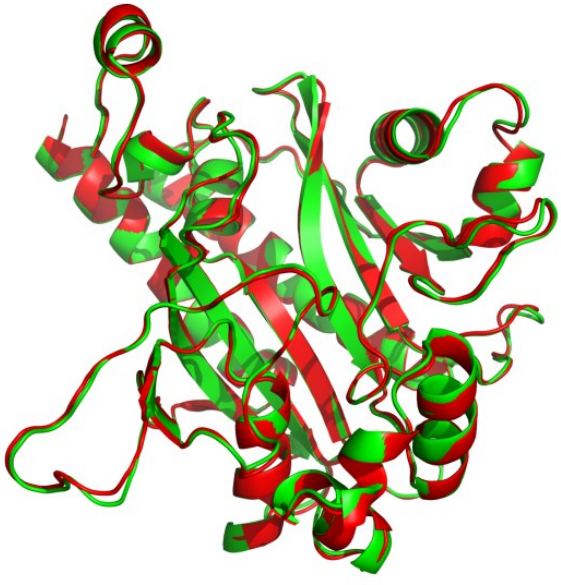
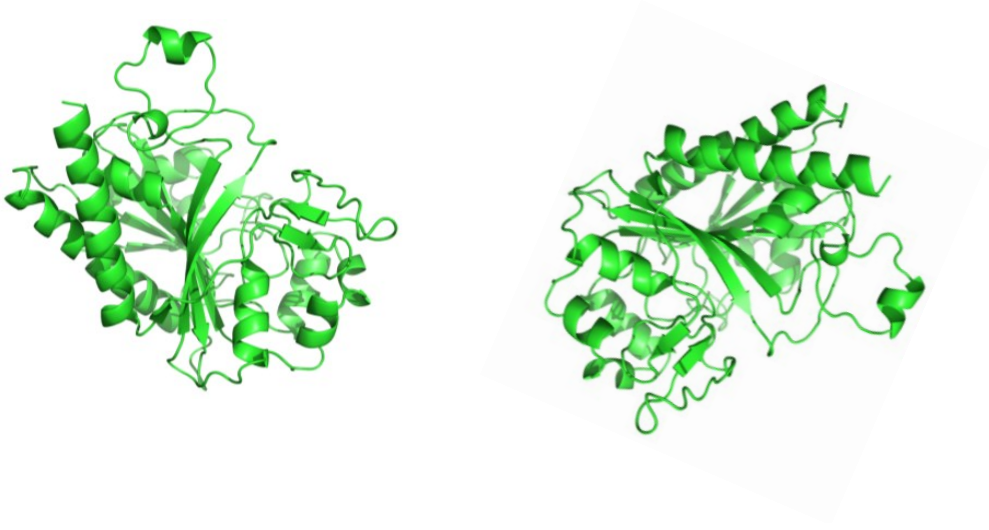
Lecture 18

Structure Alignment

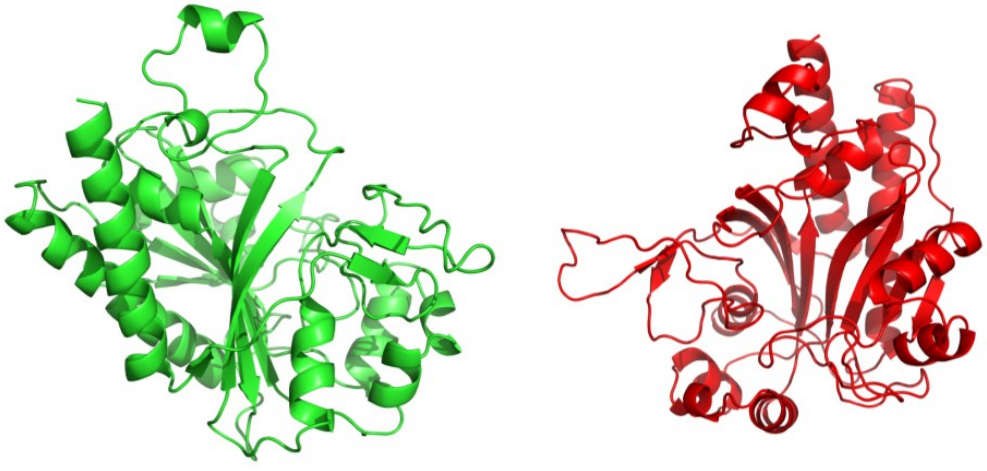


Why do we need structure alignment?



Frame of reference is different.

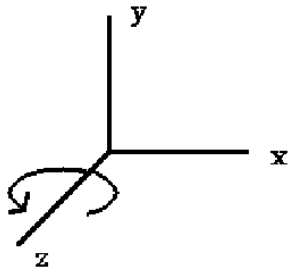
Why do we need structure alignment?



Frame of reference is different.

How?

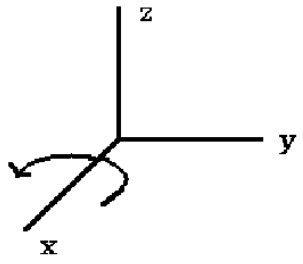
Rotation About Z-Axis in 3D



$$x' = x\cos q - y\sin q$$
$$y' = x\sin q + y\cos q$$
$$z' = z$$

$$R_z(q) = \begin{pmatrix} \cos q & -\sin q & 0 & 0 \\ \sin q & \cos q & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

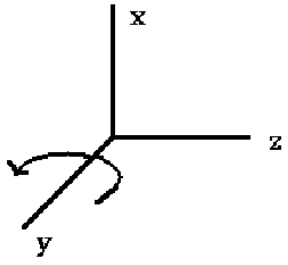
Rotation About X-Axis in 3D



y' = y*cos q - z*sin q
z' = y*sin q + z*cos q
x' = x

$$Rx(q) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos q & -\sin q & 0 \\ 0 & \sin q & \cos q & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Rotation About Y-Axis in 3D



z' = z*cos q - x*sin q
x' = z*sin q + x*cos q
y' = y

$$Ry(q) = \begin{pmatrix} \cos q & 0 & \sin q & 0 \\ 0 & 1 & 0 & 0 \\ -\sin q & 0 & \cos q & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Rotation About an Arbitrary Axis in 3D

(1) Translate space so that the rotation axis passes through the origin.

(2) Rotate space about the z axis so that the rotation axis lies in the xz plane.

(3) Rotate space about the y axis so that the rotation axis lies along the z axis.

(4) Perform the desired rotation by θ about the z axis.

(5) Apply the inverse of step (3).

(6) Apply the inverse of step (2).

(7) Apply the inverse of step (1).

