doi:10.1016/j.jmb.2006.03.017 *J. Mol. Biol.* (2006) **359**, 215–237

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Structural Classification of Small, Disulfide-rich Protein Domains

Sara Cheek[1], S. Sri Krishna[2] and Nick V. Grishin[1,3]*

[1]*Department of Biochemistry*
University of Texas
Southwestern Medical Center
5323 Harry Hines Blvd., Dallas
TX 75390, USA

[2]*Joint Center for Structural*
Genomics, University of
California, San Diego, La Jolla
CA 92093-0314, USA

[3]*Howard Hughes Medical*
Institute, University of Texas
Southwestern Medical Center
5323 Harry Hines Blvd., Dallas
TX 75390, USA

**Corresponding author*

Disulfide-rich domains are small protein domains whose global folds are stabilized primarily by the formation of disulfide bonds and, to a much lesser extent, by secondary structure and hydrophobic interactions. Disulfide-rich domains perform a wide variety of roles functioning as growth factors, toxins, enzyme inhibitors, hormones, pheromones, allergens, etc. These domains are commonly found both as independent (single-domain) proteins and as domains within larger polypeptides. Here, we present a comprehensive structural classification of approximately 3000 small, disulfide-rich protein domains. We find that these domains can be arranged into 41 fold groups on the basis of structural similarity. Our fold groups, which describe broader structural relationships than existing groupings of these domains, bring together representatives with previously unacknowledged similarities; 18 of the 41 fold groups include domains from several SCOP folds. Within the fold groups, the domains are assembled into families of homologs. We define 98 families of disulfide-rich domains, some of which include newly detected homologs, particularly among knottin-like domains. On the basis of this classification, we have examined cases of convergent and divergent evolution of functions performed by disulfide-rich proteins. Disulfide bonding patterns in these domains are also evaluated. Reducible disulfide bonding patterns are much less frequent, while symmetric disulfide bonding patterns are more common than expected from random considerations. Examples of variations in disulfide bonding patterns found within families and fold groups are discussed.

Keywords: protein classification; disulfides; small proteins; homology; fold

Introduction

The structures of very small proteins often lack an extensive hydrophobic core and possess secondary structure elements that are small and irregular. These proteins are generally stabilized either by binding a metal ion, most commonly, zinc,[1] or by the formation of disulfide bonds. Disulfide bonds have traditionally been presumed to stabilize protein structures by reducing the conformational freedom of the protein in the unfolded state, therefore reducing the entropy of the unfolded state relative to the folded state.[2–4] Another theory proposes that the stabilizing influence of these cross-links is enthalpic, whereby the presence of the disulfide bonds destabilize the denatured form of the protein by sterically inhibiting certain potential hydrogen bonding groups from forming satisfied donor–acceptor pairs.[5] It has also been suggested that both entropic and enthalpic effects contribute to the stabilizing capacity of disulfide bonds.[6] Although these cross-links are, in most cases, responsible mainly for maintaining the proper fold of the protein and are, therefore, only indirectly essential for protein function, there are also examples in which reduction or oxidation of these bonds alters protein activity.[7,8]

Small protein domains in which disulfide bonds form the scaffold of the protein are often referred to as disulfide-rich. We describe a typical disulfide-rich domain by the following characteristics: small (usually <100 residues), lacking an extensive

Abbreviations used: BPTI, bovine pancreatic trypsin inhibitor; PDB, Protein Data Bank; BBIs, Bowman–Birk inhibitors; TSP, thrombospondin; SMB, somatomedin B; PAI-1, plasminogen activator inhibitor type-1; TAP, tick anticoagulant protein.

E-mail address of the corresponding author: grishin@chop.swmed.edu hydrophobic core, having few secondary structure elements, and fold stabilization primarily due to two or more disulfide bonds in close proximity. These proteins encompass a wide variety of functions, such as growth factors, toxins, enzyme inhibitors, and structural or ligand-binding domains within larger polypeptides. Several classes of disulfide-rich proteins, such as insulin and related growth factors or ion channel-inhibiting toxins, have been of interest to researchers for medical reasons. Other disulfide-rich proteins have been the focus of folding experiments, with bovine pancreatic trypsin inhibitor (BPTI) being the most thoroughly studied example.[9,10] These folds have also been proposed as scaffolds for drug design,[11,12] and mimetics of protein-interacting surfaces.[13]

Protein classification on the basis of structural similarity and evolutionary relatedness is a common means of organizing biological data for the purpose of studying various aspects of sequence/structure/function relationships in proteins, such as structure prediction or identification of functionally important residues. Evolutionary and structural neighbors of large (>100 residues), globular proteins can often be identified using popular sequence and structure comparison tools such as PSI-BLAST,[14] and Dali.[15] However, automatic methods generally tend to be unreliable for small proteins, due to the shortness of these polypeptide chains. Classification of small protein domains is consequently a non-trivial task and one that frequently requires considerable manual analysis.

Classification schemes for disulfide-rich domains have been constructed using automated tools that compare the geometry and topology of disulfide bonds. The KNOT-MATCH program clusters proteins on the basis of the structural superposition of the disulfide bonds.[16,17] Another approach classifies proteins according to their “disulfide signature”, which considers disulfide connectivity and the loop lengths between cysteine residues.[18,19] However, the evolutionary relatedness among protein groupings identified by these approaches must be interpreted carefully, as these methods do not address established indicators of homology or biologically relevant factors, such as sequence similarity, protein function, fold topology, or other structural features beyond disulfide bonding patterns. A number of other studies have examined specific subsets of disulfide-rich domains, focusing on a particular family (e.g. toxins from snails[20] or spiders[21]), structural motif (e.g. the KNOTTIN website[22]), or function (e.g. protease inhibitors; MEROPS[23]). Although nearly all disulfide-rich domains are included in the comprehensive SCOP[24] (structural classification of proteins) database, this is not a convenient tool for studying this group of proteins as a whole, because the disulfide-rich domains are distributed among several structural classes (small proteins, all-α proteins, peptides, etc.).

In order to understand the structural and functional diversity among all available small disulfide-rich proteins, we have performed a comprehensive classification of these domains. The hierarchy of this classification is comprised of two levels, such that the disulfide-rich domains are evaluated in terms of both their structural and evolutionary relatedness. On the basis of this survey, we examine the variety of structural folds adopted by disulfide-rich domains, and describe the distant homology between previously unlinked domains. Disulfide bonding patterns among these domains are evaluated, and we identify examples of convergent and divergent evolution of functions performed by these proteins. This classification should be useful for studying the evolution of the folds and functions of disulfide-rich domains in general, as well as for investigating the structural and evolutionary neighbors of specific disulfide-rich proteins in particular.

Results and Discussion

Results of the disulfide-rich domain classification

Structures of 2945 small disulfide-rich protein domains were detected in the RCSB Protein Data Bank (PDB) as described in Materials and Methods. These domains are found in 2578 individual PDB chains from 1596 PDB structures. However, there is a high degree of redundancy within this set due to identical chains within one PDB structure or multiple structures of the same protein. Upon clustering the sequences of these 2945 domains at 95% identity with 95% length coverage, the number of representatives is reduced to 963 domains. Although the “unique” representatives comprise only ~33% of the original set, a similar reduction is not achieved by further decreasing the identity among clusters: clustering at 50% identity with 95% coverage results in 696 disulfide-rich domains (~24% of the original set). The protein domains in this classification are an average of 57(±29) residues in length and contain an average of 3(±1) disulfide bonds. Most of these domains (>96%) are from eukaryotic organisms.

Disulfide-rich domains are classified into fold groups and families

The 2945 disulfide-rich protein domains are arranged into 41 fold groups according to structural similarity (Table 1). Domains within the same fold group share a common structural core comprised of secondary structure elements found in the same spatial arrangement with topology that is either identical or related by circular permutation. One objective of this study was to bring together disulfide-rich domains whose structural similarities were previously unappreciated. Thus, the degree of structural similarity described by the fold group

Table 1. Small disulfide-rich protein domains in fold groups and families

			No. members		
Fold group	Common structural core	Families	All domains	95% identity	Representative
1	Small, distorted α-hairpin	κ-Hefutoxin-like	4	3	1hp9, A1-A22
		Immunodominant domain of attachment protein G	3	2	1brv, 171-189
		Endothelin-like	8	6	1srb, 1-21
		Integrin αVβ3 subdomain	4	2	1jv2, B601-B636
		Cellulase subdomain	44	6	1a39, 41-74
		IgE receptor antagonist	6	2	1kcn, A1-A21
2	α-Hairpin	Cytochrome *c* oxidase, subunit VIb	14	2	1ocr, H7-H85
		Cytochrome *bc*1 complex, non-heme 11 kDa protein ("hinge")	18	5	1bcc, H13-H78
		Cytochrome *c* oxidase copper chaperone	1	1	1z2g, A1-A69
		Enterotoxin B	1	1	1ehs, 1-48
		Ole e 6 pollen allergen	1	1	1ss3, A1-A50
		Attractin	1	1	1t50, A1-A58
		Neurotoxin B-IV	1	1	1vib, 1-55
		Vanabin 2	1	1	1vfi, A1-A95
3	Three-helix bundle, right-handed	Protozoan pheromones, ER-1-like	6	5	1erp, 1-38
		Anaphylotoxin C5a	3	3	1cfa, A1-A71
		P8-MTCP1	3	2	1hp8, 1-68
		Notch/DSL/LNR domain	1	1	1pb5, A1-A35
		Sea anemone toxin K	4	3	1bgk, 1-37
		CRISP family, helical bundle subdomain	5	3	1rc9, A181-A221
4	Three-helix bundle, left-handed	Insulin-like	242	15	7ins, B1-B30, A1-A21
		Helical subdomain of serine carboxypeptidase-like	5	2	1cpy, 181-252
		Molt-inhibiting hormone	1	1	1j0t, A0-A77
5	Three-helix irregular bundle with disulfide bonds to N and C-terminal extensions	Frizzled family	8	2	1ijx, D2-D123
6	Four small α-helices, non-globular array	Domain II of osmotin-like family	18	4	1aun, 129-177
7	Five-helix globular array I	Protozoan pheromone, ER-23	1	1	1ha8, A1-A51
8	Five-helix globular array II	Tetraspanin family ectodomain	4	1	1g8q, A113-A202
9	Five-helix "hollow" array	Elicitins	6	2	1bxm, -1-98
		GFRα1 domain 3	1	1	1q8d, A239-A346
10	Two antiparallel disulfide-linked α-helices and a Ca^{2+}-binding loop	Phospholipase A2	271	52	1hn4, A-5-A124
11	β-Hairpin	Antimicrobial β-hairpin	11	10	1hvz, A1-A18
		Arylsulfatase, β-hairpin subdomain	7	1	1auk, 151-177
		Subdomain of Fr-MLV envelope glycoprotein receptor-binding domain	1	1	1aol, 67-95
12	Three-strand β-sheet, antiparallel, strand order 123	Locust serine protease inhibitors	10	8	1pmc, 1-36
		Fibronectin type I module	13	8	1qgb, A61-A109
		Midkine	2	2	1mkn, A1-A59
		Thrombospondin type I repeat	4	4	1lsl, A416-A472
		Hormone binding domain of CRF receptor	1	1	1u34, A15-A133
		β-Microseminoprotein, N-terminal domain	1	1	1xhh, A1-A49
13	Three-strand β-sheet, antiparallel, strand order 132	Mammal defensin-like/sea anemone toxin-like	57	22	1dfn, A2-A31
		Bowman–Birk inhibitor/bromelain inhibitor	38	25	1bbi, 25-51
		Amb V ragweed allergen	3	1	1bbg, 1-40

(*continued on next page*)

Table 1 (*continued*)

Fold group	Common structural core	Families	No. members		Representative
			All domains	95% identity	
14	Three-strand β-sheet, mixed, strand order 132	Domain III of malarial parasite apical membrane antigen I	2	2	1w8k, A387-A453
		anti-HIV peptide RP 71955	2	1	1rpb, 1-21
		Chordin-like cysteine-rich repeat	1	1	1u5m, A44-A73
		IGFBP family, N-terminal domain	1	1	1wqj, B3-B39
15	Three-strand β-sheet, mixed, strand order 312	Crambin-like (α-hairpin inserted in β-sheet)	24	9	1cbn, 1-46
		Fungal pathogen protein NIP1	2	2	1kg1, A29-A60
16	Three-"strand" bundle; two strands form β-hairpin	TNF receptor family repeats	147	32	1d0g, T102-T130
		Vascular endothelial growth factor	6	2	1vgh, 1-27
17	Two parallel β-hairpins	Domain III of osmotin-like family	18	5	1aun, 49-80
18	Two perpendicular hairpins	Cellulose binding/docking domains	4	2	1e8r, A20-A69
19	Five β-strands in two parallel layers; each layer is antiparallel	CCP modules/SCR domains/Sushi domains	170	37	1g40, B65-B125
20	Five β-strands in two perpendicular layers; each layer is antiparallel	Kringle-like/fibronectin type II module	99	31	1ks0, A1-A59
21	Interconnected 3-"strand" subdomains	Disintegrins	15	9	1fvl, 1-70
22	Irregular β-sandwich	Methylamine dehydrogenase, L chain	16	2	2bbk, L7-L131
23	Knottin-like I: Four cysteine residues forming disulfide crossover are located on four structure elements; elements 1-2-3 with right-handed connection, 2-3-4 with left-handed connection	Spider toxin/ω-conotoxin/IGFBP knottin-like domain/VHv1.1 viral protein subdomain/plant enzyme inhibitor/gurmarin/agouti-like	111	75	1lmm, A1-A40
		Scorpion toxin-like/insect and plant defensin-like	124	82	1agt, 1-38
		Kalata-like cyclotides	14	12	1kal, 1-29
		Cellulose-binding domain of cellobiohydrolase I	6	1	1az6, 1-36
		Satiety factor CART, subdomain	1	1	1hy9, A49-A89
		Plant lectin-like	109	19	1hev, 1-43
		Colipase-like	16	7	1lpa, 42-90
		Cysteine knot cytokines	125	35	1aoc, A83-A175
24	Knottin-like II: Four cysteine residues forming disulfide crossover are located on four structure elements; left-handed connections	Snake toxin-like	158	44	1idg, A1-A74
		Leech serine protease inhibitor-like	26	11	1c9t, J7-J59
		Granulin-like repeat, N-terminal domain	5	4	1fwo, A1-A35
25	Knottin-like III: first two of four cysteine residues forming disulfide crossover are located on one irregular/bulging structure element	EGF-like	286	88	1nub, A53-A78
		Cysteine-rich repeats of EGF receptor family ectodomain	153	65	1s78, A171-A192
		CRISP family, knottin-like subdomain	6	3	1rc9, A166-A180
		DPY module	1	1	1oig, A1-A24
		Elafin-like	3	3	2rel, 1-57
		Invertebrate antimicrobial (chitin-binding) proteins	2	2	1dqc, A1-A73
		Bubble protein	1	1	1uoy, A1-A64
26	Inverted knottin: disulfide crossover is stacked in opposite direction as fold groups 23/24/25	Trefoil	20	4	1pcp, 1-52
		PSI domain	14	6	1olz, A480-A533
		Myeloperoxidase subdomain	16	1	1d2v, D113-D149
		Variant surface glycoprotein MITat1.2, C-terminal domain	1	1	1xu6, A354-A433
27	β-Hairpin and 1 α-helix disulfide-bonded to N-terminal loop	Kazal family serine protease inhibitor-like	76	21	1tbq, R1-R51
		Plant serine protease inhibitor-like	16	9	1ce3, A1-A54
28	Folded hairpin	ATI-like serine protease inhibitor	9	5	1ccv, A1-A56
29	Folded and twisted hairpin	BPTI-like serine protease inhibitor/dendrotoxin-like	156	28	1aap, A1-A56

30	Four-strand β-sheet, antiparallel, strand order 1234 and one α-helix	Neurophysin II	22	5	1npo, A5-A52
31	Four-strand β-sheet, antiparallel, strand order 2134 and two α-helices	TGFβ binding protein-like	7	4	1uzj, B2530-B2606
32	Four-strand β-barrel, strand order 1243, and one α-helix	Hydrophobin II	2	1	1r2m, A1-A70
33	One α-helix and four β-strands in flat array, meander topology	Thyroglobulin-like domain	6	3	1l3h, A1-A65
34	Two β-hairpins and two α-helices	TIMP, C-terminal subdomain	7	3	1gxd, C120-C192
35	Small disulfide-closed loop	μ/α Conotoxin-like	46	28	1tcg, 1-22
		Mini-protein 2 (synthesized)	13	5	1hqq, E1-E13
		Guanylin/heat-stable enterotoxin-like	7	5	1uya, 1-16
		Orexin A	2	1	1wso, A1-A33
		Arylsulfatase, conotoxin-like subdomain	8	3	1auk, 487-503
36	Mostly coil, N-terminal α-helix	Minicollagen-I, C-terminal domain	2	1	1sop, A1-A24
37	Mostly coil, central α-helix	Penaeidins	2	2	1ueo, A1-A63
38	Mostly coil, C-terminal α-helix	Tertiapin	1	1	1ter, 1-21
39	Mostly coil, 1 small α-helix	Somatomedin B domain	3	2	1s4g, A1-A51
40	Mostly coil, left-handed loop followed by right-handed loop	LDL receptor-like domain	20	17	1ajj, 4-40
41	Mostly coil, right-handed	Sea anemone neurotoxin III	1	1	1ans, 1-27

level of this classification is broader than that reflected in other classification schemes, such as the SCOP database. This strategy of grouping proteins at a more general structural level enables the identification of cases of convergent evolution of protein functions or structural topologies, which can in turn lend insight into protein/structure/function relationships. This approach can assist in the detection of potential distant homologs; evolutionary links between some of these structurally similar domains may be revealed in the future, although currently available sequence and function data do not substantiate a homology relationship.

