doi:10.1016/j.jmb.2003.10.077

J. Mol. Biol. (2004) **335**, 1083–1092

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

A Novel Database of Disulfide Patterns and its Application to the Discovery of Distantly Related Homologs

Herman W. T. van Vlijmen[1], Abhas Gupta[1], Lakshmi S. Narasimhan[2] and Juswinder Singh[1]*

[1] *Structural Informatics Group Biogen Inc., 14 Cambridge Center, Cambridge, MA 02142 USA*

[2] *Discovery Technologies, Pfizer Global Research and Development, Ann Arbor Laboratories, Ann Arbor, MI 48105, USA*

**Corresponding author*

Disulfide bonds are conserved strongly among proteins of related structure and function. Despite the explosive growth of protein sequence databases and the vast numbers of sequence search tools, no tool exists to draw relations between the disulfide patterns of homologous proteins. We present a comprehensive database of disulfide bonding patterns and a search method to find proteins with similar disulfide patterns. The disulfide database was constructed using disulfide annotations extracted from SwissProt, and was expanded significantly from 16,736 to 94,499 disulfide-containing domains by an inference method that combines SwissProt annotations with Pfam multiple alignments. To search the database, we define a disulfide description, called the disulfide signature, which encodes both spacings between cysteine residues and cysteine connectivity. A web tool was developed that allows users to search for related disulfide patterns and for subpatterns resulting from the removal of one or more disulfides from the pattern. We explore the possibility of using disulfide pattern conservation to identify protein homologs that are undetectable by PSI-BLAST. Examples include the homology between a sea anemone antihypertensive/antiviral protein and a sea anemone neurotoxin, and the homology between tick anticoagulant peptide and bovine trypsin inhibitor. In both examples, there is a clear structural similarity and a functional relationship. We used the database to find structural homologs for the Cripto CFC domain. The identification of a von Willebrand Factor C (VWFC)-like domain agrees with its functional role and explains mutation data. We believe that the rapid increase in structure determinations arising from structural genomics efforts and advances in mass spectrometry techniques will greatly increase the number of disulfide annotations. This information will become a valuable resource for structural and functional annotations of proteins. The availability of a searchable disulfide pattern database will thus provide a powerful new addition to existing homolog discovery methods.

Keywords: disulfide; database; protein structure; homology; structural genomics

Introduction

Disulfide bridges are ubiquitous to prokaryotic and eukaryotic proteins alike. Formed by the covalent cross-linking of cysteine residues, these structural elements are found mostly in non-reducing environments,[1,2] and have been shown to provide significant stabilization to the tertiary folds of proteins.[3–7] The stabilizing effect of disulfides on a

Abbreviations used: rTAP, recombinant tick anticoagulant protein; BPTI, bovine pancreatic trypsin inhibitor; EGF, epidermal growth factor; VWFC, von Willebrand factor C.

E-mail address of the corresponding author: juswinder.singh@biogenidec.com

protein's folded state has been described as predominantly entropic in nature,[8,9] although other explanations have been postulated.[10,11] In addition, disulfide bridges play a vital role in the folding process of many proteins.[12–14] Disulfide bridges have functional roles in proteins. A review by Yano *et al*.[15] is focused on the enzyme thioredoxin, which acts as a regulatory switch of target proteins by reducing their disulfide bonds. In two bacterial proteins, the transcription factor OxyR and the chaperone Hsp33, oxidation of cysteine residues to disulfides results in activation.[16]

Several analyses of disulfide bonds in proteins have been described in the past 20 years. Thornton[1] examined the connectivity and conformational properties of disulfides in proteins of known structure. Thornton also analyzed cysteine residues across homologous proteins and found them to be remarkably conserved. Moreover, it was noted that the loss of a disulfide was usually associated with mutation of not one but both cysteine residues. Benham & Jafri[17] demonstrated that the occurrences of disulfide connectivities are nonrandom and suggested that disulfide bond formation is a directed process. Probing the disulfide bond formation process, Harrison & Sternberg[18] analyzed disulfide connectivities in the context of sequence length and calculated relative entropic costs for disulfide bond formation. They proposed that an entropic stabilization model determines the disulfide connectivity for short proteins, whereas a diffusion model better describes the disulfide connectivities for longer sequences. Recognizing the strong conservation of disulfide frameworks, Narasimhan and collaborators[19] discussed the evolutionary relationship between snail and spider toxins, which was not obvious from the sequence similarity. Mas *et al*.[20] expanded this concept and grouped disulfide-containing protein structures based on the three-dimensional superposition of their disulfide bonds. Although applied to a small set of structures, their results reiterated the strong conservation of disulfide bonds even in the absence of significant sequence homology.

There has been an explosive growth of protein sequence databases[21] and a large number of sequence search tools†, but as yet no tool exists to draw relations between the disulfide bonding patterns of homologous proteins. Here, we present a novel approach for finding proteins with similar disulfide patterns. In addition, we present an extensive database of experimentally determined and inferred disulfide patterns. We highlight several cases of protein structural homologies where PSI-BLAST searches failed and functional or structural insights into protein sequences could be made through disulfide pattern searches.

The number of experimentally determined disulfide patterns is likely to increase significantly through structural genomics efforts[22] and recent advances in mass spectrometry techniques for disulfide determination.[23] The disulfide pattern database described here will become an increasingly powerful tool in the discovery of protein structural homologs.

Results

Database statistics

The SwissProt database release 41 (Feb 2003) contains a total of 41,846 annotated disulfide bonds, of which 5045 are determined experimentally and 34,968 are inferred by sequence similarity. A total of 1694 disulfides were annotated as interchain, which connect separate protein chains, and were thus not included in our database. For 139 disulfides, the annotations are ambiguous or erroneous, e.g. the disulfide residue numbers do not correspond to cysteine residues. The number of proteins with annotated disulfides is 10,568, which constitutes 8.6% of the total number of proteins in SwissProt. Of the 10,568 proteins, 1689 are annotated with experimentally determined disulfides, 8739 with inferred disulfides, and 140 with a combination of experimental and inferred disulfides. The structures of many of the proteins with annotated disulfides in SwissProt have been determined with X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR). A detailed analysis of the diversity of disulfide patterns and their relation to protein structure is being investigated (A.G. *et al*., unpublished results).

