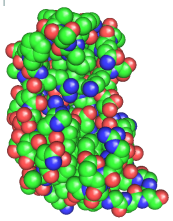
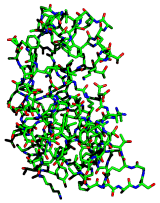
Protein Structure Comparison

ECS129
Patrice Koehl

Protein Structure Representation



CPK: hard sphere model

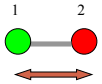


Ball-and-stick

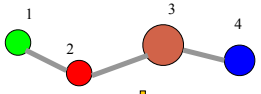
Cartoon

Degrees of Freedom in Proteins

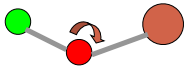
Bond length

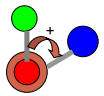


Dihedral angle

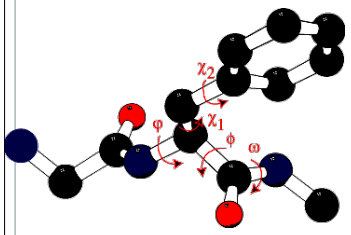


Bond angle





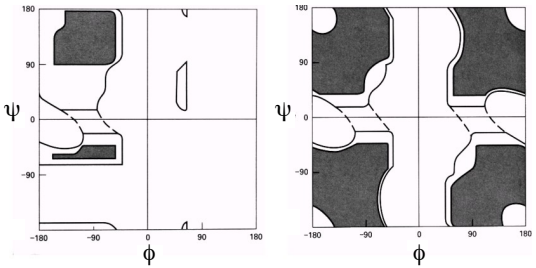
Protein Structure: Variables



Backbone: 3 angles per residue : φ, ϕ and ω

Sidechain: 1 to 7 angles, χ; each χ has 3 favored values: 60º, -60º, 180º.

Ramachandran Plots



All residues, but glycine

Glycine

Acta Cryst. (2002). D**58**, 768-776

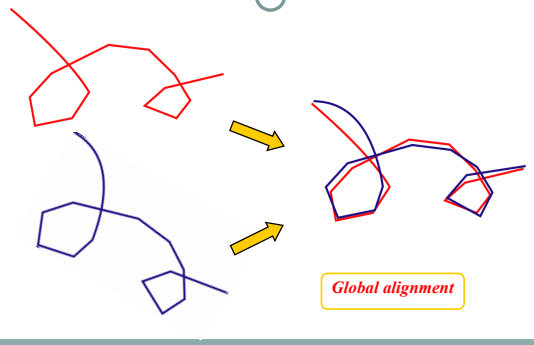
Sequence versus Structure

- ***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.

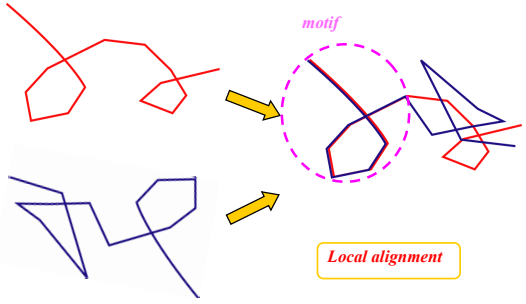
Protein Structure Comparison

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

Global versus Local



Global versus Local (2)



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)*

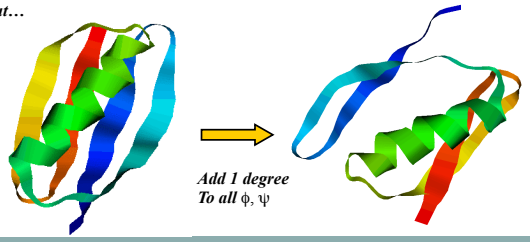
Is the resulting distance (similarity measure) D a metric?

$D(A,B) \leq D(A,C) + D(C,B)$

Comparing dihedral angles

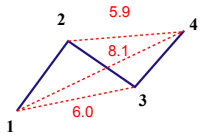
Torsion angles (ϕ,ψ) are:
- local by nature
- invariant upon rotation and translation of the molecule
- compact (O(n) angles for a protein of n residues)

But…



Add 1 degree
To all ϕ, ψ

Distance matrix



	1	2	3	4
1	0	3.8	6.0	8.1
2	3.8	0	3.8	5.9
3	6.0	3.8	0	3.8
4	8.1	5.9	3.8	0

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(n2) 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, …, a_N with b_N
- The error to minimize is defined as:

Old problem, solved in Statistics, Robotics, Medical Image Analysis, …

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

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:

$$\varepsilon = \min_{t,R} \sum_{i=1}^{N} \| Ra_i - b_i + t \|^2$$

The translation part

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

The rotation part (1)

Let μ_A and μ_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:

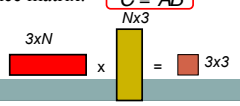
$$\mu_A = \frac{1}{N} \sum_{i=1}^{N} a_i$$

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

$$A = \begin{bmatrix} a_1 - \mu_A & a_2 - \mu_A & \dots & a_N - \mu_A \end{bmatrix}$$

$$B = \begin{bmatrix} b_1 - \mu_B & b_2 - \mu_B & \dots & b_N - \mu_B \end{bmatrix}$$

Build covariance matrix: $C = AB^T$



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:

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

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:

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

5. Compute rotation matrix

$$R = USV^T$$

6. Compute RMSD:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} a_i^2 + \sum_{i=1}^{N} b_i^2 - 2\sum_{i=1}^{3} d_i s_i}{N}}$$

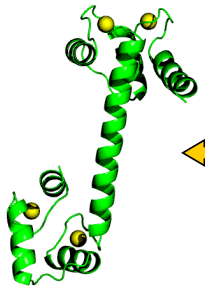
O(N) in time!

