasize that it was not our intention to reinvent the SCOP hierarchy but merely to show that ostensibly elusive information about sequence-based families can be extracted from secondary structure alone.

Taking SCOP as the 'gold standard', a 'false positive' is defined as a pair of proteins found to have the same fold by our method but not by SCOP. If the same ordered set of secondary-structure elements can be arranged in multiple ways, then many false positives are expected. Conversely, a 'false negative' is a pair of proteins with the same fold in SCOP but not in our tree. Intuitively, false negatives are not expected because proteins with identical tertiary structure will also have the same secondary structure. Nevertheless, false negatives are encountered when our dihedral angle-based classification misses a short helix or fails to distinguish between a strand and an extended loop, or when a given SCOP extended family includes multiple folds.

The principal organizing hypothesis we seek to explore is that secondary structure determines tertiary structure. If so, then similar ss strings will have similar folds. False positives serve as a stringent test of this hypothesis, and we now analyze them individually.

Referring to Fig. 1, false positives are examined within the two closely related distance categories, that is, black and brown. The most similar level (black) includes three pairs of proteins with different tertiary structures: <19, 37> = <4icb, 1aca>, <168, 141> = <2sn3, 1ubq>, and <94, 102> = <1fas, 4sgbI>. All are

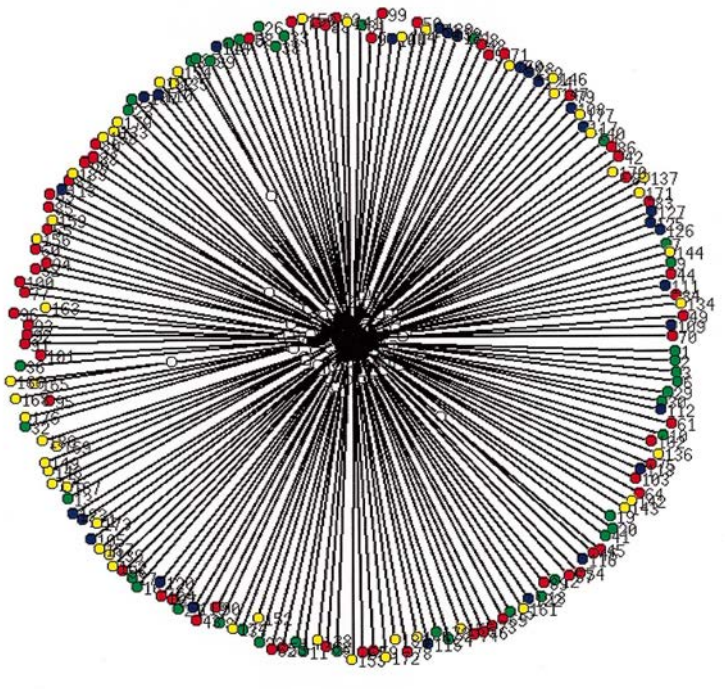

Fig. 3 Primary sequence tree, constructed using the neighbor-joining (NJ) method[34]. Distances were computed from primary sequence alignment. Branch length is proportional to the distance between nodes. Given that the test set was selected to have low sequence identity (<30%), the tree has approximate radial symmetry, as expected.

small, single-domain structures. Proteins in the first pair have four helices, and proteins in the other pairs have a multistranded sheet. In visual comparison of pair one, the all-helical proteins contain large loops, allowing for considerable variability in helix/helix packing. False positives in the other two pairs occur because an extended loop is misidentified as a strand by our dihedral angle-based method. The next level of similarity (brown) includes four false positives: 1bds (96) is clustered with <94, 102> because, again, an extended loop is misidentified as a strand. Similarly, imperfect strand recognition in 2tgi (101) causes it to be grouped with 2sicI (145). 1noa (53) and 1vqb (87) are grouped because a strand is mistakenly divided into two strands at a β-bulge. Finally, 1fiaA (39), a four-helix protein, is clustered with the two λ repressors, 1lmb3(29) and 1r69 (30), each of which has a helix (but of very different lengths) followed by four additional helices. False positives are few in number, and most often, they result from failure to distinguish between a loop and a strand.

As an alternative gold standard, a tree was constructed for proteins in common to VAST (December, 1997 VAST version) and our data set (Fig. 2). The VAST similarity score between two structures is deliberately conservative. For a given pair of proteins, the maximal common three-dimensional substructure is identified, and a cutoff value is used to eliminate any fortuitious similarities that could be expected by chance. We assign a score of 0 to all common substructures smaller than this cutoff value. Consequently, in Fig. 2, proteins emanating from the same white node are classified into different clusters (even when their nodes are neighbors in the tree). In the absence of the cutoff value, VAST scores would result in larger clusters descending from nonwhite nodes, but with an increased probability of a false positive.

In the following paragraphs, we make a detailed comparison of false negatives between all SCOP categories and the clusters in Fig. 1. To assist interpretation, proteins from the same SCOP category are encircled, as described in the figure caption. All differences are examined, first individually, and then numerically, using our cluster-scoring function. Individual comparisons are broken down by class as follows.

α-Class. SCOP extended families consistent with our tree include the globins (1–8), phycocyanins (9–10), phospholipases A (32–177), membrane all-α (16–17), and ferritins (34–35), cytochrome *c*' (21–22), monodomains (11–14) — with the exception of 2mtaC (13) — short-chain cytokines (26–28), and phage repressors (29–30). At the SCOP fold level, our tree includes the four-helical up-and-down bundles (21–25), with the exception of 2tmv (25), which has long loop regions that our algorithm identifies as strands. In addition to the two false positives already mentioned, disagreement with SCOP includes the EF hand SCOP fold (18–20), represented in our test set by proteins from different SCOP families with differing structures. In addition, 2mtaC (13), a cytochrome *c* fold (monodomain extended family), is misplaced because several loop regions are interpreted as strands.