Rotation About an Arbitrary Axis in 3D

The matrices for rotation by α around the x-axis, β around the y-axis, and γ around the z-axis

$$R_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha & 0 \\ 0 & \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_y(\beta) = \begin{bmatrix} \cos\beta & 0 & \sin\beta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\beta & 0 & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_z(\gamma) = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 & 0 \\ \sin\gamma & \cos\gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Rotation About an Arbitrary Axis in 3D

The general rotation matrix depends on the order of rotations. The first matrix rotates about x, then y, then z; the second rotates about z, then y, then x.

$$R_z R_y R_x = \begin{bmatrix} \cos\beta\cos\gamma & \cos\gamma\sin\alpha\sin\beta - \cos\alpha\sin\gamma & \cos\alpha\cos\gamma\sin\beta + \sin\alpha\sin\gamma & 0 \\ \cos\beta\sin\gamma & \cos\alpha\cos\gamma + \sin\alpha\sin\beta\sin\gamma & -\cos\gamma\sin\alpha + \cos\alpha\sin\beta\sin\gamma & 0 \\ -\sin\beta & \cos\beta\sin\alpha & \cos\alpha\cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_x R_y R_z = \begin{bmatrix} \cos\beta\cos\gamma & -\cos\beta\sin\gamma & \sin\beta & 0 \\ \cos\alpha\sin\gamma + \sin\alpha\sin\beta\cos\gamma & \cos\alpha\cos\gamma - \sin\alpha\sin\beta\sin\gamma & -\sin\alpha\cos\beta & 0 \\ \sin\alpha\sin\gamma - \cos\alpha\sin\beta\cos\gamma & \sin\alpha\cos\gamma + \cos\alpha\sin\beta\sin\gamma & \cos\alpha\cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Limitations

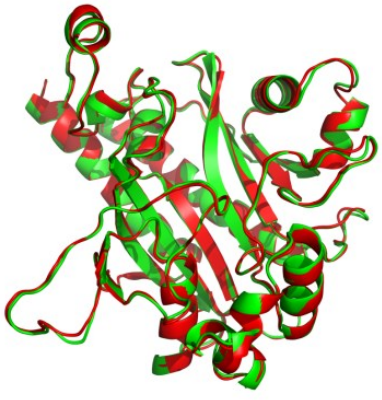
- Computationally slow
- Not recommended for large scale application
- Alternative is Quaternion based method.

$$\begin{bmatrix} c + a_x^2(1-c) & a_x a_y(1-c) - a_z s & a_x a_z(1-c) + a_y s \\ a_y a_x(1-c) + a_z s & c + a_y^2(1-c) & a_y a_z(1-c) - a_x s \\ a_z a_x(1-c) - a_y s & a_z a_y(1-c) + a_x s & c + a_z^2(1-c) \end{bmatrix}$$

Homework

- Develop a faster method compared to this.
  - Hint: Can use Quaternion.

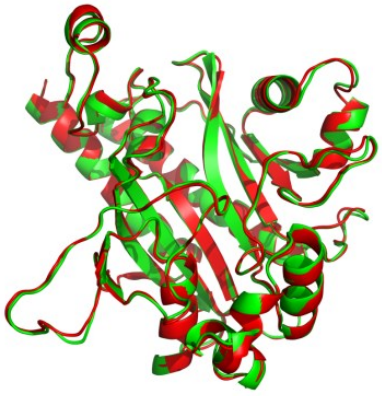
Measure of Structure Alignment



RMSD – Root Mean Square Deviation

All atom RMSD
Backbone RMSD
Backbone-trace RMSD

Measure of Structure Alignment

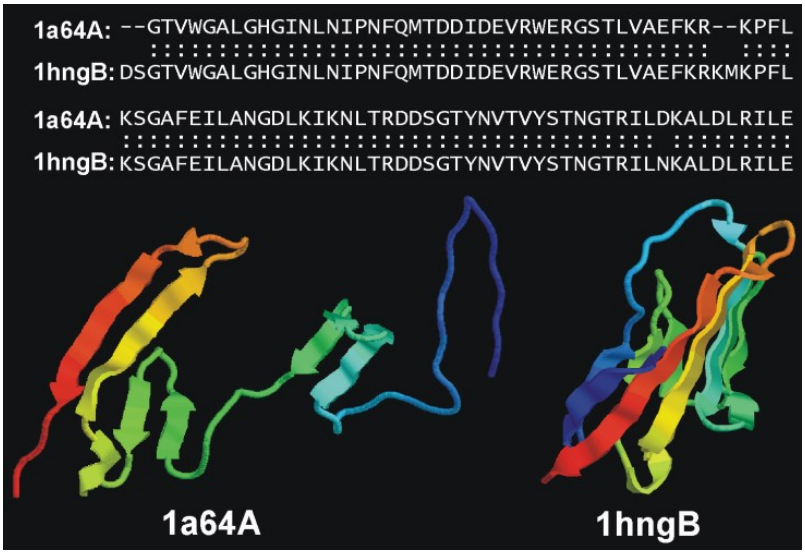


RMSD – Root Mean Square Deviation

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

All atom RMSD
Backbone RMSD
Backbone-trace RMSD

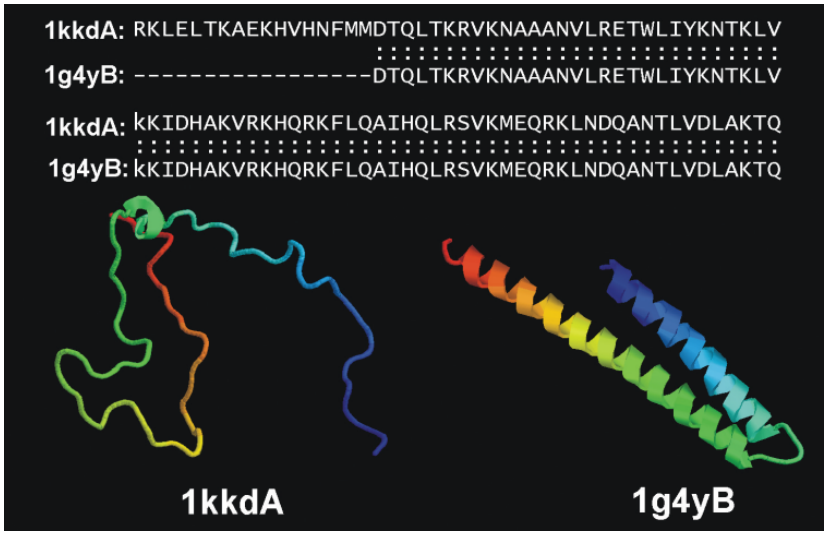
Why do we need structure alignment?



High sequence identity, but different fold.

Zhang & Skolnick (2005). NAR 33:2302–2309

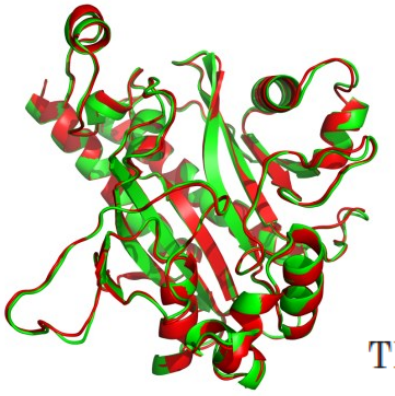
Why do we need structure alignment?



High sequence identity, but different fold.

