Protein Prediction Part 1: Structure

Structure comparison and classification

Andrea Schafferhans
(built on slides from Marco Punta, Burkhard Rost, Patrice Koehl)
Marco Punta contributed the slides

- PhD in Trieste (MD for membrane proteins)
- Postdoc @ Columbia Univ in the City of New York (contact predictions)
- Senior scientist in NYCOMPS (Target selection for membrane proteins)
- IAS Fellow @ TUM
- Project manager @ Pfam @ Sanger Inst. Hinxton (Cambridgeshire)
Notation: protein structure 1D, 2D, 3D
Structural Universe – No islands really!

B Rost 1998 Structure 6:259-263
3D classifications: goals

- Similar 3D -> Similar function
- Learn from 3D about function
- Learn about evolution
- classify
3D modules

© Christine Orengo
Fold of a protein

- some structures more often observed than others
- limited number of shapes?
- fold remains an assumption
  (that increasingly seems to be proven inappropriate)
Protein structure classification

All-alpha

3sdh

All-beta

1bww

AlphaBeta

1xne
3D classification databases

• SCOP
  http://scop.mrc-lmb.cam.ac.uk/scop/
  [A Murzin et al. (1995) JMB 247, 536-540]

• CATH
  http://www.cathdb.info/

• COPS
  http://cops.services.came.sbg.ac.at
  [SJ Suhrer et al. (2009) NAR 37, W539-W44. ]
Alexei Murzin

- Cambridge University, England
- CASP assessor
- ~90 publications
- 1 with over 3,000 quotes
- 13 with over 100 quotes (ISI 2011/05)
- H-index: 30 (ISI 2011/05)
SCOP hierarchy

Example

{All-alpha} \textit{a. class}

Structure similarity increases
SCOP classes

- alpha
- beta
- alpha and beta (a/b – interspersed)
- alpha plus beta (a+b – segregated)
- multidomain proteins
- membrane and cell-surface proteins
- small proteins
- coiled coil proteins
- low-resolution protein structures
- peptides
- designed proteins
SCOP class example

CLASS = alpha and beta (a/b)

NAD(P)-binding Rossmann-fold domains 1sw0-TIM

1sw0-TIM beta/alpha barrel
SCOP hierarchy

Example

{All-alpha}

a.

{Globin-like}

a.1

class

fold

Structure similarity increases
SCOP fold definition

• same major secondary structures
  – in the same arrangement
  – with the same topological connections

• peripheral elements may differ
  – up to 50% peripheral
  – Turns and secondary structure elements

• evolutionary relationship unclear
SCOP fold example

CLASS = alpha and beta (a/b)
FOLD = TIM beta/alpha-barrel
SCOP hierarchy

Example

{All-alpha}

{Globin-like}

{alpha-helical ferrodoxin}

a. class

a.1 fold

a.1.2 superfamily

Structure similarity increases
SCOP superfamily definition

- probable common evolutionary origin
- low similarities, but
  - share the same fold
  - have similar functions
SCOP superfamily example

- Triosephosphate isomerase (1swo)
- Phosphate aldolase (1p1x)
- Quinolinic acid phosphoribosyl transferase (1qap)
SCOP hierarchy

Example

{Alpha and beta a/b}  c  class

{TIM beta/alpha-barrel}  c.1  fold

{Triosephosphate isomerase}  c.1.1  superfamily

{Triosephosphate isomerase}  c.1.1.1  family  (sequence based)

Structure similarity increases
SCOP family definition

- clearly evolutionary relation
- Sequence identity often >30%, but not necessarily, e.g. globins: < 15% sequence identity for some members
3D classification databases

- SCOP
  http://scop.mrc-lmb.cam.ac.uk/scop/
  [A Murzin et al. (1995) JMB 247, 536-540]

- CATH
  http://www.cathdb.info/

- COPS
  http://cops.services.came.sbg.ac.at
  [SJ Suhrer et al. (2009) NAR 37, W539-W44.]
Christine A. Orengo

- UCL, England
- CASP assessor
- over 120 publications
- 1 with over 1,500 quotes
- 16 with over 100 quotes
- H-index >40 (ISI 2011/05)

- SSAP (with Willy Taylor)
- CATH
Dame Janet M. Thornton

- Director EBI (European Bioinformatics Institute, Hinxton, Cambridgeshire, England)
- BS Physics (Univ Nottingham), MS Biophysics King’s College London, PhD Biophysics UCL
- Amongst Top 100 scientists in UK
- ~400 publications
- 1 with over 11,000 quotes
- 7 with over 1,000 quotes
- 81 with over 100 quotes
- H-index >88 (ISI 2011/05)
Universe of protein structures
CATH

- Class
- Architecture
- Topology
- Homology
CATH classes

- mostly alpha, mostly beta, mixed alpha/beta, few regular secondary structure
CATH hierarchy

- **Class**
  mostly alpha, mostly beta, mixed alpha/beta, few regular secondary structure

- **Architecture**
  classification according to overall shape, *ignoring connectivity*

