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