Zhang & Skolnick (2005). NAR 33:2302–2309

Measure of Structure Alignment

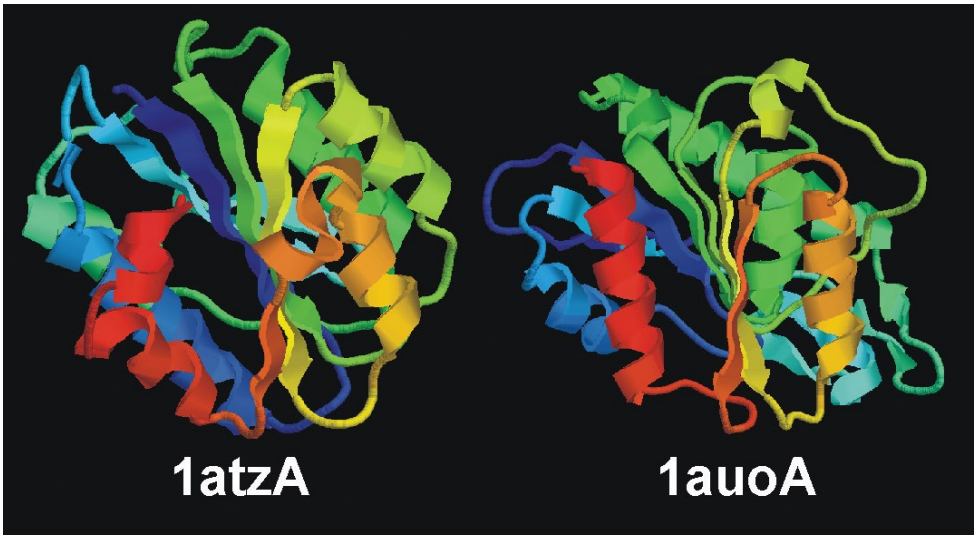


TM-Score

$$\text{TM-score} = \text{Max}\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right]$$

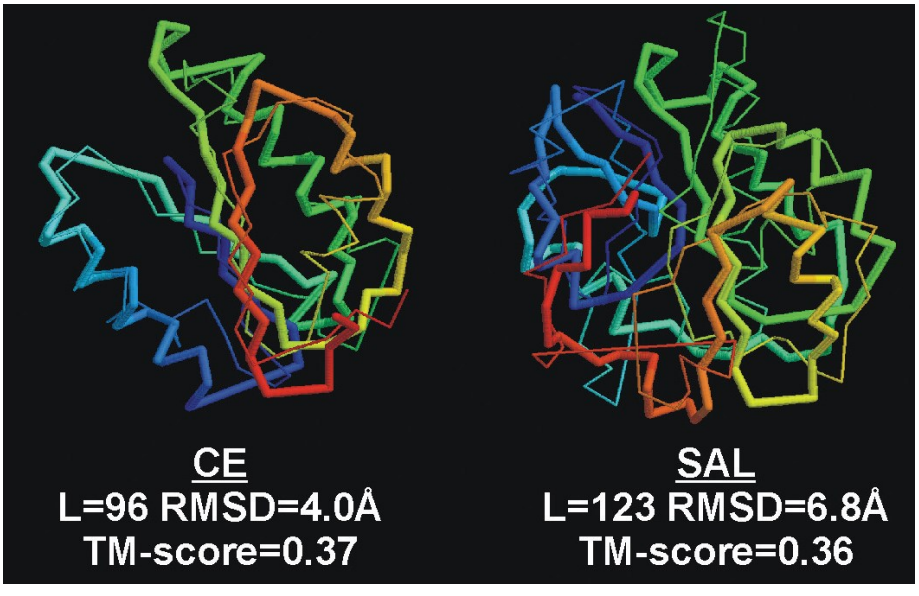
$$d_0 = 1.24\sqrt[3]{L_N - 15} - 1.8$$

Comparison



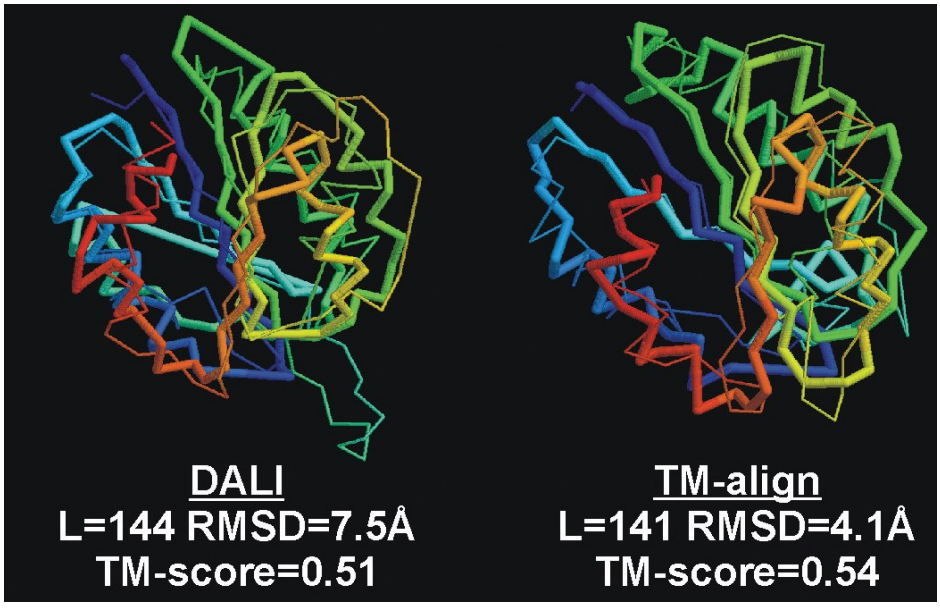
Zhang & Skolnick (2005). NAR 33:2302–2309

Comparison



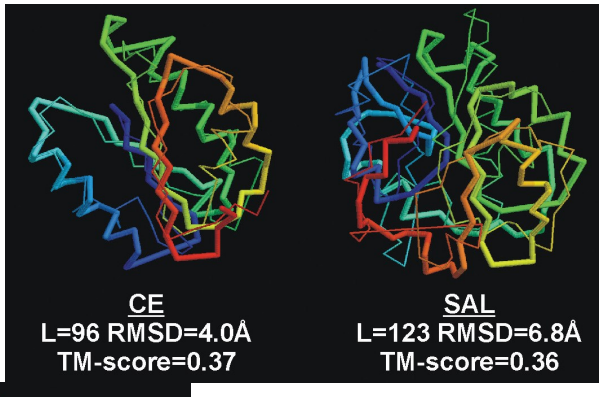
Zhang & Skolnick (2005). NAR 33:2302–2309

