BIOINFORMATIC STUDIES OF SMALL DISULPHIDE-RICH PROTEINS (SDPs)

KONG LESHENG

(M.Sc., Shanghai Jiao Tong University, China)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTORAL OF PHILOSOPHY

DEPARTMENT OF BIOCHEMISTRY

NATIONAL UNIVERSITY OF SINGAPORE

2006

Acknowledgements

My first thanks go to my supervisors, Prof. Shoba Ranganathan and Prof. Tan Tin Wee, for their inspiration, guidance and encouragement to support the accomplishment of the project, especially for their many enlightening discussions of my research career. Herein I would like to extend my special appreciation to Prof. Shoba Ranganathan who provided me with very good training opportunities in every aspect and extended her consideration during my time in her group.

My heartfelt thanks also go to the chairman of my Thesis Advisory Committee, Prof. R. Manjunatha Kini for his helpful advice during the committee meetings. I sincerely wish to thank Prof. Michael James for his special help and suggestions.

It has been my privilege to work with so many good friends in the Bioinformatics Centre: Bernett Lee, Eric Tan, Justin Choo, Li Kai, Paul Tan, Victor Tong and Vivek Gopalan: thanks you for all the help whenever I needed it. Particular thanks to Mr. Mark De Silva and Mr. Lim Kuan Siong for their technical support and helpful assistance. Also, my grateful thanks go to the Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore for the award of an NSTB (now A-STAR) research scholarship for pursuing my PhD degree.

Finally, I thank my wife Meng Chunying for her love and patience during my hard times. She always takes good care of me. I am also indebted to my parents Kong Enqing and Yang Lianju for their endless love and their encouragement to me over the years. I dedicate this dissertation to them.

Table of Content

Summary

Small disulphide-rich proteins (SDPs) represent a class of proteins which include predominantly secretory proteins that have predatory, defensive or regulatory roles (such as toxins, inhibitors and hormones). SDPs are thus a rich source for therapeutic drugs and other bioactive molecules. SDPs are characterized as short polypeptides stabilized in conformation by inter-cysteine side chain bonds known as disulphide bonds (or bridges). These disulphide bridges play crucial roles in the three dimensional structure, function and evolution of SDPs.

The roles and patterns of disulphide bridges in SDPs were investigated using bioinformatics approaches. SDPs structures and relevant data were systematically gathered from public databases to form the Small Disulphide-rich Fold Database - SDFD. Systematic analyses and mining of this database suggested that the cysteine signature in the peptide sequence could facilitate the detection of distantly related homologs or convergently evolved structures. Based on the rules derived from the analyses, a software pipeline called SDPMOD was designed and implemented specifically for the automated comparative modeling of SDPs. For further in-depth investigation of the nature of SDPs, an unusual subfamily of SDPs was selected. This potato type II proteinase inhibitor family (Pot II) was comprehensively characterized for conserved patterns in 3D structure, protein sequence and gene architecture. The analysis of the ratio of non-synonymous to synonymous substitutions suggested heterogeneous selection pressure at different regions within the Pot II domains. As opposed to “purifying selection” over the cysteine scaffold that is expected, some evidence for “positive selection” on the reactive site is presented, illustrating the power and utility of bioinformatics tools in the study of SDPs.

List of Tables

List of Figures

Abbreviations

3DEE	Domain Definition Database
AA	Amino Acid
AC	Accession Number
ASTRAL	ASTRAL Compendium for Sequence and Structure Analysis
BLAST	Basic Local Alignment Search Tool
CATH	Class Architecture Topology Homology database
CDS	Coding Sequence
DALI	Distance matrix Alignment
DC	Disulphide Connectivity
DNA	Deoxyribonucleic acid
DSSP	Definition of Secondary Structure of Proteins
EBI	European Bioinformatics Institute, U.K.
EMBL	European Molecular Biology Laboratories
FSSP	Family of Structurally Similar Proteins
GO	Gene Ontology
GOA	Gene Ontology Annotation system
NCBI	National Center for Biotechnology Information, U.S.A.
PDB	Protein Data Bank
PIs	Proteinase Inhibitors
Pot II	Potato type II proteinase inhibitor
Pfam	Protein Family database
ProDom	Protein Domain Database
RAF	ASTRAL Rapid Access Format for ATOM to SEQRES maps

RMSD	Root Mean Square Deviation
RUs	Repeat Units
SCOP	Structural Classification of Proteins
SDPs	Small Disulphide-rich Proteins
SDFs	Small Disulphide-rich Folds
SPACI	Summary PDB ASTRAL Check Index
SQL	Structured Query Language

Chapter 1 Introduction

Among the 20 standard amino acids, cysteine residues in secreted proteins have a unique property since they may pair to form disulphide bridges which contribute to the thermodynamic stability of the 3D structure. The disulphide bond is formed by the post-translational oxidation of two thiol (-SH) groups leading to the forming of a covalent S-S bond between the cysteine residues. This property was first highlighted by the pioneering work of Anfinsen on ribonuclease. According to Anfinsen's results fully denatured proteins can recover their native structure and restore the correct disulphide connectivity *in vitro* (Anfinsen and Haber 1961; Anfinsen *et al.* 1961; Anfinsen 1973). Disulphide bridges can increase the conformational stability of proteins mainly by constraining the unfolded conformation (Wedemeyer *et al.* 2000), and this effect is more significant for small proteins (Harrison and Sternberg 1994). Therefore small disulphide-rich proteins (SDPs) are good candidates for understanding the structure, conservation and evolution effects of cysteines and disulphide bridges in disulphide-bonded proteins. This thesis describes our effort to understand the roles of cysteines and disulphide bridges in SDPs through bioinformatics approaches.

The initial aim of this study is to develop automated comparative modeling methods specifically for SDPs to narrow the sequence-structure gap and thereby assign functionality to the large number of SDPs that have no structural or functional information. Building such a modeling method requires: (1) a high quality non-redundant template repository; (2) rules for the comparative modeling of SDPs. These requirements and distinct features of SDPs have inspired us to build a comprehensive database for small disulphide-rich folds (SDFs) and then carry out the systematic analysis of SDFs to study the roles and patterns of cysteines and disulphide bridges in SDPs (Chapter 2). The results of database curation and data analysis provide a non-redundant template dataset as well as rules for designing the modeling method. Based on the above, an automated comparative modeling method, SDPMOD, has been developed (Chapter 3) and applied to large scale comparative modeling of conotoxins, a family of SDPs. Moreover, the topology and parameter definition libraries for non-standard residues occurring in conotoxins have also been developed to overcome the bottlenecks of conotoxin modeling (Chapter 3).

Comparative modeling is dependent on homologous proteins adopting similar folds, which are indicative of their underlying function. Among the SDPs, we noted that domain duplication is a frequent occurrence and these duplicated domains fold into architectures with tandem repeat structures. The only exception to this observation is the Potato II (Pot II) proteinase inhibitor family. During SDF analysis and comparative modeling of SDPs, a specific family of SDPs, Pot II, came to our attention due to its multiple disulphide connectivities for the same fold and to the numerous evolutionary phenomena found in this family. To ensure that we understand how all SDPs fold, a comprehensive computational analysis was done on the Pot II family and interesting findings are reported in Chapter 4. Of them, one of the most interesting findings is that the cysteine scaffold in Pot II domain is under "purifying selection" (Kondrashov *et al.* 2002) to maintain the fold and the reactive sites under positive selection to target a broad range of proteinases from pathogens. This provides a perfect example how small disulphide-rich folds can be used to design novel proteins for drug or other bioactive molecules.

In Chapter 1, I will firstly review the background knowledge on disulphide bridges, including its formation and its roles in biological systems. Then I will define the focal theme of this thesis: small disulphide-rich proteins (SDPs) and small disulphide-rich folds (SDFs) and their features, applications and comparative modeling of SDPs. Since the comparative modeling of SDPs requires specific rules derived from systematic analysis of cysteines and disulphides in SDPs, the current databases and studies related to disulphide and disulphide-bonded proteins are briefly described. Using the domain as the basic unit to study SDPs and SDFs, the definition for domain is discussed and available structure-based domain databases are reviewed. At the end of Chapter 1, the bioinformatics problems in the study of SDPs are introduced and the objectives and contributions of this thesis are described.

1.1 Introduction to disulphide bonds

Before describing disulphide bridges, I would like to discuss the cysteine residue first. Cysteine is one of the special amino acids among the 20 standard amino acids. It has a hydrophobic methylene group ($–CH_2–$) group and a terminal sulfhydryl groups (-SH), also known as thiol group. The thiol group makes cysteine the most reactive amino acid side chain, participating in various reactions. For example, thiols of cysteine reisdues can form complexes of varying stability with a variety of metal ions (such as copper, zinc, iron), which is the basis of the high–affinity binding of metal ions (e.g. by zinc-finger transcription factors). The sulphur atom of cysteine residues can exist in diverse oxidation states, but the disulphide bond is most likely to be the end product in an oxidative milieu. Because of the special features of cysteine, this residue is hard to be substituted by other amino acids and remains one of the most conserved residues in proteins.

Disulphide bonds (also called disulphide bridges) are formed by the oxidization of thiol group of two cysteine residues. The disulphide bond covalently crosslinks regions which might be far apart in the protein's primary sequence. It can occur intra-molecularly (within a single polypeptide chain) and inter-molecularly (between two polypeptide chains). Intra-molecular disulphide bonds stabilize the tertiary structures of proteins while inter-molecular disulphide bonds are involved in stabilizing quaternary structure. Not all proteins contain disulphide bridges as these occur almost exclusively in extracytoplamic proteins.

In the following section, I will briefly introduce how disulphide bonds are formed in prokaryotic or eukaryotic cells, which is indispensable for understanding the roles and patterns of the disulphide in proteins.

1.1.1 Formation of disulphide bonds

In 1960s, Anfinsen and coworkers showed the native disulphide bonding of fully denatured ribonuclease A can be restored spontaneously *in vitro* with presence of molecular oxygen (Anfinsen *et al.* 1961). These studies led to the assumption that the disulphide bond formation is a spontaneous process *in vivo*. However, the formation of native disulphide bonds *in vitro* required hours or even days of incubation, while disulphide bond formation in the cell usually occurs within seconds or minutes after protein synthesis. The discovery of the DsbA gene in *E. coli* revealed that disulphide bond formation is actually a catalyzed process *in vivo* (Bardwell *et al.* 1991). Later a group of thiol-disulphide oxidoreductases were identified both in prokaryotic or eukaryotic organisms (Dailey and Berg 1993; Missiakas *et al.* 1995; Frand and Kaiser 1998). Currently, the pathways for disulphide bond formation have been characterized in both prokaryotic and eukaryotic organisms.

In prokaryotes, disulphide bonds are formed by the oxidation of thiol-disulphide oxidoreductase DsbA. Non-native disulphide connectivity can be rearranged by the isomerization of thiol-disulphide oxidoreductase DsbC. Disulphide bonds are generally formed in the periplasm. This is due to the reducing environment of the cytoplasm and the oxidative environment of the periplasm. Similarly, in eukaryotic cells, disulphide bonds are generally formed in the lumen of the ER (endoplasmic reticulum) and not in the cytosol because of the oxidative milieu of the ER and the reducing milieu of the cytosol. Thus, disulphide bonds are mostly found in secretory proteins, lysosomal proteins, and the exoplasmic domains of membrane proteins.

In eukaryotic cells, oxidizing equivalents for disulphide-bond formation are introduced into the ER by two parallel pathways. In the first pathway, oxidizing equivalents flow from Ero1 (ER oxidoreduction) to the thiol-disulphide oxidoreductase protein disulphide isomerase (PDI), and from PDI to secretory proteins through a series of direct thiol-disulphide exchange reactions. In the second pathway, the ER oxidase, Erv2 transfers disulphide bonds to PDI before substrate oxidation. Erv2 obtains oxidizing equivalents directly from molecular oxygen through its flavin cofactor.

From the pathways and locations of disulphide bond formation, several points are worthy to of notice for computational studies.

(1) Depending on the organism and cellular location of cysteine-containing proteins, cysteines can be oxidized to form disulphide bonds or reside in the reduced state as free cysteines. Prior to cysteine bonding state prediction and disulphide connectivity prediction, information related to the organism and the

cellular location of the protein should be considered. For example, signal peptides generally determine the cellular location of the protein and thus signal peptides may help in the prediction of cysteine-bonding states.

(2) Although there are many possible disulphide connectivities for multi-disulphide proteins, only one of them is the native connectivity. Non-native connectivities are possible under some circumstance or conditions and they can be rearranged to native disulphide connectivity by isomerization *in vivo*.

1.1.2 Roles of disulphide bridges

Disulphide bonds can be divided into two classes:

(1) **stabilizing disulphide**

Most disulphide bonds belong to this class and form the stable part of folded protein structures, especially in small proteins.

(2) **reactive disulphide**

Disulphide bonds in some proteins can alternate between the reduced and oxidized states to participate specific oxidation-reduction functions.

Disulphide bonds of the first class may contribute to the folding pathway of the protein and to the stability of its native fold. Researchers have applied this feature to design and engineer new disulphide bonds in proteins to improve their thermostability (Perry and Wetzel 1984; Mansfeld *et al.* 1997; Robinson and Sauer 2000; Martensson *et al.* 2002).

Besides stabilization of protein structures, disulphide bonds also have been reported to have other roles. In bacteria, disulphide bonds can play an important protective role as a reversible switch that turns a protein on or off when bacterial cells are exposed to oxidation reactions by hydrogen peroxide (H_2O_2), which could severely damage DNA and kill the bacterium at low concentrations if not for the protective action of the disulphide bonds. In some eukaryotic cells, it is reported that specific cleavage of one or more disulphide bonds can control the function of some secreted soluble proteins and cell-surface receptors (Hogg 2003).

1.2 Small Disulphide-rich Proteins (SDPs) and Small Disulphide-rich Folds (SDFs)

1.2.1 The definitons of SDPs and SDFs

All proteins can be classified into disulphide-containing proteins (also called disulphide-bonded proteins) and non-disulphide proteins according to the occurrence of disulphide bond. Among disulphide-bonded proteins, this thesis particularly focuses on small disulphide-rich proteins.

Before exploring further, I would like to clarify two concepts used in this study: Small Disulphide-rich Proteins (SDPs) and Small Disulphide-rich Folds (SDFs). These are highly similar and closely related but they also have minor differences. Both concepts has been used by scientists in previous studies (Harrison and Sternberg 1996; Mas *et al.* 2001). Generally disulphide-rich proteins are defined as having more than two disulphide bonds. And for small proteins, there are no widely accepted criteria. Harrison and Sternberg reported that different physical models should be used to describing disulphide connectivities for short sequences and longer sequences (Harrison and Sternberg 1994). They suggested that for short sequences as (less than 75 residues) native disulphide connectivities tend to have entropically greater-stabilising arrangement features (entropic model), while longer sequences (longer than about 200 residues) are better described by diffusive contact in the unfolded states (diffusive model). In their later research on disulphide β-Cross, they defined small disulphide-rich folds as ≤ 100 residues and with ≥ 2 disulphides (Harrison and Sternberg 1996).

In this study, both concepts are used in different situations. SDFs are practically defined as small domains (size less than 100 residues) and have at least two disulphide bonds (same as Harrison's), while SDPs are defined as proteins which are composed of SDF domains.

Generally, SDFs have broader scope since they may include small disulphide-rich domains from large proteins which also contain non-SDF domains, while SDPs are always composed of SDFs.

1.2.2 The applications of SDPs

Small disulphide-rich proteins (SDPs) are a special class of proteins with diverse functions. They include many secretory proteins, which serve predatory, defensive or regulatory roles (such as toxins, inhibitors and hormones). SDPs are involved in various biological functions and pathways and therefore many important applications:

(1) They are a "gold mine" for therapeutic drugs (Shen et al. 2000). For example, ancrod and angiotensin converting enzyme inhibitor, Captopril, from snake venom can be used for treatment of heart attack patients (von Segesser *et al.* 2001).

(2) SDPs are also very useful tools in protein-protein interaction research. For example, conotoxins are used as research tools to characterize different ion channels subtypes and molecular isoforms of receptors (Lewis 2004; Li

and Tomaselli 2004) where analyses of toxin-channel/receptor complex interfaces can expedite drug discovery.

(3) Some SDPs also serve as pesticides, such as plant proteinase inhibitors which can block insect gut proteases (Richardson 1977).

Despite the biomedical importance of SDPs, the three-dimensional structures are not available for many such proteins. This deficiency requires to be addressed by comparative modeling of SDPs, discussed in the following section.

1.2.3 Comparative modeling of SDPs

To understanding the functional roles of SDPs and exploit their applications in drug design, structural information is always essential. Studies on protein function, especially interactions between proteins, often require the availability of 3D structures. To comprehend complex biological functions, structure information is indepensable. Single amino acid mutations may result in significant changes in 3D structures and affect the function of a protein. For example, α-conotoxin ImI is a highly specific antagonist for the neuronal α7 nicotinic acetylcholine receptor (nACh receptor). The activity of its single-residue mutant (with residue 5 changed from aspartic acid to asparagine) was reduced by at least two orders of magnitude in comparison to the wild type ImI (Rogers *et al.* 2000). 3D structures are essential in drug design to improve ligand characteristics, *in silico* mutation and protein-protein interaction studies.

However, 3D structural information is only available for a small subset of proteins. Structure determination through experimental methods such as X-ray crystallography and Nuclear Magnetic Resonance Spectroscopy (NMR) are still both time-consuming and expensive although the advances of techniques and structural genomics projects. With the rapid growth of sequence data, it is impractical to experimentally solve 3D structures for all known protein sequences. This results in a huge gap between the number of known 3D structures and the number of primary sequences. According to the latest statistics (07-Feb-2006) of the UniProt database (Wu *et al.* 2006) and the Protein Data Bank (Kouranov *et al.* 2006), TrEMBL Release 32.0 contains 2,605,584 entries and SwissProt Release 49.0 (07-Feb-2006) holds 207,132 proteins whereas PDB has only 32,009 protein structures (1.23% and 15.4%, respectively of the protein sequence databases). However, this enormous structure-sequence information gap can be narrowed using large-scale automated protein structure prediction.

Currently protein structure prediction methods can be classified into three major classes: comparative structure prediction (homology modeling), fold recognition (also called threading) and *de novo* prediction (or *ab initio* modeling) (Baker and Sali 2001). Comparative modeling methods produce 3D models of given sequences based on the target-template alignment to one or more related protein structures. Fold recognition methods scan protein sequences against known 3D structures and evaluate the sequence-structure fitness, which can sometimes reveal more distant relationships than purely sequence-based methods. *De novo* methods are based on the assumption that the native structure of a protein is at the global free energy minimum, and do not require known any protein structure information. These methods carry out a large-scale search of conformational space for protein tertiary structures that are particularly low in free energy for the given amino acid sequence. These structure prediction methods are compared in Table 1.

Table 1 Comparison of protein structure prediction methods

Method	**Comparative modeling (Homology modeling)**	**Fold recognition (Threading)**	***De novo* prediction (ab initio modeling)**
Requirement of related 3D structure(s)	Yes	Yes	No
Sequence similarity	ID% ≥ 30%	< 30%	N.A
Computational time	Fast and scaleable	Slow	Extremely slow
Applicable size of protein	Almost no limits, provided a homologous template is available	Single domain	Small or medium size proteins
Model accuracy	High	Medium	Low

Among these structure prediction methods, *de novo* methods are extremely computationally intensive and are not applicable to large-scale structural modeling even though they do not require known related structures. Threading methods are less restrained by detectable sequence similarity but they are not as accurate as comparative modeling methods. Comparative modeling methods are the most reliable and accurate for generating 3D models among the three classes. They are also relatively fast and can be used for large-scale modeling. Comparative modeling methods have been applied at genomic scales to generate 3D models for proteins in *Saccharomyces cerevisiae* genomes (Sanchez and Sali 1998) or the entire SwissProt database (Guex *et al.* 1999). Structural Genomics projects worldwide are currently addressing the issue of determining all the representative structures so that most structure prediction problems will be reduced to comparative modeling (Rost 1998; Brenner and Levitt 2000; Chandonia and Brenner 2005; Xie and Bourne 2005).

Comparative modeling of protein structures often requires expert knowledge and proficiency in specialized methods. In the mid-1990s, Peitsch and co-workers developed the first automated modeling server SWISS-MODEL (Peitsch 1996), which is currently the most widely-used server of this genre. Recently, several other automated comparative modeling servers have emerged, such as CPHmodels (Lund *et al*. 1997), 3D-JIGSAW (Bates *et al.* 2001), ModWeb (Pieper *et al.* 2002) and ESyPred3D (Lambert *et al*. 2002).

Although so many automated comparative modeling servers are available, most of them do not work well on SDPs due to two reasons. Most of the automated servers are primarily designed for globular protein domains, making it difficult to discriminate SDPs with relatively small sizes, from background noise. Taking as an example the sequence of α-conotoxin PnIA (Hu *et al.* 1996) (PDB ID: 1PEN; 16 residues; 2 disulphide bridges in its structure), we note that both SWISS-MODEL and ModWeb report that they do not cover the modeling of sequences length less than 25 or 30 amino acids, respectively, while the other three servers state that no suitable templates can be identified for this sequence.

The second reason is that SDPs have distinct characteristics from medium and large globular proteins. They usually do not have a compact hydrophobic core, which is a major factor in stabilizing globular protein structure. SDPs tend to have less secondary structures and more solvent-exposed hydrophobic residues compared to larger proteins. Comparative modeling techniques tend to rely on the characteristics of assembling secondary structural units, which are only present to a limited extent in small peptides and/or small proteins such as SDPs; and burying hydrophobic residues while exposing charged residues. The 3D structures of small proteins are usually dominated by disulphide bridges, metal or ligands, according to their SCOP classification (Murzin *et al*. 1995), and tend to bind or interact with globular proteins.

In small disulphide-rich proteins, the effects of disulphide bridges and constrained residues such as prolines are more significant in determining their 3D structures. Unlike short peptides which are flexible enough to be able to adopt many conformations, SDPs are sufficiently constrained to form stable structures. For comparative modeling of such small structures, rules will have to be highly specific and different from those adopted for large globular proteins. The distinct features of SDPs require specific methodology to be developed for comparative modeling.

The development of such a modeling method further requires the availability of high quality non-redundant template repository and systematic analysis of SDPs to derive rules for automated comparative modeling. The following section will review currently available databases and related studies on disulphide and disulphide-bonded proteins.

1.3 Databases related to disulphide bridges

Disulphide bridge information can be obtained from a variety of resources, mainly public databases and literatures. These public databases can be classified into primary (where biologists deposit their data) and secondary databases (database derived from primary database).

1.3.1 Primary databases on disulphide information

The primary databases can be further classified into sequence and structure databases. Among the sequence databases, SwissProt database (Boeckmann *et al.* 2003) provides the largest number of annotated disulphide information. It contains both experimentally determined disulphides and inferred disulphides (annotated "By similarity"). Inferred disulphide annotations are assigned only when a protein sequence has a clear sequence homology to another protein with experimentally determined disulphide information. These inferred disulphide annotations should be used with caution since they may contain incorrect information.

Among the structure databases, Protein Data Bank (PDB) (Berman *et al.* 2000) is the most abundant resource for disulphide information. Beside disulphide connectivity, much more related information, such as secondary structure, solvent accessibility and dihedral angles, can be derived from PDB structures. The unambiguous and rich disulphide information available from PDB provides both accurate and comprehensive information for the study of disulphide bonds or disulphide-bonded proteins.

In consideration of data quality and features available for further in-depth investigation, PDB was selected as the main data source for the analysis of disulphides in this study.

1.3.2 Secondary databases on disulphide information

Several secondary databases (Table 2) centered on disulphide bridges were developed (Chuang *et al.* 2003; Tessier *et al.* 2004; van Vlijmen *et al.* 2004; Vinayagam *et al.* 2004). These databases have different foci and are suitable for different applications, as described below.

Table 2 Secondary databases on disulphide bonds

Database	**Data source**	**Basic unit**	**Feature**	**URL**
SSDB	PDB	PDB chain	Classification	http://e106.life.nctu.edu.tw/~ssbond/
DSDBASE	PDB	Disulphide	Protein engineering	http://caps.ncbs.res.in/dsdbase/dsdbase.html
DisulphideDB	PDB	PDB chain	Cysteine-bondng state prediction	Not available
Disulphide pattern DB	SwissProt	Pfam domain	Disulphide patterns	Not available

SSDB is a disulphide classification database that clusters disulphide-bonded proteins based on a hierarchical clustering scheme (Chuang *et al.* 2003). The curators collected 3,134 disulphide-bonded (disulphide number ≥ 2) proteins chains from PDB and treated each PDB chains as separate units. In SSDB, protein chains are classified hierarchically in three levels: disulphide-bonding numbers, disulphide-bonding connectivity and disulphide-bonding patterns. They reported that disulphide-bonding patterns could be used to detect the structural similarities of proteins of low sequence identities (<25%).

DSDBASE is a database of native and modeled disulphide bonds in proteins (Vinayagam *et al.* 2004), which provides information on native disulphides and those that are stereochemically possible between pairs of residues for all PDB structures. The modeled disulphides are obtained using MODIP (Sowdhamini *et al.* 1989), by the identification of residues pairs that can host a covalent cross-link without strain. The main application of DSDBASE is to design site-directed mutants in order to improve the thermal stability of a protein. DSDBASE can also be used for the modeling of disulphide-rich proteins.

The DisulphideDB database collected disulphide information with structural, evolutionary and neighborhood information on cysteines in proteins (Tessier *et al.* 2004). The data collection is based on a representative selection of PDB structures – PDBSELECT <http://bioinfo.tg.fh-giessen.de/pdbselect/> and only retains PDB chains from eukaryotic cells with at least one disulphide bond annotation in the PDB files. The disulphide information is used to derive rules for cysteine-bonding state prediction.

A database of disulphide patterns was developed by van Vlijmen and coworkers for analyzing disulphide patterns in proteins (van Vlijmen *et al.* 2004). The database was constructed using disulphide annotations from SwissProt, and was expanded by an inference method that combines SwissProt annotations with Pfam multiple sequence alignments. This database contains 94,999 disulphide-bonded domains and was used to detect distantly related homologs.

Although several disulphide-related databases have been constructed, all of them cannot fulfil the needs of this study due to the following reasons:

(1) Focus. None of these databases are specifically focused on SDPs.

(2) Availability. Neither DisulphideDB nor Disulphide pattern database (van Vlijmen *et al.* 2004) are available on the Internet.

Structural domains. None of these databases are based on structural domains. SSDB and DisulphideDB use PDB chains as the basic unit, which is unsuitable to the analyses of cysteine and disulphide patterns of multi-domain proteins. For example, SSDB has classified the proteinase inhibitor C1-T1 from *Nicotiana alata* (PDB ID: 1FYB, Chain A; Figure 2) in the eight-disulphide group according to its disulphide number in its structure.

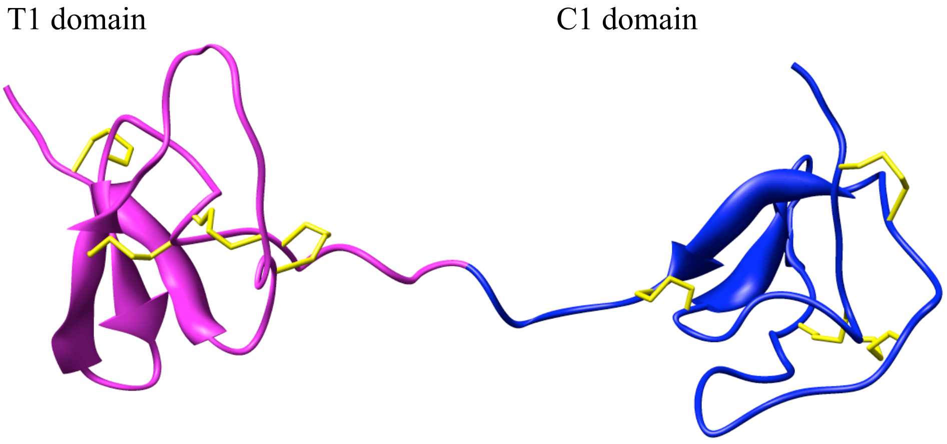


Figure 1 The structure and disulphide connectivity of C1-T1 (PDB ID: 1FYB, Chain A), a two-domain proteinase inhibitor derived from the six-domain precursor protein Na-ProPI. The structure is in ribbon representation, with disulphide bridges depicted in stick mode. Domain C1 (1-55) is colored in blue and domain T1 (56-111) in magenta.

Figure 1 shows the structure and disulphide connectivity of C1-T1 (PDB ID: 1FYB). Both domain C1 (Chymotrypsin-specific domain-1) and domain T1 (Trypsin-specific domain-1) have the same structural features (an anti-parallel β-sheet) and the same disulphide connectivity. Both of them are classified into the SCOP family Plant Proteinase Inhibitors. This example clearly shows the weakness of PDB chains as basic unit to analyze patterns of cysteines and disulphides. Based on such considerations, the domain was selected as basic unit for this study. In the section 1.4, protein domains and structure-based domain databases are described.

1.4 Reviews on domain and structure-based domain databases

The concept of protein domains is very important for studies on structure, function, and evolution of proteins. The modular architecture of proteins has been widely recognized for over a decade now (Wetlaufer 1973; Baron *et al.* 1991; Henikoff *et al.* 1997; Schultz *et al.* 1998). Proteins are composed of smaller building blocks, which are called "domain" or "modules". These building blocks are distinct regions of 3D structure resulting in protein architectures assembled from modular segments that have evolved independently. The modular nature of proteins has many advantages, offering new cooperative functions and enhanced stability. As a result of the duplication and mutational evolution of these building blocks through various gene rearrangement and stabilizing selection mechanisms, respectively, a large proportion of proteins in higher organisms especially eukaryotic extracellular proteins, consist of multiple domains (Apic *et al.* 2001). Knowledge of protein domain architecture and domain boundaries is essential for the characterization and understanding of protein function.

There are a number of databases providing domain definition and information. These domain databases can be classified into sequence-based domain databases and structure-based databases according to their data resource. Structure-based databases contain domain information derived from PDB structure while sequence-based databases are mainly based on sequence information. Domain databases and their web address are listed in Table 3.

Table 3 List of databases that contain domain information.

Database	**URL**
Sequence-based domain databases	
ProDom	http://prodes.toulouse.inra.fr/prodom/current/html/home.php
DOMO	http://www.infobiogen.fr/services/domo
BLOCKS	http://blocks.fhcrc.org/blocks/blocks_search.html
COGs	http://www.ncbi.nlm.nih.gov/COG
SMART	http://smart.embl-heidelberg.de
Pfam	http://www.sanger.ac.uk/Software/Pfam
SBASE	http://www.icgeb.trieste.it/sbase
Interpro	http://www.ebi.ac.uk/interpro
Structure-based domain databases	
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop
CATH	http://www.biochem.ucl.ac.uk/bsm/cath
3Dee	http://www.compbio.dundee.ac.uk/3Dee
DALI/FSSP	http://www.ebi.ac.uk/dali/fssp
MMDB	http://www.ncbi.nih.gov/Structure/MMDB/mmdb.shtml

Since PDB is selected as the main data source for this study, only structure-based domain databases are described as follows.

1.4.1 SCOP

The SCOP (Structural Classification Of Proteins) database is a comprehensive classification of all structures in PDB according to their evolutionary and structural relationship (Murzin *et al.* 1995; Lo Conte *et al.* 2000; Andreeva *et al.* 2004). The domain assignment in SCOP is based on both evolutionary relationship and structure features. Therefore some of the domain definitions are different from other structure-based domain databases. All the domains in SCOP are classified according to a four-level hierarchy: Family, Superfamily, Fold and Class.

(1) *Family.*

Proteins are clustered together into families on the basis of one of two criteria that imply their having a common evolutionary origin: first, all proteins that have residue identities of 30% and greater; second, proteins with lower sequence identities but whose functions and structures are very similar; for example, globins with sequence identities of 15%.

(2) *Superfamily.*

Families, whose proteins have low sequence identities but whose structures and, in many cases, functional features suggest that a common evolutionary origin is probable, are placed together in superfamilies; for example, the variable and constant domains of immunoglobulins.

(3) *Common Fold.*

Superfamilies and families are defined as having a common fold if their proteins have the same major secondary structures in the same arrangement and with the same topological connections. The structural similarities of proteins in the same fold category probably arise from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

(4) *Class*.

The different folds have been grouped into classes. Most of the folds are assigned to one of the five structural classes:

- All-α, structures essentially formed by helices.
- All-β, structures essentially formed by β-sheets.
- α/β (Mainly parallel β sheets), structures with α-helices and β-strands
- α+β (Mainly anti-parallel β sheets), structures with α-helices and β-strands are largely segregated.
- Multi-domain, structures with domains of different folds and no homologues are known at present.
- Membrane and cell surface proteins and peptides.
- Small proteins. Usually dominated by metal ligand, heme, and/or disulphide bridges.

Other classes have been assigned for Peptides, Designed proteins, Coiled coil proteins and Low resolution protein structures.

1.4.2 CATH

CATH (Pearl *et al.* 2003) is also a hierarchal classification database of protein domain structures, which clustered protein domain in five principal levels: Class (C), Architecture (A), Topology (T), Homologous superfamily (H) and Sequence family (S). The domain definitions were assigned by a consensus procedure based on three domain recognition algorithms: DETECTIVE (Swindells 1995), PUU (Holm and Sander 1994) and DOMAK (Siddiqui and Barton 1995)) as well as manual assignment. CATH domains are classified manually at C- and A-levels and automatically at T-, H- and S-levels.

(1) *Class, C-level.*

Class is determined from the protein structure secondary structure composition and its packing within the structure. Three major classes are recognized: mainly-α, mainly-β and α-β. This last class (α-β) includes both alternating α/β structures and α+β structures. A fourth class is also identified which contains protein domains, which have low secondary structure content. The class number is assigned using the automatic method of Michie *et al*. (Michie *et al.* 1996).

(2) *Architecture, A-level*

This describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures. It is currently assigned manually using a simple description of the secondary structure arrangement e.g. barrel or 3-layer sandwich. Reference is made to the literature for well-known architectures (e.g the β-propeller or α-helix bundle). Procedures are being developed for automating this step.

(3) *Topology (Fold family), T-leve*l

Structures are grouped into fold families at this level depending on both the overall shape and connectivity of the secondary structures. This is done using the structure comparison algorithm SSAP (Orengo and Taylor 1996). Parameters for clustering domains into the same fold family have been determined by empirical trials throughout the development of this databank. Structures having an SSAP score of 70 with at least 60% of the larger protein matching the smaller protein are assigned to the same T level or fold family.

(4) *Homologous Superfamily, H-level*

This level groups together, the protein domains that are thought to share a common ancestor and can therefore be described as homologous. Similarities are identified first by sequence comparisons and subsequently by structure comparison using SSAP. Structures are clustered into the same homologous super-family if they satisfy one of the following criteria:

- Sequence identity >= 35%, 60% of larger structure equivalent to smaller
- SSAP score >= 80.0 and sequence identity >= 20%
- 60% of larger structure equivalent to smaller
- SSAP score >= 80.0, 60% of larger structure equivalent to smaller, and domains that have related functions.

(5) *Sequence families, S-level*

Structures within each H-level are further clustered on sequence identity. Domains clustered in the same sequence families have sequence identities >35% (with at least 60% of the larger domain equivalent to the smaller), indicating highly similar structures and functions.

1.4.3 DALI/FSSP

DALI/FSSP database presents a fully automatic classification of all the known protein structures (Holm and Sander 1998). The classification is derived from using an all-against-all comparison of all the structures in PDB by an automatic structural alignment method DALI (Holm and Sander 1993). The structural domains are defined by a modified version of ADDA algorithm (Heger and Holm 2003). The criteria of recurrence and compactness are used for finding the domain boundaries and each domain is assigned a Domain Classification number DC_I_m_n_p represention:

- Fold space attractor region (I) represents the architecture of the proteins. There are now six fold space attractors defined based on the secondary structure composition and the supersecondary structural motifs. Attractor 1 consists of α/β, attractor 2 consists of all-α, attractor 3 consists of all-β, attractor 4 consists of anti parallel β barrels and attractor 5 contains α/β meander.

- Globular folding topology (m) represents all the domains with the same topology but having with shifts in the relative orientation of the secondary structures. They are obtained empirically based on a tree constructed by average linkage clustering of the structural similarity score. The folds are classified based on the DALI Z score levels of 2, 4, 8, 16, 32 and 64. The first level ($Z > 2$) has been used as an operational definition of folds. The higher the Z score, the higher the structural similarities among the protein structures.
- Functional family (n) represents inferred plausible evolutionary relationships from strong structural similarities, which are accompanied by functional or sequence similarities. Functional families are branches of the fold dendrogram where all pairs have a high average neural network prediction for being homologous. The neural network weighs evidence coming from: overlapping sequence neighbors as detected by PSI-BLAST, clusters of identically conserved functional residues, Enzyme Commission (E.C.) numbers, SwissProt keywords. The threshold for functional family unification was chosen empirically and is conservative; in some cases the automatic system finds insufficient numerical evidence to unify domains, which are believed to be homologous by human experts.
- Sequence family (p) represents subsets of protein structures that have proteins with sequence identity greater than 25%.

1.4.4 3Dee

3Dee (Database of Protein Domain Definitions) is a comprehensive collection of protein structural domain definitions (Siddiqui *et al.* 2001). The domains in 3Dee are defined on a purely structural basis. DOMAK algorithm (Siddiqui and Barton 1995) was used to define all domains when the database was first built. For later updates, the domains were defined by sequence alignment to existing domain definitions or manually. All the domains in 3Dee were organized a hierarchy of three levels: Domain families (sequence redundant domains), Domain sequence families (structure redundant domains) and Domain structure families (non-redundant on structure) (Dengler *et al.* 2001).

1.4.5 MMDB

MMDB (Molecular Modeling Database) is NCBI (National Center for Biotechnology Information) Entrez's 3D-structure database (Chen *et al.* 2003) derived from the PDB. MMDB contains two kinds of domains: "3D domain" and "Conserved Domain"(Chen *et al.* 2003). 3D Domains in MMDB are structural domains, which are assigned automatically using an algorithm that searches for one or more breakpoints such that the ratio of intra- to inter-domain contacts falls above a set threshold(Madej *et al.* 1995). Conserved domains in MMDB are recurrent evolutionary modules defined by Entrez's CDD (Conserved Domain Database) (Marchler-Bauer *et al.* 2003) where the domains are derived from SMART (Letunic *et al.* 2004), Pfam (Heger and Holm 2003) and COGs (Tatusov *et al.* 2003).

1.4.6 The selection of domain database for this study

As described above, there are several structure-based domain databases available.

They are derived by different methods and therefore the domain definition and classification for the same domain is different among these databases. Figure 2 illustrates an example of different domain boundary assignments for the same protein in different domain databases.

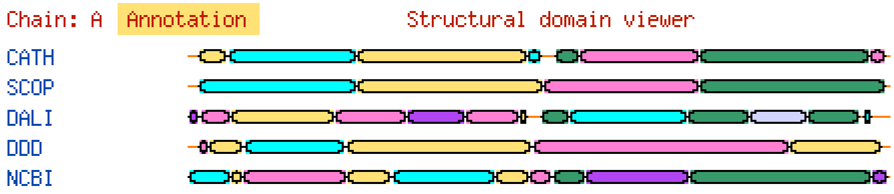


Figure 2 Domain definitions for D-Glucose 6-Phosphotransferase (PDB ID: 1HKB, Chain A) are dissimilar in different structure-based domain databases. The domain assignments are collated and visualized by XdomView (Vivek *et al.* 2003). Segments with the same color or number are assigned to the same domain.

Figure 2 shows the different domain definitions in different domain databases for the same protein. Among the five databases, DALI tends to divide protein structures into small and compact domains while SCOP is reluctant to split the domains unless there is some evidence to support to do so. In this study, SCOP is selected to be the major source for domain definition because of the following reasons:

(1) SCOP considers both evolutionary and structure information for assigning domains, while other databases mainly based on structure information to define domain. Since disulphides are always conserved during evolution to stabilize the structure and fold, SCOP domain definition will better represent the evolutionary relationship between homologous disulphide-bonded proteins.

(2) SCOP is manually curated by experts with visual inspection thus is likely

the most reliable resource for domain definition and classification. DALI, 3Dee and MMDB are generated by computer program automatically. CATH is built based on semi-automated method: manually at Class (C) and Architecture (A) levels and automated at Topology (T), Homologous superfamily (H) and Sequence family (S) levels. Therefore, for some low level classification, CATH may not be as accurate as SCOP. For example, both domains of C1-T1 (PDB ID: 1FYB, Chain A) and PCI-1 (PDB ID: 4SGB, Chain I) clearly belongs to the same sequence family, but they are classified into two sequence families (3.30.60.30.6: complex (serine proteinase-inhibitor) and 3.30.60.30.7: hydrolase) in CATH. While in SCOP, all the Pot II domains were correctly classified into SCOP family labeled plant proteinase inhibitors.

For these reasons, in this study, SCOP is selected as the major source for domain definition and domain classification and CATH is used for reference and in-depth analysis.

1.5 Objectives of this thesis

SDPs have great potential as therapeutic drugs, diagnostic agents and pesticides. The most important characteristic of SDPs is their cysteines and disulphides patterns. Due to the unique features of SDPs, applications of SDPs require an in-depth understanding of the nature of SDPs and the availability of corresponding computational resources, such as a high quality dataset and approaches specifically tailored for SDPs. The objectives of this thesis is to address these demands by systematic investigation of SDPs from the following specific aspects:

(1) Build a high quality and comprehensive dataset for the researches of SDPs;

(2) Analyze the roles and patterns of cysteines and disulphide bridges of SDPs and derive rules for further investigations and applications of SDPs;

(3) Develop computational methods specifically for SDPs, particularly on the comparative modeling of SDPs;

(4) Investigate SDP families for the in-depth understanding of structure, function and evolution of SDPs.

1.6 Contributions of this thesis

This thesis provides several novel contributions that are briefly described below:

(1) SDFD – a database of Small Disulphide-rich Folds (SDFs) has been curated to facilitate the research of SDPs and SDFs.

(2) A hierarchal classification scheme for SDFs is proposed based on disulphide number, disulphide connectivity and cysteine signature.

(3) Systematic analysis of SDFD reveals that the cysteine signature can help in detecting distantly related homologs and convergently evolved structures that are difficult to identify by sequence similarity searches.

(4) SDPMOD – a novel method for the automated comparative modeling of SDPs has been developed, specific rules for dealing with SPDs. The CHARMM22 forcefield topologies and parameters for non-standard residues has been generated and tested on large scale comparative modeling of conotoxins;

(5) The unique property of the Potato II (Pot II) proteinase inhibitor family to form structural repeats different from sequence repeats has been identified and investigated. A comprehensive analysis revealed that this family exhibits

“purifying” selection on the cysteine scaffold and positive selection on the reactive sites. The evolution of Pot II family showed the feasibility of using SDFs as scaffolds for drug design and protein engineering.

Chapter 2 Small Disulphide-rich Fold Database (SDFD)

Small disulphide-rich folds (SDFs) constitute a large group of proteins with diverse functions and have many important applications as discussed in Chapter 1. The most important characteristics of SDFs are their cysteine patterns and disulphide-bonding patterns.

To better understand the features of SDFs and facilitate the applications of SDFs, a comprehensive analysis of the roles and patterns of cysteines and disulphide bridges in SDFs is essential. Such an analysis requires the availability of a complete and accurate structural SDF dataset. Although several databases centered on disulphide proteins are available (Chuang *et al.* 2003; van Vlijmen *et al.* 2004; Vinayagam *et al.* 2004), they have different emphases and cannot fulfill the needs of this study (details in Chapter 1).

To facilitate the analysis of roles and patterns of cysteine residues and disulphide bridges in SDFs, a comprehensive database for SDFs and SDPs was built. SDF Database (SDFD) provides the clean and complete dataset for the analysis of SDFs and also serves as the template repository for comparative modeling of SDPs.

2.1 Data sources and data extraction

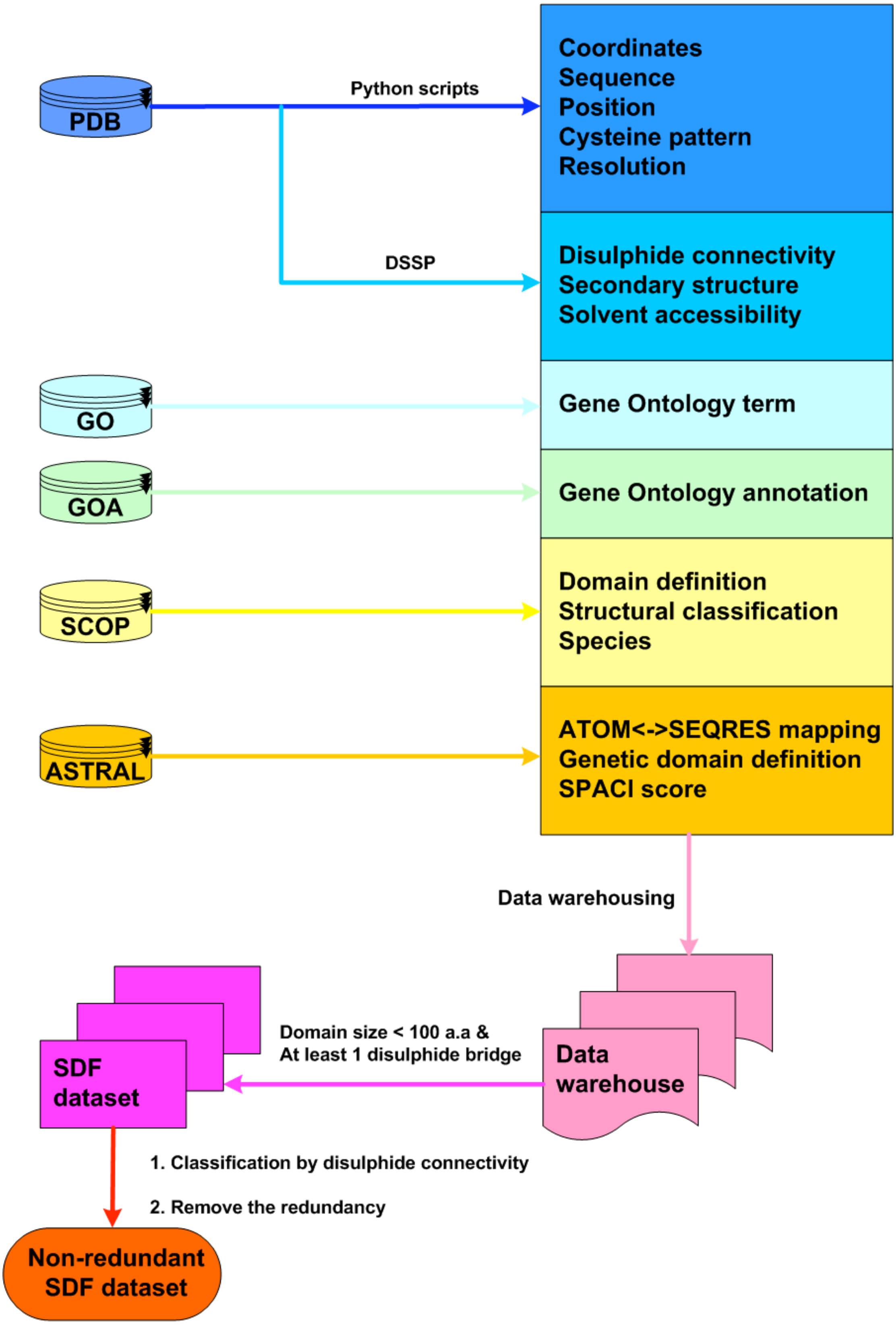


Figure 3 Flowchart shows data resources and data flow in SDFD.