Despite their structural similarity, homology between all domains within a fold group is not implied. Within each fold group, we classify the disulfide-rich domains into families on the basis of evolutionary relationships between members, which are inferred from the similarity of protein sequences, structures, and functions. The 2945 domains in our classification are arranged into 98 families of homologs.

Most of the different families within one fold group are likely the result of convergent evolution of unrelated proteins to a similar structural fold. However, some families within the same fold group may contain distantly related homologs for which there is currently insufficient sequence and functional information to confidently support an inference of homology between them. For example, members of fold group 3 are structurally characterized by a disulfide-bonded three-helix bundle with right-handed connections between the α-helices (Figure 1(b)). There are currently six distinct families of disulfide-rich domains within this fold group. In addition to this general structural similarity, two of these families, the CRISP (cysteine-rich secretory protein) family helical bundle subdomain and sea anemone toxin K, share similar disulfide-bonding patterns and have an N-terminal extension that is disulfide-bonded to the third α-helix in the bundle (Figure 1(b)). Representatives of these families (pdb|1bgk 1-37, *Bunodosoma granulifera* toxin;[25] pdb|1rc9 A181-A221, *Trimeresurus stejnegeri* stecrisp C-terminal domain[26]) superimpose with an RMSD of 2.0 Å over 31 C^{α} atoms. Additionally, it has been demonstrated that some members of the CRISP family inhibit a variety of different ion channels, including voltage-gated calcium channels,[27] calcium-activated potassium channels,[28] cyclic nucleotide-gated ion channels,[29] and ryanodine receptors.[30] Although it has not been directly established whether the helical bundle subdomain is the region of this protein responsible for the channel-blocking activity, this is an attractive hypothesis, because the sea anemone toxin K family members perform a similar function of blocking potassium channels. Considering the high level of structural similarity and potential functional similarity, a homology relationship between these two families seems plausible. However, due to the low level of sequence similarity between these proteins

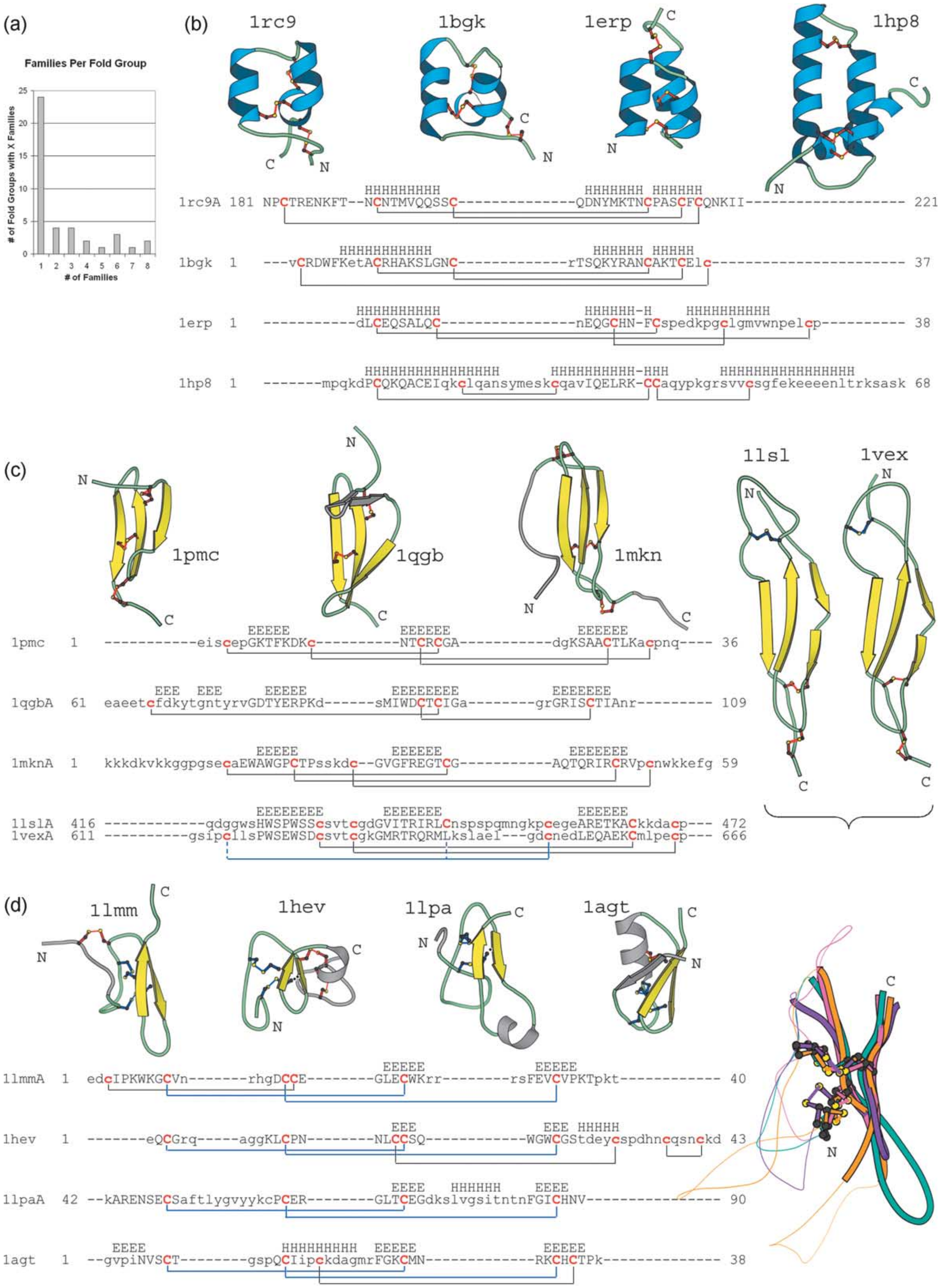

(a)
Families Per Fold Group
# of Fold Groups with X Families
# of Families
0
5
10
15
20
25
1 2 3 4 5 6 7 8
(b)
1rc9
1bgk
1erp
1hp8
N
C
HHHHHHHHH HHHHHHH HHHHHH
1rc9A 181 NPCTRENKFT--NCNTMVQQSSC--------------QDNYMKTNCPASCFCQNKII------------------- 221
HHHHHHHHHHH HHHHHH HHHHH
1bgk 1 ---vCRDWFKetACRHAKSLGNC-------------rTSQKYRANCAKTCElc----------------------- 37
HHHHHHHHHH HHHHHH-H HHHHHHHHHH
1erp 1 -----------dLCEQSALQC----------------nEQGCHN-FCspedkpgclgmvwnpelcp---------- 38
HHHHHHHHHHHHHHHH HHHHHHHHHH-HHH HHHHHHHHHHHHHHHH
1hp8 1 -------mpqkdPCQKQACEIqkclqansymeskcqavIQELRK-CCaqypkgrsvvcsgfekeeeenltrksask 68
(c)
1pmc
1qgb
1mkn
1lsl
1vex
EEEEE EEEEEE EEEEEE
1pmc 1 -----------eiscepGKTFKDKc----------NTCRCGA-----------dgKSAACTLKacpnq---- 36
EEE EEE EEEEE EEEEEEEE EEEEEEE
1qgbA 61 eaeetcfdkytgntyrvGDTYERPKd------sMIWDCTCIGa--------grGRISCTIAnr--------- 109
EEEEE EEEEEE EEEEEEE
1mknA 1 kkkdkvkkggpgsecaEWAWGPCTPsskdc--GVGFREGTCG-----------AQTQRIRCRVpcnwkkefg 59
EEEEEEEE EEEEEEE EEEEEE
1lslA 416 ------------qdggwsHWSPWSScsvtcgdGVITRIRLCnspspqmngkpcegeARETKACkkdacp--- 472
1vexA 611 ----------gsipcllsPWSEWSDcsvtcgkGMRTRQRMLkslael---gdcnedLEQAEKCmlpecp--- 666
(d)
1lmm
1hev
1lpa
1agt
EEEEE EEEEE
1lmmA 1 edcIPKWKGCVn-------rhgDCCE--------GLECWKrr--------rsFEVCVPKTpkt----------- 40
EEE EEE HHHHH
1hev 1 -------eQCGrq-----aggKLCPN--------NLCCSQ------------WGWCGStdeycspdhncqsnckd 43
EEEE HHHHHH EEEE
1lpaA 42 --kARENSECSaftlygvyykcPCER--------GLTCEGdkslvgsitntnFGICHNV--------------- 90
EEEE HHHHHHHHH EEEEE EEEEE
1agt 1 --gvpiNVSCT--------gspQCIipckdagmrFGKCMN-------------RKCHCTPk------------- 38

Figure 1. Common folds adopted by unrelated disulfide-rich families. The common structural core of each fold group is shown in color (blue, α-helices; yellow, β-strands; green, coils); additional elements are shown in grey. In this and other MOLSCRIPT[88] Figures, disulfide bonds are colored red or blue and are shown in ball-and-stick format. A structural alignment for the representatives is shown below the structure figures. In these and other alignments, upper case letters denote residues that are structurally aligned; lower case letters are residues that do not align structurally with the other fold group members. Red bold text indicates cysteine residues that are highly conserved within a family and are involved in disulfide bonds. PDB (pdb|) identifiers and chain names are shown at the far left. The numbers before and after the sequence denote the PDB residue number of the first and last residues in the domain sequence, respectively. Secondary structure elements are noted above the alignment: H, α-helix; E, β-strand. Disulfide bonding patterns are

(average identity ∼13%) and the unconfirmed function of the CRISP family helical bundle subdomain, we cannot confidently assert the merging of these two families.

Distribution of families within fold groups

Each of the 41 fold groups in this classification contains between one and eight distinct families (Figure 1(a)). There is a subset of topologies that seem to be quite common among small, disulfide-rich domains. In fact, nearly half of the 98 families belong to fold groups that consist of five or more presumably non-homologous families each. Examples of recurring structural motifs in disulfide-rich domains are depicted in Figure 1. Typically, these common folds have a simple topology that could easily have arisen multiple times by chance, such as α-hairpins (fold groups 1 and 2) or three-strand β-sheets with meander topology (fold group 12). Knottin-like topology, found in nearly 40% of the disulfide-rich domain structures currently available (fold groups 23–25), is the most commonly observed structural motif. It is characterized by two adjacent disulfide bonds (one bond is formed by the first and third cysteine residues in the primary sequence, and the other by the second and fourth cysteine residues) and a conserved β-hairpin, on which the third and fourth cysteine residues are located (Figure 1(d)). Because the two disulfide bonds are roughly perpendicular, so that they form an × or cross, the knottin-like core is also known as the disulfide β-cross. This motif has been suggested as a stable protein folding nucleus,[31] which would confer an evolutionary advantage to proteins with this particular fold and explain the convergence of a large number of families to this common core.

On the other hand, approximately half of the 41 fold groups currently include only a single protein family. Some of these proteins have more complicated architectures (e.g. irregular α-helical arrays; fold groups 5–8), while others are mostly coil proteins with little or no standard secondary structure (fold groups 36–41). This large number of unique fold groups reflects the wide conformational variety available to proteins stabilized by disulfide bonds. Because a structural scaffold that is not entirely reliant upon secondary structure and hydrophobic interactions allows for much more conformational irregularity (e.g. little or no α-helix and β-strand character, non-globular shapes, etc.), disulfide bonds can potentially stabilize numerous protein conformations that otherwise would not exist. For example, *Ascaris suum* chymotrypsin/elastase inhibitor (fold group 28) is a 60-residue protein with a structural fold maintained by five conserved disulfide bonds. As this protein contains only a few small secondary structure elements and lacks a hydrophobic core,[32] it is highly unlikely that a non-disulfide structural analog of this fold would exist in nature.

It should be noted that the currently available disulfide-rich domains are not expected to represent all disulfide-stabilized proteins that exist in nature. It is likely that this classification will require the addition of several new families and fold groups when novel disulfide-rich protein structures are revealed in the future.

Comparison to SCOP database

Most of our disulfide-rich domains are classified also by the SCOP database. Approximately 13% of our 2945 domains are not assigned a SCOP classification (version 1.69), in most cases because the structure of the protein was released quite recently and has not yet been incorporated into the SCOP database. Another 4% of domains in our study are regions within larger proteins that are not distinguished as distinct subdomains by SCOP (for example, the N-terminal EGF-like domain of alliinase; pdb|1lk9,[33] residues A2–A60). Of the remaining 2446 domains (i.e. those that are classified by SCOP), 84% are found in the Small proteins class of SCOP, 11.5% in the All-α class, 3% in the Peptides class, 1% in the All-β class, and the remaining 0.5% in the Coiled coil proteins and Designed proteins classes.

Fold group level describes broader structural similarity than SCOP folds

In SCOP, the "fold" level is intended to reflect the traditional definition of a fold, where protein structures "have the same major secondary structures in the same arrangement and with the same

depicted by lines connecting cysteine residues. (a) Distribution of disulfide-rich families in fold groups. (b) Representatives of four families with right-handed three-helix bundle fold (fold group 3): *T. stejnegeri* stecrisp helical subdomain (pdb|1rc9), *B. granulifera* toxin K (pdb|1bgk), *E. raikovi* pheromone ER-10 (pdb|1erp), and *Homo sapiens* p8 protein from oncogene MTCP1 (pdb|1hp8). (c) Representatives of four families with antiparallel three-stranded β-sheet with meander topology (fold group 12): *Locusta migratoria* protease inhibitor PMP-C (pdb|1pmc), *Homo sapiens* fibronectin type 1 module (pdb|1qgb), *Homo sapiens* midkine N-terminal domain (pdb|1mkn), and TSP-1 repeats from *Homo sapiens* thrombospondin (pdb|1lsl) and *Rattus norvegicus* F-spondin (pdb|1vex). The bracket indicates homologous domains. Disulfide bonds shown in blue involve a cysteine residue that does not align among all members of that family. (d) Representatives of four families with knottin-like topology (fold group 23): *Psalmopoeus cambridgei* psalmotoxin 1 (pdb|1lmm), *Hevea brasiliensis* hevein (pdb|1hev), *Sus scrofa* colipase C-terminal domain (pdb|1lpa), and *Leiurus quinquestriatus hebraeus* agitoxin (pdb|1agt). Disulfide bonds shown in blue form the distinctive disulfide cross. The key knottin-like features (disulfide cross and β-hairpin) are superimposed for these four representatives (pdb|1lmm, green; pdb|1hev, purple; pdb|1lpa, orange; pdb|1agt, pink). The connecting backbone is shown as a thin coil.

topological connections"†. The fold group level in our classification is comparable, in that domains within a fold group have a common structural core, but describes broader similarities between protein structures. The most fundamental difference is that our fold groups bring together structures with "similar arrangements" of elements, rather than only the "same arrangement". This typically means that members of the same fold group might have different relative orientations of secondary structure elements, but still share a common spatial layout; one such example for right-handed three-helix bundles is described below. The sizes of structural elements involved in the common structural core also vary more significantly among our fold group members than is generally observed within SCOP folds. Our fold groups include circular permutations of topologies. The key distinctions between our more lenient fold group criteria and the traditional fold definition are highlighted in the following example. *Eurplotes raikovi* pheromone ER-10 (pdb|1erp[34]) and mature T-cell proliferation oncogene-encoded protein p8$^{\mathrm{MTCP1}}$ (pdb|1hp8)[35] are two unrelated, small α-helical proteins. Because the structures of both proteins are right-handed three-helix bundles, we assign them to the same fold group (fold group 3; Figure 1(b)). However, SCOP assigns these proteins to separate folds (a.10: protozoan pheromone proteins and a.17: p8-MTCP1), most likely because of the different sizes and relative orientations of the α-helices. The broad nature of our fold groups is reflected by the distribution of SCOP folds in our classification: 18 of our fold groups include proteins from more than one SCOP fold, and six of our fold groups include representatives of more than one SCOP class.

The one exception to this trend is the set of knottin-like domains. SCOP assigns all of these domains to the same fold, g.3: Knottins (small inhibitors, toxins, lectins), on the basis of the presence of two adjacent disulfide bonds and a β-hairpin. By our definition of a fold group, however, all secondary structure elements in the structural core of a domain should be considered. Therefore, in our classification, knottin-like structures are arranged according to the topology of the backbone contributing all four cysteine residues that make up the disulfide cross, rather than only the β-hairpin that contains the third and fourth cysteine residues. The knottin-like domains comprise fold groups 23, 24, and 25 (Table 1).

Disulfide-rich families are approximately equivalent to SCOP superfamilies

Our disulfide-rich families are, for the most part, consistent with the superfamily level of SCOP, which is the broadest level of homology conveyed in the SCOP classification hierarchy. SCOP, however, is a fairly conservative database and, after careful examination of the domains in our study, a few additional homology relationships were identified. These newly linked families are described in the following section.