As described in Materials and Methods, we expanded the list of inferred disulfides by combining SwissProt annotations with Pfam multiple alignments. Since the Pfam multiple alignments (Revision: Feb 2003) contain SwissProt protein identifiers, the mapping of disulfide-containing proteins to Pfam domains was relatively straightforward. The 10,568 disulfide-containing proteins from SwissProt mapped to 13,408 domains in the Pfam-A database, corresponding to 345 different Pfam protein families, and 814 in the Pfam-B database, corresponding to 288 families. Many proteins contain multiple Pfam domains, which explains why the number of Pfam domains is larger than the number of SwissProt entries. We found 2514 disulfide-containing SwissProt annotated sequences whose disulfide-containing portion was absent from Pfam. Combining the Pfam-A, Pfam-B, and unassigned domains resulted in a total of 16,736 domains, which can be regarded as the publicly annotated number of disulfide-containing protein domains.

Applying the inferring algorithms outlined in Materials and Methods increased the disulfide database with 77,763 additional Pfam protein domains, expanding the database from 16,736 to 94,499 disulfide-containing protein domains. Figure 1 shows the distribution of the database contents by number of disulfides. In all, 2934

† See, for example, http://www.ebi.ac.uk/biocat/

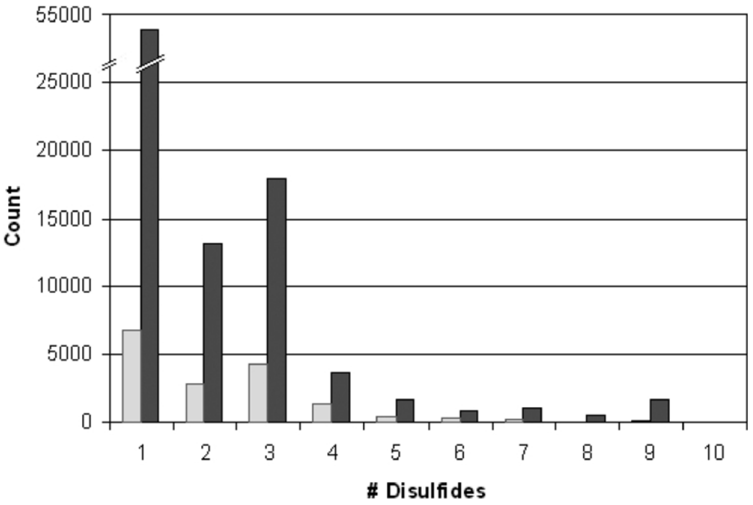

Figure 1. The count of disulfide-containing domains in the disulfide database, sorted by the number of disulfides per domain. Light, SwissProt annotations; dark, contents of our expanded database.

patterns generated in the inferring process correspond to SwissProt sequences that are either partially or completely lacking in their disulfide annotation. We were therefore able to add a significant number of annotations to the SwissProt database. The remaining patterns correspond to TrEMBL sequences that have very limited structural annotation. Figure 2 shows a view of the process leading to the final disulfide database.

Our inference method also revealed 65 domain families in which the disulfides could not be assigned unambiguously. This situation occurs, for instance, when a given cysteine residue is involved in multiple disulfide bonds across different proteins in a Pfam domain family. Preliminary analysis showed that in several cases the disulfide annotation in SwissProt was incorrect, but in other cases this ambiguity may be caused by a true plasticity of disulfides within the Pfam profile. A detailed analysis of these special cases is ongoing.

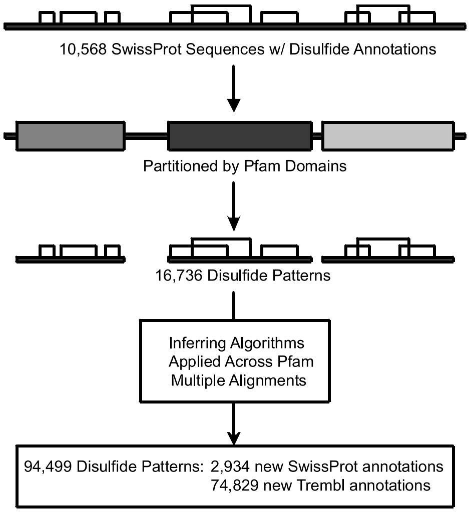
10,568 SwissProt Sequences w/ Disulfide Annotations
Partitioned by Pfam Domains
16,736 Disulfide Patterns
Inferring Algorithms Applied Across Pfam Multiple Alignments
94,499 Disulfide Patterns: 2,934 new SwissProt annotations
74,829 new Trembl annotations

Figure 2. Diagram of the disulfide database construction process.

Search parameters

The ability of the disulfide distance d_{mn} to distinguish between related and unrelated proteins is illustrated in Figure 3(a) for proteins containing three disulfides. The black and gray bars correspond to related and unrelated protein pairs, respectively. Proteins were defined as related if they belong to the same Pfam domain family. To determine the statistical relevance of a given distance d_{mn} between two disulfide patterns, we calculated false-positive score distributions using randomized disulfide patterns, as described in Materials and Methods. Cumulative distributions for the comparisons between random and related disulfide patterns correspond to the false-positive and false-negative values as a function of disulfide distance d_{mn}, respectively (Figure 3(b)). In order to define the best distance cutoff for a disulfide database search, the sum of false-positive and false-negative probabilities should ideally be at a minimum. In our distributions, we often did not observe a well-defined minimum of this sum, so we could not use this approach to define the optimal distance cutoff. Instead, we used the cumulative false-positive distributions to assign P-values to disulfide distances obtained in a database search. Figure 4 shows the dependence of the distance cutoff for a P-value of 0.01 on the pattern length L for proteins with three to five disulfides. As the pattern length increases, the cutoff value d_{mn} at $P = 0.01$ increases linearly. We report the P-value corresponding to the pattern length L of the search query in our search results.