Figure 3 shows the data flow involved in the creation of this database. SDFD is a heterogeneous database, incorporating information on protein structures and disulphide connectivity from PDB (Kouranov *et al.* 2006), protein domain definition and classification from SCOP (Andreeva *et al.* 2004), PDB ATOM—SEQRES correspondence maps, genetic domain definition and SPACI (Summary PDB ASTRAL Check Index) from ASTRAL (Brenner *et al.* 2000; Chandonia *et al.* 2002; Chandonia *et al.* 2004), Gene Ontology terms from the Gene Ontology Consortium (Ashburner *et al.* 2000) and functional annotation from GOA@EBI <http://www.ebi.ac.uk/GOA/index.html> (Camon *et al.* 2004). These data resources are described briefly in the following section. In this study, SDFs were collected according to the criteria of domain size ≤ 100 residues and with at least two disulphide bridges.

The basic unit for SDFD database is the "domain" as defined by SCOP (Andreeva *et al.* 2004), while most previous studies on disulphide bonding pattern use PDB chains as basic units (Harrison and Sternberg 1996; Chuang *et al.* 2003). Such consideration was due to an obvious problem during the analysis of cysteine patterns and disulphide-bonding patterns. For example, some multi-domain SDPs (such as the Pot II family discussed in Chapter 4) contain tandem domain duplication, so that extracting cysteine patterns or disulphide-bonding patterns based on PDB chains will introduce inaccuracies due to the repetition of a single unique pattern.

2.1.1 The Protein Data Bank

In this study, all the small disulphide-rich proteins were collected from Protein Data Bank (Kouranov *et al.* 2006). Protein 3D structures are the most accurate and informative resource for disulphide connectivity information. Although disulphide connectivity information can also be obtained from other resources, such as the annotation of Swiss-Prot (Boeckmann *et al.* 2003), some important features such as secondary structures and solvent accessibility are absent in sequence databases.

For each PDB structure, general information (such as experimental method, resolution, r-value and deposition date) and features for each protein chain (protein sequence from SEQRES and ATOM records) were extracted.

Disulphide connectivity, secondary structure and solvent accessibility were calculated using the DSSP algorithm (Kabsch and Sander 1983), which is a widely used program to calculate secondary structural features for PDB structures. Although the disulphide connectivity information was initially extracted from the SSBOND records in the PDB files, further study showed that for some PDB entries SSBOND records are incomplete or incorrect. For example, for the pancreatic trypsin inhibitor (PDB ID: 1B0C) chain E, there are six cysteines in the primary sequence (at positions: 5, 14, 30, 38, 51, 55), while the SSBOND record in the PDB files only reported 5-55, 30-51 as disulphide bridges. In fact, the distance between sulfur atoms of residues 14 and 38 is 2.04 Å, which was annotated as disulphide bonds by most structure analysis software, such as DSSP (Kabsch and Sander 1983), WHATIF (Vriend 1990) and PROMOTIF (Hutchinson and Thornton 1996). To obtain complete and accurate disulphide bonding information, DSSP was used to calculate disulphide connectivity as well as secondary structure and solvent accessibility for SDFD. Python scripts were written to extract all the useful information from the DSSP output files and populate the appropriate fields in the database.

2.1.2 SCOP and CATH

The domain is used as basic unit for SDFD. There are several public databases available for domain definition and domain classification (described in Chapter 1), such as DALI (Holm and Sander 1993), CATH (Pearl *et al.* 2003) and SCOP (Andreeva *et al.* 2004). In this study, SCOP was used for domain definition and classification since it is manually curated and is widely used as the "gold standard" for structural domain classification. SCOP 1.69 release (Aug. 2005) splits 25,973 PDB structures into 70,859 domains and classifies domains into four hierarchical levels: class, fold, superfamily and family. The domain definitions and classifications are retrieved from SCOP. CATH version 2.6.0 (Apr. 2005) was also downloaded as a reference structure classification database. FSSP is mainly derived from automatic domain classification programs and hence DALI data was not used in this study.

2.1.3 ASTRAL

ASTRAL (Chandonia *et al.* 2004) offers high quality curated data about PDB structures and SCOP domains in the following aspects:

(1) *ASTRAL RAF Sequence Maps* provide the mapping between the protein primary sequences defined in the SEQRES record of a PDB file to the actually reported atomic coordinates, found in the ATOM records. It is possible that the sequence from SEQRES records and the sequence in the ATOM records may be slightly different for some PDB entries, which is mainly due to the nature of structure determination techniques (especially X-ray crystallography). The coordinates of some residues cannot be completely determined so that the sequence from the ATOM record may vary from the biological sequence of the protein (available from the

SEQRES record). Such differences may cause problems during analysis. ASTRAL provides the PDB ATOM to SEQRES correspondence maps in Rapid Access Format to solve this problem.

(2) *Genetic domains*: a SCOP domain may include several fragments from different PDB chains. In most cases, these fragments are the product of a single gene. ASTRAL reassembles the fragments in the order found in the original gene sequence. Such information is valuable for the analysis of intra and inter-chain disulphide bridges. The definitions and sequences of genetic domains are retrieved from ASTRAL.

(3) *SPACI scores*: before data analysis, any redundancy should always be removed first from the dataset. Structure quality is one of the most important criteria for removing redundancy in structural data. SPACI scores from ASTRAL (Brenner *et al.* 2000) provide a reliable evaluation parameter for structure quality and the scores are calculated by combining three components: the quality of the experimental data (the resolution), how well the model fits the collected data (the R-factor), and the theoretical quality of the model, determined by stereochemical checks from PROCHECK (Laskowski *et al.* 1993) and WHAT_CHECK (Vriend 1990).

2.1.4 Gene Ontology (GO) and GOA@EBI

The analyses on function and common properties of biological molecules are always complicated by wide variations in terminology. The use of a common vocabulary will greatly facilitate the identification of relationships and common properties between biological molecules from different species. The Gene Ontology (GO) approach addresses such a demand by developing a structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner (Ashburner *et al.* 2000). Usually, a gene product can have one or more molecular functions and be used in one or more biological processes, while associated with one or more cellular components. Therefore, GO terms are organized in structures called directed acyclic graphs (DAGs), in which a child term (which is more specialized) can have several parent terms (which are typically less specialized).

GOA@EBI (GO Annotation@EBI) (Camon *et al*. 2004) is a project run by the European Bioinformatics Institute (EBI) that aims to provide assignments of gene products to the Gene Ontology (GO) resource. In the GOA project, GO terms are applied to all proteins described in the UniProt (Swiss-Prot/TrEMBL) (WU *et al*. 2006) and Ensembl databases (Birney *et al*. 2004) that collectively provide complete proteomes. GOA also provides annotations for all entries in PDB database (PDB-GOA).

To assist the analysis of function variation of SDFs, GO terms were downloaded from the Gene Ontology Consortium (Ashburner *et al.* 2000). The GO annotations for each PDB chain were retrieved from PDB-GOA project under GOA@EBI <http://www.ebi.ac.uk/GOA> (Camon *et al*. 2004). So the combination of GO annotation (in form of GO term ID) and GO term definition provides information on molecular function, biological processes and cellular components of each PDB chain in standard vocabularies.

2.1.5 Software packages used during the curation of SDFD

1. PostgreSQL. PostgreSQL <http://www.postgresql.org> is currently the most advanced open source relational database system. It is renowned for its reliability, data integrity and correctness. It supports a large part of the SQL (Structural Query Language, a standard computer language for accessing and manipulating databases) standard and offers many modern features, such as complex queries, foreign keys, views and transactional integrity. PostgreSQL is used as the relational database system in this study.
2. Python and Biopython. Python is an interpreted, interactive, object-oriented programming language. Python combines remarkable power with very clear syntax. It has modules, classes, exceptions, very high-level data types and dynamic typing. Python scripts are portable across almost all platforms, including all major Unix systems, Linux, Windows and Mac OS. All these features make Python an ideal language for bioinformatics tasks. The Biopython Project <http://www.biopython.org> is an international association of developers of freely available Python tools for computational molecular biology. All the scripts in this study are written in Python with the facilitation of Biopython, especially the PDB module (Hamelryck and Manderick 2003).

2.1.6 Database schema

SDFD features were organized into 7 entities (Figure 4): (1) Structure; (2) Protein_chain; (3) Domain; (4) Disulphide; (5) Pro_chain_segment; (6) Residue and (7) GO.

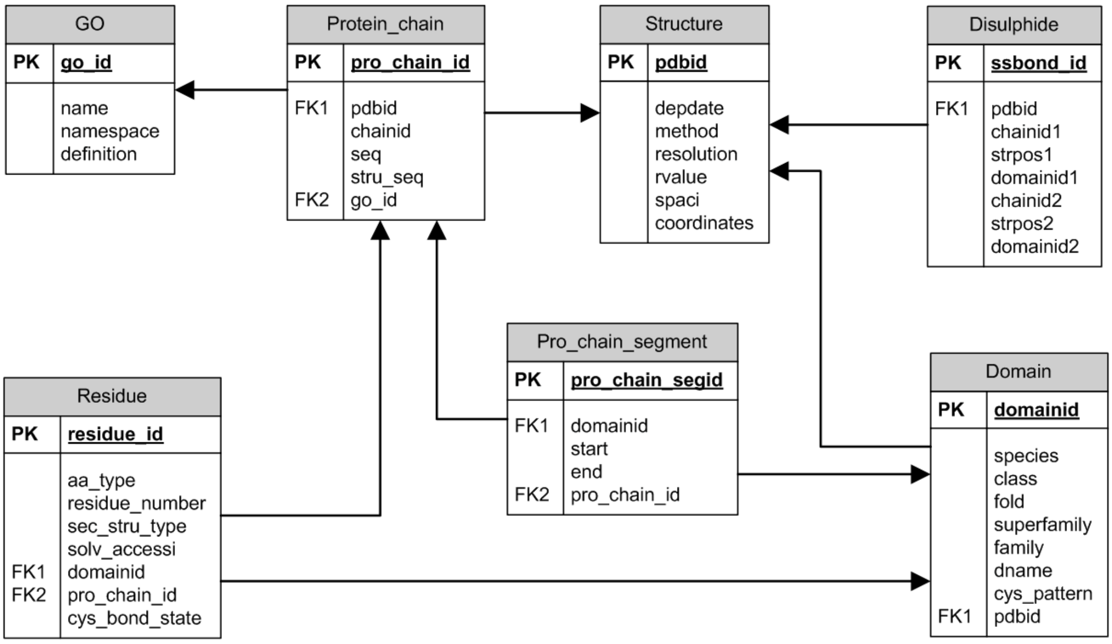


Figure 4 Schematic entity relationship of SDFD. PK represents the primary key for each entity and FK stands for foreign key that connects different entities, establishing the links between them.

SDFD is implemented on top of the open source database system PostgreSQL. It integrates all data from the primary data sources as shown in Figure 4. The data from the original sources are available in different formats, such as flat files, database dump files, or HTML pages. Parsers were written in Python and Biopython to populate SDFD with the data obtained in non-relational representation.

2.2 Classification of SDFs

SDFs are highly redundant and variable and in order to systematically classify them, we propose a hierarchical classification scheme (Appendices: Poster 2), inspired by the SCOP classification but based on the disulphide bond number and disulphide connectivity. In order to compare and classify disulphide bond connectivity, the specific cysteine residues involved in disulphide bond formation are extracted and the links between them are numbered sequentially. For example, in an SDF with four cysteines forming two disulphide bonds, the connectivity may be described as 1221, where the first disulphide link, represented by the number "1" is between the first and fourth cysteine residues and the second disulphide bond (labeled "2") describes the link between the second and third cysteines. If the connectivity is 1212, then the first cysteine is connected to the third and the second to the fourth. Similarly, for proteins with six cysteines, for instance, one of the 15 possible disulphide connectivities where three disulphide bridges are formed, the first (1) between the first cysteine (1) and the third cysteine (1), the second (2) between the second (2) and the fifth cysteine (2), and the third (3) between the fourth (3) and the sixth (3), would thus be ordered sequentially and abbreviated as "121323". Such a notation provides an easy way to discriminate between different disulphide connectivities for both human inspection and machine calculation and comparison.

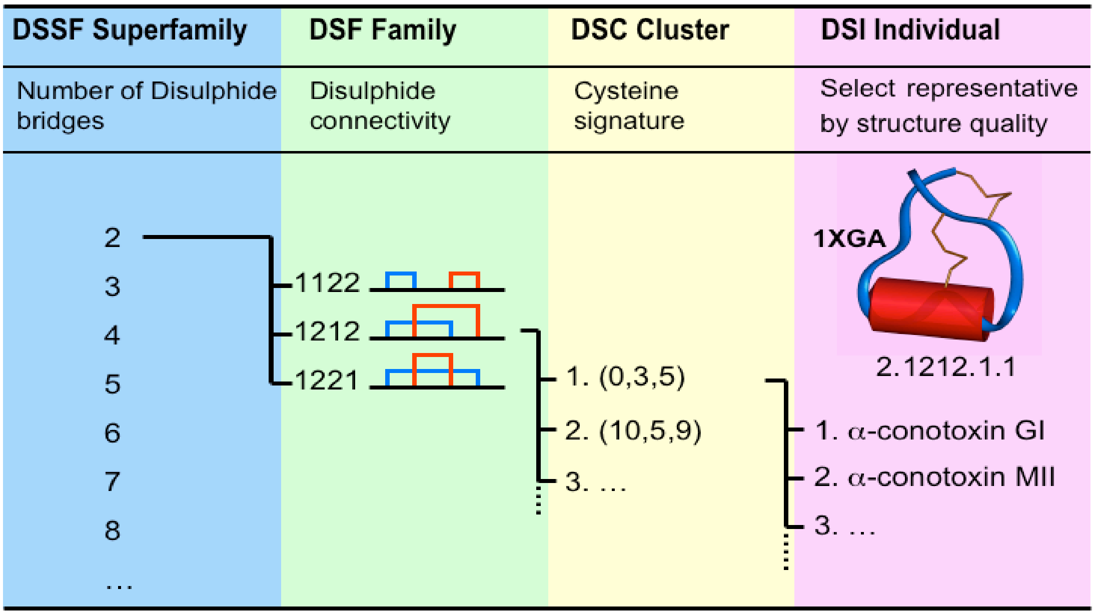


Figure 5 The classification hierarchy of SDFD. The top level is the superfamily, followed by the family, cluster and then the individual domains.

Figure 5 shows the SDFD numbering scheme for the representative structure α-conotoxin GI (PDB code: 1XGA) that has two disulphide bonds and 1212 connectivity. All the SDFs in the database were classified into four levels: Disulphide Superfamily (DSSF), Disulphide Family (DSF), Disulphide Cluster (DSC) and Disulphide Individual (DSI). These levels are described in detail as follows:

(1) DSSF (Disulphide Superfamily). DSSF is the highest level in the classification hierarchy and depends on the number of disulphide bridges in the domain. For example, α-conotoxin GI (PDB ID: 1XGA) has two disulphide bridges in its structures and therefore is classified into DSSF 2 superfamily.

(2) DSF (Disulphide Family). Each DSSF can be classified into disulphide families (DSF) according to the disulphide connectivity. In case of DSSF II, there are three possible disulphide connectivities: 1122, 1212 and 1221. α-conotoxin GI has a disulphide connectivity (C2-C7, C3-C13, where 2, 3, 7 and 13 respectively are the residue numbers of the cysteines), therefore belongs to DSF 2.1212.

(3) DSC (Disulphide Cluster). The domains in each DSF are further grouped into clusters by the "cysteine signature". In this study, the cysteine signature is represented as a vector of sequential distance between cysteines participating in the disulphide bonds. For example, the cysteine pattern occurring along the primary amino acid sequence of an SDP can be written as CX3CX9CX2C, where C is a cysteine residue involved in a disulphide bridge and X is any other non-disulphide bridge residue, and a number following X indicates the number of consecutive X residues

between two cysteine residues. This pattern can be represented as a vector (3,9,2). For example, for two 2-disulphide proteins, their cysteine signatures can be (x_1, y_1, z_1) and (x_2, y_2, z_2), respectively. And the pairwise distance (*d*) between the two cysteine signatures can be calculated by the following formula (Equation 1):

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \quad \text{(Equation 2)}$$

Likewise, for *n*-disulphide proteins, the cysteine signatures can be represented as *2n-1* dimensional vectors and their pairwise distances can be calculated in a similar way.

The pairwise distance (*d*) can be further normalized by the number of separation (*s=2n-1*) to obtain a normalized pairwise distance d_s=*d*/*s*=*d*/(*2n*-1). In this study, the members of each Disulphide Family were clustered into Disulphide Cluster according to an empirical threshould (ds ≤ 1.0). The clusters are numbered consecutively from 1. For each cluster, a representative domain is selected according to the structure quality. α-conotoxin GI is selected as the representative domain for DSC 2.1212.1.

(4) DSI (Disulphide Individual). The domains in each DSC are numbered consecutively from 1. So every domain in SDFD has a unique classification identifier. In the case of α-conotoxin GI, its classification identifier is 2.1212.1.1.

2.3 Data analysis of SDFD

2.3.1 Database content of SDFD

SDFD incorporates data retrieved and carefully extracted from PDB, SCOP, CATH, ASTRAL, GO and GOA. The data were organized into seven entities as shown in Figure 4. SDFD was further classified into hierarchal levels according to their disulphide numbers, disulphide connectivity and sequence similarities.

Table 4 The current content of SDFD database

Entities	Number
Structures	849
PDB chains	999
Domain	1,035
Disulphide bridges	3,307
Residues	58,054

Table 4 shows the current content of non-redundant SDFD database (as of March 2006). Currently SDFD contains 999 PDB chains and 1,044 domains from 849 PDB structures. More than 81% of PDB chains only contain a single domain, which suggests that most SDPs are single-domain proteins. This also indicates that domain is the functional unit for most SDPs and interactions or cooperations between multiple domains are rare in SDPs. Therefore, in Chapter 3, structural modeling of SDPs, there is very few demand for model building for multi-domain SDPs and SDPMOD didn't include a step to predict domain boundary for input sequences before the modeling. This will be further described in Chapter 3.

2.3.2 SDF distribution in SCOP classes

SCOP contains seven major classes: (1) All α proteins; (2) All β proteins; (3) α and β proteins (α/β); (4) α and β proteins (α+β); (5) Multi-domain proteins (α and β); (6) Membrane and cell surface proteins and peptides; (7) Small proteins. The distribution of SDFD entries in each SCOP classes was tabulated in Table 5.

Table 5 The distribution of SDFs among SCOP classes

SCOP Classes	Number of entries
All α proteins	34
All β proteins	44
α/β proteins	0
(α+β) proteins	90
Multi-domain proteins (α and β)	0
Membrane and cell surface proteins and peptides	0
Small proteins	843
Peptides	24
Total	1,035

The majority of SDFs (> 80%) belong to the Small Proteins Class of SCOP, which is not unexpected as this class predominantly comprises proteins that are usually dominated by metal ligand, heme, and/or disulphide bridges. In this SCOP class, the fold family labeled Knottins (small inhibitors, toxins, lectins) has the largest collection (353 entries) of SDFs. A number of SDFs are also present in the all α, all β and (α+β) protein classes. Significantly, there are no examples of SDFs among class α/β proteins, class Multi-domain proteins and class Membrane and cell surface proteins and peptides. This should be due to the small sizes and less secondary structures of SDFs.

2.3.3 SDF Distribution among SDFD superfamilies and families

All the SDF domains were collected into DSSF superfamilies according to the number of disulphide bonds. In current version of SDFD, disulphide bond number ranges from two to eight. For example, the antifreeze protein from beetle (PDB code: 1EZG, Chain B and SCOP ID: d1ezgb_) has eight disulphides within a chain of 84 residues. The distribution of domains amongst DSSF superfamilies and families is presented in Table 6.

Table 6 The distribution of entries among SDFD superfamilies and families. The most populous DSF family in each DSSF Superfamily is highlighted in bold font.

DSSF Superfamily	Number	DSF family	Number
DSSF 2	260	1122	30
		1212	**184**
		1221	46
DSSF 3	467	112233	8
		112323	11
		112332	1
		121233	73
		121323	63
		122133	1
		122313	2
		123123	**128**
		123132	26
		123213	60
		123231	84
		123312	2
		123321	18
DSSF 4	149	11223344	1
		11223443	2
		11232344	5
		11234234	1
		11234432	1
		12123344	31
		12123434	1
		12134234	3
		12231434	3
		12234134	4
		12312344	18
		12312434	1
		12312443	6
		12314234	7
		12314342	1
		12314432	1
		12321344	2
		12324134	5
		12331244	1
		12332144	2
		12341234	4
		12341342	2
		12342314	5
		12342341	**33**
		12343124	5
		12344321	4

DSSF Superfamily	Number	DSF family	Number
DSSF 5	51	1212334554	2
		1212344355	7
		1212345345	7
		1213324455	**8**
		1213434525	1
		1223134455	7
		1231245345	1
		1231425534	1
		1231452345	6
		1232145435	1
		1233245451	2
		1234134525	1
		1234215534	1
		1234253451	1
		1234321554	5
DSSF 6	7	121233445656	1
		123214543656	**5**
		121234535646	1
DSSF 7	14	12344256577631	7
		12324431565776	**6**
		12123344565677	1
DSSF 8	2	**1213234455667788**	**2**

Table 6 clearly shows the uneven distribution of SDFD entries among superfamilies and families. DSSFs 2 and 3 are most abundant superfamilies in SDFD, contributing 25% and 45% of the whole dataset, respectively. Only those families with structural examples in the PDB have been listed, although combinatorially, a large number of families are possible.

For proteins with n-disulphide bonds, the number of possible connectivity patterns can be calculated by the follow formula (Fariselli and Casadio, 2001):

$C_p = (2n - 1)!! = \Pi_{(i \leq n)} (2i - 1)$ (Equation 3)

In Equation 1, C_p represents the number of possible disulphide connectivities, n stands for the number of disulphide bridges in the protein. So theoretically the possible disulphide connectivities for each DSSF superfamily can be calculated for n

= 2,3,…,8 (Table 7).

Table 7 The theoretic number and observed number of disulphide connectivity for each disulphide superfamily (DSSF).

DSSF	Disulphide number (*n*)	Theoretical number	Observed number
DSSF 2	2	3×1 = 3	3
DSSF 3	3	5×3×1 = 15	13
DSSF 4	4	7×5×3×1 = 105	26
DSSF 5	5	9×7×5×3×1 = 945	15
DSSF 6	6	11×9×7×5×3×1 = 10,395	3
DSSF 7	7	13×11×9×7×5×3×1 = 135,135	3
DSSF 8	8	15×13×11×9×7×5×3×1 = 2,027,025	1

Table 7 shows the enormous difference between theoretical and observed number of disulphide connectivities in each DSSF. Although the number of possible connectivities for DSSF superfamily ($n \geq 4$) is huge, only a small proportion is observed in SDFD. Such gaps can be explained by the following reasons:

(1) Not every kind of disulphide connectivity is possible in nature. Obviously, observed protein sequences only account for a tiny portion of possible sequence space. Given the protein sequence, the disulphide connectivity will be restrained steric factors since cysteines close enough can possibly form disulphide bridges.

(2) The observed disulphide-rich structures are only a small fraction of known protein sequences. With the rapid development of genome sequencing projects and structural genomics projects, more and more disulphide-rich proteins will be identified and more disulphide connectivities will be found.

(3) Nature displays preferences for some disulphide connectivities over others. The observed number in each DSF clearly supports this tendency. For DSSF 2, there are three possible disulphide connectivities. According to the nomenclature

proposed for the arrangement of disulphide bridges (Harrison and Sternberg 1994), three relationships can occur between two disulphide bridges (Figure 6): (1) independence (DSF 1122); (2) overlap (DSF 1212); or (3) enclosure (DSF 1221).

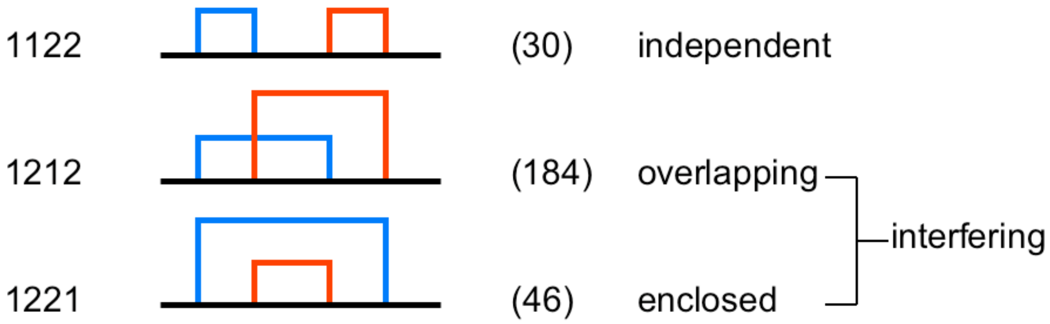


Figure 6 Three relationships between two disulphide bridges as described by Harrison and Sternberg 1994. Beside each connectivity diagram the number observed in SDFD is given. Note that this terminology does not take into consideration the 3D structure of the protein and simply describes the relationship between disulphide bridges at the level of the primary sequence. In a structural study such as this, in a number of instances, such a description may be a misnomer, e.g. a sequentially “overlapping” set of disulphide bridges do not necessarily have “overlaps” structurally. However, they have the utility of being concise and are used in this thesis on that basis.

Clearly the overlapping topology (DSF 1212) has the largest observed population (over 70%), which suggests the preference of the overlapping topology over independent and enclosed topologies. Similarly in DSSF 3, the overlapping topology of DSF family 123123 has the biggest number among 15 possible disulphide connectivities. A possible explanation for such this preference is that the overlapping topology of disulphide connectivities will help to anchor constituent protein fragments and improve the intra-domain interactions, thereby making the protein thermodynamically more stable.

2.3.4 Disulphide distance distribution

Disulphide distance is defined as the sequential distance between the two bonded cysteines of a disulphide bridge, measured as the difference between their residue numbers. Tsai and coworkers reported that a descriptor derived from the disulphide distance could improve the accuracy of disulphide connectivity predictions (Tsai *et al.* 2005; Zhao *et al.* 2005). Therefore, the distribution of disulphide distance may provide useful information for predictioning disulphide connectivity of newly sequenced SDPs. Figure 7 shows the distribution of the distances between two bonded cysteines.

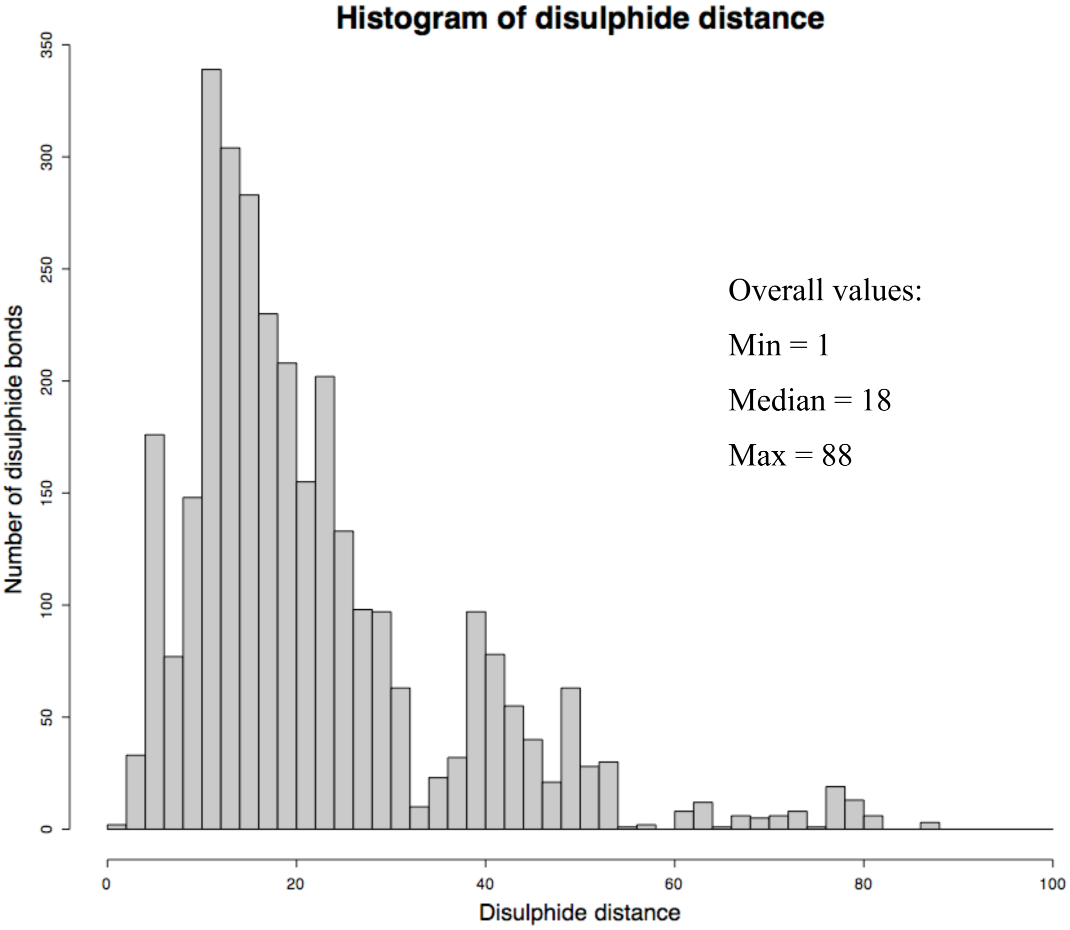


Figure 7 The distribution of disulphide distance in SDFD. The unit for disulphide distance is residues.

The maximum disulphide distance observed in SDFD is 88 residues, for the disulphide bond formed between C3A and C91A in tomato serine proteinase inhibitor II (TI-II, PDB ID: 1PJU, Chain A). The minimal disulphide distance observed in SDFD is 1, with only one occurrence, belonging to a disulphide formed between adjacent cysteines, C13A and C14A, in the insecticidal neurotoxin, J-ACTXs (PDB ID: 1DL0, Chain A). Such vicinal disulphide bridges are rare and they may have special functional roles. In the case of 1DL0, the vicinal disulphide bridge is critical for insecticidal activity of J-ACTXs (Wang *et al.* 2000).

Figure 7 shows that the distribution of disulphide distance is double-humped with two maxima at 18 (the main peak) and 40 (the secondary peak), respectively. The frequency for short disulphide distance (less than 4) and long distance (greater than 54) is very low, while the intermediate distance of 32-38 is also not preferred. This distribution should be useful for disulphide connectivity prediction programs which could use this parameter for screening out false positives.

2.3.5 Inter-domain *vs.* intra-domain disulphide bridges

The disulphide bridges in protein structures can be classified into inter-domain or intra-domain disulphide bonds, based on whether the two bonded cysteine residues belong to the same domain or not. Most of the disulphide bridges in SDFD are intra-domain disulphides. Among the 3,307 disulphide bridges in SDFD, only 93 of them connect two different domains defined by SCOP. But if the CATH domain definition is used instead, 68 of these 93 inter-domain disulphides belong to the same domain. Detailed analysis suggested that the domain boundaries for some SCOP domains might be stringently assigned, causing bonded cysteine residues to be assigned to different domains. For example, wheat germ agglutinin (PDB ID: 9WGA, Chain A) contains an inter-domain disulphide bond (C46A-C61A) linking domain d9wgaa1 and d9wgaa2 (according to SCOP domain definitions, Figure 8A). But this disulphide bond becomes an intra-domain disulphide bond in domain 9wgaA2 (based on CATH domain definition, Figure 8B).

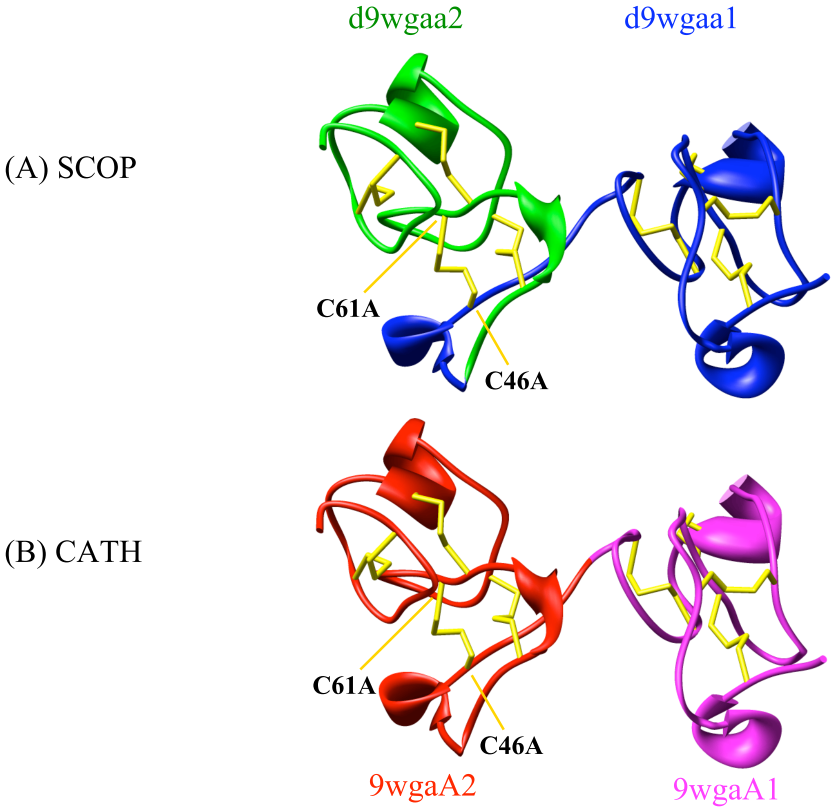
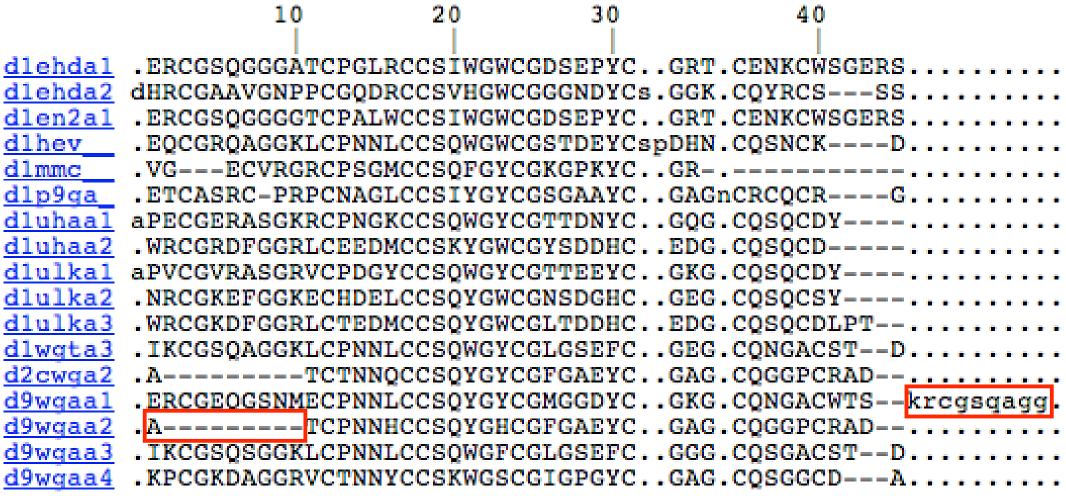


Figure 8 The comparison of SCOP and CATH domain boundaries of wheat germ agglutinin (PDB ID: 9WGA, Chain A: 1-86). (A) SCOP domain boundaries for 9WGA, domain d9wgaa1 (blue): 1A-52A, domain d9wgaa2 (green): 53A-86A; (B) CATH domain boundaries for 9WGA, domain 9wgaA1 (magenta): 1A-42A,domain 9wgaA2 (red): 43A-86A. The structures are in ribbon representation and disulphide bridges are shown in stick representation, colored in yellow. Two cysteine residues,

46A and 61A, forming the disulphide bond analyzed, are labeled.

From Figure 8, the CATH domain boundary definition is more reasonable than the SCOP definition in the case of 9WGA from the viewpoint of structure. Another evidence (Figure 9) for such misclassification is obtained from the domain sequence alignment provided by the Superfamily server (Gough *et al*. 2001).



```
                   10        20        30            40
                    |         |         |             |
d1ehda1  .ERCGSQGGGATCPGLRCCSIWGWCGDSEPYC..GRT.CENKCWSGERS..........
d1ehda2  dHRCGAAVGNPPCGQDRCCSVHGWCGGGNDYCs.GGK.CQYRCS---SS..........
d1en2a1  .ERCGSQGGGGTCPALWCCSIWGWCGDSEPYC..GRT.CENKCWSGERS..........
d1hev__  .EQCGRQAGGKLCPNNLCCSQWGWCGSTDEYCspDHN.CQSNCK----D..........
d1mmc__  .VG---ECVRGRCPSGMCCSQFGYCGKGPKYC..GR-.-----------..........
d1p9ga_  .ETCASRC-PRPCNAGLCCSIYGYCGSGAAYC..GAGnCRCQCR----G..........
d1uhaa1  aPECGERASGKRCPNGKCCSQWGYCGTTDNYC..GQG.CQSQCDY----..........
d1uhaa2  .WRCGRDFGGRLCEEDMCCSKYGWCGYSDDHC..EDG.CQSQCD-----..........
d1ulka1  aPVCGVRASGRVCPDGYCCSQWGYCGTTEEYC..GKG.CQSQCDY----..........
d1ulka2  .NRCGKEFGGKECHDELCCSQYGWCGNSDGHC..GEG.CQSQCSY----..........
d1ulka3  .WRCGKDFGGRLCTEDMCCSQYGWCGLTDDHC..EDG.CQSQCDLPT--..........
d1wgta3  .IKCGSQAGGKLCPNNLCCSQWGYCGLGSEFC..GEG.CQNGACST--D..........
d2cwga2  .A---------TCTNNQCCSQYGYCGFGAEYC..GAG.CQGGPCRAD--..........
d9wgaa1  .ERCGEQGSNMECPNNLCCSQYGYCGMGGDYC..GKG.CQNGACWTS--krcgsqagg.
d9wgaa2  .A---------TCPNNHCCSQYGHCGFGAEYC..GAG.CQGGPCRAD--..........
d9wgaa3  .IKCGSQSGGKLCPNNLCCSQWGFCGLGSEFC..GGG.CQSGACST--D..........
d9wgaa4  .KPCGKDAGGRVCTNNYCCSKWGSCGIGPGYC..GAG.CQSGGCD---A..........
```

Figure 9 The multiple sequences alignment of SCOP superfamily plant lectin by Superfamily. The regions marked by rectangles delineate the incorrect domain boundary between domains d9wgaa1 and d9wgaa2.

Figure 9 clearly shows that the unaligned sequence segment at the end of domain d9wgaa1 should go to the N-terminus of domain d9wgaa2, correctly picked up by the CATH domain definition. Therefore, the inter-domain disulphide bond 46A-61A is actually an intra-domain disulphide bond. Similarly, the other 67 of the 93 inter-domain disulphides have been reclassified as intra-domain disulphide bonds, based on the CATH domain definitions for these domains.

Only 25 disulphide bonds were classified as inter-domain disulphide bonds by both SCOP and CATH domain definitions. These inter-domain disulphide bonds help to fix the relative position of functional domains. For example, the inter-domain disulphide bond (C122C-C141L) in Gla-domainless activated protein C (PDB ID: 1AUT, Chain C, Chain L) linked to the light chain, is the catalytic domain (Mather *et al*. 1996).

The low frequency (25/3,307=0.7%) of inter-domain disulphide bond is not surprising since a domain tends to be a compact, independent unit of protein structure. Inter-domain disulphide bonds anchoring the relative positions between domains are thus uncommon in small proteins, such as SDPs.

This observation has potential application in domain boundary delineation, disulphide connectivity prediction and molecular modeling. When determining the domain boundary, the two cysteines forming a disulphide bond are more likely to be in the same domain unless there is evidence that the role of that disulphide is to help stabilizing the relative position of two domains. The same rule is also applicable to disulphide connectivity prediction. Before the prediction of disulphide connectivity, the sequence should be split into domains and disulphide connectivity predicted for each domain. This will greatly reduce the prediction search space of possible connectivities and improve the prediction accuracy.

2.3.6 Inter-chain disulphide *vs*. intra-chain disulphide bridges

Among 3307 disulphide bonds in SDFD, 148 of them are inter-chain disulphide bridges, which is only a small fraction (4.4%) of the total dataset. These 148 inter-chain disulphides can be further classified into 52 inter-domain and 96 intra-domain disulphides. According to genetic domain definitions from ASTRAL, all 96 inter-chain intra-domain disulphides belong to genetic domains, so that the multiple protein chains are actually the product of a single gene. In other words, these inter-chain disulphide bridges are actually intra-chain disulphide bridges before the protein precursors were processed into mature proteins. Insulin is the best example of such inter-chain disulphides. Insulin is derived from a single-chained precursor, proinsulin, by the removal of a segment from the middle of the precursor protein. The active hormone, insulin, is composed of two protein chains and contains two inter-chain and one intra-chain disulphide bridges. Fully denatured insulin cannot recover its native structure and disulphide bridges (Anfinsen 1973), which suggests that the complete sequence information is essential for the folding of this protein and the formation correct disulphide bonds. This result provides an indication that for disulphide connectivity prediction, the precursor sequence information may be a better descriptor and should be more informative than the mature sequence alone.

Of the 52 inter-chain inter-domain disulphide bonds, 24 of them belong to the SCOP fold of cysteine-knot cytokines. Detailed analysis shows that these inter-chain disulphide bridges link two identical sequences (monomers) derived from the same gene.

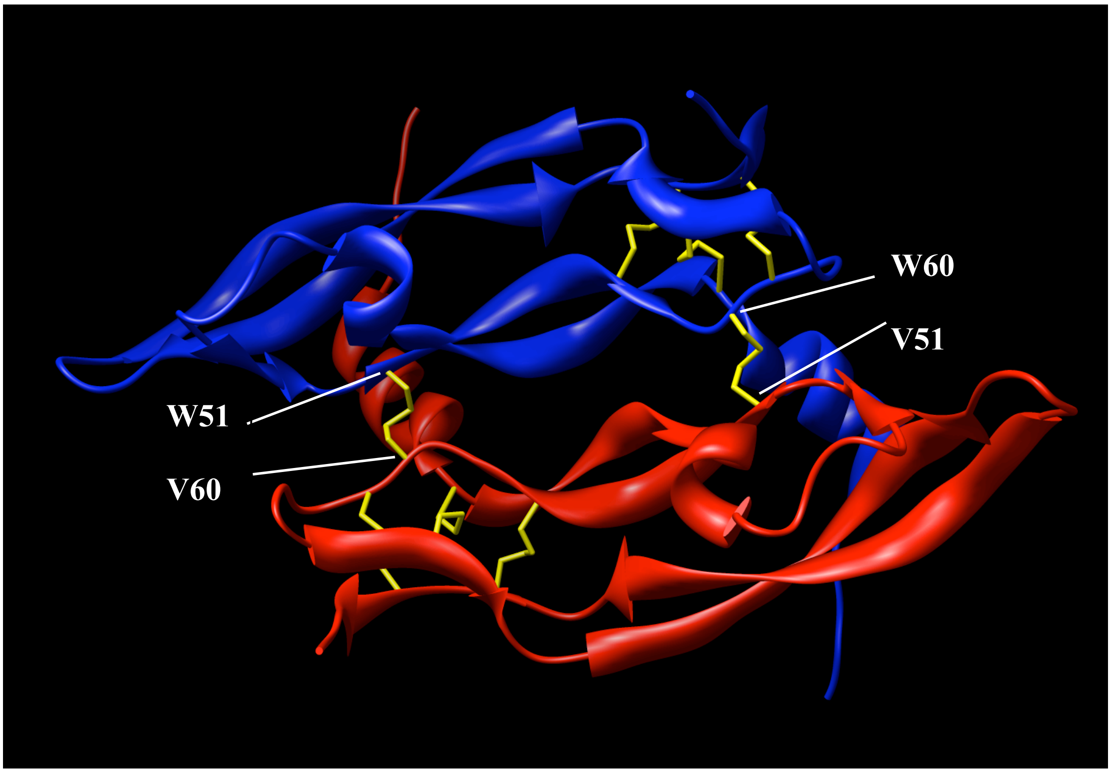


Figure 10 Inter-chain inter-domain disulphide bonds in the structure of Vascular Endothelial Growth Factor (PDB ID: 1KAT). Chain V (color in red) forms one domain (SCOP ID: d1katv_) and chain W (color in blue) forms another domain (SCOP ID: d1katw_). The structure was rendered in ribbon represenation and the disulphide bridges are shown in stick and colored in yellow.

