In the Name of God, the Merciful, the Compassionate

Introduction to Bioinformatics
12 - Protein Structure Visualization, Comparison, and Classification

Instructor: Hossein Zeinali

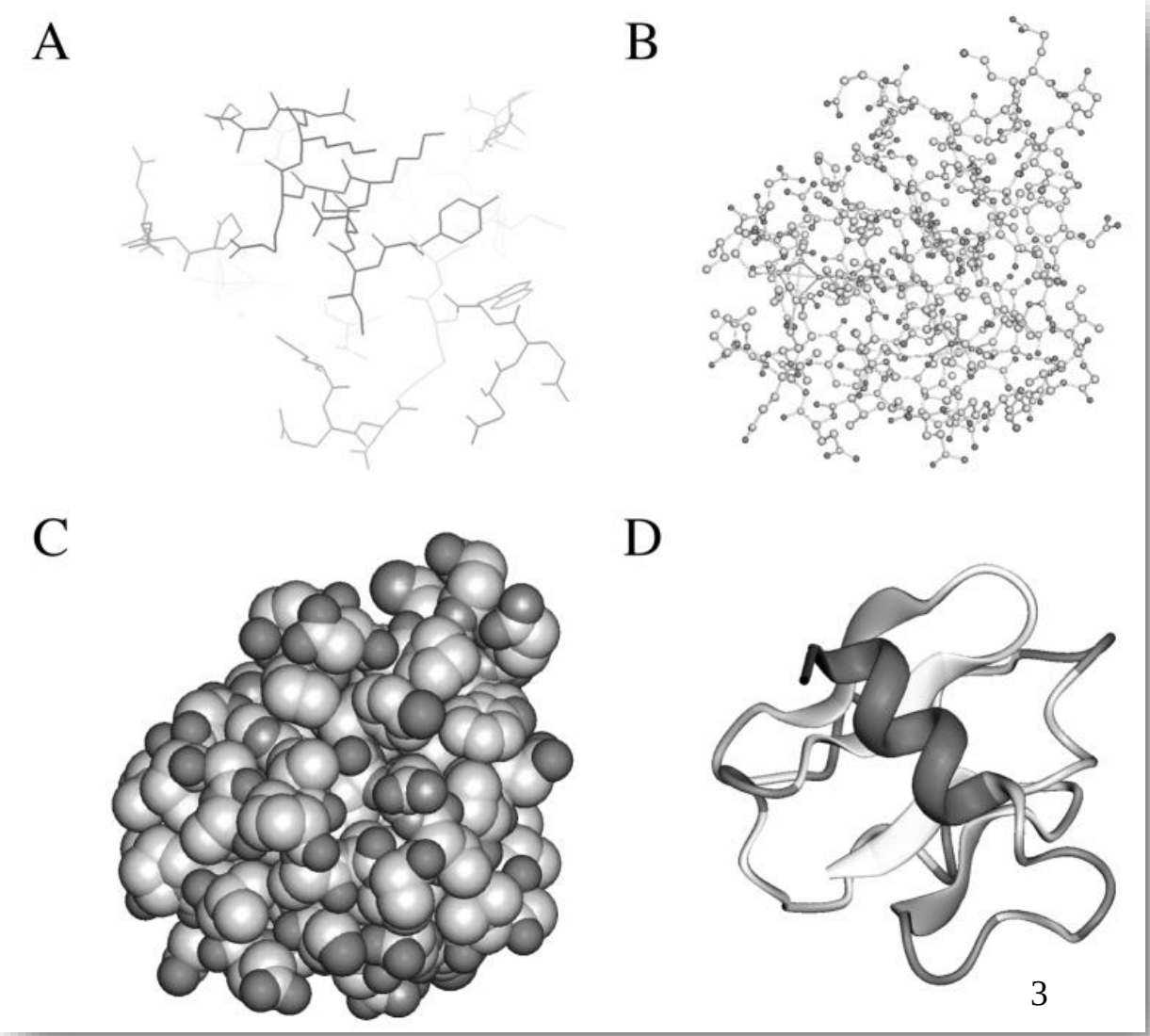
Amirkabir University of Technology

Protein Structure Visualization

- RASMOL & decendents: PyMol, MolMol
  - http://www.umass.edu/microbio/rasmol/index2.htm
- Cn3D - esp. good for structural alignments
  - http://www.biosino.org/mirror/www.ncbi.nlm.nih.gov/Structure/cn3d/
- CHIME (Protein Explorer)
  - http://www.umass.edu/microbio/chime/getchime.htm
- MolviZ.Org
  - http://www.umass.edu/microbio/chime
- Deep View = Swiss-PDB Viewer
  - http://www.expasy.org/spdbv

Molecular Structure Visualization Forms

- **(A)** Wireframes
  - A line drawing representing bonds between atoms
- **(B)** Balls and sticks
  - Solid spheres and rods, representing atoms and bonds
- **(C)** Space-filling spheres
  - each atom is described using large solid spheres
- **(D)** Ribbons
  - use cylinders or spiral ribbons to represent *α-helices* and broad, flat arrows to represent *β-strands*



PyMol

PyMOL Home Page

http://pymol.sourceforge.net/

GDCB Webmail ISU BCB Dobbs Lab BCB 444/544 Schedule of Classes Entrez PSB 2008 Conferences Seminars FadiWiki PPIDB Google Image

Download Documentation Mailing List Community Wiki F.A.Q. Plugins Links Project

PyMOL Home Page

PyMOL v1.0 Released!

PyMOL Runs On:

Linux (or any Unix)