- **Topology**
  fold groups = shape & connectivity

- **Homology**
  evolutionarily related superfamily
CATH statistics (Version 3.4 – 2010)

• Architectures 40
• Topologies (folds) 1,282
• Homologous superfamilies 2,549
• sequence families (S35) 11,330
• Total domains 152,920

• new domains (ca. 1 year) 24,232
• new topologies 49

Martin,A. et al. (1998)
Protein folds and functions.
Structure, 6, 875-884.
CATH: steps involved

- Find domains
  - *ab initio*: consensus of three methods:
    - DETECTIVE: hydrophobic interior
    - PUU: likely separation motion
    - DOMAK: count internal and external contacts
  - Problem: only 20% consistent!
  - Based on prior knowledge: CATHEDRAL
    - GT: secondary structure matching
    - DDP: structural alignment

CATH: steps involved

- Find domain
- From domain to superfamily

PDB id: 1gcq
(SH3 domains)

© CATH tutorial (www.cathdb.info)
CATH: steps involved

- Find domain
- From domain to superfamily

CATH Domain: 1gcqA00

PDB 1gcq, Chain A, Domain 0

<table>
<thead>
<tr>
<th>CATH Code</th>
<th>Level Description</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Mainly Beta</td>
<td></td>
</tr>
<tr>
<td>2.30</td>
<td>Roll</td>
<td></td>
</tr>
<tr>
<td>2.30.30</td>
<td>SH3 type barrels.</td>
<td></td>
</tr>
<tr>
<td>2.30.30.40</td>
<td>SH3 Domains</td>
<td>Gene3D</td>
</tr>
<tr>
<td>2.30.30.40.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.30.30.40.22.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.30.30.40.22.1.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.30.30.40.22.1.1.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© CATH tutorial (www.cathdb.info)
## SCOP / CATH method comparison

<table>
<thead>
<tr>
<th>SCOP</th>
<th>CATH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>Class</td>
</tr>
<tr>
<td>Fold</td>
<td>Architecture (independent of connectivity)</td>
</tr>
<tr>
<td>Superfamily (evolutionary relation)</td>
<td>Topology</td>
</tr>
<tr>
<td>Family</td>
<td>Homologous Superfamily</td>
</tr>
<tr>
<td>Manual domain definition (larger domains)</td>
<td>Automatic domain definition</td>
</tr>
</tbody>
</table>
SCOP / CATH numerical comparison

At 80% residue overlap for a domain:
70% of proteins have similar domain definitions

### Table 4: Detailed mappings of domain pairs in percent from SCOP onto CATH

<table>
<thead>
<tr>
<th></th>
<th>outer</th>
<th>class</th>
<th>fold</th>
<th>superfamily</th>
<th>family</th>
</tr>
</thead>
<tbody>
<tr>
<td>outer</td>
<td>79.38%</td>
<td>8.31%</td>
<td>0.99%</td>
<td>0.40%</td>
<td>0.03%</td>
</tr>
<tr>
<td>class</td>
<td>18.16%</td>
<td>56.15%</td>
<td>2.55%</td>
<td>1.88%</td>
<td>0.87%</td>
</tr>
<tr>
<td>arch</td>
<td>2.42%</td>
<td>24.90%</td>
<td>2.80%</td>
<td>1.27%</td>
<td>0.09%</td>
</tr>
<tr>
<td>top</td>
<td>0.04%</td>
<td>10.50%</td>
<td>81.99%</td>
<td>4.44%</td>
<td>0.66%</td>
</tr>
<tr>
<td>hom</td>
<td>0.002%</td>
<td>0.14%</td>
<td>11.66%</td>
<td>92.01%</td>
<td>98.34%</td>
</tr>
</tbody>
</table>

Classification caveats

- Similar secondary structure arrangements across different classes
- Similar binding sites
3D classification databases

- SCOP
  http://scop.mrc-lmb.cam.ac.uk/scop/
  [A Murzin et al. (1995) JMB 247, 536-540]

- CATH
  http://www.cathdb.info/

- COPS
  http://cops.services.came.sbg.ac.at
  [SJ Suhrer et al. (2009) NAR 37, W539-W44.]
Manfred J. Sippl

- CAME, Univ. Salzburg
- CASP assessor
- over 54 publications
- 1 with over 800 quotes
- 10 with over 100 quotes
- H-index >27 (ISI 2011/05)

- “Sippl” potentials of pairwise energies ("Knowledge-based potentials")
COPS hierarchy

- COPS = Classification Of Protein Structures
- Based on quantified structural comparison
- 2007: additional info for SCOP domains: qSCOP
- 2009: workbench based on PDB chains: TopSearch
  http://topsearch.services.came.sbg.ac.at/

COPS metric

Axioms / Definitions:

\[ S_{a,a} = L_a \]
\[ S_{a,b} \geq 0 \]
\[ S_{a,b} = S_{b,a} \]
\[ S_{b,c} \geq S_{a,b} + S_{a,c} - L_a \]
\[ D_{a,b} = L_a + L_b - 2S_{a,b} \]

\[ S_{a,b}: \text{similarity (}\#\ \text{equiv. res.}) \]
\[ L_a: \text{length} \]

\[ s_{a,b} = \frac{2S_{a,b}}{L_a + L_b} \]
\[ c_{a,b} = \frac{S_{a,b}}{L_a} \]