To illustrate the potential of the disulfide database and search tool, we highlight three cases in which structural and functional homologies between proteins were uncovered, none of which was apparent from conventional sequence-based methods. In the first two cases, the homologies have been noted only after structural elucidation of the proteins. In the third case, we propose a structure and function for a protein domain for

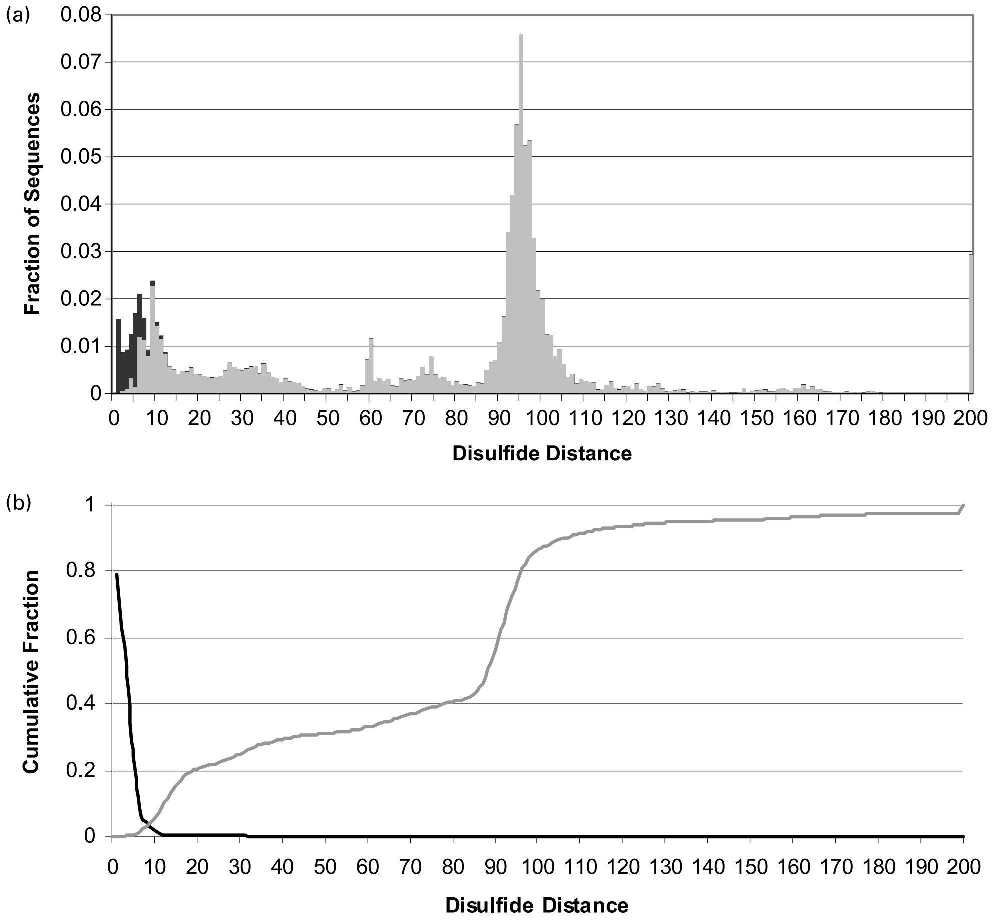

Figure 3. (a) Histogram of related (black) and unrelated (gray) disulfide pairs as a function of disulfide distance d, for proteins containing three disulfides. For this plot, the query disulfide pattern had a vector length L between 0 and 39. (b) Cumulative false positive curve (gray) of disulfide pair distances d, using randomized disulfide patterns of the set of proteins used in (a). The false negative curve based on related disulfide pairs is shown in black.

which no structural information exists with the exception of the disulfide bonding information.

Case 1: ion channel blockers

ATX Ia is a 46 residue neurotoxin of the sea anemone *Anemonia sulcata* that exerts its toxicity by blocking sodium channels. Its structure was solved by NMR, which revealed a four-stranded β-sheet structure containing three disulfide bridges.[24] The structural elucidation showed that ATX Ia was structurally similar to the 43 residue antihypertensive and antiviral protein BDS-I from the same species.[25] BDS-I operates by blocking potassium channels. Widmer *et al*.[24] noted that the homology between the two proteins was not obvious from a comparison of the amino acid sequences. This absence of observable sequence

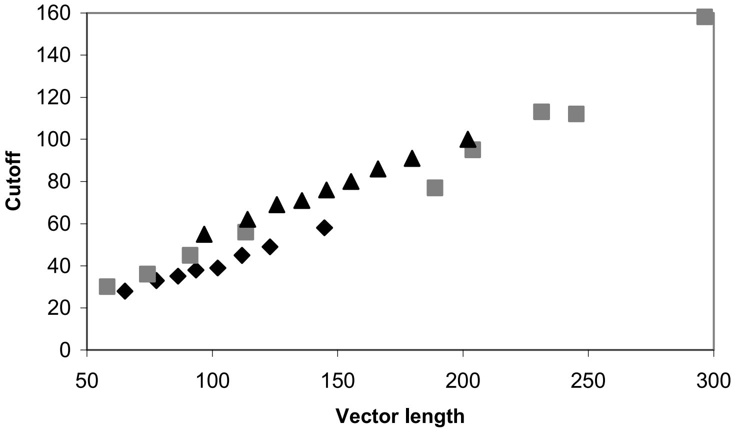

Figure 4. Disulfide pattern distance cutoff required for a false positive rate of <0.01 as a function of the query sequence vector length L. The data corresponding to patterns with three, four, and five disulfides are represented by diamonds, squares, and triangles, respectively.

homology is still present today, despite significant advances in sequence homology search methods and protein sequence databases. A PSI-BLAST search (five iterations, *E*-value cutoff 0.01) of the ATX-Ia protein sequence in both the SwissProt/TrEMBL and the non-redundant NCBI NR databases did not find the BDS-I protein, and *vice versa*. In contrast, a disulfide-based search in our database readily finds the BDS-I protein when the ATX-Ia disulfide pattern is used as the query (Table 1). The Structural Classification of Proteins (SCOP) database[26] classifies these proteins in the same structural family. The structural similarity between these proteins is illustrated in Figure 5. In this case, structural homology does translate directly to functional homology, since both proteins bind to and inhibit ion channels of a similar structure.