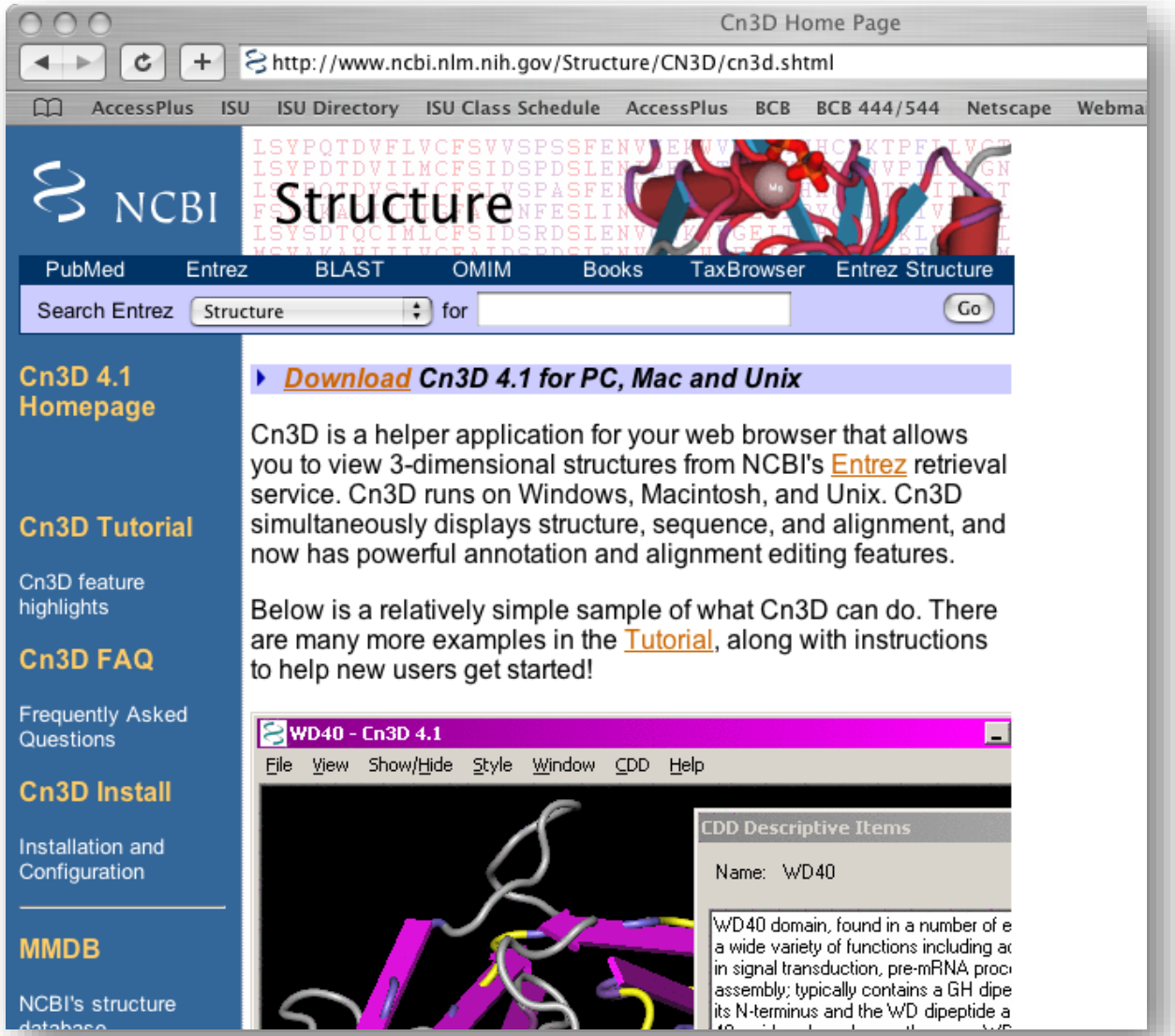
MacOS and Windows

PyMOL is a USER-SPONSORED molecular visualization system on an OPEN-SOURCE foundation. *Please support our mission to create open, effective, and affordable tools for research by* ***purchasing a subscription*** *to maintenance and/or support. Thanks!*

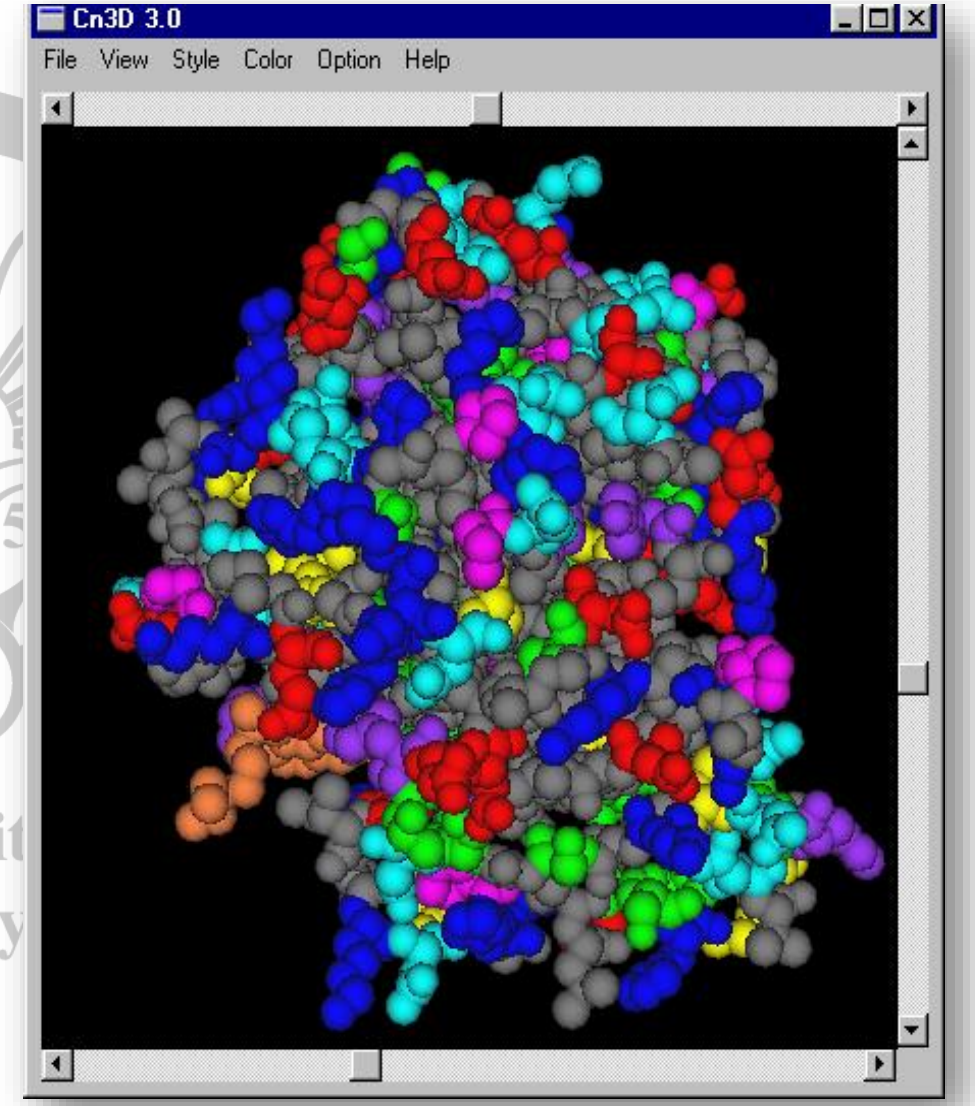
Latest News

- **October 4th, 2007: DeLano Scientific has moved!** We have relocated from temporary space into our new office at 540 University Ave., Suite 325, Palo Alto, CA 94301. Although this should be a long-term home, for certain delivery, please continue to send mail to P.O. Box 1118, Palo Alto, CA, 94302-1118, USA.

