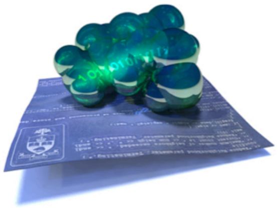
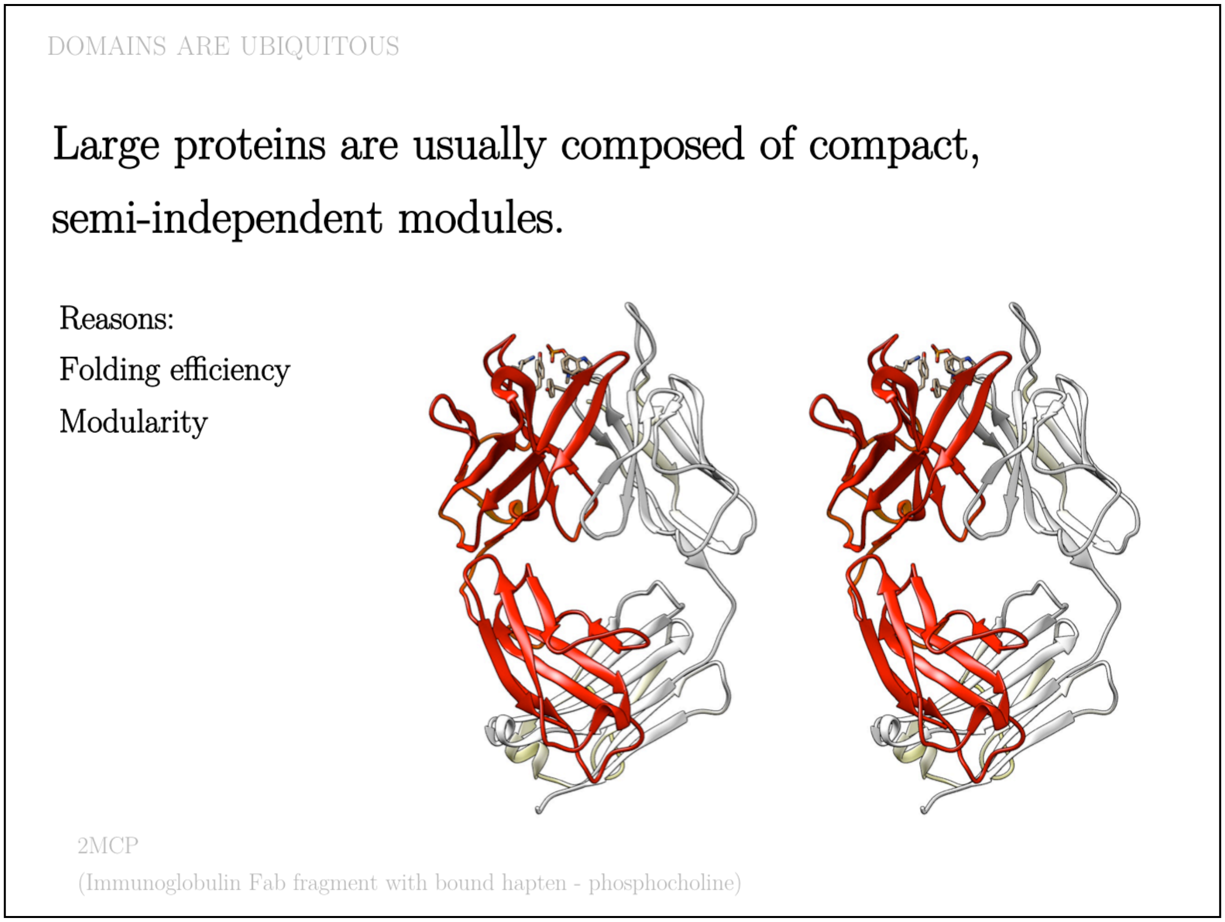
A
Bioinformatics
Course

Protein Structure Domains



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto



DOMAINS ARE UBIQUITOUS

Large proteins are usually composed of compact, semi-independent modules.

Reasons:

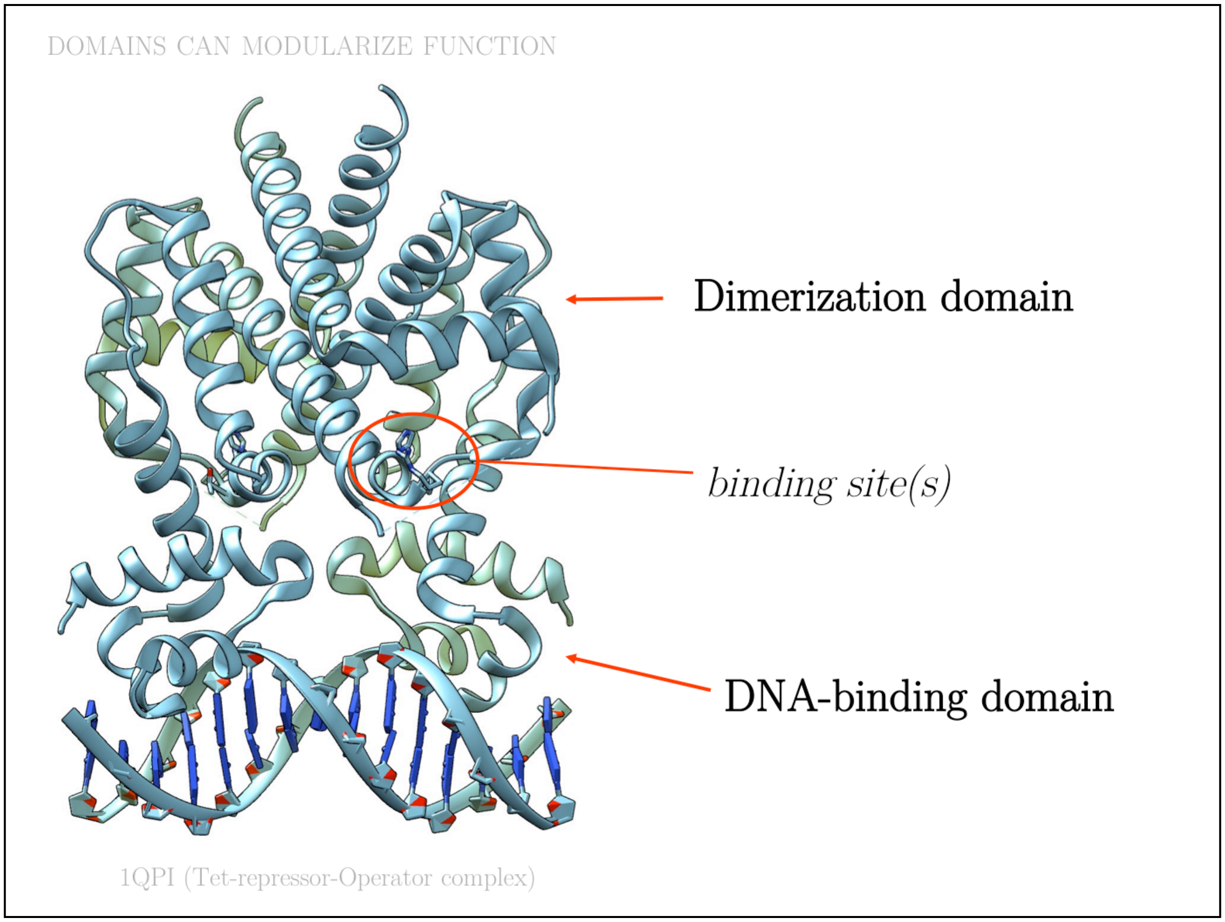
Folding efficiency

Modularity

2MCP

(Immunoglobulin Fab fragment with bound hapten - phosphocholine)

Domains are ubiquitous in proteins and – although the idea of a domain is purely conceptual – they supply an intriguing link between sequence and structure in evolution.