The aboving figure (Figure 10) shows an example of inter-chain inter-domain disulphide bonds in the structure of Vascular Endothelial Growth Factor (PDB ID: 1KAT) from SCOP fold of cysteine-knot cytokines. From Figure 10, we can see that chain V and chain W are two identical monomer structures and each chain forms one domain. The two chains were linked together by two inter-chain inter-domain disulphide bonds (between Cys60 in chain V and Cys51 in chain W, Cys51 in chain V and Cys60 in chain W, respectively). The structure is symmetrical and chain V and chain W have identical sequences.

Overall, among all inter-chain disulphides identified in SDFD, no example connects chains from different genes, indicating that such “inter-gene” disulphide bonding is difficult to form in nature.

2.3.7 The cysteine signature for the detection of structural similarity

Our results show that cysteine signatures can facilitate the detection of structural similarity, at a finer level than the DSF classification, and can be used for grouping structures as redundant or non-redundant folds. Figure 11 shows an example of two highly similar structures identified by cysteine signature clustering.

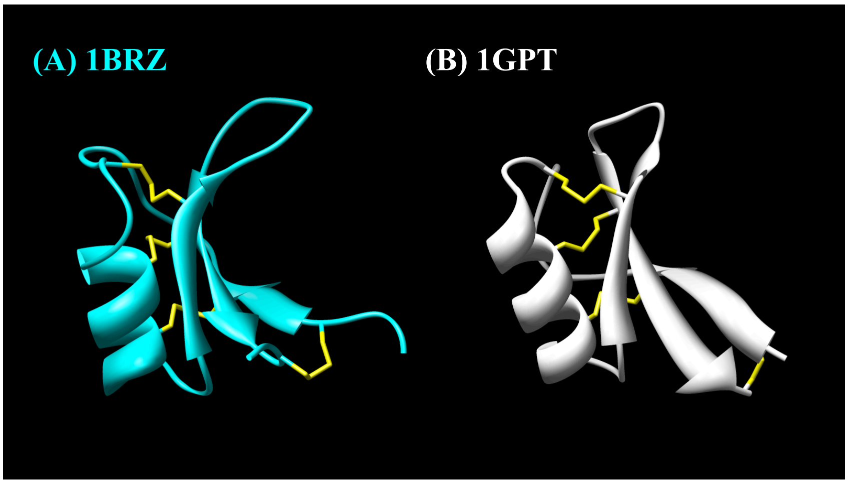


Figure 11 The structure comparison between sweet-tasting protein brazzei (PDB ID: 1BRZ) and plant toxin γ 1-hordothionin (PDB ID: 1GPT) (A) 1BRZ, colored in cyan; (B) 1GPT, in grey. Both structures are in ribbon representation. Disulphide bonds are represented in stick and colored in yellow.

The cysteine signature of a sweet-tasting protein brazzei from fruits of *Pentadiplandra brazzeana* (PDB ID: 1BRZ) is (11,5,3,10,9,1,2), is very close to the signature (10,5,3,9,6,1,3) from a plant toxin, γ 1-hordothionin from barley (PDB ID: 1GPT). The two proteins have been clustered into a single group in SDFD, although there is no functional relationship between these two proteins and the sequence identity between them is only 14.8%. Structure comparison of these two proteins (Figure 11) shows their structures are highly similar (RMSD 1.2Å) and they have the same disulphide connectivity. This suggested that the cysteine signature might be a useful feature to detect distantly related homologs or convergently evolved proteins especially when structural information is unavailable.

2.4 Conclusion

Chapter 2 described the database curation, classification and data analysis of the small disulphide fold database (SDFD). The principal findings are as follow:

(1) SDFD is a curated database, containing 1,044 non-redundant SDF domains. These domains have been used collectively as the template repository for comparative modeling of SDPs, described in Chapter 3.

(2) A hierarchical classification scheme for systematically sorting and grouping SDF domains has been developed. The SDFD database is classified into four hierarchical levels according to disulphide bond number, disulphide connectivity and cysteine signature. The hierarchical classification method is able to detect the structural similarities of proteins of low sequence identity.

(3) The distribution of SDFs among DSSF superfamily and DSF family suggested a preference of disulphide connectivity on overlapped topology.

(4) The analysis of intra- and inter-domain disulphide bonds identified some mis-assigned domain boundaries by SCOP. The low frequency of inter-domain disulphide in SDPs suggests that prior to disulphide connectivity prediction and molecular modeling, the protein sequence should be split into domains to improve the accuracy of disulphide connectivity prediction and reduce computational time.

(5) The analysis of intra and inter-chain disulphide bonds showed the most inter-chain disulphides are actually intra-chain disulphides, which connect fragments that originally belong to the same gene before the processing of protein precursors.

Chapter 3 Structural modeling of SDPs

3.1 Introduction

This chapter focuses on the comparative modeling of SDPs and contains two parts: the first part describes the design, modeling procedure and benchmarking of SDPMOD and implementation of SDPMOD as a web server; the second part will illustrate a high throughput comparative modeling technique based on conotoxins, a large species-specific SDP family, using SDPMOD with the topology and parameter libraries that we have developed for constructing proteins with non-standard residues.

3.2 The automated comparative modeling method for SDPs - SDPMOD

3.2.1 Curation of template repository

Before commencing the comparative modeling of SDPs, a non-redundant template repository needs to be created. The SDFD database served as such repository after redundancy is removed at two levels as follows.

Firstly, most structures determined by the NMR method contain an ensemble of monomer models. These structures represent models that fit all restraints determined from the NMR experiment. During comparative modeling, only one monomer needs to be used as a template, as the structural information available from the different NMR structures is redundant (Marti-Renom *et al.* 2000). Different researchers use different strategies to treat the NMR structure ensemble. While some groups simply use the first monomer as the representative structure, others utilize the mean structure by averaging atom positions of all monomers and minimizing the energy. But the mean structure results in a fictitious molecule with some abnormal bond lengths and bond angles. In this study, NMRCLUST (Kelley *et al.* 1996) is used to the select representative monomer from NMR structures. NMRCLUST can cluster monomers into groups and select the monomer that is closest to other monomers as the representative monomer.

Secondly, when multiple structures exist for the same sequence, the representative structure is chosen according to its structural qualities. The structural qualities are ranked by the following criteria (adopted from PDB): (1) X-ray structures over NMR structures, (2) higher quality factor (1/resolution - R-value) for X-ray structures and higher restraint per residue for NMR, (3) better geometry, (4) fewer missing atoms and non-standard residues and (5) later deposition date. Based on the above strategy, a non-redundant template structure dataset for SDPs was generated and loaded into a PostgreSQL database. Currently it contains more than 1,000 non-redundant SDF domain and their coordinates.

3.2.2 The Modeling procedure

SDPMOD is designed specifically for the comparative modeling of SDPs. The special features of SDPs are considered and incorporated into the model building method. Traditional comparative modeling methods usually contain four steps: (1) template selection; (2) target-template sequence alignment; (3) model building; (4) model evaluation (Marti-Renom *et al.* 2000). Among these four steps, the first two (template selection and target-template) are the most crucial steps for the quality of comparative modeling. The major problems within the comparative modeling of SDPs are also confined to these two steps, as explained in the introductory part of this chapter.

Available comparative modeling servers and methods have problems in identifying the best templates for SDP sequences and in aligning the SDP sequence and template sequences correctly, as they are parameterized and benchmarked for larger globular proteins. SDPMOD is designed with special consideration to the template selection and target-template alignment steps, according to the features of SDPs, each step of which is presented in the following sections.

3.2.2.1 Template selection

Template selection is the most crucial step in comparative modeling. Comparative modeling is based on the assumption that two structures would be similar to each other if their sequences are homologous (Sander and Schneider 1991). Template selection usually begins with sequence searching programs, such as BLASTP (Altschul *et al.* 1990) or PSI-BLAST (Altschul *et al.* 1997) against protein sequences from PDB (Berman *et al.* 2000). In most cases, multiple hits will be retrieved and the best template(s) need not necessarily be the top ones.

Traditionally the best template(s) are selected according the following criteria in the order of decreasing importance:

(1) **Similarity of domain.** The template(s) are expected to belong to the same fold as the target sequence. For the modeling of multi-domain proteins, domain boundary prediction methods, such as domain-fishing (Contreras-Moreira and Bates 2002), should be used first to split the target sequence into domains before searching unless the structural database search can identify highly similar (ID% > 70%) templates with almost full coverage of the target sequence;

(2) **Similarity of Sequence.** The template with higher sequence similarity and fewer gaps is preferred as the structural model generated from this is expected to be

closer to the real structure of the target sequence;

(3) **Better structural quality**. If there are multiple templates available from the PDB with almost the same sequence, structural quality will be an important consideration, as described in Section 3.2.1.

In SDPMOD, the strategies for template selection are different due to the unique features of SDPs.

(1) **Cysteine numbers**. Cysteine numbers defined here refer to the total number of cysteine residues in the sequence. In the case of SDPs, the total number of cysteines tend to be conserved during evolution and can serve as a reliable filter to exclude proteins that apparently belong to different folds so as to reduce the computational time. One possible problem here is that this step may filter out some good templates if there are free cysteines (cysteine which do not form disulphide bond) in the sequence. However, while this may be significant in other proteins, in SDPs, only about 0.8% (53 of 6,667) of cysteine residues are free cysteines. For novel sequences, the bonding states of cysteine residues (free or forming disulphide bridges) can be easily predicted at high accuracy (up to 88%) (Fiser and Simon 2000; Martelli *et al.* 2002). Furthermore, if both the target sequence and the template structure contain free cysteines in similar positions (which is very likely since the highly conserved nature and functional importance of cysteine residues), this would not cause problem when searching for templates. Therefore, cysteine number can still serve as a reliable and safe filter for most modeling jobs. For input sequences with free cysteines, if SDPMOD cannot identify the template automatically, the manual mode to choose the correct template has to be used.

(2) **Cysteine signatures**. Cysteine signatures are sequence motifs that are

composed of cysteines and the distances between cysteines. For small proteins, the spacing between cysteine residues is more important than sequence similarity for template selection. Drakopoulou and co-workers reported that cysteine spacing govern specific disulphide bond formation (Drakopoulou *et al.* 1998). Also, it is difficult to compensate for the distortion of structures when deletion and insertion events occur in small proteins. Fewer gaps have higher priority over sequence similarity, especially if sequence similarity is not significant, as is usually the case among SDPs. Therefore, a higher gap open penalty of 15 was used in the alignment step, rather than the default penalty of 11.

The structure redundancy problem has been solved during the curation of the template repository. The modeling of multi-domain proteins is not a major problem for the modeling of SDPs. Firstly, SDPs are small molecules that are usually single domain proteins or processed into functional single-domain proteins. So there is little or no demand for modeling multi-domain SDPs. Secondly, there is currently no good method available to predict the relative positions of protein domains. This is actually a protein-protein docking problem that is extremely computationally intensive, with multiple solutions. So the modeling of multi-domain proteins is beyond the scope of this study unless there are corresponding multi-domain templates available from PDB.

3.2.2.2 Target-template alignment

After the best template(s) are selected, alignment of the target and template sequences will be the most critical factor to determine the quality of comparative modeling. Although the final alignments for the modeling only contain sequences of the target protein and template(s), it is a good idea to include multiple homologous sequences during the alignment since they can help in the identification of conserved regions and residues. There are a number of programs to do this with the most popular one being CLUSTAL (Higgins and Sharp 1988). Generally the alignment algorithm will try to maximize the aligned regions and minimize the gap regions. But the optimal computational alignment does not always correspond to the optimum biological alignment. Therefore, the alignment generated by a sequence alignment program may not be the best alignment for the modeling purposes. The alignments usually can be improved by manual adjustment (aligning the conserved regions and residues and positioning gaps to loop regions or the ends of secondary structure elements) using freely available software such as Jalview (Clamp *et al*. 2004). Due to the importance of cysteine residues in SDPs, SDPMOD uses a modified scoring matrix, in which the value of cysteine match was doubled, to force the cysteine residues to be well aligned.

3.2.2.3 Model building

Given the template structure(s) and the target-template sequence alignment, comparative modeling programs can generate 3D models. One of the most reputable programs used in comparative model building is MODELLER (Sali and Blundell 1993). Assuming a good choice of template and an optimum target-template sequence alignment, this step is almost automatic and not much human intervention is required. However, researchers can still adjust the level of molecular dynamics (MD) optimization and start with multiple initial models to overcome the local energy minimal.

SDPMOD uses MODELLER to build 3D models based on the template structure and target-template alignment from previous two steps. Multiple models were built using different initial models and the best model is selected according to the least MODELLER objective function score, based on the in-built potential energy function including stereochemical violations.

3.2.2.4 Model evaluation

After 3D models are generated, model evaluation has to be done to check the quality of models. The models are often evaluated in two ways:

(1) **Stereochemical quality evaluation.** The most frequently used program for this purpose is PROCHECK (Laskowski *et al.* 1993). PROCHECK can assess how normal or how unusual, the geometry of each residue in a given protein structure is, as compared with stereochemical parameters derived from well-refined, high-resolution structures. Stereochemically strained regions highlighted by PROCHECK are not necessarily errors, but may correspond to unusual features for which there is a reasonable explanation (e.g. distortions due to ligand-binding at the protein's active site). Nevertheless, they are regions that should be checked manually. SDPMOD utilizes PROCHECK to evaluate the stereochemical quality of models. Generally, an ideal model should have high G-factors scores and most residues cluster in the most favorable regions of the Ramachandran plot. A reasonable model should have overall G-factors greater than –0.5 with less than 5% of the residues in the most unfavorable Ramachandran regions.

(2) **Comparison between template structure and generated 3D models.** If the RMSD (Root Mean Square Deviation) between them is too large (greater than 3Å for most proteins (Schwede *et al.* 2000) and 2 Å for SDPs), the models should be carefully re-examined. It usually indicates that selected template may not belong to the same fold with the target protein or the alignment between template and target sequence is poor. The modeling parameters should be checked and the first two steps, template selection and target-template alignment, should be re-checked. The models

with large RMSD values should only be used subsequently with caution, if at all. SDPMOD calculates the RMSD between template and model using MODELLER as an indicator for modeling reliability.

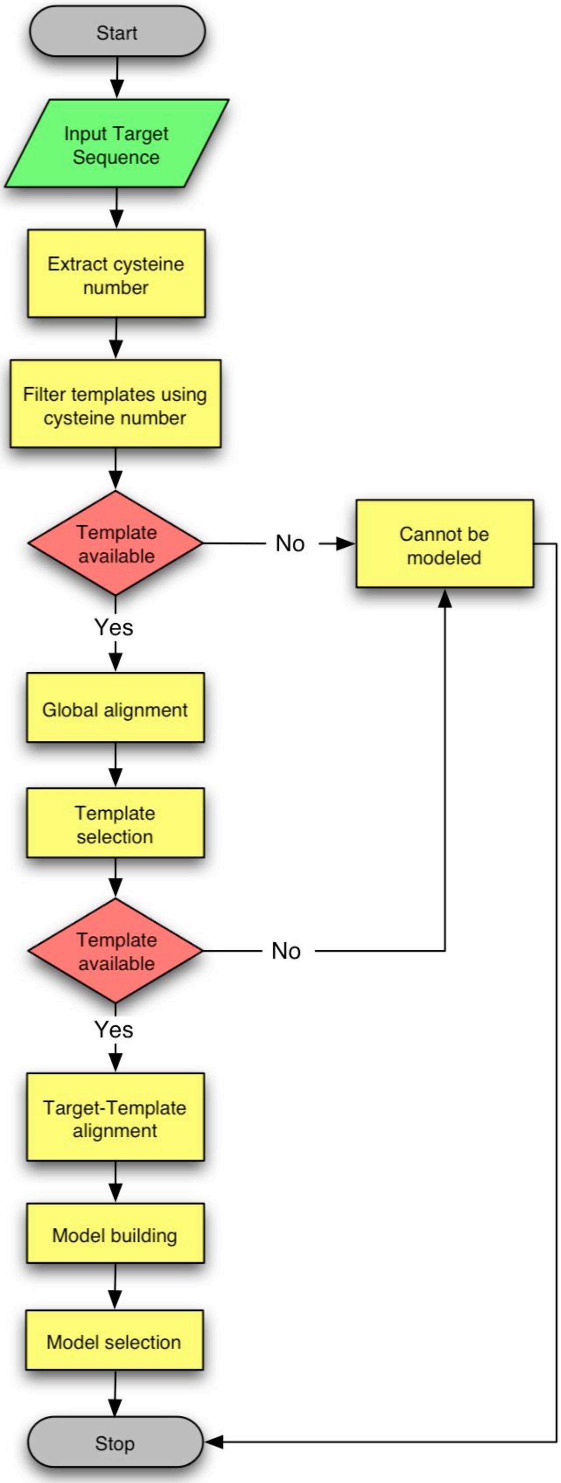


Figure 12 The flowchart of SDPMOD

Figure 12 shows the detailed modeling procedure for automated modeling of SDPs in SPDMOD. The non-redundant SDF dataset is first filtered using the number of cysteine residues, and the resulting template sequences are globally aligned to the target sequence using a modified scoring matrix. The best templates are then selected based on the highest alignment scores. A dynamic minimum threshold for the alignment scores was used in this step because in some cases, the sequence similarities can be very low. For example, the sequence identity between sweet-tasting protein brazzei (PDB ID: 1BRZ) and γ 1-hordothionin (PDB ID: 1GPT) is only 15% but the two structures are highly similar (RMSD 1.2Å, shown in Chapter 2 Figure 11). Also cysteine number and cysteine signature already serve as realiable filters prior to this step. With the selected best template, target-template alignment and model building are achieved by MODELLER (Sali and Blundell 1993), using a customized matrix to ensure that all the cysteine residues are well aligned. The final models are chosen according to the MODELLER objective function score, which reflects lower energy and fewer stereochemical violations. Finally, the overall structural quality of the generated models is evaluated against stereochemical parameters derived from high quality experimental structures using PROCHECK (Laskowski *et al.* 1993).

3.2.3 Benchmarking and Evaluation

A large-scale benchmarking was completed using the fully automated mode of the SDPMOD method. A control set of 664 sequences (a subset of our SDFD non-redundant database) with known structures was used to evaluate the reliability of the method. Prior to the modeling of each sequence, its corresponding PDB structure was removed from the template dataset. The Cα RMSD values between models and their actual experimental structures were calculated. The results are summarized in Table 8.

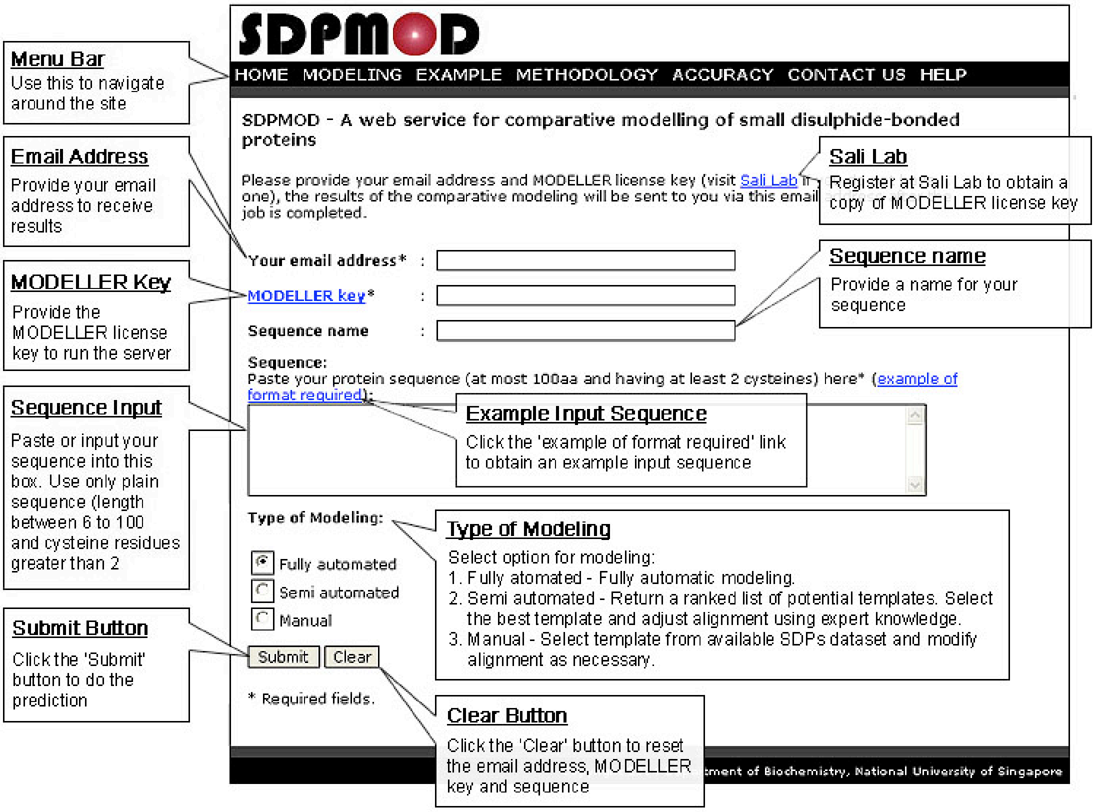
Table 8 SDPMOD results for the benchmarking dataset. D represents the RMSD.

Sequence identity (%)	**Number of Models**					
	Total	**D<0.5Å**	**0.5Å≤D<1Å**	**1Å≤D<1.5Å**	**1.5Å≤D<2Å**	**2Å≤D**
20-30	172	0/0.0%	0/0.0%	23/13.4%	105/61.0%	44/25.6%
30-40	93	0/0.0%	3/3.2%	34/36.6%	46/49.5%	10/10.7%
40-50	56	0/0.0%	5/8.9%	29/51.8%	20/35.8%	2/3.6%
50-60	55	0/0.0%	11/20.0%	24/43.7%	16/29.1%	4/7.2%
60-70	53	0/0.0%	13/5.7%	24/45.3%	15/28.3%	1/1.9%
70-80	54	4/7.4%	12/22.2%	18/33.3%	16/29.6%	4/7.4%
80-90	91	9/9.9%	19/20.9%	32/35.1%	28/30.8%	3/3.3%
90-95	90	13/14.4%	19/21.1%	32/35.5%	23/25.6%	3/3.3%
Total number	664	26/3.9%	82/12.3%	216/32.5%	253/38.1%	71/10.7%

Table 8 shows the SDPMOD results for the benchmarking dataset, based on the target-template sequence identity values. The values of RMSD (based on Cα atoms) between generated model and its template were calculated by MODELLER and were used to evaluate the accuracy of modeling. Generally, the models are considered as reasonable models if the RMSD value is less than 1.5Å. The RMSD values in each sequence identity range are calculated and tabulated. It is clear that the accuracy tends to be better in higher sequence identity ranges and become quite poor if sequence identities between target and template sequences are below 40%. The sequence identity value required here is much greater than 25% set as the threshold (“twilight zone”) for globular proteins (Sander and Schneider 1991). Overall, in the 40-70% sequence identity range, 64% of models have Cα RMSD values less than 1.5Å. The benchmarking results show SDPMOD can predict 3D models with an accuracy comparable to other automated methods (Schwede *et al.* 2000).

3.2.4 The implementation of SDPMOD as a web server

To facilitate the use of the SDPMOD methodology, a web server has been developed, which is freely accessible to academic or non-profit users via a web interface (shown in Figure 13) at <http://proline.bic.nus.edu.sg/sdpmod>. SDPMOD is primarily designed as a fully automated procedure for ease of use. However, due to the complexity of comparative modeling, human intervention and expert knowledge may be required for optimal modeling of some proteins at two critical stages, namely template selection and target-template alignment (Bates *et al.* 2001). To allow for human intervention, the current version of the SDPMOD server provides three modes of access (fully automated, semi-automated and manual) to meet the different needs of the expert users.



SDPMOD - A web service for comparative modelling of small disulphide-bonded proteins

Please provide your email address and MODELLER license key (visit Sali Lab i one), the results of the comparative modeling will be sent to you via this email job is completed.

Your email address* :

MODELLER key* :

Sequence name :

Sequence:
Paste your protein sequence (at most 100aa and having at least 2 cysteines) here* (example of format required):

Type of Modeling:

● Fully automated

○ Semi automated

○ Manual

* Required fields.

Menu Bar
Use this to navigate around the site

Email Address
Provide your email address to receive results

MODELLER Key
Provide the MODELLER license key to run the server

Sequence Input
Paste or input your sequence into this box. Use only plain sequence (length between 6 to 100 and cysteine residues greater than 2

Submit Button
Click the 'Submit' button to do the prediction

Sali Lab
Register at Sali Lab to obtain a copy of MODELLER license key

Sequence name
Provide a name for your sequence

Example Input Sequence
Click the 'example of format required' link to obtain an example input sequence

Type of Modeling
Select option for modeling:
1. Fully atomated - Fully automatic modeling.
2. Semi automated - Return a ranked list of potential templates. Select the best template and adjust alignment using expert knowledge.
3. Manual - Select template from available SDPs dataset and modify alignment as necessary.

Clear Button
Click the 'Clear' button to reset the email address, MODELLER key and sequence

Figure 13 The web interface of SDPMOD

The 'fully automated' mode presents an easy-to-use interface. User can simply submit a target protein sequence with their email address and their MODELLER license key, obtained from the MODELLER registration page <http://salilab.org/modeller/registration.shtml>. The modeling will be carried out automatically according to the procedure described in Figure 12. In the 'semi-automated' mode, a ranked list of potential templates will be returned after the target sequence is submitted. Users can then choose the best template and adjust the target-template sequence alignment using their knowledge. In the 'manual' mode, users are allowed to propose a template from our non-redundant SDP structure dataset and modify the target-template alignment where necessary.

After the modeling process is completed, a link with the prediction results will be returned via email. Users can refer to the link to view the prediction results and download the models. The prediction results consist of: (i) a summary of the selected template(s), (ii) the predicted model based on each template in PDB format and (iii) a brief report for each modeling attempt that includes the target-template alignment used in modeling building, a comparison of the model against the template as measured by RMSD and a PROCHECK report on the stereochemical quality of the models.

3.3 Comparative modeling of conotoxins

SDPMOD has been widely used for comparative modeling of SDPs. Till now according to the server log, more than one hundred users have submitted more than 1,000 modeling jobs to SDPMOD web server since July 2004 (Kong *et al.* 2004). SDPMOD was also used to do large-scale comparative modeling for SDP families, e.g. over 540 homology models for native and mutant scorpion toxins were built and incorporated into SCORPION2 database (Tan *et al.* 2006). But when SDPMOD was used for the comparative modeling of conotoxins, we encountered a new problem. Cone peptides contain non-standard amino acid residues, which will affect the accuracy of modeling. After a general introduction to conotoxins, their unique features and potential as drugs, comparative modeling of conotoxins will be discussed and with the solution to non-standard amino acid residues and an evaluation of the results obtained.

3.3.1 Introduction to conotoxins

Conotoxins (or conopeptides) are a vast array of peptide toxins secreted by cone snails for capturing prey and as a defense against predators. They form distinct families among SDPs and notable for their unprecedented selectivity and specificity for varieties of neuronal receptors and ion channels (Lewis 2004). These properties make conotoxins great tools in studies aimed at identifying receptors and their ligands (McIntosh *et al.* 1999a), as well as potential therapeutic drugs (Shen *et al.* 2000). Conopeptides have been reported to attack a wide variety of pharmacological targets, making them an invaluable source of ligands for studying the properties of these targets in normal and diseased states. A number of these peptides have shown efficacy *in vivo*, including as inhibitors of calcium channels, nicotinic acetylcholine receptors, NMDA receptors and neurotensin receptors, with several having undergone pre-clinical or clinical development for the treatment of pain.

3.3.1.1 Diversity of conopeptides

Conopeptides mainly come from the predatory cone snails (genus Conus). These Conus comprise one of the largest living genus of marine animals (~500 living species) (Olivera *et al.* 1990). Cone snails can be classified into three subgroups according to their prey preference: (1) piscivorous (fish-hunting); (2) vermivorous (worm-hunting); (3) molluscivorous (hunting on other marine snails). All conus use complex venoms to capture prey, defend predators and for other biological purposes. Most biologically active components of these venoms are small peptides (6-40 amino acid in length), called conopeptides, and the majority of those are in the range of 12-30 amino acids. It is estimated that there are 50~200 peptides in the venom of a single Conus species. So in all Conus venoms, the total number of conopeptides is anticipated to be in excess of 50,000 (Olivera *et al.* 1999).

Conopeptides are organized to multiple families according to disulphide bridge pattern and homologous target sites. The various conopeptide families are further grouped into superfamilies based on a surprising fact: within each superfamily, the conopeptides share a common highly conserved signal sequence in their precursors. Up to now conotoxins have six superfamilies (A, M, O, P, S, T).

The development of such potent and chemically diverse conopeptides, which simultaneously target multiple components of nervous system in their prey, is probably caused by natural selection pressure. For example, for fish-hunting cones, the slow-moving snails have to immobilize the fast-moving fish immediately.

Consequently, some fantastic mechanisms were developed in Conus to diversify the components of venom to increase efficiency. All known conopeptides are derived from precursors about 70-80 amino acids. In these precursors, the N-terminal prepropeptides sequence within a given superfamily is highly conserved. The C-terminus, which contains the mature conopeptides, represents a hypervariable region that is readily mutated. Mutation frequencies vary by more than one order-of-magnitude across these precursor sections, with the mature toxin region undergoing the highest mutation rate (Olivera *et al.* 1999). The rate of conopeptide evolution is higher than that of most other known proteins (Duda and Palumbi 1999). Post-translational modifications also contribute to the diversity of conopeptides.

3.3.1.2 The potential of conopeptides as drugs

As potential therapeutic drugs, conopeptides show their advantages from several aspects. (1) After more than 50 million years' evolution, conopeptides have been optimized to target specific ion channels and receptors with high affinities and selectivities. The diversity of conopeptides makes it possible to target wide range types of ion channel and neuronal receptors. Presently three types of targets have been identified. These are ligand-gated (Nicotinic, $5HT_3$, NMDA) and voltage-gated ion channels (Ca++, Na+, K+), and G protein-linked receptors (Vasopression, Phospholipid) (McIntosh *et al.* 1999b). (2) As conopeptides can be highly selective between closely related receptors subtypes, they could meet specific therapeutic needs with a reduced likelihood of side effects. Conus peptides are the most specific ligands known for several ion channel targets. For example, among ligands that target voltage-gated sodium channels, μ-conotoxin GIIIA has unprecedented specificity for the skeletal muscle subtype. This isoform is among the set of sodium channels that are etrodotoxin- and saxitoxin-sensitive. However, μ-conotoxin GIIIA is much more specific than either of these guanidinium toxins and has a preference for the skeletal muscle isoform by at least three orders of magnitude over other tetrodotoxin-sensitive subtypes. This high subtype selectivity is proving to be a general feature of conopeptides (McIntosh *et al.* 1999a).

These properties enable conopeptides as valuable drug candidates. For example, conantokin peptides, targeting mammalian NMDA receptors, are being considered as potential therapies for CNS disorders. Conopeptide MVIIA, which selectively blocks N-type calcium channels, is a potent analgesic drug in the treatment of neuropathic pain, as it can reduce pain with no development of tolerance (Shen *et al.* 2000).

3.3.1.3 The unique features of conotoxins

Conopeptides have several unique features: (1) signal sequences peptides within the same superfamily are extraordinarily conserved; in contrast, the mature toxin regions are hypermutated; (2) high percentage of cysteines that form structurally constrained disulphide bridges; (3) the abundance of post-translationally modified residues in conotoxins. The following section will discuss the post-translational modifications resulting in non-standard residues and their important roles in the structure and function of conotoxins.

3.3.1.4 Post-translational modifications in conotoxins

Post-translational modifications are very common in conotoxins and include hydroxylation of proline, γ-carboxylation of glutamate, bromination of tryptophan and C-terminal amidation. Post-translational modifications and their products are shown in Table 9.

Table 9 post-translational modifications in conotoxins

Post-translational modification	Standard residues	Resulted non-standard residues	Non-standard residues
C-terminal amidation	N.A.	N.A.	NH2
γ-carboxylation of glutamate	GLU	γ-carboxy-glutamic acid	CGU
hydroxylation of proline	PRO	4-hydroxyproline	HYP
Epimerization of tyrosine	TYR	D-tyrosine	DTY
C-terminal amidation of cysteine	CYS	2-amino-3-mercapto-propionamide	CY3
Epimerization of tryptophan	TRP	D-tryptophan	DTR
bromination of tryptophan	TRP	brominated tryptophan	BTR
Glycosylation of threonine	THR	glycosylated threonine	GTH

Some post-translational modifications are crucial to the structures of conopeptides. For example, research on structures of conantokin G reveals that upon binding calcium ions to γ-carboxyglutamic acid, conantokin G undergoes a conformation transition from a distorted 3^{10} helix to a linear α-helix (Rigby *et al.* 1997). Craig and coworkers also reported that γ-carboxylation of glutamate residues may play an essential role for the function of conantokins where the presence of γ-carboxyglutamate residues promotes formation of an α-helix (Craig *et al.* 1999).

Some non-standard residues play important roles in the affinity and toxicity of these toxins. A structure-activity relationship study of μ-conopeptide GIIIA showed that hydroxyl groups are essential for blocking the sodium channel, with the replacement of HYP17 with PRO17 decreasing the activity by a factor of 5 (Wakamatsu *et al.* 1992).

Overall, post-translation modifications and non-standard residues are important for the structure and function of conotoxins.

3.3.1.5 Why comparative modeling?

GenBank (up to Jan 2006) lists 1,301 conopeptides (1,072 non-redundant sequences), with only 64 of them having 3D structures in the PDB. The total number of conopeptides is anticipated to exceed 50,000. Structure determination by experimentation cannot meet the demand of so many sequences. So using bioinformatics methods to automatically predict 3D structures for conopeptide sequences is a reasonable solution.

Among the 1,072 conopeptide sequences, many sequences share common disulphide bridges scaffolds but have different residues in loop regions, which determine their specificities. So comparative modeling should be a promising tool to predict 3-D structure for native and mutant conotoxins. Although structures from modeling may contain errors, they can still provide us an insight to investigate structure-activity relationships. Furthermore, the generated homology model could be a repository of potential drug candidates.

3.3.1.6 Non-standard residues in the comparative modeling of conopeptides

When SDPMOD was first used for comparative modeling of conotoxins, non-standard residues became a serious problem that interrupted the modeling process and affected the model accuracy. Non-standard residues affected the comparative modeling from several aspects:

(1) SDPMOD has difficulty in template selection due to the high percentage of post-translational modification of residues. For example, the 8-residue-long sequence of Contryphan-Sm (PDB code: 1DFY) contains three non-standard residues: HYP, DTR and CY3.

(2) These non-standard residues cannot be recognized by the in-built

CHARMM22 forcefield (MacKerell *et al.* 1998) used by MODELLER. A quick-and-dirty solution is replacing non-standard residues with the most similar standard residues. But this will introduce inaccuracies since non-standard residues are crucial for the structure and function of conopeptides.

To address these problems, a solution had to be developed for the comparative modeling of conotoxins. In this study, the CHARMM22 forcefield topology and parameter libraries for non-standard residues in conotoxins were developed and incorporated into the library of MODELLER so that MODELLER can recognize and make use of non-standard residues for comparative modeling of conopeptides.

3.3.2 Topology and parameter development for non-standard residues

3.3.2.1 Topology definition

Currently there are eight kinds of non-standard residues (NH2, CGU, HYP, DTY, CY3, DTR, BTR and GTH) (see Table 9) found in conopeptides. Among these residues, topology and parameter files were developed for six non-standard residues (HYP, NH2, CGU, BTR, DTY, and DTR). The structures of these six non-standard residues are shown in Figure 14. The libraries for the remaining two residues (CY3 and GTH) were not developed for specific reasons. CY3 (2-amino-3-mercapto-propionamide) is actually cysteine with C-terminal amidation, identical to NH2 described earlier, and therefore, this termination does not require the development of a library for CY3. For GTH (glycosylated threonine), it is difficult to define the parameters due to the flexible nature of the sugar moiety. GTH is also rarely encountered as in the entire PDB database, there is only one entry with GTH. The lack of specific topology and parameter files for this residue do not affect the modeling of conotoxins, as it can be substituted by threonine.

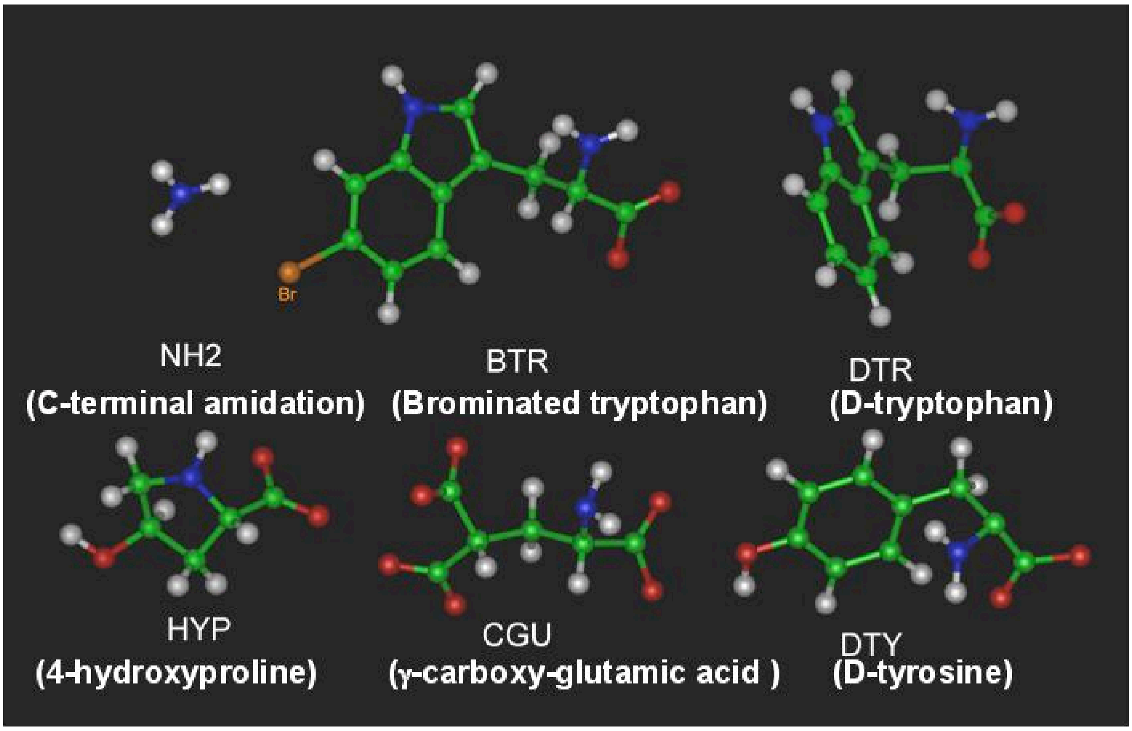


Figure 14 non-standard residues in conopeptides

Topologies and parameters for non-standard residues were developed based on high-resolution structures available from PDB database. The detailed procedure is as follows:

(1) Get the coordinates of non-standard residues from high-resolution crystal structures and then read them into Insight II (Accelrys).

(2) Check the structure to make sure there are no error or missing atoms and then add hydrogen atoms, as required.

(3) Select the CHARMM22 forcefield, assign the potential and charges, and fix partial charges. (Where the CHARMM22 forcefield cannot assign charge parameters for some atoms, BOND-INCREMENT charges were used.)

(4) Accept the assigned charges.

(5) Write “RTF” files using Insight II CHARMM RTF writer.

(6) Manual inspection and modification of the RTF file are essential. Improper dihedral angles should be added if necessary.

(7) If there are new CHARMM atom types, they need be added into the CHARMM22 topology file explicitly and their atomic radii need be defined.

3.3.2.2 Parameter estimation

Parameters are derived from similar entries in the CHARMM22 forcefield and high-resolution structures. There are several kinds of parameters that need to be defined.

(1) Bond length. The energy function for bond length is $V_{bond} = K_b\ (b-b_0)^2$. The force constant K_b is estimated from similar entries and equilibrium bond length b_0 is calculated from selected structures.

(2) Bond angle. The energy function for bond length is $V_{angle} = K_\theta\ (\theta-\theta_0)^2$. The force constant K_θ is estimated from similar entries and equilibrium bond angle θ_0 is calculated from selected structures.

(3) Dihedral angles. The energy function is: $V_{dihedral} =K_\phi\ (1+\cos(n\phi-\delta))$

K_ϕ is the force constant; n is the periodicity; δ is the phase. For dihedral angle, the force constant K_ϕ and periodicity n are basically determined by atom types of two middle atoms (X-A-A-X). When the torsion angle has the lowest energy, $\cos(n\phi-\delta)$ should be equal to -1. So the following equation can be derived: $\delta= (n\phi_0\pm180)$. ϕ_0 is the equilibrium dihedral angle and can be calculated from selected structures.

(4) Improper dihedral angles. The energy function is $V_{improper} = K_\varphi\ (\varphi-\varphi_0)^2$. The force constant K_φ is basically determined by atom types of two outer atoms (A-X-X-A) and estimated from similar entries. The periodicity n for improper dihedral

angles is always equal to 0. The equilibrium improper dihedral angle φ_0 is calculated from selected structures.

3.3.2.3 Topology and parameter evaluation method

After new topology and parameter files for each non-standard residue were generated, 3D models were built using these library files to evaluate their correctness and quality. Before the modeling can proceed, the newly developed topologies and parameters need to be incorporated into the MODELLER library by the following steps.

(1) Add new entry into restyp.lib

(2) Add new entry into model.lib

(3) Modify radii.lib and radii14.lib if necessary.

For the purpose of modeling with non-standard residues, lowercase single characters are used to represent non-standard sequences. For example, "o" stands for HYP, "k" for DTY, "m" for DTR and "v" for NH2. The scoring matrix is modified to include these non-standard residues and the values of their corresponding standard residues are used.

There are several considerations on dataset selection for the benchmarking.

(1) Only conopeptides, which have structures available in PDB, can be used to evaluate the quality of our models.

(2) The selected conopeptide sequences should include non-standard residues, so the effect of new modeling method with non-standard residues can be evaluated.

(3) To eliminate the effect of other factors such as gap, this dataset only include no gap alignment.

There are totally 19 conotoxins suitable for benchmarking according to above criteria. Among these 19 sequences, 13 sequences carry post-translation modifications of only NH2 (C-terminal amidation), while 6 sequences include HYP and 2 sequences include DTR.

The modeling is carried out with and without newly developed libraries using SDPMOD. The modeling procedure is the same, and the only difference between the two methods is that in the new method non-standard residues are introduced while in the traditional method only standard residues are used.

Models by both methods are compared to their experimentally determined structures in PDB, respectively. The RMSD (by Cα) between models and their cognate PDB structure were calculated.

3.3.2.4 Topology and parameter benchmarking results

Table 10 Comparison of models with or without non-standard residues with template structures

PDB ID	Standard model (Å)	Non-standard model (Å)	RMSD Difference	Sequence Identity	Non-standard residues and their positions
1A0M	0.59	0.46	0.13	75%	NH2 (17)
1AKG	0.51	0.36	0.15	87%	NH2 (17)
1AV3	1.50	1.55	-0.05	42%	HYP (4)
1B45	0.79	0.65	0.14	75%	NH2 (15)
1CNN	1.24	1.25	-0.01	80%	NH2 (27)
1D7T	**1.88**	**0.49**	**1.39**	**56%**	**HYP (3), DTY (4), NH2 (9)**
1DFY	**1.81**	**0.98**	**0.83**	**88%**	**HYP (3), DTR (4), NH2 (9)**
1DG2	0.89	0.74	0.15	75%	NH2 (16)
1GIB	1.30	1.21	0.09	81%	HYP (6,7,17)
1IEN	0.91	1.00	-0.09	40%	NH2 (20)
1IMI	1.34	1.28	0.06	91%	NH2 (13)
1MII	1.20	0.99	0.21	43%	NH2 (17)
1MVJ	1.20	1.19	0.01	80%	NH2 (27)
1NOT	0.66	0.64	0.02	83%	NH2 (14)
1OMN	1.29	1.28	0.01	76%	NH2 (27)
1PEN	0.56	0.48	0.08	87%	NH2 (17)
1QFB	**2.05**	**1.04**	**1.01**	**88%**	**HYP (3), DTR (4), NH2 (9)**
1QMW	0.81	0.89	-0.08	83%	NH2 (14)
1TCG	0.77	0.68	0.09	95%	HYP (6,7,17), NH2 (23)
Average	1.12	0.91	0.22		

The results in Table 10 showed that 3 models (in bold) were significantly improved after incorporating new topologies and parameters. They are 1D7T, 1DFY and 1QFB. It is reasonable because there are DTR (D-tryptophan) residues in 1DFY and 1QFB sequences and DTY (D-tyrosine) in 1D7T. D-residues will change the direction of the backbone and the use of non-standard residue templates will significantly affect the resultant structures.

For other non-standard residues such as HYP (4-hydroxyproline), NH2 (C-terminal amidation) models do not show any significant difference between the two modeling methods. The difference between HYP and PRO lies in the side chain and NH2 is only C-terminal amidation and therefore they do not affect backbone conformation as expected, although the structural models will be biologically more accurate.

Figure 15 shows the superimposition of the model with non-standard residues (1DFYnons, in green) and the standard model (1DFYstan, in red) onto its original structure (1DFY, in blue). Clearly the backbone orientation of the model with traditional standard residues (1DFYstan) is significantly different from the experimentally determined structure (1DFY), while the model with non-standard residues (1DFYnons) is very similar to 1DFY.