Distant homology between disulfide-rich domains

Bowman–Birk inhibitors and bromelain inhibitors

Bowman–Birk inhibitors (BBIs) are serine protease inhibitors specific for trypsin and chymotrypsin. These proteins are found in many plant seeds, and structures have so far been solved for BBIs from soybean, lima bean, mung bean, adzuki bean, garden pea, barley, peanut, and snail medic seeds. Also, seven isoinhibitors of the cysteine protease bromelain have been identified from the stem of pineapple. The structure of bromelain inhibitor VI (BI-VI) is currently available. The bromelain inhibitors are somewhat unique, in that they are formed by a heavy chain (41 residues) and a light chain (11 residues) that originated from a single-chain precursor.[36,37] The BBI and BI-VI proteins have very clear structural similarity (Figure 2(a)). First, both proteins contain a tandem repeat of a small, antiparallel three-stranded β-sheet with strand order 231 and three highly conserved disulfide bonds. In both BBIs and BI-VI, one domain is contiguous, while the second domain is related by circular permutation. Furthermore, while the sequence similarity between these proteins is not overwhelming, the percentage identity between the BBI and BI-VI domains is comparable with the identity between the two subdomains of the same protein. For example, the average identity between the circularly permuted domain of BI-VI (1bi6HL in Figure 2) and the BBI representative domains is 27%, while the identity between 1bi6HL and the non-permuted BI-VI domain (1bi6H1) is only marginally higher at 28%. Furthermore, BBIs and BI-VI are identified as homologs by the MEROPS database of proteases and protease inhibitors (clan IF).[23] On the basis of the sequence, structural, and functional similarity between these proteins, we have merged the BBIs and the bromelain inhibitor VI into a single family, which is included in fold group 13.

EGF-like subdomain of garlic alliinase

Garlic alliinase is a lyase that cleaves carbon–sulfur bonds to produce the sulfur-containing garlic components to which this plant's pharmacological properties are attributed. The structure of this protein confirmed the predicted presence of a cysteine-rich N-terminal subdomain similar to EGF-like domains.[33,38] The function of this subdomain is unknown. This domain lacks one of the three highly conserved disulfide bonds that are typical of EGF-like domains, and has one additional disulfide bond that is not seen among other family members (Figure 3(a)). Despite these differences in

† http://scop.mrc-lmb.cam.ac.uk/scop/intro.html

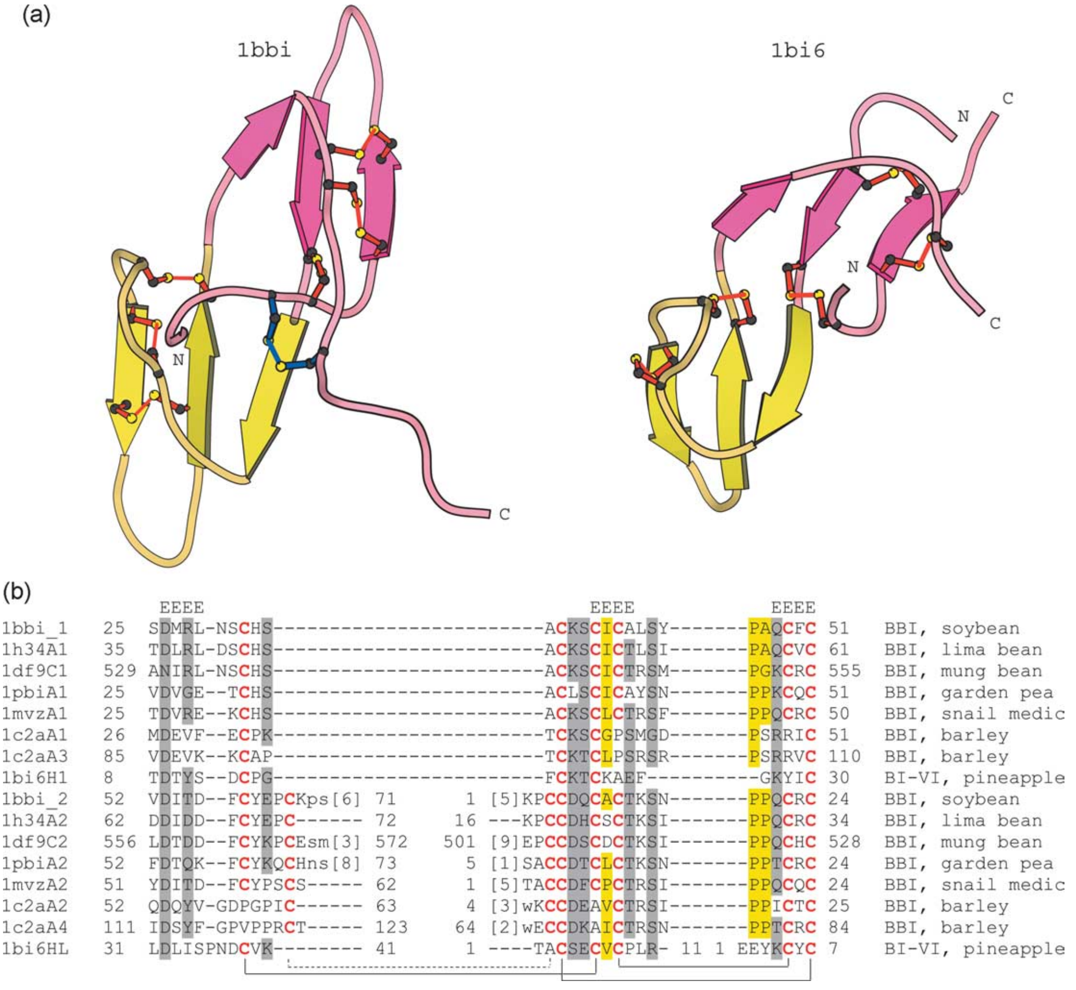

(a)
1bbi
1bi6
N
C
N
C
N
C
(b)
EEEE EEEE EEEE
1bbi_1 25 SDMRL-NSCHS-------------------------ACKSCICALSY-------PAQCFC 51 BBI, soybean
1h34A1 35 TDLRL-DSCHS-------------------------ACKSCICTLSI-------PAQCVC 61 BBI, lima bean
1df9C1 529 ANIRL-NSCHS-------------------------ACKSCICTRSM-------PGKCRC 555 BBI, mung bean
1pbiA1 25 VDVGE--TCHS-------------------------ACLSCICAYSN-------PPKCQC 51 BBI, garden pea
1mvzA1 25 TDVRE--KCHS-------------------------ACKSCLCTRSF-------PPQCRC 50 BBI, snail medic
1c2aA1 26 MDEVF--ECPK-------------------------TCKSCGPSMGD-------PSRRIC 51 BBI, barley
1c2aA3 85 VDEVK--KCAP-------------------------TCKTCLPSRSR-------PSRRVC 110 BBI, barley
1bi6H1 8 TDTYS--DCPG-------------------------FCKTCKAEF----------GKYIC 30 BI-VI, pineapple
1bbi_2 52 VDITD--FCYEPCKps[6] 71 1 [5]KPCCDQCACTKSN-------PPQCRC 24 BBI, soybean
1h34A2 62 DDIDD--FCYEPC------ 72 16 ---KPCCDHCSCTKSI-------PPQCRC 34 BBI, lima bean
1df9C2 556 LDTDD--FCYKPCEsm[3] 572 501 [9]EPCCDSCDCTKSI-------PPQCHC 528 BBI, mung bean
1pbiA2 52 FDTQK--FCYKQCHns[8] 73 5 [1]SACCDTCLCTKSN-------PPTCRC 24 BBI, garden pea
1mvzA2 51 YDITD--FCYPSCS----- 62 1 [5]TACCDFCPCTRSI-------PPQCQC 24 BBI, snail medic
1c2aA2 52 QDQYV-GDPGPIC------ 63 4 [3]wKCCDEAVCTRSI-------PPICTC 25 BBI, barley
1c2aA4 111 IDSYF-GPVPPRCT----- 123 64 [2]wECCDKAICTRSN-------PPTCRC 84 BBI, barley
1bi6HL 31 LDLISPNDCVK-------- 41 1 ----TACSECVCPLR- 11 1 EEYKCYC 7 BI-VI, pineapple

Figure 2. Bowman–Birk and bromelain inhibitors are homologs. (a) MOLSCRIPT diagrams of Bowman–Birk inhibitor from soybean (pdb|1bbi) and bromelain inhibitor VI from pineapple (pdb|1bi6). Each protein is comprised of two homologous domains that adopt an antiparallel three-stranded β-sheet fold; one domain is continuous (yellow), while the other is circularly permuted (pink). Highly conserved disulfide bonds are shown in red; additional disulfide bonds are shown in blue. (b) Structure-based multiple alignment of BBI and BI-VI representatives. Continuous lines indicate highly conserved disulfide bonds (red in (a)); broken lines indicate disulfide bonds that are found in only one of the two tandem repeats (blue in (a)). In this and other multiple alignments, red bold text indicates conserved cysteine residues involved in disulfide bonds, yellow highlighting indicates uncharged residues in mostly hydrophobic positions, grey highlighting indicates mostly polar positions, and cyan highlighting indicates mostly aromatic positions.

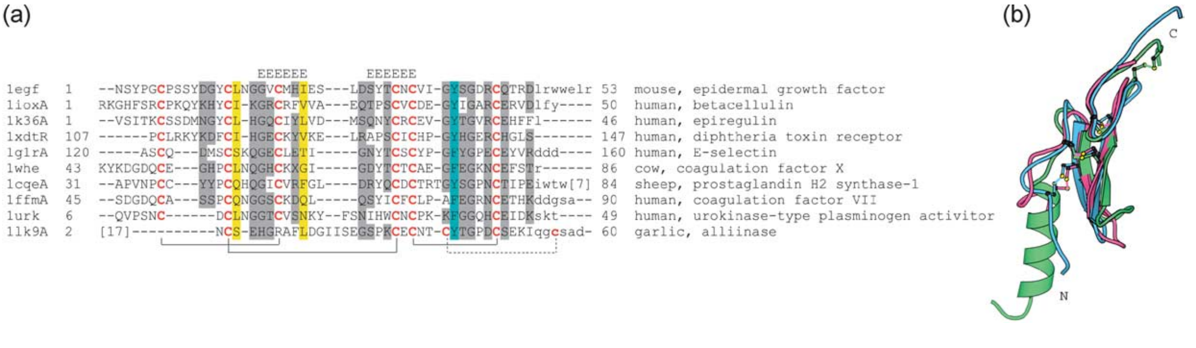

(a)
(b)
EEEEEE EEEEEE
1egf 1 --NSYPGCPSSYDGYCLNGGVCMHIES---LDSYTCNCVI-GYSGDRCQTRDlrwwelr 53 mouse, epidermal growth factor
1ioxA 1 RKGHFSRCPKQYKHYCI-KGRCRFVVA---EQTPSCVCDE-GYIGARCERVDlfy---- 50 human, betacellulin
1k36A 1 --VSITKCSSDMNGYCL-HGQCIYLVD---MSQNYCRCEV-GYTGVRCEHFFl------ 46 human, epiregulin
1xdtR 107 ------PCLRKYKDFCI-HGECKYVKE---LRAPSCICHP-GYHGERCHGLS------- 147 human, diphtheria toxin receptor
1g1rA 120 -----ASCQ---DMSCSKQGECLETI-----GNYTCSCYP-GFYGPECEYVRddd---- 160 human, E-selectin
1whe 43 KYKDGDQCE---GHPCLNQGHCKXGI-----GDYTCTCAE-GFEGKNCEFSTr------ 86 cow, coagulation factor X
1cqeA 31 --APVNPCC---YYPCQHQGICVRFGL----DRYQCDCTRTGYSGPNCTIPEiwtw[7] 84 sheep, prostaglandin H2 synthase-1
1ffmA 45 --SDGDQCA---SSPCQNGGSCKDQL-----QSYICFCLP-AFEGRNCETHKddgsa-- 90 human, coagulation factor VII
1urk 6 --QVPSNC-------DCLNGGTCVSNKY--FSNIHWCNCPK-KFGGQHCEIDKskt---- 49 human, urokinase-type plasminogen activitor
1lk9A 2 [17]-----------NCS-EHGRAFLDGIISEGSPKCECNT-CYTGPDCSEKIqgcsad- 60 garlic, alliinase
C
N

Figure 3. EGF-like subdomain of garlic alliinase. (a) Structure-based multiple alignment of EGF-like family members. Continuous lines indicate highly conserved disulfide bonds; broken lines indicate the additional disulfide bond in the garlic alliinase EGF-like subdomain. (b) Garlic alliinase subdomain (green; pdb|1lk9, A2-A60) superimposed with EGF-like domains from human coagulation factor VII (blue; pdb|1ffm, A45-A90) and human diphtheria toxin receptor (pink; pdb|1xdt, R107-R147).

disulfide bonding patterns, the alliinase N-terminal domain nonetheless shares striking structural similarity with the EGF-like family. Among the closest structural neighbors of the alliinase subdomain are EGF-like domains from human coagulation factor VII (pdb|1ffm; 2.1 Å RMSD, 33 C^{α} atoms) and human diphtheria toxin receptor (pdb|1xdt; 2.1 Å RMSD, 31 C^{α} atoms) (Figure 3(b)). Although there is limited sequence similarity between the alliinase subdomain and the other EGF-like family members (typically <20% identity), this distant homology relationship is recognized by the Pfam[39] database, which places the N-terminal domain of garlic alliinase (PF04863) into the same clan as other EGF-like domains (CL0001: EGF superfamily). We have included the N-terminal subdomain of garlic alliinase with the family of EGF-like knottins in fold group 25.

Cellulose-binding/docking domains

The third example includes domains from two different polysaccharide hydrolases that are involved in recycling carbohydrates by breaking down plant cell walls. The cellulose-binding domain of *Pseudomonas fluorescens* xylanase A (CBDx) binds carbohydrate polymers of plant cell walls.[40] The cellulose-docking domain of *Piromyces equi* endoglucanase Cel45A (CDDe) does not interact directly with cellulose but instead binds to small protein domains (cohesin domains) that are found in the same polypeptide chain as domains that do bind cellulose directly.[41] Thus, while CBDx and CDDe perform different molecular functions, they are responsible for the same role, on a more general level, of bringing the catalytic domains of these enzymes and the carbohydrate polymers into close proximity so that the hydrolysis reaction can proceed. The structural features shared by these domains include two hairpins that lie approximately perpendicular to each other (Figure 4(a)). This region of these domains superimposes with 2.7 Å RMSD (25 C^{α} atoms). Additionally, the key residues involved in binding are presented on the same face of both structures[42,43] (Figure 4(a)). Although these domains share only limited sequence similarity, the distant homology relationship between them is also identified by the Pfam database (PF02013: cellulose or protein-binding domain). We have assigned CBDx and CDDe to the cellulose-binding/docking domain family (fold group 18).

Knottin-like domains

The knottin-like proteins are a large group of structurally similar proteins, as described previously. SCOP assigns these domains to 19 superfamilies in a single fold. Our classification includes 19 knottin-like families, although some of these proteins are found outside of the SCOP knottin-like fold. Furthermore, while most of our knottin-like families are in close agreement with SCOP, we have made subtle rearrangements of some families within this large class.

