Protein Structure Comparison

ECS129
Patrice Koehl
Protein Structure Representation

CPK: hard sphere model  Ball-and-stick  Cartoon
Degrees of Freedom in Proteins

**Bond length**

1 \quad 2

**Bond angle**

**Dihedral angle**

1 \quad 2 \quad 3 \quad 4

1 \quad 2 \quad +
Protein Structure: Variables

Backbone: 3 angles per residue: $\phi, \varphi$ and $\omega$

Sidechain: 1 to 7 angles, $\chi$; each $\chi$ has 3 favored values: $60^\circ$, $-60^\circ$, $180^\circ$. 
Ramachandran Plots

All residues, but glycine

Glycine

Sequence versus Structure

- **The protein sequence is a string of letters**: there is an optimal solution (DP) to the problem of string matching, given a scoring scheme.

- **The protein structure is a 3D shape**: the goal is to find algorithms similar to DP that finds the optimal match between two shapes.
Protein Structure Comparison

- Global versus local alignment
- Measuring protein shape similarity
- Protein structure superposition
- Protein structure alignment
Global versus Local

Global alignment
Global versus Local (2)

Local alignment
Measuring protein structure similarity

Given two “shapes” or structures A and B, we are interested in defining a distance, or similarity measure between A and B.

- **Visual comparison**
- **Dihedral angle comparison**
- **Distance matrix**
- **RMSD (root mean square distance)**

Is the resulting distance (similarity measure) D a metric?

\[ D(A,B) \leq D(A,C) + D(C,B) \]
Comparing dihedral angles

_Torsion angles (ϕ,ψ) are:_
- local by nature
- invariant upon rotation and translation of the molecule
- compact (O(n) angles for a protein of n residues)

*But…*

_Add 1 degree To all ϕ, ψ*
Distance matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>3.8</td>
<td>6.0</td>
<td>8.1</td>
</tr>
<tr>
<td>2</td>
<td>3.8</td>
<td>0</td>
<td>3.8</td>
<td>5.9</td>
</tr>
<tr>
<td>3</td>
<td>6.0</td>
<td>3.8</td>
<td>0</td>
<td>3.8</td>
</tr>
<tr>
<td>4</td>
<td>8.1</td>
<td>5.9</td>
<td>3.8</td>
<td>0</td>
</tr>
</tbody>
</table>
Distance matrix (2)

- **Advantages**
  - invariant with respect to rotation and translation
  - can be used to compare proteins of different sizes

- **Disadvantages**
  - the distance matrix is $O(n^2)$ for a protein with $n$ residues
  - comparing distance matrix is a hard problem
  - insensitive to chirality
Root Mean Square Distance (RMSD)

To compare two sets of points (atoms) $A = \{a_1, a_2, \ldots, a_N\}$ and $B = \{b_1, b_2, \ldots, b_N\}$:

- Define a 1-to-1 correspondence between $A$ and $B$

  for example, $a_i$ corresponds to $b_i$, for all $i$ in $[1,N]$

- Compute RMS as:

\[
RMS(A, B) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} d(a_i, b_i)^2}
\]

$d(A_i, B_i)$ is the Euclidian distance between $a_i$ and $b_i$. 
Protein Structure Superposition

- Simplified problem: we know the correspondence between set A and set B
- We wish to compute the rigid transformation $T$ that best align $a_1$ with $b_1$, $a_2$ with $b_2$, $\ldots$, $a_N$ with $b_N$
- The error to minimize is defined as:

$$\varepsilon = \min_T \sum_{i=1}^{N} \left\| T(a_i) - b_i \right\|^2$$

Old problem, solved in Statistics, Robotics, Medical Image Analysis, \ldots
Protein Structure Superposition

- A rigid-body transformation $T$ is a combination of a translation $t$ and a rotation $R$: $T(x) = Rx + t$
- The quantity to be minimized is:

$$
\varepsilon = \min_{t,R} \sum_{i=1}^{N} \|Ra_i - b_i + t\|^2
$$
E is minimum with respect to t when:

\[
\frac{\partial \varepsilon}{\partial t} = 2 \sum_{i=1}^{N} (Ra_i - b_i + t) = 0
\]

Then:

\[
t = -R \left( \sum_{i=1}^{N} a_i \right) + \sum_{i=1}^{N} b_i
\]

If both data sets A and B have been centered on 0, then t = 0 !