Example 1: NMR structures

Superposition of NMR Models

1AW6

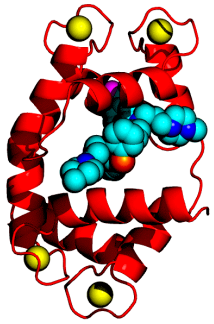
Example 2: Calmodulin



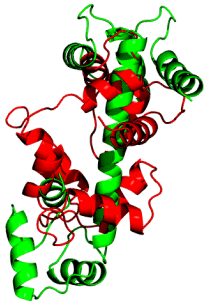
Two forms of calcium-bound Calmodulin:

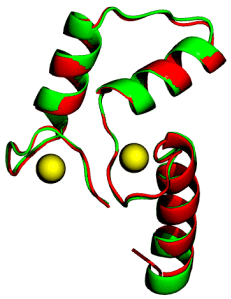
Ligand free

Complexed with trifluoperazine



Example 2: Calmodulin

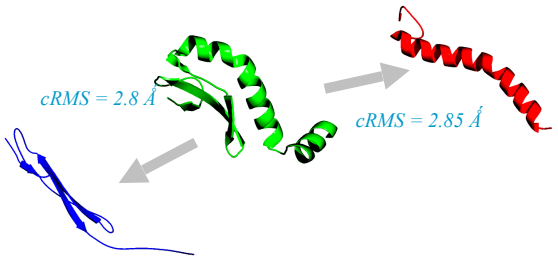




Global alignment:
RMSD =15 Å /143 residues

Local alignment:
RMSD = 0.9 Å/ 62 residues

RMSD is not a Metric



Protein Structure Alignment

Protein Structure Superposition Problem:

Given two sets of points *A=(a1, a2, …, an)* and *B=(b1,b2,…bm)* in 3D space, find the **optimal** subsets *A(P)* and *B(Q)* with |A(P)|=|B(Q)|, and find the **optimal** rigid body transformation *Gopt* 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

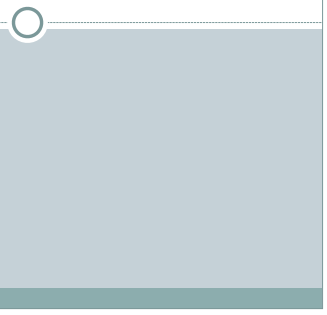
- DALI (Holm and Sander, 1993)
- SSAP (Orengo and Taylor, 1989)
- STRUCTAL (Levitt et al, 1993)
- VAST [Gibrat et al., 1996]
- LOCK [Singh and Brutlag, 1996]
- CE [Shindyalov and Bourne, 1998]
- SSM [Krissinel and Henrik, 2004]
- …

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

Protein Structure Classification



Why Classifying ?

- *Standard in biology:*

  Aristotle: Plants and Animal

  Linnaeus: binomial system

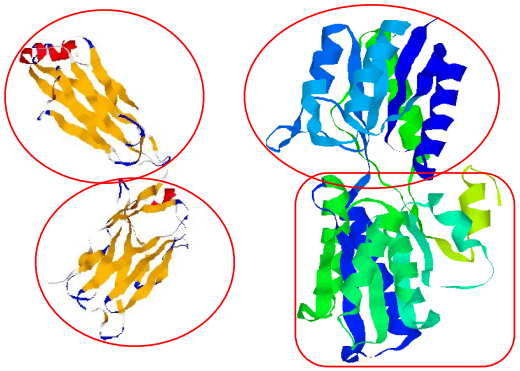
  Darwin: systematic classification that reveals phylogeny