The omega toxin-like superfamily is among the largest knottin-like superfamilies in SCOP. While most of the members are ω-conotoxins and spider toxins, this superfamily includes some insect toxins, scorpion toxins, spider lectins, and antimicrobial

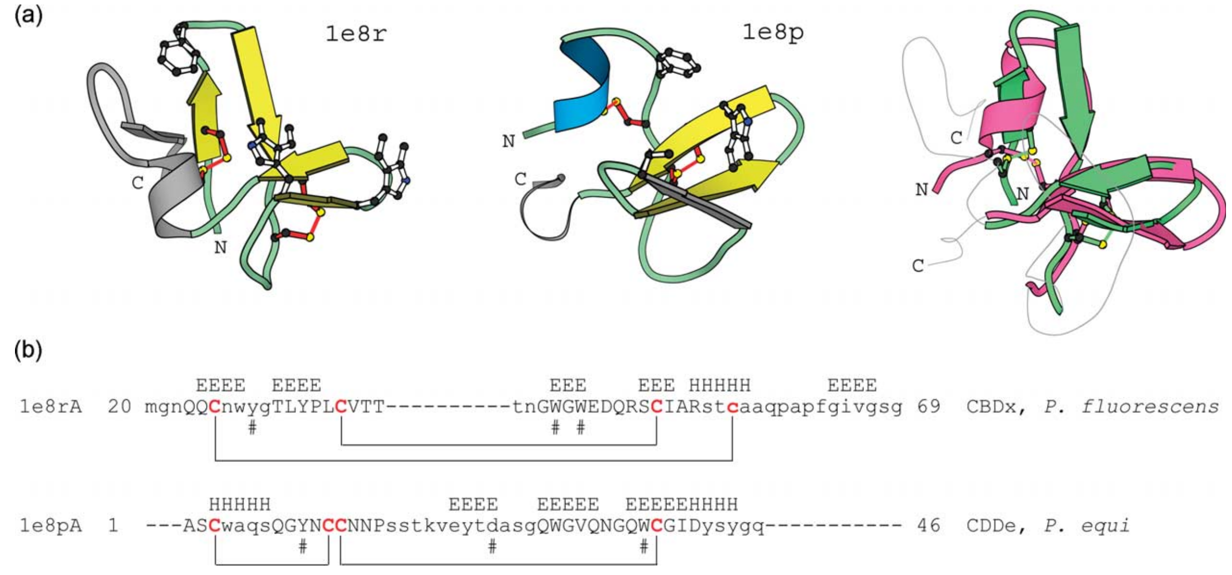

(a)
1e8r
C
N
1e8p
N
C
C
N
N
C
(b)
EEEE EEEE EEE EEE HHHHH EEEE
1e8rA 20 mgnQQCnwygTLYPLCVTT----------tnGWGWEDQRSCIARstcaaqpapfgivgsg 69 CBDx, P. fluorescens
# # #
HHHHH EEEE EEEEE EEEEEHHHH
1e8pA 1 ---ASCwaqsQGYNCCNNPsstkveytdasgQWGVQNGQWCGIDysygq---------- 46 CDDe, P. equi
#

Figure 4. Cellulose binding/docking domains. (a) The cellulose-binding domain of xylanase A (CBDx) from *P. fluorescens* (pdb|1e8r) and the cellulose-docking domain of endoglucanase Cel45 (CDDe) from *P. equi* (pdb|1e8p) share a similar structural core of two roughly perpendicular hairpins. Shared structural features are shown in color: CBDx has two β-hairpins, while CDDe has one β-hairpin and one hairpin formed by an α-helix and a coil region. Other elements are shown in grey. Putative binding residues are shown in ball-and-stick format and are white. In the superimposed view of CBDx (green) and CDDe (pink), the two hairpins are shown in color and the rest of the backbone is shown as a thin grey coil. (b) Structural alignment of the CDDe and CBDx domains. Putative binding residues are indicated by #.

proteins. We refer to the omega toxin-like superfamily of SCOP as the ωTL superfamily. This set of proteins comprise the bulk of one family in our classification (the spider toxin-like family), although we have included proteins from several other SCOP superfamilies. These additional members include gurmarin, antifungal peptides PAFP-S and Alo3, the C-terminal domain of agouti-related signaling proteins, the knottin-like domain of insulin-like growth factor-binding proteins (IGFBPs), plant enzyme (α-amylase, carboxypeptidase, trypsin) inhibitors, and the C-terminal subdomain of the VHv1.1 polydnaviral gene product. With the exception of the viral subdomain, which is not yet incorporated into the SCOP database, each of these domains is found outside of the ωTL SCOP superfamily, perhaps on the basis of functional dissimilarity. However, all of these new members display significant sequence and structural similarity with the ωTL domains (Figure 5(b)). Each of these additional subsets includes at least one domain that shares $>33\%$ sequence identity and <2.5 Å RMSD (>25 C^α atoms) with a member of the ωTL SCOP superfamily.

Furthermore, there are several cases in which these new members share greater levels of sequence and structural similarity with a representative of the ωTL SCOP superfamily than with other members of its SCOP-assigned superfamily. For example, antifungal protein PAFP-S (pdb|1dkc) is highly similar to ωTL representative conotoxin TxVIa (pdb|1fu3) with 37% sequence identity and 2.1 Å RMSD (27 C^α atoms). Meanwhile, the sweet-taste suppressor signaling protein gurmarin (pdb|1c4e) shares 38% sequence identity and 1.9 Å RMSD (26 C^α atoms) with ωTL representative conotoxin TxVII (pdb|1f3k). Although PAFP-S and gurmarin are assigned to the same SCOP superfamily, they share less similarity with each other (20% sequence identity; 3.2 Å RMSD, 33 C^α atoms) than with the ωTL representatives (Figure 5(c)). Likewise, plant inhibitors of trypsin, carboxypeptidase A, and α-amylase inhibitor belong to the same SCOP superfamily (plant inhibitors of proteinases and amylases) despite the limited sequence and structural similarity among these proteins. The α-amylase inhibitor shares only 16% sequence identity and 3.2 Å RMSD (24 C^α atoms) with carboxypeptidase A inhibitor, and only 22% sequence identity and 4.0 Å RMSD (27 C^α atoms) with the trypsin inhibitors. However, the α-amylase inhibitor (pdb|1clv) shares 36% sequence identity and 1.9 Å RMSD (27 C^α atoms) with the ωTL representative covalitoxin-I (pdb|1v5a). In a similar example, *Conus gloriamaris* conotoxin GmIXa (pdb|1ixt), the first structural representative of the P-superfamily conotoxins,[44] is assigned to the ωTL of SCOP, presumably on the basis of putative function and species of origin. However, this protein shares more obvious similarity with carboxypeptidase A inhibitor (pdb|4cpa) from the plant inhibitors of proteinases and amylases SCOP superfamily (41% sequence identity; 2.3 Å RMSD, 25 C^α atoms) than with the other ωTL domains (average sequence identity 23%; average RMSD 2.5 Å, 24 C^α atoms). Thus, there are members of the ωTL SCOP superfamily with greater similarity to other SCOP superfamilies than to other ωTL domains, and there are members of other SCOP superfamilies with greater similarity to ωTL domains than to each other. On the basis of such links, we have merged four different SCOP superfamilies with the ωTL domains. The distribution of pairwise sequence identity among spider toxin-like family members in our classification, compared to sequence identity between these proteins and non-homologous disulfide-rich domains, is shown in Supplementary Data (Supplementary Figure 3).

In most of these cases, entire SCOP superfamilies are merged with the ωTL domains. The sole exception is the knottin-like domain of IGFBP. In SCOP, this domain is classified with the cysteine-rich repeats of EGF receptors (also known as ErbBs). The extracellular domain of ErbB proteins includes two regions of cysteine-rich repeats, each of which is comprised of 13 homologous knottin-like domains in tandem, and two regions of leucine-rich repeats. The cysteine-rich regions are involved in dimerization.[45] On the other hand, three non-similar disulfide-rich domains are found within IGFBPs. The second domain adopts a knottin-like fold, and is involved in binding insulin-like growth factors.[46] The knottin-like domains of IGFBPs and ErbBs share only 19% sequence identity on average and are structurally alignable over only 12 C^α atoms, with an average RMSD of 2.1 Å over these residues. Thus, the knottin-like domains of IGFBPs and ErbBs share neither significant sequence, structural, nor functional similarity and are therefore unlikely to be evolutionarily related. The knottin-like domain of IGFBPs has been added to the spider toxin-like family. Meanwhile, the ErbB knottin-like repeats are very short (average size of 20 residues), and are structurally most similar to other very small knottin-like domains, such as DPY modules and the knottin-like subdomain of CRISP family proteins (Figure 5(d)). However, due to the limited sequence similarity between these domains (average sequence identity of 25% between knottin-like domains of ErbBs and CRISPs; average sequence identity of 19% between ErbB knottin-like domains and DPY modules), they remain as three separate families within fold group 25.

Disulfide bonding patterns and protein topology

Disulfide bonds and protein structure

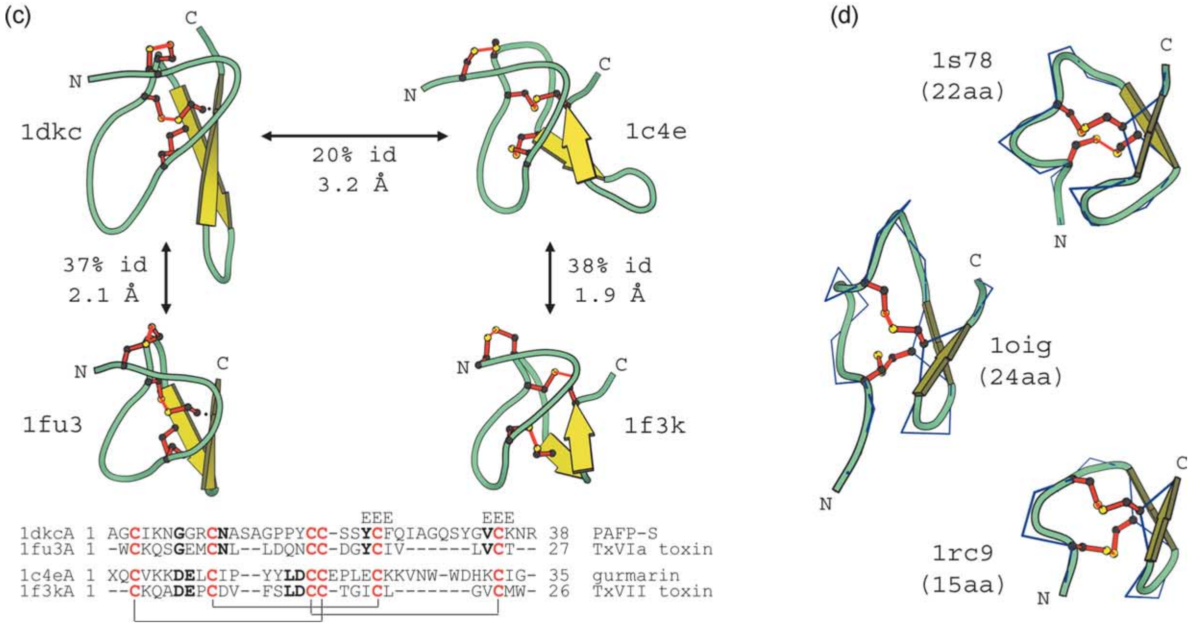
While there is debate about the physical mechanism by which disulfide bonds act as a stabilizing influence,[2,5,6] the general importance of these covalent bonds in small protein domains is demonstrated clearly by their extremely high level of conservation. However, a comprehensive view of the numerous cysteine-mutation studies performed on these proteins indicates that the extent to which the structure and function of a domain are

(a)

```
                                                EEE         EEE
1lmmA  1    ---EDCIP-KWKGCVN---RHGDCCE--GL-ECWKRR-RSFEVCVPKTPKT------ 40  psalmotoxin-1, Trinidad chevron tarantula
1omb   4    -----CIAEDYGKCTW---GGTKCCR--GR-PCRCSMIGTNCECTP----------- 38  ω-agatoxin-IVb, funnel web spider
1axh   1    --SPTCIP-SGQPCPY----NENCC---SQ-SCTFK[6]TVKRCD------------ 37  ω-atracotoxin-Hv1, Australian funnel web spider
1nixA  1    ----ECKG-FGKSCVP---GKNECCS--GY-ACNSR----DKWCKVLL--------- 33  hainantoxin-1, Chinese bird spider
1d1hA  1    ----ECRY-LFGGCKT----TSDCCK--HL-GCKFR----DKYCAWDFTFS------ 35  hanatoxin-1, Chilean rose tarantula
1f3kA  1    -----CKQ-ADEPCDV---FSLDCC---TG-ICL-------GVCMW----------- 26  TxVII toxin, textile cone snail
1fu3A  1    ----WCKQ-SGEMCNL---LDQNCC---DG-YCIV------LVCT------------ 27  TxVIa toxin, textile cone snail
1rmkA  1    ----ACSK-KWEYCIVPILGFVYCCP--GL-ICGP------FVCV------------ 31  MrVIb toxin, marble cone snail
1fygA  1    -----CKA-AGKPCSR---IAYNCC---TG-SCRS------GKC------------- 25  S03 toxin, striated cone snail
1ixtA  1    ----sc----NNSCQ----SHSDCAS--HC-ICTF------RGCGAVN--------- 27  GmIXa toxin, glory-of-the-sea cone snail
1cixA  1    --YSRCQL-QGFNCVV[5]PTIPCCR--GL-TCRSYF[4]YGRCQRY---------- 44  tachystatin, Japanese horseshoe crab
1i26A  1    -AEKDCIA-PGAPCFG---TDKPCCNP-RA-WCSSY----ANKCL------------ 34  PTU1 toxin, assassin bug
1ju8A  1    ---adc----NGACSP--FEVPPCRSR-DC-RCVPIGL-FVGFCIHPTG-------- 37  leginsulin, soybean

1c4eA  1    ---XQCVK-KDELCIP---YYLDCCE--PL-ECKKVNW-WDHKCIG----------- 35  gurmarin, gurmar plant
1dkcA  1    ---AGCIK-NGGRCNAS-AGPPYCC---SS-YCFQIAGQSYGVCKNR---------- 38  antifungal protein PAFP-S, American pokeweed
1q3jA  1    -----CIK-NGNGCQPN-GSQGNCC---SG-YCHKQPGWVAGYCRRK---------- 36  antifungal protein Alo3, harlequin beetle

1hykA  1    -----CVR-LHESCLG---QQVPCCDP-CA-TCYCRFFNAFCYCRKLGTAMNPCSRT 46  agouti-related protein, human
1mr0A  1    -----CVR-LHESCLG---qQVPCCDP-AA-TCYCRFFNAFCYC------------- 33  agouti-related protein (mutant), human

1h9iI  1    ----gcpr-ILIRCK----QDSDCLA--GC-VCTN-----NKFCGSP---------- 30  trypsin inhibitor EETI-II, squirting cucumber
4cpaI  3    hadpic----NKPCK----THDDCSGAWFCQACWNS----ARTCGPYV--------- 38  carboxypeptidase inhibitor, potato
1clvI  501  -----CIP-KWNRCGPK-MDGVPCCE--PY-TCTSD---YYGNCS------------ 532 α-amylase inhibitor, amaranth plant

1wqjB  40   ------LG-LGMPCGV---YTPRCGS--GL-RCYPP[13]QGVCMEL---------- 82  IGFBP4, human

1xi7A  6    ----TCIG-HYQKCVN---ADKPCC[14]F-ICDRD---GEGVCVPFDG-------- 52  VHv1.1 gene product, wasp polydnavirus
```

(b)

additional ωTL family members (SCOP superfamily)	*similarity to ωTL (g.3.6) domains*				
	sequence identity		*RMSD*		*linking pair example (sequence identity, RMSD)*
	average	*best*	*average*	*best*	
gurmarin and antifungal proteins (g.3.4)	26%	40%	2.7 Å	1.2 Å	gurmarin signaling protein (pdb|1c4e) & conotoxin TxVII (pdb|1f3k) 38%, 1.9 Å (26 Cα)
plant enzyme inhibitors (g.3.2)	24%	41%	2.3 Å	1.0 Å	α-amylase inhibitor (pdb|1clv) & covalitoxin-I (pdb|1v5a) 36%, 1.9 Å (27 Cα)
agouti-related, C-terminal domain (g.3.5)	24%	38%	2.7 Å	1.7 Å	agouti-related protein (pdb|1mr0) & hainantoxin-I (pdb|1nix) 33%, 2.1 Å (30 Cα)
IGFBP knottin-like domain (g.3.9)	21%	35%	2.3 Å	1.3 Å	IGFBP4 subdomain (pdb|1wqj, B40-B82) & conotoxin TxVII (pdb|1f3k) 35%, 2.5 Å (25 Cα)
C-terminal domain, VHv1.1 viral gene product (N/A)	25%	37%	2.8 Å	1.6 Å	viral subdomain (pdb|1xi7) & conotoxin TVIIa (pdb|1eyo) 37%, 1.6 Å (26 Cα)
similarity among ωTL domains	28%	--	2.4 Å	--	--



```
                                      EEE          EEE
1dkcA 1 AGCIKNGGRCNASAGPPYCC-SSYCFQIAGQSYGVCKNR 38  PAFP-S
1fu3A 1 -WCKQSGEMCNL--LDQNCC-DGYCIV------LVCT-- 27  TxVIa toxin

1c4eA 1 XQCVKKDELCIP--YYLDCCEPLECKKVNW-WDHKCIG- 35  gurmarin
1f3kA 1 --CKQADEPCDV--FSLDCC-TGICL-------GVCMW- 26  TxVII toxin
```

Figure 5. Additional members of the spider toxin-like family. (a) Structure-based multiple alignment of ωTL domains and added family members. Continuous lines indicate highly conserved disulfide bonds; broken lines indicate disulfide

dependent upon the presence of conserved disulfide bonds varies. Mutagenesis studies of several different proteins have shown that eliminating a specific disulfide bond significantly alters neither the protein structure nor function: the mutated protein adopts a native-like fold (although it is usually less stable and more susceptible to denaturation) and retains all or most of its wild-type function (which is verified experimentally in some cases and hypothesized because of the lack of structural change in others). Small protein examples include bovine pancreatic trypsin inhibitor (fold group 29),[47–49] charybdotoxin (scorpion toxin-like/insect and plant defensin-like family, fold group 23),[50] kalata B1 (kalata-like cyclotides family, fold group 23),[51]and vascular endothelial growth factor (cysteine knot cytokines family, fold group 23).[52]

In other studies, however, elimination of a single disulfide bond resulted in drastic changes in protein structure and/or function. For example, mutating the disulfide bond between the first and third cysteine residues (i.e. the 1–3 disulfide bond) in murine epidermal growth factor (EGF-like family, fold group 25) results in a structural fold that is highly similar to the native fold except at the N-terminal tail, but causes a dramatic reduction in both mitogenic activity and receptor binding.[53] The opposite situation is seen upon mutation of the 2–3 disulfide bond in an endothelin-1 analog (endothelin-like family, fold group 1): the protein retains agonist activity but the native tertiary fold is completely destroyed.[54] Eliminating the 2–4 disulfide bond of α-conotoxin GI (μ/α conotoxin-like family, fold group 35) results in both a non-native structural fold and loss of toxicity.[55] A similar disruption of both structure and function is seen when the 2–4 or 3–5 disulfide bonds of toxin ShK (sea anemone toxin K family, fold group 3) are deleted. Interestingly, the ShK variant with a mutated 1–6 disulfide bond retains potassium channel inhibitor activity despite adopting a structure significantly different from the native fold.[56] Unsurprisingly, very small proteins (<35 residues) tend to be highly intolerant of the mutation of cysteine residues involved in formation of a disulfide bond.