• Alignment method not so important!
  for COPS: TopMatch

• Metric can reveal alignment problems
  (e.g. via triangle inequality)
COPS hierarchy

• COPS = Classification Of Protein Structures
• Based on quantified structural comparison
• 2007: additional info for SCOP domains: qSCOP
• 2009: workbench based on PDB chains: TopSearch
  http://topsearch.services.came.sbg.ac.at/

L80 – similar
L60 – related
L40 – distant
L30 – remote

PDB updates 2008/08/19-2009/04/14

SJ Suhrer et al. (2009) NAR 37:W539-W544
PDB diversity in light of COPS

SJ Suhrer et al. (2009) NAR 37:W539-W544
COPS domain exploration

1z6t – human apoptotic protease- activating factor 1 bound to ADP; constituent domains

1z6t domains aligned to domains of 2a5y (CED-4-CED-9 complex)

SJ Suhrer et al. (2009) NAR 37:W539-W544
COPS exploration – different alignments

(a) Superimposition of chains 1z6t-A and 2a5y-B.

SJ Suhrer et al. (2009) NAR 37:W539-W544
COPS ➔ TopSearch

- No domain decomposition
- But:
  - Complete structure comparisons
  - Biological units
  - New metric

TopMatch – new score (2012)

$$S = \sum_{i=1}^{L} e^{-r_i^2 / \sigma^2}$$

with

$$r_i^2 = (x_i - y_i)^2$$

$$s = \frac{S}{L} = \frac{1}{L} \sum_{i=1}^{L} e^{-r_i^2 / \sigma^2}$$

$$E_r = \sqrt{\frac{1}{L} \sum r_i^2}$$

$$S_r = \sigma \sqrt{-\ln s}$$

TopMatch:

- Query 1z6t@2  S 466
- Target 3shf@1  S_r 1.14
Structure alignment

Slides taken from Patrice Koehl, UC Davis

Structure alignment: two steps

1. Identify equivalent positions (residues that match in 3D)
2. find superposition independent of domain movements
Root mean square displacement (rmsd)

\[ \text{rmsd}(A,B) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} d_i^2} \]

d_i are the distances between all corresponding points (typically: C_\text{alpha}, all atoms)
RMSD is not a metric

cRMS = 2.8 Å

cRMS = 2.85 Å
## Structure alignment methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Authors</th>
<th>Year, Journal, Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSAP</td>
<td>WR Taylor &amp; CA Orengo</td>
<td>1989, JMB 208:1-22</td>
</tr>
<tr>
<td>DALI</td>
<td>L Holm &amp; C Sander</td>
<td>1993, JMB 233:123-38</td>
</tr>
<tr>
<td>CE</td>
<td>IN Shindyalov &amp; P Bourne</td>
<td>1998, Prot Engng 1:739-47</td>
</tr>
<tr>
<td>SKAN</td>
<td>A Yan, D Petrey &amp; B Honig</td>
<td>unpublished</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Idea**: use distance matrix and apply dynamic programming
SSAP – objective function

- Optimize:

\[ S_{ik} = \sum_{m=\pm n}^{m=+n} a \left| d_i^{A} - d_k^{B} \right| + b \]

**Problem:** loss of information about direction

- Replace distances by interatomic vectors (V)

Optimize:

\[ S_{ik} = a + w D_{ik} \]

\[ \left| V_{ij}^{A} - V_{kl}^{B} \right| + b \]

- Include sequence information
  (D(xy): Dayhoff)
SSAP – structure description

\[ \delta = v_{i \rightarrow i-2} - v_{k \rightarrow k-2} \]
VOROLIGN

Fabian Birzele, Ralf Zimmer et al.

VOROLIGN

Dynamic programming on Voronoi environments

F Birzele, JE Gewehr, G Csaba & R Zimmer (2006) Bioinformatics 23:e205-11: Fig. 2
DALI
Liisa Holm & Chris Sander

Liisa Holm
Univ of Helsinki
Finland

Chris Sander
SKCC New York

DALI – algorithm

Distance matrix Alignment

Monte Carlo on all-against-all for hexapeptides:

• Divide into overlapping peptides
• Generate contact map for each peptide
• Compare hexapeptide maps from A and B
• Combine matches based on simulated annealing
Comparison of structure alignments

Rachel Kolodny, Patrice Koehl, Michael Levitt

Comparison of structure alignments

dashed lines: original method
solid lines: SAS measure
2 forms of calcium-bound Calmodulin

Two forms of calcium-bound Calmodulin:

- Ligand free
- Complexed with trifluoperazine
Global alignment:
RMSD = 15 Å / 143 residues

Local alignment:
RMSD = 0.9 Å / 62 residues
Take home message / Questions