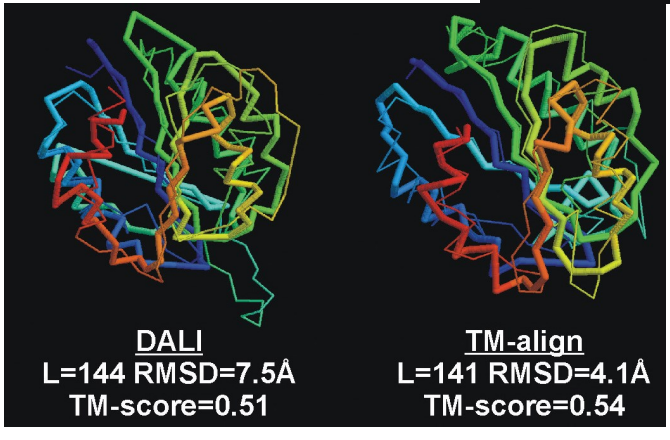
Comparison



Zhang & Skolnick (2005). NAR 33:2302–2309

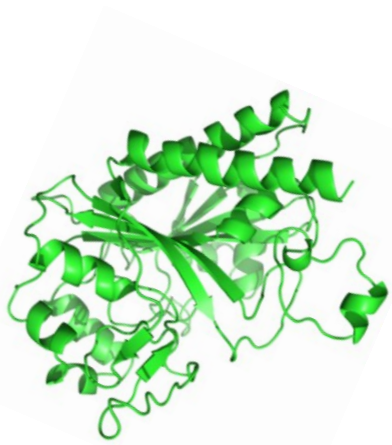
Comparison



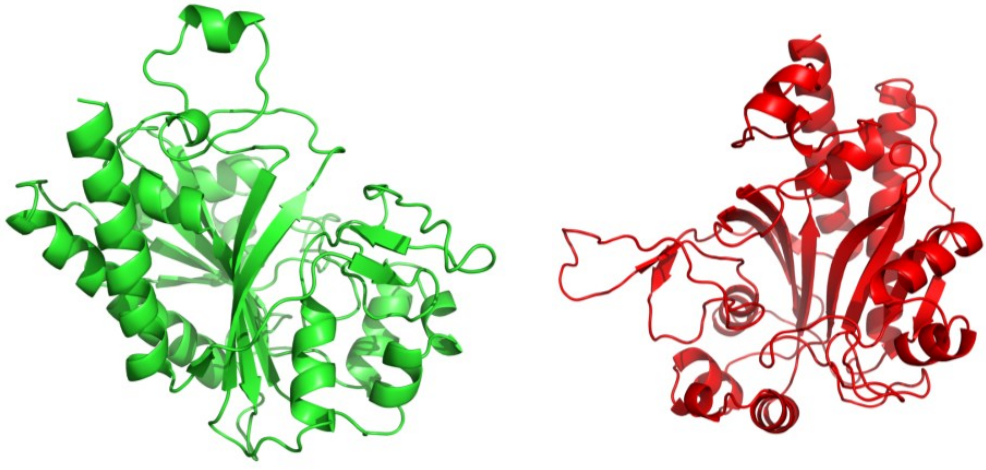


Zhang & Skolnick (2005). NAR 33:2302–2309

Protein Structure



Protein Structure



SCOP

- SCOP – Structural Classification of Proteins
- Website - http://scop.mrc-lmb.cam.ac.uk/scop/
- Utility: Evolutionary and Structural Classification of proteins.
- Method used for classification: Visual inspection and comparison of structures through various automated tools.

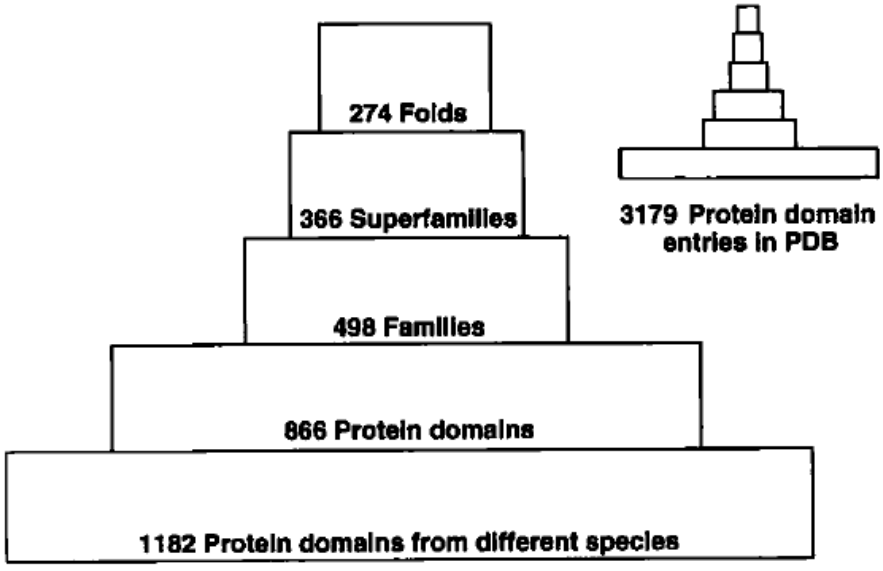
SCOP

- ***Family:*** Proteins are clustered together into families on the basis of one of two criteria that imply their having a common evolutionary origin: first, all proteins that have residue identities of 30% and greater; second, proteins with lower sequence identities but whose functions and structures are very similar; for example, globins with sequence identities of 15%.
- ***Superfamily:*** Families, whose proteins have low sequence identifies but whose structures and, in many cases, functional features suggest that a common evolutionary origin is probable, are placed together in superfamilies; for example, actin, the ATPase domain of the heat-shock protein and hexokinase.
- ***Common folds:*** Superfamilies and families are defined as having a common fold if their proteins have same major secondary structures in same arrangement with the same topological connections.
- ***Class:*** Most of the folds are assigned to one of the five structural classes on the basis of the secondary structures of which they composed.

SCOP class

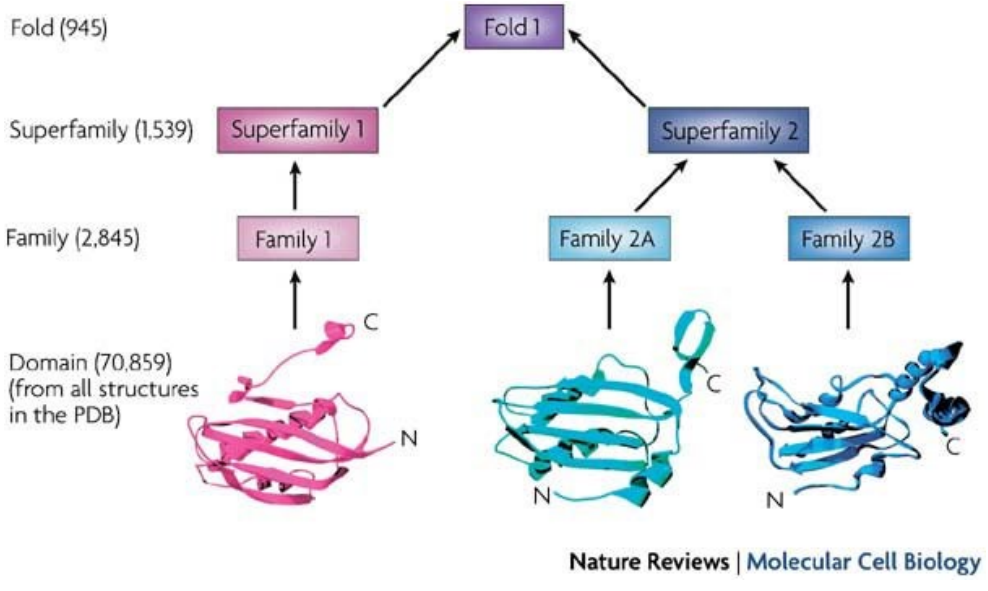
1. All alpha-(for proteins whose structure is essentially formed by α-helices),
2. All beta (for those whose structure is essentially formed by β-sheets),
3. Alpha and beta (for proteins with α-helices and β-strands that are largely interspersed),
4. Alpha plus beta (for those in which α-helices and β-strands are largely segregated)
5. Multi-domain (for those with domains of different fold and for which no homologues are known at present).