Native variations in disulfide bonds

It has long since been recognized that the cysteine residues in disulfide bonds are nearly always conserved as pairs.[4] That is to say, loss of a disulfide bond in a protein is typically due to mutation of both contributing cysteine residues rather than only one. However, cases in which some members of a disulfide-rich family have fewer (or more) disulfide bonds relative to others are not uncommon. Potassium channel inhibitor conkunitzin-S1 (pdb|1yl2[57]) is a member of the BPTI-like/dendrotoxin-like family (fold group 29). The structure is essentially identical with other family members (the closest structural neighbor is the Kunitz-type domain from human type VI collagen, pdb|1knt, 0.96 Å RMSD, 55 C^{α} atoms), despite the fact that conkunitzin-S1 contains only two of the three disulfide bonds that are otherwise highly conserved in the BPTI-like family. Similarly, the C-terminal region of merozoite surface protein 1 contains a tandem repeat of EGF-like domains, but the first repeat in some *Plasmodium* species has only two of the three highly conserved disulfide bonds of the EGF-like domain family, while the second repeat has all three.[58] Numerous examples are seen in which a few proteins have additional disulfide bonds relative to the majority of family members. Some spider toxins (for example ω-agatoxin IVa, pdb|1iva[59] or μ-agatoxin I, pdb|1eit[60]) have an additional disulfide bond on the key β-hairpin of their knottin-like fold. Likewise, antifungal peptide 2 of the hardy rubber tree (pdb|1p9g[61]) contains a fifth disulfide bond that is not seen in any other structures of plant lectin-like family members (fold group 23). Why certain family members seem to require fewer (or more) disulfide bonds than their homologs is not clear. Potential explanations might include different functional constraints, or different folding pathway requirements resulting from sequence variations between family members or the environment in which the organisms thrive†.

Less common variations within families are seen when a cysteine residue in a disulfide bond is contributed from similar spatial positions but different regions of the protein sequence. One such example of a migrated cysteine residue is found in the thrombospondin type 1 family of fold group 12. Members of this family have three disulfide bonds, two of which are conserved in the sequence (i.e. formed by cysteine residues that align by sequence). The third disulfide bond (shown in blue in Figure 1(c)) is formed by the

† Alignments of these four examples can be found at ftp://iole.swmed.edu/pub/disulf_aln

bonds found in few family members. (b) Calculated similarity between ωTL domains and added family members. Average values were calculated using representatives at 95% identity. A "linking pair" is defined as a pair of domains from different SCOP superfamilies that share >33% sequence identity and <2.5 Å RMSD (>25 C^{α}). (c) American pokeweed antifungal protein PAFP-S (pdb|1dkc) and gurmar plant signaling protein gurmarin (pdb|1c4e) are assigned to the same SCOP superfamily but share a lower level of similarity to each other than to ωTL conotoxins TxVIa (pdb|1fu3) and TxVII (pdb|1f3k). Boldface letters in the alignment indicate identical residues between the domains assigned to different SCOP superfamilies. (d) Very small knottin-like domains: cysteine-rich repeat from human ErbB2 (pdb|1s78 A171-A192), DPY module from *Drosophila melanogaster* dumpy protein (pdb|1oig), and the knottin-like subdomain of *T. stejnegeri* stecrisp (pdb|1rc9 A166-A180). The C^{α} traces are shown in order to clarify which β-strand contributes each "hanging" cysteine side-chain.

third and fourth cysteine residues in thrombospondin (TSP) (pdb|1lsl[62]) but by the first and fourth cysteine residues in F-spondin (pdb|1szl, 1vex; K. Paakkonen *et al.*, unpublished results). Although these residues (the third cysteine residue of TSP and the first cysteine residue of F-spondin) are separated by ∼25 amino acid residues in the sequence, they are located in approximately equivalent spatial locations. When the TSP and F-spondin domains are superimposed (average RMSD: 3.4 Å, 49 C^{α} atoms), the S atoms of the migrated cysteine residues are ∼4.2 Å apart. Cases such as these are intriguing because they suggest that maintaining the fold of a particular family (in this case, a very oblong three-strand meander β-sheet) may require additional stabilization in a specific region of the structure.

More generally, however, examples of shared disulfide-bonding requirements are not seen among fold groups. Families within a fold group are often structurally stabilized by different numbers of disulfide bonds that cross-link different pairs of structure elements (for example, see Figure 1(b)–(d)). Variations among bonding patterns suggests that while these domains do require disulfide bonds to maintain the protein fold, the specific arrangement of those bonds within the structure may not be particularly important. In fact, of the 17 fold groups in this classification that include more than one family, the only cases in which all members share the same disulfide bonding patterns are the simple α-hairpins and β-hairpins. The disulfide bond connectivity patterns within each fold group are summarized in Supplementary Data (Supplementary Figure 2).

Homologs with different disulfide-bonding patterns

Among the most interesting examples are homologous or even identical proteins shown to have different disulfide-bonding patterns. The family of disintegrins (fold group 21) includes proteins from *Viperidae* and *Crotalidae* snake venoms, which inhibit biological processes such as platelet aggregation and tumor invasion by binding to integrins of the β1 and β3 classes.[63] Despite their high level of sequence similarity (typically >40% identity among family members), the solved structures of representative disintegrins have revealed that these proteins have quite different topologies and disulfide-bonding patterns (Figure 6(a)). These representatives can be divided into four groups on the basis of disulfide connectivities. Grouping these proteins on the basis of similarity of structural fold roughly parallels the groupings by disulfide-bonding patterns. Members of the kistrin/trimestatin/flavoridin subset superimpose reasonably well with schistatin and three novel *Echis carinatus* disintegrin polypeptides (average RMSD between the subsets: 2.31 Å, 61 C^{α} atoms), which is not unexpected considering the four disulfide bonds they have in common. Conversely, echistatin and obtustatin also have four disulfide bonds in common, as well as nearly 50% sequence identity, but have quite different folds. Obtustatin has a compact, globular shape, while echistatin adopts a more extended conformation, similar to the other disintegrin structures. Salmosin is unlike all other disintegrins with solved structure both in structural fold and disulfide-bonding pattern. The most conserved structural feature among disintegrins is a β-hairpin containing the RGD motif, which is involved in binding integrin. Although the disintegrins all perform the same general function (integrin-binding), they are relatively selective for different integrin–ligand interactions.[64] Additionally, some of these proteins function as homodimers (schistatin),[65] others as heterodimers (*E. carinatus* novel disintegrin),[66] and still others as monomers (kistrin).[67] It has been suggested that the different integrin inhibition specificities may result from the variations in surface charge distribution or the striking conformational differences observed between family members.[68,69] This family is intriguing, in that, despite such clear sequence and functional similarity, the disintegrins exhibit significant variations in disulfide-bonding patterns, structural fold, and dimerization state.

In a related example, different disulfide-bonding patterns are seen for the same protein domain. The somatomedin B (SMB) domain of human vitronectin, an adhesive glycoprotein found in blood, contains binding sites for plasminogen activator inhibitor type-1 (PAI-1), urokinase-type plasminogen activator receptor, and integrins. The three structures currently available of the SMB domain all have different disulfide-bonding patterns (Figure 6(b)).[70–72] Interestingly, it was observed that several alternative bonding patterns would be compatible with the same fold.[71] For example, two of the SMB domain structures in the PDB (pdb|1oc0 and pdb|1ssu) superimpose with 2.0 Å RMSD over 36 C^{α} atoms, despite having only one of four disulfide bonds in common. Furthermore, it was demonstrated that the PAI-1 binding function of this domain was retained by these dissimilar folds. As the only shared feature of these folds was a short α-helix containing the previously identified key functional residues (Figure 6(b)), it was suggested that function is maintained because each of the disulfide-bonding patterns is compatible with the formation of this essential secondary structure element.[72] While only one native bonding pattern apparently exists in human blood PAI-1 (*aabcdbcd*),[72,73] it is nonetheless interesting to note the dramatic variations in global fold and disulfide-bonding patterns that are tolerated by this domain without sacrificing function.

Disulfide-bonding patterns observed in small protein domains

Occurrences of disulfide-bonding patterns in proteins were analyzed by Benham & Jafri.[74] In their study, the bonding patterns observed in 186 non-identical protein chains from the PDB structure database and National Biomedical Research

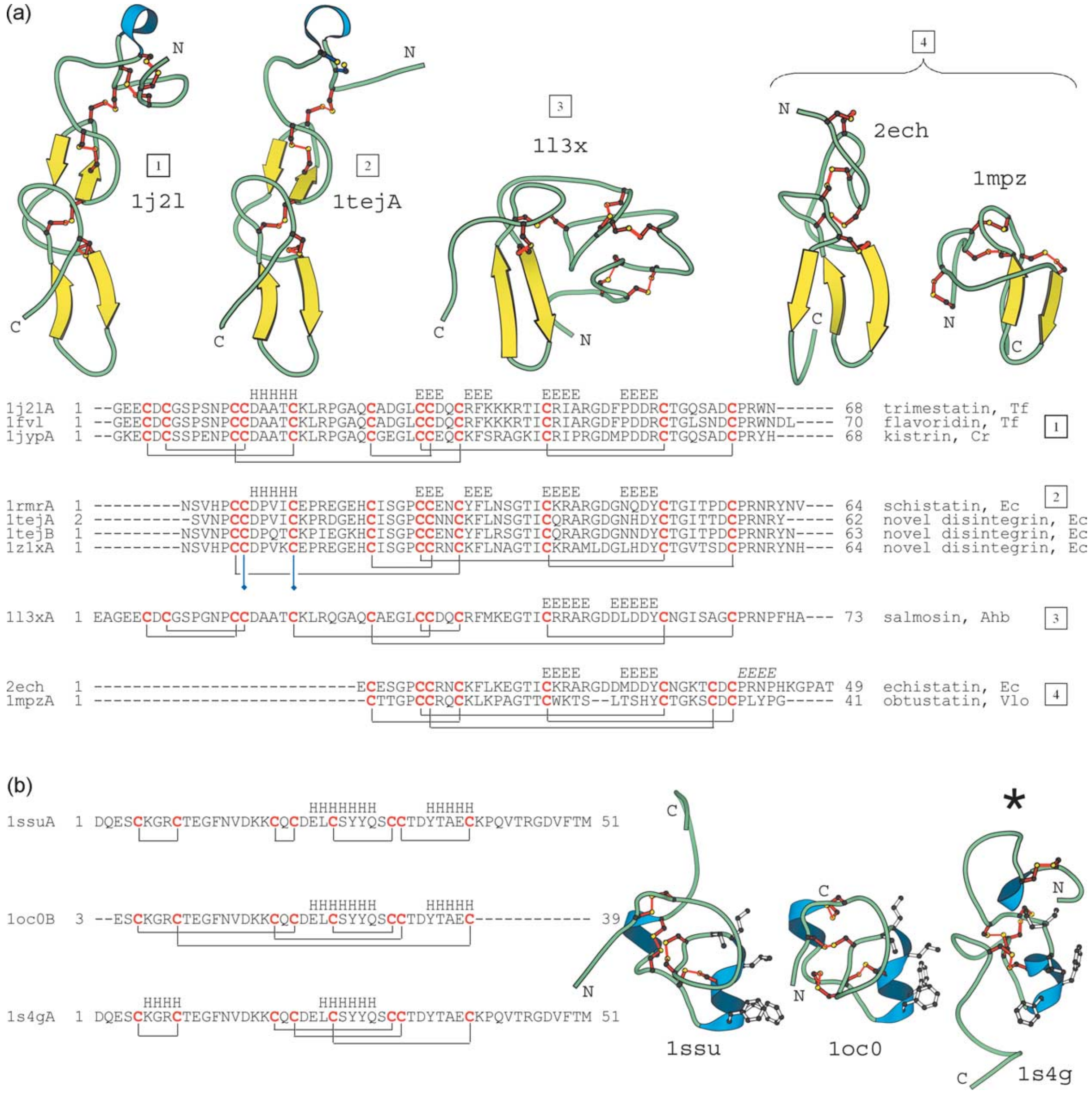

(a)
N
1
1j2l
C
N
2
1tejA
C
3
1l3x
C
N
4
N
2ech
1mpz
C
N
C
HHHHH EEE EEE EEEE EEEE
1j2lA 1 --GEECDCGSPSNPCCDAATCKLRPGAQCADGLCCDQCRFKKKRTICRIARGDFPDDRCTGQSADCPRWN------ 68 trimestatin, Tf
1fvl 1 --GEECDCGSPSNPCCDAATCKLRPGAQCADGLCCDQCRFKKKRTICRIARGDFPDDRCTGLSNDCPRWNDL---- 70 flavoridin, Tf
1jypA 1 --GKECDCSSPENPCCDAATCKLRPGAQCGEGLCCEQCKFSRAGKICRIPRGDMPDDRCTGQSADCPRYH------ 68 kistrin, Cr
1
HHHHH EEE EEE EEEE EEEE
1rmrA 1 ---------NSVHPCCDPVICEPREGEHCISGPCCENCYFLNSGTICKRARGDGNQDYCTGITPDCPRNRYNV--- 64 schistatin, Ec
1tejA 2 ----------SVNPCCDPVICKPRDGEHCISGPCCNNCKFLNSGTICQRARGDGNHDYCTGITTDCPRNRY----- 62 novel disintegrin, Ec
1tejB 1 ---------NSVNPCCDPQTCKPIEGKHCISGPCCENCYFLRSGTICQRARGDGNNDYCTGITPDCPRNRYN---- 63 novel disintegrin, Ec
1z1xA 1 ---------NSVHPCCDPVKCEPREGEHCISGPCCRNCKFLNAGTICKRAMLDGLHDYCTGVTSDCPRNRYNH--- 64 novel disintegrin, Ec
2
EEEEE EEEEE
1l3xA 1 EAGEECDCGSPGNPCCDAATCKLRQGAQCAEGLCCDQCRFMKEGTICRRARGDDLDDYCNGISAGCPRNPFHA--- 73 salmosin, Ahb
3
EEEE EEEE EEEE
2ech 1 ---------------------------ECESGPCCRNCKFLKEGTICKRARGDDMDDYCNGKTCDCPRNPHKGPAT 49 echistatin, Ec
1mpzA 1 ---------------------------CTTGPCCRQCKLKPAGTTCWKTS--LTSHYCTGKSCDCPLYPG----- 41 obtustatin, Vlo
4
(b)
HHHHHHH HHHHH
1ssuA 1 DQESCKGRCTEGFNVDKKCQCDELCSYYQSCCTDYTAECKPQVTRGDVFTM 51
HHHHHHH HHHHH
1oc0B 3 --ESCKGRCTEGFNVDKKCQCDELCSYYQSCCTDYTAEC------------ 39
HHHH HHHHHH
1s4gA 1 DQESCKGRCTEGFNVDKKCQCDELCSYYQSCCTDYTAECKPQVTRGDVFTM 51
C
N
C
N
*
N
C
1ssu
1oc0
1s4g

Figure 6. Variations in disulfide-bonding patterns. (a) Disintegrins with solved structures are grouped into four sets on the basis of disulfide bonding patterns. In some cases, similar bonding patterns result in similar topologies; for example, trimestatin (pdb|1j2l) and chain A of *E. carinatus* disintegrin heterodimer (pdb|1tej). In another case, disintegrins adopt non-similar topologies despite identical bonding patterns: echistatin (pdb|2ech) and obtustatin (pdb|1mpz). Salmosin (pdb|1l3x) differs from the other structure representatives in both bonding pattern and fold. Cysteine residues involved in intramolecular disulfide bonds are shown in red; cysteine residues involved in intermolecular disulfide bonds are shown in blue. In the alignment, secondary structure elements not conserved in all proteins of a subset are shown in italics. Intramolecular disulfide bonds are shown as black lines; intermolecular disulfide bonds are shown as blue diamond arrows. Species abbreviations are as follows: Tf, *Trimeresurus flavoviridis*; Cr, *Calloslasma rhodostoma*; Ec, *E. carinatus*; Ahb, *Agkistrodon halys brevicaudus*; Vlo, *Vipera lebetina obtusa*. (b) Three structures of the SMB domain of human vitronectin have different disulfide bonding patterns. Functional residues are shown in ball-and-stick format. The native bonding pattern is indicated by an asterisk (*).

Foundation protein sequence database (now a part of UniProt)[75] were specified and evaluated in terms of two intrinsic properties, symmetry and reducibility. We have repeated this analysis with the domains in our classification as described in Materials and Methods, and the results are shown in Table 2.

