Lesheng Kong is a graduate student at the Department of Biochemistry, National University of Singapore, working on structural bioinformatics analysis of small disulphide-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

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
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: 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: 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.

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 (e.g., 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
A 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).

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 and PubMed.

Evolutionary domains can also be called ‘protein modules’. Modules are subsets of domains that can be found in functionally diverse proteins as building blocks (e.g., the Src-homology 2 or SH2 domain). In the early 1990s, it was hypothesised that modules often correspond to single exons with the same phase at their intron/exon boundaries. 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 and DOMO databases are derived from automated homologous sequence clustering and are rich sources of modules. The domains in the SCOP database were assigned according to evolutionary information and therefore comprise evolutionary domains.

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.

**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. 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...
data source: structure or sequence. The main sequence-based domain databases include ProDom, DOMO, Pfam, SMART, COGs, BLOCKS, SBASE, and Interpro. The major structure-based domain databases are SCOP, CATH, 3Dee, Dali/FSSP, and MMDB. XdomView provides a quick and easy interface to compare the structural domain definitions from these different databases. The only reported linker database is LinkerDB 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**

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

**DOMO**
DOMO 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. The latest update (1998) of DOMO contains 99,058 domains which are clustered into 8,877 multiple sequence alignments.

**BLOCKS**
The BLOCKS 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.

**COGs**
COGs (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. 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 (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 domains.
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

Pfam
Pfam35 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
SBASE24 is a collection of annotated protein domain sequences. The data sources for SBASE include Swiss-Prot+TrEMBL,13 PIR,36 Pfam,35 SMART5 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).

Table 1: Databases that contain domain or linker information

<table>
<thead>
<tr>
<th>Database</th>
<th>URL</th>
<th>Stored information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence-based</strong></td>
<td><strong>domain databases</strong></td>
<td></td>
</tr>
<tr>
<td>DOMO</td>
<td><a href="http://www.infobiogen.fr/services/domai/">http://www.infobiogen.fr/services/domai/</a></td>
<td>Evolutionary domain</td>
</tr>
<tr>
<td>BLOCKS</td>
<td><a href="http://blocks.fhcrc.org/blocks/blocks_search.html">http://blocks.fhcrc.org/blocks/blocks_search.html</a></td>
<td>Evolutionary and functional domain</td>
</tr>
<tr>
<td>SMART</td>
<td><a href="http://smart.embl-heidelberg.de">http://smart.embl-heidelberg.de</a></td>
<td>Evolutionary, functional and structural domain</td>
</tr>
<tr>
<td>Pfam</td>
<td><a href="http://www.sanger.ac.uk/Software/Pfam/">http://www.sanger.ac.uk/Software/Pfam/</a></td>
<td>Evolutionary, functional and structural domain</td>
</tr>
<tr>
<td>SBASE</td>
<td><a href="http://www.icgeb.trieste.it/sbase/">http://www.icgeb.trieste.it/sbase/</a></td>
<td>Evolutionary, functional and structural domain</td>
</tr>
<tr>
<td>InterPro</td>
<td><a href="http://www.ebi.ac.uk/interpro/">http://www.ebi.ac.uk/interpro/</a></td>
<td>Evolutionary, functional and structural domain</td>
</tr>
<tr>
<td><strong>Structure-based</strong></td>
<td><strong>domain databases</strong></td>
<td></td>
</tr>
<tr>
<td>SCOP</td>
<td><a href="http://scop.mrc-lmb.cam.ac.uk/scop/">http://scop.mrc-lmb.cam.ac.uk/scop/</a></td>
<td>Evolutionary and structural domain</td>
</tr>
<tr>
<td>CATH</td>
<td><a href="http://www.biochem.ucl.ac.uk/bsm/cath/">http://www.biochem.ucl.ac.uk/bsm/cath/</a></td>
<td>Structural domain</td>
</tr>
<tr>
<td>3Dec</td>
<td><a href="http://www.compbio.dundee.ac.uk/3Dec/">http://www.compbio.dundee.ac.uk/3Dec/</a></td>
<td>Structural domain</td>
</tr>
<tr>
<td>Dali/FSSP</td>
<td><a href="http://www.ebi.ac.uk/dali/fssp/">http://www.ebi.ac.uk/dali/fssp/</a></td>
<td>Structural domain</td>
</tr>
<tr>
<td>XdomView*</td>
<td><a href="http://surya.bio.nus.edu.sg/xdom/">http://surya.bio.nus.edu.sg/xdom/</a></td>
<td>Structural and evolutionary domains</td>
</tr>
<tr>
<td><strong>Linker database</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LinkerDB</td>
<td><a href="http://ibivu.cs.vu.nl/programs/linkerdbww/">http://ibivu.cs.vu.nl/programs/linkerdbww/</a></td>
<td>Linker derived from 3D structure</td>
</tr>
</tbody>
</table>

*Although not strictly a database, XdomView integrates domain data from all five structure-based domain databases.
binding sites and 20 post translational modifications.

**STRUCTURE-BASED DOMAIN DATABASES**

**SCOP**
The SCOP (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 is also a hierarchy 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, PUU and DOMAK) 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 (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 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). 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 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). The structural domains of the current release (May 2003) are defined by a modified version of ADDA algorithm.

**MMDB**
MMDB (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’. 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. Conserved domains in MMDB are recurrent evolutionary modules defined by Entrez’s CDD (Conserved Domain Database), where the domains are derived from SMART, Pfam and COGs.