**Step 1:** Translate point sets A and B such that their centroids coincide at the origin of the framework
Let $\mu_A$ and $\mu_B$ be then barycenters of A and B, and $A'$ and $B'$ the matrices containing the coordinates of the points of A and B centered on O:

$$\mu_A = \frac{1}{N} \sum_{i=1}^{N} a_i$$

$$\mu_B = \frac{1}{N} \sum_{i=1}^{N} b_i$$

$$A = [a_1 - \mu_A \ a_2 - \mu_A \ ... \ a_N - \mu_A]$$

$$B = [b_1 - \mu_B \ b_2 - \mu_B \ ... \ b_N - \mu_B]$$

Build covariance matrix: $C = AB^T$

$3xN \times Nx3 = 3x3$
The rotation part (2)

Compute SVD (Singular Value Decomposition) of $C$:

$$ C = UDV^T $$

$U$ and $V$ are orthogonal matrices, and $D$ is a diagonal matrix containing the singular values. $U$, $V$ and $D$ are 3x3 matrices.

Define $S$ by:

$$ S = \begin{cases} 
I & \text{if } \det(C) > 0 \\
\text{diag}\{1,1,-1\} & \text{otherwise}
\end{cases} $$

Then

$$ R = USV^T $$
The algorithm

1. Center the two point sets A and B

2. Build covariance matrix:
   \[ C = AB^T \]

3. Compute SVD (Singular Value Decomposition) of C:
   \[ C = UDV^T \]

4. Define S:
   \[ S = \begin{cases} 
   I & \text{if } \det(C) > 0 \\
   \text{diag}\{1,1,-1\} & \text{otherwise} 
   \end{cases} \]

5. Compute rotation matrix
   \[ R = USV^T \]

6. Compute RMSD:
   \[ \text{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} a_i^2 + \sum_{i=1}^{N} b_i^2 - 2 \sum_{i=1}^{3} d_i s_i}{N}} \]

\[ O(N) \text{ in time!} \]
Example 1: NMR structures

Superposition of NMR Models

1AW6
Example 2: Calmodulin

Two forms of calcium-bound Calmodulin:

- Ligand free
- Complexed with trifluoperazine
Example 2: Calmodulin

Global alignment:  
RMSD = 15 Å / 143 residues

Local alignment:  
RMSD = 0.9 Å / 62 residues
RMSD is not a Metric

cRMS = 2.8 Å
cRMS = 2.85 Å
Protein Structure Superposition Problem:

Given two sets of points $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_m)$ in 3D space, find the optimal subsets $A(P)$ and $B(Q)$ with $|A(P)| = |B(Q)|$, and find the optimal rigid body transformation $G_{opt}$ between the two subsets $A(P)$ and $B(Q)$ that minimizes a given distance metric $D$ over all possible rigid body transformation $G$, i.e.

$$\min_{G} \{ D(A(P) - G(B(Q))) \}$$

The two subsets $A(P)$ and $B(Q)$ define a “correspondence”, and $p = |A(P)| = |B(Q)|$ is called the correspondence length.
Two Subproblems

1. Find correspondence set

2. Find alignment transform
   (protein superposition problem)
Existing Software

- **DALI** (Holm and Sander, 1993)
- **SSAP** (Orengo and Taylor, 1989)
- **STRUCTAL** (Levitt et al, 1993)
- **VAST** [Gibrat et al., 1996]
- **LOCK** [Singh and Brutlag, 1996]
- **CE** [Shindyalov and Bourne, 1998]
- **SSM** [Krissinel and Henrik, 2004]
- …
Trial-and-Error Approach to Protein Structure Alignment

**Iterate N times:**

1. Set Correspondence $C$ to a **seed** correspondence set (small set sufficient to generate an alignment transform)
2. Compute the alignment transform $G$ for $C$ and apply $G$ to the second protein $B$
3. Update $C$ to include all pairs of features that are close apart
4. If $C$ has changed, then return to Step 2
Why Classifying?

