title: Computational Biology 1 - Protein structure:

**Intro: 3D comparisons**

short title: cb1_intro_3dcompare

lecture: Protein Prediction 1 - Protein structure
Computational Biology 1 - TUM Summer 2015
Protein Prediction - Part 1: Structure

1 Introduction (contd.)
Notation: protein structure 1D, 2D, 3D

Notation: protein structure 1D, 2D, 3D
Notation: protein structure 1D, 2D, 3D
Comparing 3D structures
Blue and red similar?

Doyle et al. (1998) Science 280:69-77 - The structure of the potassium channel: molecular basis of K+ conduction and selectivity
Similarity now clear?

Doyle et al. (1998) Science 280:69-77 - The structure of the potassium channel: molecular basis of K+ conduction and selectivity
3D comparisons: how to?
Matching shapes

How to match?
How to match?
Differences for corresponding points
Differences for corresponding points
Differences for corresponding points
Differences for corresponding points
Differences for corresponding points
Differences for corresponding points

Difference
= d1+d2+d3...+d8
= |r1a-r1b|+...+|r8a-r8b|

RMSD (root mean square deviation)
=SQRT [ (r1a-r1b)^2+...(r8a-r8b)^2 ]

\[ RMSD(A, B) = \sqrt{\sum_{i} (r_i^A - r_i^B)^2} \]
Differences for corresponding points

\[ \text{RMSD}(A, B) = \sqrt{\sum_i \left( r_i^A - r_i^B \right)^2} \]
Actual algorithm inversed

☐ 1st: find corresponding points
☐ 2nd: superimpose

$$RMSD(A, B) = \sqrt{\sum_i (r_i^A - r_i^B)^2}$$
fit now?
Scaling easy for simple shapes

\[ x^2 + y^2 = r^2 \]
Proteins: points are defined->no scaling

Global vs. local comparisons
Global vs. local comparisons
Global vs. local comparisons

global solution 1:

global solution 2:
cut into "units"
cut into “units”
trouble: where to stop?

valid “unit” for comparison?
How to decide what is a valid unit?
Decision upon validity

valid “unit” for comparison?
Valid or not?

☐ Scientifically significant: some expert says
Valid or not?

- Scientifically significant: some expert says
- Statistically significant: background

[Graph showing distribution of counts with signal and background peaks]
Cut, match, compare by RMSD

\[ \text{RMSD}(A, B) = \sqrt{\sum_{i} (r_i^A - r_i^B)^2} \]
Only Cartesian RMSD comparison?

\[ \text{RMSD}(A, B) = \sqrt{\sum_i (r_i^A - r_i^B)^2} \]
## 2D: difference matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Diagram

```
 8 7
 6
 5
 4
 3
 2
 1
```

© Burkhard Rost

ROSTLAB

TUM
Comparison 2D: differences of differences

Total of 8 x 8 differences
3D comparisons: biology
Structure alignment

Slides taken from Patrice Koehl, UC Davis

Patrice Koehl
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)

- Step 1: find corresponding points in proteins A and B
- $d(i)$ are the distances between all corresponding points (typic: C$\alpha$, all atoms)

$$\text{rmsd}(A,B) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} d^2_{\text{C}\alpha, \text{all atoms}}}$$
RMSD is not a metric

A similar B
B similar C
NOT implying:
A similar C

cRMSD = 2.8 Å = 0.28 nm
cRMSD = 2.85 Å = 0.285 nm
SSAP
3D alignment
Taylor & Orengo
SSAP concept

WR Taylor & CA Orengo (1989)
Protein structure alignment
JMB 208:1-22

\[ S_{ik} = \sum_{m=-n}^{m=+n} \left| d_{i,i+m}^A - d_{k,k+m}^B \right| + b \]
SSAP concept

WR Taylor & CA Orengo (1989)
Protein structure alignment
JMB 208:1-22

\[ S_{ik} = \sum_{m=-n}^{m=n} \frac{a}{|d_{i,i+m}^A - d_{k,k+m}^B|} + b \]

Problem: loss of information about direction
DALI
3D alignment
Holm & Sander
Structural alignment: **DALI**