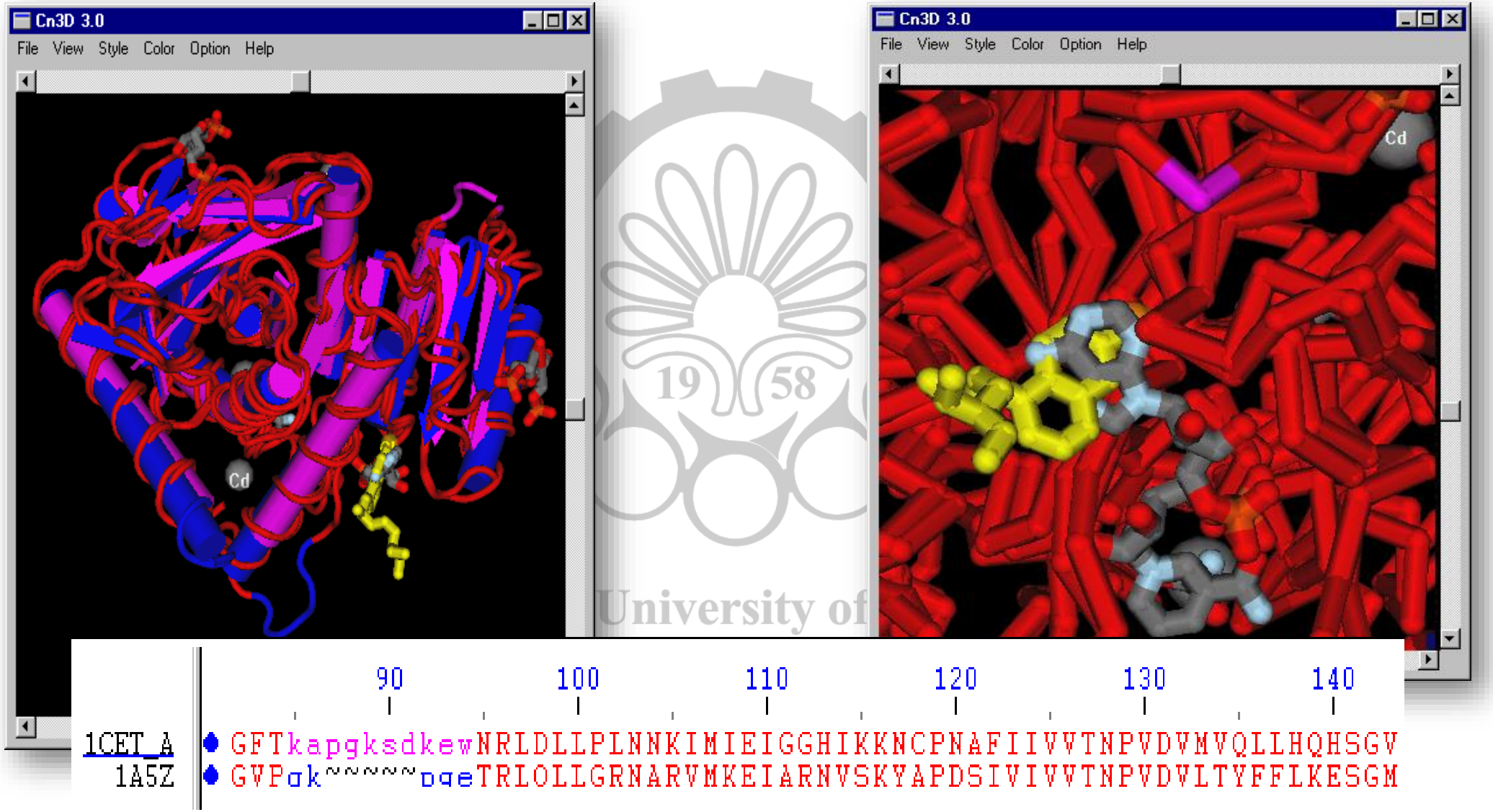
Cn3D


Cn3D Home Page
http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
AccessPlus ISU ISU Directory ISU Class Schedule AccessPlus BCB BCB 444/544 Netscape Webma
NCBI
Structure
PubMed Entrez BLAST OMIM Books TaxBrowser Entrez Structure
Search Entrez Structure for Go
Cn3D 4.1 Homepage
Download Cn3D 4.1 for PC, Mac and Unix
Cn3D is a helper application for your web browser that allows you to view 3-dimensional structures from NCBI's Entrez retrieval service. Cn3D runs on Windows, Macintosh, and Unix. Cn3D simultaneously displays structure, sequence, and alignment, and now has powerful annotation and alignment editing features.
Cn3D Tutorial
Cn3D feature highlights
Below is a relatively simple sample of what Cn3D can do. There are many more examples in the Tutorial, along with instructions to help new users get started!
Cn3D FAQ
Frequently Asked Questions
Cn3D Install
Installation and Configuration
MMDB
NCBI's structure
WD40 - Cn3D 4.1
File View Show/Hide Style Window CDD Help
CDD Descriptive Items
Name: WD40