For the α-class, most, though not all, VAST classifications are consistent with our tree. VAST does not cluster the monodomain extended family (11–14) or short-chain cytokines (26–28).

β-Class. SCOP extended families consistent with our tree include the fatty acid-binding lipocalins (74–77), retinol-binding lipocalins (78–80), acid proteases (83–84), streptavidins (85–86), sialidases (88–90) and porins (103–104). The V-set domains (43–46) cluster well though not perfectly. Imperfect clusters are formed by both extended families (animal virus, 56–62; and plant virus, 63–66), representing viral coat and capsid SCOP fold (58–66); plastocyanin/azurin-like (67–70); and snake venom toxins (92–94). The defensin extended family members (95–96), which have differing numbers of strands, are close but not immediate neighbors.

At the SCOP fold level, immunoglobulin-like β-sandwiches (43–53) cluster well, though imperfectly; C1 and C2 set domains (48, 49) cluster separately, and 1vqb (87) is a false negative, as mentioned before. The extended family containing the EGF-type modules (97–98) is not grouped because the protein is quite small, and clustering is dominated by a short hairpin in one member that is a loop in the other. Nor is the Con A (54–57) grouped, as expected, given that different members have a differing number of strands. The two trypsins (81, 82) fail to cluster; again, the structures are dissimilar. Unsurprisingly, OB folds (87, 180–183) do not cluster either; some are α+β, but, in keeping with the SCOP convention, we classified them into all-β.

VAST classifications agree with our own, for the most part, but with some exceptions. Unlike in Fig. 1, the SCOP viral coat and capsid fold does form a cluster in the VAST tree in Fig. 2. However, immunoglobulins fail to cluster perfectly in VAST, as they do by our method, and the Con A cluster is incomplete also. The OB fold is partially recovered in VAST. Other VAST clusters are consistent with ours.

α/β-Class. SCOP extended families consistent with our tree include flavodoxins (114–116) and dihydrofolate reductases (131–133). Thioredoxins (122–125) cluster imperfectly. The isocitrate and isopropylmalate dehydrogenase extended fami-

Table 1*a* Mostly α proteins used in this study (green)

	PDB ID	Family		PDB ID	Family
1.	1eca	Heme binding	23.	256bA	Cytochrome *b*562
2.	1mbd	Heme binding	24.	2hmqA	Hemerythrin
3.	3sdhA	Heme binding	25.	2tmv	Viral coat proteins
4.	1hbg	Heme binding	26.	3inkC	Short-chain cytokines
5.	1thbA	Heme binding	27.	1rcb	Short-chain cytokines
6.	1mba	Heme binding	28.	2gmfA	Short-chain cytokines
7.	1ithA	Heme binding	29.	1lmb3	Phage repressors
8.	1lh1	Heme binding	30.	1r69	Phage repressors
9.	1cpcA	Phycocyanin	31.	1lfb	Homeodomain
10.	1cpcB	Phycocyanin	32.	1ppa	Vertebrate phospholipase A2
11.	3c2c	Monodomain cytochrome *c*	177.	1poa	Vertebrate phospholipase A2
12.	351c	Monodomain cytochrome *c*	176.	1poc	Insect phospholipase A2
13.	2mtaC	Monodomain cytochrome *c*	33.	1utg	Uteroglobin
14.	1cc5	Monodomain cytochrome *c*	34.	1fha	Ferritin
15.	1colA	Colicin membrane	35.	1rci	Ferritin
16.	1prcM	Membrane all-α	37.	1aca	Acyl-CoA binding protein
17.	1prcL	Membrane all-α	38.	1ropA	ROP protein
18.	1osa	Calmodulin	39.	1fiaA	FIS protein
19.	4icb	Calbindin D9K α	40.	2wrpR	Trp repressor
20.	1rro	Parvalbumin	41.	1aep	Apolipophorin-III
21.	2ccyA	Cytochrome c'	42.	1prcC	Photosynthetic reaction center
22.	1bbhA	Cytochrome c'	179.	1csc	Citrate synthase

Table 1*b* Mostly β proteins used in this study (red)