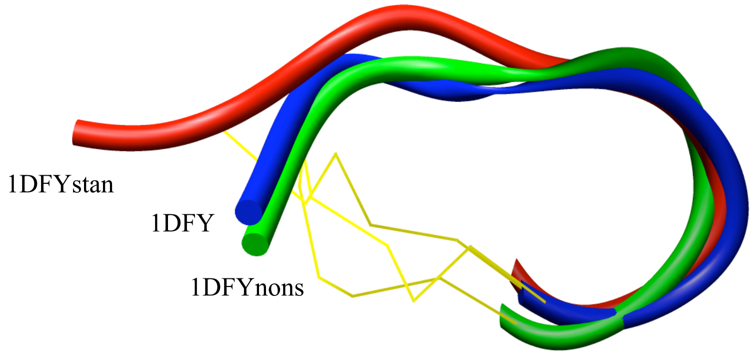


Figure 15 The superimposition of standard (1DFYstan, in red) and non-standard model (1DFYnons, in green) to the PDB structure (1DFY, in blue). The structures are in ribbon representation and disulphide bonds in wire representation (yellow).

For other non-standard residues, there were no significant improvements in the backbone RMSD values of the models. If the three models involving D-residues were removed from the list, the average RMSD difference between the two methods is only 0.05 Å. This result is reasonable because the modification in residue side chains will not significantly affect the backbone conformation.

Overall, the benchmarking results showed that the incorporation of D-residues (such as DTR, DTY) will significantly improve the quality of models. While for those non-standard residues (HYP and NH2) which only had difference in side chain with standard ones, there was only little improvement in backbone of models. The new topology and parameter libraries facilitate and improve the modeling of conotoxins. These library files also can be incorporated into other programs using the CHARMM22 forcefield.

Using the modified version of SDPMOD, homology models for 125 conopeptide sequences were built (Table 11) and the generated models had been incorporated into the MOLLUSK database <http://research.i2r.a-star.edu.sg/MOLLUSK/>.

Table 11 Statistics of homology models for conotoxin families and

their disulphide connectivities (SDFD DSF)

Conotoxin family	Conotoxin superfamily	Number of models	SDFD disulphide family (DSF)
α-conotoxin	A	22	2.1212
αA-conotoxin	A	3	3.122313
μ-conotoxin	M	12	3.123123
ϖ-conotoxin	C	78	3.123123
τ-conotoxin	T	6	2.1212
Contryphan	Others	4	N.A. (only one disulphide)

3.4 Conclusion

Chapter 3 focuses on structural modeling of SDPs. The major results are as follows:

(1) An automated comparative modeling method specifically for SDPs, SDPMOD, has been developed. The benchmarking results showed that SDPMOD can reliably generate homology models for SDPs with reasonable accuracy.

(2) A web server version of SDPMOD, with three modes of access (fully automated, semi-automated and manual) has been implemented to provide to the different needs of the users.

(3) CHARMM22 topology and parameter libraries for non-standard residues in conotoxins have been developed and incorporated into the MODELLER library, accessed by SDPMOD with validation results suggesting improved modeling accuracy, especially for conotoxins which contain D-residues.

(4) Homology models for conotoxins which contains non-standard residues have been successfully built with the updated version of SDPMOD which incorporated CHARMM22 topology and parameter for non-standard residues.

Chapter 4 Computational analysis of Pot II proteinase inhibitor family

4.1 Introduction

Proteinase inhibitors are one of the most well studied classes of proteins within SDPs and they widely exist in almost all known organisms and in various tissues of these organisms. They play critical roles in organisms in various ways: regulating the activities of endogenous proteinases and inhibiting exogenous proteinases. Proteinase inhibitors have received intensive research interests because of their potential applications in medicine and agriculture, e.g. designing effective inhibitors targeting HIV (Human Immunodeficiency Virus) proteinase or constitutive expression of inhibitors in transgenic crops to control the pests.

To better understand the important roles of proteinase inhibitors, firstly let us have a quick look at the functions of proteinases. Proteinases are ubiquitous and they have a cradle-to-grave relationship with proteins. They aid the maturation of the proteins by removing the initiating Met residues and removing the signal peptides. They also convert both exogenous proteins (food digestion) and endogenous proteins (protein turnover) to amino acids, which are then utilized for new protein synthesis or in other metabolic pathways. More importantly, proteinases process proteins to turn on or off numerous cellular regulatory activities which are responsible for many biological phenomena such as blood clotting, clot dissolution, protein hormone action, differentiation, cell death and apoptosis (Neurath 1989). Although controlled proteolysis is essential to life, unrestricted proteolysis is lethal. If our blood clots uncontrollably, or our pancreas are digested by self-secreted proteinases, the consequences would be lethal. Along with so many important functions, the proteolysis processes must be tightly controlled in time and place in order to be effective. One of the most important measures developed during protein evolution is the creation of proteinase inhibitors.

Proteinase inhibitors (PIs) can be classified into four categories according to the classification of their targeted proteinase: serine-, cysteine-, metallo- and aspartyl-proteinase inhibitors (Laskowski and Kato 1980). Among them, serine proteinase inhibitors have the largest number of well characterized members because of the dominant role of serine proteinases and their inhibitors in fundamental life processes.

Serine proteinase inhibitors from plants are reported to be major constituents of seeds, tubers and leaves of members of the *Solanaceae* (e.g. potatoes, tomatoes, eggplant, sweet peppers, chili peppers, tobacco and petunias) and *Leguminosae* (e.g. legumes, pea or bean) families (5-15% of the total protein) (Richardson 1977). These PIs are an integral part of the constitutive and inducible defensive mechanisms that protect plants from attacking pests (bacteria, fungi and insects) (Bowles 1990). These defensive mechanisms involve the systemic synthesis of serine PIs that accumulate in distal tissue and can inhibit the digestive trypsin- and chymotrypsin-like enzymes of insects and other related serine proteinases of plant pathogens (Johnson *et al.* 1989). The inhibitory properties towards serine proteinases of these PIs have already been exploited for the production of transgenic plants over-expressing specific PIs in an attempt to control pests (Duan *et al.* 1996).

Potato type II proteinase inhibitor family (Pot II) is one of the major serine proteinase inhibitor families which are mainly found in higher plants from the *Solanaceae* family (Greenblatt *et al.* 1989). Pot II accumulation is always in response to stress, infection and wounding, and constitute an important measure for defense against predators or diseases. Intensive research has been conducted on proteinase inhibitors (PIs) from this family.

This family of PIs is interesting in that it exhibits domain duplication resulting in 2-8 copies of the ancestral single domain protein, of which we have evidence only from genome sequences. More interestingly, the structure adopted by these proteins is a permutation of the ancestral fold, so that the structural repeat does not correspond to the sequence repeat. The correlation between sequence and structural repeats within this family and the evolution and molecular adaptation of Pot II genes has been investigated through computational analysis, using the putative ancestral domain sequence as the basic repeat unit.

4.1.1 Origin and function of Pot II PIs

Previous research suggests that there are mainly three kinds of physiological functions for Pot II PIs:

(1) defense against predators or diseases. Members of the Pot II have been reported to inhibit a wide spectrum of serine proteinase, such as trypsin, chymotrypsin, subtilisin, oryzin and elastase (Pearce *et al.* 1982; Plunkett *et al.* 1982);

(2) endogenous regulatory role. Reports on their developmental regulation and their tissue-specific accumulation suggest they have endogenous functions such as regulating proteolysis (Xu *et al.* 2001);

(3) storage proteins in tuber or seeds. For example, Potato Inhibitor II (PI-II) is one of the major proteins in Russet Burbank potato tubers, representing about 5% of the soluble proteins (Greenblatt *et al.* 1989). The

concentrations of PIs in potato tubers are dramatically lowered during their sprouting, which suggests that PIs may serve as storage proteins during the development of plants (Richardson 1977).

PIs of the Potato II (Pot II) inhibitor family have been isolated from various plants and organs: wounded tomato and tobacco leaves, green tomatoes, potato tubers, eggplant fruits, paprika seeds and ornamental tobacco flower stigma. Pot II PIs can accumulate systemically in plant tissue as a result of wound, stress or pathogen attacks. But some PIs are expressed constitutively or regulated in a developmental- and tissue-specific manner. The systemic response to attack in the *Solanaceae* family has been attributed to a complex signaling cascade that is initiated by the binding of systemin to a cell-surface receptor and leads to the release of linolenic acid which is then converted to 12-oxophytodienoic acid and jasmonic acid (Li *et al.* 2002). The release of jasmonic acid leads to the activation of several signaling pathways that in turn lead to the production of more jasmonic acid, H_2O_2 and the synthesis of PIs (Ryan and Moura 2002). Within 48 hours of insect attack or wounding, PIs can accumulate to levels of 2% or more of the total soluble protein in the leaves of tomato and potato plants and are thought to have adverse effects on the digestive physiology of insects (Lee *et al.* 1986). The wide distribution and inducible expression of Pot II PIs in plants strongly suggest the fundamental importance of these proteins to the pest defense strategies of many commercially important crops. Table 12 below summarizes the distribution of Pot II PIs on species, tissues and expression patterns.

Table 12 The source and expression profile of Pot II PIs

Species	Tissue	Expression profile
Capsicum annuum (paprika, bell pepper)	Seeds (Antcheva *et al.* 1996)	Constitutive expression
	Pericarp (unpublished, Swiss-Prot entry name: IP22_CAPAN)	Development

	Flower, green fruits (rich); leaves, red fruits (little); root, stems (absent) (Shin *et al.* 2001)	TMV-P0;
		SA, MeJA, ethephon;
		Wound

Nicotiana alata (persian tobacco, ornamental tobacco)	Stigmas of flower(Nielsen *et al.* 1995; Lee *et al.* 1999; Miller *et al.* 2000)	Development
Nicotiana attenuate (coyote tobacco)	Leaves (Hui *et al.* 2003)	Wound
Nicotiana glutinosa (tobacco)	Young leaves and floral organs (Choi *et al.* 2000)	Development
	Mature leaves(Choi *et al.* 2000)	Wound, pathogen
Nicotiana tabacum (common tobacco)	Leaves (Pearce *et al.* 1993; Hara *et al.* 2000)	Wound (not by systemin)
	Flower (Pearce *et al.* 1993)	Development
Lycopersicon esculentum (tomato)	Leaves (Graham *et al.* 1985) aerial tissues (Gadea *et al.* 1996), Seedling root (Taylor *et al.* 1993)	Wound (systemin) (Graham *et al.* 1985), viroid infection and ethephon treatment (Gadea *et al.* 1996), auxin (Taylor *et al.* 1993)
	Green fruits (Pearce *et al.* 1988), shoot apex and developing flower (Brandstadter *et al.* 1996)	Development
	Roots of healthy plants(Gadea *et al.* 1996)	Constitutive expression (Gadea *et al.* 1996)
Solanum americanum (black nightshade)	Phloem of stems, roots and leaves, flowers (Xu *et al.* 2001)	Development
Solanum melongena (eggplant)	Fruits(Richardson 1979)	Constitutive expression
Solanum phureja	leaves (unpublished, GenBank Accession No.: AAO88244)	Wound-induced
Solanum tuberosum (potato)	Root, leaves (Dammann *et al.* 1997)	Systemin Wound->Abscisic acid -> jasmonic acid (Dammann *et al.* 1997)
	Tuber (Bryant *et al.* 1976)	Constitutive expression (Bryant *et al.* 1976)

4.1.2 Domain repeats in Pot II

Interesting phenomena in Pot II family (such as tandem duplication, domain swapping and fold circular permutation (Scanlon *et al.* 1999)) make this family an excellent example to study gene family evolution and protein folding. Members within this family have been identified with different numbers of tandem sequence repeat units (RUs), such as two (Keil *et al.* 1986), three (Balandin *et al.* 1995), four (Miller *et al.* 2000), six (Atkinson *et al.* 1993), seven (GenBank Accession No.: AAO85558) and eight (Choi *et al.* 2000) RUs. Each RU can be characterized as a ~50-residue-long 8-cysteine polypeptide, which includes a reactive site targeting serine proteinases. The evolution of several members of this multi-domain family, at the gene duplication level, has been recently reported (as the Pin2 family (Barta *et al.* 2002)). However, the complex correspondence between sequence repeats and their 3D structure has not been well investigated.

Several 3D structures of the Pot II family are known, belonging to the SCOP (Lo Conte *et al.* 2000) fold family of plant proteinase inhibitors. Pot II family RUs adopts a variety of structural repeats, by circular permutation of the same fold (Greenblatt *et al.* 1989; Lee *et al.* 1999; Scanlon *et al.* 1999). Structures exhibited by naturally occurring proteins are single or double chain permutated domains composed of N- and C-termini segments from sequence repeats. The engineered putative ancestral domain protein alone has a fold corresponding to the sequence repeat (Scanlon *et al.* 1999).

The complex correlation between sequence and structural repeats within this family has been investigated using sequence, structural and phylogenetic analyses, with the putative ancestral domain sequence as the basic repeat unit. Systematic analysis of Pot II family using bioinformatic approaches has revealed many interesting findings.

(1) The sequence repeats cluster into distinct phylogenetic groups depending on the repeat number and the species. The conservation patterns between repeat units in available genes suggest variation of duplication history and mechanism in different species.

(2) The permutated domains appear more stable than original repeat domain, from available structural information. Therefore, a multiple-repeat sequence (up to eight in *Nicotiana*) is likely to adopt the permuted fold from contiguous sequence segments, with the N- and C-termini forming a single non-contiguous structural domain, linking the bracelet of tandem repeats.

(3) Two 3-repeat sequences from *Capsicum annuum* have evolved to tailor the sequence repeats to correspond with the structural repeats thus eliminating the bracelet link. The repeat unit for this group is a circular permutation of the ancestral domain, making this group the late entrant to the Pot II family.

(4) The analysis of nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) in Pot II domain revealed heterogeneous selective pressures among amino acid sites: the reactive site is under position selection (providing different specificity to target varieties of proteinases) while the cysteine scaffold is under purifying selection (essential for maintaining the fold).

(5) For multi-RU Pot II genes from *Nicotiana* genus, the proteolytic processing site is under positive selection to achieve higher efficiency for cleavage.

This chapter provides comprehensive analysis and characterization of the Pot II family, and aims to enlighten our understanding of the strategies (gene and domain duplication, structural circular permutation and molecular adaptation) of *Solanaceae* plants for defense against pest attacks through the evolution of Pot II genes.

4.2 Materials and Methods

4.2.1 Collection of Pot II Family Members: structures, gene and protein sequences

To identify 3D structures in Pot II family, PSI-BLAST (Altschul *et al.* 1997) was used to search against PDB (Berman *et al.* 2000) database with the Potato Inhibitor II sequence (PI-II, SwissProt Accession No.: P01080; Keil *et al.* 1986). The default parameters were used with four iterations (to convergence) and manual selection of homologues. There were seven significant hits, 4SGBI, 1FYB, 1CE3, 1TIH, 1QH2, 1OYV and 1PJU. For NMR structures where the PDB entry comprises multiple conformers, NMRCLUST (Kelley *et al.* 1996) has been used to choose the representative structure (details available in Chapter 3).

The gene structure of Pot II family will provide clues about its evolution. DNA sequences of Pot II genes were retrieved through a search of the non-redundant GenBank database (Benson *et al*. 2006) with TBLASTN (Altschul *et al.* 1990) using PI-II. Only complete DNA sequences were retrieved. TBLASTN searches were also performed against *Arabidopsis thaliana* and *Oryza sativa* genomes available from TIGR (The Institute for Genomic Research, http://www.tigr.org/) and single domain Pot II genes were located. The final dataset for Pot II genes was derived from the combination of the results of all these searches followed by redundancy removal and manual checking. The GenBank accession numbers of 13 significant hits are

AB110700, AK105387, AY007240, AY129402, L25128, M15186, NM_105864, U45450, X04118, X78275, Z12753, Z13992 and Z29537.

PSI-BLAST was used to search against NCBI non-redundant protein database to retrieve protein sequences of the Pot II family while TBLASTN facilitated searching the NCBI dbEST database (Boguski *et al.* 1993). The search results were combined with the collection of Pfam (Bateman *et al.* 2002) entry Prot_inhib_II, which contains 94 Pfam domains (as of December 2005). Partial sequences and redundancies were removed. The final sequence dataset includes 40 protein sequences. The IDs (Swiss-Prot names, accession numbers and GenBank accession numbers are used whenever possible.) for these sequences were listed as follows: AAF14181, AAF18450, AAF18451, AAF25496, AAO85558, AAL36458, AAO88244, AAR37362, AAX84035, AAX84036, AC096689, AI724716, AY105802, AW616253, BE033392, BE033653, BE033692, BE942349, BE943304, BI421162, BI434643, BI436259, CAA27409, CAA27730, CN847229, CO516657, IP22_CAPAN, IP27_SOLTU, IP2Y_SOLTU, IP25_SOLTU, IP2K_SOLTU, IP2T_SOLTU, IP2X_SOLTU, IP21_LYCES, IP23_LYCES, IP22_LYCES, IP21_TOBAC, JQ2153, NP_177351 and X99095.

4.2.2 Protein Structure Analysis

The alignments of 3D structures were performed using MULTI-GAFIT (May and Johnson 1995) and MALIGN3D algorithm in the MODELLER package (Sali and Blundell 1993). The structures were displayed using RASMOL (Sayle and Milner-White 1995) and Swiss PDB Viewer (Guex and Peitsch 1997). Structural images were generated using YASARA (available from http://www.yasara.org). MODELLER (Sali and Blundell 1993) was used to build homology models for PI-II. The different

Pot II topologies were compared to evaluate their structural qualities by using several structure validation methods, WHATIF Packing Quality Control (Vriend and Sander 1993), ProQ (Cristobal *et al.* 2001) and ERRAT (Colovos and Yeates 1993).

4.2.3 Gene Structure Analysis

The analysis of Pot II family gene structure (exon/inton boundary, organization and splicing phase) were facilitated by Xpro (Gopalan *et al.* 2004) and EMBOSS (Rice *et al.* 2000). The *Arabidopsis thaliana* and *Oryza sativa* genomes were downloaded from TIGR. The FASTA format genomic sequences were formatted and queried using NCBI standalone BLAST package (Altschul *et al.* 1990).

4.2.4 Protein Sequence Analysis

The sequences of Pot II proteins were extracted and then split into single Repeat Units (RUs) according to the putative ancestral domain sequence from 1CE3. The multiple sequence alignments were carried out with CLUSTAL_X (Thompson *et al.* 1997) , followed by manual inspection and adjustment, to maximize the alignment of identical and similar residues and minimize the number of gaps. The consensus sequences were represented using Sequence Logos (Schneider and Stephens 1990). The degree of conservation of each amino acid was assessed by the maximum-likelihood method (Armon *et al.* 2001) and mapped onto the surface of the putative ancestral 3D structure (1CE3) using ConSurf (Glaser *et al.* 2003).

4.2.5 Phylogenetic Tree Building

Nucleotide sequences were retrieved from NCBI Entrez server and split into single RUs corresponding to putative ancestral domain sequence from 1CE3. The alignment of nucleotide sequences was facilitated by protal2dna server (http://bioweb.pasteur.fr/seqanal/interfaces/protal2dna.html), based on the aligned amino acid sequences. The phylogeny was estimated using Neighbor-Joining method (Saitou and Nei 1987a) and bootstrapped for 1000 replicates. The trees were displayed using TreeView (Page 1996).

4.2.6 Analyses of Selective Pressure

To examine the selective pressure acting on genes from Pot II family, only sequences from *Solanaceae* plants were used and the single-RU Pot II genes were excluded from the dataset since they are not well annotated and their inhibition functions are uncertain. The dataset included 83 RUs sequences from multi-RU Pot II genes after removing 12 single-RU genes. All the analyses were performed using the CODEML module of the PAML 3.15 package (Yang 1997).

4.2.6.1 Site-based Analysis

Codon-substitution Models of the variable ω (dN/dS, nonsynonymous and synonymous substitution ratio) among sites were used to test for the existence of amino acid sites under positive selection (with ω >1) and to identify these sites. Several models (M0, M1, M2, M3, M7 and M8) were used for this analysis, as recommended by Yang *et al.* (Yang *et al.* 2000a; Wong *et al.* 2004) and implemented in the CODEML module of the PAML 3.15 package (Yang 1997). For this analysis, the tree topology generated by the previous phylogenetic (Section 4.2.6) analysis is used, with the exclusion of the single-RU Pot II genes since these are not well annotated and their function and proteinase inhibitory activities are putative.

Among the model used, Model M0 (one ratio) assumes an invariable ω for all sites. Model M1 (NearlyNeutral) assumes two classes of sites in the protein: the conserved sites at which $0 < \omega < 1$ and the neutral sites at which $\omega = 1$. In addition to the classes mentioned for M1, the M2 Model (PositiveSelection) adds a third class of sites with ω as a free parameter, thus allowing for sites with $\omega > 1$. Model M3 (discrete) uses a general discrete distribution with three site classes, with proportions (p_0, p_1, and p_2) and the ω ratios (ω_0, ω_1, and ω_2) estimated from the data. Model M7 (β) assumes a β distribution between 0 and 1 depending on the parameters p and q. Finally, Model M8 (β and ω) adds an extra class of sites to the β (M7) model, with ω values and proportions estimated from the data. Among the above models, only Models M2, M3, and M8 can detect sites under positive selection.

From these models, Likelihood Ratio Test (LRT) can be done to test the positive selection hypothesis by comparing the simpler null hypothesis (M0, M1 and M7) with their more complex alternative models (M3, M2 and M8). All analyses were checked for convergence by performing the analysis with different starting ω values (0.3, 1 and 1.7). When the estimation of the parameters was finished, both naive empirical Bayes (NEB) (Nielsen and Yang 1998; Yang *et al.* 2000b) and Bayes empirical Bayes (BEB) (Yang *et al.* 2005) approaches were used to calculate the posterior probability for site classes. All statistics analyses were performed using the CODEML module in the PAML package (Yang 1997).

4.2.6.2 Branch-based Analysis

To test whether there is significant difference in selective pressure among different clades, branch models have been used, which allow for variable ω ratios among branches in the tree (Yang 1998). The null hypothesis model assumed the same ω for all lineages in the tree. The alternative hypothesis model assigns different ω ratios for different clades in the tree (discussed in Section 4.3.4). An LRT has been carried out to compare the null and the alternative hypothesis models.

4.2.6.3 Clade-wise Site-based Analyses

Site models assume the same ω ratios for all branches while branch models assume no variation among amino site sites. These site and branch models might not detect lineage-specific changes in selective pressure at specific amino acid sites. The branch-site model (Yang and Nielsen 2002; Zhang *et al.* 2005) allows the ω ratio to vary both among lineage and among sites but the current implementation of branch-site model only supports two branch types and cannot be used to detect different positive selection sites among different clades. Clade-wise site-based analyses in selective pressure have been conducted on Clade 3 (1st RUs of 2-RU or 3-RU PIs), Clade 4 (2nd RUs of 2-RU or 3-RU PIs) and Clade 7 (Similar RUs of multi-RU PIs from Nicotiana genus). Other clades cannot be analyzed separately since they contain very few sequences.

4.2.7 Codon Usage Analysis

The codon usage analyses were carried out to check whether there is codon bias in the Pot II gene family for domain duplication. The single-RU Pot II genes were removed from the dataset. Codon usage tables of Pot II genes were calculated using the CUSP module of EMBOSS package (Rice *et al.* 2000). Codon usage tables for individual species were retrieved from the Codon Usage Database (Nakamura *et al.* 2000), which is available from http://www.kazusa.or.jp/codon. Codon usage tables were compared with the Graphical Codon Usage Analyser (GCUA, http://gcua.schoedl.de/) (Fuhrmann *et al.* 2004).

4.3 Results and Discussion

4.3.1 Protein 3D Structure Analysis of the Pot II Family

PSI-BLAST identified seven structures for the Pot II family. These are 4SGB (Greenblatt *et al.* 1989), 1CE3 (Scanlon *et al.* 1999), 1FYB (Schirra *et al.* 2001), 1QH2 (Lee *et al.* 1999), 1TIH (Nielsen *et al.* 1995), 1OYV (Barrette-Ng *et al.* 2003b), and 1PJU (Barrette-Ng *et al.* 2003a). Among them, 1TIH, 1QH2 and 1FYB are one or two domains (T1, C2 and C1-T1 domains, respectively) of the *Nicotiana alata* Pot II PI (Na-PI) (Atkinson *et al.* 1993) a 6-domain precursor protein. The engineered single domain proteinase inhibitor, 1CE3, is the putative ancestral protein of Na-PI, which corresponds to the single domain RU putative sequences identified by genome searching (Section 4.2.1). The representative structures for 1CE3, 1FYB and 1TIH were selected by NMRCLUST as models 9, 4 and 5, respectively. These monomers are named 1CE39, 1FYB4 and 1TIH5. The structure of PCI-1, from the I chain of 4SGB, is referred to as 4SGBI. 1OYV is a 2:1 complex of Subtilisin Carlsberg and the two-domain tomato inhibitor II (TI-II), while 1PJU is actually the unbound form of TI-II. All these structures belong to the SCOP (Lo Conte *et al.* 2000) family of plant proteinase inhibitors. Among these structures, only 1FYB and 1PJU are two-domain PIs while the rest have a single domain. All these structures have little secondary structure and are restrained principally by four disulphide bridges in each domain, and the main secondary structure in their folds is an anti-parallel 3-stranded β-sheet on the face opposite to the reactive site loop.

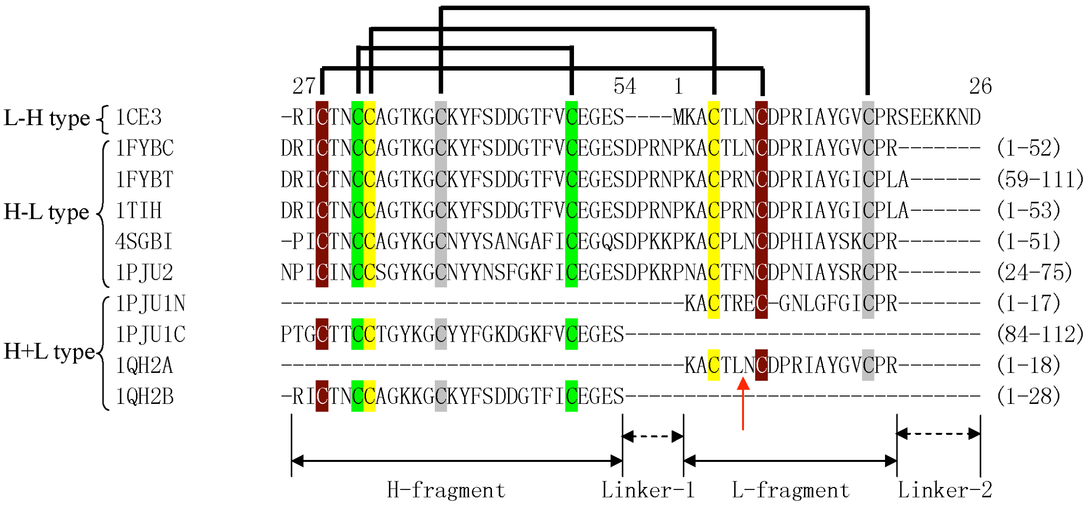

27
54
1
26
L-H type
1CE3 -RICTNCCAGTKGCKYFSDDGTFVCEGES----MKACTLNCDPRIAYGVCPRSEEKKND
H-L type
1FYBC DRICTNCCAGTKGCKYFSDDGTFVCEGESDPRNPKACTLNCDPRIAYGVCPR------- (1-52)
1FYBT DRICTNCCAGTKGCKYFSDDGTFVCEGESDPRNPKACPRNCDPRIAYGICPLA------ (59-111)
1TIH DRICTNCCAGTKGCKYFSDDGTFVCEGESDPRNPKACPRNCDPRIAYGICPLA------ (1-53)
4SGBI -PICTNCCAGYKGCNYYSANGAFICEGQSDPKKPKACPLNCDPHIAYSKCPR------- (1-51)
1PJU2 NPICINCCSGYKGCNYYNSFGKFICEGESDPKRPNACTFNCDPNIAYSRCPR------- (24-75)
H+L type
1PJU1N --------------------------------KACTREC-GNLGFGICPR------- (1-17)
1PJU1C PTGCTTCCTGYKGCYYFGKDGKFVCEGES---------------------------- (84-112)
1QH2A --------------------------------KACTLNCDPRIAYGVCPR------- (1-18)
1QH2B -RICTNCCAGKKGCKYFSDDGTFICEGES---------------------------- (1-28)
H-fragment
Linker-1
L-fragment
Linker-2

Figure 16 Multiple sequence alignment of domains of all structures in the Pot II family. The arrow marks out the positions of the reactive sites. Pairs of cysteines forming disulphide bridges are linked by lines. Abbreviations used: 1FYBC, chymotrypsin-specific domain of 1FYB (Domain I); 1FYBT, trypsin-specific domain of 1FYB (Domain II); 1PJU2, Domain II of 1PJU; 1PJU1N, N-terminal segment of 1PJU (Domain I); 1PJU2C, N-terminal segment of 1PJU (Domain I); 1QH2A, chain A of 1QH2; 1QH2B, chain B of 1QH2.

The sequence alignment of domains of the Pot II family structures (Figure 16) suggests that the sequences of all domains can mainly be divided into two parts, named here as the H- and L-fragments (for heavy and light fragments) connected by Linker-1 or Linker-2. In most structures, the L-fragment forms the reactive loop and one strand of the β-sheet, while the H-fragment forms a loop and two strands.

From Figure 16, it is clear that all the structures share the same disulphide connectivity although the combination of the H- and L-fragments is different. These domains can be divided into three types based on their sequences and structures: (1) H-L type (H- and L-fragment joined by Linker-1): with structural examples, 4SGBI, 1TIH, 1FYBC, 1FYBT and 1PJU2; (2) L-H type (L- and H-fragment linked by Linker-2): the engineered ancestral protein 1CE3; (3) H+L type (No Linker-1 or Linker-2 between the two fragments): 1QH2 and 1PJU1. The three structures shown in Figure 17 are actually circular permutations of the same fold. All three topologies have the β-sheet and the functional proteinase inhibitory site conserved, although the intra-chain connectivities are different. The H+L structure (1PJU1) can be considered the basic fold, with Linker-1 between C2 and N1 in 4SGBI and Linker-2 between C1 and N2 in 1CE3. The existence of the H+L structure shows the viability of a two-chain protease inhibitor in this fold family.

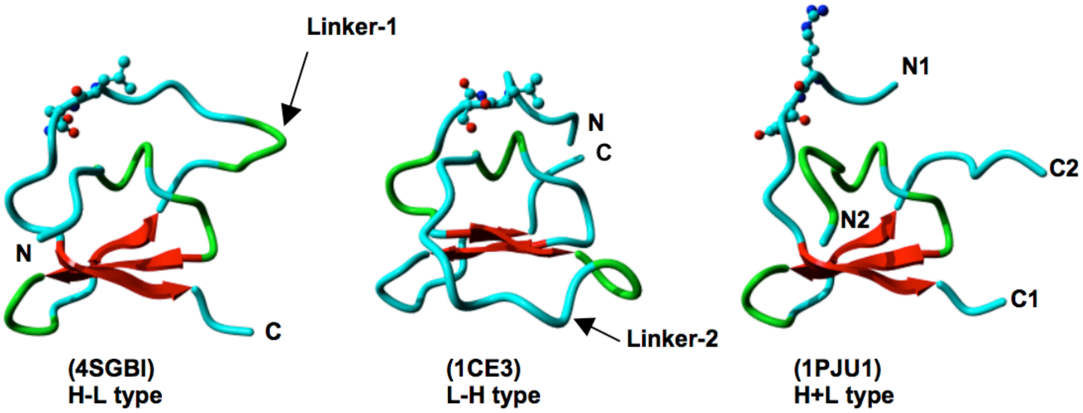


Figure 17 Structural comparison of three types of Pot II PI topologies: H-L, L-H and H+L. The structures are in ribbon representation, with the N- and C-termini marked and the reactive sites depicted in ball-and-stick mode. The β-strands are shown in red, with the linker regions marked.

Based on the structure analysis of the plant proteinase inhibitor family, it is obvious that the same fold is possibly formed by different topologies by circular permutation of sequence information. In a protein with multiple repeated regions, such as PI-II (with two domains) and the ornamental tobacco (*Nicotiana alanta*) Na-

PI-II (with six domains), theoretically there are two possible domain organizations: (1) tandem repeat domain organization. Each domain is equivalent to the sequence repeat and adopts L-H topology; (2) circularly permuted domain organization. The domains do not correspond to the sequence repeats. The domain formed by N- and C-terminal sequence segments adopts H+L topology while the other internal domains adopt the H-L topology.

So the problem is: given a multi-RU Pot II protein, which domain organization will it naturally prefer? Based on the observation of the current data set, all experimentally determined multi-domain structures have circularly permuted two-domain organization (an H+L domain and an H-L domain). And most single-domain Pot II PIs (often derived from processing of multi-domain PIs) adopt the H-L type topology which also suggests that the multi-domain PIs have circularly permuted domain organization before they were processed. The only exception is 1CE3, which has only one RU in its primary sequence and thus can only adopt the L-H topology alone, and moreover it is the product of an engineered gene (Scanlon *et al.* 1999). The abundance of the H-L topology suggests it is more favorable in nature than the L-H topology.

So the next question is: does the H-L topology have an advantage (e.g. greater stability or better packing) over the L-H topology? To evaluate the structural quality of different topologies and domain organization, several structure validation methods (WHATIF packing quality control (Vriend and Sander 1993), ERRAT (Colovos and Yeates 1993) and ProQ (Cristobal *et al.* 2001)) were used, to compare representative structures of each type. In the Pot II family, there is only one 2-domain structure available namely Tomato Proteinase Inhibitor II, (TI-II, PDB ID: 1PJU), which adopts a circularly permuted 2-domain domain organization. To compare the structure quality between two types of domain organizations, 3D models were built for PI-II of Type 1 (tandem 2-domain) and Type 2 (circularly permuted 2-domain) for the purpose of further analysis, named PI2t1 (Type 1, based on template 1CE3) and PI2t2 (Type 2, based on template 1PJU), respectively. The comparison results are summarized in Table 13.

Table 13 Quality comparison of representative structures using different structure validation methods.

Structure	Domain organization	Domain topology	WHATIF quality control		ERRAT Score	ProQ	
			Coarse	Fine		LGscore	MaxSub
1PJU	Permuted 2D	H-L, H+L	-1.59	-0.95	92.16	1.69	0.09
PI2t1	Tandem 2D	L-H, L-H	-2.11	-4.60	57.28	1.42	0.07
PI2t2	Permuted 2D	H-L, H+L	-1.54	-2.36	86.41	2.02	0.13
1PJU2	1D	H-L	-1.55	-0.43	88.10	0.92	0.08
1CE3	1D	L-H	-1.93	-3.43	47.86	0.09	-0.09
1QH2	1D	H+L	-2.20	-2.73	NA	0.20	-0.10

The results (shown in Table 13) of WHATIF packing quality control showed that the coarse scores (-1.59 and -1.54) for both of permuted 2D structures (1PJU and PI2t2) are better than the score (-2.11) of the tandem 2D structure PI2t1. According to WHATIF documentation, a molecule is certain to be incorrectly folded if the average coarse packing quality score is below -3.0, while poorly refined molecules, very well energy minimized mis-threaded molecules and low homology models give values between -2.0 and -3.0. The fine packing quality control suggests that permuted structures and H-L type structures have better packing quality than tandem repeat and L-H type structures, based on the fine packing quality control criteria. ERRAT and ProQ also recommend permuted structures and H-L type topologies have better structure qualities with fewer packing errors. Overall, the structure quality comparison of representative structures using different structure validation methods suggested that H-L type topology is the most favorable topology and that multi-domain Pot II proteins tend to fold as H-L topology domains.

4.3.2 The Gene Structure of Pot II Family

Gene structures can potentially provide clues for the evolution of Pot II family. To this end, the gene structures of the Pot II PIs were investigated. Firstly the exon/intron organization information for all available Pot II family members was collected. TBLASTN searches were carried out with PI-II against the GenBank non-redundant database as well as the *Oryza sativa* genome and the assembled *Arabidopsis thaliana* genome from TIGR. The searches retrieved DNA/RNA records which include Pot II repeat units. All the results were combined, and only records which have complete coding sequence (CDS) information were retained. All the 30 significant hits come from plants. More specifically, most of them were from *Solanaceous* family species except one entry each from *Arabidopsis thaliana, Oryza sativa* and *Zea mays.* Only 13 entries from the 30 significant hits have intron information available. Among these 13 records, six are from *Solanum tuberosum*, four from *Lycopersicon esculentum* and one each from *Nicotiana tabacum, Oryza sativa* and *Arabidopsis thaliana*.

The locus and distribution of Pot II gene in the *A. thaliana* genome can be investigated using the assembled whole genome sequence for *A. thaliana,* available from TIGR *Arabidopsis thaliana* Database (http://www.tigr.org/tdb/e2k1/ath1), using TBLASTN searches. The results show that there is only one copy of the Pot II gene (labeled here as AT-PI) in the entire *A. thaliana* genome, with one RU. The locus for this gene is 26,718,284-26,718,630 of chromosome 1 and it was composed of two exons (26,718,284-26,718,326, 26,718,435-26,718,630) and a 108-bp intron

(26,718,327-26,718,434).

The putative Pot II gene in *Oryza sativa* (OS-PI) is from whole genome shotgun sequence (GenBank Accession No.: AAAA01000128) (Yu *et al.* 2002). The locus for single-domain OS-PI is 15,645-15,297 (on the reverse strand) with two exons (15,645-15,600, 15,496-15,294) and a 103-bp intron (15,599-15,497). As with *A. thaliana*, rice has a single copy of the 1-RU Pot II gene.

The exon and intron information for all records with available intron information were collected and their gene structures were investigated with the assistance of the Xpro database (Gopalan *et al.* 2004) (http://origin.bic.nus.edu.sg/xpro/). Interestingly, all the records had the same gene structure including the putative Pot II genes from *A. thaliana* and *Oryza sativa*.

(1) All the records have two exons. The first exon encodes a part of the signal peptide (12-17 residues). The second exon encodes the remaining part of the signal peptide (7-12 residues) and the mature polypeptide. There is no intron between the RUs in the genes of multi-RU.

(2) The splice phases for all records are conserved as phase 1. The last nucleotide of the exon 1 and the first two nucleotides of exon 2 always encode a Gly residue.

(3) The splicing motif is also conserved and found to be GT…AG.

Overall, the conservation of exon/intron organization, splice phase, splice motif and Gly residues all confirm the homologous relationship between the identified Pot II family members. The same gene structure features are also found in AT-PI and OS-PI, which are strongly indicative of these two are also members of the Pot II family. Moreover, it is found that in all the Pot II family members lacking intron information, there is a conserved Gly in a similar location in their signal peptides. These records came from a range of species of the *Solanaceae* family, such as *Solanum americanum*, *Solanum nigrum*, *Nicotiana glutinosa, Nicotiana alata* and *Capsicum annuum.*

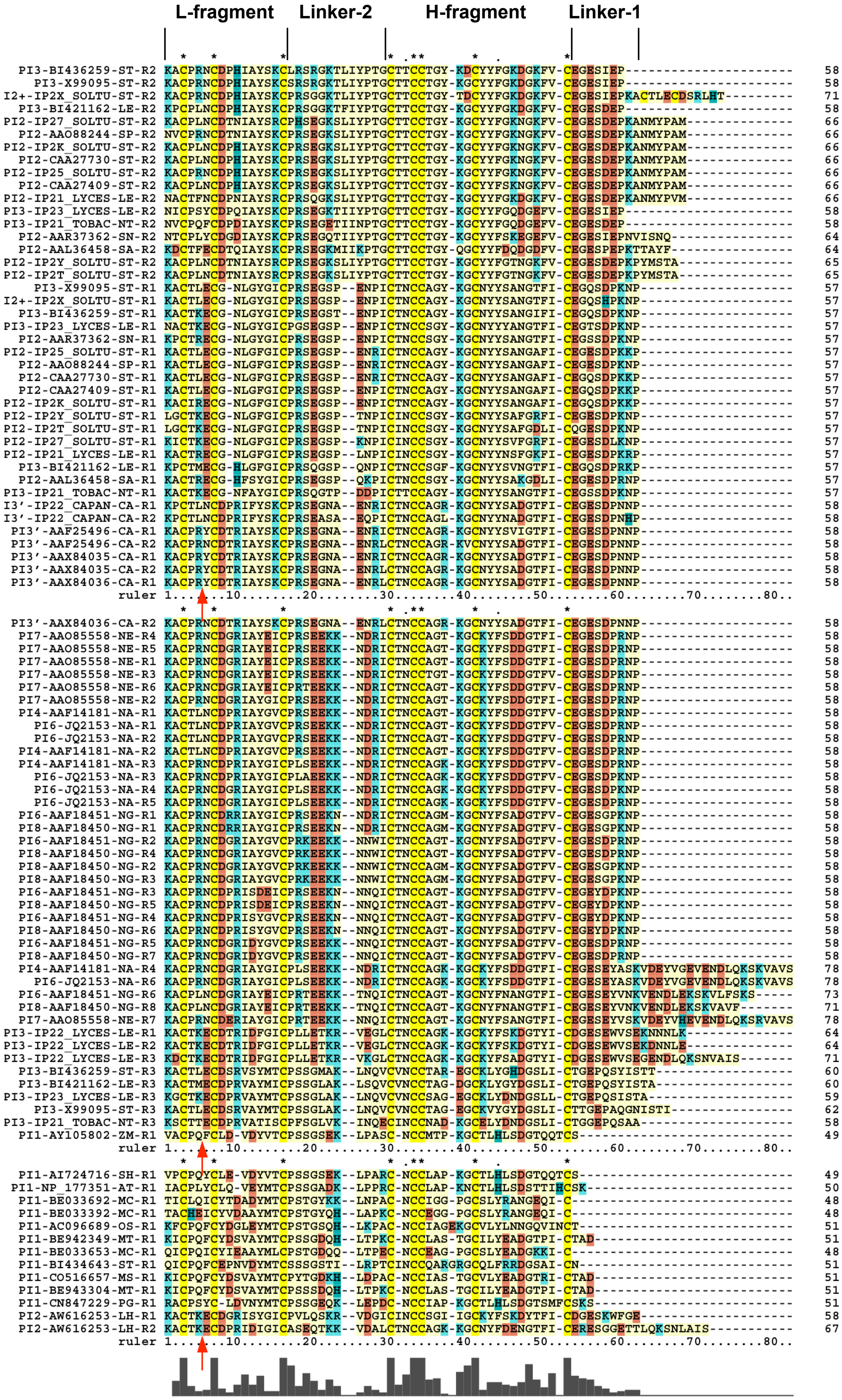
Both AT-PI and OS-PI had only one L-H type RU. Although more than ten single-domain PIs have been reported, none of them was found to be the direct translation product of a single-RU gene. On the contrary, most of them are identical to a part of multiple-domain PI precursors, indicating that these single-domain PIs are proteolytic products of multiple-domain PIs. Considering the range of multiple-domain PIs found in *Solanaceae,* gene duplication mechanism has been suggested to play an important role in the evolution of the Pot II family members, with the ancestral gene having only one RU (Scanlon *et al.* 1999). The characteristics of AT-PI and OS-PI strongly support this hypothesis.

Generally, the existence of introns between exons are regarded as facilitators of domain duplication events, since without introns there would be only a few sites in the original gene at which a recombination could duplicate the domain (Alberts *et al.* 2002). The mechanism for tandem domain duplication in Pot II family, however, remains unclear. Although the multi-RU proteins can be regarded as a result of a series of unequal crossovers (UECOs) (Barta *et al.* 2002), it is not sufficient to explain how the domain duplication has occurred accurately without the assistance of introns. For example, in the animal Kazal family, which shares the same SCOP superfamily as the Pot II PIs, there is an intron between each inter-repeat region (Scott *et al.* 1987). With the present dataset, there is very little information on gene structures to enable us to arrive at a hypothesis on the evolution of multi-RU Pot II members. Further investigation of the duplication mechanism requires the availability of more sequenced plant genomes.

4.3.3 Protein Sequence Analysis

The protein sequences of all Pot II family members were collected and putative Pot II PIs from the NCBI non-redundant protein database and dbEST database. After removing duplicates, 40 non-redundant protein sequences remained, with 95 RUs. The RUs were named according to the following convention: Total_number_repeats-Accession-Species-RU_number. For example, PI3-IP22_LYCES-LE-R1 represents the first repeat unit (R1) of the 3-RU (PI3) protein, IP22_LYCES (Swiss-Prot names, accession numbers and GenBank accession numbers are used whenever possible.) from *Lycopersicon esculentum* (LE). (Abbreviations for other species are: AT, *Arabidopsis thaliana*; CA, *Capsicum annuum*; LE, *Lycopersicon esculentum;* LH, *Lycopersicon hirsutum*; MC, *Mesembryanthemum crystallinum*; MT, *Medicago truncatula*; NA, *Nicotiana alata*; NE, *Nicotiana attenuate*; NG, *Nicotiana glutinosa*; NT, *Nicotiana tabacum;* OS, *Oryza sativa*; SA, *Solanum americanum;* SH, *Sorghum halepense*; SM, *Solanum melongena;* SN, *Solanum nigrum*; SP, *Solanum phureja;* ST, *Solanum tuberosum*; ZM, *Zea mays*).