- **Standard in biology:**
  - **Aristotle**: Plants and Animal
  - **Linnaeus**: binomial system
  - **Darwin**: systematic classification that reveals phylogeny

- *It is easier to think about a representative than to embrace the information of all individuals*
Protein Structure Classification

- Domain Definition
- 3 Major classifications
  - SCOP
  - CATH
  - DDD
Protein Domain: Definitions

1) Regions that display significant levels of sequence similarity
2) The minimal part of a gene that is capable of performing a function
3) A region of a protein with an experimentally assigned function
4) Region of a protein structure that recurs in different contexts and proteins
5) A compact, spatially distinct region of a protein
## Web services for domain identification

<table>
<thead>
<tr>
<th>Program</th>
<th>Web access</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIAL</td>
<td><a href="http://www.ncbs.res.in/~faculty/mini/ddbase/dial.html">http://www.ncbs.res.in/~faculty/mini/ddbase/dial.html</a></td>
</tr>
<tr>
<td>DomainParser</td>
<td><a href="http://compbio.ornl.gov/structure/domainparser">http://compbio.ornl.gov/structure/domainparser</a></td>
</tr>
<tr>
<td>DOMAK</td>
<td><a href="http://www.compbio.dundee.ac.uk/Software/Domak/domak.html">http://www.compbio.dundee.ac.uk/Software/Domak/domak.html</a></td>
</tr>
<tr>
<td>PDP</td>
<td><a href="http://123d.ncifcrf.gov/pdp.html">http://123d.ncifcrf.gov/pdp.html</a></td>
</tr>
</tbody>
</table>
Protein Structure Space

1CTF

68 AA

1TIM

247 AA

1K3R

268 AA

1A1O

384 AA

1NIK

4504 AA

1AON

8337 AA
PDB Statistics: Overall Growth of Released Structures Per Year

- Number of Structures Released Annually
- Total Number Available

Year: 1971 to 2018
Number of Entries: 0 to 140,000
## Current state of the PDB

### PDB Data Distribution by Experimental Method and Molecular Type

<table>
<thead>
<tr>
<th>Experimental Method</th>
<th>Proteins</th>
<th>Nucleic Acids</th>
<th>Protein/NA Complex</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-Ray</td>
<td>115135</td>
<td>1905</td>
<td>5872</td>
<td>10</td>
<td>122922</td>
</tr>
<tr>
<td>NMR</td>
<td>10626</td>
<td>1234</td>
<td>247</td>
<td>8</td>
<td>12115</td>
</tr>
<tr>
<td>Electron Microscopy</td>
<td>1422</td>
<td>30</td>
<td>497</td>
<td>0</td>
<td>1949</td>
</tr>
<tr>
<td>Other</td>
<td>204</td>
<td>4</td>
<td>6</td>
<td>13</td>
<td>227</td>
</tr>
<tr>
<td>Multi Method</td>
<td>103</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>109</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>127490</strong></td>
<td><strong>3176</strong></td>
<td><strong>6624</strong></td>
<td><strong>32</strong></td>
<td><strong>137322</strong></td>
</tr>
</tbody>
</table>
Welcome to SCOP: Structural Classification of Proteins.

1.75 release (June 2009)

38221 PDB Entries. 1 Literature Reference. 110800 Domains. (excluding nucleic acids and theoretical models).


Authors. Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. scop@mrc-lmb.cam.ac.uk


http://scop.mrc-lmb.cam.ac.uk/scop/
http://scop.berkeley.edu/
SCOP is organized into 4 hierarchical layers:

(1) Classes:

1. All alpha proteins [46456] (284)
2. All beta proteins [48724] (174)
3. Alpha and beta proteins (a/b) [51349] (147)
   Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) [53931] (376)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) [56572] (66)
   Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides [56835] (58)
   Does not include proteins in the immune system
7. Small proteins [56992] (90)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins [57942] (7)
   Not a true class
9. Low resolution protein structures [58117] (26)
   Not a true class
10. Peptides [58231] (121)
    Peptides and fragments. Not a true class
11. Designed proteins [58788] (44)
    Experimental structures of proteins with essentially non-natural sequences. Not a true class
Classification of Protein Structure: SCOP

(2) Folds: *Major structural similarity*

Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections.

3) Superfamily: *Probable common evolutionary origin*

Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies.