The most striking result is that the number of observed reducible patterns is much lower than what is predicted when assuming all patterns are equally probable. This clearly suggests that irreducible disulfide-bonding patterns offer some kind of evolutionary advantage over reducible patterns. The most obvious explanation would be that irreducible patterns result in more stable structures than their reducible counterparts. Because a reducible pattern can, by definition,

Table 2. Frequencies of bonding patterns for small protein domains with *N* disulfide bonds

	Families					Representative domains (95%)				
		Symmetric patterns		Reducible patterns			Symmetric patterns		Reducible patterns	
N	Families with *N* bonds	Obs	Pred	Obs	Pred	Domains with *N* bonds	Obs	Pred	Obs	Pred
2	38	38	38.0	1	12.7	267	267	267.0	5	89.0
3	36	21	16.8	5	12.0	505	294	235.7	129	168.3
4	10	2	2.4	3	3.0	84	2	20.0	64	24.8
5	5	0	0.4	0	1.3	11	0	0.9	0	2.8
6	1	0	3.2×10^{-2}	0	0.2	2	0	6.4×10^{-2}	0	0.4
9	1	0	7.8×10^{-4}	1	0.1	1	0	7.8×10^{-4}	1	0.1
Σ		**61**	**57.6**	**10**	**29.3**		**563**	**523.7**	**199**	**285.4**

be divided into independent cross-linked regions, each subpattern found within a reducible pattern could be responsible for locally stabilizing areas within a protein, but still allow for undesirable flexibility between those regions. Furthermore, entropic stabilization is predicted to be greater for irreducible patterns.[76] In the case of small protein domains that lack a hydrophobic core, the greater complexity of irreducible patterns may often be essential for maintaining the protein fold.

In contrast, occurrences of symmetric patterns in the set are slightly higher than expected if all patterns are equally probable. The total number of symmetric patterns observed in our set (61 families or 563 representative domains) is 6–8% greater than the sum predicted from random. This minor overrepresentation may indicate that some kind of biological advantage is gained by symmetric bonding patterns as well. This may also be a reflection of the high frequency of symmetric fold topologies seen in proteins.

Thus, we find that symmetry was slightly overrepresented, while reducibility was highly underrepresented in the disulfide-bonding patterns of small protein domains. Notably, Benham & Jafri found that both symmetry and reducibility were greatly overrepresented in their dataset.[74] There are several explanations that account for these conflicting results. One factor is the difference in sample size: Benham & Jafri's work was completed about 12 years ago when the number of proteins with confidently established disulfide-bonding patterns was quite small. Additionally, Benham & Jafri's study was not limited to small protein domains. It is likely that the biological and physical forces guiding disulfide-bonding patterns are different for larger proteins, which could contribute to the differences between these results. Furthermore, their analysis considered entire polypeptide chains rather than individual domains. The inclusion of multi-domain proteins would greatly increase the observed occurrences of reducible bonding patterns relative to our survey of only single-domain representatives.

A related analysis was performed by Hartig *et al.*, who examined occurrences of each specific two and three-bond pattern.[77] Their observed frequencies of those patterns are very well correlated with the bonding pattern frequencies in our set of disulfide-rich domains (data not shown).

Functions of disulfide-rich domains

General domain functions

Disulfide-rich domains have been demonstrated to accomplish a wide variety of cellular roles, which can be divided into three functional categories: communication, structural, and enzymatic. By far the most prevalent of these three is communication. Popular functions of disulfide-rich domains in this category are hormones, growth factors, pheromones, enzyme inhibitors, ligand-binding domains of extracellular receptors, etc. A related set of functions includes tasks of an offensive (e.g. immobilizing prey by interfering with ion channel activity) or defensive (e.g. inducing cell lysis of microbial predators) nature. With the exception of the ligand-binding domains, most disulfide-rich domains with communication roles are single-domain (i.e. not subdomains of larger polypeptides). Furthermore, these domains are predominantly extracellular.

Other disulfide-rich domains are theorized to play structural roles. Most of these examples are subdomains within larger proteins, such as the PSI domain of the human Met receptor (fold group 26), which is proposed to serve as a wedge to properly orient the propeller-like and immunoglobulin domains of this protein.[78] There are also a few single-domain disulfide-rich proteins with structural roles, including the hinge protein (non-heme 11 kDa protein) of the cytochrome bc_1 complex (fold group 2), which is essential for complex formation.[79]

Additionally, there are two disulfide-rich proteins that have been demonstrated to perform enzymatic functions. These are phospholipase A2 (fold group 10) and the light chain of methylamine dehydrogenase (fold group 22).

It should be noted, however, that many disulfide-rich domains have not been functionally characterized. In many cases, a cellular or physiological role has been established but the molecular target is not yet identified, and then there are some domains for which the function is completely unknown.

Functional convergence among disulfide-rich domains

There are many examples of similar functions that are performed by a number of unrelated disulfide-rich domains. Cases of similar functions performed by domains within different families and/or fold groups are most likely examples of convergent evolution. Of course, it is possible that some examples may reflect remote homology that cannot be established with confidence, given the currently available sequence, structure, and functional data.

The most prevalent function among the domains in this classification is inhibition of the activity of many different types of ion channels. Disulfide-rich toxins have been demonstrated to block channels that conduct a variety of different ions (including Na^+, K^+, Ca^{2+}, Cl^-, or non-specific cations) with a variety of different gating mechanisms (voltage-gated, ligand-gated, or mechanosensitive). In our classification, nine fold groups and ten families include at least one protein that is a known or putative ion channel inhibitor. Among these, there are several examples of disulfide-rich toxins from related species found in different families and in different fold groups: e.g. sea anemone toxins with right-handed three-helix bundle or three-strand antiparallel β-sheet folds (fold groups 3 and 13), scorpion toxins with short α-hairpin or knottin-like folds (fold groups 1 and 23), and conotoxins with knottin-like or small, disulfide-closed loop folds (fold groups 23 and 35). Another common function is the inhibition of various serine proteases, including trypsin, chymotrypsin, elastase, plasmin, thrombin, factor Xa, factor VIIa, etc. Despite their different specificities and global folds, many of these inhibitors are believed to share a common mechanism. Comparison of the backbone angles of the inhibitory loops of serine protease inhibitors from unrelated families has shown that these regions adopt very similar conformations.[80] Serine protease inhibitors are found in eight fold groups and ten families of our classification. Also, many disulfide-rich domains are annotated as antimicrobial or defensin proteins. The presumed mechanism of these domains is to induce cell lysis of a microbial predator by disrupting the cell membrane, although the details of such a mechanism are unclear. Moreover, some of these proteins are thought to target specific extracellular receptors rather than interact directly with the membrane. Putative antimicrobial or defensin proteins are found in six fold groups and nine families of our classification. Membrane disruption is also the suggested mechanism for a number of non-defensive proteins, such as snake venom cardiotoxins. The functions described in the preceding examples are most commonly performed by whole (i.e. single-domain) proteins. Disulfide-rich subdomains, on the other hand, are frequently involved in the binding of molecules found in abundance on or near the cell surface, such as heparin, chitin, integrins, and TGFβ superfamily members.

Functional divergence of disulfide-rich domains

Examples of the divergent evolution of homologous disulfide-rich proteins to various molecular or cellular functions are common as well. Often, these domains perform related functions, such as spider toxins that block various types of ion channels (Figure 7(a)), disintegrins that inhibit the function of different integrin receptors with high selectivity, or α-conotoxins that bind to and inhibit assorted subtypes of nicotinic acetylcholine receptors. In these rapidly evolving families, numerous highly similar proteins are frequently found in the same species.

In other cases, the homologous proteins perform more distant functions. In the BPTI-like family, for example, some members inhibit serine proteases, while others block K^+ or Ca^{2+} channels. The BPTI-like family includes an interesting example of mechanistic divergence, while cellular function is retained. As previously mentioned, many serine protease inhibitors appear to share a common mechanism. In these canonical inhibitors, including the majority of inhibitors in this family, the inhibitory loop forms one β-strand of a distorted antiparallel β-sheet at the active site of the protease.[80] However, in a small number of BPTI-like

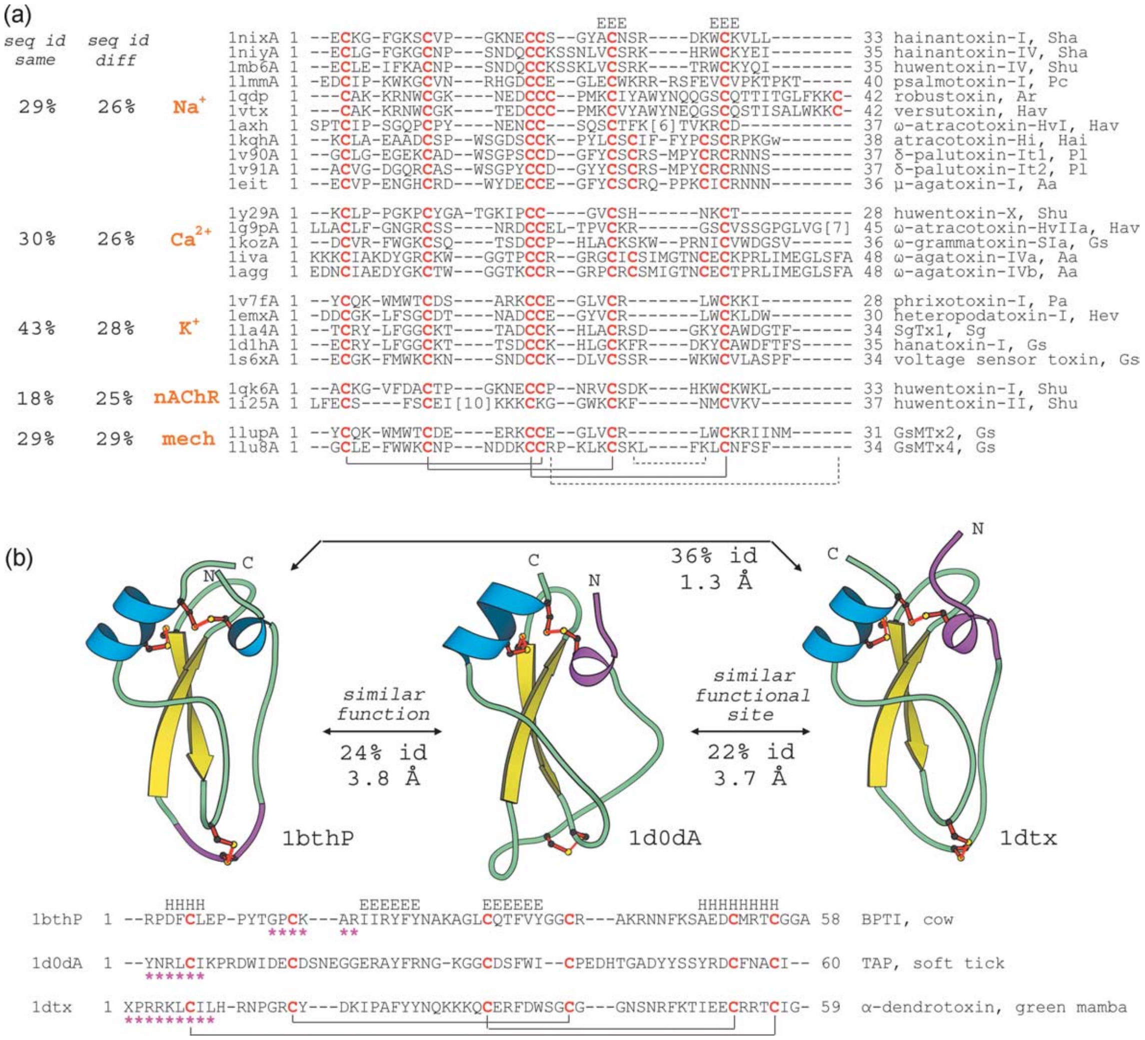

(a)
seq id same
seq id diff
EEE EEE
29% 26% Na+
1nixA 1 --ECKG-FGKSCVP---GKNECCS--GYACNSR----DKWCKVLL-------- 33 hainantoxin-I, Sha
1niyA 1 --ECLG-FGKGCNP---SNDQCCKSSNLVCSRK----HRWCKYEI-------- 35 hainantoxin-IV, Sha
1mb6A 1 --ECLE-IFKACNP---SNDQCCKSSKLVCSRK----TRWCKYQI-------- 35 huwentoxin-IV, Shu
1lmmA 1 -EDCIP-KWKGCVN---RHGDCCE--GLECWKRR-RSFEVCVPKTPKT----- 40 psalmotoxin-I, Pc
1qdp 1 ---CAK-KRNWCGK----NEDCCC--PMKCIYAWYNQQGSCQTTITGLFKKC- 42 robustoxin, Ar
1vtx 1 ---CAK-KRNWCGK----TEDCCC--PMKCVYAWYNEQGSCQSTISALWKKC- 42 versutoxin, Hav
1axh 1 SPTCIP-SGQPCPY----NENCC---SQSCTFK[6]TVKRCD---------- 37 ω-atracotoxin-HvI, Hav
1kqhA 1 --KCLA-EAADCSP--WSGDSCCK--PYLCSCIF-FYPCSCRPKGw------- 38 atracotoxin-Hi, Hai
1v90A 1 --GCLG-EGEKCAD--WSGPSCCD--GFYCSCRS-MPYCRCRNNS-------- 37 δ-palutoxin-It1, Pl
1v91A 1 --ACVG-DGQRCAS--WSGPYCCD--GYYCSCRS-MPYCRCRNNS-------- 37 δ-palutoxin-It2, Pl
1eit 1 --ECVP-ENGHCRD---WYDECCE--GFYCSCRQ-PPKCICRNNN-------- 36 μ-agatoxin-I, Aa
30% 26% Ca2+
1y29A 1 --KCLP-PGKPCYGA-TGKIPCC----GVCSH------NKCT----------- 28 huwentoxin-X, Shu
1g9pA 1 LLACLF-GNGRCSS----NRDCCEL-TPVCKR------GSCVSSGPGLVG[7] 45 ω-atracotoxin-HvIIa, Hav
1kozA 1 --DCVR-FWGKCSQ----TSDCCP--HLACKSKW--PRNICVWDGSV------ 36 ω-grammatoxin-SIa, Gs
1iva 1 KKKCIAKDYGRCKW---GGTPCCR--GRGCICSIMGTNCECKPRLIMEGLSFA 48 ω-agatoxin-IVa, Aa
1agg 1 EDNCIAEDYGKCTW---GGTKCCR--GRPCRCSMIGTNCECTPRLIMEGLSFA 48 ω-agatoxin-IVb, Aa
43% 28% K+
1v7fA 1 --YCQK-WMWTCDS----ARKCCE--GLVCR-------LWCKKI--------- 28 phrixotoxin-I, Pa
1emxA 1 -DDCGK-LFSGCDT----NADCCE--GYVCR-------LWCKLDW-------- 30 heteropodatoxin-I, Hev
1la4A 1 --TCRY-LFGGCKT----TADCCK--HLACRSD----GKYCAWDGTF------ 34 SgTx1, Sg
1d1hA 1 --ECRY-LFGGCKT----TSDCCK--HLGCKFR----DKYCAWDFTFS----- 35 hanatoxin-I, Gs
1s6xA 1 --ECGK-FMWKCKN----SNDCCK--DLVCSSR----WKWCVLASPF------ 34 voltage sensor toxin, Gs
18% 25% nAChR
1qk6A 1 --ACKG-VFDACTP---GKNECCP--NRVCSDK----HKWCKWKL-------- 33 huwentoxin-I, Shu
1i25A 1 LFECS----FSCEI[10]KKKCKG--GWKCKF------NMCVKV--------- 37 huwentoxin-II, Shu
29% 29% mech
1lupA 1 --YCQK-WMWTCDE----ERKCCE--GLVCR-------LWCKRIINM------ 31 GsMTx2, Gs
1lu8A 1 --GCLE-FWWKCNP---NDDKCCRP-KLKCSKL----FKLCNFSF-------- 34 GsMTx4, Gs
(b)
36% id
1.3 Å
similar function
24% id
3.8 Å
similar functional site
22% id
3.7 Å
C
N
1bthP
1d0dA
1dtx
HHHH EEEEEE EEEEEE HHHHHHHH
1bthP 1 --RPDFCLEP-PYTGPCK---ARIIRYFYNAKAGLCQTFVYGGCR---AKRNNFKSAEDCMRTCGGA 58 BPTI, cow
**** **
1d0dA 1 --YNRLCIKPRDWIDECDSNEGGERAYFRNG-KGGCDSFWI--CPEDHTGADYYSSYRDCFNACI-- 60 TAP, soft tick
******
1dtx 1 XPRRKLCILH-RNPGRCY---DKIPAFYYNQKKKQCERFDWSGCG---GNSNRFKTIEECRRTCIG- 59 α-dendrotoxin, green mamba
****

Figure 7. Functional divergence of disulfide-rich homologs. (a) Alignment of spider toxins, grouped by channel type targeted: sodium (Na^+), calcium (Ca^{2+}), potassium (K^+), nicotinic acetylcholine receptors (nAChR), mechanosensitive (mech). Seq id same refers to the average sequence identity among spider toxins that block the same type of channels. Seq id diff refers to the average sequence identity between spider toxins that block a specific channel type and the rest of the spider toxins (inhibitors of other channel types). Percentage identities were calculated using a non-redundant set of spider toxins with solved structures and known channel types. Abbreviations of the spider species are as follows: Aa, *Agelenopsis aperta*; Ar, *Atrax robustus*; Gs, *Grammostola spatulata*; Hav, *Hadronyche versuta*; Hai, *Hadronyche infensa*; Hev, *Heteropodidae venatoria*; Pc, *Psalmopoeus cambridgei*; Pl, *Paracoelotes luctuosus*; Pa, *Phrixotrichus auratus*; Sha, *Selenocosmia hainana*; Shu, *Selenocosmia huwena*; Sg, *Scodra griseipes*. Highly conserved disulfide bonds are indicated by continuous lines; additional disulfide bonds found in some family members are indicated by broken lines. (b) Serine protease inhibitors in the BPTI-like family perform the same function using different parts of the protein fold: bovine pancreatic trypsin inhibitor (BPTI, pdb|1bth, chain P) and factor Xa inhibitor (also called TAP or tick anticoagulant protein, pdb|1d0d, chain A). Other BPTI-like family members utilize the same region of the protein to perform different functions: the N-terminal residues are the key functional residues of TAP (pdb|1d0d; chain A) and green mamba snake α-dendrotoxin (pdb|1dtx). Functional sites are indicated in purple in the MOLSCRIPT diagrams and with asterisks (*) in the alignment.

family members, such as tick anticoagulant protein (TAP) and ornithodorin, the inhibitory activity is accomplished by the N-terminal residues, which run parallel to the protease active site.[81,82] Notably, the N-terminal region of this structure also contributes the key functional residues of α-dendrotoxin,[83] a non-protease-inhibitor member of this family. Interestingly, the toxin members of this family share higher levels of sequence and structural similarity with the canonical-type inhibitors rather than the TAP-like inhibitors with which they share a common functional site (Figure 7(b)).