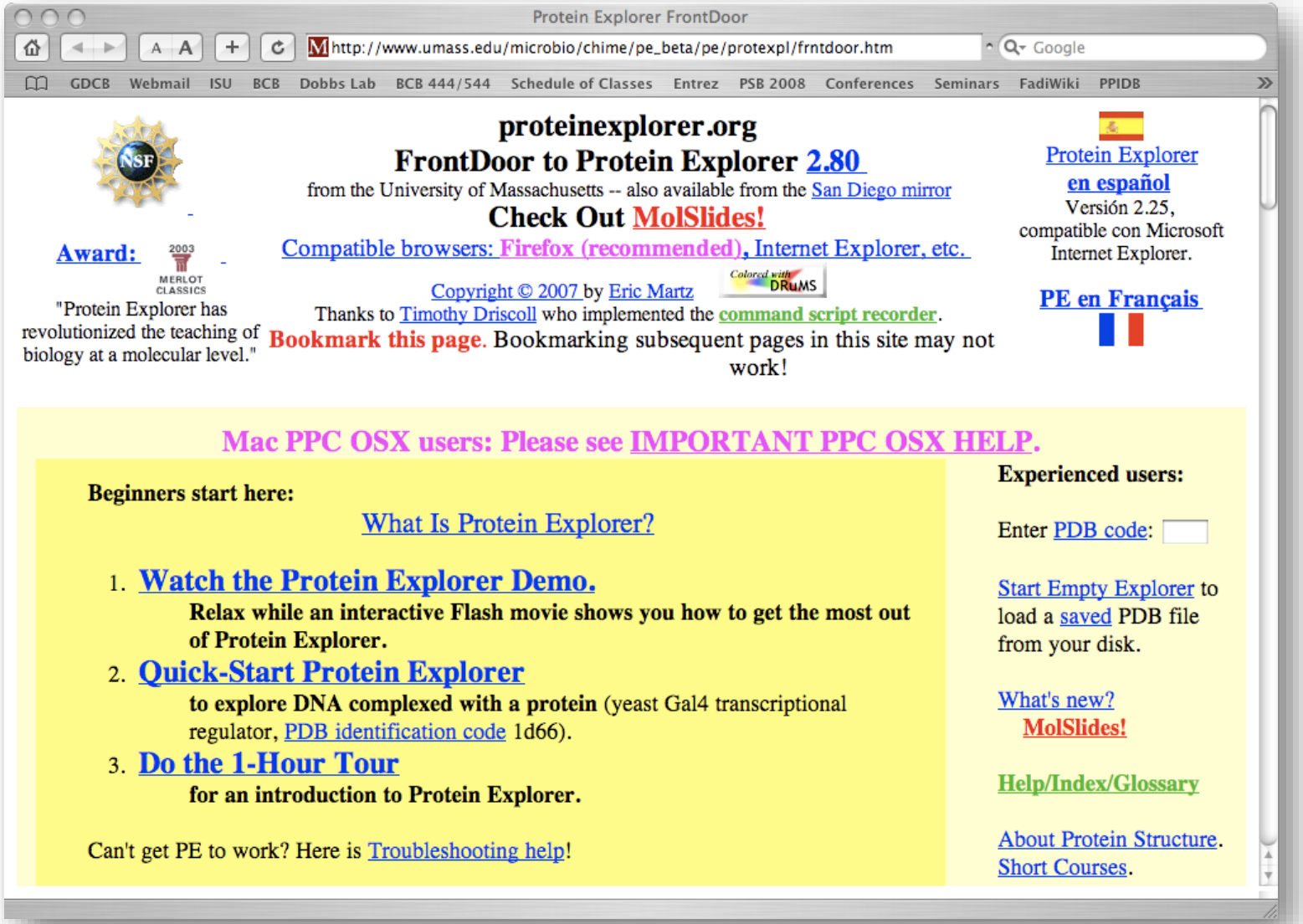
Cn3D : Displaying 3' Structures



Cn3D: Structural Alignments



Protein Explorer (Chime)



Protein Structure Comparison

- The comparative analysis involves the *direct alignment* and *superimposition of structures* in a three-dimensional space.
- Structure comparison is one of the fundamental techniques in protein structure analysis.
  - The comparative approach is important in finding remote protein homologs.
- Proteins can share common structures even without sequence similarity
  - Structures have a much higher degree of conservation than the sequences
- Structure comparison can often reveal distant evolutionary relationships between proteins
- Protein structure comparison is a prerequisite for protein structural classification

Protein Structure Comparison Methods

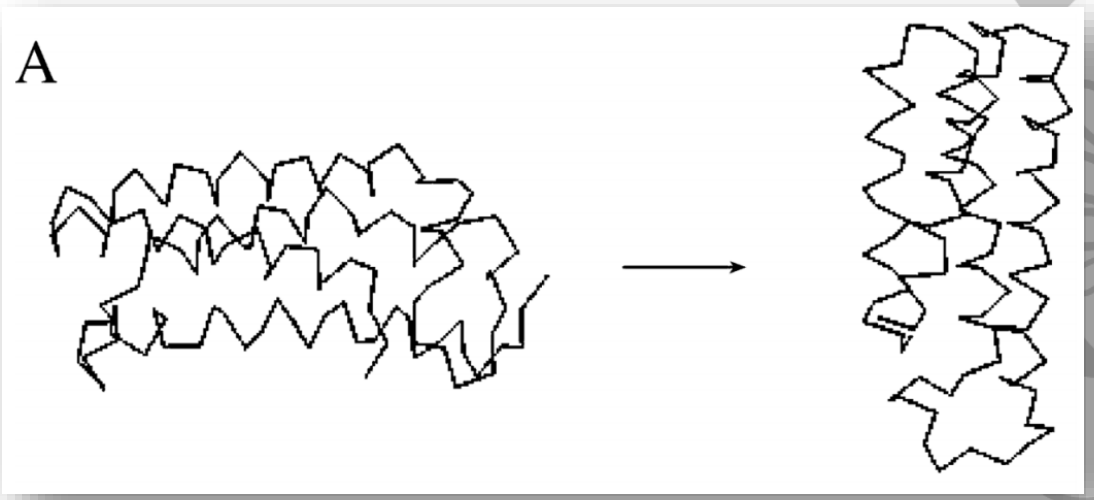
- Approaches to comparing protein geometric properties can be divided into three categories
  - Minimizing intermolecular distances
  - Measuring intramolecular distances
  - Combine both intermolecular and intramolecular

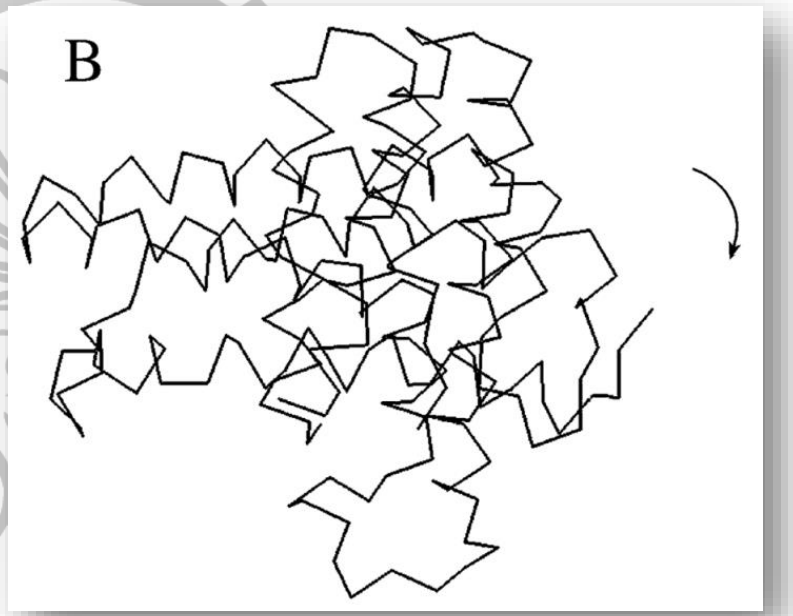
- DALI/FSSP (most commonly used)
  - Fully automated structure alignments using intramolecular distance method
  - DALI server: http://www.ebi.ac.uk/dali/index.html
  - DALI Database (fold classification): http://ekhidna.biocenter.helsinki.fi/dali/start