	PDB ID	Family		PDB ID	Family
43.	1cd8	V-set domain, immunoglobulin	78.	1bbpA	Retinol binding
44.	2rhe	V-set domain, immunoglobulin	79.	1hbq	Retinol binding
45.	1cdb	V-set domain, immunoglobulin	80.	1mup	Retinol binding
46.	1cdc	V-set domain, immunoglobulin	81.	2sga	Prokaryotic proteases, trypsin
47.	1cid	1–105: C2-set domain, immunoglobulin	82.	2pkaA	Eukaryotic proteases, trypsin
48.	1fc2D	238–341: C1-set domain, immunoglobulin	83.	2rspA	Acid protease
50.	1tlk	I-set domain, immunoglobulin	84.	1hivA	Acid protease
51.	1ten	Fibronectin type III	85.	1stp	Avidin/streptavidin
52.	1cobA	Cu, Zn Superoxide dismutase	86.	1aveA	Avidin/streptavidin
53.	1noa	Actinoxantin	88.	1nscA	Sialidase
54.	2ltnA	Legume lectins	89.	2bat	Sialidase
55.	2ayh	Bacillus 1-3,1-4-β-glucanase	90.	2sim	Sialidase
56.	1sltA	Galectin	91.	1hcc	Complement control
57.	2ctvA	Galectin	92.	3ebx	Snake venom toxin
58.	1bbt3	Animal virus, viral coat and capsid	93.	1cdtA	Snake venom toxin
59.	2plv1	Animal virus, viral coat and capsid	94.	1fas	Snake venom toxin
60.	2plv2	Animal virus, viral coat and capsid	95.	1atx:	Defensin
61.	2plv3	Animal virus, viral coat and capsid	96.	1bds	Defensin
62.	1bbt2	Animal virus,viral coat and capsid	97.	1egf	EGF-type module, Knottin
63.	4sbvA	Plant virus, viral coat and capsid	98.	2tgf	EGF-type module, Knottin
64.	2tbvA	Plant virus, viral coat and capsid	99.	1tpm	Fibronectin type I module
65.	2stv	Plant virus, viral coat and capsid	100.	1pdg	Platelet-derived growth factor
66.	1bmv1	Plant virus, viral coat and capsid	101.	2tgi	Transforming growth factor
67.	1plc	Plastocyanin/azurin	102.	4sgbI	Plant proteinase inhibitors
68.	1aaj	Plastocyanin/azurin	103.	2por	Porin
69.	1paz	Plastocyanin/azurin	104.	2omf	Porin
70.	1aizA	Plastocyanin/azurin	105.	1ttaA	Transthyretin
71.	1hoe	Amylase inhibitor sandwich	106.	4fgf	Fibroblast growth factors
72.	1tnfA	Tumor necrosis factor sandwich	175.	1zsa	Carbonic anhydrase
73.	1gpr	Glucose permease	87.	1vqb	Bacteriophage ssDNA-binding
74.	1ifc	Fatty acid binding	180.	1pyp	Organic pyrophosphatase
75.	1ifb	Fatty acid binding	181.	2sns	Staphylococcal nuclease
76.	1mdc	Fatty acid binding	182.	1bovA	Bacterial AB5 toxins, B-subunit
77.	1opaA	Fatty acid binding	183.	1ltsD	Bacterial AB5 toxins, B-subunit

Table 1*c* α/β **proteins used in this study (blue)**

	PDB ID	Family		PDB ID	Family
107.	1xis	Xylose isomerase (TIM)-barrel	121.	7icd	Isocitrate and isopropylmalate dehydrogenase
108.	5timA	Triosephosphate isomerase (TIM)-barrel	122.	2trxA	Thioltransferase
109.	1nar	type II chitinase (TIM)-barrel	123.	1aba	Thioltransferase
110.	1fbaA	Aldolase (TIM)-barrel	124.	3trx	Thioltransferase
111.	1gox	FMN-linked oxidoreductases (TIM)-barrel	125.	1ego	Thioltransferase
112.	2mnr	133–359: Enolase and muconate-lactonizing enzyme, C-terminal domain, (TIM)-barrel	126.	1nipA	Nitrogenase iron protein-like
113.	1chrA	127–370: Enolase and muconate-lactonizing enzyme, C-terminal domain, (TIM)-barrel	127.	5p21	G protein
114.	1ofv	Flavodoxin	128.	1etu	G protein
115.	4fxn	Flavodoxin	129.	2ak3A	Nucleotide and nucleoside kinase
116.	3chy	CheY-like	130.	3adk	Nucleotide and nucleoside kinase
117.	2ctc	Pancreatic carboxypeptidases	131.	4dfrA	Dihydrofolate reductase
118.	2cmd	Lactate and malate dehydrogenase	132.	3dfr	Dihydrofolate reductase
119.	1cseE	Subtilase	133.	8dfr	Dihydrofolate reductase
120.	1ipd	Isocitrate and isopropylmalate dehydrogenase			

Table 1*d* $\alpha+\beta$ **proteins used in this study (yellow)**

	PDB ID	Family		PDB ID	Family
134.	1pba	Pancreatic carboxypeptidase	156.	3b5c	Cytochrome *b*
135.	2bopA	viral DNA-binding domain	157.	2msbA	C-type lectin
136.	1fxd	Short-chain ferredoxin	158.	2tscA	Thymidylate synthase
137.	2nckL	Nucleoside diphosphate kinase	159.	1fus	Microbial ribonuclease
138.	3rubS	22–147: Ribulose 1,5-bisphosphate carboxylase-oxygenase	160.	1brnL	Microbial ribonuclease
139.	1pgx	Immunoglobulin-binding	161.	1gmpA	Microbial ribonuclease
140.	1frrA	2Fe-2S ferredoxin	162.	1rveA	Restriction endonuclease *Eco*RV
141.	1ubq	Ubiquitin	163.	1adn	Ada DNA repair protein
142.	1apa	Ribosome-inactivating proteins	164.	1cbn	Crambin-like
143.	1shaA	SH2-like	165.	5pti	Small Kunitz-type inhibitors and BPTI-like toxins
144.	2pna	SH2-like	166.	1hev	Agglutinin
145.	2sicI	Subtilisin inhibitor	167.	wgaA	Agglutinin
146.	1ptf	Histidine-containing phosphocarrier protein	168.	2sn3	Long-chain scorpion toxins
147.	2rn2	Ribonuclease H	169.	1gps	Long-chain scorpion toxins
148.	1hrhA	Ribonuclease H	170.	2ovo	Animal Kazal-type inhibitor
149.	1onc	Ribonuclease A	171.	8rxnA	Rubredoxin
150.	7rsa	Ribonuclease A	172.	1gatA	Glucocorticoid receptor
151.	1aak	Ubiquitin-conjugating enzyme	36.	1gluA	Hormone receptor
152.	3il8	Interleukin 8-like chemokine	173.	3cla	CoA-dependent acetyltransferase
153.	1ctf	Ribosomal protein L7/12	174.	1eaf	CoA-dependent acetyltransferase
154.	1fkb	FKBP-like	178.	1cmbA	Met repressor-like
155.	1ltsA	ADP-ribosylation toxin	49.	1tmcA	MHC antigen

ly, which can be considered 'open barrels', are grouped with the barrel SCOP fold (107–113) because our methods fail to distinguish between open and closed barrels. The P-loop superfamily (126–130) actually includes members of three extended families: only one forms a cluster (129–130), and, in fact, each of the three has a different number of helices and strands.