- Distance matrix Alignment
- Algorithm: Monte Carlo on all-against-all for hexapeptides (5)
Vorolign
3D alignment
Birzele & Zimmer
Structural alignment: VOROLIGN

- Dynamic programming on Voronoi environments

F Birzele, JE Gewehr, G Csaba & R Zimmer (2006) Bioinformatics 23:e205-11: Fig. 2
3D comparisons: others
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

Structure alignment methods

- **SSAP**
  WR Taylor & CA Orengo 1989 JMB 208:1-22

- **DALI**
  L Holm & C Sander 1993 JMB 233:123-38

- **STRUCTAL**

- **CE**
  IN Shindyalov & P Bourne 1998 Prot Engng 1:739-47

- **VAST**

- **LSQMAN**

- **SSM**

- **SKAN**
  A Yan, D Petrey & B Honig, unpublished

- ...
Comparison of structure alignments

Rachel Kolodny, Patrice Koehl, Michael Levitt

Rachel Kolodny
Univ of Haifa

Patrice Koehl
UC Davis

Michael Levitt
Stanford Univ

Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures
JMB 346:1173-88
How to assess 3D comparisons? standard-of-truth?
Comparison of structure alignments

R Kolodny, P Koehl & M Levitt (2004) JMB 346:1173-88 (Fig. 1A)

dashed lines: original method
solid lines: SAS measure
Comparison of structure alignments

dashed lines: original method
solid lines: SAS measure

JMB 346:1173-88
(Fig. 1A)

Best-of-All
3D comparisons: protein space and databases
Structural universe
Unit of the universe: a domain
3D modules

Multiple 3D alignment identifies consensus secondary structure
Guessing domains from sequence

protein A
protein B
protein C
protein D
protein E
protein F

domain 1  domain 2
Structural universe

B Rost 1998 *Structure* 6:259-263
Evolution of pieces

Fig 1: Proteins from pieces. Panel A shows two building block types, the αα-hairpin and the ββ-hairpin, and panel B protein domains formed by their repetition.

© Andrei Lupas MPI Tuebingen
Structure evolves without leaps?

Fig. 1: NV Grishin 2001 NAR 29:638-43

© Nick V Grishin HHMI + Univ Dallas
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

Multiple 3D alignment identifies consensus secondary structure

© 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 comparisons

All-alpha

All-beta

AlphaBeta

3sdh

1bww

1xne
How to recognize the similarity?
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 - QSCOP - TopMatch**
  
  http://cops.services.came.sbg.ac.at
  [SJ Suhrer et al. (2009) NAR 37, W539-W44.]
Classify protein structure: SCOP
3D classification databases

- SCOP
  - http://scop.mrc-lmb.cam.ac.uk/scop/
  - [Murzin et al. J. Mol. Biol. 247, 536-540]

- hierarchy
Protein structure comparisons

All-alpha
3sdh

All-beta
1bww

AlphaBeta
1xne
SCOP hierarchy

Example

{All-alpha}

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

CLASS = alpha and beta (a/b)

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

1sw0-TIM beta/alpha barrel
SCOP hierarchy

Example

{All-alpha}

{Globin-like}

a.

class

a.1

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

CLASS = alpha and beta (a/b)
FOLD  = TIM beta/alpha-barrel
Structural universe: no islands, really

B Rost 1998 Structure 6:259-263
SCOP hierarchy

Example

{All-alpha}

{Globin-like}

{alpha-helical ferredoxin}

a. class

a.1 fold

a.1.2 superfamily

Structure similarity increases
probable common evolutionary origin
low similarities, but
• share the same fold
• have similar functions
SCOP hierarchy

TRIOSEPHOSPHATE ISOMERASE (1swo)

QUINOLINIC ACID PHOSPHORIBOSYLTRANSFERASE (1gap)

PHOSPHATE ALDOLASE (1p1x)
SCOP hierarchy

Example

{Alpha and beta a/b}

{TIM beta/alpha-barrel}

{Triosephosphate isomerase}

\[ \text{c} \quad \text{class} \]

\[ \text{c.1} \quad \text{fold} \]

\[ \text{c.1.1} \quad \text{superfamily} \]

\[ \text{c.1.1.1} \quad \text{family} \]