- *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 Structural Domains



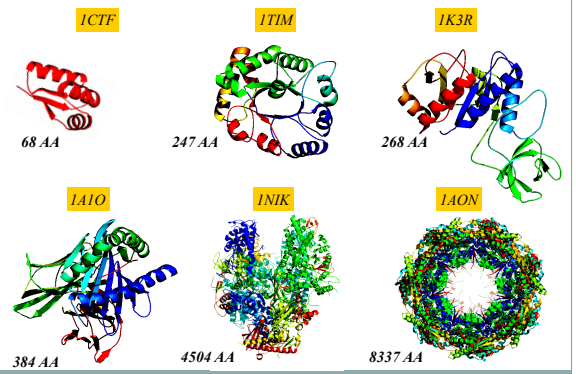
Protein Domain: Definitions

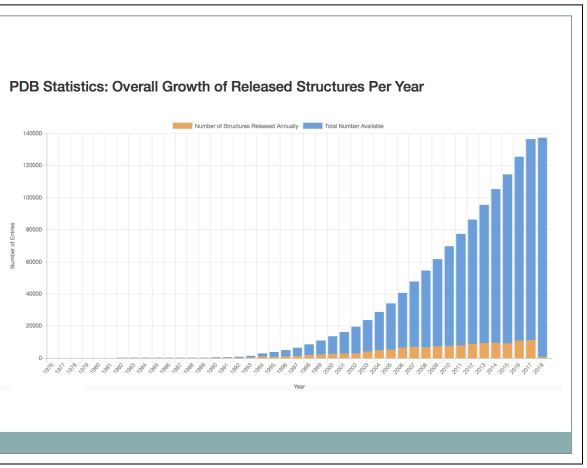
1) Regions that display significant levels of sequence similarity
2) The minimal part of a gene that is capable of performing a function
3) A region of a protein with an experimentally assigned function
4) Region of a protein structure that recurs in different contexts and proteins
5) A compact, spatially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

Web services for domain identification

Program	Web access
DIAL	http://www.ncbs.res.in/~faculty/mini/ddbase/dial.html
DomainParser	http://compbio.ornl.gov/structure/domainparser
DOMAK	http://www.compbio.dundee.ac.uk/Software/Domak/domak.html
PDP	http://123d.ncifcrf.gov/pdp.html

Protein Structure Space





Current state of the PDB

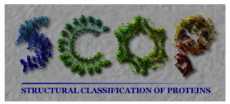
PDB Data Distribution by Experimental Method and Molecular Type

Copy CSV

Experimental Method	Proteins	Nucleic Acids	Protein/NA Complex	Other	Total
X-Ray	115135	1905	5872	10	122922
NMR	10626	1234	247	8	12115
Electron Microscopy	1422	30	497	0	1949
Other	204	4	6	13	227
Multi Method	103	3	2	1	109
Total	127490	3176	6624	32	137322

Classification of Protein Structure: SCOP

Structural Classification of Proteins



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).
Folds, superfamilies, and families statistics here.
New folds superfamilies families.
List of obsolete entries and their replacements.

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
Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [PDF]
Recent changes are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [PDF],
Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229. [PDF], and
Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2007). Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.* 2008 36: D419-D425; doi:10.1093/nar/gkm993 [PDF].

http://scop.mrc-lmb.cam.ac.uk/scop/
http://scop.berkeley.edu/

Classification of Protein Structure: SCOP

SCOP is organized into 4 hierarchical layers:

(1) Classes:

Root: scop

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

Classification of Protein Structure: SCOP

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)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	284	507	871
All beta proteins	174	354	742
Alpha and beta proteins (a/b)	147	244	803
Alpha and beta proteins (a+b)	376	552	1055
Multi-domain proteins	66	66	89
Membrane and cell surface proteins	58	110	123
Small proteins	90	129	219
Total	1195	1962	3902

Classification of Protein Structure: CATH

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.

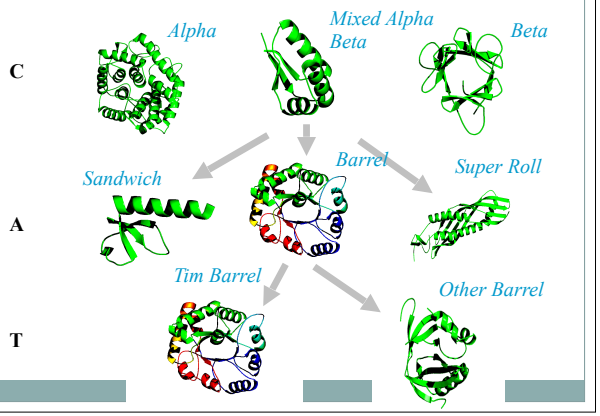
Latest Release Statistics

	CATH-Plus 4.2.0	CATH (daily snapshot)
PDB Release	17-05-2017	5 days ago
Domains	434857	462567
Superfamilies	6119	6892
Annotated PDBs	131091	137491

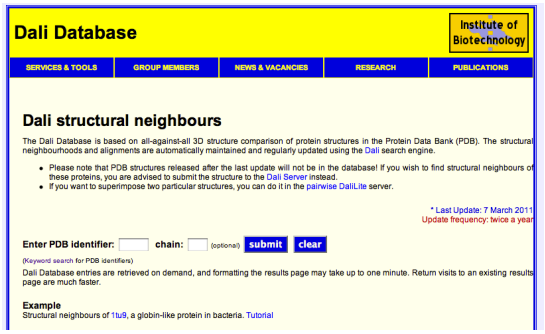
	Gene3D v16
Protein Sequences	52,073,853
CATH Domain Predictions	95,665,487

http://www.cathdb.info

Classification of Protein Structure: CATH



The DALI Database



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!)

Summary

- 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