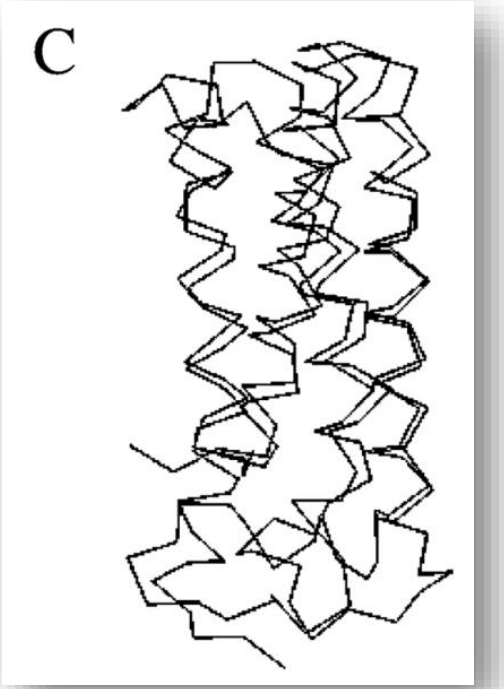
Intermolecular Method

- Normally applied to relatively similar structures
- One of the structures has to be moved with respect to the other in such a way that the two structures have a maximum overlap.
- Procedure:
  - Identifying equivalent residues or atoms by sequence based alignment
  - *Translation* : one of the structures is moved laterally and vertically toward the other structure
  - The structures are further rotated relative to each other around the three-dimensional axes
    - The distances between equivalent positions are constantly measured
  - The rotation continues until the shortest intermolecular distance is reached: *optimal superimposition*
    - Equivalent residue pairs can be identified

Intermolecular Method







Root Mean Square Deviation (RMSD)

- An important measurement of the structure fit during superposition is the distance between equivalent positions on the protein structures.
- *Root mean square deviation*:

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

where D is the *distance between coordinate data points* and N is the total number of corresponding residue pairs.

- In practice, only the distances between Cα carbons of corresponding residues are measured.

Root Mean Square Deviation (RMSD)

- The goal of structural comparison is to achieve a minimum RMSD:
  - The problem with RMSD is that it depends on the size of the proteins being compared
  - For the same degree of sequence identity, large proteins tend to have higher RMSD values
- Correct this size-dependency by logarithmic factor:

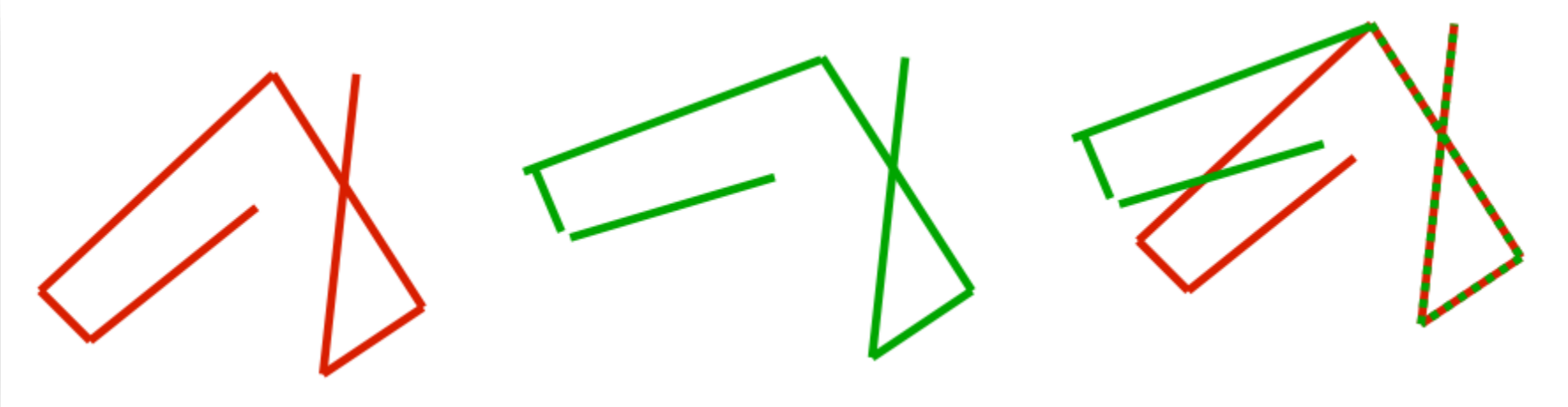
$$RMSD_{100} = \frac{RMSD}{-1.3 + 0.5\ln(N)}$$

Intermolecular Method (Cont.)

- The most challenging part of using the intermolecular method is to identify equivalent residues in the first place.
  - Depends on sequence alignment methods
- Solutions to compare more distantly related structures:
  - Delete sequence variable regions outside secondary structure elements to reduce the search time
  - Divide the proteins into small, match similar regions is then done fragment by fragment. Finally, a joint superposition for the entire structure is performed.
  - Using *iterative optimization*

Problems with RMSD

- A small local alignment error can propagate and the quality of alignment may be underestimated



Intramolecular Method

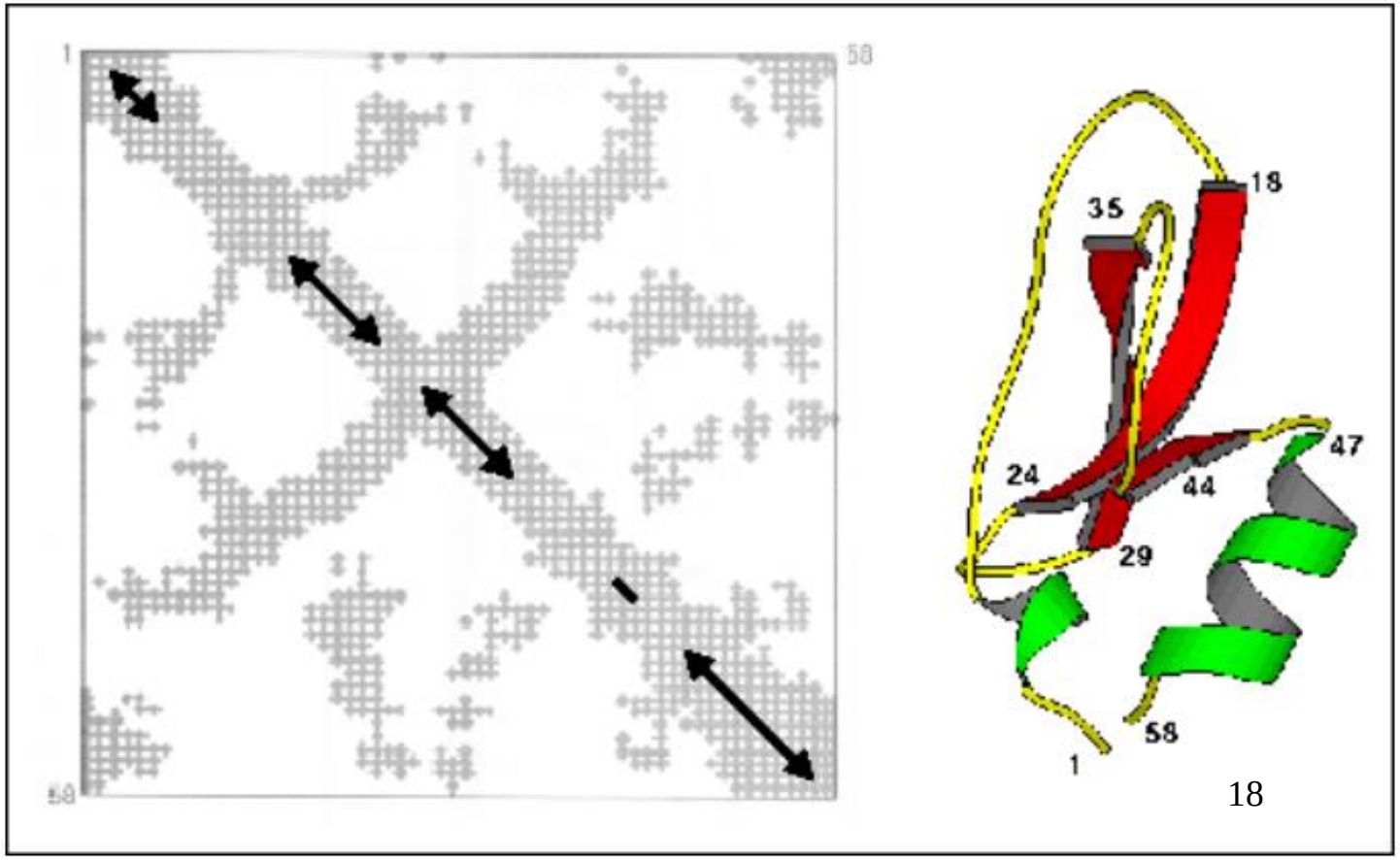
- Relies on structural internal statistics
  - Does not depend on sequence similarity between the proteins.
- Does not generate a physical superposition of structures
  - Provides a quantitative evaluation of the structural similarity between corresponding residue pairs.
- The method works by generating a distance matrix between residues of the same protein.
  - The distance matrices from the two structures are moved relative to each other to achieve maximum overlaps.
    - Similar intramolecular distance patterns representing similar structure folding regions can be identified.
- For the ease of comparison, each matrix is decomposed into smaller submatrices.

Contact matrix: the Used Distance Matrix

- Contact matrix n x n matrix where n = #residues

$$d(i,j) = distance\,(C\alpha_i, C\alpha_j)$$

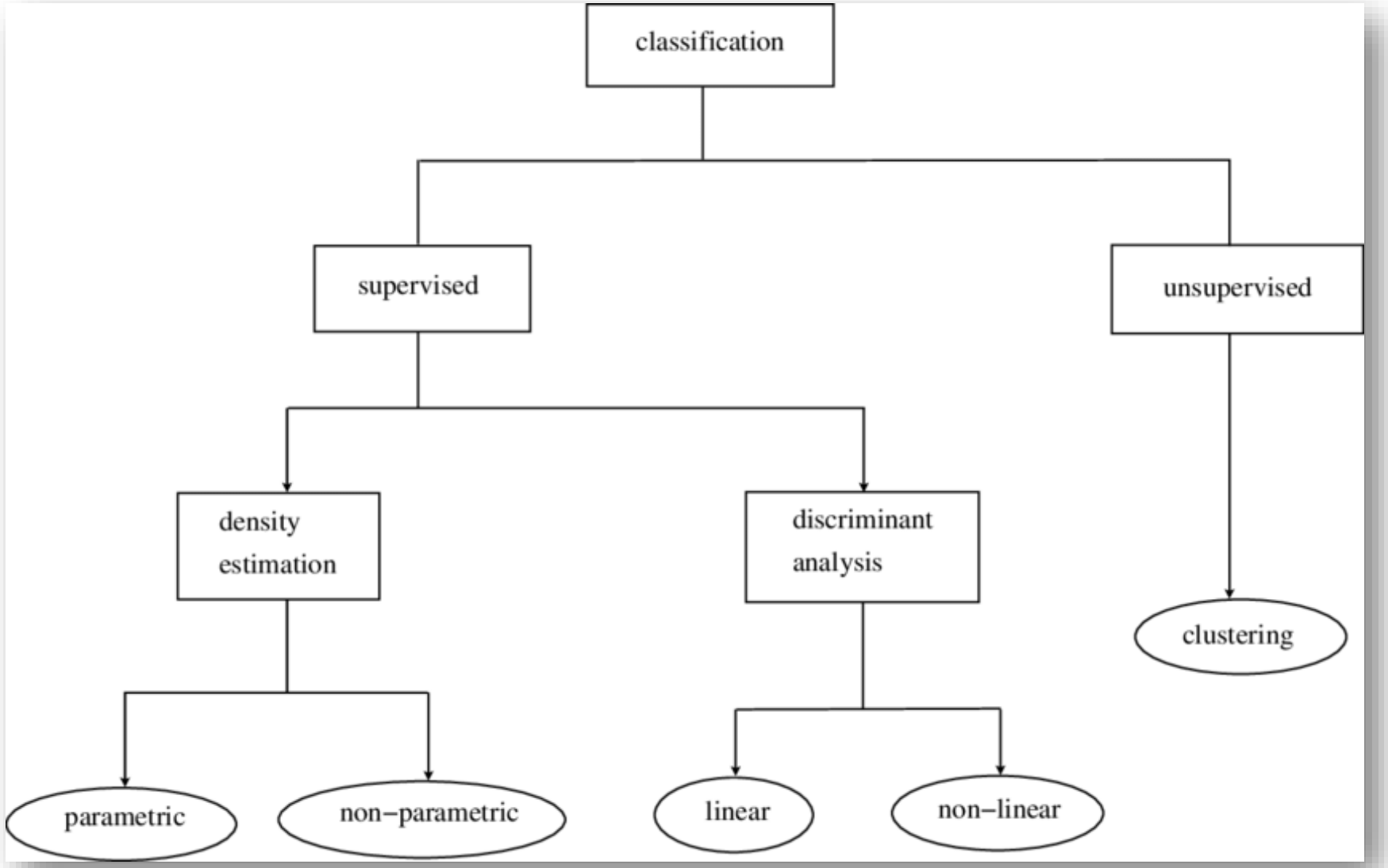
- Example: pairs with d(i,j) below a certain threshold are gray and the rest is white
- **Idea**: Similar structures have similar contact matrices



Comparison Tools

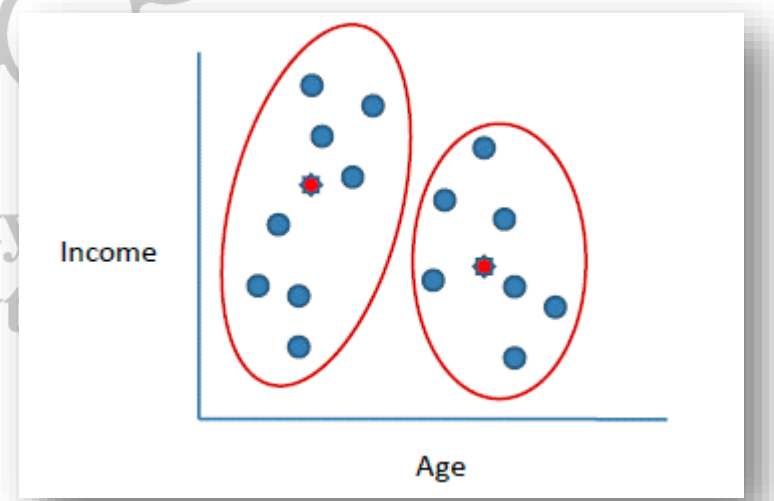
- Combinatorial Extension (CE)
  - http://cl.sdsc.edu/ce.html
  - Uses the intramolecular distance
- Vector Alignment Search Tool (VAST)
  - www.ncbi.nlm.nih.gov:80/Structure/VAST/vast.shtml
  - Uses both the inter- and intramolecular approaches
- SSAP
  - www.biochem.ucl.ac.uk/cgi-bin/cath/GetSsapRasmol.pl
  - Uses an intramolecular distance–based method
- STAMP
  - www.compbio.dundee.ac.uk/Software/Stamp/stamp.html
  - Uses the intermolecular approach by *iterative alignment*

Classification Hierarchy



K-Means Clustering

1. Clusters the data into *k* groups where *k* is predefined.
2. Select *k* points at random as cluster centers.
3. Assign objects to their closest cluster center according to the *Euclidean distance* function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.



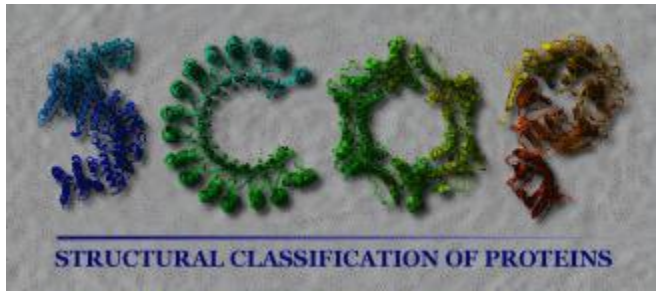
Protein Structure Classification

- One of the applications of protein structure comparison is structural classification.
- The ability to compare protein structures allows classification of the structure data and identification of relationships among structures.
- Usage: establish hierarchical relationships among protein structures and to provide a comprehensive and evolutionary view of known structures.
- Method & Databases:
  - Structural Classification of Proteins (SCOP)
  - Class, Architecture, Topology and Homologous (CATH)

Protein Structure Classification (Cont.)

- The first step in structure classification is to remove redundancy from databases.
  - The redundancy can be removed by selecting representatives through a sequence alignment–based approach.
- The second step is to separate structurally distinct domains within a structure.
  - Proteins with multiple domains must be subdivided before a sensible structural comparison can be carried out.
    - Can be done either manually or based on special algorithms for domain recognition.
  - Structure comparison can be conducted at the domain level, either through manual inspection, or automated, or a combination of both.
- The last step involves grouping proteins/domains of similar structures and clustering them based on different levels.

Structural Classification of Proteins (SCOP)

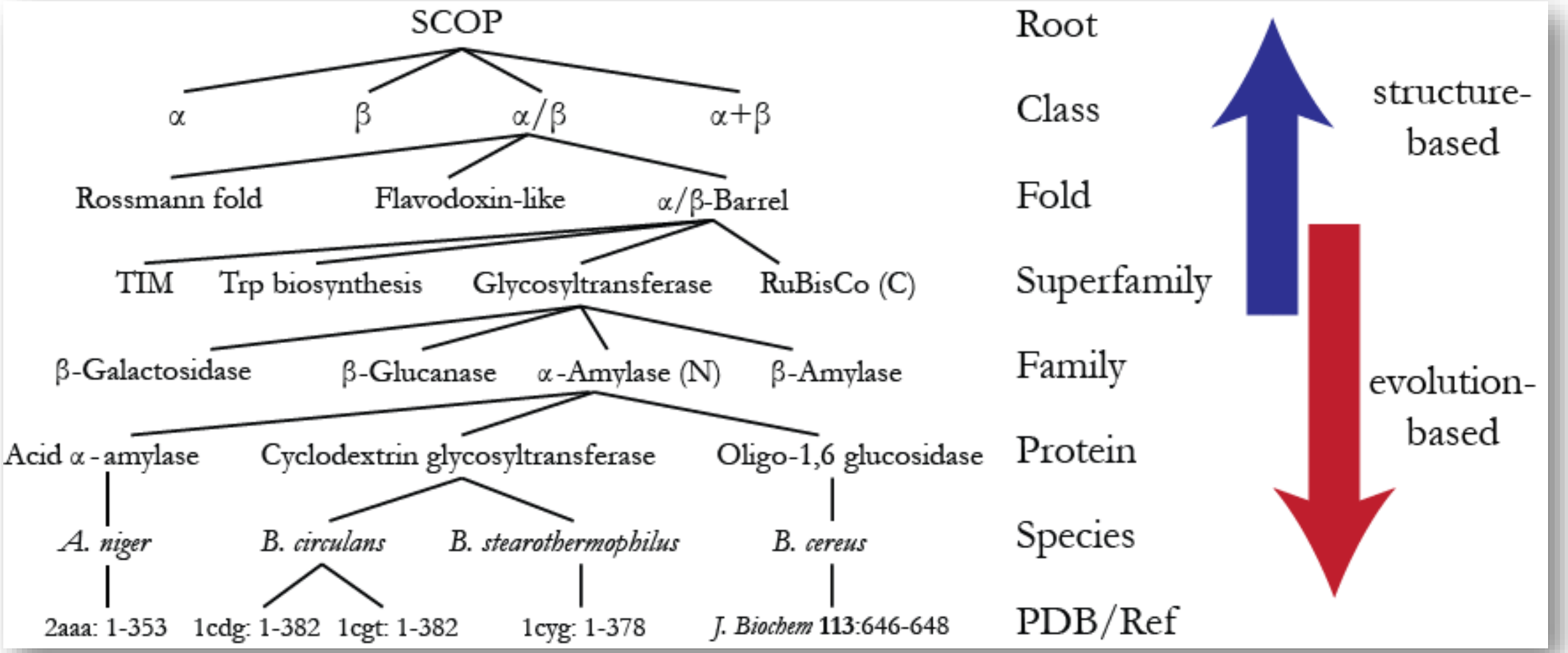


- SCOP is a database for comparing and classifying protein structures.
  - http://scop.mrc-lmb.cam.ac.uk/scop/
  - Is constructed based on manual examination of protein structures.
  - The proteins are grouped into hierarchies of classes, folds, super-families, and families.
  - Includes 44,769 non-redundant domains representing 540,282 protein structures.
- The SCOP families consist of proteins having high sequence identity (>30%).

Structural Classification of Proteins (SCOP)

- Super-families consist of families with similar structures, but weak sequence similarity.
- Folds consist of super-families with a common core structure, which is determined manually.
  - Fold's members do not always have evolutionary relationships.
- Classes consist of folds with similar core structures.
  - The highest level of the hierarchy
  - Distinguishes groups of proteins by secondary structure compositions

Structural Classification of Proteins (SCOP)

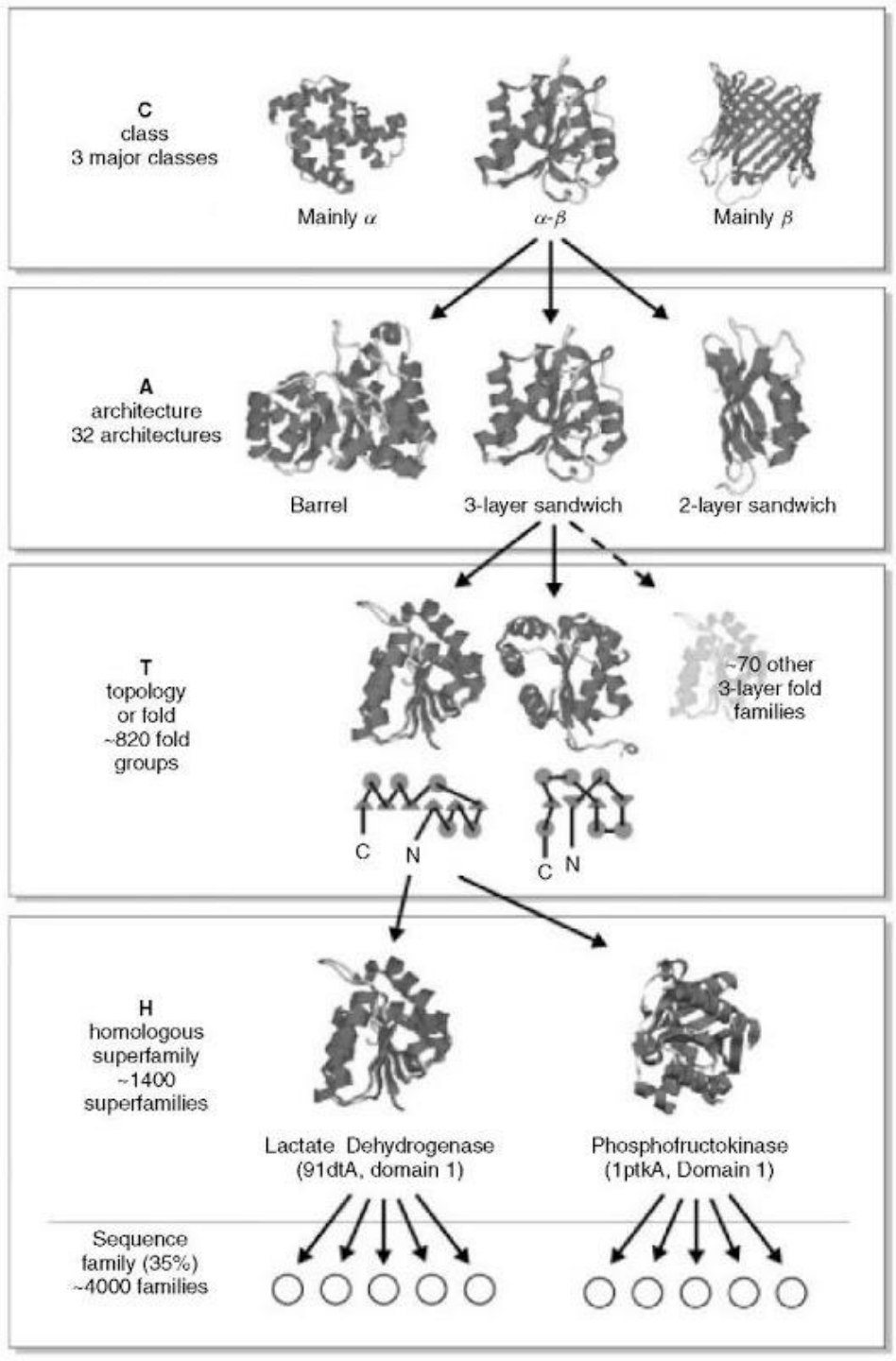


Class, Architecture, Topology and Homologous (CATH)

- CATH classifies proteins based on the automatic structural alignment program SSAP as well as manual comparison.
  - https://www.cathdb.info/
  - Structural domain separation is carried out as a combined effort of a human expert and computer programs.
- Individual domain structures are classified at five major levels: class, architecture, fold/topology, homologous superfamily, and homologous family.
- Architecture describes the overall packing and arrangement of secondary structures independent of connectivity between the elements.

Class, Architectu ologous (CATH)

- CATH classifies p structural alignment progran parison.
  - https://www.cathd
  - Structural domain bined effort of a human expert and
- Individual domain e major levels: class, architecture, uperfamily, and homologous famil
- Architecture descr rrangement of secondary structur y between the elements.



Comparison of SCOP and CATH

- SCOP is almost entirely based on manual comparison of structures by human experts
  - CATH is a combination of manual curation and automated procedure
- The classification results from both systems are quite similar.
  - The results from the two systems converge at about 80% of the time.
  - Only about 20% of the structure fold assignments are different.

References

- Mostly used:
  - Essential bioinformatics, Chapter 13 (Protein Structure Visualization, Comparison, and Classification)

- IP notice: some slides were selected from Drena Dobbs' slides.

Thanks for your attention