(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/
[Murzin et al. J. Mol. Biol. 247, 536-540]
Classify protein structure: CATH
3D classification databases

- **SCOP**
  - [http://scop.mrc-lmb.cam.ac.uk/scop/](http://scop.mrc-lmb.cam.ac.uk/scop/)
  - [Murzin et al. J. Mol. Biol. 247, 536-540]

- **CATH**
  - [http://www.cathdb.info/](http://www.cathdb.info/)
  - [AL Cuff et al. (2009) NAR 37, D310-314; CA Orengo et al. (1997) Structure 15, 1093-1108]
Universe of protein structures

Christine Orengo et al. 1997 *Structures* 5 1093-1108
CATH

☐ Class
☐ Architecture
☐ Topology
☐ Homology
Class:
mostly alpha, mostly beta, mixed alpha/beta, few regular secondary structure
CATH

☐ 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
<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>folds</td>
<td>1,282</td>
</tr>
<tr>
<td>superfamilies</td>
<td>2,549</td>
</tr>
<tr>
<td>sequence families</td>
<td>11,330</td>
</tr>
<tr>
<td>domains</td>
<td>24,232</td>
</tr>
</tbody>
</table>
CATH: steps involved

☐ Find domain

Multiple 3D alignment identifies consensus secondary structure
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)

PDB id: 1gcqA0
(SH3 domain)

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

Find domain

PDB id: 1gcqA0
(SH3 domain)

CATH Domain: 1gcqA0

<table>
<thead>
<tr>
<th>CATH Code</th>
<th>Level Description</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.30.30</td>
<td>SH3 type barrels</td>
<td>Gene3D</td>
</tr>
<tr>
<td>2.30.40.22</td>
<td>SH3 Domains</td>
<td></td>
</tr>
</tbody>
</table>

http://www.cathdb.info/domain/1gcqA0

© CATH tutorial (www.cathdb.info)
CATH vs SCOP

- 70% of proteins in PDB have similar domains at 80% residue domain overlap.

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>

CATH: 50 structures - 1 superfamily

superfamily 3.40.640.10

Type I PLP-dependent aspartate aminotransferase-like (Major domain)
Rank by family size

Percentage of all domain sequences

Remaining families (new BIG)

BIG families (currently Pfam)

Structural families
(i.e. one or more solved structures CATH/SCOP)
3D classification databases

- SCOP
  http://scop.mrc-lmb.cam.ac.uk/scop/
  [Murzin et al. J. Mol. Biol. 247, 536-540]

- CATH
  http://www.cathdb.info/
Classify protein structure: COPS/QSCOP/TopMatch
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

Triangle inequality/transitivity:

\[ A \approx B \]  (A similar to B)
\[ B \approx C \]  (B similar to C)

\[ \not\Rightarrow B \approx C \]  (does not imply: A similar to C)
PDB updates 2008/08/19-2009/04/14

novelty:

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 parsing

Apaf-1
PDB id 1z6t

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

COPS c1z6tA1 (CARD domain) - c2a5yB1

COPS c1z6tA2 (α/β domain) - c2a5yB

COPS c1z6tA3 (helical domain I) - c2a5yB3

COPS c1z6tA4 (winged-helix domain) - c2a5yB4
COPS domain parsing

PDB id
1z6t-A
with 2a5y-B

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

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

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.]
3D classification databases

- **SCOP**
  
  http://scop.mrc-lmb.cam.ac.uk/scop/  
  A Murzin et al. 1995 JMB:247, 536-540
  
  Alexey Murzin et al, Cambridge UK

- **CATH**
  
  http://www.cathdb.info/  
  AL Cuff et al. 2009 NAR 37:D310-314;  
  CA Orengo et al. 1997 Structure 15:1093-1108
  
  Christine Orengo & Janet Thornton et al, UCL UK

- **COPS**
  
  http://cops.services.came.sbg.ac.at  
  SJ Suhrer et al. 2009 NAR 37:W539-W44
  
  Manfred Sippl et al. Salzburg