The Multiple sequence alignment of 95 Pot II RUs are shown in Figure 18. The eight cysteines are fully conserved in all the 95 RUs.



```
                       *   *        *             * .**      *  .        *
  PI3-BI436259-ST-R2 KACPRNCDPHIAYSKCLRSRGKTLIYPTGCTTCCTGY-KDCYYFGKDGKFV-CEGESIEP----------------------   58
    PI3-X99095-ST-R2 KACPRNCDPHIAYSKCSRSRGKTLIYPTGCTTCCTGY-KGCYYFGKDGKFV-CEGESIEP----------------------   58
I2+-IP2X_SOLTU-ST-R2 KACPRNCDPHIAYSKCPRSGGKTLIYPTGCTTCCTGY-TDCYYFGKDGKFV-CEGESIEPKACTLECDSRLHT---------   71
  PI3-BI421162-LE-R2 KPCPLNCDPHIAYSKCPRSGGKTFIYPTGCTTCCTGY-KGCYYFGKDGKFV-CEGESDEP----------------------   58
PI2-IP27_SOLTU-ST-R2 KACPLNCDTNIAYSRCPHSEGKSLIYPTGCTTCCTGY-KGCYYFGKNGKFV-CEGESDEPKANMYPAM--------------   66
  PI2-AAO88244-SP-R2 NVCPRNCDTNIAYSKCPRSEGKSLIYPTGCTTCCTGY-KGCYYFGKNGKFV-CEGESDEPKANMYPAM--------------   66
PI2-IP2K_SOLTU-ST-R2 KACPLNCDPHIAYSKCPRSEGKSLIYPTGCTTCCTGY-KGCYYFGKNGKFV-CEGESDEPKANMYPAM--------------   66
  PI2-CAA27730-ST-R2 KACPLNCDPHIAYSKCPRSEGKSLIYPTGCTTCCTGY-KGCYYFGKNGKFV-CEGESDEPKANMYPAM--------------   66
PI2-IP25_SOLTU-ST-R2 KACPRNCDPHIAYSKCPRSEGKSLIYPTGCTTCCTGY-KGCYYFGKNGKFV-CEGESDEPKANMYPAM--------------   66
  PI2-CAA27409-ST-R2 KACPLNCDPHIAYSKCPRSEGKSLIYPTGCTTCCTGY-KGCYYFSKNGKFV-CEGESDEPKANMYPAM--------------   66
PI2-IP21_LYCES-LE-R2 NACTFNCDPNIAYSRCPRSQGKSLIYPTGCTTCCTGY-KGCYYFGKDGKFV-CEGESDEPKANMYPVM--------------   66
PI3-IP23_LYCES-LE-R2 NICPSYCDPQIAYSKCPRSEGKTIIYPTGCTTCCTGY-KGCYYFGQDGEFV-CEGESIEP----------------------   58
PI3-IP21_TOBAC-NT-R2 NVCPQFCDPDIAYSKCPRSEGETIINPTGCTTCCTGY-KGCYYFGQDGEFV-CEGESDEP----------------------   58
  PI2-AAR37362-SN-R2 NTCPLYCDGDIAYSKCPRSEGQTIIYPTGCTTCCTGY-KGCYYFSKEGEFV-CEGESIEPNVISNQ----------------   64
  PI2-AAL36458-SA-R2 KDCTFECDTQIAYSKCPRSEGKMIIKPTGCTTCCTGY-QGCYYFDQDGDFV-CEGESPEPKTTAYF----------------   64
PI2-IP2Y_SOLTU-ST-R2 KACPLNCDTNIAYSRCPRSEGKSLIYPTGCTTCCTGY-KGCYYFGTNGKFV-CEGESDEPKPYMSTA---------------   65
PI2-IP2T_SOLTU-ST-R2 KACPLNCDTNIAYSRCPRSEGKSLIYPTGCTTCCTGY-KGCYYFGTNGKFV-CEGESDEPKPYMSTA---------------   65
    PI3-X99095-ST-R1 KACTLECG-NLGYGICPRSEGSP--ENPICTNCCAGY-KGCNYYSANGTFI-CEGQSDPKNP--------------------   57
I2+-IP2X_SOLTU-ST-R1 KACTLECG-NLGYGICPRSEGSP--ENPICTNCCAGY-KGCNYYSANGTFI-CEGQSHPKNP--------------------   57
  PI3-BI436259-ST-R1 KACTKECG-NLGYGICPRSEGST--ENPICTNCCAGY-KGCNYYSANGIFI-CEGQSDPKNP--------------------   57
PI3-IP23_LYCES-LE-R1 NACTKECG-NLGYGICPGSEGSP--ENPICTNCCSGY-KGCNYYYANGTFI-CEGTSDPKNP--------------------   57
  PI2-AAR37362-SN-R1 KPCTRECG-NLGYGICPRSEGSP--ENPICTNCCSGY-KGCNYYSANGTFI-CEGSSDPKNP--------------------   57
PI2-IP25_SOLTU-ST-R1 KACTLECG-NLGFGICPRSEGSP--ENRICTNCCAGY-KGCNYYSANGAFI-CEGESDPKKP--------------------   57
  PI2-AAO88244-SP-R1 KACTLECG-NLGFGICPRSEGSP--ENRICTNCCAGY-KGCNYYSANGAFI-CEGESDPKNP--------------------   57
  PI2-CAA27730-ST-R1 KACTLECG-NLGFGICPRSEGSP--ENRICTNCCAGY-KGCNYYSANGAFI-CEGQSDPKKP--------------------   57
  PI2-CAA27409-ST-R1 KACTLECG-NLGFGICPRSEGSP--ENPICTNCCAGY-KGCNYYSANGAFI-CEGQSDPKKP--------------------   57
PI2-IP2K_SOLTU-ST-R1 KACIRECG-NLGFGICPRSEGSP--ENPICTNCCAGY-KGCNYYSANGAFI-CEGQSDPKKP--------------------   57
PI2-IP2Y_SOLTU-ST-R1 LGCTKECG-NLGFGICPRSEGSP--TNPICINCCSGY-KGCNYYSAFGRFI-CEGESDPKNP--------------------   57
PI2-IP2T_SOLTU-ST-R1 LGCTKECG-NLGFGICPRSEGSP--TNPICINCCSGY-KGCNYYSAFGDLI-CQGESDPKNP--------------------   57
PI2-IP27_SOLTU-ST-R1 KICTKECG-NLGFGICPRSEGSP--KNPICINCCSGY-KGCNYYSVFGRFI-CEGESDLKNP--------------------   57
PI2-IP21_LYCES-LE-R1 KACTRECG-NLGFGICPRSEGSP--LNPICINCCSGY-KGCNYYNSFGKFI-CEGESDPKRP--------------------   57
  PI3-BI421162-LE-R1 KPCTMECG-HLGFGICPRSQGSP--QNPICTNCCSGF-KGCNYYSVNGTFI-CEGQSDPRKP--------------------   57
  PI2-AAL36458-SA-R1 KACTRECG-HFSYGICPRSEGSP--QKPICTNCCSGY-KGCNYYSAKGDLI-CEGESDPRNP--------------------   57
PI3-IP21_TOBAC-NT-R1 KACTKECG-NFAYGICPRSQGTP--DDPICTTCCAGY-KGCNYYSANGTFI-CEGSSDPKNP--------------------   57
I3'-IP22_CAPAN-CA-R1 KPCTLNCDPRIFYSKCPRSEGNA--ENRICTNCCAGR-KGCNYYSADGTFI-CEGESDPNNP--------------------   58
I3'-IP22_CAPAN-CA-R2 KPCTLNCDPRIFYSKCPRSEASA--EQPICTNCCAGL-KGCNYYNADGTFI-CEGESDPNHP--------------------   58
 PI3'-AAF25496-CA-R1 KACPRYCDTRIAYSKCPRSEGNA--ENRICTNCCAGR-KGCNYYSVIGTFI-CEGESDPNNP--------------------   58
 PI3'-AAF25496-CA-R2 KACPRNCDTRIAYSKCPRSEGNA--ENRLCTNCCAGR-KVCNYYSADGTFI-CEGESDPNNP--------------------   58
 PI3'-AAX84035-CA-R1 KACPRYCDTRIAYSKCPRSEGNA--ENRICTNCCAGR-KGCNYYSADGTFI-CEGESDPNNP--------------------   58
 PI3'-AAX84035-CA-R2 KACPRYCDTRIAYSKCPRSEGNA--ENRLCTNCCAGR-KGCNYYSADGTFI-CEGESDPNNP--------------------   58
 PI3'-AAX84036-CA-R1 KACPRYCDTRIAYSKCPRSEGNA--ENRLCTNCCAGR-KGCNYYSADGTFI-CEGESDPNNP--------------------   58
               ruler 1.......10........20........30........40........50........60........70........80..
                       *   *        *             * .**      *  .        *
 PI3'-AAX84036-CA-R2 KACPRNCDTRIAYSKCPRSEGNA--ENRLCTNCCAGR-KGCNYYSADGTFI-CEGESDPNNP--------------------   58
  PI7-AAO85558-NE-R4 KACPRNCDGRIAYEICPRSEEKK--NDRICTNCCAGT-KGCKYFSDDGTFI-CEGESDPRNP--------------------   58
  PI7-AAO85558-NE-R5 KACPRNCDGRIAYEICPRSEEKK--NDRICTNCCAGT-KGCKYFSDDGTFI-CEGESDPRNP--------------------   58
  PI7-AAO85558-NE-R1 KACPRNCDGRIAYEICPRSEEKK--NDRICTNCCAGT-KGCKYFSDDGTFI-CEGESDPRNP--------------------   58
  PI7-AAO85558-NE-R3 KACPRNCDGRIAYEICPRSEEKK--NDRICTNCCTGT-KGCKYFSDDGTFV-CEGESDPRNP--------------------   58
  PI7-AAO85558-NE-R6 KACPRNCDGRIAYEICPRTEEKK--NDRICTNCCAGT-KGCKYFSDDGTFV-CEGESDPRNP--------------------   58
  PI7-AAO85558-NE-R2 KACPRNCDGRIAYGICPRSEEKK--NDRICTNCCAGT-KGCKYFSDDGTFI-CEGESDPRNP--------------------   58
  PI4-AAF14181-NA-R1 KACTLNCDPRIAYGVCPRSEEKK--NDRICTNCCAGT-KGCKYFSDDGTFV-CEGESDPRNP--------------------   58
    PI6-JQ2153-NA-R1 KACTLNCDPRIAYGVCPRSEEKK--NDRICTNCCAGT-KGCKYFSDDGTFV-CEGESDPRNP--------------------   58
    PI6-JQ2153-NA-R2 KACTLNCDPRIAYGVCPRSEEKK--NDRICTNCCAGT-KGCKYFSDDGTFV-CEGESDPRNP--------------------   58
  PI4-AAF14181-NA-R2 KACTLNCDPRIAYGVCPRSEEKK--NDRICTNCCAGT-KGCKYFSDDGTFV-CEGESDPRNP--------------------   58
  PI4-AAF14181-NA-R3 KACPRNCDPRIAYGICPLSEEKK--NDRICTNCCAGK-KGCKYFSDDGTFV-CEGESDPRNP--------------------   58
    PI6-JQ2153-NA-R3 KACPRNCDPRIAYGICPLAEEKK--NDRICTNCCAGK-KGCKYFSDDGTFV-CEGESDPKNP--------------------   58
    PI6-JQ2153-NA-R4 KACPRNCDGRIAYGICPLSEEKK--NDRICTNCCAGK-KGCKYFSDDGTFV-CEGESDPKNP--------------------   58
    PI6-JQ2153-NA-R5 KACPRNCDGRIAYGICPLSEEKK--NDRICTNCCAGK-KGCKYFSDDGTFV-CEGESDPRNP--------------------   58
  PI6-AAF18451-NG-R1 KACPRNCDRRIAYGICPRSEEKN--NDRICTNCCAGM-KGCNYFSADGTFV-CEGESGPKNP--------------------   58
  PI8-AAF18450-NG-R1 KACPRNCDRRIAYGICPRSEEKN--NDRICTNCCAGM-KGCNYFSADGTFV-CEGESGPKNP--------------------   58
  PI6-AAF18451-NG-R2 KACPRNCDGRIAYGVCPRKEEKK--NNWICTNCCAGT-KGCNYFSADGTFV-CEGESDPRNP--------------------   58
  PI8-AAF18450-NG-R4 KACPRNCDGRIAYGVCPRKEEKK--NNWICTNCCAGT-KGCNYFSADGTFV-CEGESDPRNP--------------------   58
  PI8-AAF18450-NG-R2 KACPRNCDGRIAYGVCPRKEEKK--NNWICTNCCAGM-KGCNYFSADGTFV-CEGESGPKNP--------------------   58
  PI8-AAF18450-NG-R3 KACPRNCDGRIAYGVCPRKEEKK--NNWICTNCCAGM-KGCNYFSADGTFV-CEGESGPKNP--------------------   58
  PI6-AAF18451-NG-R3 KACPRNCDPRISDEICPRSEEKN--NNQICTNCCAGT-KGCNYFSADGTFV-CEGEYDPKNP--------------------   58
  PI8-AAF18450-NG-R5 KACPRNCDPRISDEICPRSEEKN--NNQICTNCCAGT-KGCNYFSADGTFV-CEGEYDPKNP--------------------   58
  PI6-AAF18451-NG-R4 KACPRNCDPRISYGVCPRSEEKN--NNQICTNCCAGT-KGCNYFSADGTFV-CEGEYDPKNP--------------------   58
  PI8-AAF18450-NG-R6 KACPRNCDPRISYGVCPRSEEKN--NNQICTNCCAGT-KGCNYFSADGTFV-CEGEYDPKNP--------------------   58
  PI6-AAF18451-NG-R5 KACPRNCDGRIDYGVCPRSEEKK--NNQICTNCCAGT-KGCNYFSADGTFV-CEGESDPRNP--------------------   58
  PI8-AAF18450-NG-R7 KACPRNCDGRIDYGVCPRSEEKK--NNQICTNCCAGT-KGCNYFSADGTFV-CEGESDPRNP--------------------   58
  PI4-AAF14181-NA-R4 KACPRNCDGRIAYGICPLSEEKK--NDRICTNCCAGK-KGCKYFSDDGTFI-CEGESEYASKVDEYVGEVENDLQKSKVAVS   78
    PI6-JQ2153-NA-R6 KACPRNCDGRIAYGICPLSEEKK--NDRICTNCCAGK-KGCKYFSDDGTFI-CEGESEYASKVDEYVGEVENDLQKSKVAVS   78
  PI6-AAF18451-NG-R6 KACPLNCDGRIAYEICPRTEEKK--TNQICTNCCAGT-KGCNYFNANGTFI-CEGESEYVNKVENDLEKSKVLFSKS-----   73
  PI8-AAF18450-NG-R8 KACPLNCDGRIAYEICPRTEEKK--TNQICTNCCAGT-KGCNYFNANGTFI-CEGESEYVNKVENDLEKSKVAVF-------   71
  PI7-AAO85558-NE-R7 KACPRNCDERIAYGICPRTEEKK--NNQICTNCCAGT-KGCNYFSANGTFI-CEGESEYVSKVDEYVHEVENDLQKSRVAVS   78
PI3-IP22_LYCES-LE-R1 KACTKECDTRIDFGICPLLETKR--VEGLCTNCCAGK-KGCKYFSKDGTYI-CDGESEWVSEKNNNLK--------------   64
PI3-IP22_LYCES-LE-R2 KACTKECDTRIDFGICPLLETKR--VEGLCTNCCAGK-KGCKYFSKDGTYI-CDGESEWVSEKDNNLE--------------   64
PI3-IP22_LYCES-LE-R3 KDCTKECDTRIDFGICPLLETKR--VKGLCTNCCAGK-KGCKYFSADGTYI-CDGESEWVSEGENDLQKSNVAIS-------   71
  PI3-BI436259-ST-R3 KACTLECDSRVSYMTCPSSGMAK--LNQVCVNCCTAR-EGCKLYGHDGSLI-CTGEPQSYISTT------------------   60
  PI3-BI421162-LE-R3 KACTMECDPRVAYMICPSSGLAK--LSQVCVNCCTAG-DGCKLYGYDGSLI-CTGEPQSYISTA------------------   60
PI3-IP23_LYCES-LE-R3 KGCTKECDPRVAYMTCPSSGLAK--LNQVCVNCCSAG-EGCKLYDNDGSLL-CTGEPQSISTA-------------------   59
    PI3-X99095-ST-R3 KACTLECDSRVAYMTCPSSGLAK--LNQVCVNCCTAG-EGCKVYGYDGSLI-CTTGEPAQGNISTI----------------   62
PI3-IP21_TOBAC-NT-R3 KSCTTECDPRVATISCPFSGLVK--INQECINCCNAD-KGCELYDNDGSLI-CTGGEPQSAA--------------------   58
  PI1-AY105802-ZM-R1 VACPQFCLD-VDYVTCPSSGSEK--LPASC-NCCMTP-KGCTLHLSDGTQQTCS----------------------------   49
               ruler 1.......10........20........30........40........50........60........70........80..
                       *   *        *             * .**      *  .        *
  PI1-AI724716-SH-R1 VPCPQYCLE-VDYVTCPSSGSEK--LPARC-NCCLAP-KGCTLHLSDGTQQTCS----------------------------   49
 PI1-NP_177351-AT-R1 IACPLYCLQ-VEYMTCPSSGADK--LPPRC-NCCLAP-KNCTLHLSDSTTIHCSK---------------------------   50
  PI1-BE033692-MC-R1 TICLQICYTDADYMTCPSTGYKK--LNPAC-NCCIGG-PGCSLYRANGEQI-C-----------------------------   48
  PI1-BE033392-MC-R1 TACHEICYVDAAYMTCPSTGYQH--LAPKC-NCCEGG-PGCSLYRANGEQI-C-----------------------------   48
  PI1-AC096689-OS-R1 KFCPQFCYDGLEYMTCPSTGSQH--LKPAC-NCCIAGEKGCVLYLNNGQVINCT----------------------------   51
  PI1-BE942349-MT-R1 KICPQFCYDSVAYMTCPSSGDQH--LTPKC-NCCLAS-TGCILYEADGTPI-CTAD--------------------------   51
  PI1-BE033653-MC-R1 QICPQICYIEAAYMLCPSTGDQQ--LTPEC-NCCEAG-PGCSLYEADGKKI-C-----------------------------   48
  PI1-BI434643-ST-R1 QICPQFCEPNVDYMTCSSSGSTI--LRPTCINCCQARGRGCQLFRRDGSAI-CN----------------------------   51
  PI1-CO516657-MS-R1 KICPQFCYDSVAYMTCPYTGDKH--LDPAC-NCCIAS-TGCVLYEADGTRI-CTAD--------------------------   51
  PI1-BE943304-MT-R1 KICPQFCYDSVAYMTCPSSSDQH--LTPKC-NCCLAS-TGCILYEADGTPI-CTAD--------------------------   51
  PI1-CN847229-PG-R1 RACPSYC-LDVNYMTCPSSGEQK--LEPDC-NCCIAP-KGCTLHLSDGTSMFCSKS--------------------------   51
  PI2-AW616253-LH-R1 KACTKECDGRISYGICPVLQSKR--VDGICINCCSGI-IGCKYFSKDYTFI-CDGESKWFGE--------------------   58
  PI2-AW616253-LH-R2 KACTKECDPRIDIGICASEQTKK--VDALCTNCCAGK-KGCNYFDENGTFI-CERESGGETTLQKSNLAIS-----------   67
               ruler 1.......10........20........30........40........50........60........70........80..
```

Figure 18 Multiple sequence alignment of 95 Pot II RUs. Full conserved residues are

marked with “*” and highly conserved residues by “.” The reactive site was marked by arrows.

Sequence Logo representation of the consensus sequence of the 95 RUs from the entire Pot II family was shown in Figure 19, with the eight Cys residues fully conserved. Besides these, other residues that are highly conserved are two Gly residues and a Pro residue (marked by arrows in Figure 19), probably having important roles in stabilizing the 3D structure of the protein.

The degree of conservation of the amino acid sites of Pot II RUs were estimated by a Maximum Likelihood method (Armon *et al.* 2001) and mapped to a reference 3D structure (1CE3) to identify functionally important regions by the program ConSurf (Glaser *et al.* 2003). The result was shown in Figure 20 below.

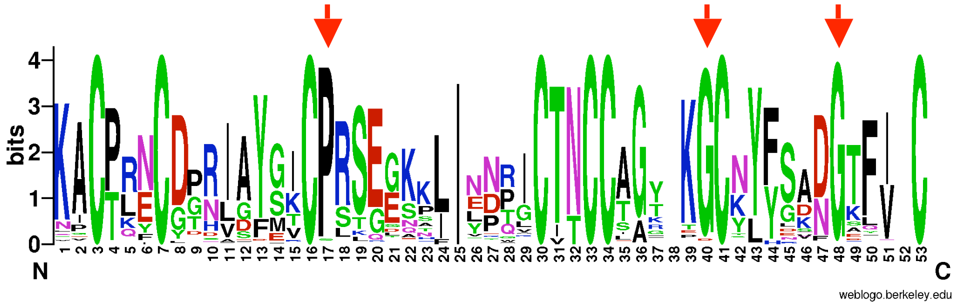


Figure 19 Sequence Logo representation of the consensus sequence of the 95 RUs from the entire Pot II family. The highly conserved residues besides the eight cysteines were marked by arrows.

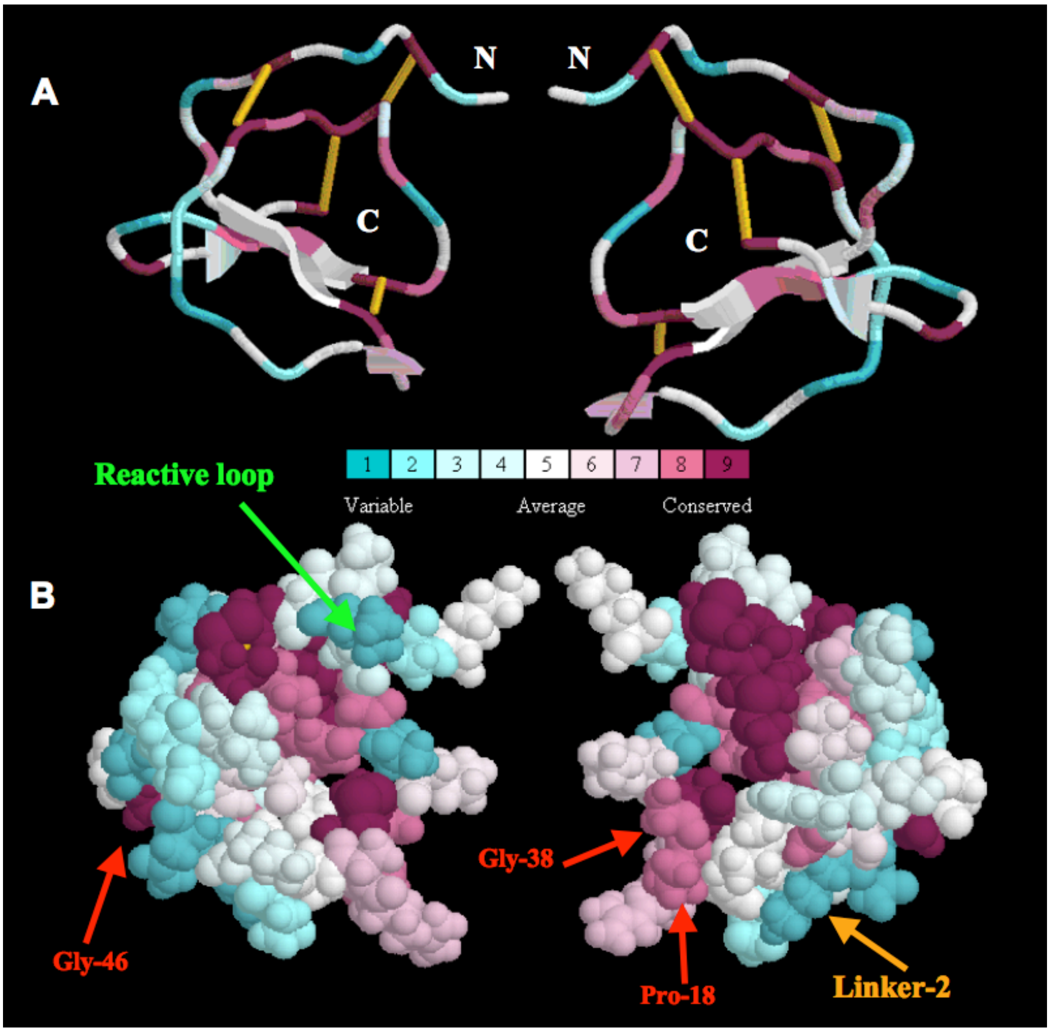


Figure 20 Residue conservation analysis for the Pot II family RUs from ConSurf,

mapped onto the structure, 1CE3 (residues K2-C50). LHS and RHS are different views of the same structure, rotated by 180°, in (A) ribbon and (B) CPK representations. Residues are shaded from cyan (highly variable) through white (moderate conservation) to purple (highly conserved).

Figure 20 show that distinct regions in the RUs of Pot II PIs have very different conservation degrees. Besides the eight fully conserved cysteines as structural scaffold in the core region, a few highly conserved residues are also important for maintaining the fold, such as Pro-18, Gly-38 and Gly-46 (numbering according to 1CE3). The detailed analysis reveals that they belong to three β-turns, respectively. For example, the *i+3* position of a type I β-turn is favored by a Gly residue, which is Gly-46, in 1CE3. Its ϕ and ψ angles (80.3° and 63.7°, respectively) falls into the region that is not favored by other residues, and makes it hard to be replaced by other residues without distorting the fold. These 11 residues including the eight cysteines, are structurally important residues. Unlike most globular proteins, the reactive loop in this domain is highly variable. The variability of the reactive loop may allow the inhibitor to target a variety of different proteinases from invading organisms efficiently. The two linker regions between the H- and the L-fragments (Figure 16), are also hypervariable which suggests that they are less critical for the functionality of the Pot II domain. 1CE3 has only linker region 2 (Linker-2, shown in Figure 20) and does not have the linker region 1, which is present in 4SGBI.

4.3.4 Phylogenetic Analysis of Pot II Family

To investigate the evolution of Pot II family genes, the phylogenetic tree was constructed using the Neighbor-Joining method (Saitou and Nei 1987b). The taxa in the tree can be clustered into several clades, by repeat number and species. All single-

RU PIs cluster into one group, and they are widely distributed in non-solanaceous plants. They are more distantly related to other members of the Pot II family and are more likely the ancestral single domain Pot II proteins. With only one RU, the sequence and the structural units are identical, with the L-H topology of 1CE3. All these single-domain PIs were defined as outgroup and the tree was rerooted.

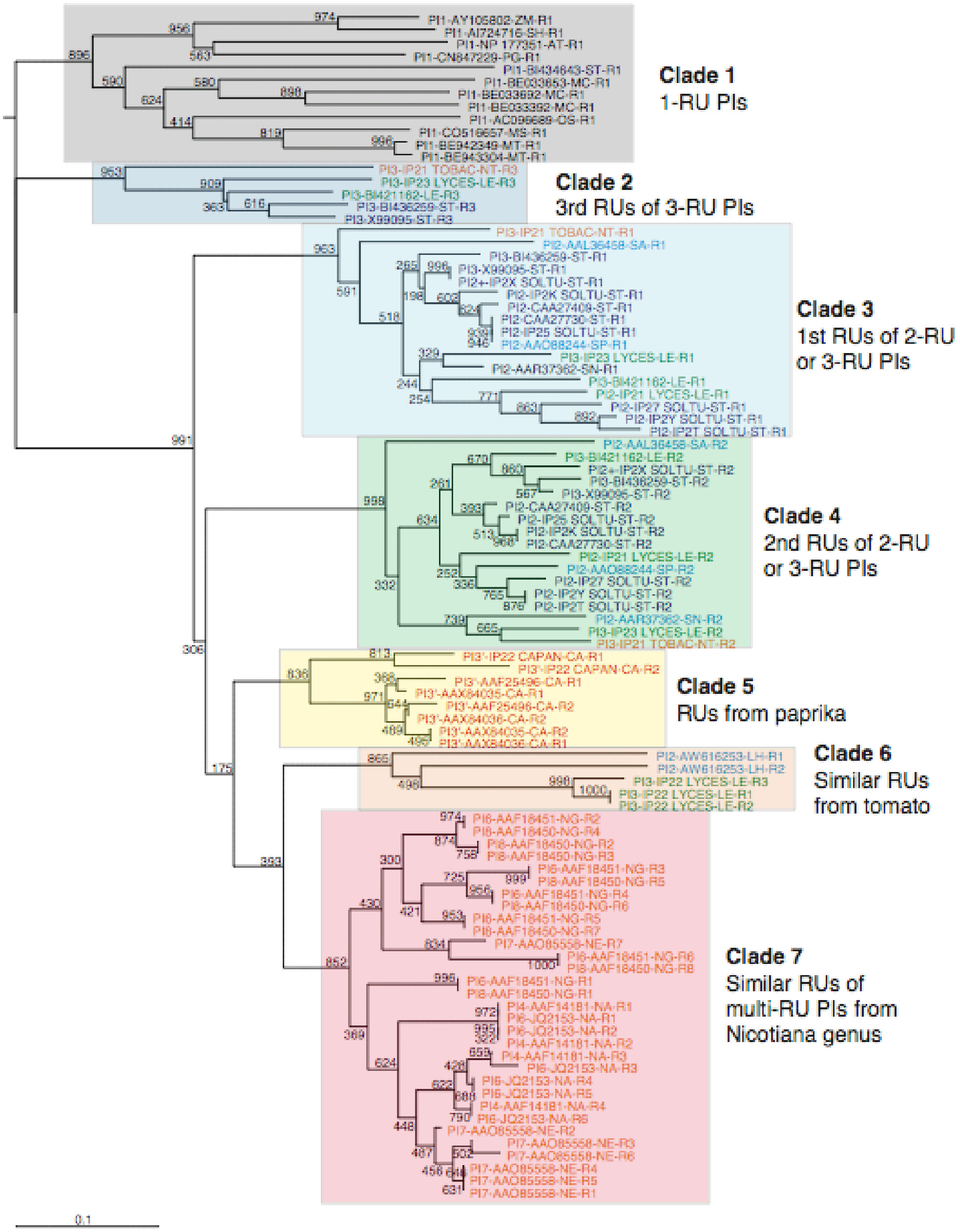


Figure 21 Phylogenetic tree of Pot II PIs repeat units. PIs from different species were colored into different colors. Green, tomato; dark blue, potato; red, paprika; orange, *Nicotiana* genus; blue, *Solanum* genus (except potato and tomato); black, non-solanaceous plants.

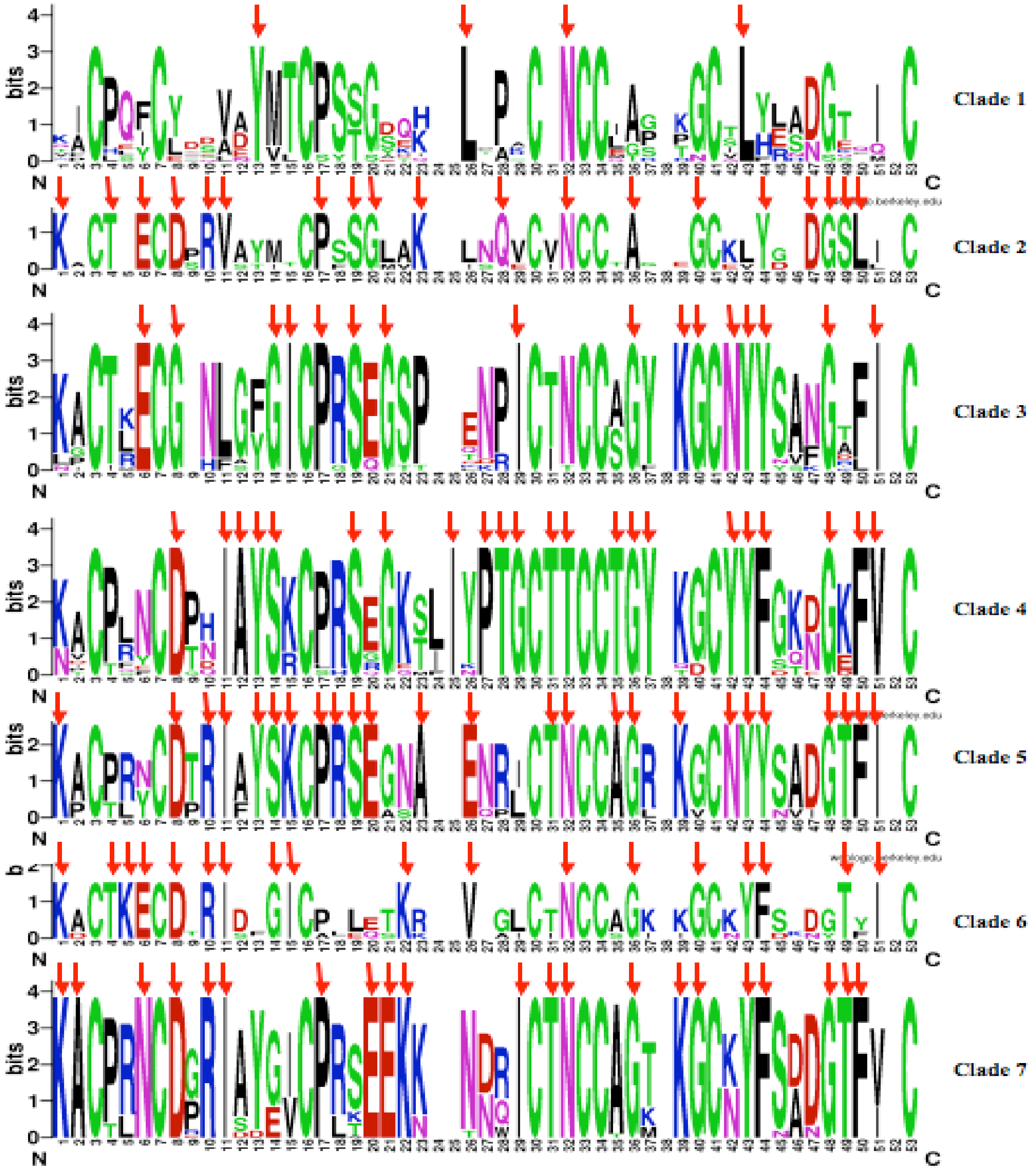


Figure 22 Clade-wise Sequence Logo representation of the consensus sequences for each clades. The arrows make out the full conserved residues except the cysteine residues.

Figure 21 shows the inferred phylogenetic tree of 95 Pot II RUs. All RUs are clustered into clades, according to repeat number, species or total RU number. This clustering of RUs within each clade is strongly supported by the high bootstrap proportions (BP) where the relative positions between clades is tentative because their BP values are low. Clade 1 contains all (12 taxa) single-RU Pot II PIs, which exist in a wide range of species and are more likely the ancient genes in Pot II family. The Sequence Logo representation of the consensus sequences (Figure 22) showed the single-RU PIs are quite diverse. The functionality or inhibitor activity of these genes is unknown because of the lack of experimental information. Clade 2 (5 taxa) comprises the third RUs of 3-RU PIs while Clade 3 (17 taxa) and 4 (17 taxa) consist of the first and second RUs of 2-RU and 3-RU PIs, respectively. Most of RUs in Clade 2, 3 and 4 are from *Solanum* genus. Clade 5 includes 8 taxa from paprika, and the repeat unit sequences in this clade are H-L type, which is different with RUs from all other members of Pot II family. Clade 6 (5 taxa) contains one 2-RU and one 3-RU PIs from *Solanum* genus. Clade 7 (31 taxa) includes 4-RU, 6-RU, 7-RU and 8-RU PIs from *Nicotiana* genus.

There are mainly three features observed in the conservation patterns (Figure 21):

(1) RUs with the same repeat numbers are most similar. The 2-RU and 3-RU PIs from the *Solanum* genus (Clade 2, 3 and 4) have 17 sequences, from 7 species with total 39 RUs, and are the largest group in this family. Here, the first RU clusters into one group as do the second RU and the third RU. This suggests the RU tandem duplication events happened before the speciation, although this level of sequence similarity cannot be detected at the DNA

sequence level between different repeats.

(2) Clade 5, 6, 7 contain repeats that are striking similar to each other within the same genes. The similarity is even clearly detectable at the DNA level. Such pattern cannot be explained by purifying selection since the domain duplications usually loose the functional constraints and allow more mutations. The remarkable similarity suggests the existence of concerted evolution which usually can be resulted by unequal crossing over and gene conversion (Dover 1982; Schlotterer and Tautz 1994; Santoyo and Romero 2005).

(3) In Clade 5, RUs from paprika is very different to other members of the *Solanacae* species. Unlike all the other groups, the RUs of the Pot II inhibitor from *Capsicum annuum* are of the H-L type. The sequence repeat is thus identical to the structural repeat observed in potato and tomato and in *Nicotiana* (H-L type in Figure 16) and has no N- and C-terminal sequence segments, which form the "bracelet" link domain in other multi-RU PIs (H+L type in Figure 16). As each domain adopts the H-L domain topology, multiple-domain PIs from *Capsicum annuum* are likely to adopt tandem structural domains with a "beads-on-a-string" domain organization, which is different from all other multiple-domain PIs in Pot II family. Strong sequence similarity exists in this cluster at both protein and nucleotide sequence levels.

4.3.5 Analysis of Selective Pressure

4.3.5.1 Site-based Analysis of Selective Pressure

Codon substitution models of were used to analyze Pot II genes to identify amino acid sites under diversifying selection. The models used the nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) as an indicator of selection pressure and allowed the ratio to vary among sites. The ω ratio of a site <1 indicates that the nonsynonymous mutations at this site are deleterious and the site is under purifying selection while ω >1 suggests that the nonsynonymous mutations at this site are beneficial and the site should be under positive selection.

The results of site-based analysis of Pot II genes were summarized in Table 14 and Table 15. *p* is the number of parameters in each model and *l* is the Likelihood values estimated under each model.

Table 14 Likelihood values and parameter estimates for Pot II genes

Models	***p***	***l***	**kappa**	**d_N/d_S**	**Estimates of parameters**	**Positive Selected Site**
M0 (one-ratio)	1	-3281.13	1.706	0.262	ω=0.262	None
M1 (NearlyNeutral)	2	-3219.57	1.986	0.551	p_0=0.513, ω_0=0.126 p_1=0.487, ω_1=1.000	Not Allowed
M2 (PostiveSelection)	4	-3201.55	2.045	0.714	p_0=0.499,ω_0=0.128 p_1=0.480, ω_1=1.000 p_2=0.021, ω_2=8.001	Site 5
M3 (discrete)	5	-3163.57	1.762	0.372	p_0=0.363, ω_0=0.041 p_1=0.616, ω_1=0.420 p_2=0.021, ω_2=4.621	Site 5
M7 (β)	2	-3169.60	1.745	0.323	p=0.525, q= 1.095	Not Allowed
M8 (β and ω)	4	-3156.75	1.791	0.387	p_0=0.979, (p_1=0.021) p=0.599, q=1.450, w=4.791	Site 5

Table 15 Likelihood Ratio Test Statistics (2Δ*l*)

Comparison	**2Δ*l***	**d.f.**	**$\chi^2_{1\%}$**	***p* value**
M0 (one-ratio) vs. M3 (discrete)	2×[-3163.57–(-3281.13)]= 235.12	4	13.28	<0.0001
M1 (NearlyNeutral) vs. M2 (PostiveSelection)	2×[-3201.55–(-3219.57)]= 36.04	2	9.21	<0.0001
M7 (β) vs. M8 (β and ω)	2×[-3156.75-(-3169.60)]= 25.70	2	9.21	<0.0001

Table 14 shows the parameters estimated under variable selective pressure among sites using the unrooted tree topology of Figure 21 without the outgroup (PI1, single-RU Pot II genes). The average ω ratio ranges from 0.32 to 0.38 among all but the worst-fitting models. The Likelihood Ratio Test (LRT) statistics (Table 15) suggested the highly variable ω ratio among amino acid sites. For example, the model of one ω ratio for all sites (M0) is rejected by a big margin when compared with model M3 (discrete), which allows for three classes of sites with different ω ratios. The LRT statistic for this comparison is 235.12, much greater than critical values from a χ^2 distribution with d.f. = 4. The discrete model (M3) suggests a small proportion of sites (p_2=2.1%) under positive selection, with $\omega_2 = 4.621$. This models fits the data significantly better than M0 (one-ratio) or M1 (NearlyNeutral). Similarly, Model M8 (β and ω) also suggests 2.1% of sites under diversifying selection with $\omega_1 = 4.791$. The LRT statistic for comparing M7 (β) and M8 (β and ω) is 25.70. The P-value for this comparison is 0.1×10^{-4}, in comparison with the χ^2 distribution with d.f. = 2. M7 is thus rejected in favor of M8. In sum, among all the models tested, all models designed to detect positive selection sites (M2, M3 and M8) were significantly better than their counterpart null hypothesis (M0, M1 and M7), which provide consistent evidence for the presence of heterogeneous selection pressure among amino acid sites within Pot II domains.

Furthermore, all models allowed positive selection (M2, M3 and M8) converged to the same site, site 5. And site 5 had a high posterior probability (above the 99% level) of being in the positively selected class in all models allowed positive selection (M2, M3 and M8).

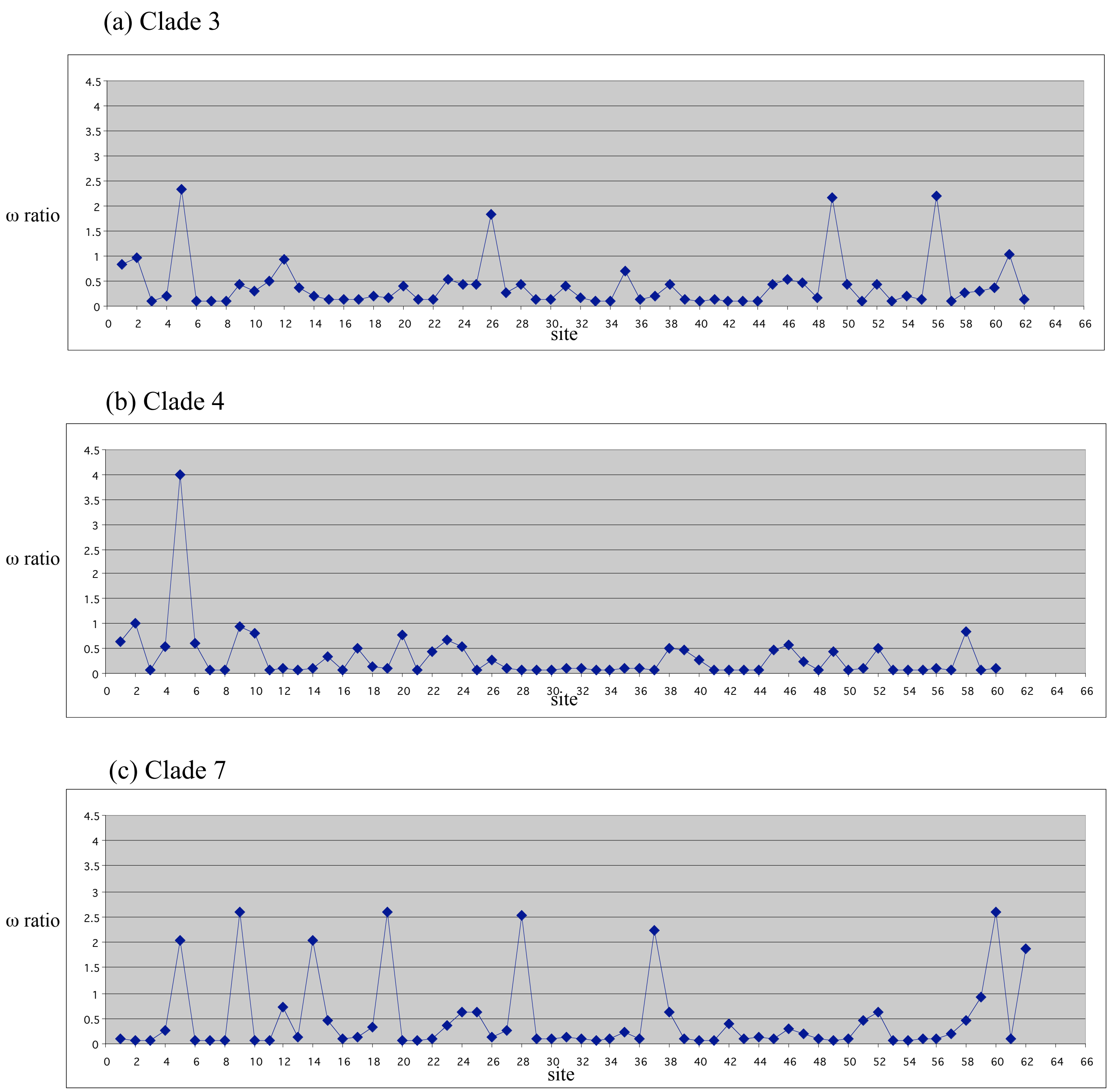
Statistics analyses of variation of ω among sites provide strong evidence of the positive selection. Interestingly, the positively selected site 5 locates at P_1 position of the reactive site of Pot II domains according the nomenclature of the Schechter and

Berger (Schechter and Berger 1968). For standard mechanism, canonical proteinaceous PIs of serine proteinases, the specificity of the inhibitors is determined, at least in part, by a single residue at the P_1 position (Laskowski and Kato 1980). In Pot II PI structures, the P_1 residue contribute the largest number of contacts (Schirra and Craik 2005). Therefore, the hypervariability and positive selection of the P_1 residue in reactive site can be easily understood since they allow the Pot II inhibitors to provide inhibition activity to a wide range of proteinases which help *Solanaceae* to combat pathogenic attacks.

4.3.5.2 Clade-wise site-based analyses in selective pressure

Clade-wise site-based analyses in selective pressure were also conducted on Clade 3 (1st RUs of 2-RU or 3-RU PIs), Clade 4 (2nd RUs of 2-RU or 3-RU PIs) and Clade 7 (Similar RUs of multi-RU PIs from Nicotiana genus) in order to detect the short episode of positive Darwinian selection within each clades.

For all three clades, LRT tests support the existence of positive selected sites, but selective pressures among sites are quite different between Clade 3, Clade 4 and Clade 7. For Clades 3, 4 and 7 separately, the approximate posterior mean of ω ratio at each site was plotted (Figure 23).


(a) Clade 3

(b) Clade 4

(c) Clade 7

Figure 23 Approximate posterior mean of the ω ratio by Bayes Empirical Bayes (BEB) method for each site calculated under model M8 (β and ω) for the (a) Clade 3 (1^{st} RUs of 2-RU or 3-RU PIs); (b) Clade 4 (2^{nd} RUs of 2-RU or 3-RU PIs); (c) Clade 7 (Similar RUs of multi-RU PIs from Nicotiana genus).

