Protein classification

SCOP
TOPS

Contact maps

domains
To a **cell biologist** a **domain** is a sequential unit within a gene, usually with a specific function.

To a **structural biologist** a **domain** is a compact globular unit within a protein, classified by its 3D structure.
A domain is...

- ... an autonomously-folding substructure of a protein.
- ... > 30 residues, but typically < 200. May be bigger.
- ...usually has a single hydrophobic core
- ... usually composed of one chain (occasionally composed of multiple chains)
- ...is usually composed on one contiguous segment (occasionally made of discontiguous segments of the same chain)
SAR-2 spike protein — a multi domain protein
SCOP -- a hierarchy

http://scop.berkeley.edu

(1) class
(2) fold
(3) superfamily
(4) family
(5) protein
(6) species

- similar secondary structure content
- vague structural homology
- Clear structural homology
- Clear sequence homology
- nearly identical sequences
- individual structures
SCOP -- class

1. all α (289)
2. all β (178)
3. α/β (148)
4. α+β (388)
5. multidomain (71)
6. membrane (60)
7. small (98)
8. coiled coil (7)
9. low-resolution (25)
10. peptides (148)
11. designed proteins (44)
12. artifacts (1)

Proteins of the same class conserve secondary structure content

classes of domains

Not true classes of globular protein domains
SCOP -- fold level

within α/β proteins -- Mainly parallel beta sheets (beta-alpha-beta units)

TIM-barrel (22)
swivelling beta/beta/alpha domain (5)
spoIIa-like (2)
flavodoxin-like (10)
restriction endonuclease-like (2)
ribokinase-like (2)
chelatase-like (2)

Many folds have historical names. “TIM” barrel was first seen in TIM. These classifications are done by eye, by experts.

Proteins of the same Fold conserve topology.
SCOP fold level jargon

example: $\alpha/\beta$ proteins: flavodoxin-like

SCOP Description: 3 layers, $\alpha/\beta/\alpha$; parallel beta-sheet of 5 strand, order 21345

Note the term: “layers”
Rough arrangements of secondary structure elements.

Note the term: “order”
The sequential order of beta strands in a beta sheet.
Fold-level similarity

7-stranded alpha/beta barrel

SSE are in the same order along the chain, and trace roughly the same path through space. Similarity is evident when viewed side-by-side.

But the SSE do not superpose. Some superposition algorithms fail to superpose proteins of the same fold.
Superfamily level similarity

FAD-linked reductases

Members of the same superfamily cannot usually be found in a BLAST search. But can be identified by structural superposition.

Proteins in the same superfamily may look completely different, but upon close inspection they contain a superposable domain of secondary structure elements.
Different members of the same family superimpose well. At this level, a structure may be used as a molecular replacement model for Xray crystallography.

A BLAST search using one family member finds all other family members.
A **Family** is the set of homologs we can find by BLAST sequence database search.

A **Superfamily** is a set of distant homologs that cannot be easily found by BLAST search, but can be recognized by sophisticated fold recognition algorithms.

A **Fold** is an even more distant homologous relationship, recognizable only when the structure is known.

A **Class** is not a homologous relationship but just a statement of the gross secondary structure content.
Contact maps and TOPS diagrams
TOPS topology cartoons

Secondary structure elements (SSE)

beta strand pointing up
beta strand pointing down
alpha helix
connections

A parallel beta sheet
An anti-parallel beta sheet
TOPS topology cartoons

A right-handed $\beta\alpha\beta$ unit

A left-handed $\beta\alpha\beta$ unit (rarely seen)
How to draw TOPS

To do this on your own, find the link "TOPS practice" (tops_practice.moe) on the course web site. Download. Open it in moe.

Or just follow along as I guide you through it. Get pen and paper.
How to draw TOPS

Line up the molecule along the beta sheet, if present.

Otherwise choose a direction so that secondary structures are mostly perpendicular to the screen.
TOPS diagram

- Draw secondary structures first.
TOPS diagram

- number them and connect

Be careful to draw connections to the center or side, when it is in front or in back, respectively.
Name it. SCOP-style.

- 3 layers, 2-4-2 $\alpha\beta\alpha$, all parallel, 1234
Exercise 16.2: contact map and TOPS cartoon

Open MOE

File | Open: RCSB PDB: codes: 2pt1

Ribbon | Style: oval

Ribbon | Color : structure

Identify SSEs. Draw triangles and circles

Ribbon | Color : terminus

Number and connect SSEs.
2ptl contact map

H-bonds  
Distance cutoff
To draw a barrel, determine strand neighbors, up or down, arrange triangles in a circle. Draw connector lines in front, or in back, of triangles.
Exercise 16.3: TOPS cartoon of beta barrel

Open MOE. Open Green Fluorescent Protein
File | Open: RCSB PDB: code: 2b3p
Ribbon | Style: oval
Ribbon | Color : structure
Identify SSEs. Draw triangles and circles
Ribbon | Color : terminus
Number SSEs. Draw connections. Label termini.
• Mostly anti-parallel barrel, closed, containing a helix; \( n=11 \)
• sheet order 1 2 3 11 10 7 8 9 4 5 6

GFP-like fluorescent proteins
Contact maps: proteins in 2D

In a Contact Map: “1” = $D_{ij} < 8\text{Å}$

“0” = $D_{ij} > 8\text{Å}$

- Hairpin
- Helix
- Parallel strands
- Anti-parallel strands
TOPS and contact maps