Case 2: proteinase inhibitor

The solution structure of the 60 residue recombinant tick anticoagulant protein (rTAP) was solved by NMR and shown to be structurally similar to Kunitz-type proteinase inhibitors such as bovine pancreatic trypsin inhibitor (BPTI).[27] Both structures contain a two-stranded β-sheet and a C-terminal α-helix, stabilized by three disulfide bonds (Figure 6). There is also a strong functional similarity between TAP and BPTI: both proteins are inhibitors of proteinases (factor Xa and trypsin, respectively). The absence of significant sequence homology between TAP and BPTI was noted by Antuch *et al.*[27] and PSI-BLAST searches in the current versions of SwissProt/TrEMBL and NR were unsuccessful in identifying the similarity between these two proteins. The disulfide-based search identified the structural relationship between these proteins readily, as shown in Table 2. The SCOP database classified these proteins in the same category at the superfamily level.

Cripto CFC domain

In a recent study of the CFC domain of human Cripto, we employed the disulfide database search tool to obtain structural information on the protein.[28] Cripto is a protein involved in early embryonic development and was shown to be overexpressed in a number of human cancers.[29] Cripto family proteins are characterized by two cysteine-rich structural motifs: an epidermal growth factor (EGF)-like domain and a CFC domain, the latter of which is considered unique to this family. We determined the disulfide pattern of the CFC domain, which contains three disulfides, and searched for related proteins with known structure in our disulfide database. The search revealed two small, structurally related serine protease inhibitors, PMP-D2 and PMP-C. Both proteins are classified as VWFC (von Willebrand factor C)-like domains. BLAST searches with the CFC domain sequence on SwissProt/TrEMBL and NCBI NR databases do find the VWFC domains, albeit with very low confidence (*E*-values >1).

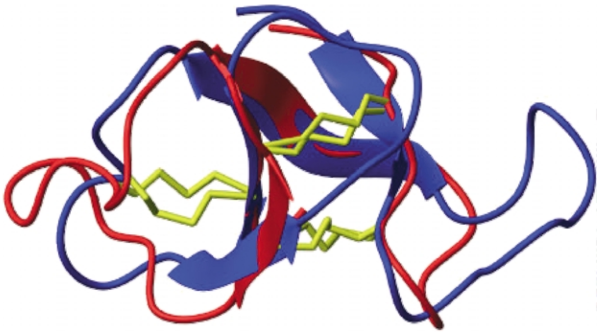

Figure 5. Structural superposition of ATX-Ia (blue, PDB code 1atx) and BDS-I (red, PDB code 1bds). Disulfide bonds are shown in yellow.

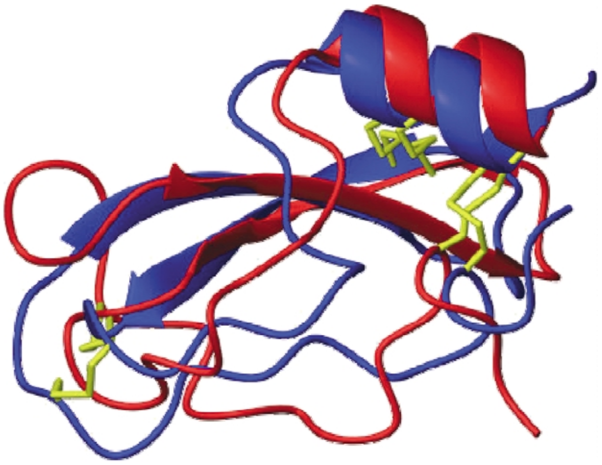

Figure 6. Structural superposition of BPTI (blue, PDB code 5pti) and TAP (red, PDB code 1tap).

The annotation of the CFC domain as a VWFC domain resulted in the identification of a number of proteins that have the same modular structure of an EGF-like domain followed by a VWFC domain, including NELL1, NELL2, JAGGED1, and JAGGED2. This inferred structural relationship suggested functional similarities among the proteins. A comparison between Cripto and JAGGED2 showed that they have distinct similarities at the sequence level (undetectable by sequence search algorithms), that they are both involved in signal transduction, and that both play roles in patterning and morphogenesis in early embryonic development.

The NMR structure of PMP-C (Protein Data Bank (PDB)[30] code 1pmc) was used to build a three-dimensional model of the Cripto CFC domain. The model was consistent with data from functional studies on mutants of the CFC domain, since two very important residues for interaction of the CFC domain with the Alk4 receptor, H120 and W123, were both located in the same area on the solvent-accessible surface of the structural model.

Table 1. Disulfide database search results for ATX-Ia (SwissProt code TXA1_ANESU)

Score	$P(x)$	Class	Chain	Search pattern	Cysseq	Expseq	Top	Bounds	Length	Structure
0.00	0	PF00706	TXA1_ANESU	39-2-28-21-17	2-21-7-9-1	4-43-6-34-27-44	1-5_2-4_3-6	3-44	3	1atx
1.41	0	PF00706	TXA2_RADMA	40-2-28-21-18	2-21-7-10-1	3-43-5-33-26-44	1-5_2-4_3-6	2-44	3	[1atx,]
4.24	0	PF00706X	CLX1_CALPA	39-2-28-18-20	2-18-10-9-1	36-75-38-66-56-76	1-5_2-4_3-6	35-76	3	[1atx,...]
4.24	0	PF00706X	CLX2_CALPA	39-2-28-18-20	2-18-10-9-1	36-75-38-66-56-76	1-5_2-4_3-6	35-76	3	[1atx,...]
4.24	0	PF00706	TXAB_ANTXA	42-2-30-23-18	2-23-7-10-1	4-46-6-36-29-47	1-5_2-4_3-6	3-47	3	[1atx,...]
4.24	0	PF00706	TXAA_ANTXA	42-2-30-23-18	2-23-7-10-1	4-46-6-36-29-47	1-5_2-4_3-6	3-47	3	[1atx,...]
6.78	0.000413	NULL	BDS1_ANESU	35-2-26-16-18	2-16-10-7-1	4-39-6-32-22-40	1-5_2-4_3-6	0-0	3	1bds
10.95	0.001032	PF00321	THN_PYRPU	38-1-27-12-11	1-12-11-4-10	3-41-4-31-16-27	1-6_2-5_3-4	1-47	3	[1cnb,...]
11.62	0.001652	PF00321X	Q9S980	37-1-28-12-10	1-12-10-6-8	27-64-28-56-40-50	1-6_2-5_3-4	25-70	3	[1cnb,...]
13.78	0.002375	PF01549X	Q9M0K1	40-7-26-9-21	7-9-17-4-3	275-315-282-308-291-312	1-6_2-4_3-5	274-315	3	[1roo,...]