**XdomView**
XdomView 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 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, 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. A systematic investigation of linker regions has been reported by George and Heringa, resulting in a curated linker database (LinkerDB).

**LinkerDB**

LinkerDB is derived from the non-redundant structure data set available from NCBI. Linker regions are assigned by extending the domain boundaries determined by Taylor algorithm. 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 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, SUPERFAMILY and Domain Fishing) 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 and HMM. Since most comparative methods are quite similar in principle, only one method is reviewed here.

**Domain Fishing**

Domain Fishing is targeted to predict domain architecture and identify structural templates for each domain for comparative modelling. PDB, Pfam and
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 is used to search dPFAM_PDB to predict domain content and boundaries are defined by dSCOP.

### Clustering methods for domain boundary prediction

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, DOMO 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, MKDOM, GeneRAGE and GEANFAMMER, of which MKDOM is described below.

### MKDOM

MKDOM (version 2) is an automatic clustering algorithm used to generate the current release of the ProDom 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.

### Table 2: Domain boundary prediction methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>URL or availability</th>
<th>Server or standalone</th>
<th>Features</th>
<th>Input</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparative methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domain Fishing</td>
<td><a href="http://www.bmm.icnet.uk/servers/3djigsaw/dom_fish/">http://www.bmm.icnet.uk/servers/3djigsaw/dom_fish/</a></td>
<td>Server</td>
<td>PSI-BLAST</td>
<td>Single</td>
</tr>
<tr>
<td>SBASE</td>
<td><a href="http://www3.icgeb.trieste.it/~sbasesrv/main.html">http://www3.icgeb.trieste.it/~sbasesrv/main.html</a></td>
<td>Server</td>
<td>BLAST</td>
<td>Single</td>
</tr>
<tr>
<td>SUPERFAMILY</td>
<td><a href="http://supfam.org">http://supfam.org</a></td>
<td>Server</td>
<td>HMM</td>
<td>Single</td>
</tr>
<tr>
<td><strong>Clustering methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneRAGE</td>
<td><a href="http://www.ebi.ac.uk/research/cgg/services/rage/">http://www.ebi.ac.uk/research/cgg/services/rage/</a></td>
<td>Standalone</td>
<td>Clustering</td>
<td>Large data set</td>
</tr>
<tr>
<td>GEANFAMMER</td>
<td><a href="http://www.mrc-lmb.cam.ac.uk/genomes/geanfammer.html">http://www.mrc-lmb.cam.ac.uk/genomes/geanfammer.html</a></td>
<td>Standalone</td>
<td>Clustering</td>
<td>Large data set</td>
</tr>
<tr>
<td><strong>Ab initio methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UMA (Linker prediction)</td>
<td>Available upon request from C. Townsend</td>
<td>Standalone</td>
<td>Hydrophobicity and amino acid conservation</td>
<td>MSA</td>
</tr>
<tr>
<td>SnapDRAGON</td>
<td>Available upon request from J. Heringa</td>
<td>Standalone</td>
<td>Ab init 3D models</td>
<td>MSA</td>
</tr>
<tr>
<td>DomSSEA</td>
<td><a href="http://bioinf.cs.ucl.ac.uk/dompred/">http://bioinf.cs.ucl.ac.uk/dompred/</a></td>
<td>Server</td>
<td>Secondary structure alignment</td>
<td>MSA</td>
</tr>
<tr>
<td>PASS</td>
<td><a href="http://www.bio.gsc.riken.go.jp/PASS/pass_query_sample.htm">http://www.bio.gsc.riken.go.jp/PASS/pass_query_sample.htm</a></td>
<td>Server</td>
<td>Similarity plot</td>
<td>MSA</td>
</tr>
<tr>
<td>DomCut (Linker prediction)</td>
<td><a href="http://www.bork.embl-heidelberg.de/~suyama/domcut/">http://www.bork.embl-heidelberg.de/~suyama/domcut/</a></td>
<td>Server, standalone</td>
<td>Amino acid composition</td>
<td>Single</td>
</tr>
<tr>
<td><strong>Combination method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DomPred</td>
<td><a href="http://bioinf.cs.ucl.ac.uk/dompred/">http://bioinf.cs.ucl.ac.uk/dompred/</a></td>
<td>Server</td>
<td>Pfam search followed by DomSSEA</td>
<td></td>
</tr>
</tbody>
</table>
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 (DGS), entropy profiles or differential amino acid composition have been selected as discriminatory criteria. Predicted secondary structure and ab initio simulation of 3D structure are also used to make informed boundary predictions. The followings are the most popular ab initio domain boundary prediction methods.

UMA

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

• 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:

• 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 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 using multiple sequence alignment and predicted secondary structures as input. Secondly, domain boundaries of these models are assigned using the method of Taylor. 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 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**
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).

**DomCut**
DomCut[^7] predicts inter-domain linkers regions using sliding-windows average of linker index derived from a domain/linker data set collected from the 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 the 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**

Galzitskaya and Melnik report a method that predicts domain boundaries based on the calculation of entropy profiles. 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. 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.).

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 step-wise 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 first searches the Pfam database to identify known domains, and the ab initio method DomSSEA is used only if there are no hits in the first round.

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

- 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.

- 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.

- The complexity of domain boundary prediction is also greatly increased by rearrangements within the domain,
such as the insertion of one domain into another or domain swapping. 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.

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.

References
Delineation of modular proteins