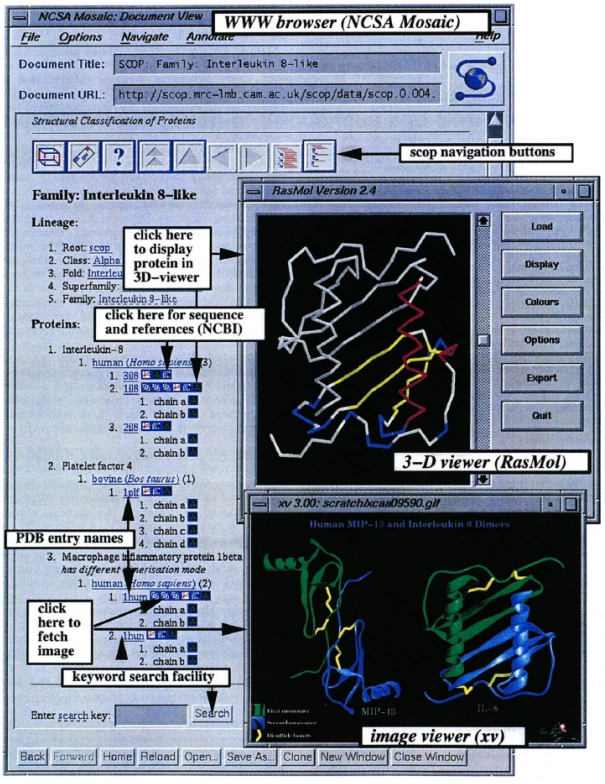
SCOP



JMB (1995) ***247****, 536-540*

SCOP - hierarchy





SCOP

SCOP – new developments

- ***Species***, representing a distinct protein sequence and its naturally occurring or artificially created variants;
- ***Protein***, grouping together similar sequences of essentially the same functions that either originate from different biological species or represent different isoforms within the same organism;
- ***Family***, containing proteins with related sequences but typically distinct functions;
- ***Superfamily***, bridging together protein families with common functional and structural features inferred to be from a common evolutionary ancestor.
- ***Folds***, structurally similar superfamilies with different characteristic features;
- ***Class***, mainly on their secondary structure content and organization.

SCOP – Class

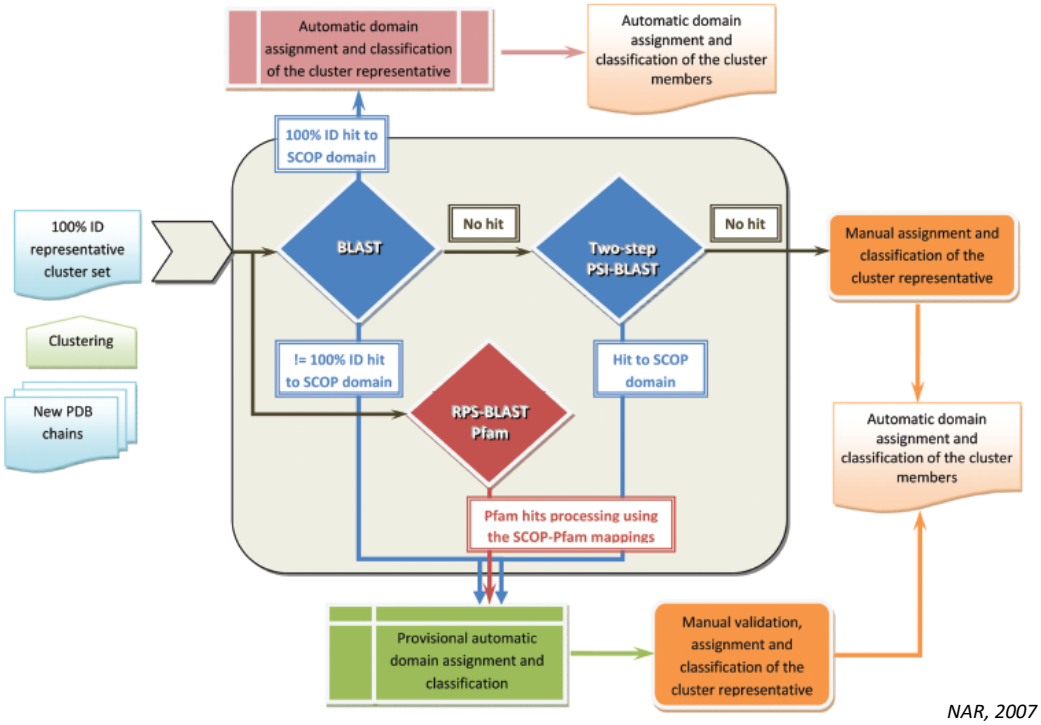
1. All alpha proteins
2. All beta proteins
3. Alpha and beta proteins (a/b) (Mainly parallel beta sheets (beta-alpha-beta units))
4. Alpha and beta proteins (a+b) (Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) (Folds consisting of two or more domains belonging to different classes)
6. Membrane and cell surface proteins and peptides (Does not include proteins in the immune system)
7. Small proteins (Usually dominated by metal ligand, heme, and/or disulfide bridges)
8. Coiled coil proteins
9. Low resolution protein structures
10. Peptides
11. Designed proteins

SCOP – new developments

- The seven main classes in the release 1.73 contain
  - 92,927 domains organized into
  - 3,464 families,
  - 1,777 superfamilies and
  - 1086 folds.

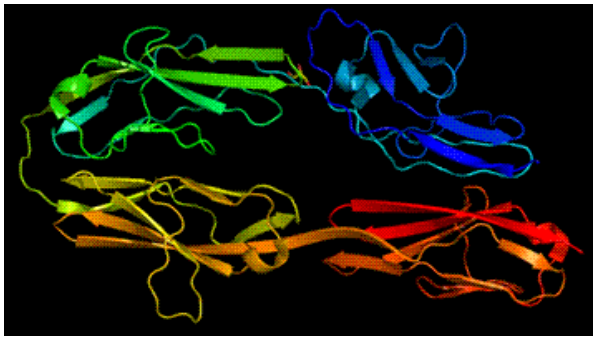
The SCOP domains correspond to 34,495 entries in the Protein Data Bank (PDB).