DOMAINS CAN MODULARIZE FUNCTION

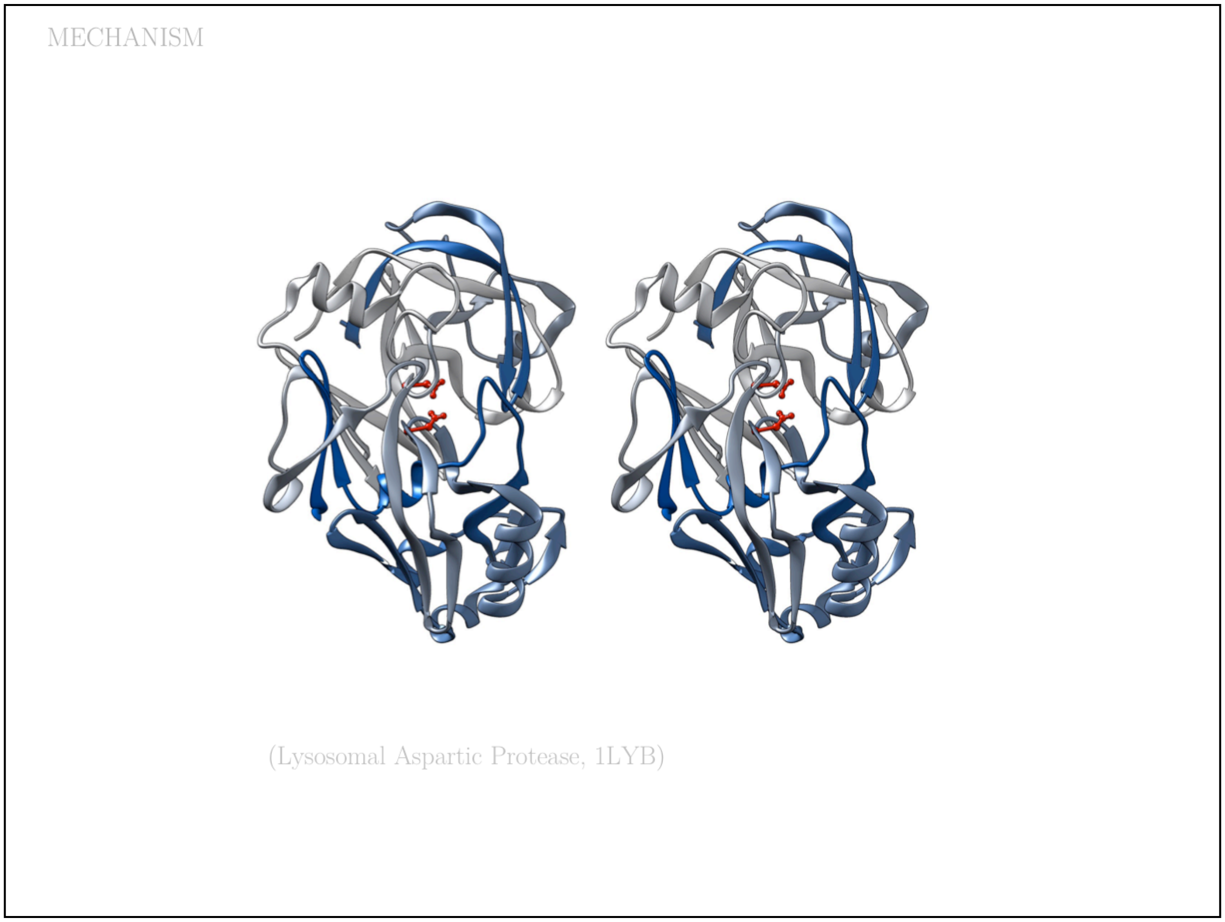
Dimerization domain

binding site(s)

DNA-binding domain

1QPI (Tet-repressor-Operator complex)

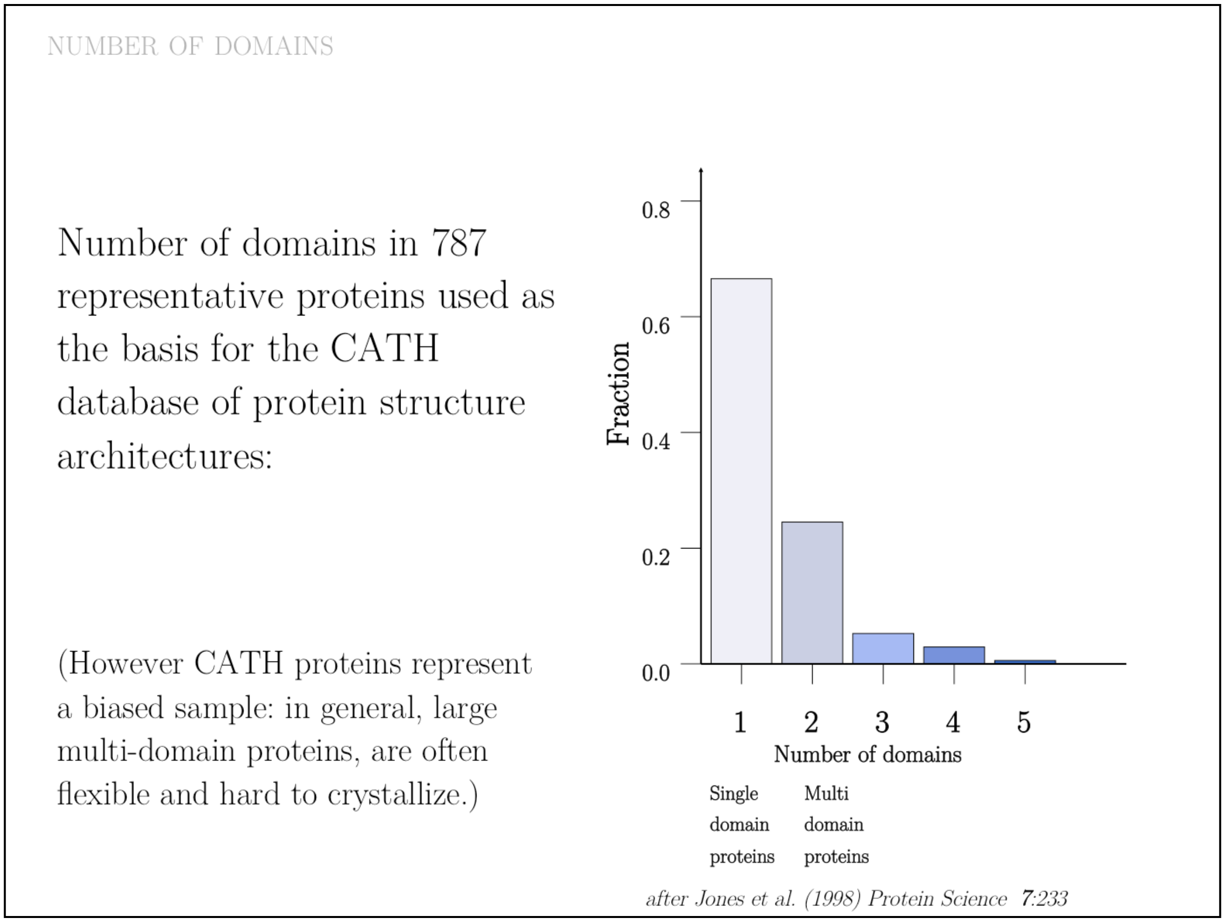
Many domains are empirically found to have distinct functions in proteins. The example above is the Tet-repressor. This observation is highly non-trivial and reflects on the nature of the process of evolution. To explain this fact requires to consider what selective *advantages* can be gained from compartmentalizing function in domains, rather than distributing it over the entirety of the structure, **or** understanding how this fact can be a consequence of the *process* of evolution, or of constraints that arise from evolutionary *mechanisms*.



MECHANISM

(Lysosomal Aspartic Protease, 1LYB)

But it is not always the case that distinct domains correspond to discrete functions: in this example, the active site of the aspartic protease cathepsin D is shared between two domains – in fact the relative motion of the domains appears to be important for the catalytic mechanism.



NUMBER OF DOMAINS

Number of domains in 787 representative proteins used as the basis for the CATH database of protein structure architectures:

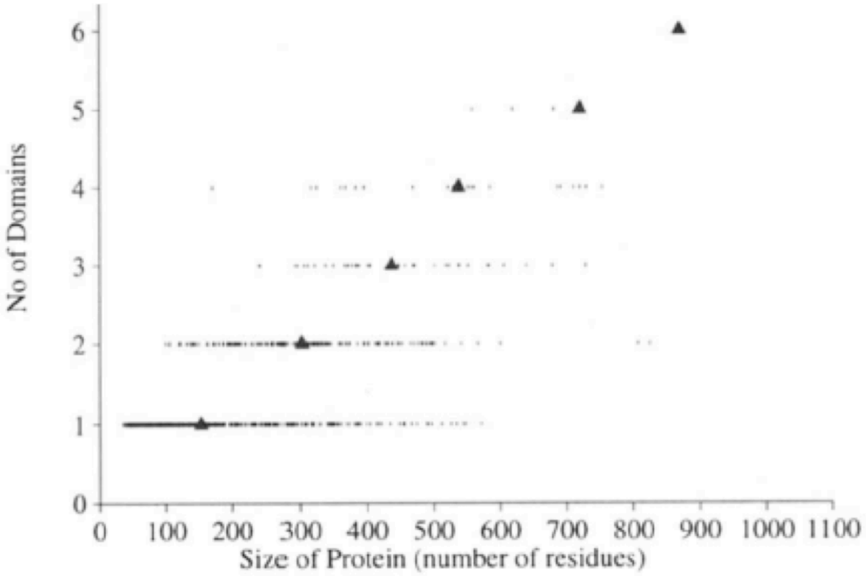
(However CATH proteins represent a biased sample: in general, large multi-domain proteins, are often flexible and hard to crystallize.)

after Jones et al. (1998) Protein Science ***7****:233*

A quarter of all PDB structuires contain multi-domain proteins. The fraction of multi-domain proteins in living organism is higher, and it is higher (in general) in eukaryotes than in prokaryotes. Multi domain proteins are less likely to crystallize than single domain proteins: frequently domains are well-defined in their internal structure, but mobile relative to erach other and this mobility is detrimental to crystal growth.

CHARACTERISTIC DOMAIN SIZE

Non-random relationship between domain number and chain length in the 787 representative proteins used as the basis for the CATH database

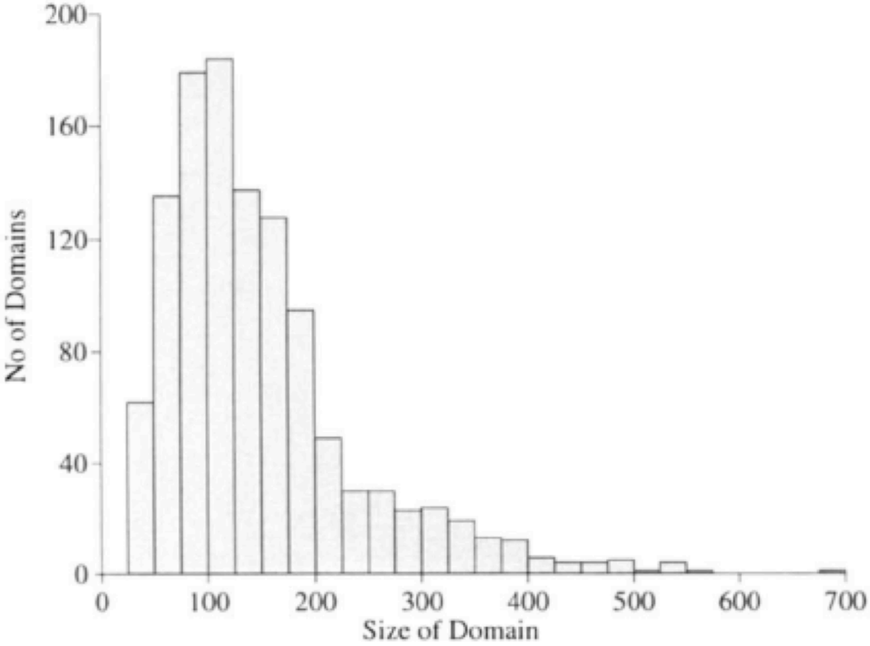


Jones *et al.* (1998)
Protein Science **7**:233

As is to be expected, a larger fraction of longer protein chains contains more than one domain. Interestingly, the distribution correlates well with a characteristic size of properly folding domains, which is – from statistical mechanical considerations – on the order of 110 amino acids.