Figure 23 shows that the majority of amino acid sites in Clade 3 and Clade 4 are under purifying or neutral selection while Clade 7 has more amino acid sites under positive selection. In Clade 3 and Clade 4, site 5 (P1 site of reactive loop) was identified as statistically significant positive selected sites by all models (M2, M3 and M8), which is consistent with the previous analysis. While in Clade 7, all models support strong positive selection over site 19, which is the ending residue after the proteolytic processing removing the Linker 1 region (highly conserved linker "EEKKN" in multi-RU Pot II PIs from Nicotiana genus).

Such differences in variable selective pressure between Clade 3 and Clade 4 and Clade 7 may be due to the number of RUs. For two-domain Pot II PIs, the two domains can bind to two proteinases simultaneous without steric interference since the two binding sites are at the opposite ends of two inhibitor domains (e.g. the bound form of TI-II) (Barrette-Ng *et al.* 2003c). For Pot II PIs with more than two domains, it becomes more and more difficult for each domain to bind a proteinase without steric hindrance. Heath and co-workers reported that the six-domain precursor NA-PI has stoichiometry of only 2.6 trypsin molecules (Heath *et al.* 1995). So the efficiency of proteolytic processing of multi-domain PIs may provide evolutionary advantages by performing better inhibitory activity. This may explain why in Clade 7 the residue at the boundary of the on the cleavage sites is under positive selection.

4.3.6 Linker region analyses of Pot II genes

Schirra and Craik proposed that linker regions particularly the EEKKN linker (Linker 2 in Figure 16) determined the circular permutation of multi-RU Pot II genes in a recent review (Schirra and Craik 2005). To validate this hypothesis and investigate the features and patterns of linker regions, systematic linker region analyses were carried out on the Linker 1 (L1) and Linker 2 (L2) in each clade. The results are summarized in Table 16.

Table 16 The sequence patterns and the extent of conservation of the linker regions.

Clade	L1 (important for structural repeats)		L2 (important for sequence repeats)		%L1/%L2
All Clades after removing the ending RUs	EGESDPxNP	89%	PRSEexkxxxnxI	50%	1.8
Clade 1, single-RU	xxx------	0%	PsSGxxx--LxPx	42%	0.0
Clade 2, the third RU of 3-RU	-GEPqsxxx	44%	PsSGlaK--lnQv	62%	0.7
Clade 3, the first RU of 2-RU and 3-RU	EGxSDPKnP	83%	PRSEGSP--eNPI	81%	1.0
Clade 4, the second RU of 2-RU and 3-RU	EGESdEPkx	78%	PRSeGKxlIYPTG	85%	0.9
Clade 5, RUs from paprika	EGESDPNNP	100%	PRSEgnA--Enrx	62%	1.6
Clade 6, similar RUs	dgESxwxxe	44%	pxlxxKr--Vxgl	35%	1.2
Clade 7, Similar RUs of multi-RU PIs from Nicotiana genus	EGESDPxNP	89%	PRsEEKK--NdxI	69%	1.3

For an estimation of % conservation, we have used a simple metric with 1 for fully conserved, 0.5 for partly conserved and 0 for unconserved positions.

From Table 16, the ranking of L1 conservation is 5 > 7 > 3 > 4 > 6 = 2 (>>1). This clearly reflects the tendency to nucleate permuted domains as structural units, with clade 5 showing maximum propensity, closely followed by clade 7.

For L2 conservation, the ranking (4 > 3 > 7 > 5 = 2 (> 1) > 6) indicates propensity for domain duplication at the sequence level: obviously clade 7 has greatest tendency in this respect. What is surprising is clade 6 from tomato, has a lower level of conservation than clade 1, which is made up of many organisms. This is an artifact due to the low number of domains in this clade, made up entirely of a 3-RU PI and a 2-RU PI, each of which is remarkably conserved.

L1 and L2 conservation need to be considered together, in order to understand the subtle interplay between sequence and structural repeat units in this protein family. Preference for structural repeats over sequence repeats may be measured by taking the ratio of %L1/%L2. Here the clades are in the order:

%L1/%L2: 5 > 7 > 6 > 3 > 4 > 2 (>> 1)

Thus the third RU of 3-RU PIs and the second RU of 2- and 3-RU PIs (clades 2 and 4) show more L2 than L1 conservation. In fact, it is probable that clade 2, expressed alone or in combination with the preceding domain from clade 4 might adopt the L+H topology of 1CE3. However, the first RU of these PIs (clade 3) slows a slightly higher L1 conservation, which tilts the structure towards H+L over L+H. Clades 6, 7 and 8 show progressively enhanced preference for L1 conservation over L2, shifting the equilibrium towards conserved structural repeat units of the H+L type.

Considering all clades *in toto*, the family has evolved to preferentially adopt H+L topology over L+H, culminating in clade 5 with sequence repeats that mirror the structural repeat unit of 4SGBI. This is supported by the rapidly evolving PIs as well as those of more recent origin with %L1/%L2 ratios > 1.0 (clades 5-7). The older proteins (clades 2-4) represent the cross-roads when sequence and structural repeats are vying for supremacy: the obvious choice of H+L topology is suggestive of pressures other than evolution, such as evasion of protease degradation events.

In the creation of the engineered protein of *Nicotiana alata* (1CE3), L1 and L2 segments were swapped, creating a sequence with <75% (1/1.3) probability of adopting the H+L structure over that of L+H, leading to the observed structure 1CE3.

4.3.7 Codon usage analysis of Pot II genes

Codon usage analyses were carried out on Pot II genes to evaluate whether there is codon usage bias and whether such bias is advantageous or not. The codon usage tables were calculated using CUSP module of EMBOSS package. The derived codon usage table for Pot II genes was compared with codon usage table of *Nicotiana* *tabacum* by Graphical Codon Usage Analyser (GCUA, http://www.gcua.de) and the mean difference between codon usage tables are also calculated by GCUA. The result is shown in Figure 24.

Since a general codon usage table for *Solanaceae* family is not available, *Nicotiana tabacum* was chosen as a representative organism for *Solanaceae* family for the following considerations:

(1) The difference between codon usage tables from organisms from *Solanaceae* family is subtle. For example, the mean difference of codon usage table between *Nicotiana tabacum* and *Solanum tuberosum* is only 1.7%, and the differences between *Solanaceae* plants we observed so far are all less than 2.5%. So the selection of a representative organism will not affect the analysis results significantly.

(2) The number of CDS and codons used for the codon usage calculation is very important since a small sample size will possibly introduce gloss statistics of codon usage frequency. The codon usage table of *Nicotiana tabacum* from Codon Usage Database was calculated by a large number of genes (1343 CDS and 513,897 codons, the largest dataset in plants from *Solanaceae* family), so the frequency of codon usage in this table should be quite reliable.

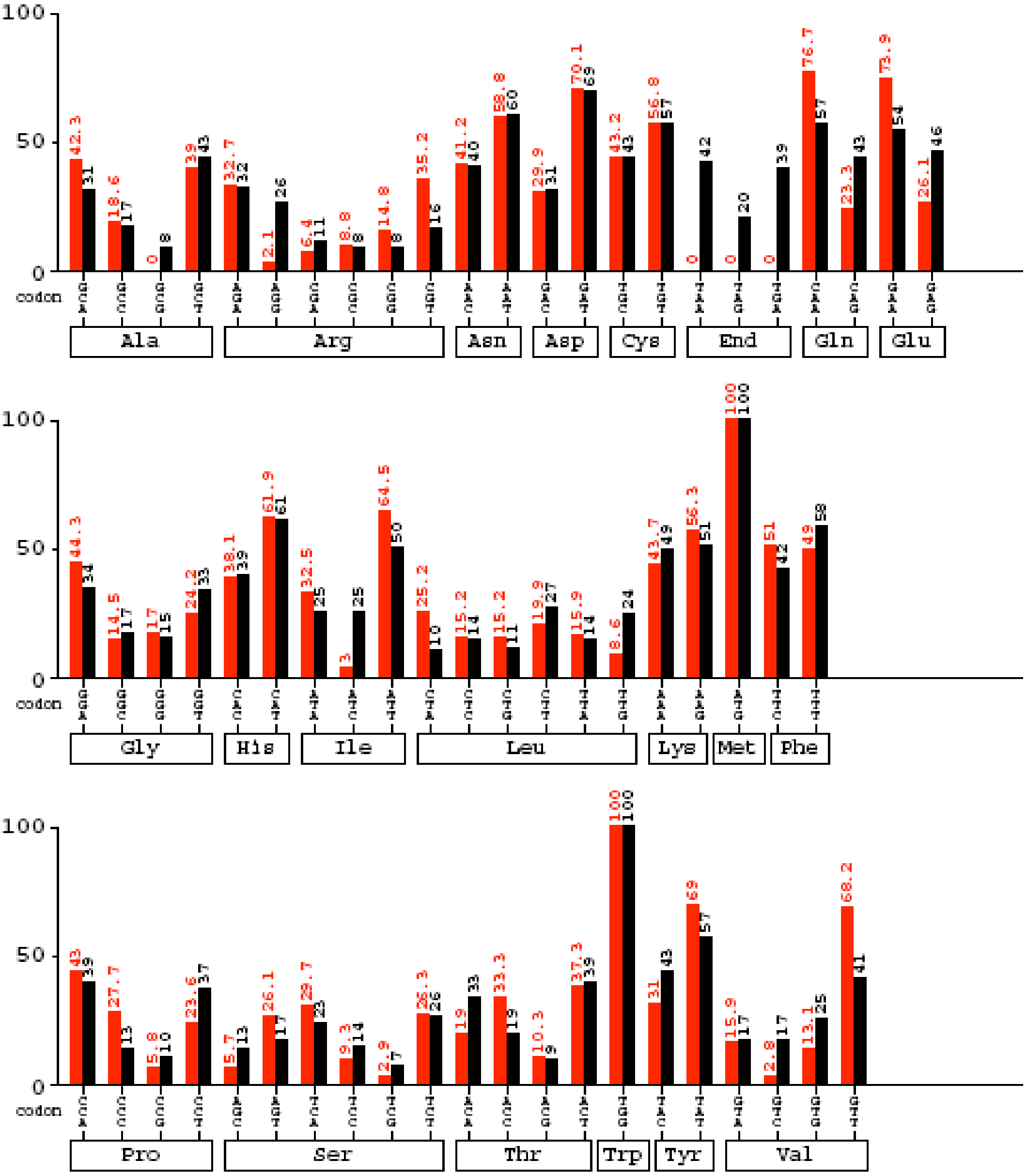


Figure 24 Codon usage tables comparison between Pot II genes and *Nicotiana tabacum*. Columns of Pot II genes are in grey (left) while columns of *Nicotiana tabacum* in black (right).

The results of codon usage table comparison (Figure 24) showed the difference of codon usage between Pot II genes and *Nicotiana tabacum*. The mean difference between two tables is in moderate level (9.18%). For most residues, the codon usages are almost the same. But there are significant codon usage differences on Gln, Glu, Ile, Tyr and Val. Interestingly, for all these residues, the codon usages in Pot II genes apparently tend to use the codons which are used more frequently in *Nicotiana tabacum,* and avoid to use the low-frequency codons. For example, ILE is encoded by three codons: ATA, ATC and ATT. The codon usage frequency for these three codons in *Solanum tuberosum* is 25%, 25% and 50%, respectively, while the frequency in Pot II genes are 32.5%, 3% and 64.5%. And these frequencies are based on a reasonable number of codon observations (335 codons for ILE). These frequencies are also consistent with the tRNA gene abundance in plant. Since the complete genome data for *Nicotiana tabacum* is unavailable, the number of tRNA genes in *Arabidopsis thaliana* was used as a reference. In *Arabidopsis thaliana*, the numbers of tRNA genes (identified so far) for codons ATA, ATC and ATT are 5, 0 and 19 copies. The codon usage frequencies for these residues are obviously advantageous since it suggests that Pot II genes utilize abundant tRNA subpopulations that facilitate the rapid expression and response to wounds and pest infestation. The codon usage tables comparison were also conducted on individual organism of Pot II genes, e.g. the codon usage table of Pot II genes from *Nicotiana glutinosa* were compared to *Nicotiana glutinosa* codon usage tables. These results are very similar to the above observation except there are apparently some fake biases due to the small sample size.

4.4 Conclusion

Chapter 4 described systematic analyses of Pot II family using a range of bioinformatics analysis tools, leading to several interesting findings:

(1) The database search has identified new putative single-RU Pot II PIs from non-solanaceous species such as *Arabidopsis thaliana, Oryza sativa* and *Zea Mays,* which are representative of the ancestral Pot II domain, synthesized by Craik *et al.* (Scanlon *et al.* 1999) with the 3D structure 1CE3, having L-H topology.

(2) The gene structure analysis reveals conserved features including: (a) similar exon/intron organization; (b) conserved splice phase and splice motif; (c) conserved Gly residues across splice sites.

(3) The protein sequence alignment suggests the consensus sequence of Pot II family to be:

CX(3)CX(7,8)CPX(9,12)CX(1,2)CCX(4,5)GCX(6)GX(3,4)C,

with C, G and P representing Cys, Gly and Pro residues, X is any residue and (m,n) represents residue repeat numbers ranging from m to n, where m and n are integers.

(4) Based on observed domain organization or all known sequences in Pot II family, there is a propensity in Pot II PIs domain's topology to adopt the H-L topology (representative structure being 4SGBI). Given that the repeat unit for most multiple-RU Pot II PIs is of the L-H type, such PIs will therefore fold into contiguous permuted structural domains, linked by bracelet-like structures formed by the N- and C-terminal segments from the first and the last repeat units.

(5) For PIs from paprika alone, the repeat unit is of the H-L type, so that multiple-domain PIs from paprika should adopt a simple tandem permuted domain architecture, with no linking bracelet structure, which is unique to the Pot II PI family. Naturally isolated L-H type single-domain PIs can only be derived from single-RU genes, which are present in Clade 1, so far recognized in rice, maize, etc.

(6) The degree of conservation for each residue in the Pot II PIs repeat units was evaluated and mapped onto the molecular surface of the structure for the putative ancestral protein, 1CE3. The result shows that different regions of the protein sequences, have very different mutation rates. Eight fully conserved cysteines form the scaffold in the protein core, with the reactive loop and linker region being highly variable. The rapid mutation of the reactive site is consistent with the PIs possessing the ability to adopt different specificities to target a wide range of proteinases. Three other highly conserved residues (two Gly's and a Pro) are located at structurally important sites β-turns and are thus critical for maintaining the overall PI structure.

(7) Phylogenetic analysis shows that the repeat units cluster into several groups according to repeat number and species. The different similarities patterns between repeat units in genes suggest that in different species the duplication history and mechanism should be different.

Overall, the evolution of Pot II serine proteinase inhibitors brings obvious advantages to *Solanaceae* plants for fighting against pests. The duplications in both gene level and domain level enable rapid and efficient expression of Pot II genes.

Codon usage analysis suggests that Pot II genes utilize abundant tRNA subpopulations that facilitate the rapid expression and response to wounds and pest infestation. On the structure level, the multi-RU precursors can acquire circularly permutated structures which have a more stable and thermodynamic favorable folding. The molecular adaptation particular the positive selection over reactive sites provides various inhibition activities targeting the broad range of pathogenic proteinases.

In our quest to build 3D structural models for SDPs, new SDP proteins resulting from single-domain genes of the Pot II family will adopt the ancestral fold (with L-H topology), while all multi-domain Pot II sequences will adopt the permuted fold (H-L topology), with the termini arranged as H+L. Normally, for all the SDPs, repeated sequence units fold into repeated structural units, each of which can be modeled using SDPMOD directly. The Pot II family is the only exception to this rule and will require manual query-template alignments to be generated prior to model building.

Chapter 5 Conclusions and future directions

5.1 Conclusions

Small disulphide-rich proteins (SDPs) are a special class of proteins with diverse functions, which mainly includes secretory proteins with predatory, defensive or regulatory roles (such as toxins, inhibitors and hormones). SDPs are rich sources for therapeutic drugs, diagnostic agents and pesticides. SDPs are characterized as short polypeptides stabilized in conformation by disulphide bridges. Bioinformatics studies suggest the central importance of these disulphide bridges in the structure, function and evolution of SDPs. For this important class of proteins, we have developed strategies for determining single domains, for each of which a custom-designed 3D model building strategy has been devised and tested for large scale comparative modelling. While almost all SDPs are composed of tandem repeats of monomeric domains, which are conserved both in sequence and in 3D structure, we had a single example of an SDP family where, as a defensive strategy, the structural repeat is a permutated fold from the sequence repeat. We have used in-depth bioinformatics analyses to understand why this occurs and predict how a new member of this family would fold.

Overall, the specific outcomes of this study can be summarized as follows:

(1) *SDFD* – a database of Small Disulphide-rich Folds (SDFs), has been curated to host high quality and comprehensive data for the research of SDPs and SDFs. SDFD incorporated clean data from various resources and can serve as

a template repository for structural modeling of SDPs.

(2) *Classification scheme*: A hierarchal classification scheme for SDFs is proposed and applied to SDFD. The classification scheme classifies all SDFs into four levels: DSSF (Disulphide superfamily, according to the disulphide number), DSF (Disulphide family, based on the disulphide connectivity), DSC (Disulphide cluster, clustering by cysteine signature) and DSI (Disulphide individual, each SDF domain).

(3) *SDFD data analysis*: A systematic analysis of SDFD revealed the following interesting findings:

   a. The distribution of SDFs on disulphide number and disulphide connectivity is uneven. Current data suggested disulphide connectivities for two or three-disulphide SDFs have preference on overlapped topology.

   b. The analysis of intra- and inter-domain disulphide shows the low frequency of inter-domain disulphide in SDFs and this preference can be applied to improve computational methods from several fields, such as domain boundary prediction, disulphide connectivity prediction and structural modeling of SDPs.

   c. The analysis of intra- and inter-chain disulphide reported the low occurrence of inter-chain disulphide bonds. Most inter-chain disulphide in structure databases are actually intra-chain disulphide bonds according to the definition of genetic domain.

   d. Analysis shows cysteine signature can help detecting distantly related homologs and convergently evolved structures.

(4) *Modeling 3D structures of SDPs:* SDPMOD – a novel method for the automated comparative modeling of SDPs has been developed. The CHARMM22 forcefield topologies and parameters for non-standard residues in conotoxins were developed for the structural modeling of conotoxin. To the best of our knowledge, this is the only methodology available currently for building 3D models of proteins with non-standard residues.

(5) *Novel SDP family analysis:* An intriguing family of SDPs, Potato II (Pot II) proteinase inhibitor family, was investigated systematically. The main findings are listed as below:

   a. The conserved patterns and features were characterized on gene architecture, protein sequence and structural domain;

   b. The sequence repeats cluster into distinct phylogenetic groups depending on the repeat number and the species. The conservation patterns between repeat units in available genes suggest variation of duplication history and mechanism in different species;

   c. The permutated domains appear more stable than original repeat domain, from available structural information. Therefore, a multiple-repeat sequence (up to eight in *Nicotiana*) is likely to adopt the permuted fold from contiguous sequence segments, with the N- and C-termini forming a single non-contiguous structural domain, linking the bracelet of tandem repeats;

   d. Two 3-repeat sequences from *Capsicum annuum* have evolved to tailor the sequence repeats to correspond with the structural repeats thus eliminating the bracelet link. The repeat unit for this group is a circular permutation of the ancestral domain, making this group the late entrant to the Pot II family;

e. The analysis of nonsynonymous/synonymous substitution rate ratio (ω = d_N/d_S) in Pot II domain revealed heterogeneous selective pressures among amino acid sites: the reactive site is under position selection (providing different specificity to target varieties of proteinases) while the cysteine scaffold is under purifying selection (essential for maintaining the fold). This provides a prefect example for the application of SDFs in protein engineering and drug design.

5.2 Future directions

Although the roles of cysteines and disulphide bridges on the structure, function and evolution of SDPs is being studied, the effort to accurately predict the behavior of cysteines and disulphides and utilize them is still enormous as such predictions are still in their infancy. Several avenues of SDP-related research directions can be pursued in the future. A brief outline of a few of these is provide below.

5.2.1 Disulphide connectivity prediction

Disulphide connectivity prediction is one of the major topics in the research of disulphide-bonded proteins. The correct prediction of disulphide connectivity for a given protein sequence will greatly facilitate the protein structure prediction by reducing the search space. Although several methods have been developed recently (Fariselli and Casadio 2001; Vullo and Frasconi 2004; Chen and Hwang 2005; Tsai *et al*. 2005), the best reported accuracy is 55% for proteins with two to five disulphide (Chen and Hwang 2005). This area thus offers an opportunity for methodological development and improvement of prediction accuracy. SDFD provides a clean dataset for the prediction of disulphide connectivity. The findings obtained in this study (e.g. distribution of disulphide distance, the preference of disulphide connectivity, cysteine signature) can be used as features for sophisticated machining learning techniques.

5.2.2 The *de novo* modeling of SDPs

In this study, the structural modeling of SDPs was limited to comparative modeling. Although comparative modeling can provide reliable homology models, it is dependent on the availability of known related structures as templates, available only for a small fraction of known sequences. The *de novo* modeling of SDPs will be greatly simplified with known disulphide connectivity, which can be used as distance restraints during the modeling. The dataset derived from this study can be used for testing and validating the new method. The development of novel template-independent modeling methods will greatly expand the scope of protein structure prediction for SDPs.

5.2.3 Protein engineering and drug design

The analysis of Pot II family illustrated a perfect example for the fitness of small disulphide-rich fold as a scaffold for protein engineering. The multiple cross-linked disulphide bridges provide robustness for the domain while the loop regions can be designed to meet different requirements for functional specificity and affinity. Beside Pot II domain and conotoxins, several other small disulphide-rich domains have been reported as perfect scaffolds for protein engineering and drug design, such as Knottins (Rees *et al*. 1982), BPTI domains (James *et al.* 1995), three-finger domains (Menez 2004). Such small disulphide-rich scaffolds can provide both rigidity and variability which are critical for the tight binding to target molecules (Greenblatt *et al.* 1989; Barrette-Ng *et al.* 2003). Therefore protein engineering and drug design based on SDFs is a promising and attractive research area. SDFD also contains 150 protein complexes, e.g. one two-domain inhibitors (TI-II) binding to two molecules of proteinases (PDB ID: 1OYV). The protein-protein interaction studies and docking analyses of these complexes would be interesting and valuable for drug design.

Bibliography

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. 2002. *Molecular Biology of the Cell*, 4th ed. Garland Science, New York, pp. 462.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389-3402.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32:** D226-D229.

Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181:** 223-230.

Anfinsen, C.B., and Haber, E. 1961. Studies on the reduction and re-formation of protein disulfide bonds. *J. Biol. Chem.* **236:** 1361-1363.

Anfinsen, C.B., Haber, E., Sela, M., and White, F.H., Jr. 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences* **47:** 1309-1314.

Antcheva, N., Patthy, A., Athanasiadis, A., Tchorbanov, B., Zakhariev, S., and Pongor, S. 1996. Primary structure and specificity of a serine proteinase inhibitor from paprika (Capsicum annuum) seeds. *Biochim. Biophys. Acta* **1298:** 95-101.

Apic, G., Gough, J., and Teichmann, S.A. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310:** 311-325.

Armon, A., Graur, D., and Ben-Tal, N. 2001. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307:** 447-463.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25-29.

Atkinson, A.H., Heath, R.L., Simpson, R.J., Clarke, A.E., and Anderson, M.A. 1993. Proteinase inhibitors in Nicotiana alata stigmas are derived from a precursor protein which is processed into five homologous inhibitors. *Plant Cell* **5:** 203-213.

Baker, D., and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* **294:** 93-96.

Balandin, T., van der Does, C., Albert, J.M., Bol, J.F., and Linthorst, H.J. 1995. Structure and induction pattern of a novel proteinase inhibitor class II gene of tobacco. *Plant Mol. Biol.* **27:** 1197-1204.

Bardwell, J.C., McGovern, K., and Beckwith, J. 1991. Identification of a protein required for disulfide bond formation in vivo. *Cell* **67:** 581-589.

Baron, M., Norman, D.G., and Campbell, I.D. 1991. Protein modules. *Trends Biochem. Sci.* **16:** 13-17.

Barrette-Ng, I.H., Ng, K.K., Cherney, M.M., Pearce, G., Ghani, U., Ryan, C.A., and

James, M.N. 2003a. Unbound form of tomato inhibitor-II reveals interdomain flexibility and conformational variability in the reactive site loops. *J. Biol. Chem.* **278:** 31391-31400.

Barrette-Ng, I.H., Ng, K.K., Cherney, M.M., Pearce, G., Ryan, C.A., and James, M.N. 2003b. Structural basis of inhibition revealed by a 1:2 complex of the two-headed tomato inhibitor-II and subtilisin carlsberg. *Journal of Biological Chemistry.*

Barrette-Ng, I.H., Ng, K.K., Cherney, M.M., Pearce, G., Ryan, C.A., and James, M.N. 2003c. Structural basis of inhibition revealed by a 1:2 complex of the two-headed tomato inhibitor-II and subtilisin Carlsberg. *J. Biol. Chem.* **278:** 24062-24071.

Barta, E., Pintar, A., and Pongor, S. 2002. Repeats with variations: accelerated evolution of the Pin2 family of proteinase inhibitors. *Trends Genet.* **18:** 600-603.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30:** 276-280.

Bates, P.A., Kelley, L.A., MacCallum, R.M., and Sternberg, M.J. 2001. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins* **Suppl 5:** 39-46.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235-242.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31:** 365-370.

Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. 1993. dbEST--database for "expressed sequence tags". *Nat. Genet.* **4:** 332-333.

Bowles, D.J. 1990. Defense-related proteins in higher plants. *Annu. Rev. Biochem.* **59:** 873-907.

Brandstadter, J., Rossbach, C., and Theres, K. 1996. Expression of genes for a defensin and a proteinase inhibitor in specific areas of the shoot apex and the developing flower in tomato. *Mol. Gen. Genet.* **252:** 146-154.

Brenner, S.E., Koehl, P., and Levitt, M. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **28:** 254-256.

Brenner, S.E., and Levitt, M. 2000. Expectations from structural genomics. *Protein Sci.* **9:** 197-200.

Bryant, J., Green, T.R., Gurusaddaiah, T., and Ryan, C.A. 1976. Proteinase inhibitor II from potatoes: isolation and characterization of its protomer components. *Biochemistry* **15:** 3418-3424.

Chandonia, J.M., and Brenner, S.E. 2005. Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins* **58:** 166-179.

Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S.E. 2004. The ASTRAL Compendium in 2004. *Nucleic Acids Res.* **32:** D189-D192.

Chandonia, J.M., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S.E. 2002. ASTRAL compendium enhancements. *Nucleic Acids Res.* **30:** 260-263.

Chen, J., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., et al. 2003. MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.* **31:** 474-477.

Choi, D., Park, J.A., Seo, Y.S., Chun, Y.J., and Kim, W.T. 2000. Structure and stress-related expression of two cDNAs encoding proteinase inhibitor II of Nicotiana glutinosa L. *Biochim. Biophys. Acta* **1492:** 211-215.

Chuang, C.C., Chen, C.Y., Yang, J.M., Lyu, P.C., and Hwang, J.K. 2003. Relationship between protein structures and disulfide-bonding patterns. *Proteins* **53:** 1-5.

Colovos, C., and Yeates, T.O. 1993. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* **2:** 1511-1519.

Contreras-Moreira, B., and Bates, P.A. 2002. Domain fishing: a first step in protein comparative modelling. *Bioinformatics* **18:** 1141-1142.

Craig, A.G., Bandyopadhyay, P., and Olivera, B.M. 1999. Post-translationally modified neuropeptides from Conus venoms. *Eur. J. Biochem.* **264:** 271-275.

Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L., and Elofsson, A. 2001. A study of quality measures for protein threading models. *BMC Bioinformatics* **2:** 5.

Dailey, F.E., and Berg, H.C. 1993. Mutants in disulfide bond formation that disrupt flagellar assembly in Escherichia coli. *Proceedings of the National Academy of Sciences* **90:** 1043-1047.

Dammann, C., Rojo, E., and Sanchez-Serrano, J.J. 1997. Abscisic acid and jasmonic acid activate wound-inducible genes in potato through separate, organ-specific signal transduction pathways. *Plant J.* **11:** 773-782.

Dengler, U., Siddiqui, A.S., and Barton, G.J. 2001. Protein structural domains: analysis of the 3Dee domains database. *Proteins* **42:** 332-344.

Dover, G. 1982. Molecular drive: a cohesive mode of species evolution. *Nature* **299:** 111-117.

Drakopoulou, E., Vizzavona, J., Neyton, J., Aniort, V., Bouet, F., Virelizier, H., Menez, A., and Vita, C. 1998. Consequence of the removal of evolutionary conserved disulfide bridges on the structure and function of charybdotoxin and evidence that particular cysteine spacings govern specific disulfide bond formation. *Biochemistry* **37:** 1292-1301.

Duan, X., Li, X., Xue, Q., Abo-el-Saad, M., Xu, D., and Wu, R. 1996. Transgenic rice plants harboring an introduced potato proteinase inhibitor II gene are insect resistant. *Nat. Biotechnol.* **14:** 494-498.

Duda, T.F., Jr., and Palumbi, S.R. 1999. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod Conus. *Proc. Natl. Acad. Sci. U. S. A.* **96:** 6820-6823.

Fiser, A., and Simon, I. 2000. Predicting the oxidation state of cysteines by multiple sequence alignment. *Bioinformatics* **16:** 251-256.

Frand, A.R., and Kaiser, C.A. 1998. The ERO1 gene of yeast is required for oxidation of protein dithiols in the endoplasmic reticulum. *Mol. Cell* **1:** 161-170.

Fuhrmann, M., Hausherr, A., Ferbitz, L., Schodl, T., Heitzer, M., and Hegemann, P. 2004. Monitoring dynamic expression of nuclear genes in Chlamydomonas reinhardtii by using a synthetic luciferase reporter gene. *Plant Mol. Biol.* **55:** 869-881.

Gadea, J., Mayda, M.E., Conejero, V., and Vera, P. 1996. Characterization of defense-related genes ectopically expressed in viroid-infected tomato plants. *Mol. Plant. Microbe Interact.* **9:** 409-415.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19:** 163-164.

Gopalan, V., Tan, T.W., Lee, B.T., and Ranganathan, S. 2004. Xpro: database of eukaryotic protein-encoding genes. *Nucleic Acids Res.* **32:** D59-D63.

Graham, J.S., Pearce, G., Merryweather, J., Titani, K., Ericsson, L.H., and Ryan, C.A. 1985. Wound-induced proteinase inhibitors from tomato leaves. II. The cDNA-deduced primary structure of pre-inhibitor II. *J. Biol. Chem.* **260:** 6561-6564.

Greenblatt, H.M., Ryan, C.A., and James, M.N. 1989. Structure of the complex of Streptomyces griseus proteinase B and polypeptide chymotrypsin inhibitor-1 from Russet Burbank potato tubers at 2.1 A resolution. *J. Mol. Biol.* **205:** 201-228.

Guex, N., Diemand, A., and Peitsch, M.C. 1999. Protein modelling for all. *Trends in Biochemical Science* **24:** 364-367.

Guex, N., and Peitsch, M.C. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18:** 2714-2723.

Hara, K., Yagi, M., Koizumi, N., Kusano, T., and Sano, H. 2000. Screening of wound-responsive genes identifies an immediate-early expressed gene encoding a highly charged protein in mechanically wounded tobacco plants. *Plant Cell Physiology* **41:** 684-691.

Harrison, P.M., and Sternberg, M.J. 1994. Analysis and classification of disulphide connectivity in proteins. The entropic effect of cross-linkage. *J. Mol. Biol.* **244:** 448-463.

Harrison, P.M., and Sternberg, M.J. 1996. The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. *J. Mol. Biol.* **264:** 603-623.

Heath, R.L., Barton, P.A., Simpson, R.J., Reid, G.E., Lim, G., and Anderson, M.A. 1995. Characterization of the protease processing sites in a multidomain proteinase inhibitor precursor from Nicotiana alata. *Eur. J. Biochem.* **230:** 250-257.

Heger, A., and Holm, L. 2003. Exhaustive enumeration of protein domain families. *J. Mol. Biol.* **328:** 749-767.

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K., and Hood, L. 1997. Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278:** 609-614.

Higgins, D.G., and Sharp, P.M. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73:** 237-244.

Hogg, P.J. 2003. Disulfide bonds as switches for protein function. *Trends Biochem. Sci.* **28:** 210-214.

Holm, L., and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233:** 123-138.

Holm, L., and Sander, C. 1994. Parser for protein folding units. *Proteins* **19:** 256-268.

Holm, L., and Sander, C. 1998. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **26:** 316-319.

Hu, S.H., Gehrmann, J., Guddat, L.W., Alewood, P.F., Craik, D.J., and Martin, J.L. 1996. The 1.1 A crystal structure of the neuronal acetylcholine receptor antagonist, alpha-conotoxin PnIA from Conus pennaceus. *Structure* **4:** 417-

423.

Hui, D., Iqbal, J., Lehmann, K., Gase, K., Saluz, H.P., and Baldwin, I.T. 2003. Molecular interactions between the specialist herbivore Manduca sexta (lepidoptera, sphingidae) and its natural host Nicotiana attenuata: V. microarray analysis and further characterization of large-scale changes in herbivore-induced mRNAs. *Plant Physiol.* **131:** 1877-1893.

Johnson, R., Narvaez, J., An, G., and Ryan, C. 1989. Expression of proteinase inhibitors I and II in transgenic tobacco plants: effects on natural defense against Manduca sexta larvae. *Proc. Natl. Acad. Sci. U. S. A.* **86:** 9871-9875.

Keil, M., Sanchez-Serrano, J., Schell, J., and Willmitzer, L. 1986. Primary structure of a proteinase inhibitor II gene from potato (Solanum tuberosum). *Nucleic Acids Res.* **14:** 5641-5650.

Kelley, L.A., Gardner, S.P., and Sutcliffe, M.J. 1996. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.* **9:** 1063-1065.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. 2002. Selection in the evolution of gene duplications. *Genome Biology* **3:** RESEARCH0008.

Kong, L., Lee, B.T., Tong, J.C., Tan, T.W., and Ranganathan, S. 2004. SDPMOD: an automated comparative modeling server for small disulfide-bonded proteins. *Nucleic Acids Res.* **32:** W356-W359.

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E., and Berman, H.M. 2006. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.* **34:** D302-D305.

Lambert, C., Leonard, N., De Bolle, X., and Depiereux, E. 2002. ESyPred3D: Prediction of proteins 3D structures. *Bioinformatics* **18:** 1250-1256.

Laskowski, M., Jr., and Kato, I. 1980. Protein inhibitors of proteinases. *Annu. Rev. Biochem.* **49:** 593-626.

Laskowski, R.A., Moss, D.S., and Thornton, J.M. 1993. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* **231:** 1049-1067.

Lee, J.S., Brown, W.E., Graham, J.S., Pearce, G., Fox, E.A., Dreher, T.W., Ahern, K.G., Pearson, G.D., and Ryan, C.A. 1986. Molecular characterization and phylogenetic studies of a wound-inducible proteinase inhibitor I gene in Lycopersicon species. *Proc. Natl. Acad. Sci. U. S. A.* **83:** 7277-7281.

Lee, M.C., Scanlon, M.J., Craik, D.J., and Anderson, M.A. 1999. A novel two-chain proteinase inhibitor generated by circularization of a multidomain precursor protein. *Nat. Struct. Biol.* **6:** 526-530.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P., and Bork, P. 2004. SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* **32:** D142-D144.

Lewis, R.J. 2004. Conotoxins as selective inhibitors of neuronal ion channels, receptors and transporters. *IUBMB Life* **56:** 89-93.

Li, L., Li, C., Lee, G.I., and Howe, G.A. 2002. Distinct roles for jasmonate synthesis and action in the systemic wound response of tomato. *Proc. Natl. Acad. Sci. U. S. A.* **99:** 6416-6421.

Li, R.A., and Tomaselli, G.F. 2004. Using the deadly mu-conotoxins as probes of voltage-gated sodium channels. *Toxicon* **44:** 117-122.

Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G., and Chothia, C. 2000. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **28:** 257-259.

Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J., and Brunak, S. 1997. Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.* **10:** 1241-1248.

MacKerell, A.D., Bashford, J., Bellott, D.M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem.* **102:** 3586-3616.

Madej, T., Gibrat, J.F., and Bryant, S.H. 1995. Threading a database of protein cores. *Proteins* **23:** 356-369.

Mansfeld, J., Vriend, G., Dijkstra, B.W., Veltman, O.R., Van den Burg, B., Venema, G., Ulbrich-Hofmann, R., and Eijsink, V.G. 1997. Extreme stabilization of a thermolysin-like protease by an engineered disulfide bond. *J. Biol. Chem.* **272:** 11152-11156.

Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., et al. 2003. CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* **31:** 383-387.

Martelli, P.L., Fariselli, P., Malaguti, L., and Casadio, R. 2002. Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy. *Protein Sci.* **11:** 2735-2739.

Martensson, L.G., Karlsson, M., and Carlsson, U. 2002. Dramatic stabilization of the native state of human carbonic anhydrase II by an engineered disulfide bond. *Biochemistry* **41:** 15867-15875.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29:** 291-325.

Mas, J.M., Aloy, P., Marti-Renom, M.A., Oliva, B., de Llorens, R., Aviles, F.X., and Querol, E. 2001. Classification of protein disulphide-bridge topologies. *J. Comput. Aided Mol. Des.* **15:** 477-487.

May, A.C., and Johnson, M.S. 1995. Improved genetic algorithm-based protein structure comparisons: pairwise and multiple superpositions. *Protein Eng.* **8:** 873-882.

McIntosh, J.M., Olivera, B.M., and Cruz, L.J. 1999a. Conus peptides as probes for ion channels. *Methods Enzymol.* **294:** 605-624.

McIntosh, J.M., Santos, A.D., and Olivera, B.M. 1999b. Conus peptides targeted to specific nicotinic acetylcholine receptor subtypes. *Annual Review of Biochemstry* **68:** 59-88.

Michie, A.D., Orengo, C.A., and Thornton, J.M. 1996. Analysis of domain structural class using an automated class assignment protocol. *Journal of Molecular Biology* **262:** 168-185.

Miller, E.A., Lee, M.C., Atkinson, A.H., and Anderson, M.A. 2000. Identification of a novel four-domain member of the proteinase inhibitor II family from the stigmas of Nicotiana alata. *Plant Mol. Biol.* **42:** 329-333.

Missiakas, D., Schwager, F., and Raina, S. 1995. Identification and characterization of a new disulfide isomerase-like protein (DsbD) in Escherichia coli. *EMBO J.* **14:** 3415-3424.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536-540.

Nakamura, Y., Gojobori, T., and Ikemura, T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* **28:** 292.

Neurath, H. 1989. Proteolytic processing and physiological regulation. *Trends in Biochemical Science* **14:** 268-271.

Nielsen, K.J., Heath, R.L., Anderson, M.A., and Craik, D.J. 1995. Structures of a series of 6-kDa trypsin inhibitors isolated from the stigma of Nicotiana alata. *Biochemistry* **34:** 14304-14311.

Nielsen, R., and Yang, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929-936.

Olivera, B.M., Rivier, J., Clark, C., Ramilo, C.A., Corpuz, G.P., Abogadie, F.C., Mena, E.E., Woodward, S.R., Hillyard, D.R., and Cruz, L.J. 1990. Diversity of Conus neuropeptides. *Science* **249:** 257-263.

Olivera, B.M., Walker, C., Cartier, G.E., Hooper, D., Santos, A.D., Schoenfeld, R., Shetty, R., Watkins, M., Bandyopadhyay, P., and Hillyard, D.R. 1999. Speciation of cone snails and interspecific hyperdivergence of their venom peptides. Potential evolutionary significance of introns. *Ann. N. Y. Acad. Sci.* **870:** 223-237.

Orengo, C.A., and Taylor, W.R. 1996. SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **266:** 617-635.

Page, R.D. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12:** 357-358.

Pearce, G., Johnson, S., and Ryan, C.A. 1993. Purification and characterization from tobacco (Nicotiana tabacum) leaves of six small, wound-inducible, proteinase isoinhibitors of the potato inhibitor II family. *Plant Physiol.* **102:** 639-644.

Pearce, G., Ryan, C.A., and Liljegren, D. 1988. Proteinase inhibitor-I and inhibitor-II in fruit of wild tomato species - transient components of a mechanism for defense and seed dispersal. *Planta* **175:** 527-531.

Pearce, G., Sy, L., Russell, C., Ryan, C.A., and Hass, G.M. 1982. Isolation and characterization from potato tubers of two polypeptide inhibitors of serine proteinases. *Arch. Biochem. Biophys.* **213:** 456-462.

Pearl, F.M., Bennett, C.F., Bray, J.E., Harrison, A.P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., and Orengo, C.A. 2003. The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.* **31:** 452-455.

Peitsch, M.C. 1996. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.* **24:** 274-279.

Perry, L.J., and Wetzel, R. 1984. Disulfide bond engineered into T4 lysozyme: stabilization of the protein toward thermal inactivation. *Science* **226:** 555-557.

Pieper, U., Eswar, N., Stuart, A.C., Ilyin, V.A., and Sali, A. 2002. MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* **30:** 255-259.

Plunkett, G., Senear, D.F., Zuroske, G., and Ryan, C.A. 1982. Proteinase inhibitors I and II from leaves of wounded tomato plants: purification and properties. *Arch. Biochem. Biophys.* **213:** 463-472.

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16:** 276-277.

Richardson, M. 1977. The proteinase inhibitors of plants and micro-organisms.

Phytochemistry **16:** 159-169.

Richardson, M. 1979. The complete amino acid sequence and the trypsin reactive (inhibitory) site of the major proteinase inhibitor from the fruits of aubergine (Solanum melongena L.). *FEBS Lett.* **104:** 322-326.

Rigby, A.C., Baleja, J.D., Li, L., Pedersen, L.G., Furie, B.C., and Furie, B. 1997. Role of gamma-carboxyglutamic acid in the calcium-induced structural transition of conantokin G, a conotoxin from the marine snail Conus geographus. *Biochemistry* **36:** 15677-15684.

Robinson, C.R., and Sauer, R.T. 2000. Striking stabilization of Arc repressor by an engineered disulfide bond. *Biochemistry* **39:** 12494-12502.

Rogers, J.P., Luginbuhl, P., Pemberton, K., Harty, P., Wemmer, D.E., and Stevens, R.C. 2000. Structure-activity relationships in a peptidic alpha7 nicotinic acetylcholine receptor antagonist. *J. Mol. Biol.* **304:** 911-926.

Rost, B. 1998. Marrying structure and genomics. *Structure* **6:** 259-263.

Ryan, C.A., and Moura, D.S. 2002. Systemic wound signaling in plants: a new perception. *Proc. Natl. Acad. Sci. U. S. A.* **99:** 6519-6520.

Saitou, N., and Nei, M. 1987a. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406-425.

Saitou, N., and Nei, M. 1987b. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4:** 406-425.

Sali, A., and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234:** 779-815.

Sanchez, R., and Sali, A. 1998. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proceedings of the National Academy of Sciences* **95:** 13597-13602.

Santoyo, G., and Romero, D. 2005. Gene conversion and concerted evolution in bacterial genomes. *FEMS Microbiol. Rev.* **29:** 169-183.

Sayle, R.A., and Milner-White, E.J. 1995. RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **20:** 374.

Scanlon, M.J., Lee, M.C., Anderson, M.A., and Craik, D.J. 1999. Structure of a putative ancestral protein encoded by a single sequence repeat from a multidomain proteinase inhibitor gene from Nicotiana alata. *Structure* **7:** 793-802.

Schechter, I., and Berger, A. 1968. On the active site of proteases. 3. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochem. Biophys. Res. Commun.* **32:** 898-902.

Schirra, H.J., and Craik, D.J. 2005. Structure and folding of potato type II proteinase inhibitors: circular permutation and intramolecular domain swapping. *Protein and Pept. Lett.* **12:** 421-431.

Schirra, H.J., Scanlon, M.J., Lee, M.C., Anderson, M.A., and Craik, D.J. 2001. The solution structure of C1-T1, a two-domain proteinase inhibitor derived from a circular precursor protein from Nicotiana alata. *J. Mol. Biol.* **306:** 69-79.

Schlotterer, C., and Tautz, D. 1994. Chromosomal homogeneity of Drosophila ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. *Curr. Biol.* **4:** 777-783.

Schneider, T.D., and Stephens, R.M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18:** 6097-6100.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.* **95:** 5857-5864.

Schwede, T., Diemand, A., Guex, N., and Peitsch, M.C. 2000. Protein structure computing in the genomic era. *Res. Microbiol.* **151:** 107-112.

Scott, M.J., Huckaby, C.S., Kato, I., Kohr, W.J., Laskowski, M., Jr., Tsai, M.J., and O'Malley, B.W. 1987. Ovoinhibitor introns specify functional domains as in the related and linked ovomucoid gene. *J. Biol. Chem.* **262:** 5899-5907.

Shen, G.S., Layer, R.T., and McCabe, R.T. 2000. Conopeptides: From deadly venoms to novel therapeutics. *Drug Discovery Today* **5:** 98-106.

Shin, R., Lee, G.-J., Park, C.-J., Kim, T.-Y., You, J.-S., Nam, Y.-W., and Paek, K.-H. 2001. Isolation of pepper mRNAs differentially expressed during the hypersensitive response to tobacco mosaic virus and characterization of a proteinase inhibitor gene. *Plant Science* **161:** 727-737.

Siddiqui, A.S., and Barton, G.J. 1995. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.* **4:** 872-884.

Siddiqui, A.S., Dengler, U., and Barton, G.J. 2001. 3Dee: a database of protein structural domains. *Bioinformatics* **17:** 200-201.

Sowdhamini, R., Srinivasan, N., Shoichet, B., Santi, D.V., Ramakrishnan, C., and Balaram, P. 1989. Stereochemical modeling of disulfide bridges. Criteria for introduction into proteins by site-directed mutagenesis. *Protein Eng.* **3:** 95-103.

Swindells, M.B. 1995. A procedure for detecting structural domains in proteins. *Protein Sci.* **4:** 103-112.

Tan, P.T., Veeramani, A., Srinivasan, K.N., Ranganathan, S., and Brusic, V. 2006. SCORPION2: A database for structure-function analysis of scorpion toxins. *Toxicon* **47:** 1-8.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41.