A "contact map" for a $\beta\alpha\beta$ unit.

parallel sheet.

helix contacts.
Contact map for a small protein

A contact map contains enough information to build the 3D structure within ~2Å RMSD.
A simplified contact map based on SSEs

(1) Arrange the SSEs along the sequence (a line) in both directions
(2) Draw a line parallel to the diagonal for each helix
(3) For any two SSEs that touch, draw a line parallel to the diagonal if the contacts are parallel, draw a line perpendicular to the diagonal if the contacts are anti-parallel. Draw a dotted line if a helix is involved.
Simplified contact map to TOPS diagram

- $\alpha_1$, $\beta_1$
- $\alpha_2$, $\beta_2$
- $\alpha_3$, $\beta_3$
- $\alpha_4$, $\beta_4$
- $\alpha_5$
- $\beta_5$

Legend:
- alpha i- $i+4$
- beta-beta
- alpha-beta
- alpha-alpha
Exercise 16.4: TOPS from contact map

Do this on paper.

Draw a TOPS cartoon that has this contact map. SSEs are $\beta\alpha\beta\alpha\beta$. 
Most genes represent multidomain proteins

~40% of known structures (crystal, NMR) are multidomain proteins, but

**Most** of all proteins are multidomain. (~60% in unicellular organisms, ~90% in eukaryotes).

Domain boundaries can be seen as "weak" connections in the structure.

"Weak" means few contacts and few chain cross-overs.

Domain boundaries can be seen in multiple sequence alignments if the alignments are of whole genes.
Example of two, discontiguous domains seen using a contact map

Contacts are mostly within domains, not between domains. One domain consists of N and C-terminal parts.
C/N-Terminal domain, cut-and-pasted

(research)
Exercise 16.1: Superimpose by hand

Do this pair: 1WFA.A vs 1WFA.B (2 chains of the same PDB structure)

File | Open: RCSB PDB: code: 1WFA
Ribbon | Style: oval, Color: chain or terminus
Select | synchronize (check if not already checked)
In SEQ window (cntl-Q)
    Double-click on chain label to select one molecule.
In MOE window (cntl-M) practice these moves. Superpose the chains.
    Rotate selected: meta-middlemouse-drag.
    Translate selected: shift-meta-middlemouse-drag
    Rotate all: middlemouse-drag
    Translate all: shift-middlemouse-drag

Share screen to show me your superposition.
Exercise 16.2: Superimpose automatically

Same chains: 1WFA.A vs 1WFA.B

Do these steps.

1. SEQ | Alignment|Align/Superpose
2. Open setup chains. Select waters (click on chain name), set to “i” (ignore)
3. Align (sequence and structural)
4. Inspect by showing straight-line trace ribbon.
5. Superpose. (explore options). Try selecting the C-terminal half (either MOE left-mouse drag or SEQ left-mouse drag along “ruler”), in menu set Selected Residues, then Superpose again. Do same after selecting N-terminal half. What is happening?

Share screen to show me your superposition.
Exercise 16.5: domain boundaries

6vsb. — Coronavirus spike protein, a multi domain protein.

File | Open | PDB: 6vsb


Where are the domains? What kind are they?

Select atoms of each domain. Color domains differently.
Homework 1 -- domains in coronavirus spike protein

- Align and superpose the three protein chains of SAR-2 spike (6vsb)
- Why doesn't the whole molecule superpose well?
- Superpose based on the receptor domain only ACE2 binding domain, residues 330-440
- Draw a TOPS diagram.
- Some loops are missing!
- Do http://www.bioinfo.rpi.edu/bystrc/courses/biol4550/homework1.pdf
- Turn in as PDF file: http://www.bioinfo.rpi.edu/bystrc/courses/biol4550/homework.html
test drive the homework server

• Goto http://www.bioinfo.rpi.edu/bystrc/courses/biol4550/homework.html
• Upload a file for homework 1. It can be any file. (I will delete it)
• Problems? Send me email.
Review questions

• What is a domain?
• What is a sequence “family” according to SCOP?
• What does “strand order” mean w/respect to SCOP naming?
• What defines a sequence “superfamily”?
• What characterizes a “fold”?
• Draw a beta-alpha-beta unit using TOPS.
• Draw a simplified contact maps based on a TOPS diagram.
• Find domain boundaries using a contact map.
• How can we infer domain boundaries using a multiple sequence alignment?
• In a TOPS diagram, what does a triangle pointing up mean?
Supplementary slides
CATH

- Class
- Architecture
- Topology
- Homology

Architecture = conserves arrangement of SSE (secondary structural elements) but not sequential order.

Topology = like SCOP Fold.

http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html
protein structure and representation - a hierarchy or a continuum?

<table>
<thead>
<tr>
<th>Structure</th>
<th>representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary structure</td>
<td>1D, three states</td>
</tr>
<tr>
<td>Local structure</td>
<td>motifs, backbone angles</td>
</tr>
<tr>
<td>Super-secondary structure</td>
<td>TOPS.</td>
</tr>
<tr>
<td>Inter-residue distances</td>
<td>2D contact maps</td>
</tr>
<tr>
<td>Tertiary structure</td>
<td>3D backbone</td>
</tr>
<tr>
<td>Side chain conformation</td>
<td>rotamers</td>
</tr>
<tr>
<td>Domain-domain interactions</td>
<td>interface maps</td>
</tr>
<tr>
<td>Quaternary structure</td>
<td>poses, interaction maps.</td>
</tr>
</tbody>
</table>