VAST classifies the P-loop superfamily with the flavodoxin superfamily (114–116), as does our method in one case (128). However, the VAST barrel cluster is complete.

α+β-Class. SCOP extended families consistent with our tree include SH2 (143–144), ribonuclease H (147–148), ribonuclease A (149–150), microbial ribonucleases (159–161), and CoA-dependent acetyltransferases (173–174). Finally, we lack knottin clusters (166–167) and (168–169), because our secondary-structure recognition algorithm misidentifies some extended segments in loops as strands.

At the SCOP fold level, the flavodoxin fold (114–116) cluster is complete, but ferrodoxins (134–138) and β-grasp proteins (139–141) did not cluster well.

VAST did not cluster microbial ribonucleases (159–161) and SH2 (143–144) domains. Consistent with Fig. 1, VAST groups two β-grasp proteins (140–141) but not the third (139). Also

Table 2*a* Cluster scores

Fold type / Extended family	Number of Members	Members	Score	VAST score[1]
Globin-like	**10**	**1–10**	**0.9**	**0.6**
Heme-binding protein	8	1–8	1	1
Phycocyanin	2	9–10	1	1
Cytochrome *c*				
Monodomain cytochrome *c*	4	11–14	0.5	0
Membrane all-α				
Membrane all-α	2	16–17	1	1
EF Hand-like	**3**	**18–20**	**0.0**	**0**
Calmodulin-like	1	18		
Calbindin D9K	1	19		
Parvalbumin	1	20		
Four-helical up-and-down bundle	**5**	**21–25**	**0.6**	**0.1**
Cytochrome c'	2	21–22	1	0
Cytochrome *b*562	1	23		
Hemerythrin	1	24		
Viral coat proteins	1	25		
Four-helical cytokines				
Short-chain cytokines	3	26–28	1	0
Lambda repressor-like DNA-binding				
Phage repressors	2	29–30	0.6	0

similar to our tree, VAST clusters the ferredoxins incompletely. Finally, VAST leads to one false positive with respect to SCOP: 1ctf (ribosomal protein, 153) and 1pba (ferredoxin, 134).

To score SCOP categories in our tree quantitatively, we designed a conservative cluster-scoring function that assesses tree topology (that is, separation in the tree), but also includes a cutoff value (Table 2). In essence, the total score for a given SCOP category is reckoned as the sum over all pairwise scores divided by the number of pairs. The score for a pair ranges between 0 and 1, and it measures the directness of the path connecting these two nodes, that is, how often the path between two nodes in a SCOP category is interrupted by branches to proteins external to that category. The scoring function also includes a numerical cutoff value such that any path traversing a white node has a score of zero. In the Methods, we give a precise definition of the scoring function. Each SCOP category is scored both in our tree and in the VAST tree.

A few examples will provide an intuitive appreciation for the scoring-function: a set of four proteins in which three are found to cluster together but one is distant has a score of 2/3. Similarly, a set of three proteins in which two cluster but one does not has a score of 1/3 when the outlier is distant, and a score that approaches 5/6 as the outlier becomes increasingly proximate. A set that is split into two perfect but distant subclusters of equal size has a score of ~1/4. Typically, a random set of proteins has scores below 0.1.

Of the 38 extended families with more than one member (Table 2), 23 are found to have perfect scores, while five have scores below 0.3 (essentially random). Among these latter five, only two (that is, bacterial AB5 toxins and long-chain scorpion toxins) have significantly higher scores on the VAST tree. Ten extended families scored above 0.3 but did not achieve a perfect score. Among these, only one (thioredoxin-like) achieved a perfect score on the VAST tree and two more (G proteins and plant virus extended family in viral coat and capsid proteins) were considerably better. At the SCOP fold level, 18 folds are represented by at least two SCOP families; 10 of these score above 0.3. A similar number is found using VAST, although it should be noted that two large SCOP families (TIM barrels and viral coat and capsid proteins) have perfect scores in the VAST tree but not in our tree. Conversely, a number of SCOP categories have higher scores in our tree than in the VAST tree. However, a denser population of proteins, with additional intervening outliners, would lead to diminished scores in our tree.

It is important to assess the degree to which our results are sensitive to minor changes in the scoring function, secondary-structure definitions or tree construction algorithm. To address the first question, all experiments were repeated with variant scoring functions that introduce a gap penalty or change the penalty for aligning a loop and a nonloop, and so on. Although the resulting trees where not identical, the clusters are little changed within the same category.

To assess the tree construction algorithm, a library of similar algorithms was tried. Again, results were little changed within the same category. No one algorithm recommended itself above others, and we could see no rationale for abandoning our policy of using the simplest possible approach.

As a final control, a tree was constructed (Fig. 3) from sequence-based distances that were obtained using Clustal W[33], with tree branches drawn proportional to their lengths, using the neighbor-joining (NJ) method[34]. It is apparent at a glance that residue sequences do not underwrite a useful taxonomy.

Discussion

We address two related questions: whether secondary structure determines the three-dimensional fold and, if so, whether it can be used to classify proteins automatically. This initial attempt adopts a simple (and perhaps simplistic) approach to secondary-structure identification based primarily on backbone dihedral angles, reasoning that more information can be added later, if warranted.

We anticipate that a satisfactory protein taxonomy, like any scheme to classify a body of complex data, will be attained gradually, in conceptual increments. For animals and molecules[35] alike, recognition of common features gives rise to nomenclature, which can lead, in turn, to the perception of ordered classes. Ideally, one seeks a parsimonious description that spans the phenomena of interest, followed by a theory that can generate this description from first principles. The periodic table of elements is a well-known and successful example of this two-pronged approach. More complex examples that include syntactic ambiguity are represented by natural languages and their grammars.