Taylor, B.H., Young, R.J., and Scheuring, C.F. 1993. Induction of a proteinase inhibitor II-class gene by auxin in tomato roots. *Plant Mol. Biol.* **23:** 1005-1014.

Tessier, D., Bardiaux, B., Larre, C., and Popineau, Y. 2004. Data mining techniques to study the disulfide-bonding state in proteins: signal peptide is a strong descriptor. *Bioinformatics* **20:** 2509-2512.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876-4882.

Tsai, C.H., Chen, B.J., Chan, C.H., Liu, H.L., and Kao, C.Y. 2005. Improving disulfide connectivity prediction with sequential distance between oxidized cysteines. *Bioinformatics* **21:** 4416-4419.

van Vlijmen, H.W., Gupta, A., Narasimhan, L.S., and 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.

Vinayagam, A., Pugalenthi, G., Rajesh, R., and Sowdhamini, R. 2004. DSDBASE: a consortium of native and modelled disulphide bonds in proteins. *Nucleic Acids*

Res. **32:** D200-D202.

Vivek, G., Tan, T.W., and Ranganathan, S. 2003. XdomView: protein domain and exon position visualization. *Bioinformatics* **19:** 159-160.

von Segesser, L.K., Mueller, X., Marty, B., Horisberger, J., and Corno, A. 2001. Alternatives to unfractionated heparin for anticoagulation in cardiopulmonary bypass. *Perfusion* **16:** 411-416.

Vriend, G. 1990. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8:** 52-56.

Vriend, G., and Sander, C. 1993. Quality control of protein models: Directional atomic contact analysis. *Journal of Applied Crystallography* **26:** 46-70.

Wakamatsu, K., Kohda, D., Hatanaka, H., Lancelin, J.M., Ishida, Y., Oya, M., Nakamura, H., Inagaki, F., and Sato, K. 1992. Structure-activity relationships of mu-conotoxin GIIIA: structure determination of active and inactive sodium channel blocker peptides by NMR and simulated annealing calculations. *Biochemistry* **31:** 12577-12584.

Wang, X., Connor, M., Smith, R., Maciejewski, M.W., Howden, M.E., Nicholson, G.M., Christie, M.J., and King, G.F. 2000. Discovery and characterization of a family of insecticidal neurotoxins with a rare vicinal disulfide bridge. *Nat. Struct. Biol.* **7:** 505-513.

Wedemeyer, W.J., Welker, E., Narayan, M., and Scheraga, H.A. 2000. Disulfide bonds and protein folding. *Biochemistry* **39:** 4207-4216.

Wetlaufer, D.B. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **70:** 697-701.

Wong, W.S., Yang, Z., Goldman, N., and Nielsen, R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168:** 1041-1051.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., et al. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34:** D187-D191.

Xie, L., and Bourne, P.E. 2005. Functional Coverage of the Human Genome by Existing Structures, Structural Genomics Targets, and Homology Models. *PLoS Comput Biology* **1:** e31.

Xu, Z.F., Qi, W.Q., Ouyang, X.Z., Yeung, E., and Chye, M.L. 2001. A proteinase inhibitor II of Solanum americanum is expressed in phloem. *Plant Mol. Biol.* **47:** 727-738.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555-556.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15:** 568-573.

Yang, Z., and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19:** 908-917.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M. 2000a. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431-449.

Yang, Z., Swanson, W.J., and Vacquier, V.D. 2000b. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol. Biol. Evol.* **17:** 1446-1455.

Yang, Z., Wong, W.S., and Nielsen, R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22:** 1107-1118.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296:** 79-92.

Zhang, J., Nielsen, R., and Yang, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22:** 2472-2479.

Zhao, E., Liu, H.L., Tsai, C.H., Tsai, H.K., Chan, C.H., and Kao, C.Y. 2005. Cysteine separations profiles on protein sequences infer disulfide connectivity. *Bioinformatics* **21:** 1415-1420.

Appendices

Publications

1. Kong L, and Ranganathan S. Delineation of Modular Proteins: Domain Boundary Prediction from Sequence Information. **Briefings in Bioinformatics**, 2004, 5: 179-192.

2. Kong L, Lee BT, Tong JC, Tan TW, Ranganathan S. SDPMOD: an Automated Comparative Modeling Server for Small Disulphide-bonded Proteins. **Nucleic Acids Res**., 2004, 32: W356-W359.

3. Kong L, Tan TW and Ranganathan S. Tandem Duplication, Circular Permutation and Molecular Adaptation: Strategies of *Solanaceae* Plants for Fighting against Pathogens via Pot II Inhibitors. Manuscript under preparation.

Posters

1. Kong L, Tan TW and Ranganathan S. CHARMM Topology and Parameter Library Development for Non-standard Residues in Conopeptides. **InCoB 2003**, Penang, Malaysia.

2. Kong L, Tan TW and Ranganathan S. SDPS: Small Disulphide-bonded Proteins Structural Database. **ISMB 2003**, Brisbane, Australia.

3. Kong L, Lee BT, Tong JC, Tan TW and Ranganathan S. SDPMOD: a

Comprehensive Comparative Modeling Server for Small Disulphide-bonded Proteins. **ISMB2004**, Glasgow, UK.

Presentations

1. Bioinformatics and Structural Modeling of Small Disulphide-bonded Proteins. Pre-18th FAOBMB Symposium Satellite Workshop on Bioinformatics, Lahore, Pakistan, 27th October 2005.

2. SDPS: Small Disulphide-bonded Proteins Structural Database. Singapore Bioinformatics Symposium 2003, Singapore, 15th August 2003.

W356–W359 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh394

SDPMOD: an automated comparative modeling server for small disulfide-bonded proteins

Lesheng Kong[1], Bernett Teck Kwong Lee[1], Joo Chuan Tong[1], Tin Wee Tan[1] and Shoba Ranganathan[1,2,*]

[1]Department of Biochemistry, National University of Singapore, 8 Medical Drive, 117597, Singapore and [2]Biotechnology Research Institute, Macquarie University, NSW 2109, Australia

Received February 15, 2004; Revised and Accepted March 24, 2004

ABSTRACT

Small disulfide-bonded proteins (SDPs) are rich sources for therapeutic drugs. Designing drugs from these proteins requires three-dimensional structural information, which is only available for a subset of these proteins. SDPMOD addresses this deficit in structural information by providing a freely available automated comparative modeling service to the research community. For expert users, SDPMOD offers a manual mode that permits the selection of a desired template as well as a semi-automated mode that allows users to select the template from a suggested list. Besides the selection of templates, expert users can edit the target–template alignment, thus allowing further customization of the modeling process. Furthermore, the web service provides model stereochemical quality evaluation using PROCHECK. SDPMOD is freely accessible to academic users via the web interface at http://proline.bic.nus.edu.sg/sdpmod.

INTRODUCTION

Small disulfide-bonded proteins (SDPs) are a special class of proteins that are relatively small in size (length ⩽100 residues) and have disulfide bonds within their three-dimensional (3D) structures (1). SDPs include many secretory proteins which serve predatory, defensive or regulatory roles (such as toxins, inhibitors and hormones), and they are rich source for therapeutic drugs (2) and pesticides (3). The 3D structures of SDPs are essential for understanding the functions of SDPs and for drug design. However, 3D structure determination through experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are still both time-consuming and expensive. This results in a gap between the number of known 3D structures and the number of primary sequences that could be narrowed using large-scale automated protein structure prediction.

Among current structure prediction methods, comparative modeling is the most reliable method for generating 3D models. Comparative modeling of protein structures often requires expert knowledge and proficiency in specialized methods. In the mid-1990s, Peitsch and coworkers developed the first automated modeling server SWISS-MODEL (4), which is currently the most widely used server of this genre. Recently, several other automated comparative modeling servers have also been developed, such as CPHmodels (5), 3D-JIGSAW (6), ModWeb (7) and ESyPred3D (8).

Although so many automated comparative modeling servers are available, most of them do not work well on small SDPs for two reasons. Most of the automated servers are primarily designed for globular protein domains, making it difficult to discriminate small-sized SDPs from background noise. Taking as an example the sequence of α-conotoxin PnIA (9) (PDB id: 1PEN; 16 residues; 2 disulfide bridges in its structure), we note that both SWISS-MODEL and ModWeb report that they do not cover the modeling of sequences <25 or ⩽30 amino acid residues in length, respectively, while the other three servers state that no suitable templates can be identified for this sequence.

The second reason is that SDPs have distinct characteristics from medium-sized and large globular proteins. They usually do not have a compact hydrophobic core, which is a major factor in stabilizing protein structure. Their side chains are more likely to be exposed to solvent and their conformations are more flexible. The 3D structures of small proteins are usually dominated by disulfide bridges, metal or ligand (according to SCOP classification) (10) and tend to bind or interact with large molecules. In small disulfide-rich proteins, the effects of disulfide bridges and constrained residues such as prolines are more significant than sequence similarity. As such, the comparative modeling rules for such proteins are highly specific and different from those adopted for large globular proteins. These distinct features require specific methods and datasets to be developed for the comparative modeling of SDPs.

To address these problems, we have first developed special strategies and rules for large-scale automated comparative modeling of the entire family of conotoxins (L. Kong and

[*]To whom correspondence should be addressed. Tel: +61 2 9850 6262; Fax: +61 2 9850 8313; Email: shoba@els.mq.edu.au

S. Ranganathan, unpublished data). Subsequently these rules were extended to other SDPs. Here, we present SDPMOD, a comprehensive comparative modeling server that is designed specifically for SDPs with specialized rules and datasets.

MATERIALS AND METHODS

Non-redundant SDP structure dataset

Before the modeling can proceed, a non-redundant dataset for SDPs needs to be created to serve as the template repository. Structures containing protein chains of length <100 amino acids with at least two cysteines were retrieved from the Protein Data Bank (PDB) (11) and loaded into MySQL, a relational database management system for flexible query and manipulation. The redundancy in SDP structures was removed at two levels. First, for NMR structures which have multiple monomer models, the representative monomers were selected using NMRCLUST (12). Second, when multiple structures exist for the same sequence, the representative structure was chosen according to its structural qualities. The structural qualities are ranked by the following criteria (adopted from PDB): (i) X-ray structures over NMR structures, (ii) higher-quality factor (1/resolution$-$*R*-value) for X-ray structures and higher restraint per residue for NMR, (iii) better geometry, (iv) fewer missing atoms and non-standard residues and (v) later deposition date. Based on the above strategy, a non-redundant structure database for SDPs was generated. Currently it contains >1300 non-redundant protein chains and their coordinates. The database will be automatically updated once a month.

Modeling procedure

The SDPMOD server performs comparative modeling in four steps: (i) template selection, (ii) target–template alignment, (iii) model building and (iv) model evaluation (13). Figure 1 shows the detailed modeling procedure for automated modeling. The non-redundant dataset is first filtered using the number of cysteine residues, and the resulting template sequences are globally aligned to the target sequence using a modified scoring matrix derived from the non-redundant SDP dataset. The best templates are then selected based on the alignment scores. Target–template alignment and model building are achieved by MODELLER (14) (http://salilab.org/modeller/modeller.html), using a customized matrix to ensure that all the cysteine residues are well aligned. The final models are chosen according to the MODELLER objective function score, which reflects low energy and least stereochemical violations. Finally, the overall structural quality of the generated models is evaluated against stereochemical parameters derived from high-quality experimental structures by PROCHECK (15) (http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html).

Benchmarking

A large-scale benchmarking excercise was completed using the fully automated mode of the SDPMOD server. A control set of 664 sequences (a subset of our non-redundant SDP dataset) with known structures was used to evaluate the reliability of the server. The Cα root mean square deviation (RMSD) values between models and their actual experimental structures were calculated. The benchmarking results show SDPMOD can predict 3D models with a reasonable accuracy. For example, in the 40–70% sequence identity range, 64% of models have Cα RMSD values <1.5 Å. The detailed analysis of the accuracy of our modeling protocol is available from http://proline.bic.nus.edu.sg/sdpmod/accuracy.html.

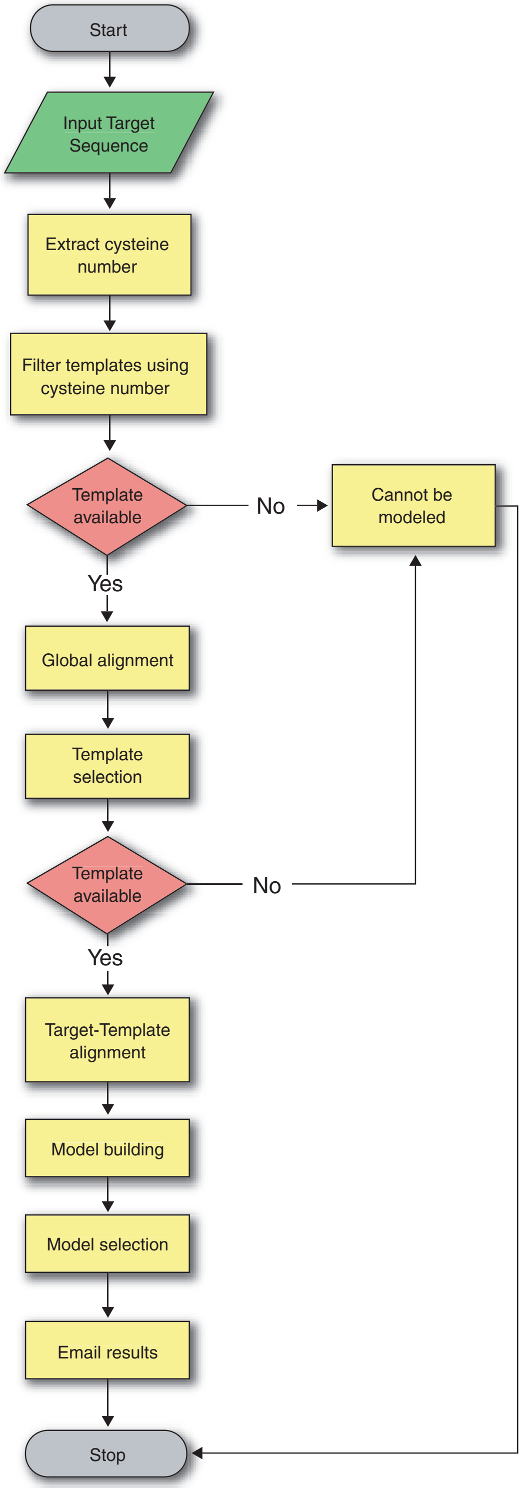
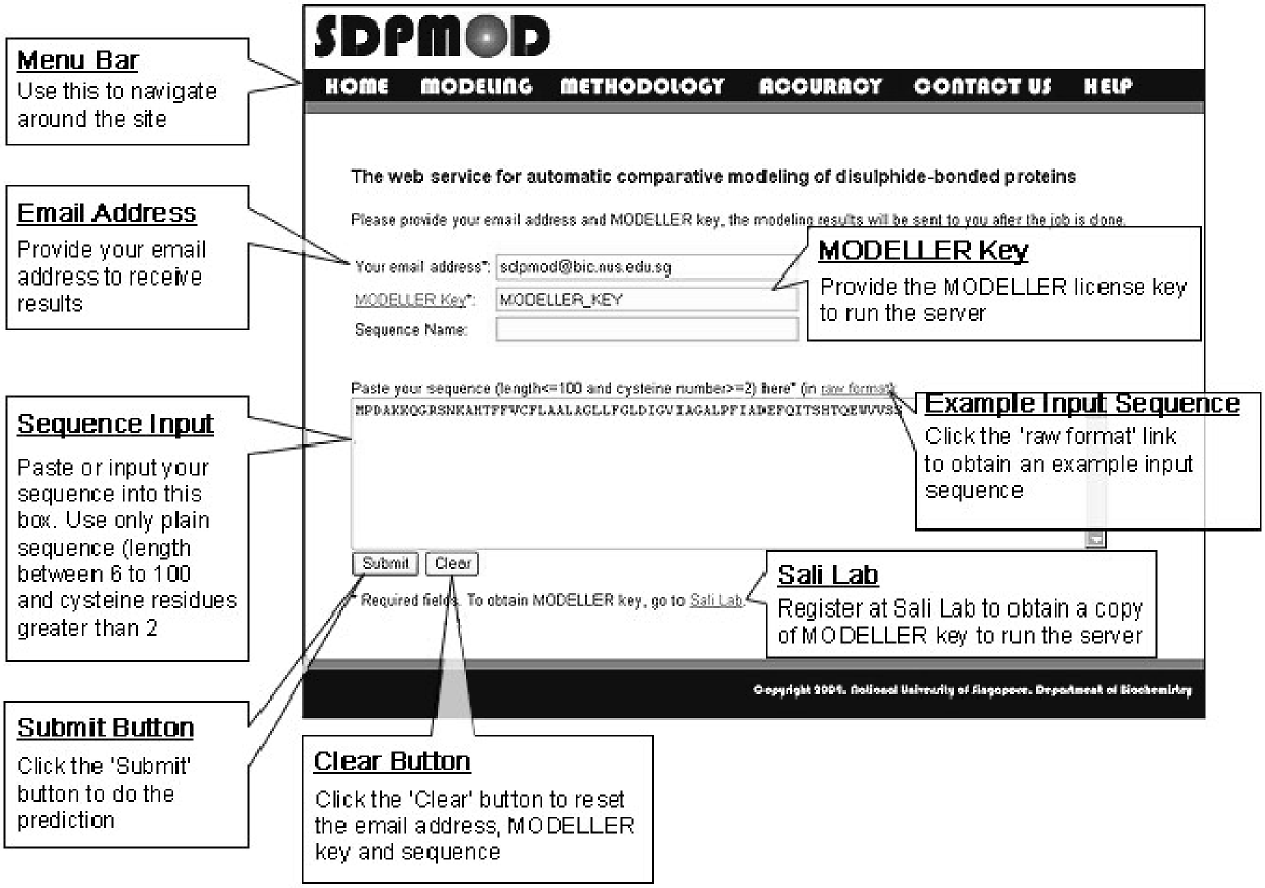


Figure 1. The SDPMOD methodology for automatic comparative modeling of small disulfide-bonded proteins.



Menu Bar
Use this to navigate around the site

Email Address
Provide your email address to receive results

Sequence Input
Paste or input your sequence into this box. Use only plain sequence (length between 6 to 100 and cysteine residues greater than 2

Submit Button
Click the 'Submit' button to do the prediction

Clear Button
Click the 'Clear' button to reset the email address, MODELLER key and sequence

MODELLER Key
Provide the MODELLER license key to run the server

Example Input Sequence
Click the 'raw format' link to obtain an example input sequence

Sali Lab
Register at Sali Lab to obtain a copy of MODELLER key to run the server

The web service for automatic comparative modeling of disulphide-bonded proteins

Please provide your email address and MODELLER key, the modeling results will be sent to you after the job is done.

Your email address*: sdpmod@bic.nus.edu.sg
MODELLER Key*: MODELLER_KEY
Sequence Name:

Paste your sequence (length<=100 and cysteine number>=2) here* (in raw format)

MPDAKKQGRSNKAHTFPWCFLAALAGLLFGLDIGVIAGALPFIADEFQITSHTQEWVVSS

* Required fields. To obtain MODELLER key, go to Sali Lab.

Copyright 2004. National University of Singapore. Department of Biochemistry

Figure 2. Example of the SDPMOD input page.

WEB SERVICE

SDPMOD is freely accessible to academic or non-profit users via a web interface (shown in Figure 2) at http://proline.bic.nus.edu.sg/sdpmod. SDPMOD is primarily designed as a fully automated procedure for ease of use. However, due to the complexity of comparative modeling, human intervention and expert knowledge may be required for optimal modeling of some proteins at two critical stages, namely template selection and target–template alignment (6). To allow for human intervention, the current version of the SDPMOD server provides three modes of modeling (fully automated, semi-automated and manual) to meet the different needs of the expert users.

The 'fully automated' mode presents an easy-to-use interface. Users can simply submit a target sequence with their email address and their MODELLER license key, obtained from the MODELLER registration page http://salilab.org/modeller/registration.shtml, and the modeling will be carried out automatically according to the procedure described in Figure 1. In the 'semi-automated' mode, a ranked list of potential templates will be returned after the target sequence is submitted. Users can then choose the best template and adjust the target–template alignment using expert knowledge. In the 'manual' mode, users are allowed to propose a template from our non-redundant SDP structure dataset and modify the target–template alignment where necessary.

After the modeling process is completed, a link with the prediction results will be returned via email. Users can refer to the link to view the prediction result and download the models. The prediction results consist of (i) a summary of the selected template(s), (ii) the predicted model based on each template in PDB format and (iii) a brief report for each modeling attempt that includes the target–template alignment used in model building, a comparison of the model against the template by means of RMSD and a PROCHECK report on the stereochemical quality of the models.

ACKNOWLEDGEMENTS

We would like to thank our colleagues at the Department of Biochemistry, National University of Singapore for their helpful comments and discussions. We are especially grateful to Professor Andrej Sali for permitting us to use MODELLER as a part of the server and Dr Ben Webb for useful suggestions. L.K. and B.L. would also like to thank the National University of Singapore for the award of Agency for Science, Technology and Research, Singapore (ASTAR) scholarships that made this work possible.

REFERENCES

1. Harrison,P.M. and Sternberg,M.J. (1996) The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. *J. Mol. Biol.*, **264**, 603–623.
2. Shen,G.S., Layer,R.T. and McCabe,R.T. (2000) Conopeptides: from deadly venoms to novel therapeutics. *Drug Discov. Today*, **5**, 98–106.

3. Richardson,M. (1977) The proteinase inhibitors of plants and micro-organisms. *Phytochemistry*, **16**, 159–169.
4. Peitsch,M.C. (1996) ProMod and Swiss-Model: internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.*, **24**, 274–279.
5. Lund,O., Frimand,K., Gorodkin,J., Bohr,H., Bohr,J., Hansen,J. and Brunak,S. (1997) Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.*, **10**, 1241–1248.
6. Bates,P.A., Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins*, **Suppl 5**, 39–46.
7. Pieper,U., Eswar,N., Stuart,A.C., Ilyin,V.A. and Sali,A. (2002) MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.*, **30**, 255–259.
8. Lambert,C., Leonard,N., De Bolle,X. and Depiereux,E. (2002) ESyPred3D: prediction of proteins 3D structures. *Bioinformatics*, **18**, 1250–1256.
9. Hu,S.H., Gehrmann,J., Guddat,L.W., Alewood,P.F., Craik,D.J. and Martin,J.L. (1996) The 1.1 Å crystal structure of the neuronal acetylcholine receptor antagonist, alpha-conotoxin PnIA from *Conus pennaceus*. *Structure*, **4**, 417–423.
10. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
11. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
12. Kelley,L.A., Gardner,S.P. and Sutcliffe,M.J. (1996) An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.*, **9**, 1063–1065.
13. Marti-Renom,M.A., Stuart,A.C., Fiser,A., Sanchez,R., Melo,F. and Sali,A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.
14. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
15. Laskowski,R.A., Moss,D.S. and Thornton,J.M. (1993) Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.*, **231**, 1049–1067.

Lesheng Kong
is a graduate student at the Department of Biochemistry, National University of Singapore, working on structural bioinformatics analysis of small disulphick-bonded proteins. He has received several ISCB travel awards to present his work at international bioinformatics conferences in Australia, Malaysia and the USA.

Shoba Ranganathan
is the Chair Professor of Bioinformatics at the Biotechnology Research Institute, Macquarie University, and an Adjunct Professor, Department of Biochemistry, National University of Singapore. Her professional appointments include Director, ISCB; the Chair, S* Life Science Informatics Alliance; and Vice-President, APBioNet. Her research work focuses on computational structural biology and comparative genome sequence analysis.

Keywords: *protein domain boundary prediction, protein modules, domain databases, domain architecture*

Shoba Ranganathan,
Biotechnology Research Institute,
Macquarie University,
NSW 2109, Australia

Tel: +61 2 9850 6262
Fax: +61 2 9850 8313
E-mail: shoba@els.mq.edu.au

Delineation of modular proteins: Domain boundary prediction from sequence information

Lesheng Kong and Shoba Ranganathan

Date received (in revised form): 12th March 2004

Abstract

The delineation of domain boundaries of a given sequence in the absence of known 3D structures or detectable sequence homology to known domains benefits many areas in protein science, such as protein engineering, protein 3D structure determination and protein structure prediction. With the exponential growth of newly determined sequences, our ability to predict domain boundaries rapidly and accurately from sequence information alone is both essential and critical from the viewpoint of gene function annotation. Anyone attempting to predict domain boundaries for a single protein sequence is invariably confronted with a plethora of databases that contain boundary information available from the internet and a variety of methods for domain boundary prediction. How are these derived and how well do they work? What definition of 'domain' do they use? We will first clarify the different definitions of protein domains, and then describe the available public databases with domain boundary information. Finally, we will review existing domain boundary prediction methods and discuss their strengths and weaknesses.

INTRODUCTION

Studies on conformation, function and evolution of proteins have revealed the central importance of protein domains as fundamental units of organisation.[1] The modular architecture of protein has been widely recognised for over a decade now.[2–5]

Proteins are composed of smaller building blocks, which are called 'domains' or 'modules'. These building blocks are distinct regions in 3D structure resulting in protein architectures assembled from modular segments that have evolved independently. The modular nature of proteins has many advantages, offering new cooperative functions and enhanced stability. As a result of the duplication and mutational evolution of these building blocks through various gene rearrangement and purifying selection mechanisms, respectively, a large proportion of proteins in higher organisms especially eukaryotic extracellular proteins, consist of multiple domains.[6]

Knowledge of protein domain architecture and domain boundaries is essential for the characterisation and understanding of protein function, particularly in the post-genome era. Domain boundary prediction has applications in many areas of protein science:

- Protein engineering: the knowledge of protein domain boundaries facilitates the engineering and design of new proteins, such as the creation of chimeric proteins which are composed of multifunctional domains and downsizing of proteins without loss of their functions.[7]

- Protein 3D structure determination: the 3D structures of large proteins are difficult to determine using standard X-ray crystallography and nuclear

Protein structure determination

magnetic resonance (NMR) spectroscopic methods owing to problems associated with crystallisation, solubility or limitations on protein size. In such case, domain boundary prediction methods can be used to split the proteins into distinct domains and then the structure of each constituent domain can be determined independently.[3]

Protein structure prediction

- Protein structure prediction: for comparative modelling, the delineation of domain boundary can optimise the search for templates, which are classified on the basis of domains;[8] and for threading, the domain boundary prediction can improve the performance by enhancing the signal-to-noise ratio.[9]

Multiple sequence alignment

- Multiple sequence alignment: accurate delineation of boundaries for homologous domains is important for reliable multiple sequence alignment,[10] which in turn serves as input to phylogenetic and other bioinformatic analyses.

Our current knowledge of domain boundaries is entirely dependent on 3D structure determination and multiple sequence alignment of protein families with the same or related function. With the exponential growth of newly determined sequences, our ability to predict domain boundaries rapidly and accurately from sequence information alone is both essential and critical from the viewpoint of gene function annotation.

Different domain definitions

In the area of protein domains, there are several databases, providing different numbers of domains with varying domain boundaries for the same protein structure.[11] When attempting to predict domain boundaries for a query protein sequence, the number of WWW servers and methods available today overwhelms the unwary user. Indeed, even the definition of the word 'domain' can differ depending on the database or method used. In this review, we attempt to separate the available definitions for the protein 'domain' into structural, functional and evolutionary classes. We then present a collection of the most frequently used and current databases and methods available for the domain boundary prediction problem. The prediction methods have been categorised depending on their methodology and applicability, with references to the databases they derive from, with our assessment of the pros and cons of choosing a particular method over others of the genre.

DIFFERENT DOMAIN DEFINITIONS: STRUCTURAL, FUNCTIONAL AND EVOLUTIONARY DOMAINS

The concept of domains plays an important role in protein science. However, this concept is defined differently under different circumstances. The term 'domain' was initially introduced in structural biology for those globular proteins that are composed of several distinct structural regions that fold independently.[2] It was also observed that specific regions of proteins are involved in effecting a specific biological task such as catalytic activity or binding a ligand (eg a DNA-binding domain). The occurrence of similar functional segments in diverse proteins led to the concept of modular building blocks which are believed to have evolved independently. Depending on the identification method and the focus of the investigation, the domain names and boundaries attributed to a single protein sequence can be quite different. Here, we summarise the usage of the word 'domain' in three main categories – structural domains, functional domains and evolutionary domains – to distinguish between different domain definitions and to facilitate comparisons of similarly defined domains.

A structural domain is a substructure formed by specific regions of a

Structural domains

polypeptide chain, capable of folding independently into a compact, stable entity. A structural domain usually contains between 40 and 350 amino acids, and is the modular unit from which many larger proteins are constructed. The domain boundary information mainly comes from domain assignment of known 3D structures available from the Protein Data Bank (PDB).[12]

Functional domains

A functional domain refers to particular regions in proteins that are responsible for a specific biological function. Functional domains are, in the main, identified by deletion experiments through whittling down proteins to their smallest active fragments using proteinases and recombinant technology. The information on functional domains is scattered in many primary databases such as Swiss-Prot[13] and PubMed.[14]

Evolutionary domains

Evolutionary domains can also be called 'protein modules'. Modules are subsets of domains that can be found in functionally diverse proteins as building blocks (eg the *Src*-homology 2 or SH2 domain).[15] In the early 1990s, it was hypothesised that modules often correspond to single exons with same phase at their intron/exon boundaries.[3] But with the growing body of information, we observe that intron/exon boundaries need not correspond to domain boundaries (Figure 1). The identification of modules usually results from comparative sequence alignment. ProDom[18] and DOMO[10] databases are derived from automated homologous sequence clustering and are rich sources of modules. The domains in the SCOP database[19] were assigned according to evolutionary information and therefore comprise evolutionary domains.

Domain/linker databases

Modules represent contiguous segments of protein sequence, while structural domains are independently folded parts that are not necessarily contiguous. Although the three kinds of domain are identical in many cases, structural domains are not necessarily exactly the same as functional domains, and may not correspond to evolutionary domains. So when we wish to assign domains to a protein sequence, it is critical to decide which category of domains we are interested in and then choose the appropriate databases and methods.

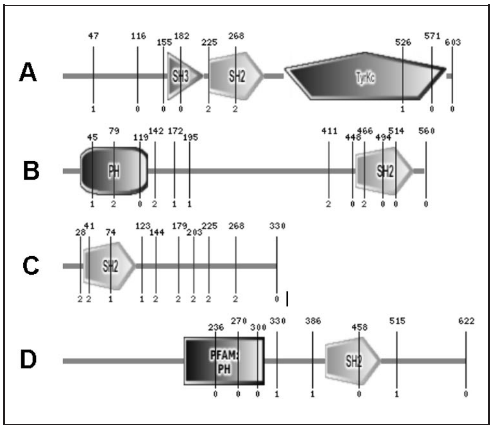


Figure 1: SMART[16] representation of SH2 domain in several proteins shows that module is not necessary to correspond to a single exon. Intron positions are indicated with vertical lines showing the intron phase and exact position in the amino acid sequence. The Ensembl[17] ID for four sequences containing SH2 domains are: (A) CG8049-PA; (B) ENSMUSP00000001110; (C) ENSMUSP00000002216; (D) ENSMUSP00000005188

DOMAIN AND LINKER DATABASES

Before rushing into domain boundary prediction methods, a good understanding of existing domain/linker databases is indispensable. These databases can provide both rich domain boundary information as well as the validation data set for the evaluation of prediction methods. But different databases use different methods to delineate the domain boundary, so that domain boundaries for the same protein can be vastly different.[20] Figure 2 illustrates an example of different domain boundaries assignment for the same protein in different domain databases.

In this paper, we will briefly review the available domain and linker databases. All domain databases can be classified into two categories according to their primary

Sequence-based domain databases

Structure-based domain databases

data source: structure or sequence. The main sequence-based domain databases include ProDom,[18] DOMO,[10] Pfam,[21] SMART,[16] COGs,[22] BLOCKS,[23] SBASE[24] and Interpro.[25] The major structure-based domain databases are SCOP,[18] CATH,[26] 3Dee,[27] Dali/FSSP[28] and MMDB.[29] XdomView[11] provides a quick and easy interface to compare the structural domain definitions from these different databases. The only reported linker database is LinkerDB[30] which contains information on inter-domain linkers. The WWW addresses of these databases and the type of domain information they contain is available from Table 1.

SEQUENCE-BASED DOMAIN DATABASES

Domain databases from sequence alignments

Orthologous sequences

ProDom

The ProDom[18] database is a comprehensive set of protein domain families automatically generated from Swiss-Prot and TrEMBL[13] databases using MKDOM2,[31] which is based on position-specific iterative BLAST (PSI-BLAST).[32] The current release (2003.1) contains 556,964 domain families. Among them, 144,444 have at least two sequence members.

Hidden Markov models

DOMO

DOMO[10] is a database of aligned protein domains constructed from sequence information alone by a fully automated process that involves detection and clustering of similar sequences, domain delineation and multiple sequence alignment. The domain boundaries were inferred from the relative positions of homologous segments.[33] The latest update (1998) of DOMO contains 99,058 domains which are clustered into 8,877 multiple sequence alignments.

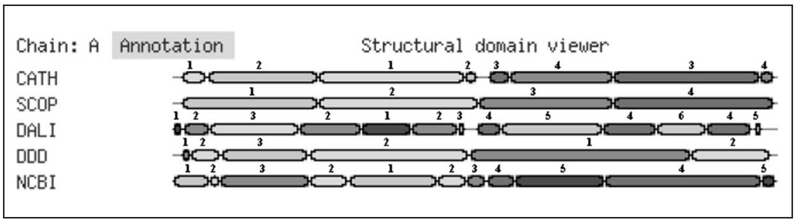


Figure 2: Domain boundaries for D-glucose 6-phosphotransferase (PDB ID: 1HKB, chain A) are dissimilar in different structure-based domain databases. The domain assignments are collated and visualised by XdomView.[11] Segments with the same number are assigned to the same domain

BLOCKS

The BLOCKS[23] database consists of blocks which are ungapped multiple sequence alignments of the most conserved regions of proteins. It is built by automated PROTOMAT system from documented families of related proteins. The current BLOCKS release (Version 14.0, October 2003) includes 24,294 sequence blocks representing 4,944 groups documented in InterPro.[25]

COGs

COGs[22] (Clusters of Orthologous Groups of proteins) database is the delineation of protein sequences encoded in 43 complete genomes by clustering of orthologues, which present 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogues from at least three lineages and thus corresponds to an ancient conserved domain. The COGs database initially contained only the sequenced genome of prokaryotes and unicellular eukaryotes.[34] A recent update to include multicellular eukaryote genomes has enlarged the database to 74,059 COGs and 104,101 proteins from 43 completed genomes.

SMART

SMART[5] (a Simple Modular Architecture Research Tool) is a tool for protein domain identification and annotation and domain architecture representation. The database consists of a library of hidden Markov models (HMMs) which are derived mainly from refined multiple sequence alignment primarily collected from published papers. The domain boundaries are verified with 3D structure, wherever possible, in conjunction with protein N- and C-termini and the known extents of adjacent

Table 1: Databases that contain domain or linker information

Database	URL	Stored information
Sequence-based domain databases		
ProDom	http://prodes.toulouse.inra.fr/prodom/current/html/home.php/	Evolutionary domain
DOMO	http://www.infobiogen.fr/services/domo/	Evolutionary domain
BLOCKS	http://blocks.fhcrc.org/blocks/blocks_search.html	Evolutionary and functional domain
COGs	http://www.ncbi.nlm.nih.gov/COG/	Evolutionary and functional domain
SMART	http://smart.embl-heidelberg.de	Evolutionary, functional and structural domain
Pfam	http://www.sanger.ac.uk/Software/Pfam/	Evolutionary, functional and structural domain
SBASE	http://www.icgeb.trieste.it/sbase/	Evolutionary, functional and structural domain
InterPro	http://www.ebi.ac.uk/interpro/	Evolutionary, functional and structural domain
Structure-based domain databases		
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/	Evolutionary and structural domain
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/	Structural domain
3Dee	http://www.compbio.dundee.ac.uk/3Dee/	Structural domain
Dali/FSSP	http://www.ebi.ac.uk/dali/fssp/	Structural domain
MMDB	http://www.ncbi.nih.gov/Structure/MMDB/mmdb.shtml	Structural and evolutionary domain
XdomView*	http://surya.bic.nus.edu.sg/xdom/	Structural and evolutionary domains
Linker database		
LinkerDB	http://ibivu.cs.vu.nl/programs/linkerdbwww/	Linker derived from 3D structure

*Although not strictly a database, XdomView integrates domain data from all five structure-based domain databases.

domains. The release 4.0 (January 2004) of SMART contains 685 protein domains with extensive annotation for each domain. The latest update for SMART allows the combined representation of detailed gene structure (exon/intron boundaries and phases) and domain architecture, which facilitates investigation of the correlation between exon/intron boundaries and protein domain boundaries.[16]

SBase and InterPro are integrated resources that include domain annotations

Pfam

Pfam[35] is a comprehensive collection of protein domains and families represented by multiple sequence alignments and HMMs. Pfam has two parts: Pfam-A and Pfam-B. Pfam-A includes manually curated families while Pfam-B is derived from ProDom database domains that are not in Pfam-A. To obtain more accurate domain definitions, Pfam makes use of structure information and compares its domain definition with structural domain databases such as SCOP and CATH.[21] The recent release 11.0 (December 2003) of Pfam contains 7,255 families.

SBASE

SBASE[24] is a collection of annotated protein domain sequences. The data sources for SBASE include Swiss-Prot+TrEMBL,[13] PIR,[36] Pfam,[35] SMART[5] and PRINTS.[37] The boundaries of domains are defined by experiment report or homology to known domains. The current version (release 10) includes 1,052,904 protein domain sequences, all of which are clustered into 4,340 functionally or structurally well-characterised domains (SBASE-A) and 1863 less well-characterised groups (SBASE-B).

InterPro

InterPro[25] is an integrated documentation resource for protein families, domains, patterns and functional sites. It is a comprehensive resource that includes information from PROSITE,[38] Pfam, PRINTS, ProDom, SMART and TIGRFAMs.[39] The latest release 7.1 (December 2003) contains 10,403 entries, representing 2,239 domains, 7,901 families, 197 repeats, 26 active sites, 20 binding sites and 20 post translational modifications.

STRUCTURE-BASED DOMAIN DATABASES

SCOP

Domain databases from structural alignments include SCOP, CATH, 3Dee, Dali/FSSP and MMDB

The SCOP[40] (Structural Classification Of Proteins) database is a comprehensive classification of all structures in PDB according to their evolutionary and structural relationship. The domain assignments in SCOP are mainly based on evolutionary relationship and therefore some of the domain definitions are different from other structure-based domain databases. All the domains in SCOP are manually classified according to a four-level hierarchy: Family, Superfamily, Fold and Class. The 1.65 release of SCOP (December 2003) contains 20,619 structures, 54,745 domains, 2,327 families, 1,294 superfamilies, 800 folds and 7 classes.

CATH

CATH[26] is also a hierarchal classification database of protein domain structures, which clustered protein domain in five principal levels: Class (C), Architecture (A), Topology (T), Homologous superfamily (H) and Sequence family (S). The domain definitions were assigned by a consensus procedure based on three algorithms for domain recognition (DETECTIVE,[41] PUU[42] and DOMAK[43]) as well as manual assignment. CATH domains are classified manually at C- and A-level and automatically at T-, H- and S-level. The current available release (v2.5.0, August 2003) of CATH includes 43,299 domains, grouped into 4,036 sequence families, 1,467 superfamilies, 813 topologies, 37 architectures and 4 main classes.

3Dee

3Dee[27] (Database of Protein Domain Definitions) is a comprehensive collection of protein structural domain definitions. The domains in 3Dee are defined on a purely structural basis. DOMAK algorithm[43] was used to define all domains when the database was first built. For later updates, the domains were defined by sequence alignment to existing domain definitions or manually. All the domains in 3Dee were organised in a hierarchy of three levels: Domain families (sequence-redundant domains), Domain sequence families (structure-redundant domains) and Domain structure families (non-redundant on structure).[44] The last release of 3Dee (November 1999) contained 13,767 protein chains and 18,896 domains. These domains were further clustered into 1,715 domain sequence families and 1,199 domain structure families.

Dali/FSSP

Dali/FSSP[28] database presents a fully automatic classification of all known protein structures. The classification is derived using all-against-all comparison of all structures in PDB by an automatic structural alignment method (Dali[45]). The structural domains of the current release (May 2003) are defined by a modified version of ADDA algorithm.[46]

MMDB

MMDB[29] (Molecular Modeling Database) is NCBI Entrez's 3D-structure database derived from the PDB. MMDB contains two kinds of domains: '3D domain' and 'Conserved Domain'.[29] 3D Domains in MMDB are structural domains, which are assigned automatically using an algorithm that searches for one or more breakpoints such that the ratio of intra- to inter-domain contacts falls above a set threshold.[47] Conserved domains in MMDB are recurrent evolutionary modules defined by Entrez's CDD (Conserved Domain Database),[48] where the domains are derived from SMART, Pfam and COGs.

XdomView

XdomView[11] is a Chime-based visualisation tool that integrates and maps the domain boundaries of the input PDB chain obtained from protein structure classification databases (SCOP, CATH, 3Dee, Dali/FSSP and MMDB) to its tertiary structure. It also runs BLAST2 for the input PDB chain sequence against all protein sequences in the ExInt[49] database and maps the intron positions and phases of aligned search results on the input protein's 3D structure. XdomView, a useful visualisation tool for scientists working on gene and protein evolution and structural modelling and classification, is able to provide domain boundary information on a PDB structure simultaneously from the five different structure-based domain databases listed above.

LINKER DATABASE

Linkers are sequence regions between defined structural domains. Linker regions have usually been regarded as unstructured, non-globular or low-complexity segments that are flexible in 3D space,[50] but recent studies show linker regions may significantly affect the cooperation and interaction between domains and therefore alter the overall functionality and efficiency of multiple-domain proteins.[51] A systematic investigation of linker regions has been reported by George and Heringa,[30] resulting in a curated linker database (LinkerDB).

LinkerDB

LinkerDB is derived from the non-redundant structure data set available from NCBI.[30] Linker regions are assigned by extending the domain boundaries determined by Taylor algorithm.[52] All the linkers in LinkerDB were grouped by several criteria: length (small, medium and large); the numbers of intervening linkers separating two domains (1-linker, 2-linker, 3-linker and >3-linker sets); secondary structure type for linkers (helix, strand and loops). Two main types of linkers were identified: helical and non-helical, with distinct properties such as rigidity or amino acid composition. Statistics from the linker database reveal that certain residues (Pro, Arg, Phe, Thr, Glu and Gln) are preferred by linker regions while others (Cys and Gly) are preferentially located within domains. The analysis by George and Heringa[30] suggested the amino acid propensity of inter-domain linkers is distinct from intra-domain loops. The accurate amino acid propensity and other properties of linkers derived from LinkerDB may benefit domain boundary prediction methods.

DOMAIN BOUNDARY PREDICTION METHODS

Currently there are many domain boundary prediction methods available. All these methods can be classified into three categories: comparative methods, clustering methods and *ab initio* methods. Table 2 lists major domain boundary prediction methods.

Comparative domain boundary prediction methods

Each of these methods (SBASE,[24] SUPERFAMILY[53] and Domain Fishing[8]) uses exhaustive sequence searches against known domain definitions within the associated domain database(s). They predict domain boundaries as well as domain content and thus can be used for the identification of protein domain architecture. Their predictions are reliable if a known homologous domain can be detected within their internal database. Comparative methods need prior knowledge about domains. As more and more domains are identified and characterised, it is expected that comparative methods will perform better with novel sequences. Generally, standard sequence database search protocols are used to identify domains, eg PSI-BLAST[32] and HMM. Since most comparative methods are quite similar in principle, only one method is reviewed here.

Domain Fishing

Domain Fishing[8] is targeted to predict domain architecture and identify structural templates for each domain for comparative modelling. PDB, Pfam and

Integrated viewer for domain contents and exon/interon boundaries

Predicting domain boundaries

Linker regions have different properties from domain regions

Comparative methods

Domain architecture prediction

Table 2: Domain boundary prediction methods

Methods	URL or availability	Server or standalone	Features	Input
Comparative methods				
Domain Fishing	http://www.bmm.icnet.uk/servers/3djigsaw/dom_fish/	Server	PSI-BLAST	Single
SBASE	http://www3.icgeb.trieste.it/~sbasesrv/main.html	Server	BLAST	Single
SUPERFAMILY	http://supfam.org	Server	HMM	Single
Clustering methods				
MKDOM	ftp://ftp.toulouse.inra.fr/pub/xdom/	Standalone	Clustering	Large data set
GeneRAGE	http://www.ebi.ac.uk/research/cgg/services/rage/	Standalone	Clustering	Large data set
GEANFAMMER	http://www.mrc-lmb.cam.ac.uk/genomes/geanfammer.html	Standalone	Clustering	Large data set
***Ab initio* methods**				
UMA (Linker prediction)	Available upon request from C. Townsend (ctownsend@jhu.edu)	Standalone	Hydrophobicity and amino acid conservation	MSA
SnapDRAGON	Available upon request from J. Heringa (jhering@nimr.mrc.ac.uk)	Standalone	*Ab initio* 3D models	MSA
DomSSEA	http://bioinf.cs.ucl.ac.uk/dompred/	Server	Secondary structure alignment	MSA
PASS	http://www.bio.gsc.riken.go.jp/PASS/pass_query_sample.htm	Server	Similarity plot	MSA
DomCut (Linker prediction)	http://www.bork.embl-heidelberg.de/~suyama/domcut/	Server, standalone	Amino acid composition	Single
DGS	http://www.ncbi.nlm.nih.gov/Structure/dgs/DGSWeb.cgi	Server, standalone	Sequence length	Single
Entropy profile	ftp://ftp.ncbi.nlm.nih.gov/pub/wheelan/DGS	Standalone	Entropy profile	Single
Combination method				
DomPred	http://bioinf.cs.ucl.ac.uk/dompred/	Server	Pfam search followed by DomSSEA	

SCOP databases have been combined and two sequence databases, dPFAM_PDB and dSCOP, generated, which serve as template domain repositories. Given a query sequence, PSI-BLAST[32] is used to search dPFAM_PDB to predict domain content and boundaries are defined by dSCOP.