Materials and Methods

Identification of disulfide-rich protein domains

We consider a protein to be potentially disulfide-rich if the structure contains two disulfide bonds within 23 Å. This distance cutoff was determined empirically on the basis of protein domains previously noted as disulfide-rich in the Small proteins class of the SCOP database. A locally mirrored version of the Protein Data Bank (PDB; current through August 2, 2005) was searched for structures containing two or more disulfide bonds within 23 Å.[84] A disulfide bond was assumed to exist between two cysteine residues if their gamma sulfur atoms were less than 3.5 Å apart. The sequences of individual PDB chains from those structures identified by this automated search were extracted and clustered on the basis of sequence identity using the BLASTCLUST program (I. Dondoshansky & Y. Wolf, unpublished results†) using a 50% identity threshold and a length coverage threshold of 90% on each sequence. A representative of each cluster was examined in order to identify and exclude non-disulfide-rich chains within PDB structures and proteins in which the cysteine side-chains contribute to metal-binding rather than disulfide bonds.

For the purposes of this study, we were interested in disulfide-rich protein domains with structural folds stabilized primarily by the formation of disulfide bonds rather than by the hydrophobic core of the protein. Such proteins typically have a very small hydrophobic core and few secondary structure elements. Therefore, protein structures with a significant hydrophobic core and many secondary structure elements were removed from the set of structures that had been identified in the automated search. For example, the structure of *Macadamia integrifolia* antimicrobial protein MiAMP1 (pdb|1c01),[85] which contains three disulfide bonds and a substantial hydrophobic core (eight-stranded β-sandwich with Greek key topology), was excluded from our classification because the disulfide bonds appear to be incidental to the stability of the protein's structural fold. Likewise, proteins for which non-disulfide-rich homologs or structural analogs are known were also excluded, such as the *Aspergillus giganteus* antifungal protein AGAFP (pdb|1afp),[86] which adopts an OB-like fold. We consider structures such as these to be better described by their fold topology than by their disulfide bonds, and are therefore classified more appropriately with their non-disulfide-rich structural neighbors. Such cases were identified by manual examination of cluster representatives. In general, the presence of a substantial hydrophobic core and many secondary structure elements corresponds to the size or length of the domain in question. The average size of domains set aside in this manual filtering step is 215 residues, compared to the average size of 57 residues for the disulfide-rich domains included in this study. The size distribution of the domains included *versus* excluded in this classification can be seen in Supplementary Data (Supplementary Figure 1).

Classification of disulfide-rich protein domains

Fold groups of structurally similar disulfide-rich domains

The disulfide-rich protein domains identified from the PDB were classified according to a two-tier hierarchy. The first tier is the fold group level, which is based on structural similarity between protein domains. Domains in the same fold group share a common structural core with topology that is either identical or related by circular permutation. Circular permutation is usually defined as an evolutionary event that results in the generation of folds that share a similar packing of secondary structural elements, while differing in their topological connectivities. In some cases, there is known homology between circularly permuted domains (e.g. the two subdomains of Bowman–Birk inhibitors described previously); these domains are placed in the same family (see definition of the family level below). Disulfide-rich domains from different families in the same fold group signify cases in which the protein structures are related by circular permutation but for which an evolutionary relationship cannot currently be supported, although it is possible that future sequence or function data will reveal that these domains are in fact distant evolutionary relatives. Therefore, it should be noted that when proteins from different families are related by permutation we mean an "artificial permutation" of the structures so that they can be better superimposed than if they were not permuted, and an evolutionary connection is not implied in these cases.

Thus, within a fold group, all members share a common structural core comprised of secondary structure elements found in the same spatial arrangement with either identical or circularly permuted topology. Each fold group can be differentiated from all other fold groups by either the architecture of secondary structure elements, topological connectivity, or both. For example, fold groups 23 and 24 are both characterized by a knottin-like domain core (as described in detail in another section) with four separate structure elements contributing the cysteine residues that form the disulfide cross, but these two fold groups can be distinguished by the topological connections of these four elements: domains in fold group 24 have left-handed connectivity between the four elements while domains in fold group 23 have right-handed connectivity between the first three elements but left-handed connectivity between the last three elements. Generally, all fold group members also share the same secondary structure composition in the domain core. Therefore, the description of the common structural core in Table 1 is usually enough to visualize the structural pattern and to show a distinction between different fold groups. In many cases, it would also be possible to attribute a new protein to a fold group just by this description; for example, such patterns as helical bundles or hairpins are well-known and recognizable super-secondary structures. However, only abbreviated descriptions are given in a few instances where the topology of a fold group is more complicated than can be explained succinctly. For example, *E. raikovi* pheromone ER-23 (pdb|1ha8, fold group 7) adopts a fold comprised of five helices, where helices 1 and 5 are adjacent and roughly parallel, helices 3 and 4 are antiparallel and lie approximately perpendicular to helices 1 and 5, helix 2 is found on a connecting loop, and the connections between elements are mostly right-handed with the exception of left-handed connectivity between helices 2, 3, and 4; instead, this fold group is simply described as a five-helix globular array. Although most fold group members share the same core, some exceptions and deviations from strict definitions are observed among the disulfide-rich proteins due to the atypical nature of small protein domains (i.e. the high incidence of small and irregular secondary structure elements). Broken or bulging β-strands and single-turn α-helices and 3_{10} helices are quite common in these proteins, and can result in non-identical secondary structure content within fold groups. For example, short 3_{10} helices are often found in place of α-helices. Similarly, irregularities in β-strands can result in predominantly coil regions with only very short segments retaining the hydrogen bonding patterns characteristic of true β-strand elements. Another interesting example is seen in the cellulose-binding/

† ftp://ftp.ncbi.nih.gov/blast/

docking domains of fold group 18. Two roughly perpendicular hairpins comprise the structural core of both members of this family, the cellulose-binding domain of xylanase A (CBDx) from *P. fluorescens* and the cellulose-docking domain of endoglucanase Cel45 (CDDe) from *P. equi*. However, CBDx contains two β-hairpins, while CDDe includes one β-hairpin and one hairpin formed by an α-helix and a coil region (Figure 4(a)). In these homologous proteins, an equivalent region of the protein adopts quite different secondary structure conformations. It should also be noted that our fold group definition does not require all members of the same fold group to share the same disulfide-bonding connectivity. The relationship between disulfide-bonding patterns and structural topology is discussed in another section.

Fold groups were determined by visual inspection. The significance of these fold groups is evaluated in Supplementary Data (Supplementary Figure 2), which details the average structural similarity calculated for representatives of the same fold group compared to the structural similarity between representatives from different fold groups. One traditional method for assessing similarity between protein structures is by calculating RMSD values on the basis of structural alignments. The average RMSD within our disulfide-rich protein families is 2.20(±0.84) Å, while the average RMSD within our fold groups (excluding comparisons of members of the same family) is only slightly higher at 2.34(±0.22) Å. However, members of the same family can be superimposed over 39 C^{α} atoms on average, while members of the same fold group (but different families) are alignable over only 18 C^{α} atoms on average. This example highlights one negative aspect of RMSD: due to the length-dependent nature of the calculation, the comparison of different results is not always straightforward. There are other drawbacks to RMSD analysis as well. First, structural similarity statistics in general (including RMSD) tend to be unreliable for small proteins due to the shortness of their polypeptide chains. Also, large RMSD values do not necessarily ensure a lack of structural similarity, most notably in cases of proteins with internal rearrangements such as circular permutations or domain swaps. Furthermore, while protein structures are more resilient to change than their sequences, they are not immutable and are intrinsically flexible. Nevertheless, while RMSD is not an infallible method, it does often give a reasonable indication of the structural similarity between proteins. RMSD calculations for each family and fold group can be found in Supplementary Data (Supplementary Figure 2).

Families of homologous disulfide-rich domains

The second tier is the family level, which reflects an evolutionary relationship between domains. The superfamily level of SCOP was used as the starting point for establishing the disulfide-rich families in this study. The ~500 disulfide-rich domains (17% of domains in this study) that were not classified in SCOP were merged with the existing SCOP superfamilies if significant overall sequence similarity was recognized (typically >40% identity). This resulted in 110 preliminary groupings of homologs (87 associated with SCOP superfamilies and 23 additional sets of domains not classified in SCOP). These preliminary groupings were then further studied to identify more distant homology links by the following process. First, manual structural alignments were generated for potentially related groupings of domains. These alignments were used to determine whether the groupings in question share significant sequence similarity (typically >20% identity), and to recognize similarity among their cysteine conservation and disulfide-bonding patterns. Indications of functional similarity (e.g. domains performing related molecular or cellular functions, or domains with key functional residues found in similar spatial locations) that were identified by searching the literature were considered as further verification of these more distant homology relationships. After merging certain preliminary groupings on the basis of these criteria (sequence and functional similarity), a final total of 98 families of disulfide-rich domains were established. Examples of disulfide-rich families in this study that differ from SCOP superfamilies are described in Results and Discussion. Sequence similarity within the disulfide-rich families is summarized in Supplementary Data (Supplementary Figure 2). An example of the distribution of sequence identity among family members compared to sequence identity among non-homologous disulfide-rich domains is shown in Supplementary Data (Supplementary Figure 3).

Multiple alignments of disulfide-rich families and fold groups

The InsightII package was used to visualize and superimpose the structures of the disulfide-rich protein domains. Multiple structure-based alignments were constructed manually for each family and fold group based on the superpositions made in InsightII†.

Evaluation of disulfide bonding patterns

Disulfide bonding patterns in the set of small, disulfide-rich domains identified in this study were analyzed according to the intrinsic properties of symmetry and reducibility, as defined by Benham & Jafri.[74] A bonding pattern has symmetry if it reads the same from N→C and C→N. For example, if a domain has three disulfide bonds where the first cysteine residue is bonded to the second, the third to the fourth, and the fifth to the sixth, then it will read as *aabbcc* from both directions. A bonding pattern is reducible if a single cut can separate it into two discrete subpatterns, where no disulfide bond is split between the subpatterns. For example, the pattern *ababcc* is reducible because it can be cut into *abab* and *cc*, but the pattern *abcabc* is irreducible because it cannot be split into self-contained subpatterns.

The number of symmetric and reducible patterns has been tabulated in two ways: by the bonding pattern of each family (defined as the pattern of the disulfide bonds conserved in >80% of the family members) and by the bonding pattern of each representative domain after clustering all of the domains in the classification (95% identity and 95% length coverage). Both measures are considered because neither method alone provides an ideal sample: the number of families provides a very small sample size, but the counts given by representatives are biased in favor of overpopulated families such as scorpion toxins and epidermal growth factor-like domains. Furthermore, seven families from our classification with members having different disulfide bonding patterns were excluded from this analysis. For example, two domains in the cellulose-binding/docking domain family (fold group 18) have bonding pattern *aabb*

† These alignments are available at ftp://iole.swmed.edu/pub/disulf_aln

(pdb|1e8p, 1e8q)[43] while the other two domains in this family have bonding pattern *abba* (pdb|1e8r, 1qld).[87] For consistency, members of these seven families are also excluded from the calculations with representative domains.

The predicted fraction of reducible patterns with *N* disulfide bonds expected by random, equiprobable connections between cysteines can be described by the number of possible reducible patterns with *N* bonds divided by the total number of possible patterns with *N* bonds. Similar calculations give the predicted number of symmetric patterns with *N* disulfide bonds.

Acknowledgements

This work was supported by NIH grants GM67165 to N.V.G. S.C. was supported by NIH training grant T32 GM08297 to the University of Texas Southwestern Graduate Program in Molecular Biophysics.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.03.017

References

1. Krishna, S. S., Majumdar, I. & Grishin, N. V. (2003). Structural classification of zinc fingers: survey and summary. *Nucl. Acids Res.* **31**, 532–550.
2. Flory, P. J. (1956). Theory of elastic mechanisms in fibrous proteins. *J. Am. Chem. Soc.* **78**, 5222–5235.
3. Anfinsen, C. B. & Scheraga, H. A. (1975). Experimental and theoretical aspects of protein folding. *Advan. Protein Chem.* **29**, 205–300.
4. Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.* **151**, 261–287.
5. Doig, A. J. & Williams, D. H. (1991). Is the hydrophobic effect stabilizing or destabilizing in proteins? The contribution of disulphide bonds to protein stability. *J. Mol. Biol.* **217**, 389–398.
6. Betz, S. F. (1993). Disulfide bonds and the stability of globular proteins. *Protein Sci.* **2**, 1551–1558.
7. Aslund, F. & Beckwith, J. (1999). Bridge over troubled waters: sensing stress by disulfide bond formation. *Cell*, **96**, 751–753.
8. Yano, H., Kuroda, S. & Buchanan, B. B. (2002). Disulfide proteome in the analysis of protein function and structure. *Proteomics*, **2**, 1090–1096.
9. Creighton, T. E. (1992). Protein folding pathways determined using disulphide bonds. *BioEssays*, **14**, 195–199.
10. Creighton, T. E. (1997). Protein folding coupled to disulphide bond formation. *Biol. Chem.* **378**, 731–744.
11. Menez, A. (1998). Functional architectures of animal toxins: a clue to drug design? *Toxicon*, **36**, 1557–1572.
12. Craik, D. J., Simonsen, S. & Daly, N. L. (2002). The cyclotides: novel macrocyclic peptides as scaffolds in drug design. *Curr. Opin. Drug Discov. Dev.* **5**, 251–260.
13. Vita, C., Drakopoulou, E., Vizzavona, J., Rochette, S., Martin, L., Menez, A. *et al.* (1999). Rational engineering of a miniprotein that reproduces the core of the CD4 site interacting with HIV-1 envelope glycoprotein. *Proc. Natl Acad. Sci. USA*, **96**, 13091–13096.
14. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
15. Holm, L. & Sander, C. (1995). Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478–480.
16. Mas, J. M., Aloy, P., Marti-Renom, M. A., Oliva, B., Blanco-Aparicio, C., Molina, M. A. *et al.* (1998). Protein similarities beyond disulphide bridge topology. *J. Mol. Biol.* **284**, 541–548.
17. Mas, J. M., Aloy, P., Marti-Renom, M. A., Oliva, B., de Llorens, R., Aviles, F. X. & Querol, E. (2001). Classification of protein disulphide-bridge topologies. *J. Comput. Aided Mol. Des.* **15**, 477–487.
18. van Vlijmen, H. W., Gupta, A., Narasimhan, L. S. & Singh, J. (2004). A novel database of disulfide patterns and its application to the discovery of distantly related homologs. *J. Mol. Biol.* **335**, 1083–1092.
19. Gupta, A., Van Vlijmen, H. W. & Singh, J. (2004). A classification of disulfide patterns and its relationship to protein structure and function. *Protein Sci.* **13**, 2045–2058.
20. Espiritu, D. J., Watkins, M., Dia-Monje, V., Cartier, G. E., Cruz, L. J. & Olivera, B. M. (2001). Venomous cone snails: molecular phylogeny and the generation of toxin diversity. *Toxicon*, **39**, 1899–1916.
21. Escoubas, P., Diochot, S. & Corzo, G. (2000). Structure and pharmacology of spider venom neurotoxins. *Biochimie*, **82**, 893–907.
22. Gelly, J. C., Gracy, J., Kaas, Q., Le-Nguyen, D., Heitz, A. & Chiche, L. (2004). The KNOTTIN website and database: a new information system dedicated to the knottin scaffold. *Nucl. Acids Res.* **32**, D156–D159.
23. Rawlings, N. D., Tolle, D. P. & Barrett, A. J. (2004). MEROPS: the peptidase database. *Nucl. Acids Res.* **32**, D160–D164.
24. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
25. Dauplais, M., Lecoq, A., Song, J., Cotton, J., Jamin, N., Gilquin, B. *et al.* (1997). On the convergent evolution of animal toxins. Conservation of a diad of functional residues in potassium channel-blocking toxins with unrelated structures. *J. Biol. Chem.* **272**, 4302–4309.
26. Guo, M., Teng, M., Niu, L., Liu, Q., Huang, Q. & Hao, Q. (2005). Crystal structure of the cysteine-rich secretory protein stecrisp reveals that the cysteine-rich domain has a K^+ channel inhibitor-like fold. *J. Biol. Chem.* **280**, 12405–12412.
27. Nobile, M., Noceti, F., Prestipino, G. & Possani, L. D. (1996). Helothermine, a lizard venom toxin, inhibits calcium current in cerebellar granules. *Expt. Brain Res.* **110**, 15–20.
28. Wang, J., Shen, B., Guo, M., Lou, X., Duan, Y., Cheng, X. P. *et al.* (2005). Blocking effect and crystal structure of natrin toxin, a cysteine-rich secretory protein from *Naja atra* venom that targets the BK_{Ca} channel. *Biochemistry*, **44**, 10145–10152.
29. Brown, R. L., Haley, T. L., West, K. A. & Crabb, J. W. (1999). Pseudechetoxin: a peptide blocker of cyclic nucleotide-gated ion channels. *Proc. Natl Acad Sci. USA*, **96**, 754–759.