SCOP – current semi automatic update protocol

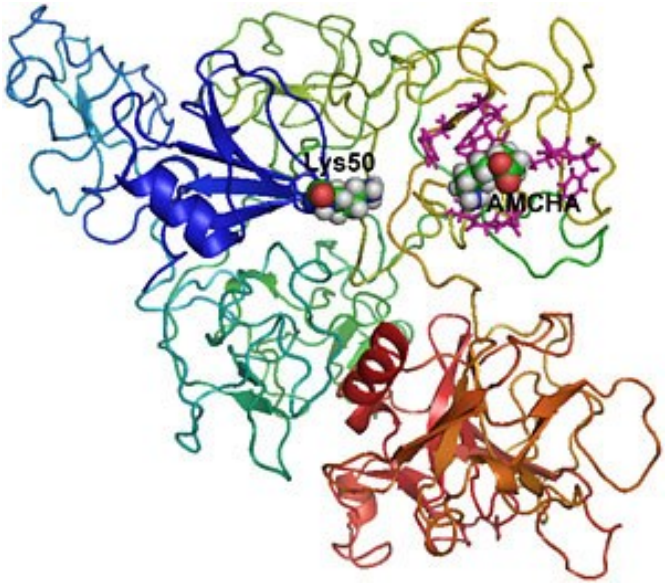


NAR, 2007

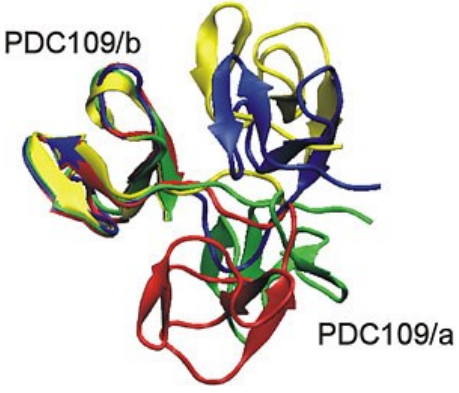
Multi domain proteins

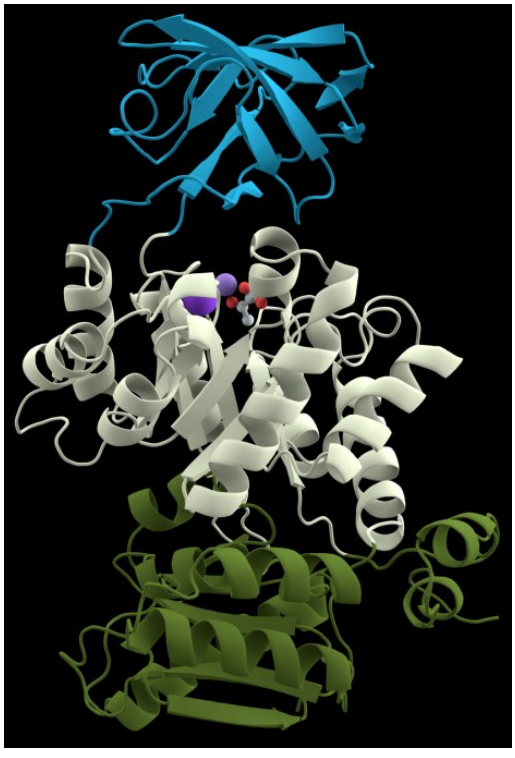


Multi domain proteins – identify the domains



Multi domain proteins – identify the domains

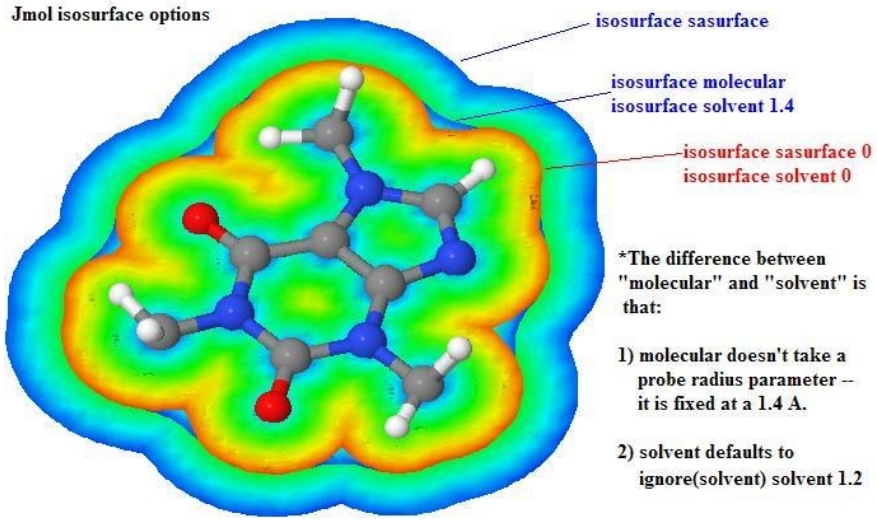




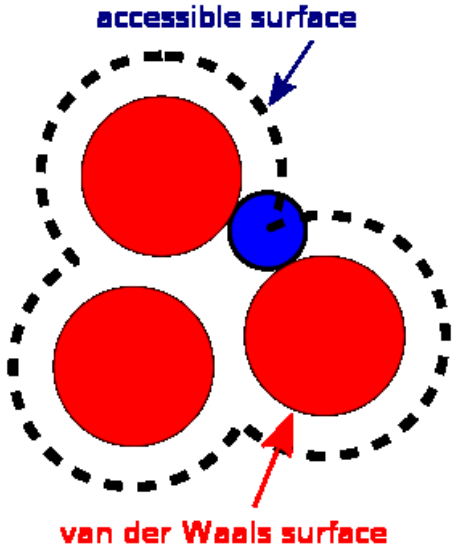
Lecture 22

Molecular Surface

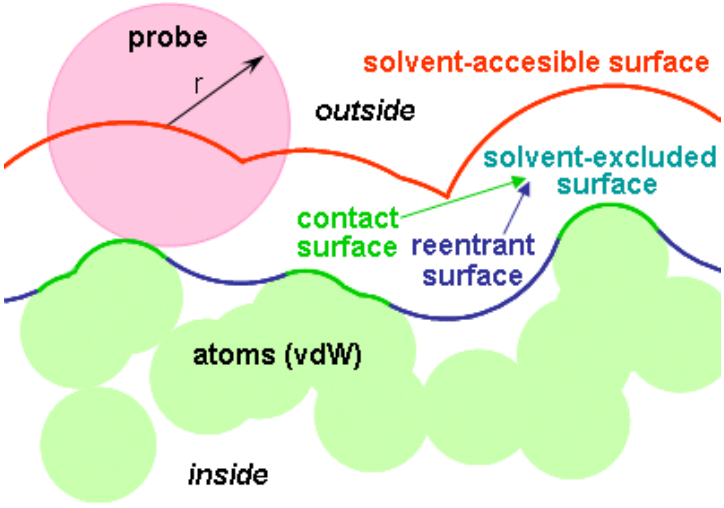
Molecular Surface



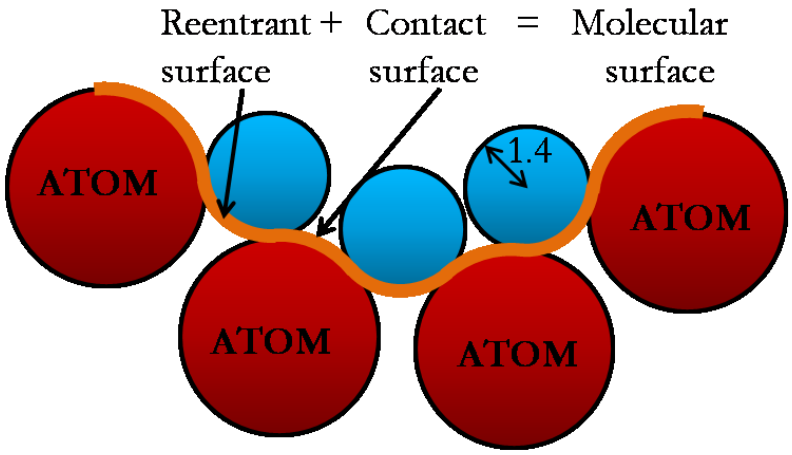
Molecular Surface



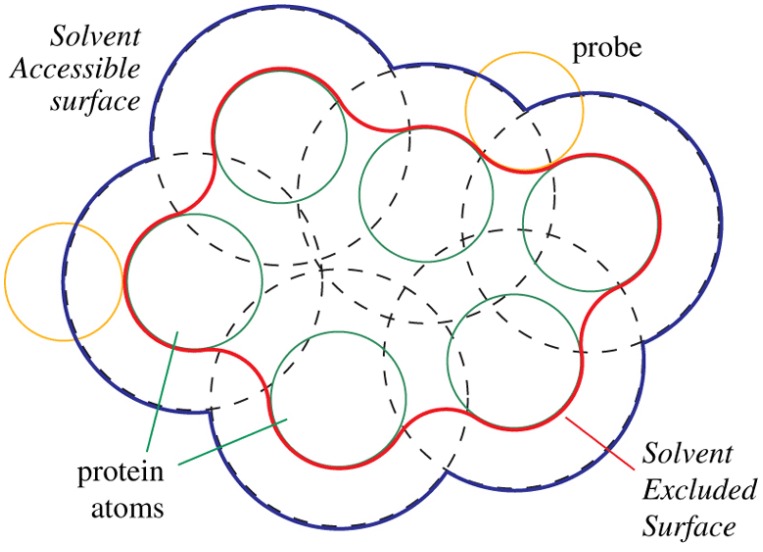