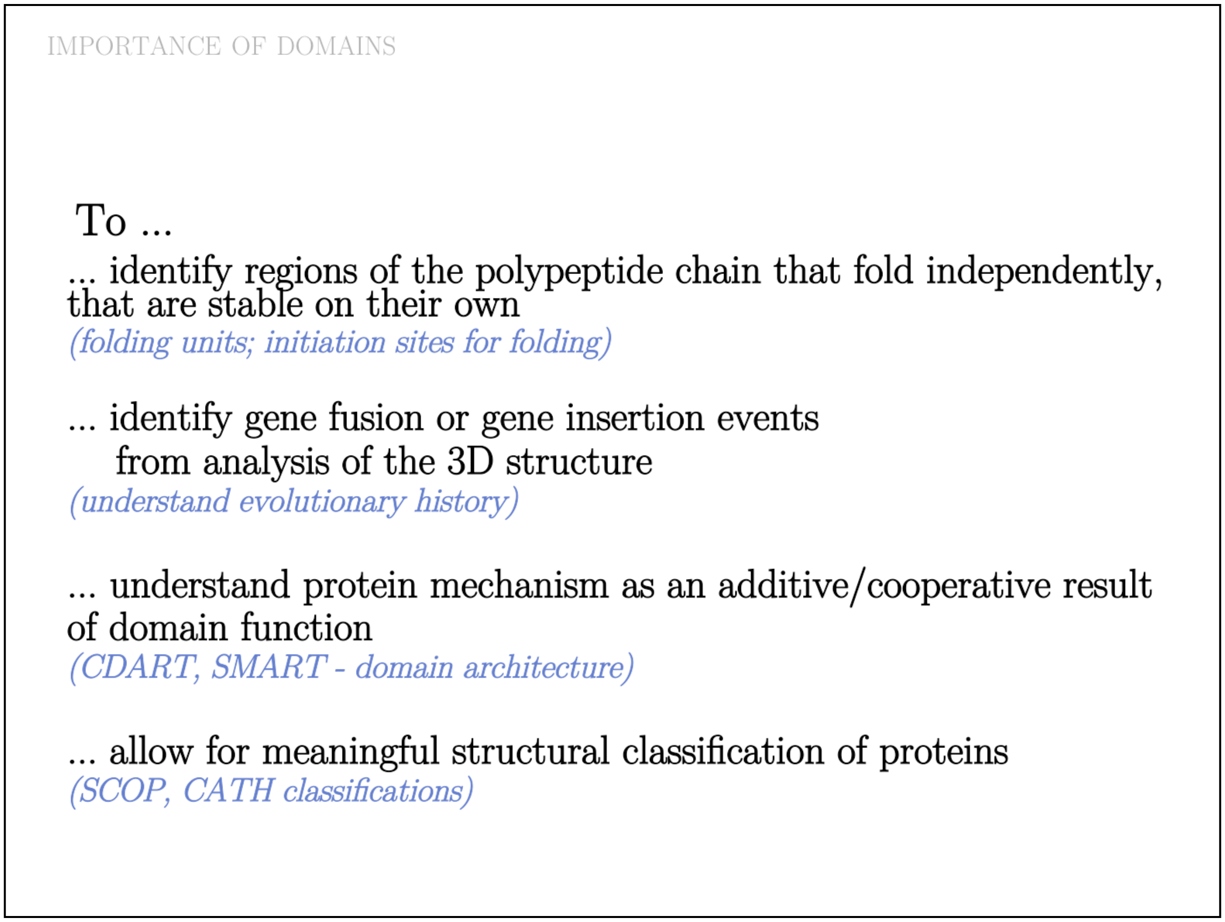
CHARACTERISTIC DOMAIN SIZE

Domain size in the 787 representative proteins used as the basis for the CATH database



Jones *et al.* (1998)
Protein Science **7**:233

Indeed, this length of 110 amino acids is approximately what we observe in nature. It results from a balance between the need to make multiple stabilizing interactions (enthalpy) and not loose too much entropy upon folding from a disordered unfolded state into the single conformation of the native state.



IMPORTANCE OF DOMAINS

To ...

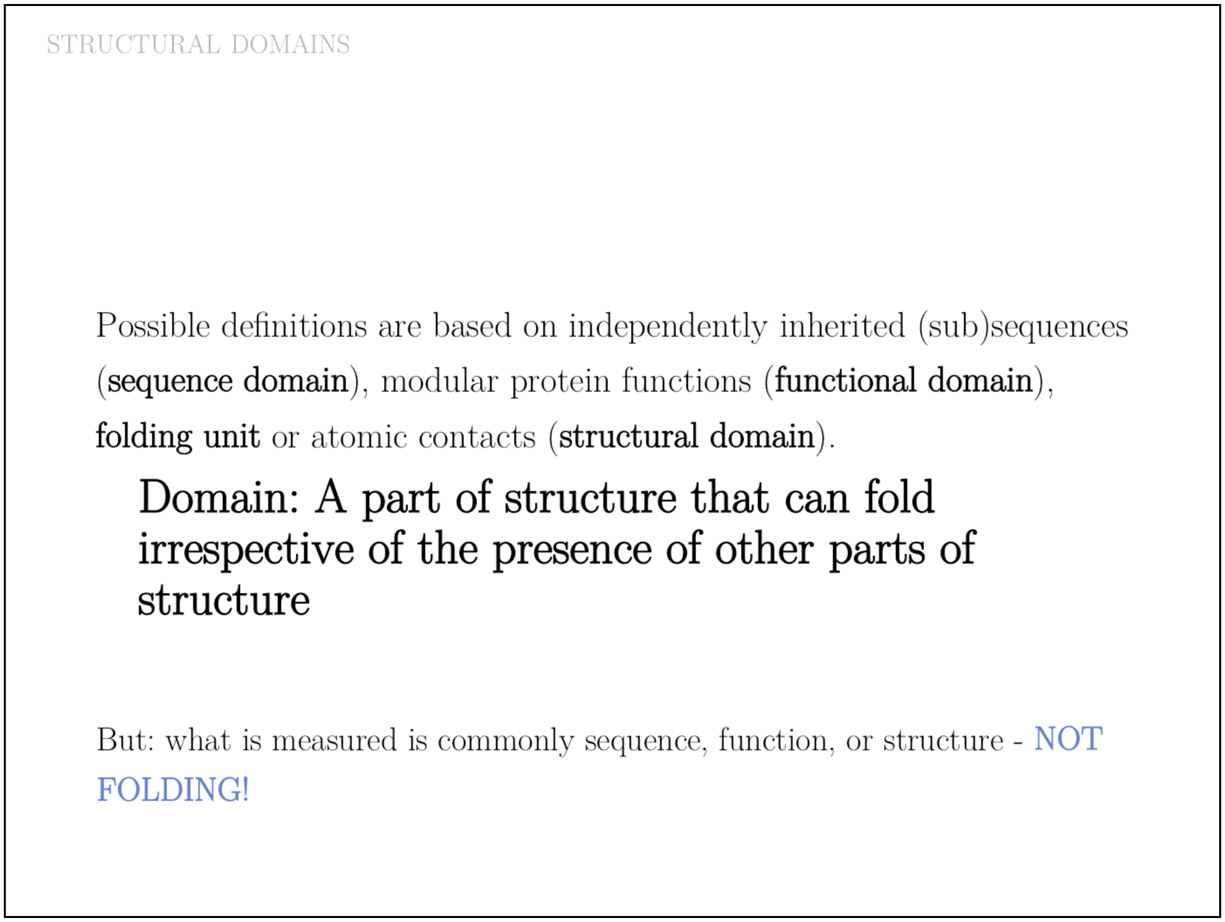
... identify regions of the polypeptide chain that fold independently, that are stable on their own
(folding units; initiation sites for folding)

... identify gene fusion or gene insertion events
from analysis of the 3D structure
(understand evolutionary history)

... understand protein mechanism as an additive/cooperative result of domain function
(CDART, SMART - domain architecture)

... allow for meaningful structural classification of proteins
(SCOP, CATH classifications)

Domains can be used for sequence analysis in many ways.