Here, the hypothesis that secondary structure implies tertiary structure is assessed as a generating theory for the SCOP taxonomy. In that SCOP includes evolutionary information, a successful generator would have implications for the evolution of protein structure. SCOP is not the only available taxonomy for protein molecules. Other notable examples include CATH[36] and DALI[37].

False positives (that is, incorrectly grouped proteins) serve as a stringent test of the hypothesis that secondary structure determines tertiary structure; few are found despite the simplicity of our approach. If secondary structure can successfully differentiate among protein folds, then sequences of secondary-structure elements will correspond to unique substructures (folding units) at some level. A natural extension of our analysis would be to classify such secondary structure-based folding units. The existence of such assemblies has been postulated in the literature repeatedly[32,38,39].

False negatives (that is, failure to group similar folds) occur more frequently, but largely for technical reasons: loops identified as strands and *vice versa*, or long helices that are mistakenly subdivided. Of course, false negatives will also arise in cases where SCOP classifies substantially different structures into the same family (for example, OB folds and EF hands). Our dihedral angle-based algorithm for secondary-structure identification has the advantage of simplicity, but it can lack discrimination, resulting in the false negatives just noted. It is anticipated that our scoring algorithm can be extended to distinguish between parallel and antiparallel strands or between strands and extended loops. Related information could also be added without a significant increase in the complexity of the algorithm. The experiment described here demonstrates that such extensions are likely to be worthwhile.

Our quantitative scoring-function merits discussion. Ideally, a numerical score would not be needed. Instead, only the organizing hypothesis that secondary structure determines tertiary structure would be used (as in Fig. 1) to route each protein to its most suitable locale within the tree (like a hierarchic bubble sort). Such a procedure has the distinct advantage of avoiding the bias that can result from the use of a numerical score, where it is the investigator, not the protein, who establishes a minimum threshold of similarity. However, hierarchic sorting suffers from the flaw that quite dissimilar proteins can sort to proximate locations in an underpopulated tree. A uniformly dense protein distribution would overcome this problem, but there is no *a priori* way to select such a distribution. Consequently, we invented the conservative scoring function used in Table 2.

Table 2*b* Cluster scores

Fold type / Extended family	Number of Members	Members	Score	VAST score[1]
Phospholipase A2	**3**	**32, 176, 177**	**0.33**	**–**
Vertebrate phospholipase A2	2	32, 176	1	1
Insect phospholipase A2	1	177		
Ferritin-like				
Ferritin	2	34–35	1	1
Immunoglobulin-like β-sandwich	**9**	**43–48, 50–53**	**0.31**	**0.47**
V-set domains	4	43–46	0.59	0.38
C2-set domains	1	47		
C1-set domains	1	48		
I-set domains	1	50		
Fibronectin type III	1	51		
Cu, Zn superoxide dismutase	1	52		
Actinoxantin-like	1	53		
ConA-like lectins/glucanases	**4**	**54–57**	**0.27**	**0.17**
Legume lectins	1	54		
Bacillus 1-3,1-4-β-glucanase	1	55		
Galectin	2	56–57	0.01	0
Viral coat and capsid proteins	**9**	**58–66**	**0.59**	**1**
Animal virus proteins	5	56–62	0.66	0.85
Plant virus proteins	4	63–66	0.38	0.78
Cupredoxins				
Plastocyanin/azurin-like	4	67–70	0.49	1
Lipocalins	**7**	**74–80**	**0.71**	**0.85**
Fatty acid binding	4	74–77	1	1
Retinol binding	3	78–80	1	1
Trypsin-like serine proteases	**2**	**81–82**	**0**	**0**
Prokaryotic proteases	1	81		
Eukaryotic proteases	1	82		
Acid proteases				
Retroviral proteases	2	83–84	1	1
Streptavidin-like				
Avidin/streptavidin	2	85–86	1	1
Sialidases				
Sialidases	3	88–90	1	1
Snake toxin-like				
Snake venom toxins	3	92–94	0.4	0.33
Defensin-like				
Defensin-like	2	95–96	0.75	–
Knottins				
EGF-type module	2	97–98	0	–
Cystine-knot cytokines	**2**	**101–102**	**0.75**	**–**
Platelet-derived growth factor	1	101		
Transforming growth factor	1	102		
Membrane				
Porins	2	103–104	1	1
OB-fold	**5**	**87, 180–183**	**0.04**	**0.5**
Bacteriophage ssDNA-binding	1	87		
Inorganic pyrophosphatase	1	180		
Staphylococcal nuclease	1	181		
Bacterial AB5 toxins, B-subunits	2	182–183	0.07	1

In large part, our computational experiments support the conjecture that secondary structure determines the tertiary fold, and consequently, aligned elements of secondary structure can be an effective basis for protein classification. However, the resulting hierarchy is not identical to the SCOP classification, as seen in Fig. 1, where categories are circled to facilitate visual comparison. It should be noted that differences between SCOP and our clusters often involve sets of proteins that do not cluster successfully using VAST scores. Major discrepancies occur in the α+β and α/β proteins, which are intermixed. Perhaps, division into these classes should be reassessed.

The mechanism of protein evolution is a topic of keen interest. Over time, protein sequences can vary substantially, while many seemingly disparate attributes (such as those linked to function, foldability, recognition and regulation) remain highly conserved. Although the SCOP hierarchy incorporates this kind of conserved

Table 2*c* Cluster scores

Fold type / Extended family	Number of members	Members	Score	VAST score[1]
(TIM)-barrel	**7**	**107–113**	**0.32**	**1**
Xylose isomerase	1	107		
Triosephosphate isomerase	1	108		
Type II chitinase	1	109		
Aldolases	1	110		
FMN-linked oxidoreductases	1	111		
Muconate lactonizing enzyme-like	2	112–113	1	0.5
Flavodoxin-like	**3**	**114–116**	**1**	**0.5**
Flavodoxin	2	114–115	1	–
CheY-like	1	116		
Isocitrate and isopropylmalate dehydrogenases				
Isocitrate and isopropylmalate dehydrogenases	2	120–121	1	1
Thioredoxin-like				
Thioltransferase	4	122–125	0.44	1
P-loop	**5**	**126–130**	**0.28**	**0.7**
Nitrogenase iron	1	126		
G proteins	2	127–128	0.38	1
Nucleotide and nucleoside kinase	2	129–130	1	1
Dihydrofolated reductases				
Dihydrofolate reductases	3	131–133	1	1
Ferredoxin-like	**5**	**134–138**	**0.18**	**0.1**
Pancreatic carboxypeptidase	1	134		
Viral DNA-binding	1	135		
Short-chain ferredoxins	1	136		
Nucleoside diphosphate kinases	1	137		
Ribulose 1,5-bisphosphate carboxylase-oxygenase	1	138		
β-grasp	**3**	**139–141**	**0.21**	**0.33**
Immunoglobulin-binding domains	1	139		
2Fe-2S ferredoxin	1	140		
Ubiquitin	1	141		
SH2-like				
SH2 domain	2	143–144	1	0
Ribonuclease H-like motif				
Ribonuclease H-like	2	147–148	1	1
Ribonuclease A-like				
Ribonuclease A-like	2	149–150	1	1
Microbial ribonucleases				
Microbial ribonucleases	2	159–161	1	0.33
Knottins	**4**	**166–169**	**0**	**0.33**
Agglutinin (lectin) domain	2	166–167	0	–
Long-chain scorpion toxins	2	168–169	0	1
Glucocorticoid receptor-like (DNA-binding)	**2**	**172, 36**	**0.75**	**–**
Erythroid transcription factor GATA-1	1	172		
Hormone receptor	1	36		
CoA-dependent acetylotransferases				
CoA-dependent acetylotransferases	2	173–174	1	1

[1]Scores obtained using the December 1997 version of VAST with a conservative cutoff value. Without the cutoff value, the VAST scores are likely to be higher. Therefore, these scores should be viewed as a measure of similarity between members of a given category, but not as a measure of the quality of VAST clustering. Since we collected our data, a newer version of VAST has been made available, which would likely lead to different scores. A dash indicates that the data are insufficient for evaluation of the category.

evolutionary information, our tree does not. Yet, the two trees are similar. Why?

Protein structure is far better conserved than protein sequence. If a protein's secondary structure determines its three-dimensional fold, as our results suggest, then this ordered sequence of secondary-structure elements is also conserved. The information that engenders these secondary-structure elements is widely dispersed through the chain[40,41].

Implicit in such an organization is the conclusion that the universe of protein folds is limited. The argument is as follows[40]. Proteins are large molecules that enclose a solvent-shielded interior, within which hydrophobic groups can be sequestered. Shielded

hydrophobic side chains are covalently attached to the polar backbone, which is also shielded in most cases. Were this backbone unable to form hydrogen bonds within the molecular interior, then hydrogen bonding would push the conformational equilibrium far toward the unfolded state, where hydrogen bonds to solvent water could be realized. Consistent with this idea, almost all backbone groups within the interior of proteins of known structure are found to be hydrogen-bonded. There are only two structures that provide ubiquitous hydrogen bonding for interior residues: α-helix and β-sheet. While other interior structures are found occasionally, they do not lend themselves readily to routine hydrogen bonding. Thus, protein domains will be composed of helix, sheet and the tight turns that interconnect them. A typical domain of 100 residues contains ~10 elements of regular secondary structure — either α or β. Therefore, the total number of protein domains is, at most, on the order 2^{10} = 1,024, multiplied by the number of possibilities at each turn position.

The preceding argument, while approximate, demonstrates that protein evolution is not open-ended. From the outset, chemistry, in promoting structure, predetermines the universe of conceivable folds for polypeptide chains in an aqueous environment. Protein evolution can only populate folds within this realm, not invent novel ones. In contrast, function is not subject to such limitations. A protein's function is due to a comparatively small number of residues, suitably interspersed throughout the sequence. This process of imbedding functional residues within a robust framework constitutes a versatile mechanism to confer multiple functions upon a given fold. An especially compelling example, involving a family of enzymes involved in nucleotide metabolism, is described by Holm and Sander[42].

Methods

The method is entirely straightforward and completely automatic. A representative set of proteins was chosen from the Protein Data Bank[31]. Specifically, we screened PDB_select (August 1996)[30] to choose all proteins with resolutions of 2.5 Å or better and sequence identity of <30%. Of these, entries with <250 residues were identified and winnowed to produce a set of single-domain proteins. For each protein in this set, all residues are classified into secondary-structure type (helix, strand or loop), after which that protein is represented as an ordered sequence of secondary-structure elements, each of a given length. Next, a pairwise similarity score is computed for all such sequences, using a dynamic programming algorithm. Finally, these scores are arrayed in a similarity matrix, and a clustering algorithm is applied, resulting in a tree. The tree gives a taxonomic organization of proteins in the chosen set, and it is suitable for comparison with corresponding trees from other methods.

Each protein is represented by an ordered sequence of its secondary-structure elements, together with their lengths, termed the 'ss string'. Secondary-structure classification is based on backbone dihedral angles, φ and ψ, as follows:

Helix: Five or more consecutive residues with backbone dihedral angles in the range φ, ψ = (-60 ± 20°, -35 ± 30°). Solitary outliers in the range φ, ψ = (-75 ± 35°, -37 ± 37°) are allowed.

Strand: Three or more consecutive residues with backbone dihedral angles in the range φ, ψ = (-120 ± 60°, +120 ± 60°) or (-120 ± 60°, -160 ± 20°).

Loop: Residues not classified as either helix or strand default to loop. Initial and final loops are ignored.

A pairwise similarity score in the range [0, 1] is computed between all ss strings. Akin to conventional sequence alignment, a dynamic programming algorithm is used[19], except that in this case each string element corresponds to an entire segment of secondary structure, not just a single residue. (Thus, the ss string of a protein is much shorter than its residue sequence.) In detail, alignment between two elements of lengths l_i and l_j are scored as follows:

Alignment between	Score
Identical elements	$\min(l_i, l_j)$
Helix/strand and loop	$0.5\min(l_i, l_j)$
Helix and strand	0

where min(x,y) is the minimum of x or y. Scores are summed over all elements, and the total is then normalized by the mean sequence length of the two proteins. Neither helix nor strand elements can be split for alignment with multiple helices or strands (for example, a long helix cannot be aligned with two shorter helices). However, both helices and strands can be split for alignment with loops, and loop elements can be split into either two or three smaller loops. (Thus, a loop is treated like a 'wild card' for alignment.) No explicit gap penalty is used. However, given that the algorithm aligns entire secondary-structure elements — not just single residues — the implicit penalty for a gap is considerable. This alignment procedure is inherently global.

As a clarifying example, consider two secondary sequences: S_1 = h7, l3, h7 and S_2 = h5, l6, h6. For optimal alignment, l6 in the second sequence is split into l2, l3, l1. Then, h7 is aligned with h5 + l2, l3 with l3, and the second h7 is aligned with l1 + h6. The total score is: 5 + 0.5*2 + 3 + 0.5*1 + 6 = 15.5, and the normalized score is 0.91.

This alignment algorithm has been applied to each pair of proteins in the data set to obtain the matrix of similarity scores used to produce the trees shown in Figs. 1–3. We have kept the algorithm simple intentionally, reasoning that further refinement can always be added at a later stage, if initial results are sufficiently promising.

A clustering algorithm is applied to the similarity matrix. The 'distance' between two ss strings is defined as 1 - (their similarity score). (This is not a distance in the strictest sense because the triangle inequality need not be satisfied.) Clustering is achieved by constructing a 'similarity tree' using one of the several available tree construction algorithms.

We experimented with trees obtained using two such algorithms: the weighted pair group method (WPG)[43] and the neighbor-joining method (NJ)[34]. In both, the basic idea is to cluster pairs having the smallest distance, iteratively. At each iteration, a new element, corresponding to the pair, is introduced, and the original two elements are removed. Then, the distance is computed between this new element and all other elements in the set. The WPG and NJ methods differ in how this distance is computed.

In greater detail, given an n × n similarity matrix, S, the pair of elements i,j with the smallest distance d(i,j) is determined. This pair (i,j) is then removed from S and a new element (u) is introduced, where u corresponds to the newly created cluster. Then, d(u,k) is computed for each entry k in S. A simple unweighted distance can be taken as an average between the paired cluster and each k: d(u,k) = (d(i,k) + d(j,k)) / 2. However, this unweighted distance does not account for the multiplicity of clustered elements represented by each matrix element. (Initially, each element in S corresponds to one ss string, with multiplicity of unity.) To weight these distances, let n(i) be the number of ss strings represented by the i^{th} element. In the WPG method, d(u,k) is calculated as d(u,k) = (n(i)d(i,k) + n(j)d(j,k)/(n(i)+ n(j)), while in the NJ algorithm d(u,k) = (d(i,k) + d(j,k) - d(i,j) / 2. In the WPG method, distance is simply a number-weighted average between elements in the two clusters, and it is not intended to define branch lengths in the resulting tree. In the NJ method, the distance between two proteins is approximated by their summed branch lengths.

The cluster-scoring function was developed to quantify comparison between a cluster in our tree and a corresponding SCOP category. The function is intended to capture the tree topology. For a set, A, and two proteins a, b ∈ A, we measure a 'topological distance', td(a,b), between a and b in the tree. A numerical cutoff was included such that when the distance between the corre-

sponding ss strings is large enough (computed from sequence alignment), the topological distance is set to zero. Assume that A has n members and thus n(n – 1) / 2 protein pairs. The clustering score is defined as the summed topological distances between all pairs of members divided by the number of such pairs:

$$\frac{2\sum_{a,b\in A} td(a,b)}{n(n-1)}$$

The topological distance between two proteins is defined by a recursive procedure. In it, the tree can be viewed as rooted at any internal white internal node. Initially, each leaf ∈ A is assigned a weight of 1, and each leaf ∉ A is assigned a weight of 0. Then, recursively, the weight of an internal node is reckoned as the average of the weight of its children. Ultimately, td(a,b) equals the weight of the lowest common ancestor of a and b. If this lowest common ancestor is a white node, then the td(a,b) is set to 0. For example, the score for set {11,12,13,14} in Fig. 1 is ((3 × 0) + (3 × 1)) / 6 = 0.5, a stiff penalty for displacement of protein 13. The score of set {43,44,45,46} is only slightly higher: ((3 × 1) + (3 × 1/16)) / 6 = 0.53. This scoring function is quite conservative, and scores exceeding 0.5 imply substantial clustering.

Acknowledgments

We thank R. Srinivasan, V. Murthy and P. Thiessen for helpful suggestions, and J. Cohen for providing access to his tree-construction program, Tande. We are particularly indebted to an anonymous referee for assistance in bringing this paper to fruition. Supported by the Sloan Foundation (T.P.) and the NIH (G.D.R.).

Received 24 July, 1998; accepted 29 March, 1999.

1. Minor, D.L. Jr. & Kim, P.S. Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380**, 730–734 (1996).
2. Itahaki, L.S., Otzen, D.E. & Fersht, A.R. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation–condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288 (1995).
3. Shao, X. & Matthews, C.R. Single-tryptophan mutants of monomeric tryptophan repressor: optical spectroscopy reveals nonnative structure in a model for an early folding intermediate. *Biochemistry* **37**, 7850–7858 (1998).
4. Clark, P.L., Liu, Z.-P., Rizo, J. & Gierasch, L.M. Cavity formation before stable hydrogen bonding in the folding of a beta-clam protein. *Nature Struct. Biol.* **4**, 883–886 (1997).
5. Yee, D.P., Chan, H.S., Havel, T.F. & Dill, K.A. Does compactness induce secondary structure in proteins? A study of poly-alanine chains computed by distance geometry. *J. Mol. Biol.* **241**, 557–573 (1994).
6. Havel, T.F., Crippen, G.M. & Kuntz, I.D. Effects of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions. *Biopolymers* **18**, 73–81 (1979).
7. Reymond, M.T., Merutka, G., Dyson, H.J. & Wright, P.E. Folding propensities of peptide fragments of myoglobin. *Protein Sci.* **6**, 706–716 (1997).
8. Dyson, H.J. *et al.* Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding II. Plastocyanin. *J. Mol. Biol.* **226**, 819–835 (1992).
9. Srinivasan, R. & Rose, G.D. LINUS—a simple algorithm to predict the fold of a protein. *Proteins Struct. Funct. Genet.* **22**, 81–99 (1995).
10. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
11. Madej, T., Gibrat, J-F. & Bryant, S.H. Threading a database of protein cores. *Proteins Struct. Funct. Genet.* **23**, 356–369 (1995).
12. Mitchell, E.M., Artymiuk, P.J., Rice, D.W. & Willett, P. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* **212**, 151–166 (1990).
13. Di Francesco, V., Garnier, J. & Munson, P.J. Protein topology recognition from secondary structure sequences: application of the hidden markov models to the alpha class proteins. *J. Mol. Biol.* **267**, 446–463 (1997).
14. Russell, R.B., Copley, R.R. & Barton, G.J. Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259**, 349–365 (1996).
15. Rost, B., Schneider, R. & Sander, C. Protein fold recognition by prediction-based threading. *J Mol Biol* **270**, 471–480 (1997).
16. Rice, D.W. & Eisenberg, D. A 3D–1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.* **267**, 1026–1038 (1997).
17. Aurora, R. & Rose, G.D. Seeking an ancient enzyme in *Methanococcus jannaschii* using ORF, a program based on predicted secondary structure comparisons. *Proc. Natl. Acad. Sci. USA* **95**, 2818–2823 (1998).
18. Holm, L. & Sander, C. Mapping the protein universe. *Science* **273**, 595–603 (1996).
19. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
20. Sander, C. & Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct. Funct. Genet.* **9**, 56–68 (1991).
21. Doolittle, R.F. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287–314 (1995).
22. Doolittle, R.F. *Of Urfs and Orfs* 1-1-103 (University Science Books, Sausalito, California; 1986).
23. Altschul, S.F., Boguski, M.S., Gish, W. & Wootton, J.C. Issues in searching molecular sequence databases. *Nat. Genet.* **6**, 119–129 (1994).
24. Smith, H.O., Annau, T.M. & Chandrasegaran, S. Finding sequence motifs in groups of functionally related proteins. *Proc Natl Acad Sci USA* **87**, 826–830 (1990).
25. Lipman, D.J. & Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441 (1985).
26. Neuwald, A.F., Liu, J.S., Lipman, D.J. & Lawrence, C.E. Extracting protein alignment models from the sequence database. *Nucleic Acids Res.* **25**, 1665–1677 (1997).
27. Henikoff, S. & Henikoff, J.G. Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.* **6**, 698–705 (1997).
28. Luthy, R., Bowie, J.U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83–85 (1992).
29. Gibrat, J-F., Madej, T. & Bryant, S.H. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377–385 (1996).
30. Hobohm, U. & Sander, C. Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524 (1994).
31. Bernstein, F.C. *et al.* The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542 (1977).
32. Levitt, M. & Chothia, C. Structural patterns in globular proteins. *Nature* **261**, 552–558 (1976).
33. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
34. Saitou, N. & Nei, M. The neighborhood-joining method: a new method for reconstructing phylogenic trees. *Mol. Biol. Evol.* **4**, 406–424 (1987).
35. Richardson, J.S. The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.* **34**, 168–340 (1981).
36. Orengo, C.A., Michie, A.D., Jones, D.T., Swindells, M.B. & Thornton, J.M. CATH—a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
37. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
38. King, J. Genetic analysis of protein folding pathways. *Biotechnology* **4**, 297–303 (1986).
39. Lattman, E.E. & Rose, G.D. Protein folding — what's the question? *Proc. Natl. Acad. Sci. USA* **90**, 439–441 (1993).
40. Aurora, R., Creamer, T.P., Srinivasan, R. & Rose, G.D. Local interactions in protein folding: lessons from the α-helix. *J. Biol. Chem.* **272**, 1413–1416 (1997).
41. Baldwin, R.L. & Rose, G.D. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* **24**, 26–33 (1999).
42. Holm, L. & Sander, C. An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins Struct. Funct. Genet.* **28**, 72–82 (1997).
43. Waterman, M.S. *Introduction to computational biology: maps, sequences, and genomes* (Chapman & Hall, London;1995).
44. Cohen, J. & Farach, M. In *Proc. of eighth ann. ACM–SIAM symp. on discrete algorithms.* (Association for Computing Machinery, New York; 410–416; 1997).