Molecular Surface



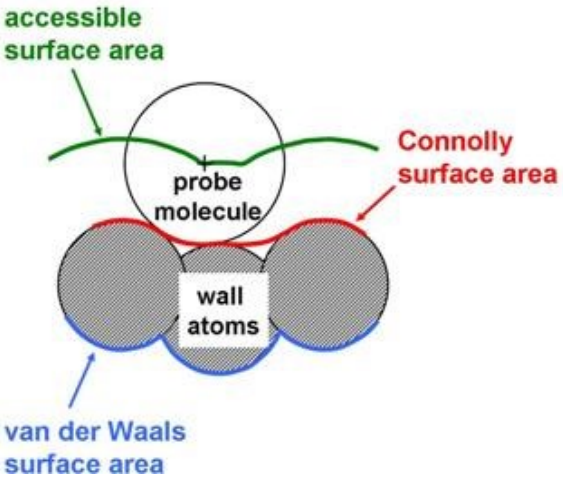
Molecular Surface



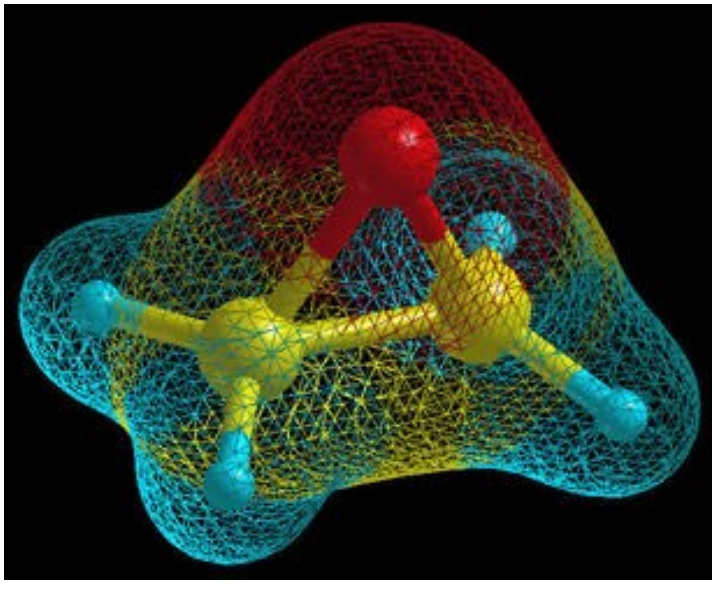
Molecular Surface



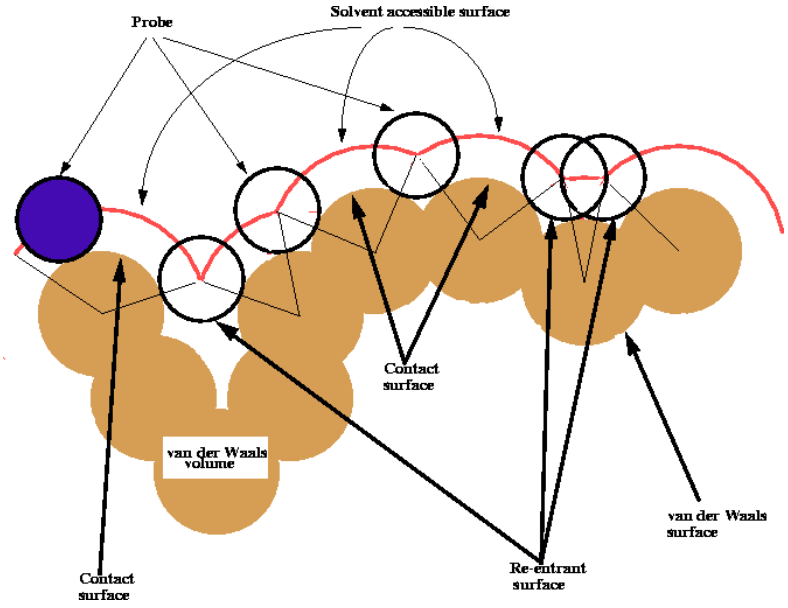
Molecular Surface



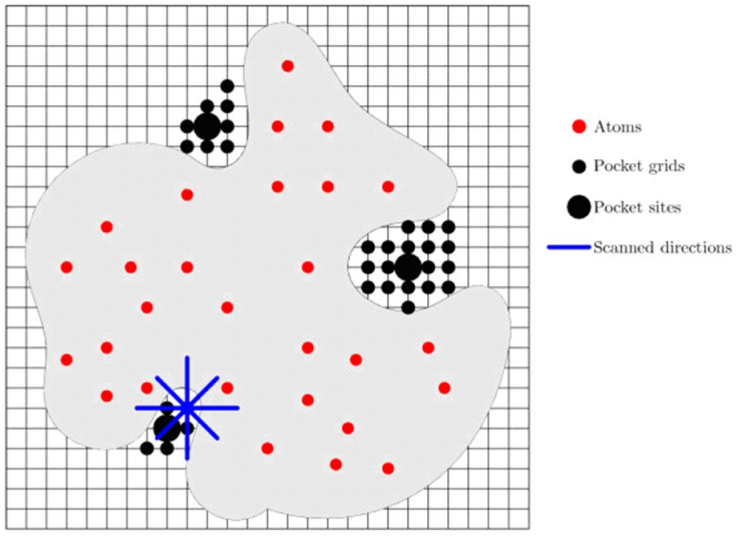
Molecular Surface



Molecular Surface



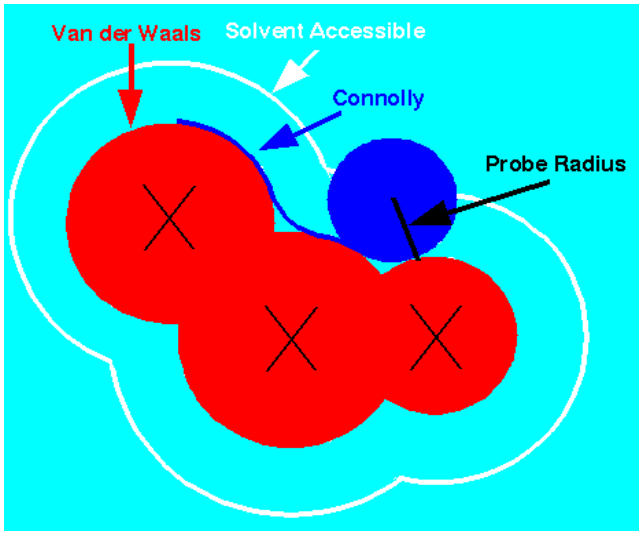
Molecular Surface



Mathematical Models of Surface Representation

- Connolly surface
- NACCESS
- Surface Normal and Critical Points
- Grid Representation
- Advantage
  - Geometric Hashing based rigid-body programs