30. Morrissette, J., Kratzschmar, J., Haendler, B., el-Hayek, R., Mochca-Morales, J., Martin, B. M. *et al.* (1995). Primary structure and properties of helothermine, a peptide toxin that blocks ryanodine receptors. *Biophys. J.* **68**, 2280–2288.
31. Harrison, P. M. & Sternberg, M. J. (1996). The disulphide β-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. *J. Mol. Biol.* **264**, 603–623.
32. Huang, K., Strynadka, N. C., Bernard, V. D., Peanasky, R. J. & James, M. N. (1994). The molecular structure of the complex of *Ascaris* chymotrypsin/elastase inhibitor with porcine elastase. *Structure*, **2**, 679–689.
33. Kuettner, E. B., Hilgenfeld, R. & Weiss, M. S. (2002). The active principle of garlic at atomic resolution. *J. Biol. Chem.* **277**, 46402–46407.
34. Brown, L. R., Mronga, S., Bradshaw, R. A., Ortenzi, C., Luporini, P. & Wuthrich, K. (1993). Nuclear magnetic resonance solution structure of the pheromone ER-10 from the ciliated protozoan *Euplotes raikovi*. *J. Mol. Biol.* **231**, 800–816.
35. Barthe, P., Yang, Y. S., Chiche, L., Hoh, F., Strub, M. P., Guignard, L. *et al.* (1997). Solution structure of human $p8^{MTCP1}$, a cysteine-rich protein encoded by the MTCP1 oncogene, reveals a new α-helical assembly motif. *J. Mol. Biol.* **274**, 801–815.
36. Hatano, K., Kojima, M., Tanokura, M. & Takahashi, K. (1995). Primary structure, sequence-specific ^{1}H-NMR assignments and secondary structure in solution of bromelain inhibitor VI from pineapple stem. *Eur. J. Biochem.* **232**, 335–343.
37. Sawano, Y., Muramatsu, T., Hatano, K., Nagata, K. & Tanokura, M. (2002). Characterization of genomic sequence coding for bromelain inhibitors in pineapple and expression of its recombinant isoform. *J. Biol. Chem.* **277**, 28222–28227.
38. Kuettner, E. B., Hilgenfeld, R. & Weiss, M. S. (2002). Purification, characterization, and crystallization of alliinase from garlic. *Arch. Biochem. Biophys.* **402**, 192–200.
39. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S. *et al.* (2004). The Pfam protein families database. *Nucl. Acids Res.* **32**, D138–D141.
40. Millward-Sadler, S. J., Davidson, K., Hazlewood, G. P., Black, G. W., Gilbert, H. J. & Clarke, J. H. (1995). Novel cellulose-binding domains, NodB homologues and conserved modular architecture in xylanases from the aerobic soil bacteria *Pseudomonas fluorescens* subsp. *cellulosa* and *Cellvibrio mixtus*. *Biochem. J.* **312**, 39–48.
41. Fanutti, C., Ponyi, T., Black, G. W., Hazlewood, G. P. & Gilbert, H. J. (1995). The conserved noncatalytic 40-residue sequence in cellulases and hemicellulases from anaerobic fungi functions as a protein docking domain. *J. Biol. Chem.* **270**, 29314–29322.
42. Ponyi, T., Szabo, L., Nagy, T., Orosz, L., Simpson, P. J., Williamson, M. P. & Gilbert, H. J. (2000). Trp22, Trp24, and Tyr8 play a pivotal role in the binding of the family 10 cellulose-binding module from *Pseudomonas* xylanase A to insoluble ligands. *Biochemistry*, **39**, 985–991.
43. Raghothama, S., Eberhardt, R. Y., Simpson, P., Wigelsworth, D., White, P., Hazlewood, G. P. *et al.* (2001). Characterization of a cellulosome dockerin domain from the anaerobic fungus *Piromyces equi*. *Nature Struct. Biol.* **8**, 775–778.
44. Miles, L. A., Dy, C. Y., Nielsen, J., Barnham, K. J., Hinds, M. G., Olivera, B. M. *et al.* (2002). Structure of a novel P-superfamily spasmodic conotoxin reveals an inhibitory cystine knot motif. *J. Biol. Chem.* **277**, 43033–43040.
45. Cho, H. S. & Leahy, D. J. (2002). Structure of the extracellular region of HER3 reveals an interdomain tether. *Science*, **297**, 1330–1333.
46. Kalus, W., Zweckstetter, M., Renner, C., Sanchez, Y., Georgescu, J., Grol, M. *et al.* (1998). Structure of the IGF-binding domain of the insulin-like growth factor-binding protein-5 (IGFBP-5): implications for IGF and IGF-I receptor interactions. *EMBO J.* **17**, 6558–6572.
47. Eigenbrot, C., Randal, M. & Kossiakoff, A. A. (1990). Structural effects induced by removal of a disulfide-bridge: the X-ray structure of the C30A/C51A mutant of basic pancreatic trypsin inhibitor at 1.6 Å. *Protein Eng.* **3**, 591–598.
48. Perona, J. J., Tsu, C. A., Craik, C. S. & Fletterick, R. J. (1993). Crystal structures of rat anionic trypsin complexed with the protein inhibitors APPI and BPTI. *J. Mol. Biol.* **230**, 919–933.
49. van Mierlo, C. P., Darby, N. J., Neuhaus, D. & Creighton, T. E. (1991). (14–38, 30–51) Double-disulphide intermediate in folding of bovine pancreatic trypsin inhibitor: a two-dimensional ^{1}H nuclear magnetic resonance study. *J. Mol. Biol.* **222**, 353–371.
50. Song, J., Gilquin, B., Jamin, N., Drakopoulou, E., Guenneugues, M., Dauplais, M. *et al.* (1997). NMR solution structure of a two-disulfide derivative of charybdotoxin: structural evidence for conservation of scorpion toxin α/β motif and its hydrophobic side chain packing. *Biochemistry*, **36**, 3760–3766.
51. Daly, N. L., Clark, R. J. & Craik, D. J. (2003). Disulfide folding pathways of cystine knot proteins. *J. Biol. Chem.* **278**, 6314–6322.
52. Muller, Y. A., Heiring, C., Misselwitz, R., Welfle, K. & Welfle, H. (2002). The cystine knot promotes folding and not thermodynamic stability in vascular endothelial growth factor. *J. Biol. Chem.* **277**, 43410–43416.
53. Barnham, K. J., Torres, A. M., Alewood, D., Alewood, P. F., Domagala, T., Nice, E. C. & Norton, R. S. (1998). Role of the 6–20 disulfide bridge in the structure and activity of epidermal growth factor. *Protein Sci.* **7**, 1738–1749.
54. Hewage, C. M., Jiang, L., Parkinson, J. A., Ramage, R. & Sadler, I. H. (1999). Solution structure of a novel ET_B receptor selective agonist ET_{1-21} [Cys(Acm)^{1,15}, Aib^{3,11}, Leu^7] by nuclear magnetic resonance spectroscopy and molecular modelling. *J. Pept. Res.* **53**, 223–233.
55. Mok, K. H. & Han, K. H. (1999). NMR solution conformation of an antitoxic analogue of α-conotoxin GI: identification of a common nicotinic acetylcholine receptor α1-subunit binding surface for small ligands and α-conotoxins. *Biochemistry*, **38**, 11895–11904.
56. Pennington, M. W., Lanigan, M. D., Kalman, K., Mahnir, V. M., Rauer, H., McVaugh, C. T. *et al.* (1999). Role of disulfide bonds in the structure and potassium channel blocking activity of ShK toxin. *Biochemistry*, **38**, 14549–14558.
57. Bayrhuber, M., Vijayan, V., Ferber, M., Graf, R., Korukottu, J., Imperial, J. *et al.* (2005). Conkunitzin-S1 is the first member of a new Kunitz-type neurotoxin family. *J. Biol. Chem.* **280**, 23766–23770.
58. Garman, S. C., Simcoke, W. N., Stowers, A. W. & Garboczi, D. N. (2003). Structure of the C-terminal domains of merozoite surface protein-1 from *Plasmodium knowlesi* reveals a novel histidine binding site. *J. Biol. Chem.* **278**, 7264–7269.
59. Reily, M. D., Holub, K. E., Gray, W. R., Norris, T. M. & Adams, M. E. (1994). Structure–activity relationships for P-type calcium channel-selective ω-agatoxins. *Nature Struct. Biol.* **1**, 853–856.

60. Omecinsky, D. O., Holub, K. E., Adams, M. E. & Reily, M. D. (1996). Three-dimensional structure analysis of μ-agatoxins: further evidence for common motifs among neurotoxins with diverse ion channel specificities. *Biochemistry*, **35**, 2836–2844.
61. Xiang, Y., Huang, R. H., Liu, X. Z., Zhang, Y. & Wang, D. C. (2004). Crystal structure of a novel antifungal protein distinct with five disulfide bridges from *Eucommia ulmoides* Oliver at an atomic resolution. *J. Struct. Biol.* **148**, 86–97.
62. Tan, K., Duquette, M., Liu, J. H., Dong, Y., Zhang, R., Joachimiak, A. *et al*. (2002). Crystal structure of the TSP-1 type 1 repeats: a novel layered fold and its biological implication. *J. Cell Biol.* **159**, 373–382.
63. Gould, R. J., Polokoff, M. A., Friedman, P. A., Huang, T. F., Holt, J. C., Cook, J. J. & Niewiarowski, S. (1990). Disintegrins: a family of integrin inhibitory proteins from viper venoms. *Proc. Soc. Expt. Biol. Med.* **195**, 168–171.
64. Calvete, J. J., Marcinkiewicz, C., Monleon, D., Esteve, V., Celda, B., Juarez, P. & Sanz, L. (2005). Snake venom disintegrins: evolution of structure and function. *Toxicon*, **45**, 1063–1074.
65. Bilgrami, S., Tomar, S., Yadav, S., Kaur, P., Kumar, J., Jabeen, T. *et al*. (2004). Crystal structure of schistatin, a disintegrin homodimer from saw-scaled viper (*Echis carinatus*) at 2.5 Å resolution. *J. Mol. Biol.* **341**, 829–837.
66. Bilgrami, S., Yadav, S., Kaur, P., Sharma, S., Perbandt, M., Betzel, C. & Singh, T. P. (2005). Crystal structure of the disintegrin heterodimer from saw-scaled viper (*Echis carinatus*) at 1.9 Å resolution. *Biochemistry*, **44**, 11058–11066.
67. Adler, M., Lazarus, R. A., Dennis, M. S. & Wagner, G. (1991). Solution structure of kistrin, a potent platelet aggregation inhibitor and GP IIb-IIIa antagonist. *Science*, **253**, 445–448.
68. Paz Moreno-Murciano, M., Monleon, D., Marcinkiewicz, C., Calvete, J. J. & Celda, B. (2003). NMR solution structure of the non-RGD disintegrin obtustatin. *J. Mol. Biol.* **329**, 135–145.
69. Shin, J., Hong, S. Y., Chung, K., Kang, I., Jang, Y., Kim, D. S. & Lee, W. (2003). Solution structure of a novel disintegrin, salmosin, from *Agkistrondon halys* venom. *Biochemistry*, **42**, 14408–14415.
70. Zhou, A., Huntington, J. A., Pannu, N. S., Carrell, R. W. & Read, R. J. (2003). How vitronectin binds PAI-1 to modulate fibrinolysis and cell migration. *Nature Struct. Biol.* **10**, 541–544.
71. Kamikubo, Y., De Guzman, R., Kroon, G., Curriden, S., Neels, J. G., Churchill, M. J. *et al*. (2004). Disulfide bonding arrangements in active forms of the somatomedin B domain of human vitronectin. *Biochemistry*, **43**, 6519–6534.
72. Mayasundari, A., Whittemore, N. A., Serpersu, E. H. & Peterson, C. B. (2004). The solution structure of the N-terminal domain of human vitronectin: proximal sites that regulate fibrinolysis and cell migration. *J. Biol. Chem.* **279**, 29359–29366.
73. Horn, N. A., Hurst, G. B., Mayasundari, A., Whittemore, N. A., Serpersu, E. H. & Peterson, C. B. (2004). Assignment of the four disulfides in the N-terminal somatomedin B domain of native vitronectin isolated from human plasma. *J. Biol. Chem.* **279**, 35867–35878.
74. Benham, C. J. & Jafri, M. S. (1993). Disulfide bonding patterns and protein topologies. *Protein Sci.* **2**, 41–54.
75. Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S. *et al*. (2004). UniProt: the universal protein knowledgebase. *Nucl. Acids Res.* **32**, D115–D119.
76. Harrison, P. M. & Sternberg, M. J. (1994). Analysis and classification of disulphide connectivity in proteins. The entropic effect of cross-linkage. *J. Mol. Biol.* **244**, 448–463.
77. Hartig, G. R., Tran, T. T. & Smythe, M. L. (2005). Intramolecular disulphide bond arrangements in nonhomologous proteins. *Protein Sci.* **14**, 474–482.
78. Kozlov, G., Perreault, A., Schrag, J. D., Park, M., Cygler, M., Gehring, K. & Ekiel, I. (2004). Insights into function of PSI domains from structure of the Met receptor PSI domain. *Biochem. Biophys. Res. Commun.* **321**, 234–240.
79. Kim, C. H. & King, T. E. (1983). A mitochondrial protein essential for the formation of the cytochrome c1-c complex. Isolation, purification, and properties. *J. Biol. Chem.* **258**, 13543–13551.
80. Laskowski, M. & Qasim, M. A. (2000). What can the structures of enzyme–inhibitor complexes tell us about the structures of enzyme substrate complexes? *Biochim. Biophys. Acta*, **1477**, 324–337.
81. van de Locht, A., Stubbs, M. T., Bode, W., Friedrich, T., Bollschweiler, C., Hoffken, W. & Huber, R. (1996). The ornithodorin-thrombin crystal structure, a key to the TAP enigma? *EMBO J.* **15**, 6011–6017.
82. St Charles, R., Padmanabhan, K., Arni, R. V., Padmanabhan, K. P. & Tulinsky, A. (2000). Structure of tick anticoagulant peptide at 1.6 Å resolution complexed with bovine pancreatic trypsin inhibitor. *Protein Sci.* **9**, 265–272.
83. Gasparini, S., Danse, J. M., Lecoq, A., Pinkasfeld, S., Zinn-Justin, S., Young, L. C. *et al*. (1998). Delineation of the functional site of α-dendrotoxin. *J. Biol. Chem.* **273**, 25393–25403.
84. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al*. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
85. McManus, A. M., Nielsen, K. J., Marcus, J. P., Harrison, S. J., Green, J. L., Manners, J. M. & Craik, D. J. (1999). MiAMP1, a novel protein from *Macadamia integrifolia* adopts a Greek key β-barrel fold unique amongst plant antimicrobial proteins. *J. Mol. Biol.* **293**, 629–638.
86. Campos-Olivas, R., Bruix, M., Santoro, J., Lacadena, J., Martinez del Pozo, A., Gavilanez, J. G. & Rico, M. (1995). NMR solution structure of the antifungal protein from *Aspergillus giganteus*: evidence for cysteine pairing isomerism. *Biochemistry*, **34**, 3009–3021.
87. Raghothama, S., Simpson, P. J., Szabo, L., Nagy, T., Gilbert, H. J. & Williamson, M. P. (2000). Solution structure of the CBM10 cellulose binding module from *Pseudomonas* xylanase A. *Biochemistry*, **39**, 978–984.
88. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.

Edited by Michael J. E. Sternberg

(Received 3 November 2005; received in revised form 5 March 2006; accepted 9 March 2006)
Available online 29 March 2006