The columns in the Table indicate the disulfide distance *d*, the false-positive score (*P*-value), the Pfam domain, the SwissProt protein code, the disulfide signature, the cysteine spacing pattern, the residue numbers of the disulfides, the disulfide topology, the sequence bounds of the Pfam domain, the number of disulfides, and the available structural information. If there is a PDB structure of the hit itself, a PDB code is listed. If there exist PDB structures of any of the Pfam members of the hit, a PDB code is shown in parentheses. A number of hits from the PF00706 family were removed to highlight the hits of interest. The BDS-I protein has the SwissProt code BDS1_ANESU.

Table 2. Disulfide database search results for TAP (SwissProt code TAP_ORNMO)

Score	$P(x)$	Class	Chain	Search pattern	Cysseq	Expseq	Top	Bounds	Length	Structure
0.00	0	NULL	TAP_ORNMO	54-10-24-18-22	10-18-6-16-4	5-59-15-39-33-55	1-6_2-4_3-5	0-0	3	1tap
2.45	0	PF00014	TFP2_HUMAN	53-10-24-16-23	10-16-8-15-4	96-149-106-130-122-145	1-6_2-4_3-5	96-149	3	[5pti,...]
3.74	0	PF00014	ISC2_BOMMO	51-10-24-16-21	10-16-8-13-4	9-60-19-43-35-56	1-6_2-4_3-5	9-60	3	[5pti,...]
4.69	0	PF00014	TFPI_RAT	50-9-24-16-21	9-16-8-13-4	124-174-133-157-149-170	1-6_2-4_3-5	124-174	3	[5pti,...]
4.69	0	PF00014	SPT2_HUMAN	50-9-24-16-21	9-16-8-13-4	133-183-142-166-158-179	1-6_2-4_3-5	133-183	3	[5pti,...]
4.69	0	PF00014	A4_HUMAN	50-9-24-16-21	9-16-8-13-4	291-341-300-324-316-337	1-6_2-4_3-5	291-341	3	1aap
4.69	0	PF00014	IVB3_VIPAA	50-9-24-16-21	9-16-8-13-4	7-57-16-40-32-53	1-6_2-4_3-5	7-57	3	[5pti,...]
4.69	0	PF00014	BPT2_BOVIN	50-9-24-16-21	9-16-8-13-4	40-90-49-73-65-86	1-6_2-4_3-5	40-90	3	[5pti,...]
4.69	0	PF00014	BPT1_BOVIN	50-9-24-16-21	9-16-8-13-4	40-90-49-73-65-86	1-6_2-4_3-5	40-90	3	5pti
4.69	0	PF00014	CA36_HUMAN	50-9-24-16-21	9-16-8-13-4	3111-3161-3120-3144-3136-3157	1-6_2-4_3-5	3111-3161	3	1knt

The columns are as described for Table 1. The BPTI protein has the SwissProt code BPT1_BOVIN.

Discussion

The disulfide database and search tool described here provide a unique means to effectively find structurally related proteins through similarity in disulfide patterns. The disulfide database was built initially from the annotations in SwissProt and expanded subsequently almost tenfold by using the Pfam multiple alignments. Our database of 94,499 disulfide patterns provides a significant coverage of disulfide space. Because any inconsistencies in the database expansion were discarded, we are confident about the additional 77,763 inferred disulfide annotations. Several assumptions were made during the generation of the disulfide database. We attempted to apply a consistent framework to SwissProt disulfide annotations, despite numerous observed ambiguities. SwissProt states that disulfides annotated with "By Similarity" are inferred from homologous proteins exhibiting a strong degree of similarity. However, the definition for homology is not explicit. Moreover, we encountered examples of disulfides annotated with "Potential" and "Probable", for which no SwissProt definition exists. Disulfides without any annotations were assumed to be determined experimentally.

We propose a restructuring of the disulfide annotation standard in SwissProt. We believe that disulfide annotations should be accompanied with references to the source from which they are obtained. In the case of disulfides that are inferred by homology with another protein, that protein should be mentioned in the annotation. In addition, annotations should indicate functional roles of disulfides such as the regulatory switch observed in thioredoxin.[15]

The disulfide signature definition and simple Euclidean distance measure provide a fast and straightforward way to find disulfide patterns with a similar topology and cysteine spacing. The large increase in the number of disulfide patterns in our database greatly improves the probability of finding matches closely related to a search pattern.

The growth of experimental disulfide information has been relatively flat compared to the increase in known protein sequences. To a large extent, this is due to the difficulties of experimental disulfide determination, which is usually done by enzymatic protein digestion followed by mass spectrometry analysis,[23] by X-ray crystallography, or by NMR.

As shown by the examples of the ion channel blockers and proteinase inhibitors, similarity in disulfide patterns may indicate structural and/or functional homology. In both examples, thorough PSI-BLAST searches did not find the homology between the proteins described. The disulfide database contains many proteins with unknown functions, which opens up the possibility of discovering as yet unknown relationships between proteins based on their disulfide pattern similarities.

The Cripto CFC domain example highlights a function of the disulfide database that may have significant applications in structural genomics. One of the main goals of structural genomics is to obtain 3D structural information on as many proteins as experimentally possible, usually by X-ray crystallography or NMR. From a structural genomics point of view, the most interesting targets are proteins with no detectable homology to any protein with known structure. It has been shown recently that this may prove to be a difficult task, since many proteins do not express well, crystallize readily or are too large to be tractable with NMR.[22] In many cases, it may be possible to experimentally determine the disulfide pattern and find proteins with similar disulfide patterns in the disulfide database, even when the disulfide pattern information is incomplete. If any of the significant hits in the database has a known structure, other experimental techniques such as circular dichroism[31] or infrared spectroscopy[32] may be used to validate the structural homology.

In the absence of experimental disulfide information, a predicted disulfide pattern could be used to discover structural and functional relationships between proteins. An accurate disulfide prediction method would allow the database mining of cysteine-rich proteins with unknown function. Unfortunately, no accurate prediction methods exist to date.[33]

In conclusion, we have compiled a searchable database of known disulfide patterns and have expanded it significantly by inference. The database can be seen as a good reflection of the currently known disulfide pattern space. The database and search tool may play a significant role in the structural assignment of cysteine-rich proteins and will provide a useful tool in structural genomics efforts.

Materials and Methods

Search tool

Similarity in disulfide patterns reflects similarity in both disulfide topology and cysteine spacing. The disulfide topology denotes the connectivity of the cysteine residues involved in disulfide bonds. For example, a protein with two disulfides has three possible topologies: 1-2_3-4 (also written as *aabb*), 1-3_2-4 (*abab*), and 1-4_2-3 (*abba*). The numbers in the disulfide topologies correspond to the sequential numbering of the cysteine residues in the protein sequence and the dashes represent bonds between those residues. The number of possible disulfide patterns rapidly increases with the number of disulfides.[17]

The cysteine spacing pattern is defined as a string of residue spacings between adjacent disulfide-linked cysteine residues of a protein, starting with the first and continuing along to the last cysteine residue in the sequence (Figure 7). Thus, the cysteine spacing pattern is a set of $(2n - 1)$ numbers, where n is the number of disulfides in the protein. This representation does not encompass any cysteine connectivity information and

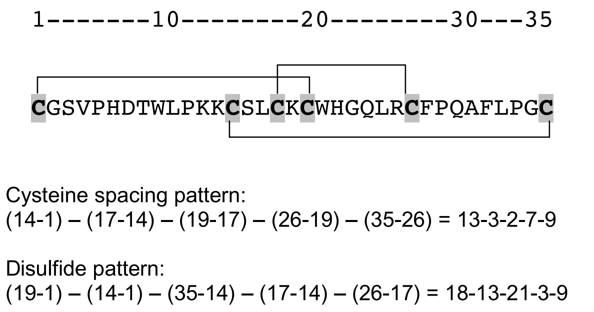

Figure 7. Example of the cysteine spacing pattern and disulfide signature of the Cripto CFC domain, a protein sequence segment with three disulfides. The disulfide topology is 1-4_2-6_3-5.

therefore captures only one aspect of the disulfide pattern. A search in a protein sequence database for a cysteine spacing pattern can be emulated using standard query methods such as FASTA or BLAST coupled with scoring matrices in which cysteine residues have strongly increased weights (e.g. see Karlin & Altshul).[34]

To incorporate disulfide topology into our search, we propose a disulfide pattern that implicitly contains disulfide topology. We call this pattern the disulfide signature. Given a protein with known disulfide topology and cysteine spacings, the disulfide signature is defined as follows: the first number is the length of the first disulfide bridge, the second number is the spacing between the first residue of the first disulfide and the first residue of the second disulfide, the third number is the length of the second disulfide bridge, and this is repeated to the last number of the pattern, which is the length of the last disulfide bridge (Figure 7). Therefore, the odd positions in the pattern correspond to disulfide bridge lengths and the even positions correspond to spacings between the first residues in neighboring disulfides. As with the cysteine spacing pattern, the disulfide signature contains $(2n-1)$ numbers, where n is the number of disulfides. The disulfide topology and the cysteine spacing pattern can be reconstructed easily from the disulfide signature.

To calculate the similarity between two disulfide signatures or two cysteine spacing patterns, we define the pairwise Euclidean distance d_{mn} between patterns of sequences m and n:

$$d_{mn} = \sqrt{\sum_i (m_i - n_i)^2} \tag{1}$$

where the index i sums over all numbers in the pattern. This definition requires that both cysteine patterns are made up of the same number of disulfide bonds. Shorter distances indicate higher degrees of similarity between the disulfide patterns.

Creation of disulfide database

The protein database with the largest number of annotated disulfides is the SwissProt database.[35] It contains both experimentally determined disulfides and inferred disulfides (annotated as By similarity). Inferred disulfide annotations are assigned only when a protein sequence has a clear sequence homology to another protein with experimentally determined disulfides. Although the number of disulfide annotations thus obtained is quite large, there exist many more proteins for which disulfide connectivity patterns can be inferred on the basis of overall sequence homology. To expand the set of inferred disulfides, we combined the annotations in SwissProt (version 41, Feb 2003) with the multiple sequence alignments in the Pfam database (version 8.0, Feb 2003).[36] The process is illustrated in Figure 8.

For every Pfam family multiple alignment, including the curated Pfam-A and the DOMAINER-generated[37] Pfam-B domain families, we identified sequences in the alignment with annotated disulfides in SwissProt. In many cases, more than one protein in a given Pfam domain family has disulfide annotations in SwissProt, sometimes at different sequence positions and/or with different connectivity patterns. The residue columns of the multiple alignments corresponding to the cysteine residues of the disulfides were determined, and a cumulative set of disulfide bonds was thus defined for the multiple alignment of the Pfam domain family. We then assigned disulfides to all proteins in the multiple

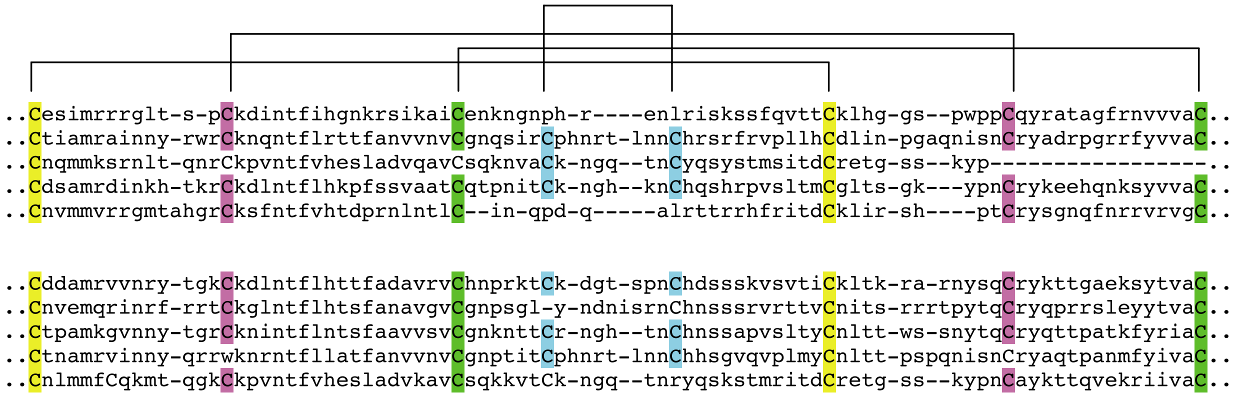
..Cesimrrrglt-s-pCkdintfihgnkrsikaiCenkngnph-r----enlrisksssfqvttCklhg-gs--pwppCqyratagfrnvvvaC..
..Ctiamrainny-rwrCknqntflrttfanvvnvCgnqsirCphnrt-lnnChrsrfrvpllhCdlin-pgaqnisnCryadrpgrrfyvvaC..
..Cnqmmksrnlt-qnrCkpvntfvhesladvqavCsqknvaCk-ngq--tnCyqsystmsitdCretg-ss--kyp-----------------..
..Cdsamrdinkh-tkrCkdlntflhkpfssvaatCqtpnitCk-ngh--knChqshrpvsltmCglts-gk---ypnCrykeehqnksyvvaC..
..CnvmmvrrgmtahgrCksfntfvhtdprnlntlC--in-qpd-q-----alrttrrhfritdCklir-sh----ptCrysgnqfnrrvrvgC..
..Cddamrvvnry-tgkCkdlntflhttfadavrvChnprktCk-dgt-spnChdssskvsvtiCkltk-ra-rnysqCrykttgaeksytvaC..
..Cnvemqrinrf-rrtCkglntflhtsfanavgvCgnpsgl-y-ndnisrnChnsssrvrttvCnits-rrrtpytqCryqprrsleyytvaC..
..Ctpamkgvnny-tgrCknintflntsfaavvsvCgnknttCr-ngh--tnChnssapvsltyCnltt-ws-snytqCryqttpatkfyriaC..
..Ctnamrvinny-qrrwknrntfllatfanvvnvCgnptitCphnrt-lnnChhsgvqvplmyCnltt-pspqnisnCryaqtpanmfyivaC..
..CnlmmfCqkmt-qgkCkpvntfvhesladvkavCsqkkvtCk-ngq--tnryqskstmritdCretg-ss--kypnCaykttqvekriivaC..

Figure 8. Illustration of the disulfide inference method used to expand the disulfide database. The top five sequences are from Pfam domain PF00074 (pancreatic ribonuclease) and have disulfide annotations, indicated by matching cysteine colors and above connecting lines. The bottom five sequences are from the same Pfam domain and have no disulfide annotations. Cysteine residues of the inferred disulfides are shown here in colors corresponding to the previously annotated disulfides. Note that in sequences 2 and 4, one of the disulfide-participating cysteine residues has mutated to a non-cysteine residue.

alignment that had cysteine residues at both positions for any of the cumulative set of disulfide bonds.

The cysteine spacing patterns and disulfide signatures of all disulfides defined in SwissProt and inferred by our approach were stored in a database. As there are many cases of proteins within the same family differing in the number of disulfides,[1] the search tool includes the option of searching against all subpatterns of every disulfide pattern in the database, and the option to search with all subpatterns of the query. A subpattern is defined as the cysteine spacing pattern or disulfide signature that results from the removal of one or more disulfide bonds from an original disulfide pattern. When a subpattern search is invoked, the complete set of subpatterns resulting from the removal of one and two disulfides is calculated at execution time, for each pattern in the database and/or for the query pattern.

Determination of search parameters

The search method and database defined thus far enable us to find the proteins with the most similar disulfide patterns to a given query pattern. To define the distance d at which the similarity is statistically significant, we generated distance distributions by calculating the distances between 100,000 pairs of random disulfide patterns. To construct the random patterns, we randomly picked the m_i and n_i values (see equation (1)) from the collection of all spacings in the set of proteins with the corresponding number of disulfides. Separate distributions were calculated for patterns of three disulfides up to ten disulfides. Furthermore, we subdivided the generated disulfide distributions into ten equally populated sets (10,000 distances each) based on the vector length L of the m_i values of the random pairs:

$$L = \sqrt{\sum_i m_i^2} \tag{2}$$

This subdivision was incorporated to account for the dependency of pattern distances on L. For example, the random distance distribution to a "short" pattern, say 5-2-6-4-8, is centered at a significantly lower value than the distribution of a "long" pattern such as 25-10-18-7-35. As the distance distributions are based on random disulfide patterns, they signify false-positive distance values. The integration of normalized distance distributions was used to assign the statistical significance values to different disulfide pattern similarity scores.

Acknowledgements

We thank Alex Lukashin for discussions on the statistics of disulfide pattern searching.

References

1. Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.* **151**, 261–287.
2. Fiser, A. & Simon, I. (2000). Predicting the oxidation state of cysteines by multiple sequence alignment. *Bioinformatics*, **16**, 251–256.
3. Creighton, T. E. (1993). *Proteins*, 2nd edit., W.H. Freeman and Company, New York.
4. Matsumura, M., Sognor, G., Thomson, J. A. & Barnett, B. J. (1989). Stabilization of phage T4 lysozyme by engineered disulfide bonds. *Proc. Natl Acad. Sci. USA*, **86**, 6562–6566.
5. Matsumura, M., Signor, G. & Matthews, B. W. (1989). Substantial increase of protein stability by multiple disulphide bonds. *Nature*, **342**, 291–293.
6. Hinck, A. P., Truckses, D. M. & Markley, J. L. (1996). Engineered disulfide bonds in staphylococcal nuclease: effects on the stability and conformation of the folded protein. *Biochemistry*, **35**, 10328–10338.
7. Pace, C. N., Grimsley, G. R., Thomson, J. A. & Barnett, B. J. (1988). Conformational stability and activity of ribonuclease T_1 with zero, one, and two intact disulfide bonds. *J. Biol. Chem.* **263**, 11820–11825.
8. Cooper, A., Eyles, S. J., Radford, S. E. & Dobson, C. M. (1992). Thermodynamic consequences of the removal of a disulphide bridge from hen lysozyme. *J. Mol. Biol.* **225**, 939–943.
9. Kauzmann, W. (1959). Relative probabilities of isomers in cystine-containing randomly coiled polypeptides. In *Sulfur in Proteins* (Benesch, R., Benesch, R. E., Boyer, P., Klotz, I., Middlebrook, W. R., Szent-Gyorgyi, A. & Schwartz, D. R., eds), pp. 93–108, Academic Press, New York.
10. 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.
11. Betz, S. F. (1993). Disulfide bonds and the stability of globular proteins. *Protein Sci.* **2**, 1551–1558.
12. Creighton, T. E. & Goldenberg, D. P. (1984). Kinetic role of a meta-stable native-like two-disulphide species in the folding transition of bovine pancreatic trypsin inhibitor. *J. Mol. Biol.* **179**, 497–526.
13. Gilbert, H. F. (1994). The formation of native disulfide bonds. In *Mechanisms of Protein Folding* (Pain, R. H., ed.), pp. 104–136, IRL Press, Oxford.
14. Wedemeyer, W. J., Welker, E., Narayan, M. & Scheraga, H. A. (2000). Disulfide bonds and protein folding. *Biochemistry*, **39**, 4207–4216.
15. Yano, H., Kuroda, S. & Buchanan, B. B. (2002). Disulfide proteome in the analysis of protein function and structure. *Proteomics*, **2**, 1090–1096.
16. Aslund, F. & Beckwith, J. (1999). Bridge over troubled waters: sensing stress by disulfide bond formation. *Cell*, **96**, 751–753.
17. Benham, C. J. & Jafri, M. S. (1993). Disulfide bonding patterns and protein topologies. *Protein Sci.* **2**, 41–54.
18. 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.
19. Narasimhan, L., Singh, J., Humblet, C., Guruprasad, K. & Blundell, T. (1994). Snail and spider toxins share a similar tertiary structure and "cystine motif". *Nature Struct. Biol.* **1**, 850–852.
20. 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.
21. Baxevanis, A. D. (2003). The molecular biology database collection: 2003 update. *Nucl. Acids Res.* **31**, 1–12.
22. Zhang, C. & Kim, S. H. (2003). Overview of structural geznomics: from structure to function. *Curr. Opin. Chem. Biol.* **7**, 28–32.
23. Gorman, J. J., Wallis, T. P. & Pitt, J. J. (2002). Protein disulfide bond determination by mass spectrometry. *Mass Spectrom. Rev.* **21**, 183–216.

24. Widmer, H., Billeter, M. & Wuthrich, K. (1989). Three-dimensional structure of the neurotoxin ATX Ia from *Anemonia sulcata* in aqueous solution determined by nuclear magnetic resonance spectroscopy. *Proteins: Struct. Funct. Genet.* **6**, 357–371.
25. Driscoll, P. C., Gronenborn, A. M., Beress, L. & Clore, G. M. (1989). Determination of the three-dimensional solution structure of the antihypertensive and antiviral protein BDS-I from the sea anemone *Anemonia sulcata*: a study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. *Biochemistry*, **28**, 2188–2198.
26. Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **28**, 257–259.
27. Antuch, W., Guntert, P., Billeter, M., Hawthorne, T., Grossenbacher, H. & Wuthrich, K. (1994). NMR solution structure of the recombinant tick anticoagulant protein (rTAP), a factor Xa inhibitor from the tick *Ornithodoros moubata*. *FEBS Letters*, **352**, 251–257.
28. Foley, S. F., van Vlijmen, H. W. T., Boynton, R. E., Adkins, H. B., Cheung, A. E., Singh, J. *et al.* (2003). The CRIPTO/FRL-1/CRYPTIC (CFC) domain of human Cripto. Functional and structural insights through disulfide structure analysis. *Eur. J. Biochem.* **270**, 3610–3618.
29. Saloman, D. S., Bianco, C., Ebert, A. D., Khan, N. I., De Santis, M., Normanno, N. *et al.* (2000). The EGF-CFC family: novel epidermal growth factor-related proteins in development and cancer. *Endocr. Relat. Cancer*, **7**, 199–226.
30. 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.
31. Kelly, S. M. & Price, N. C. (2000). The use of circular dichroism in the investigation of protein structure and function. *Curr. Protein Pept. Sci.* **1**, 349–384.
32. Jung, C. (2000). Insight into protein structure and protein-ligand recognition by Fourier transform infrared spectroscopy. *J. Mol. Recogn.* **13**, 325–351.
33. Fariselli, P. & Casadio, R. (2001). Prediction of disulfide connectivity in proteins. *Bioinformatics*, **17**, 957–964.
34. Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
35. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E. *et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**, 365–370.
36. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R. *et al.* (2002). The Pfam protein families database. *Nucl. Acids Res.* **30**, 276–280.
37. Corpet, F., Gouzy, J. & Kahn, D. (1998). The ProDom database of protein domain families. *Nucl. Acids Res.* **26**, 323–326.

Edited by F. E. Cohen

(Received 12 August 2003; received in revised form 7 October 2003; accepted 27 October 2003)

Note added in proof: While this manuscript was being reviewed, Chuang *et al*. ((2003). *Proteins: Struct. Funct. Genet.* **53**, 1–5) published a paper in which relationships between protein structures and disulfide-bonding patterns were explored.