STRUCTURAL DOMAINS

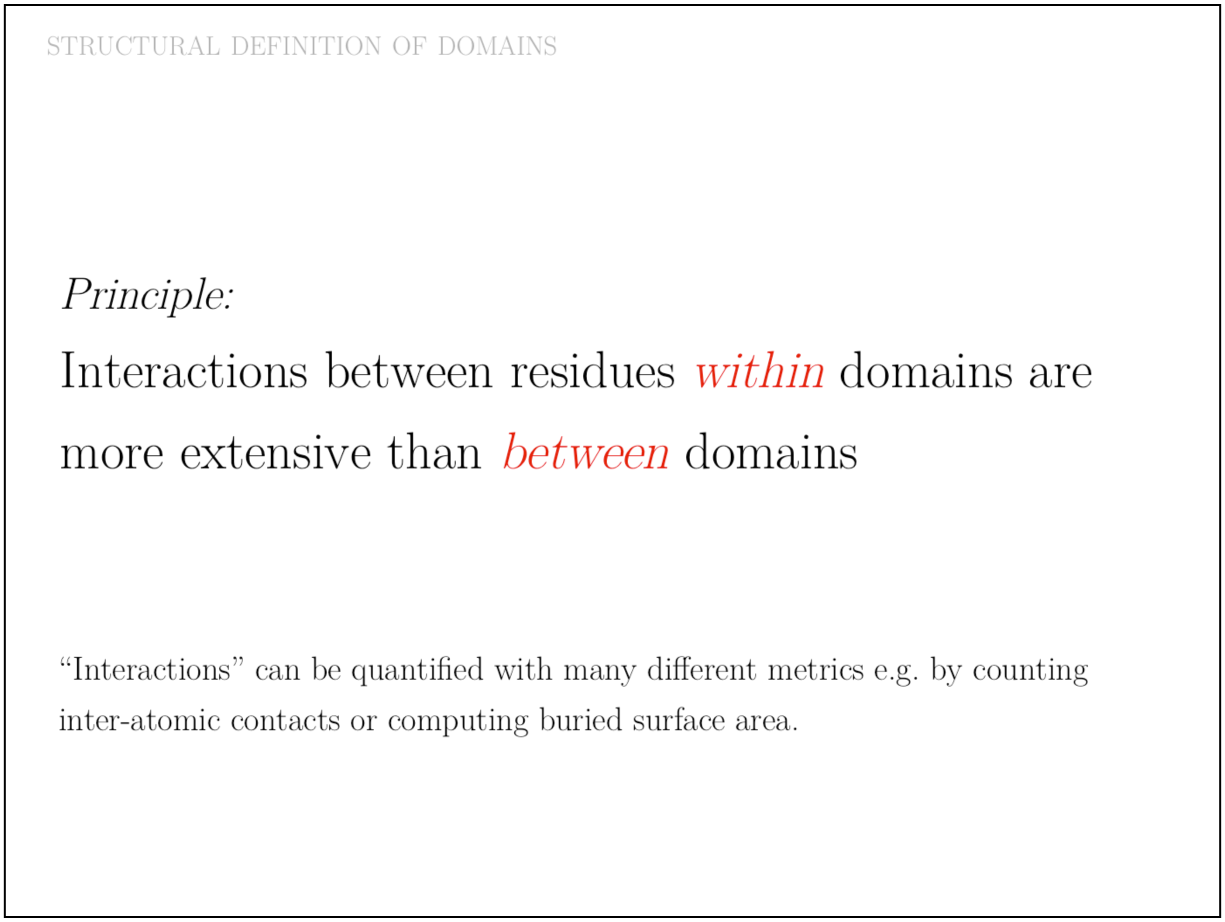
Possible definitions are based on independently inherited (sub)sequences (**sequence domain**), modular protein functions (**functional domain**), **folding unit** or atomic contacts (**structural domain**).

Domain: A part of structure that can fold irrespective of the presence of other parts of structure

But: what is measured is commonly sequence, function, or structure - NOT FOLDING!

Different domain *definitions* suggest different *algorithms* to identify domains.

The separation of a structure into domains requires the arbitrary definition of thresholds in a continuum of possibilities.



STRUCTURAL DEFINITION OF DOMAINS

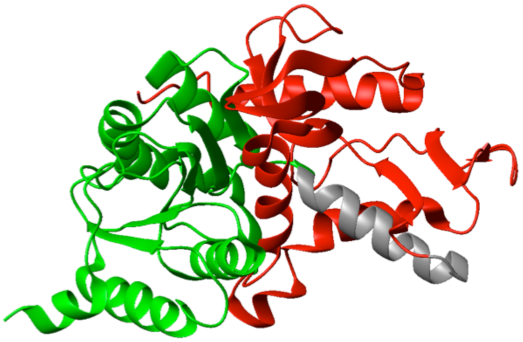
Principle:

Interactions between residues *within* domains are more extensive than *between* domains

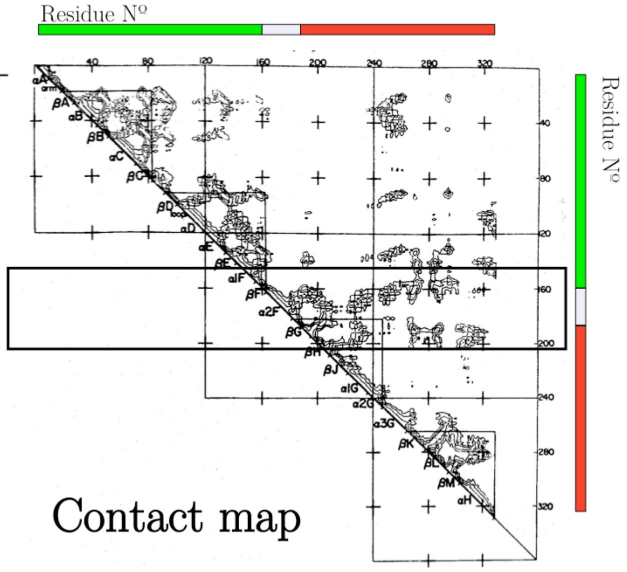
“Interactions” can be quantified with many different metrics e.g. by counting inter-atomic contacts or computing buried surface area.

The definition of domains in structures is similar to the definition of clusters.

DOMAIN DEFINITION FROM CONTACT MAPS

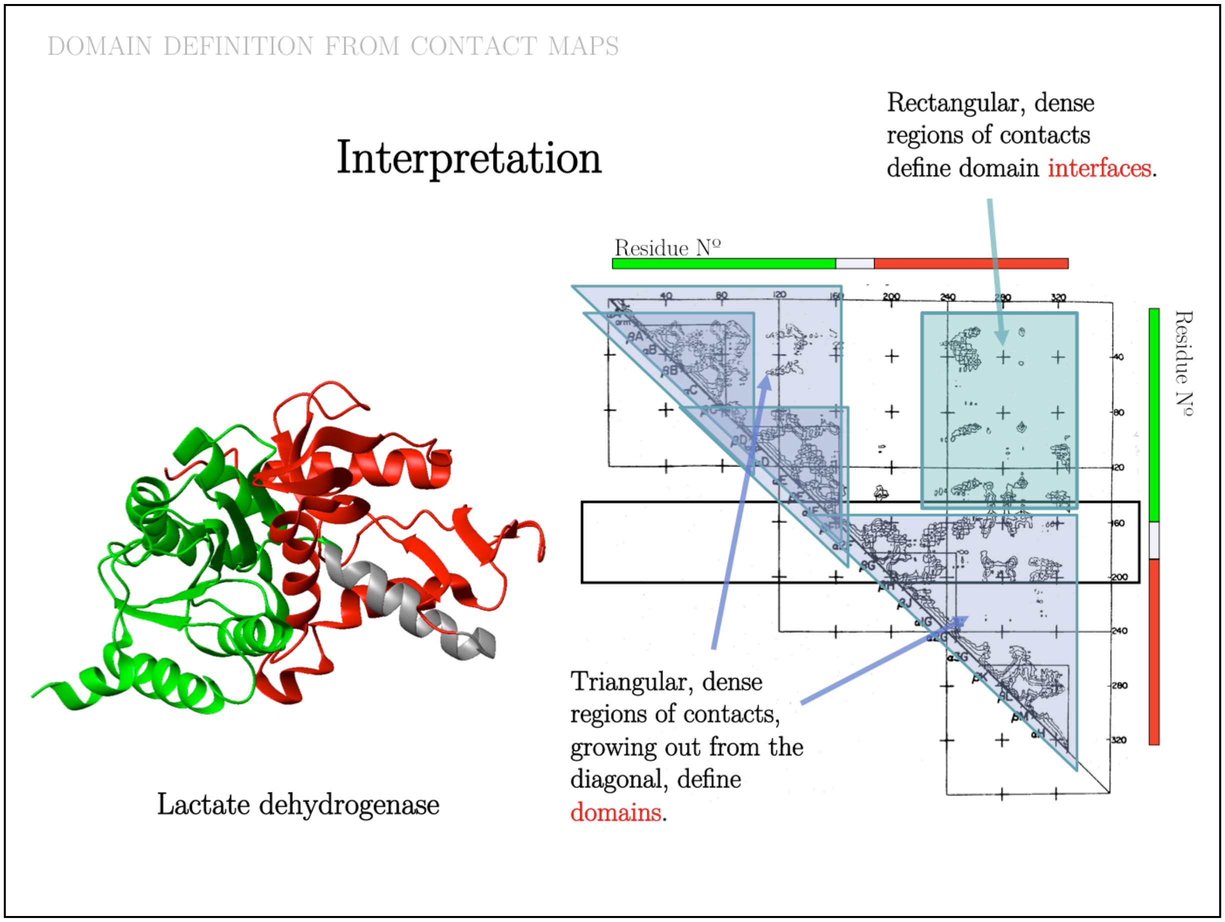


Lactate dehydrogenase



Contact map

Contact maps can illustrate this principle. A contact map plots the distance between residue pairs on a square grid.



DOMAIN DEFINITION FROM CONTACT MAPS

Interpretation

Rectangular, dense regions of contacts define domain interfaces.

Triangular, dense regions of contacts, growing out from the diagonal, define domains.