4) Family: *Clear evolutionarily relationship*

Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater.
Scop Classification Statistics

SCOP: Structural Classification of Proteins. 1.75 release
38221 PDB Entries (23 Feb 2009). 110800 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of folds</th>
<th>Number of superfamilies</th>
<th>Number of families</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha proteins</td>
<td>284</td>
<td>507</td>
<td>871</td>
</tr>
<tr>
<td>All beta proteins</td>
<td>174</td>
<td>354</td>
<td>742</td>
</tr>
<tr>
<td>Alpha and beta proteins (a/b)</td>
<td>147</td>
<td>244</td>
<td>803</td>
</tr>
<tr>
<td>Alpha and beta proteins (a+b)</td>
<td>376</td>
<td>552</td>
<td>1055</td>
</tr>
<tr>
<td>Multi-domain proteins</td>
<td>66</td>
<td>66</td>
<td>89</td>
</tr>
<tr>
<td>Membrane and cell surface proteins</td>
<td>58</td>
<td>110</td>
<td>123</td>
</tr>
<tr>
<td>Small proteins</td>
<td>90</td>
<td>129</td>
<td>219</td>
</tr>
<tr>
<td>Total</td>
<td>1195</td>
<td>1962</td>
<td>3902</td>
</tr>
</tbody>
</table>
Classification of Protein Structure: CATH

CATH / Gene3D v4.2
95 million protein domains classified into 6,119 superfamilies

What is CATH-Gene3D?

CATH is a classification of protein structures downloaded from the Protein Data Bank. We group protein domains into superfamilies when there is sufficient evidence they have diverged from a common ancestor.

- Search CATH by text, ID or keyword
- Search CATH by protein sequence
- Search CATH by PDB structure
- Browse CATH Hierarchy
- CATH Release Statistics
- CATH Tutorials

Gene3D uses the information in CATH to predict the locations of structural domains on millions of protein sequences available in public databases. This allows us to include additional annotations to the CATH-Gene3D database such as functional information and active site residues.

- Go to Gene3D
- Compare Genomes
- Download Gene3D Data
- Learn how Gene3D is created

If you have any questions, comments or suggestions please get in touch via Twitter, ask a question in our online forum or visit our support page.

Latest Release Statistics

<table>
<thead>
<tr>
<th></th>
<th>CATH-Plus 4.2.0</th>
<th>CATH (daily snapshot)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB Release</td>
<td>17-05-2017</td>
<td>5 days ago</td>
</tr>
<tr>
<td>Domains</td>
<td>434857</td>
<td>462567</td>
</tr>
<tr>
<td>Superfamilies</td>
<td>6119</td>
<td>6892</td>
</tr>
<tr>
<td>Annotated PDBs</td>
<td>131091</td>
<td>137491</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Gene3D v16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Sequences</td>
<td>52,073,853</td>
</tr>
<tr>
<td>CATH Domain Predictions</td>
<td>95,665,487</td>
</tr>
</tbody>
</table>

http://www.cathdb.info
Classification of Protein Structure: CATH

- Alpha
- Mixed Alpha Beta
- Beta
- Sandwich
- Barrel
- Super Roll
- Tim Barrel
- Other Barrel
The DALI Database

Dali Database

Dali structural neighbours

The Dali Database is based on all-against-all 3D structure comparison of protein structures in the Protein Data Bank (PDB). The structural neighbourhoods and alignments are automatically maintained and regularly updated using the Dali search engine.

- Please note that PDB structures released after the last update will not be in the database! If you wish to find structural neighbours of these proteins, you are advised to submit the structure to the Dali Server instead.
- If you want to superimpose two particular structures, you can do it in the pairwise DaliLite server.

* Last Update: 7 March 2011
Update frequency: twice a year

Enter PDB identifier:  
chain:  
(optional)  
submit  
clear  

(Keyword search for PDB identifiers)

Dali Database entries are retrieved on demand, and formatting the results page may take up to one minute. Return visits to an existing results page are much faster.

Example

Structural neighbours of 1tu9, a globin-like protein in bacteria. Tutorial

http://ekhidna.biocenter.helsinki.fi/dali/start
The DALI Domain Dictionary

- All-against-all comparison of PDB90 using DALI
- Define score of each pair as a Z-score
- Regroup proteins based on pair-wise score:
  - Z-score > 2: “Folds”
  - Z-score > 4, 6, 8, 10: sub-groups of “folds”
    (different from Families, and sub-families!)
Classification is an important part of biology; protein structures are not exempt.

Prior to being classified, proteins are cut into domains.

While all structural biologists agree that proteins are usually a collection of domains, there is no consensus on how to delineate the domains.

There are three main protein structure classification:
- SCOP (manual) *source of evolutionary information*
- CATH (semi-automatic) *source of geometric information*
- Dali (automatic) *source of raw data*