Solvent Accessible Surface – SAS Connolly Surface



Connolly's MS algorithm

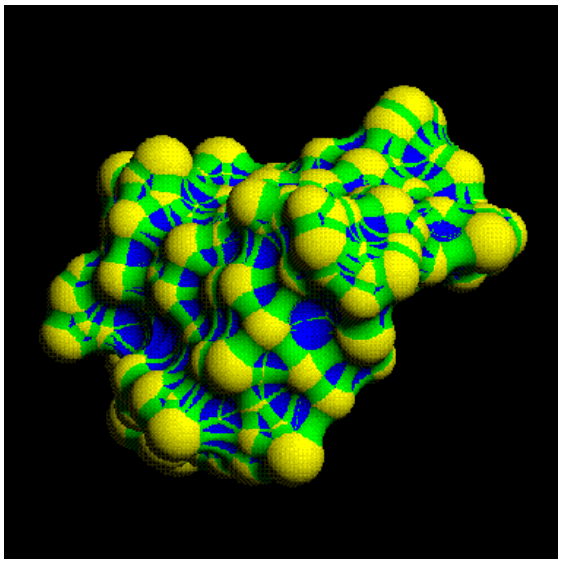
- A 'water' probe ball (1.4-1.8 A diameter) is rolled over the van der Waals surface.
- Smoothens the surface and bridges narrow 'inaccessible' crevices.

Connolly's MS algorithm - cont.

- Convex, concave and saddle patches according to the no. of contact points between the surface atoms and the probe ball.

⌗ Outputs points+normals according to the required sampling density (e.g. 10 pts/A^2).

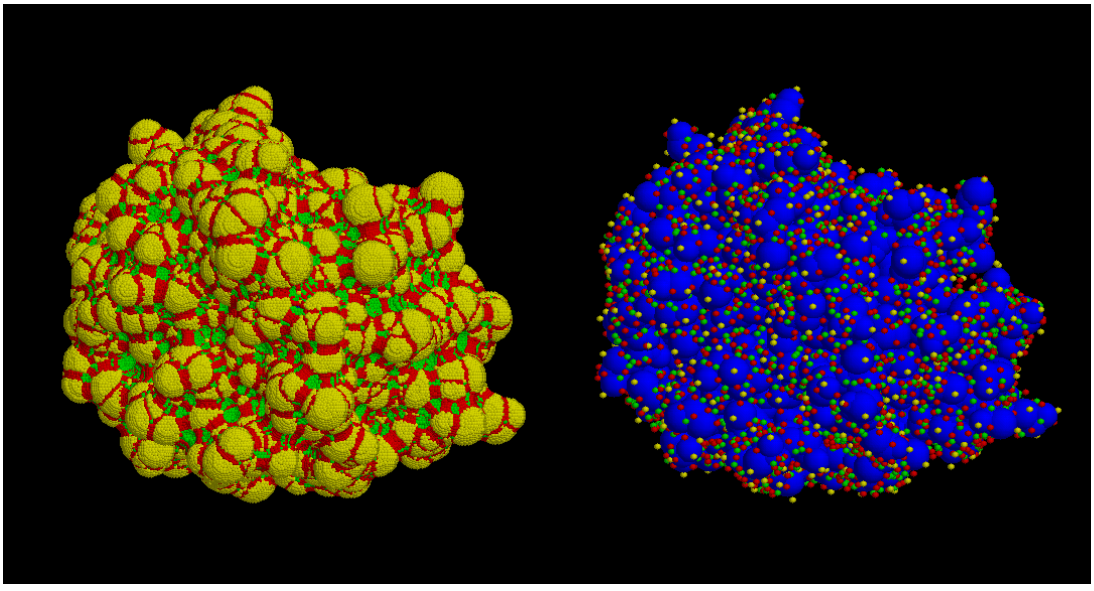
Example - the surface of crambin



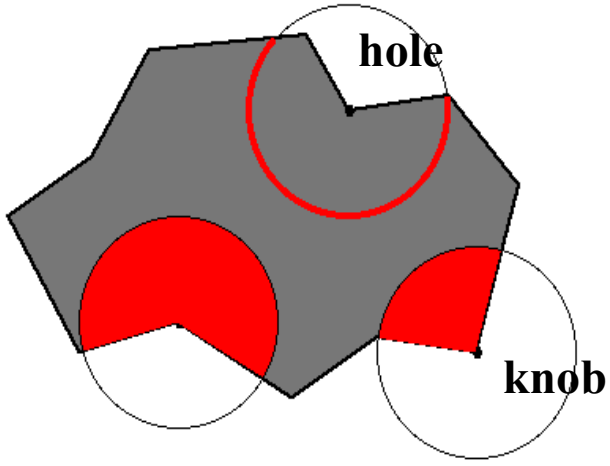
Critical points based on Connolly rep. (Lin, Wolfson, Nussinov)

- Define a single point+normal for each patch.
- Convex-caps, concave-pits, saddle - belt.

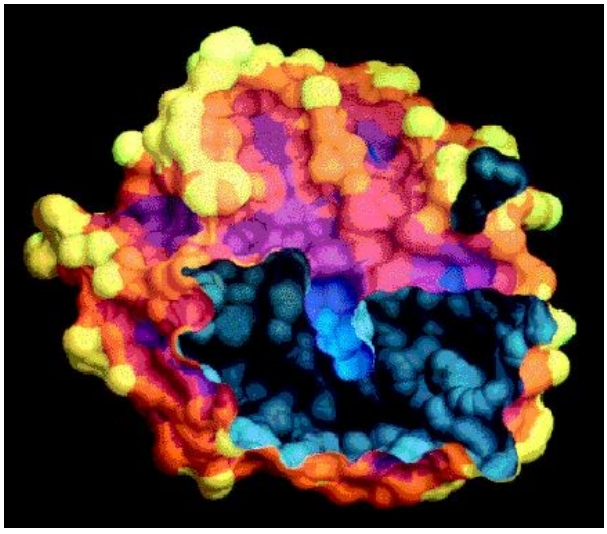
Connolly => Shou Lin



Solid Angle local extrema



Chymotrypsin surface colored by solid angle (yellow-convex, blue-concave)



Definition of Protein Surface

- Roll the probe over the protein molecule using NACCESS program to compute the accessible surface area (ASA) (also called as SASA – surface accessible surface area)

- If an atom's accessible surface area (ASA) is more that 0.1 $Å^2$ then that is define as surface atom.

Definition of Protein Interface Area

- The surface area where two protein molecules interacts during complex formation is called as the interface region.
- If an atom loses its ASA by more than 0.1 $Å^2$ then we call that as an interface atom. The total loss of ASA by all the interface atoms are the interface area.
- Sometimes, we take the average for interface area to call it as interface area per protein molecule.
- How to compute identify interface atoms? How to compute interface area?
  - Run the NACCESS on individual protein molecules
  - Run the NACCESS on the protein complex
  - Identify the atoms that loses its ASA by more than 0.1 $Å^2$

Definition of Protein Interface Area