Lactate dehydrogenase

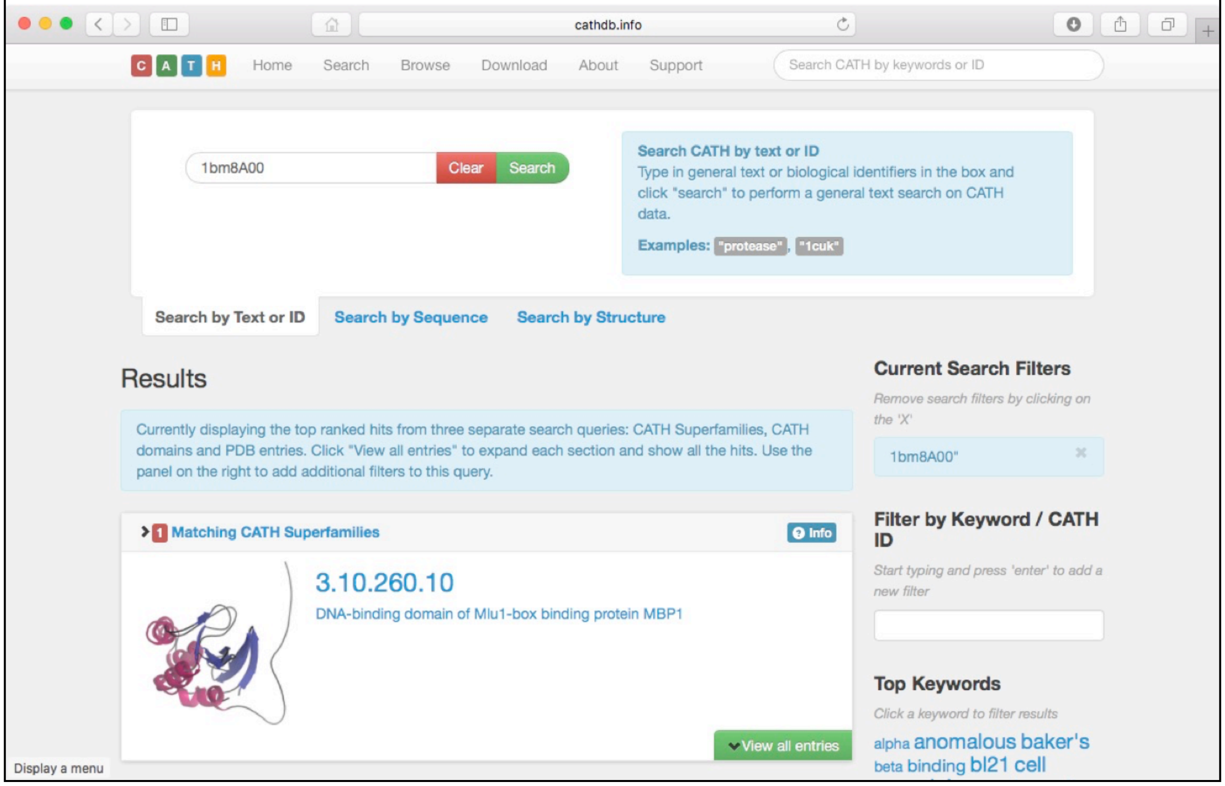
Portions of structure that have a large number of intra-domain contacts show up as density growing out from the diagonal. Here, the red and green colours identify two distinct domains of LDH. Note that both of these domains clearly show additional subdomain structure.

Protein structure is composed hierarchically!

This is one of the problems of accurately defining domains: there may not be a natural level at which the hierarchy can be decomposed into structurally or functionally meaningful units.

However, domains much smaller than 80 amino acids or so are unlikely to fold independently.

CATH



Search CATH by text or ID
Type in general text or biological identifiers in the box and click "search" to perform a general text search on CATH data.

Examples: "protease", "1cuk"

Results

Currently displaying the top ranked hits from three separate search queries: CATH Superfamilies, CATH domains and PDB entries. Click "View all entries" to expand each section and show all the hits. Use the panel on the right to add additional filters to this query.

3.10.260.10
DNA-binding domain of Mlu1-box binding protein MBP1

Current Search Filters
Remove search filters by clicking on the 'X'
1bm8A00"

Filter by Keyword / CATH ID
Start typing and press 'enter' to add a new filter

Top Keywords
Click a keyword to filter results

CATH is a (largely) automated, authoritative, hierarchical classification of all PDB domains. It uses DETECTIVE, DOMAK and PUU to “chop” full-length protein structures into domains, with some manual curation applied to conflicting cases, then it uses CATHEDRAL and SSAP to find which ...

... **Class**

... **Architecture**

... **Topology**, and

... **Homology** family the domain can be classified into.

http://www.cathdb.info

Jones *et al.* (1998) Domain assignment for protein structures using a consensus approach: Characterization and analysis. *Protein Science* 7:233-242

Domain Definition Algorithms Used by CATH

DOMAK:

Maximal interactions within each unit -
minimal interaction between units (domains).

- Assign value to each parwise type of contact
- Arbitrarily split protein and calculate sum of values for both domains
- Large split values correspond to distinct domains

Siddiqui AS & Barton GJ (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.* 4(5):872-884.

Domain Definition Algorithms Used by CATH

DETECTIVE:

Domains (as folding units) have recognizable hydrophobic cores.

- Define distinct hydrophobic cores by considering secondary structure, side-chain accessibility and side-chain side-chain contacts: residues are part of a core when they are in regular secondary structure and have buried side-chains that make non-polar contacts with each other
- Each core and its shell define a distinct domain

Swindells MB (1995) A procedure for detecting structural domains in proteins. *Protein Sci.* 4(1):103-112.

Domain Definition Algorithms Used by CATH

PUU (Parser for Protein Unfolding Units):

**Maximal interactions within each unit -
minimal interaction between units (domains).**

- Build contact matrix for residues
- Solve eigenvalue problem related to "strength" of residue interfaces (this groups residues by interactions)
- Search for all reasonable bisections (guided by physical criteria that identify units of sufficient internal stability) of residue groups.
- Recursive bisections build folding "tree".

Holm L & Sander C (1994) Parser for protein folding units. *Proteins* **19**(3):256-268

Domain Search Algorithms Used by CATH

CATHEDRAL (Sequential Structure Alignment Program)

Fast prescreening of domains against a database

- CATHEDRAL uses a fast graph-matching algorithm on graphs of secondary structure elements and their distances to prescreen a domain against a database
- Significantly matching domains are aligned and refined using SSAP

Pearl FMG *et al.* (2003) The CATH database: an extended protein family resource for structural and functional genomics
Nucleic Acids Res. 31(1):452-455.

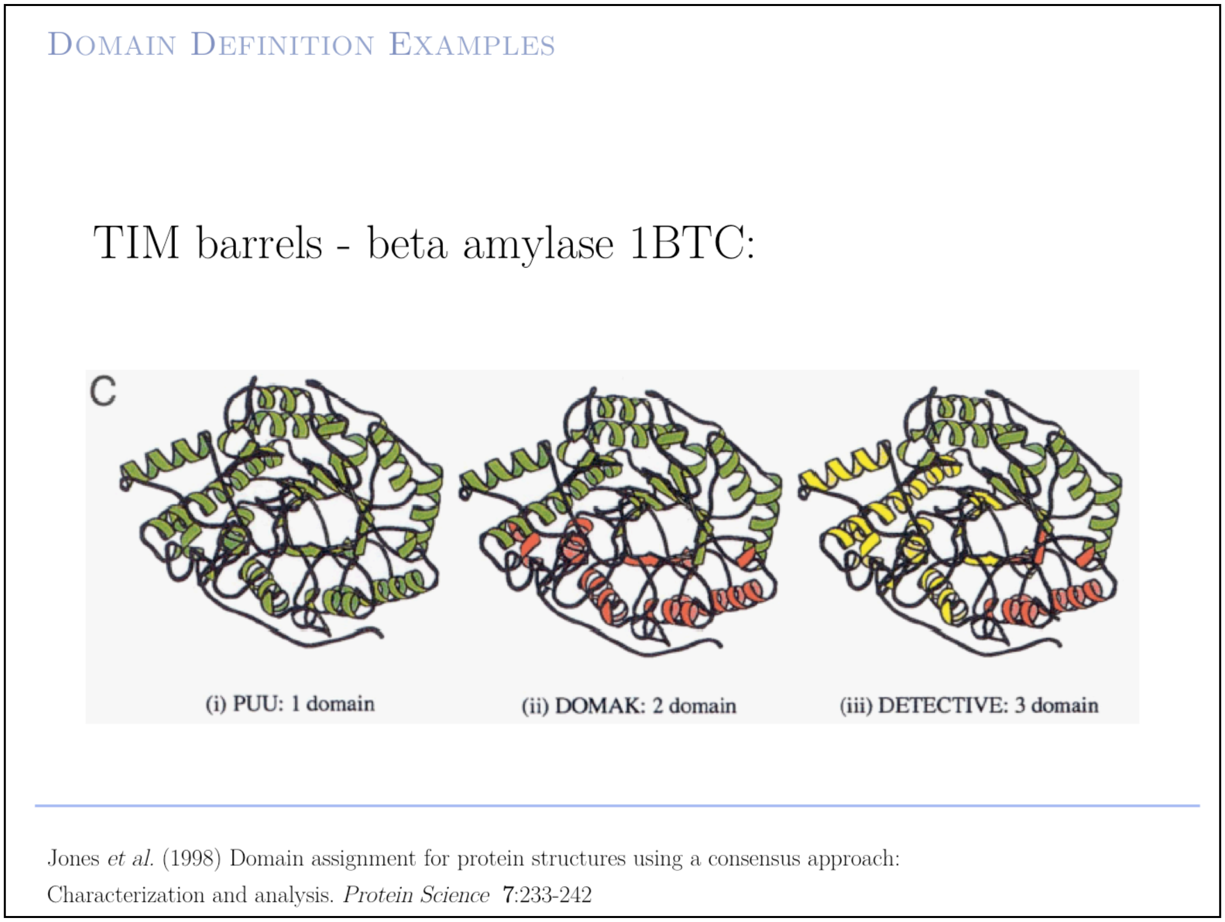
Domain Search Algorithms Used by CATH

SSAP (Sequential Structure Alignment Program)

Accurate structural alignment

- SSAP is a general-purpose structural alignment program
- Residues are described by a vector of distance to other residues
- Alignments are sought that minimize the difference between distances

Orengo CA & Taylor WR (1996) SSAP: Sequential structure alignment program for protein structure comparison
Methods in Enzymology. **266**:617-635.