Clustering methods for domain boundary prediction

Clustering methods

Iterative BLAST

Unlike comparative methods, clustering methods do not require any prior knowledge for domains. The biological basis for all clustering methods is the modular nature of proteins. Clustering methods will iteratively search against the data set and generate segment sequence clusters. Several databases such as ProDom[18] and DOMO[10] are generated in this manner. Clustering methods are usually applied to large data sets such as Swiss-Prot and TrEMBL, leading to comprehensive derived domain databases. But the biological meaning of these domains may be not clear and sometimes just be artefacts of the specific thresholds applied during clustering. Clustering methods include DOMAINER,[54] MKDOM,[31] GeneRAGE[55] and GEANFAMMER,[56] of which MKDOM is described below.

MKDOM

MKDOM (version 2)[31] is an automatic clustering algorithm used to generate the current release of the ProDom[18] database. It relies on the assumption that the shortest protein sequence corresponds to a single domain. The program iteratively searches the query sequence for matches to the database sequences, starting with the shortest entry, using PSI-BLAST. All significant hits are removed from the query sequence and the remaining fragment(s) are searched, until the database entries are exhausted. Prior to the iterative clustering process, fragmentary sequences (less than the shortest sequence in the database) are removed and low-complexity regions are masked using SEG.[50]

***Ab initio* methods**

Ab initio methods for domain boundary prediction

Ab initio methods attempt to predict domain boundaries in the absence of experimental determined 3D structures or detectable known domain definitions. Physical properties such as domain size distribution[9] (DGS), entropy profiles[57] or differential amino acid composition[7] have been selected as discriminatory criteria. Predicted secondary structure and *ab initio* simulation of 3D structure are also used to make informed boundary predictions.[58,59] The followings are the most popular *ab initio* domain boundary prediction methods.

UMA

UMA[60] (Udwary–Merski Algorithm) is a method for predicting linker regions within large multifunctional proteins. It is relies on three assumptions:

***Ab initio* 3D modelling**

Hydrophobic core

- proteins can be dissected into two kinds of regions: compact, independent folding, bioactive globular regions (domains) and unstructured, flexible regions (linkers);
- amino acids in domain regions are relatively more conserved while linker regions carry more mutations; and
- linker regions are more hydrophilic than domain regions.

According to these assumptions, the propensity of an amino acid in a sequence to be within a linker or a domain is calculated as the weighted sum of three properties (primary sequence similarity, secondary structure similarity and hydrophobicity).

The UMA algorithm provides better predictions than sequence alignments alone, but it also has several limitations:

- the criteria for linker regions based on UMA scores is loosely defined and thus the selection of linkers is subjective, based on user-defined thresholds;
- UMA depends on the availability of detectable homologous sequences of target sequence;
- the input for UMA requires at least two homologous sequences; with prediction reliability increasing with more input sequences;
- sequence alignment quality may strongly affect the reliability of linker prediction, necessitating manual inspection and adjustment of the multiple sequence alignments.

SnapDRAGON

SnapDRAGON[59] is a suite of programs used to predict domain boundaries based on the consistency of a set of *ab initio* 3D structural models. The assumption behind SnapDRAGON is that hydrophobic residues cluster together in space, forming the protein core. This algorithm includes three steps. Firstly, 100 *ab initio* models are generated by the distance-geometry based DRAGON method[61] using multiple sequence alignment and predicted secondary structures as input. Secondly, domain boundaries of these models are assigned using the method of Taylor.[52] Lastly, the final domain boundaries are determined from the consistency of the assigned domain boundaries in the set of alternative 3D models. This method was evaluated with a non-redundant 3D structure data set available from NCBI. The domain definitions of this data set were assigned by Taylor algorithm[52] and validated by SCOP and Dali. The accuracy of domain boundaries prediction is 63.9 per cent for proteins with continuous domains and 35.4 per cent for proteins with discontinuous domains, with an overall accuracy of 51.8 per cent. SnapDRAGON is a reliable method and can predict domain boundaries for protein with discontinuous domains. But it is computational intensive and therefore not suitable for large-scale sequence analysis. It also requires a set of homologous sequences, similar to the target sequence to generate a multiple sequence alignment as input.

DomSSEA

Secondary structure prediction

DomSSEA[58] predicts domain boundaries by aligning secondary structural elements. The secondary structure of a query sequence is first predicted by PSIPRED[62] and this prediction is aligned with known secondary structures of CATH domains. The best matches are reported as predicted domains for the input sequence. This method is not entirely *ab initio* since it depends on CATH domain definitions. At the same time, it differs from the comparative methods in that there is no requirement for detectable sequence similarity. The success rate of this method for assigning domain number correctly is 73.3 per cent and the correct prediction of domain number and location of boundaries is 24 per cent for multiple domain set (±20 residues).

Narrow distribution of domain sizes

DomCut

Amino acid propensity for domain/linker regions

DomCut[7] predicts inter-domain linkers regions using sliding-windows average of linker index derived from a domain/linker data set collected from Swiss-Prot annotation. DomCut uses the difference of amino acid composition between domain and linker regions, while DGS[9] (discussed below) and SnapDRAGON[59] are based on the length distribution of known 3D domain structures and *ab initio* 3D model construction, respectively. The propensity of different amino acids to be located in domain or linker regions is compiled from sequence databases, unlike LinkerDB,[30] which is based on structural data. For example, Pro, Ser and Thr are quite abundant in linker regions while Try, Gly, Cys and Trp prefer to be located within domains. At the default threshold value −0.09, the sensitivity and selectivity for DomCut are 53.5 and 50.1 per cent, respectively.

From our analysis, there are several points in the domain/linker selection criteria of DomCut that need to be addressed:

- Domain/linker definitions derived from structure may define the boundaries of domains more accurately and better represent residue preferences.

- The pre-set range for domains (50–500) and linkers (10–100) may miss some data. In protein structure, short linkers, fewer than 10 residues, are not uncommon.[29]

These changes may result in a better data set and more accurate linker preference profiles.

DGS

DGS[9] (Domain Guess by Size) is based on two observations of domain size distribution:

- Domain sizes follow a narrow distribution (peak at 100 residues).

- Most domains are formed by single continuous segment (83.6 per cent).[9]

These observations are derived from the non-redundant data set selected from PDB and domain definitions were taken from NCBI Entrez.[47] Given the length of target sequence, DGS will enumerate all possible domain boundaries (with a step size of 20 residues) and calculate their relative likelihood according to a likelihood function based on empirical distributions of domain length and segment number. The accuracy of DGS was reported to be 28 per cent for two-domain proteins (±20 residues). Wheelan *et al.*[9] suggest that DGS is more successful for protein sequences shorter than 400 residues with one or two domains. DGS can potentially predict complicated domain organisation including discontinuous domains. For DGS, several top guesses should be considered rather than the first guess, which is always a single domain, owing to the preponderance of single-domain proteins in the data set. DGS is not practical as a domain boundary prediction method alone, but it can be used together with other methods or the prior knowledge of functional regions.

CALCULATION OF ENTROPY PROFILES

Entropy profiles

Galzitskaya and Melnik report a method that predicts domain boundaries based on the calculation of entropy profiles.[57] This method is founded on the hypothesis that segments with high side chain entropy correspond to domain regions, while linker regions have relatively low side chain entropy. The data set is built through selection of SCOP structures with two continuous domains. Redundancy (sequence ID $>$ 80 per cent) and small domains (length $<$ 50 residues) have been removed from the data set. The entropy parameters for each residue have been defined by Galzitskaya *et al.*[63] A sliding window (with a 40 residue window size) is used to average the entropy profiles. The boundaries are predicted by the global minimal of the entropy. The success rate of this method on the data set is 63 per cent (±40 residues). It is worth noting that the data set includes only two-domain proteins with continuous domains, so that the complexity of prediction is significantly reduced. The current version of this method can only be applied to two-domain proteins and is not suitable for proteins with small domains. The success rate may not reflect the real accuracy of this method since the resolution of this method is ±40 residues, which is close to the average size of domain (100 residues according to Wheelan *et al.*[9]).

Among *ab initio* approaches, some methods require a multiple sequence alignment as input. Although this should improve the prediction accuracy, it also has some limitations on sequences that have no known structural homologues.

DISCUSSION

Each category of method discussed above has its own strengths and weaknesses. Comparative methods are accurate and informative but have difficulties when the target sequence has no detectable homologue with known domain information. Clustering methods are better for large data sets but are not applicable for the analysis of a single sequence. *Ab initio* methods are generally not limited by the availability of known homologous domains or data set, but their sensitivities and specificities are significantly lower than those of other methods. The combination of multiple methods may achieve a more reliable and accurate prediction for domain boundaries. So the practical procedure for domain boundary prediction is a stepwise approach. At the outset, one should try to use comparative methods to search the domain databases. If no significant hits are detected, then *ab initio* methods should be tried. Some of the available methods have already adopted such a strategy. For example, the DomPred server[58] first searches the Pfam[21] database to identify known domains, and the *ab initio* method DomSSEA is used only if there are no hits in the first round.

Stepwise strategy

Difficulties in domain boundary prediction

Although there are a variety of methods available for domain boundary prediction, there is room for improvement, especially for *ab initio* methods:

Discontinuous domains

- The boundary prediction for discontinuous domains remains very difficult, especially from *ab initio* approaches. To figure out which segments form a discontinuous domain is a great challenge. Currently the most successful *ab initio* method for predicting discontinuous domains is SnapDRAGON.[59]

- Large multiple domain proteins are more difficult targets for correct domain boundary prediction, since they are more complex and can result in several complex combinatorial domain possibilities.[7,57]

- The complexity of domain boundary prediction is also greatly increased by rearrangements within the domain,

Domain rearrangement

such as the insertion of one domain into another or domain swapping.[64] In the case of potato proteinase inhibitor II (Pot II) family, domain duplication followed by domain swapping results in three topologies for the same fold (SCOP family of plant proteinase inhibitors) in the same protein family (Figure 3). The three types of domain are circularly permuted with respect to each other and, of the three, the type 1 domain seems to be the most stable based on observed data.[65]

The currently available methods cannot discriminate between these three types of structural domains and thus are unable to provide correct prediction for domain boundaries (Kong and Ranganathan, unpublished results).

Acknowledgments
The authors would like to thank their colleagues at the Department of Biochemistry, National University of Singapore for their helpful comments, discussions and support. L.K. gratefully acknowledges the National University of Singapore for the award of an Agency for Science, Technology and Research, Singapore (A*STAR) scholarship.

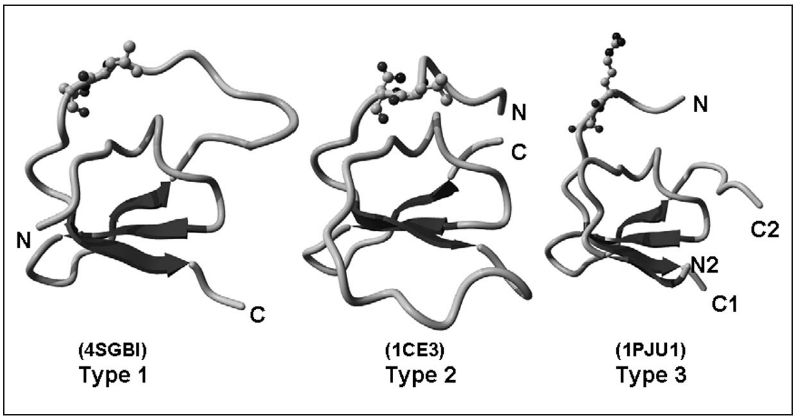


Figure 3: Structure comparison of the three types of topologies in the Pot II family. The structures are shown as a ribbon diagram, with the N- and C-termini marked and the active site residues in ball and stick representation. Type 1: potato inhibitor PCI-1 (PDB ID: 4SGB, chain I); Type 2: putative ancestral inhibitor Api (PDB ID: 1CE3); Type 3: tomato inhibitor-II (PDB ID: 1PJU, domain I)

References

1. Alberts, B., Johnson, A., Lewis, J. *et al.* (2002), 'Molecular Biology of the Cell', Garland Science, New York, pp. 140–146.
2. Wetlaufer, D. B. (1973), 'Nucleation, rapid folding, and globular intrachain regions in proteins', *Proc. Natl Acad. Sci. USA*, Vol. 70, pp. 697–701.
3. Baron, M., Norman, D. G. and Campbell, I. D. (1991), 'Protein modules', *Trends Biochem. Sci.*, Vol. 16, pp. 13–17.
4. Henikoff, S., Greene, E. A., Pietrokovski, S. *et al.* (1997), 'Gene families: The taxonomy of protein paralogs and chimeras', *Science*, Vol. 278, pp. 609–614.
5. Schultz, J., Milpetz, F., Bork, P. and Ponting, C. P. (1998), 'SMART, a simple modular architecture research tool: Identification of signaling domains', *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 5857–5864.
6. Apic, G., Gough, J. and Teichmann, S. A. (2001), 'Domain combinations in archaeal, eubacterial and eukaryotic proteomes', *J. Mol. Biol.*, Vol. 310, pp. 311–325.
7. Suyama, M. and Ohara, O. (2003), 'DomCut: Prediction of inter-domain linker regions in amino acid sequences', *Bioinformatics*, Vol. 19, pp. 673–674.
8. Contreras-Moreira, B. and Bates, P. A. (2002), 'Domain Fishing: A first step in protein comparative modelling', *Bioinformatics*, Vol. 18, pp. 1141–1142.
9. Wheelan, S. J., Marchler-Bauer, A. and Bryant, S. H. (2000), 'Domain size distributions can predict domain boundaries', *Bioinformatics*, Vol. 16, pp. 613–618.
10. Gracy, J. and Argos, P. (1998), 'DOMO: A new database of aligned protein domains', *Trends Biochem. Sci.*, Vol. 23, pp. 495–497.
11. Vivek, G., Tan, T. W. and Ranganathan, S. (2003), 'XdomView: Protein domain and exon position visualization', *Bioinformatics*, Vol. 19, pp. 159–160.
12. Berman, H. M., Westbrook, J., Feng, Z. *et al.* (2000), 'The Protein Data Bank', *Nucleic Acids Res.*, Vol. 28, pp. 235–42.
13. Boeckmann, B., Bairoch, A., Apweiler, R. *et al.* (2003), 'The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003', *Nucleic Acids Res.*, Vol. 31, pp. 365–370.
14. Roberts, R. J. (2001), 'PubMed Central: The GenBank of the published literature', *Proc. Natl Acad. Sci. USA*, Vol. 98, pp. 381–382.
15. Hegyi, H. and Bork, P. (1997), 'On the classification and evolution of protein modules', *J. Protein Chem.*, Vol. 16, pp. 545–551.
16. Letunic, I., Copley, R. R., Schmidt, S. *et al.*

(2004), 'SMART 4.0: Towards genomic data integration', *Nucleic Acids Res.*, Vol. 32, pp. D142–D144.

17. Birney, E., Andrews, D., Bevan, P. *et al.* (2004), 'Ensembl 2004', *Nucleic Acids Res.*, Vol. 32, pp. D468–D470.

18. Servant, F., Bru, C., Carrere, S. *et al.* (2002), 'ProDom: Automated clustering of homologous domains', *Brief. Bioinform.*, Vol. 3, pp. 246–51.

19. Andreeva, A., Howorth, D., Brenner, S. E. *et al.* (2004), 'SCOP database in 2004: Refinements integrate structure and sequence family data', *Nucleic Acids Res.*, Vol. 32, pp. D226–D229.

20. Hadley, C. and Jones, D. T. (1999), 'A systematic comparison of protein structure classifications: SCOP, CATH and FSSP', *Structure*, Vol. 7, pp. 1099–1112.

21. Bateman, A., Coin, L., Durbin, R. *et al.* (2004), 'The Pfam protein families database', *Nucleic Acids Res.*, Vol. 32, pp. D138–D141.

22. Tatusov, R. L., Fedorova, N. D., Jackson, J. D. *et al.* (2003), 'The COG database: An updated version includes eukaryotes', *BMC Bioinformatics*, Vol. 4, pp. 41.

23. Henikoff, J. G., Greene, E. A., Pietrokovski, S. and Henikoff, S. (2000), 'Increased coverage of protein families with the BLOCKS database servers', *Nucleic Acids Res.*, Vol. 28, pp. 228–230.

24. Vlahovicek, K., Kajan, L., Murvai, J. *et al.* (2003), 'The SBASE domain sequence library, release 10: Domain architecture prediction', *Nucleic Acids Res.*, Vol. 31, pp. 403–405.

25. Mulder, N. J., Apweiler, R., Attwood, T. K. *et al.* (2003), 'The InterPro Database, 2003 brings increased coverage and new features', *Nucleic Acids Res.*, Vol. 31, pp. 315–318.

26. Pearl, F. M., Bennett, C. F., Bray, J. E. *et al.* (2003), 'The CATH database: An extended protein family resource for structural and functional genomics', *Nucleic Acids Res.*, Vol. 31, pp. 452–455.

27. Siddiqui, A. S., Dengler, U. and Barton, G. J. (2001), '3Dee: A database of protein structural domains', *Bioinformatics*, Vol. 17, pp. 200–201.

28. Holm, L. and Sander, C. (1998), 'Touring protein fold space with Dali/FSSP', *Nucleic Acids Res.*, Vol. 26, pp. 316–319.

29. Chen, J., Anderson, J. B., DeWeese-Scott, C. *et al.* (2003), 'MMDB: Entrez's 3D-structure database', *Nucleic Acids Res.*, Vol. 31, pp. 474–477.

30. George, R. A. and Heringa, J. (2002), 'An analysis of protein domain linkers: Their classification and role in protein folding', *Protein Eng.*, Vol. 15, pp. 871–879.

31. Gouzy, J., Corpet, F. and Kahn, D. (1999), 'Whole genome protein domain analysis using a new method for domain clustering', *Computers Chem.*, Vol. 23, pp. 333–340.

32. Altschul, S. F., Madden, T. L., Schaffer, A. A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25, pp. 3389–3402.

33. Gracy, J. and Argos, P. (1998), 'Automated protein sequence database classification. II. Delineation of domain boundaries from sequence similarities', *Bioinformatics*, Vol. 14, pp. 174–187.

34. Tatusov, R. L., Galperin, M. Y., Natale, D. A. and Koonin, E. V. (2000), 'The COG database: A tool for genome-scale analysis of protein functions and evolution', *Nucleic Acids Res.*, Vol. 28, pp. 33–36.

35. Bateman, A., Birney, E., Cerruti, L. *et al.* (2002), 'The Pfam protein families database', *Nucleic Acids Res.*, Vol. 30, pp. 276–280.

36. Wu, C. H., Yeh, L. S., Huang, H. *et al.* (2003), 'The Protein Information Resource', *Nucleic Acids Res.*, Vol. 31, pp. 345–347.

37. Attwood, T. K. (2002), 'The PRINTS database: A resource for identification of protein families', *Brief. Bioinform.*, Vol. 3, pp. 252–263.

38. Falquet, L., Pagni, M., Bucher, P. *et al.* (2002), 'The PROSITE database, its status in 2002', *Nucleic Acids Res.*, Vol. 30, pp. 235–238.

39. Haft, D. H., Selengut, J. D. and White, O. (2003), 'The TIGRFAMs database of protein families', *Nucleic Acids Res.*, Vol. 31, pp. 371–373.

40. Lo Conte, L., Ailey, B., Hubbard, T. J. *et al.* (2000), 'SCOP: A structural classification of proteins database', *Nucleic Acids Res.*, Vol. 28, pp. 257–259.

41. Swindells, M. B. (1995), 'A procedure for detecting structural domains in proteins', *Protein Sci.*, Vol. 4, pp. 103–112.

42. Holm, L. and Sander, C. (1994), 'Parser for protein folding units', *Proteins*, Vol. 19, pp. 256–268.

43. Siddiqui, A. S. and Barton, G. J. (1995), 'Continuous and discontinuous domains: An algorithm for the automatic generation of reliable protein domain definitions', *Protein Sci.*, Vol. 4, pp. 872–884.

44. Dengler, U., Siddiqui, A. S. and Barton, G. J. (2001), 'Protein structural domains: Analysis of the 3Dee domains database', *Proteins*, Vol. 42, pp. 332–344.

45. Holm, L. and Sander, C. (1993), 'Protein structure comparison by alignment of distance matrices', *J. Mol. Biol.*, Vol. 233, pp. 123–138.

46. Heger, A. and Holm, L. (2003), 'Exhaustive

enumeration of protein domain families', *J. Mol. Biol.*, Vol. 328, pp. 749–767.

47. Madej, T., Gibrat, J. F. and Bryant, S. H. (1995), 'Threading a database of protein cores', *Proteins*, Vol. 23, pp. 356–369.

48. Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C. *et al.* (2003), 'CDD: A curated Entrez database of conserved domain alignments', *Nucleic Acids Res.*, Vol. 31, pp. 383–387.

49. Sakharkar, M., Long, M., Tan, T. W. and de Souza, S. J. (2000), 'ExInt: An exon/intron database', *Nucleic Acids Res.*, Vol. 28, pp. 191–192.

50. Wootton, J. C. (1994), 'Non-globular domains in protein sequences: Automated segmentation using complexity measures', *Comput. Chem.*, Vol. 18, pp. 269–285.

51. Gokhale, R. S. and Khosla, C. (2000), 'Role of linkers in communication between protein modules', *Curr. Opin. Chem. Biol.*, Vol. 4, pp. 22–27.

52. Taylor, W. R. (1999), 'Protein structural domain identification', *Protein Eng.*, Vol. 12, pp. 203–216.

53. Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001), 'Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure', *J. Mol. Biol.*, Vol. 313, pp. 903–919.

54. Sonnhammer, E. L. and Kahn, D. (1994), 'Modular arrangement of proteins as inferred from analysis of homology', *Protein Sci.*, Vol. 3, pp. 482–492.

55. Enright, A. J. and Ouzounis, C. A. (2000), 'GeneRAGE: A robust algorithm for sequence clustering and domain detection', *Bioinformatics*, Vol. 16, pp. 451–457.

56. Park, J. and Teichmann, S. A. (1998), 'DIVCLUS: An automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins', *Bioinformatics*, Vol. 14, pp. 144–150.

57. Galzitskaya, O. V. and Melnik, B. S. (2003), 'Prediction of protein domain boundaries from sequence alone', *Protein Sci.*, Vol. 12, pp. 696–701.

58. Marsden, R. L., McGuffin, L. J. and Jones, D. T. (2002), 'Rapid protein domain assignment from amino acid sequence using predicted secondary structure', *Protein Sci.*, Vol. 11, pp. 2814–2824.

59. George, R. A. and Heringa, J. (2002), 'SnapDRAGON: A method to delineate protein structural domains from sequence data', *J. Mol. Biol.*, Vol. 316, pp. 839–851.

60. Udwary, D. W., Merski, M. and Townsend, C. A. (2002), 'A method for prediction of the locations of linker regions within large multifunctional proteins, and application to a type I polyketide synthase', *J. Mol. Biol.*, Vol. 323, pp. 585–598.

61. Aszodi, A., Gradwell, M. J. and Taylor, W. R. (1995), 'Global fold determination from a small number of distance restraints', *J. Mol. Biol.*, Vol. 251, pp. 308–326.

62. Jones, D. T. (1999), 'Protein secondary structure prediction based on position-specific scoring matrices', *J. Mol. Biol.*, Vol. 292, pp. 195–202.

63. Galzitskaya, O. V., Surin, A. K. and Nakamura, H. (2000), 'Optimal region of average side-chain entropy for fast protein folding', *Protein Sci.*, Vol. 9, pp. 580–586.

64. Heringa, J. and Taylor, W. R. (1997), 'Three-dimensional domain duplication, swapping and stealing', *Curr. Opin. Struct. Biol.*, Vol. 7, pp. 416–421.

65. Scanlon, M. J., Lee, M. C., Anderson, M. A. and Craik, D. J. (1999), 'Structure of a putative ancestral protein encoded by a single sequence repeat from a multidomain proteinase inhibitor gene from *Nicotiana alata*', *Structure Fold. Des.*, Vol. 7, pp. 793–802.

SDPS: Small Disulphide-bonded Proteins Structural Database

Lesheng Kong[1], Tin Wee Tan[1] and Shoba Ranganathan[1, 2]
[1]Department of Biochemistry & [2]Department of Biological Sciences,
National University of Singapore, Singapore

Introduction

Small Disulphide-bonded Proteins (SDP) is a class of small proteins (length <100 a.a) that contains at least one disulphide bridge. Its members include varieties of proteins, such as insulin, inhibitors and toxins. They are an abundant resource of potential therapeutic drugs. A major problem in the structure prediction of SDP is to figure out their disulphide connectivity as this largely determines their fold. Their small sizes and complex disulphide connectivity make them distinct from large globular proteins, requiring specialized applications and datasets.

Our comprehensive Small Disulphide-bonded Proteins Structural (SDPS) database aims to facilitate research on SDP and disulphide connectivity. Data sources include PDB [1], SCOP [2], ASTRAL [3] and DSSP [4]. A number of features have been extracted and calculated, the most important being the disulphide connectivity, which cannot be easily obtained through public databases. SDPS database is accessible at http://origin.bic.nus.edu.sg/sdps.

A key feature of the database will be the introduction of a hierarchical classification system based on the number of disulfide bridges, connectivity and sequence similarity.

Methodology

Data flow in SDPS

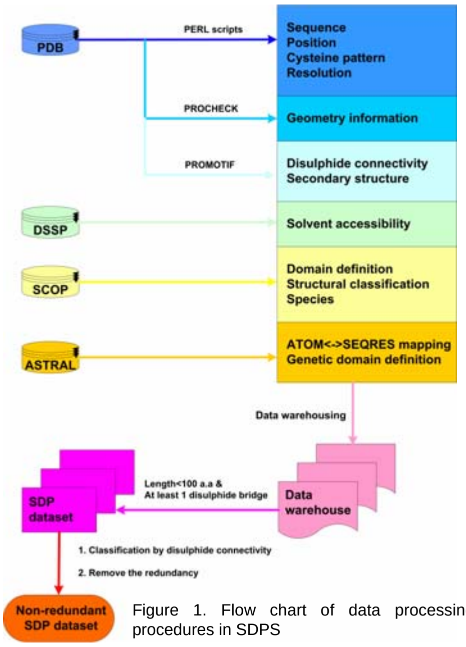


Figure 1. Flow chart of data processing procedures in SDPS

SDPS classification hierarchy

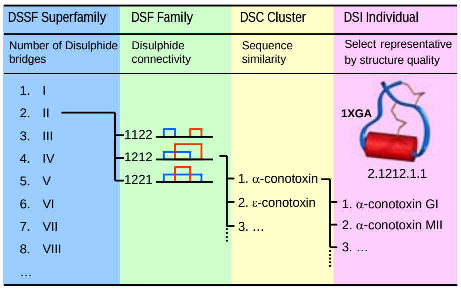


Figure 2. SDPS numbering scheme for representative structure α-conotoxin GI (PDB code: 1XGA) which has 2 disulphide bonds and 1212 connectivity. Classification into Disulphide Superfamily (DSSF), Disulphide Family (DSF), Disulphide Cluster (DSC) and Disulphide Individual (DSI)

Results and discussion

Database statistics

Entities	3-D structures	Protein chains	Protein domains	Disulphide bonds
Number	990	1285	1280	3619

Table 1. SDP dataset content

Plot of number of protein chains against number of disulphide bridges

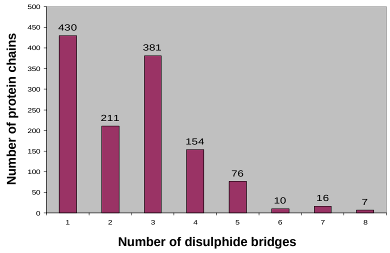


Figure 3. Histogram of the distribution of disulphide bonds number in SDP dataset

Histogram of disulphide distance

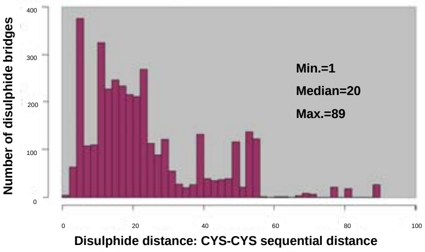


Figure 4. Histogram of the distribution of disulphide distance in SDP dataset

Preliminary analysis of the current dataset reveals some interesting statistics. For example, among 855 protein chains which have multiple disulphide bonds, 238 chains have α−ω (i.e. first and last cysteine residues) type disulphide connectivity. Furthermore only a small portion (5%) of the 1285 protein chains have free cysteines.

Analysis of DSSF VI

We analyzed DSSF VI which has six disulfide bridges and ten members. Based on sequence identities, redundant structures were removed and only seven non-redundant structures remain (PDB codes: 1FVL, 1KST, 1F5Y, 1HJ7, 1HZ8, 1DQB, 1EMN). Theoretically, there are 11 x 9 x 7 x 5 x 3 = 10,395 possible connectivities. These seven structures belonging to just six DSFs. 1EMN (fibrillin) and 1HJ7 (LDL receptor) both share the same connectivity. Interestingly, they have very similar topology although they share low sequence identities (34.1%).

(A)

```
1HJ7 1    ---GTNECLDNNGGCSH-VCNDLKIGYECLCPDGFQLVAQRRCEDI 42
             ..:||.:.: .|.| .|.:....|.|.||.|: ::|...|.|.
1EMN 1    SAVDMDECKEPD-VCKHGQCINTDGSYRCECPFGY-ILAGNECVDT 44

1HJ7 43           DECQDPDTCSQ-LCVNLEGGYKCQCEEGFQLDPHTKACK  80
                  |||...:.|.. .|.|:.||::|.|||||:..| ...|:
1EMN 45           DECSVGNPCGNGTCKNVIGGFECTCEEGFEPGP-MMTCE  82
```

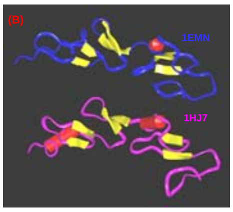


Figure 5. The structure and sequence comparison of 1EMN and 1HJ7. The RMSD between two structures is 1.68 Å.

Applications and future work

- Flexible web interface for easy access.
- Removal of redundancy using classification
- Specialized parameters to be derived:
  - A scoring matrix for SDP alignments
  - Stereochemical parameters for SDP geometrical quality evaluation
- Prediction of disulphide connectivity for SDP.
- 3D structure predictions of SDP using predicted disulphide connectivity.

Acknowledgements

LK gratefully acknowledges the award of a National Science and Technology Board research scholarship from the National University of Singapore and suggestions from Prof. Michael James, University of Alberta, Canada.

References

1 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235-242

2 Lo Conte, L. *et al.* (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.* 28, 257-259

3 Chandonia, J.M. *et al.* (2002) ASTRAL compendium enhancements. *Nucleic Acids Res.* 30, 260-263

4 Kabsch, W. and Sander, C. (1983) Dictionary of Protein Secondary Structures: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 22, 2577-2637

SDPS web address: http://origin.bic.nus.edu.sg/sdps

contact: lesheng@bic.nus.edu.sg

Topology and parameter development for non-standard residues in conopeptides

Lesheng Kong[1], Tin Wee Tan[1] and Shoba Ranganathan[1, 2]
[1]Department of Biochemistry & [2]Department of Biological Sciences,
National University of Singapore, Singapore

Introduction

Conopeptides mainly come from predatory cone snails. They are notable for their unprecedented selectivity and specificity for varieties of neuronal receptors and ion channels. These properties make conopeptides very useful in studies aimed identifying receptors and their ligands, as well as in drug development [1]. In GenBank (up to Sep. 2003), there are 881 conopeptides, among them only 61 of them have 3D structures in PDB. We developed an automatic comparative modeling method to predict structures for conopeptides. During the model development, one big obstacle was the presence of common post-translational modifications. Figure 1 shows 6 non-standard residues in conopeptides.

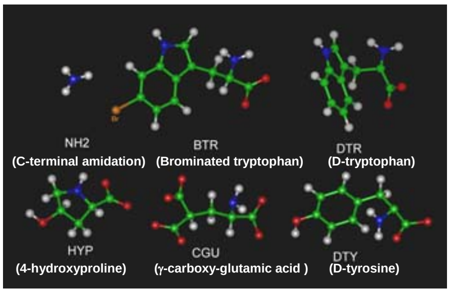


Figure 1. Non-standard residues in conopeptides

Previous studies have shown that some modifications are crucial for the observed functions of conopeptides and their affinity for ion channels [2]. Post-translational modifications may considerably contribute to the structure, affinity and specificity of conopeptides. But traditional molecular modeling methods can neither recognize templates containing non-standard residues nor generate models with these non-standard residues.

To address these problems, we tried to develop special method that can recognize and make use of non-standard residues for comparative modeling of conopeptides. Our strategy is to define the charmm22 forcefield [3] library files for non-standard residues and incorporate them into a commonly used program MODELLER [4].

Methodology

Topology definition

(1) Get the coordinates of non-standard residues from selected high resolution structures.
(2) Check the structure for errors or missing atoms and add missing atoms and hydrogens.
(3) Assign partial charges and forcefield parameters based on atom types.
(4) Inspect topology files for improper dihedral angles and fix if necessary.

Parameter development

Parameters are derived from similar entries in charmm22 forcefield and good quality structures for the following types:
(1) Bond length: $V_{bond} = K_b (b - b_0)^2$
(2) Bond angle: $V_{angle} = K_\theta (\theta - \theta_0)^2$
(3) Dihedral angles: $V_{dihedral} = K_\Phi(1 + \cos(n\Phi - \delta))$
(4) Improper dihedral angles: $V_{improper} = K_\varphi(\varphi - \varphi_0)^2$
The force constants K_b, K_θ, K_Φ, K_φ, were extrapolated from similar atom type entries and equilibrium values b_0, θ_0, δ, φ_0 were calculated from selected structures.

Benchmarking

We selected all conopeptides which have structures available in PDB as well as non-standard residues in their structures (total 19).

A CASP-like benchmarking were done to compare the modeling before and after the incorporation of non-standard residues library files. Jacknife (leave-one-out) technique was applied into our benchmarking due to low sample size. We use generated models with and without the use of non-standard residues for comparative modeling of these 19 conopeptides.

The models generated by both methods were compared to their experimental structures and the RMSDs were calculated.

Result and Discussion

Basing on the above strategy, libraries files for 6 non-standard residues were developed and incorporated into MODELLER library.

Benchmarking was done to validate these libraries files. The results were shown in Figure 2.

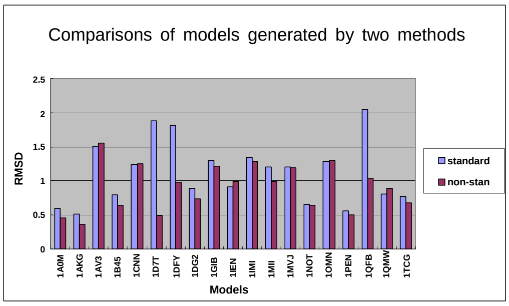


Figure 2. Comparisons of models generated by two methods

From Figure 2, we can see that the difference between two kinds of models (ΔRMSD) can be clustered into two groups: (1) ΔRMSD < 0.30Å; (2) ΔRMSD > 0.80Å. 3 models (ΔRMSD for 1D7T, 1DFY and 1QFB is 1.39Å, 0.83Å and 1.01Å, respectively) are significantly improved after the incorporation of new topologies and parameters. This is to be expected since there are D-amino acids present (DTR in 1DFY and 1QFB and DTY in 1D7T). Standard modeling packages can only deal with L-residues leading to considerable error in backbone conformation.

Figure 3 shows the dramatic backbone improvement in the model for 1QFB when non-standard residues are used.

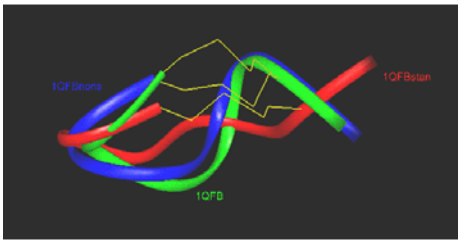


Figure 3. The overlay of two models with real structure. 1QFBstan (red, only standard residues); 1QFBnons (blue, includes non-standard residues); 1QFB (green, real structure).

For other non-standard residues such as HYP and NH2, models do not show big difference on backbone conformations between two methods. The difference between HYP and PRO lies in exposed side chain and NH2 is only c-terminal amidation. They will not affect backbone conformation apparently. But they may change the hydrophobic/hydrophilic property or electrostatic potential of the protein surface. This is subject to further analysis.

In summary, the comparison of models with non-standard residues and those with only standard residues:

Worse	Similar	Better
ΔRMSD < -0.5Å	|ΔRMSD| < 0.5Å	ΔRMSD > 0.5Å
None	16 models	3 models

The significance of residues leading to model improvement: DTR, DTY > CGU, BTR, HYP, NH2

Applications

With the facilitation of these topology and parameter files, it is possible to do comparative modeling on conopeptides that includes non-standard residues, and enhance the accuracy of modeling. The topology and parameter files for these 6 non-standard residues are available on request (Email: lesheng@bic.nus.edu.sg).

Acknowledgements

LK gratefully acknowledges the award of a NSTB scholarship from NUS and suggestions from Dr. Betty Cheng, Stanford University, USA.

References

1 Shen, G.S. *et al*. (2000) Conopeptides: From deadly venoms to novel therapeutics. Drug Discovery Today 5, 98-106
2 Craig, A.G. *et al* (1999) Post-translationally modified neuropeptides from Conus venoms. *European Journal of Biochemistry* 264, 271-275
3 MacKerell, A.D., *et al.* (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B, 102, 3586-3616.
4 Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* 234, 779-815

SDPMOD: A Comprehensive Comparative Modeling Server for Small Disulphide-bonded Proteins

Lesheng Kong[1], Bernett Teck Kwong Lee[1], Joo Chuan Tong[1], Tin Wee Tan[1] and Shoba Ranganathan[1, 2,*]

[1]Department of Biochemistry, National University of Singapore, 8 Medical Drive, 117597, Singapore
[2]Biotechnology Research Institute, Macquarie University, Sydney, NSW 2109, Australia

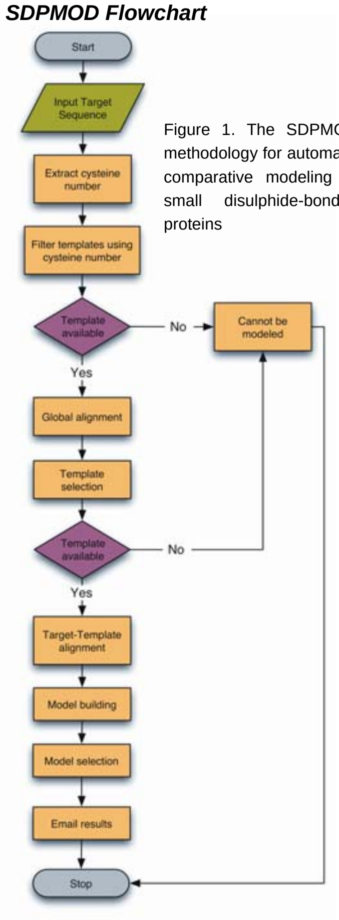
Introduction

Small Disulphide-bonded Proteins (SDPs) are a special class of proteins that are relatively small in size (length<100 residues) and have disulphide bonds within their 3D structures. SDPs include many secretory proteins which serve predatory, defensive or regulatory roles (such as toxins, inhibitors and hormones) and they are rich source for therapeutic drugs and pesticides. Designing drugs from these proteins requires 3D structural information, which is only available for a subset of these proteins.

SDPMOD addresses this deficit in structural information by providing a freely available comprehensive comparative modeling service (http://proline.bic.nus.edu.sg/sdpmod) to the research community [1].

Methodology

SDPMOD Flowchart



Start

Input Target Sequence

Extract cysteine number

Filter templates using cysteine number

Template available

No

Cannot be modeled

Yes

Global alignment

Template selection

Template available

No

Yes

Target-Template alignment

Model building

Model selection

Email results

Stop

Figure 1. The SDPMOD methodology for automatic comparative modeling of small disulphide-bonded proteins

Non-redundant SDPs structure dataset

Before the modeling can proceed, a non-redundant dataset for SDPs needs to be created to serve as the template repository. Structures containing protein chains of length less than 100 aa with at least two cysteines were retrieved from the Protein Data Bank (PDB) [2] and loaded into MySQL for flexible query and manipulation. The redundancies in SDP structures were removed according to structure quality.

Modeling Procedure

The SDPMOD server performs comparative modeling in the four steps: (i) template selection, (ii) target-template alignment, (iii) model building, and (iv) model evaluation. Figure 1 shows the detailed modeling procedure for automated modeling. Target-template alignment and model building are achieved by MODELLER [3] using a customized matrix to ensure that all the cysteine residues are well aligned. The overall structural quality of the generated models are evaluated by PROCHECK [4].

Result and Discussion

Benchmarking

A large-scale benchmarking was completed using SDPMOD server. A control set of 664 sequences (a subset of our SDPs non-redundant dataset) with known structures was used to evaluate the reliability of server. The Cα RMSD values between models and their actual experimental structures were calculated. The results are summarized in Table 1.

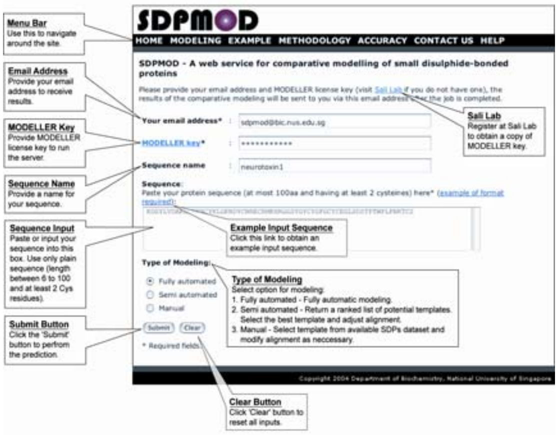
ID (%)	No. of models	No. of models RMSD<0.5Å	No. of models 0.5Å≤RMSD<1 Å	No. of models 1Å≤RMSD<1.5 Å	No. of models 1.5Å≤RMSD<2 Å	No. of models 2Å≤RMSD
20-30	172	0	0	23	105	44
30-40	93	0	3	34	46	10
40-50	56	0	5	29	20	2
50-60	55	0	11	24	16	4
60-70	53	0	13	24	15	1
70-80	54	4	12	18	16	4
80-90	91	9	19	32	28	3
90-95	90	13	19	32	23	3
Total	664	26	82	216	253	71

Table 1. Probabilities of SDPMOD accuracy for target-template identity classes.

The benchmarking results show SDPMOD can predict 3D models with a reasonable accuracy. For example, in the 40-70% sequence identity range, 64% of models have Cα RMSD values less than 1.5 Å.

Web Service

SDPMOD is freely accessible to academic or non-profit users via a web interface (shown in Figure 2) at http://proline.bic.nus.edu.sg/sdpmod.



SDPMOD - A web service for comparative modelling of small disulphide-bonded proteins

Please provide your email address and MODELLER license key (visit Sali Lab if you do not have one), the results of the comparative modeling will be sent to you via this email address after the job is completed.

Your email address* : sdpmod@bic.nus.edu.sg

MODELLER key* : ***********

Sequence name : neurotoxin1

Sequence:
Paste your protein sequence (at most 100aa and having at least 2 cysteines) here* (example of format required):

Type of Modeling:

● Fully automated
○ Semi automated
○ Manual

* Required fields

Copyright 2004 Department of Biochemistry, National University of Singapore

Menu Bar Use this to navigate around the site.

Email Address Provide your email address to receive results.

MODELLER Key Provide MODELLER license key to run the server.

Sequence Name Provide a name for your sequence.

Sequence Input Paste or input your sequence into this box. Use only plain sequence (length between 6 to 100 and at least 2 Cys residues).

Submit Button Click the 'Submit' button to perfrom the prediction.

Sali Lab Register at Sali Lab to obtain a copy of MODELLER key.

Example Input Sequence Click this link to obtain an example input sequence.

Type of Modeling Select option for modeling:
1. Fully automated - Fully automatic modeling.
2. Semi automated - Return a ranked list of potential templates. Select the best template and adjust alignment.
3. Manual - Select template from available SDPs dataset and modify alignment as neccessary.

Clear Button Click 'Clear' button to reset all inputs.

Figure 2. The web interface of SDPMOD.

SDPMOD is primarily designed as a fully automated procedure for easy of use. However due to the complexity of comparative modeling, human intervention and expert knowledge may be required for optimal modeling of some proteins. To allow for human intervention, the current version of the SDPMOD server provides three modes of modeling (fully automated, semi-automated and manual) to meet the different needs of the expert users.

The manual mode permits the expert users to specify desired template, and the semi-automated mode allows users to select the template from a suggested list. Besides the selection of templates, expert users can edit the target-template alignment thus allowing further customization of the modeling process.

After the modeling process is completed, a link with the prediction results will be returned via email. Users can refer to the link to view the prediction result and download the models. The prediction results consist of: (i) a summary of the selected template(s), (ii) the predicted model based on each template in PDB format and (iii) a brief report for each modeling attempt that include the target-template alignment used in modeling building, a comparison of the model against the template by means of RMSD and a PROCHECK report on the stereochemical quality of the models.

References

1 Kong, L., Lee, B.T., Tong, J.C., Tan, T.W. and Ranganathan, S. (2004) SDPMOD: an automated comparative modeling server for small disulphide-bonded proteins. Nucleic Acids Res., 32, W356-W359.

2 Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. Nucleic Acids Res., 28, 235-242.

3 Sali, A and Blundell, T.L. (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints. J. Mol. Biol., 234, 779-815

4 Laskowski, R.A., Moss, D.S. and Thornton, J.M. (1993) Main-chain bond lengths and bond angles in protein structures. J. Mol. Biol., 231, 1049-1067

SDPMOD web address: http://proline.bic.nus.edu.sg/sdpmod

contact: lesheng@bic.nus.edu.sg