Domain Definition Examples

TIM barrels - beta amylase 1BTC:

(i) PUU: 1 domain

(ii) DOMAK: 2 domain

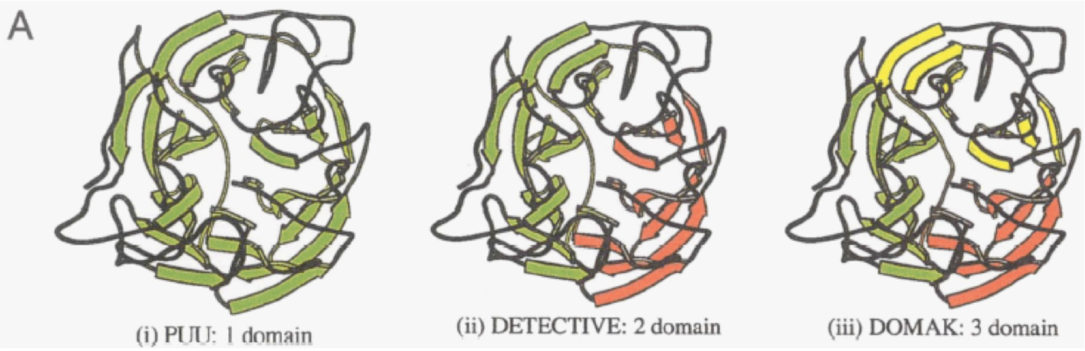
(iii) DETECTIVE: 3 domain

Jones *et al.* (1998) Domain assignment for protein structures using a consensus approach: Characterization and analysis. *Protein Science* **7**:233-242

Domain classification programs do not always agree. The three classifiers used for CATH all disagree on the 1BTC TIM-barrel structure.

Domain Definition Examples

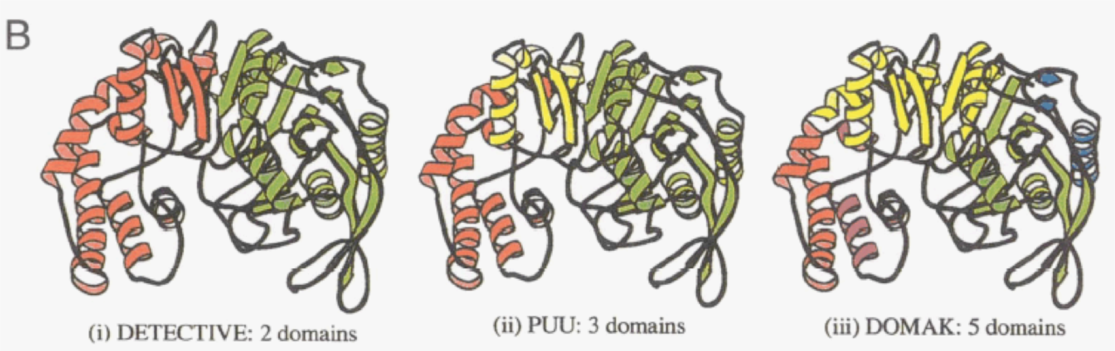
beta propeller - neuraminidase 1NSC_A:



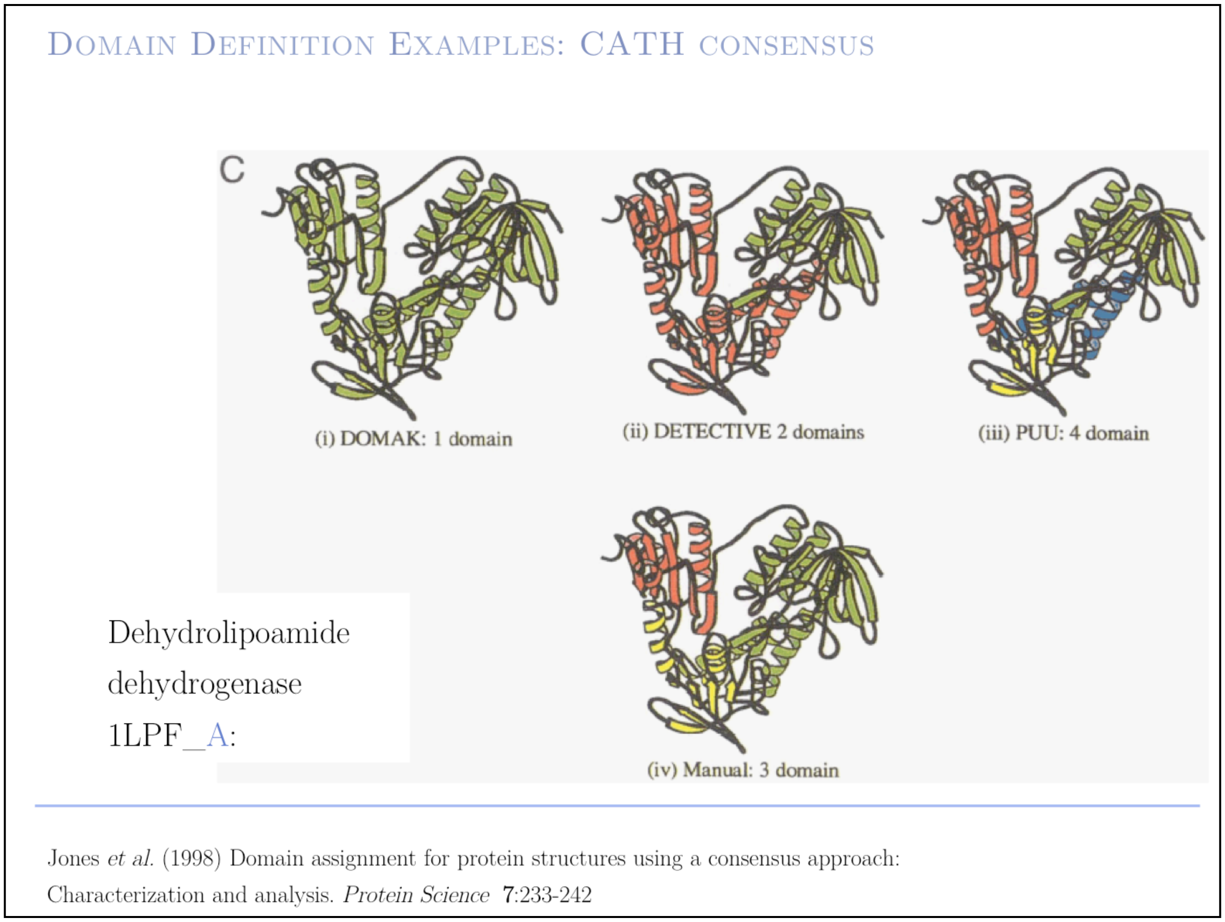
Jones *et al.* (1998) Domain assignment for protein structures using a consensus approach: Characterization and analysis. *Protein Science* **7**:233-242

Domain Definition Examples

Adenylosuccinate synthetase 1ADE_A:



Jones *et al.* (1998) Domain assignment for protein structures using a consensus approach: Characterization and analysis. *Protein Science* **7**:233-242



Domain Definition Examples: CATH consensus

(i) DOMAK: 1 domain

(ii) DETECTIVE 2 domains

(iii) PUU: 4 domain

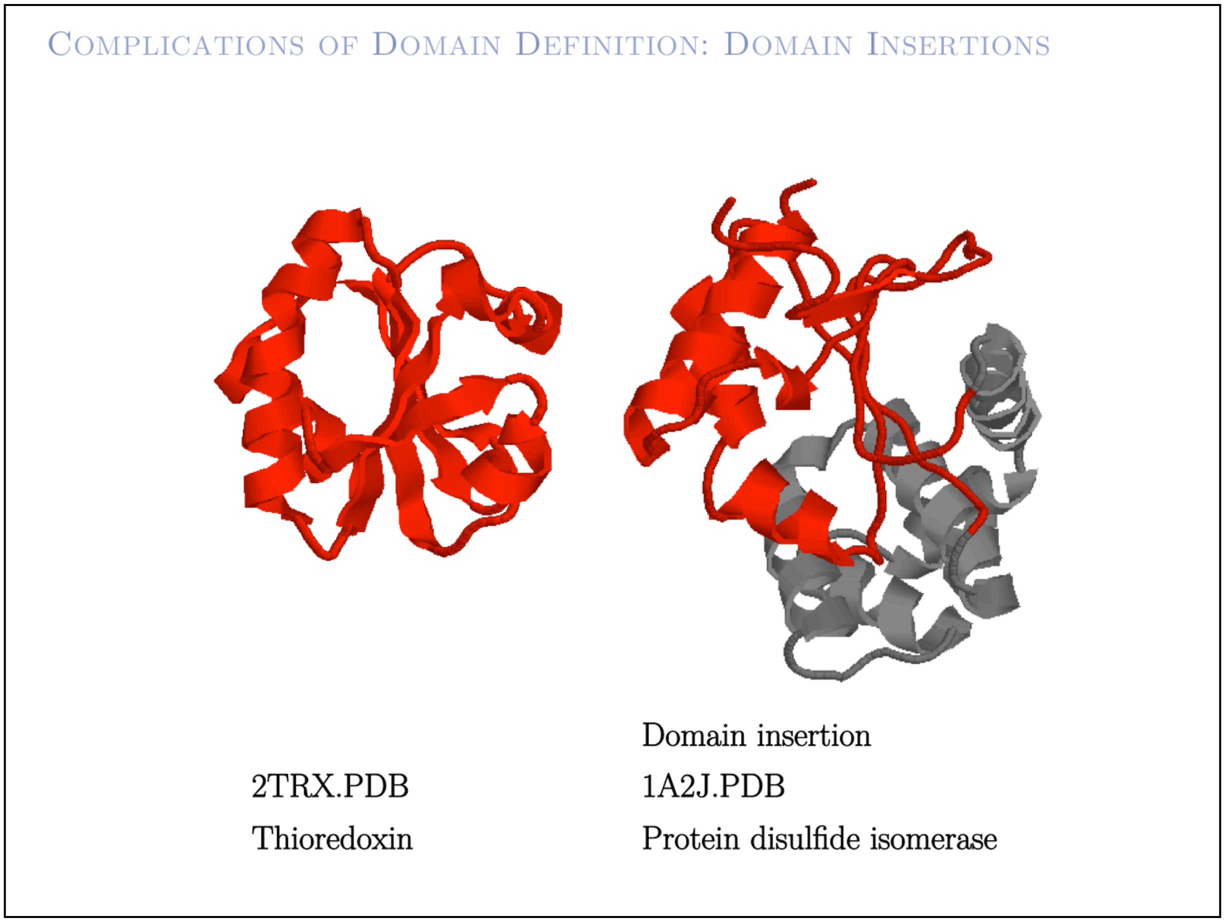
(iv) Manual: 3 domain

Dehydrolipoamide dehydrogenase 1LPF_A:

Jones *et al.* (1998) Domain assignment for protein structures using a consensus approach: Characterization and analysis. *Protein Science* **7**:233-242

CATH is built from domains that are defined from a consensus of its three core classifiers – with manual intervention if necessary.

Expert intervention and curation is extremely helpful to maintain the quality of the database – but anything that needs to be done manually **does not scale**.



Complications of Domain Definition: Domain Insertions

Domain insertion

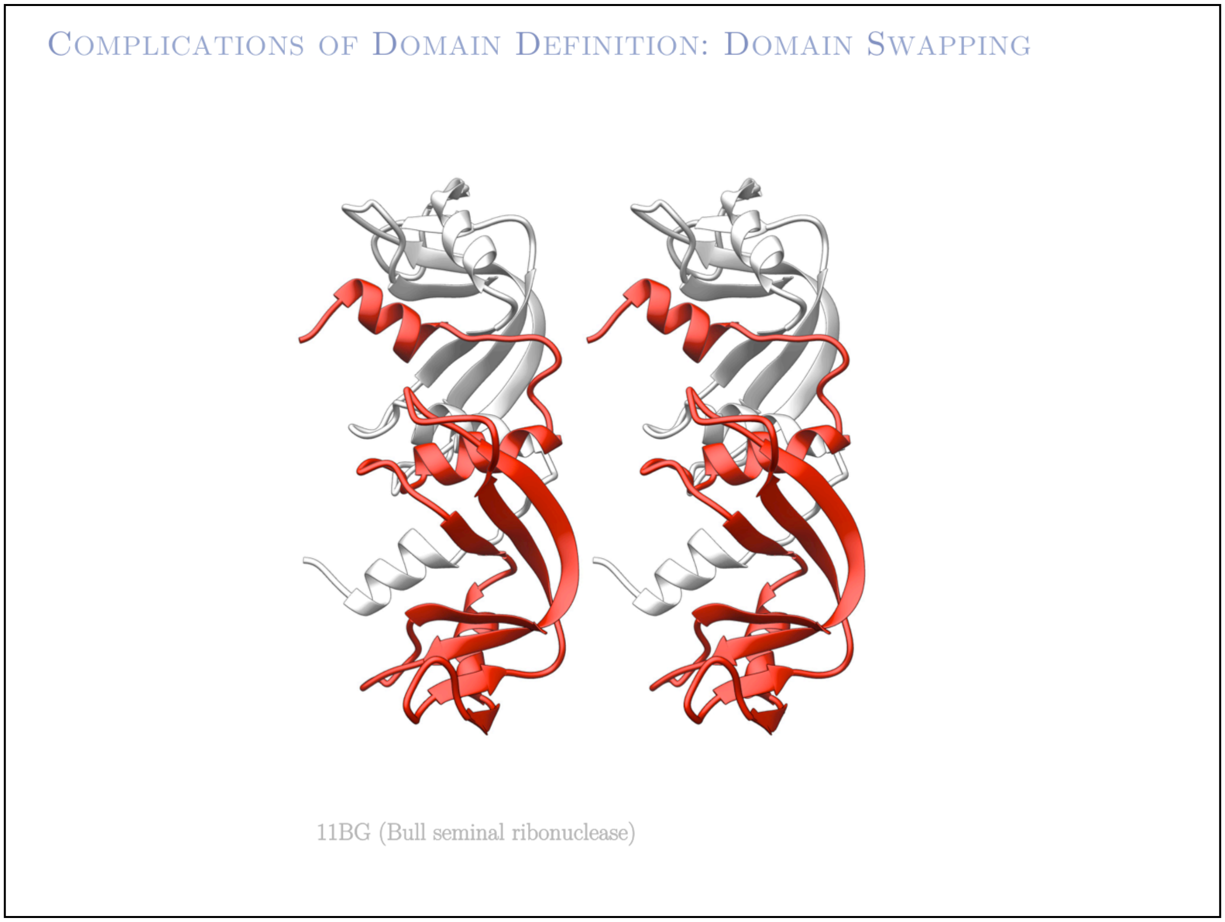
2TRX.PDB

Thioredoxin

1A2J.PDB

Protein disulfide isomerase

Sometimes the definition of domains is made additionally difficult by the complicated biology of natural proteins. Protein disulfide isomerase has a thioredoxin fold in principle – but there is a complete independent domain inserted into one of its loops.



Complications of Domain Definition: Domain Swapping

11BG (Bull seminal ribonuclease)

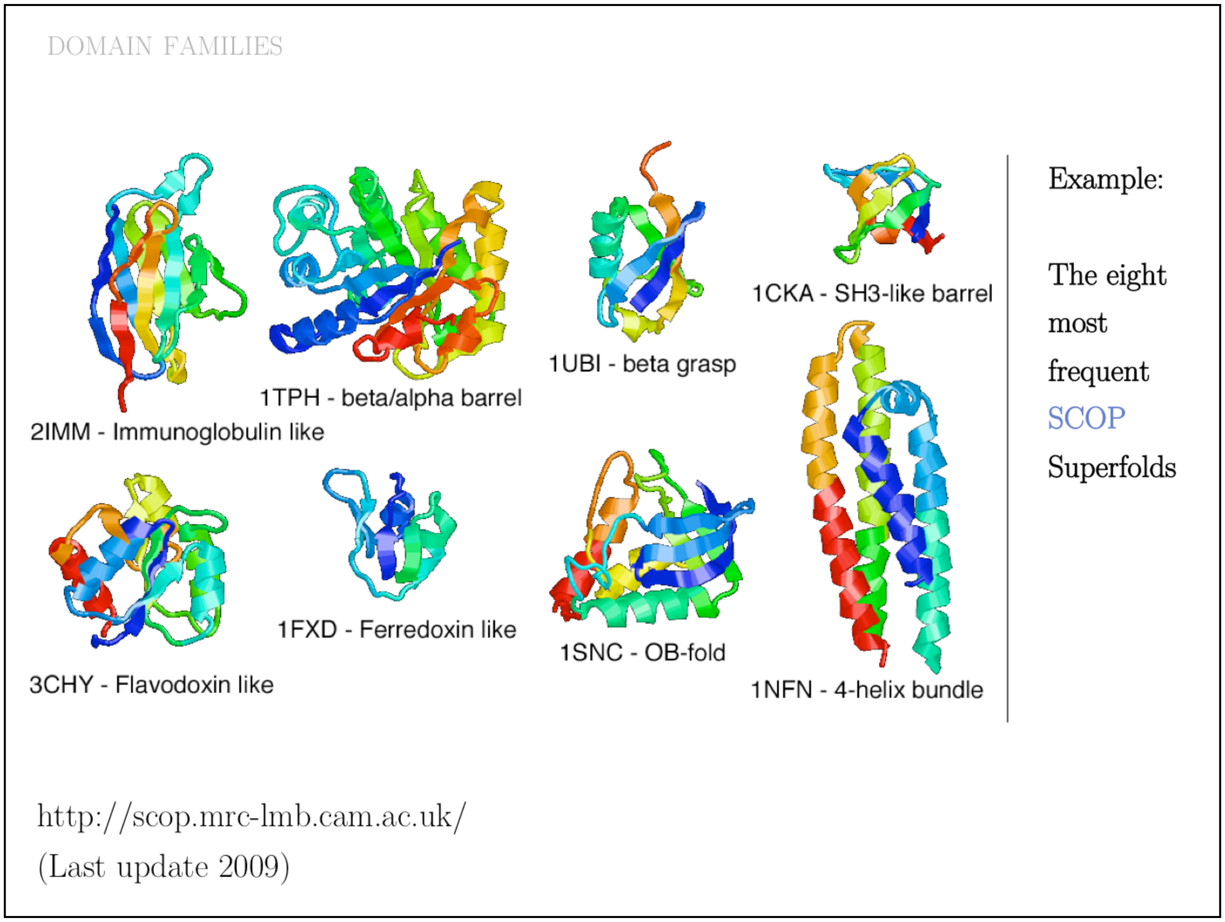
Domain swapping leads to elements of the *same sequence* being integrated into the structure of *different domains*. This requires a strained connection region, otherwise the higher local concentration of the *intra*-domain contact would override the possibility of *inter*-domain interactions.

The result is a very tight and often essentially irreversible association that requires complete unfolding of the domains to undo.

The top and bottom parts of the image correspond to well structured, compact domains. However both of these domains incorporate a helix from the respectively other sequence of this homodimer.

cf.

Wodak, Malevanets & MacKinnon (2015) The Landscape of Intertwined Associations in Homooligomeric Proteins. Biophys J 109(6):1087-1100.



DOMAIN FAMILIES

2IMM - Immunoglobulin like

1TPH - beta/alpha barrel

1UBI - beta grasp

1CKA - SH3-like barrel

3CHY - Flavodoxin like

1FXD - Ferredoxin like

1SNC - OB-fold

1NFN - 4-helix bundle

Example:

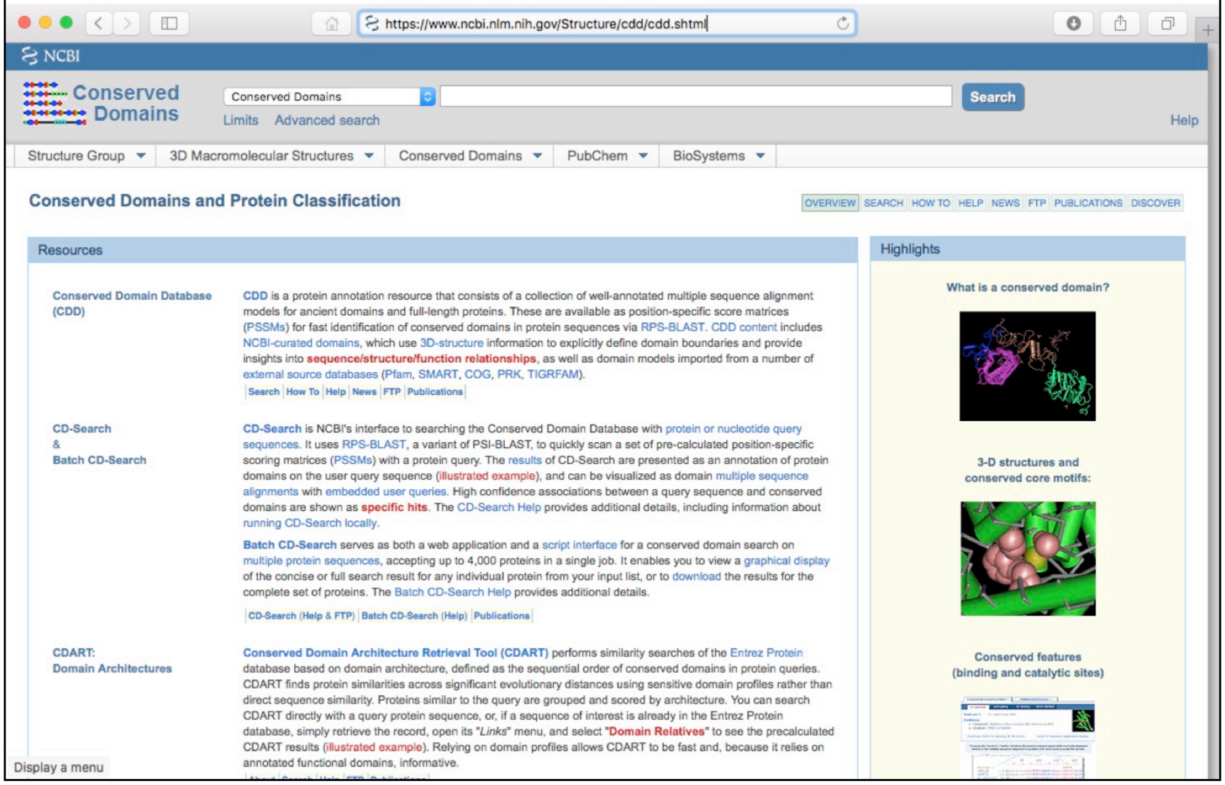
The eight most frequent SCOP Superfolds

http://scop.mrc-lmb.cam.ac.uk/
(Last update 2009)

Domains are found in many families – but the boundaries between families may be somewhat fluid as it may be not obvious when two domains are actually different. Moreover, even within families, we may find proteins that - as far as we can tell - are not actually related to each other, but have arrived at their particular fold through **convergent** evolution. Nevertheless, assigning folds to families allows us to bring some order to the zoo of possibilities and this underlies approaches to organize and retrieve domains in databases.

The examples above are taken from the highly curated SCOP database, which unfortunately has not been updated since 2009.

CDD



Conserved Domains and Protein Classification

Resources

Conserved Domain Database (CDD)

CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domains, which use 3D-structure information to explicitly define domain boundaries and provide insights into **sequence/structure/function relationships**, as well as domain models imported from a number of external source databases (Pfam, SMART, COG, PRK, TIGRFAM).

Search | How To | Help | News | FTP | Publications

CD-Search & Batch CD-Search

CD-Search is NCBI's interface to searching the Conserved Domain Database with protein or nucleotide query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query. The results of CD-Search are presented as an annotation of protein domains on the user query sequence (illustrated example), and can be visualized as domain multiple sequence alignments with embedded user queries. High confidence associations between a query sequence and conserved domains are shown as **specific hits**. The CD-Search Help provides additional details, including information about running CD-Search locally.

Batch CD-Search serves as both a web application and a script interface for a conserved domain search on multiple protein sequences, accepting up to 4,000 proteins in a single job. It enables you to view a graphical display of the concise or full search result for any individual protein from your input list, or to download the results for the complete set of proteins. The Batch CD-Search Help provides additional details.

CD-Search (Help & FTP) | Batch CD-Search (Help) | Publications

CDART: Domain Architectures

Conserved Domain Architecture Retrieval Tool (CDART) performs similarity searches of the Entrez Protein database based on domain architecture, defined as the sequential order of conserved domains in protein queries. CDART finds protein similarities across significant evolutionary distances using sensitive domain profiles rather than direct sequence similarity. Proteins similar to the query are grouped and scored by architecture. You can search CDART directly with a query protein sequence, or, if a sequence of interest is already in the Entrez Protein database, simply retrieve the record, open its "*Links*" menu, and select "**Domain Relatives**" to see the precalculated CDART results (illustrated example). Relying on domain profiles allows CDART to be fast and, because it relies on annotated functional domains, informative.

Highlights

What is a conserved domain?

3-D structures and conserved core motifs:

Conserved features (binding and catalytic sites)

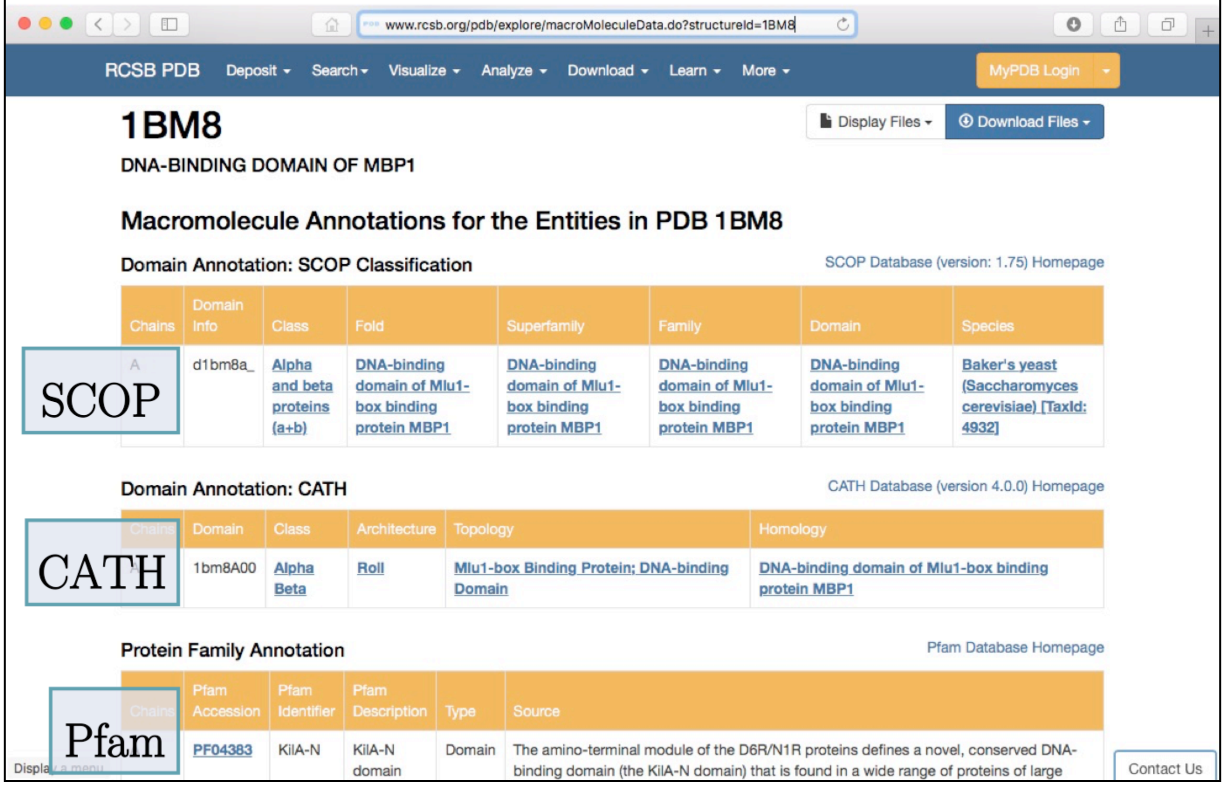
CDD is the integrated domain analysis and search tool at the NCBI. It imports structural domain definitions from SCOP and CATH, as well as sequence domains from **Pfam**, thus providing an integrated access to sequence- as well as structural domains. It's results are available via cross references in all typical sequence resources, as well as in BLAST searches *etc.*

http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

Marchler-Bauer *et al.* (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research* 39(DB issue):D225-D229

DOMAIN DATABASES

Unified access via PDB:



1BM8

DNA-BINDING DOMAIN OF MBP1

Macromolecule Annotations for the Entities in PDB 1BM8

Domain Annotation: SCOP Classification

SCOP Database (version: 1.75) Homepage

Chains	Domain Info	Class	Fold	Superfamily	Family	Domain	Species
A	d1bm8a_	Alpha and beta proteins (a+b)	DNA-binding domain of Mlu1-box binding protein MBP1	DNA-binding domain of Mlu1-box binding protein MBP1	DNA-binding domain of Mlu1-box binding protein MBP1	DNA-binding domain of Mlu1-box binding protein MBP1	Baker's yeast (Saccharomyces cerevisiae) [TaxId: 4932]

Domain Annotation: CATH

CATH Database (version 4.0.0) Homepage

Chains	Domain	Class	Architecture	Topology	Homology
	1bm8A00	Alpha Beta	Roll	Mlu1-box Binding Protein; DNA-binding Domain	DNA-binding domain of Mlu1-box binding protein MBP1

Protein Family Annotation

Pfam Database Homepage

Chains	Pfam Accession	Pfam Identifier	Pfam Description	Type	Source
	PF04383	KilA-N	KilA-N domain	Domain	The amino-terminal module of the D6R/N1R proteins defines a novel, conserved DNA-binding domain (the KilA-N domain) that is found in a wide range of proteins of large

The PDB provides cross references into the domain databases for each of its entries. SCOP and CATH are databases for structural domains, Pfam defines sequence domains.

http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA