

Structural Bioinformatics (Part 3)

Presenter
Jose Lugo-Martinez
Lab Meeting

"The golden rule is that there are no golden rules." George Bernard Shaw

April 5, 2011





Outline

- Structure Comparison and Alignment
 - DALI
- (Structural) Domains Overview
- Protein Structure Classification DB
 - SCOP
 - CATH



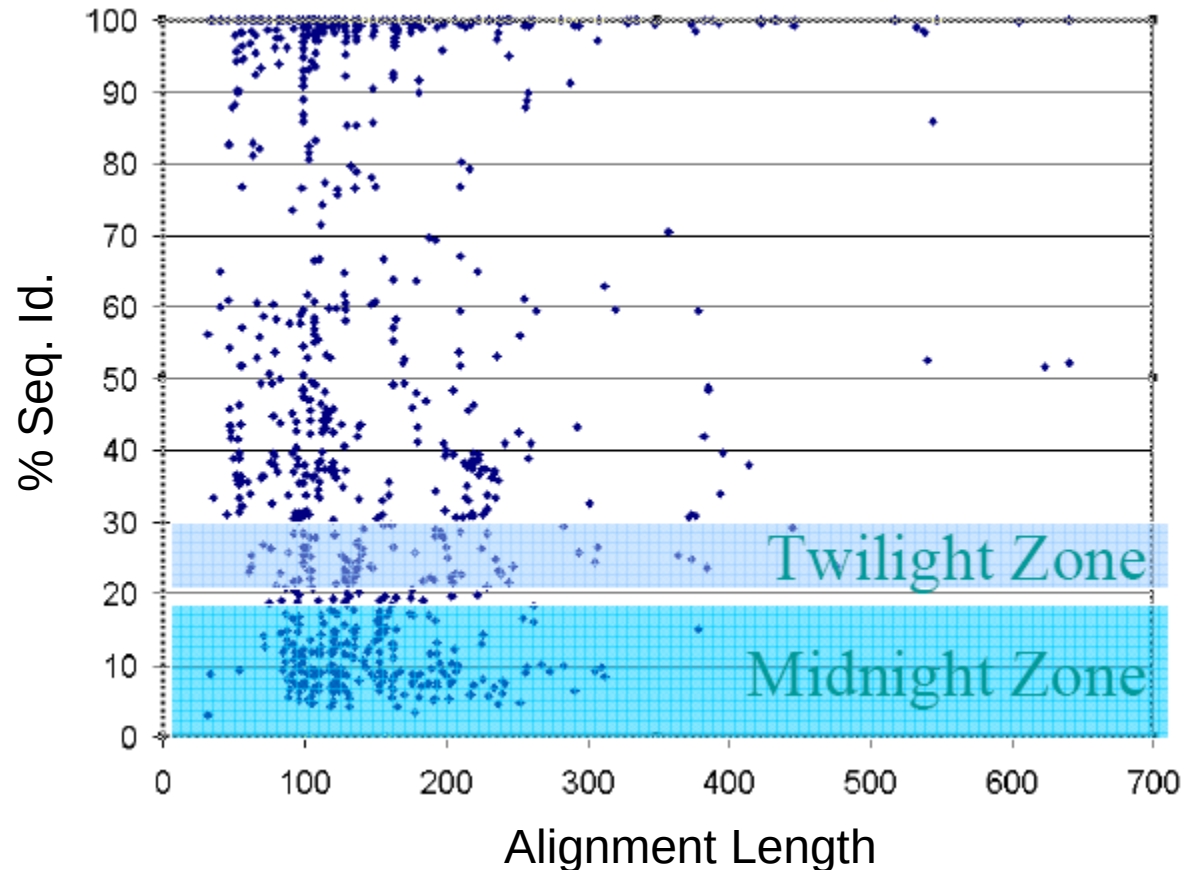
Structural Comparison and Alignment

- Analogous to sequences comparison and alignment
 - Inferring Biological Meaning by Similarity
 - Similarity implies homology
 - Transfer biological properties (e.g. function)
- **Why bother?**
 - <30% sequence identity of homologous relationship not guaranteed by sequence alone, thus, enabling improved sequence relationships
 - Structure is much more conserved than sequence
 - Humans: ~25,000 protein sequences vs 1,000-5,000 protein folds
 - Distinguishing between divergent versus convergent evolution is an issue
 - Automatically classify new protein structures
 - Functional assignment



Sequence vs Structure Similarity

Random 1000 structurally similar PDB polypeptide chains from CE with $z > 4.5$





Sequence-Structure-Function Relationships

- >40% seq. id. sequences tend to have the same structure and function (there are exceptions)
- Structure and function tend to diverge at same level of seq. id.

Similar Sequence Similar Structure Similar Function

Yes	Yes	Yes/No
Yes	No	Yes/No
No	Yes	Yes/No
No	No	Yes/No

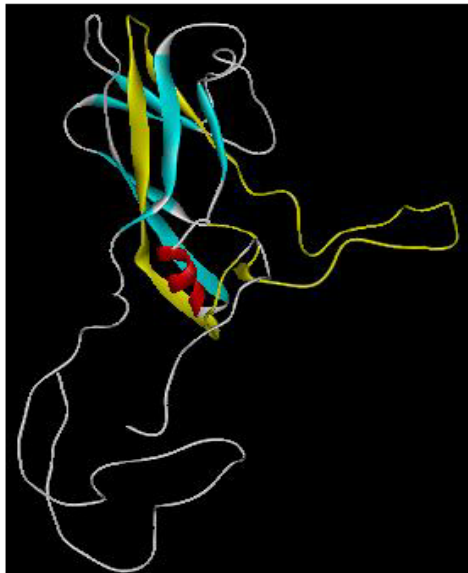


INDIANA UNIVERSITY

Similar Sequences – Different Structures

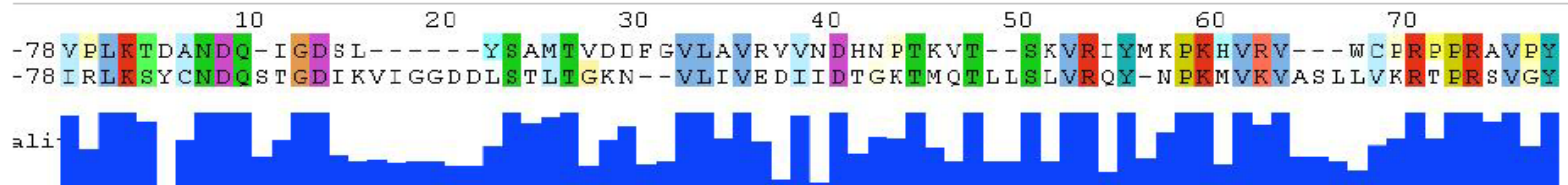
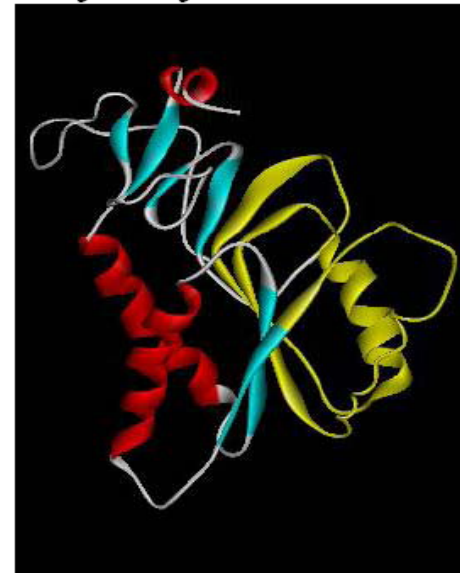
1PIV:1

Viral Capsid Protein



1HMP:A

Glycosyltransferase

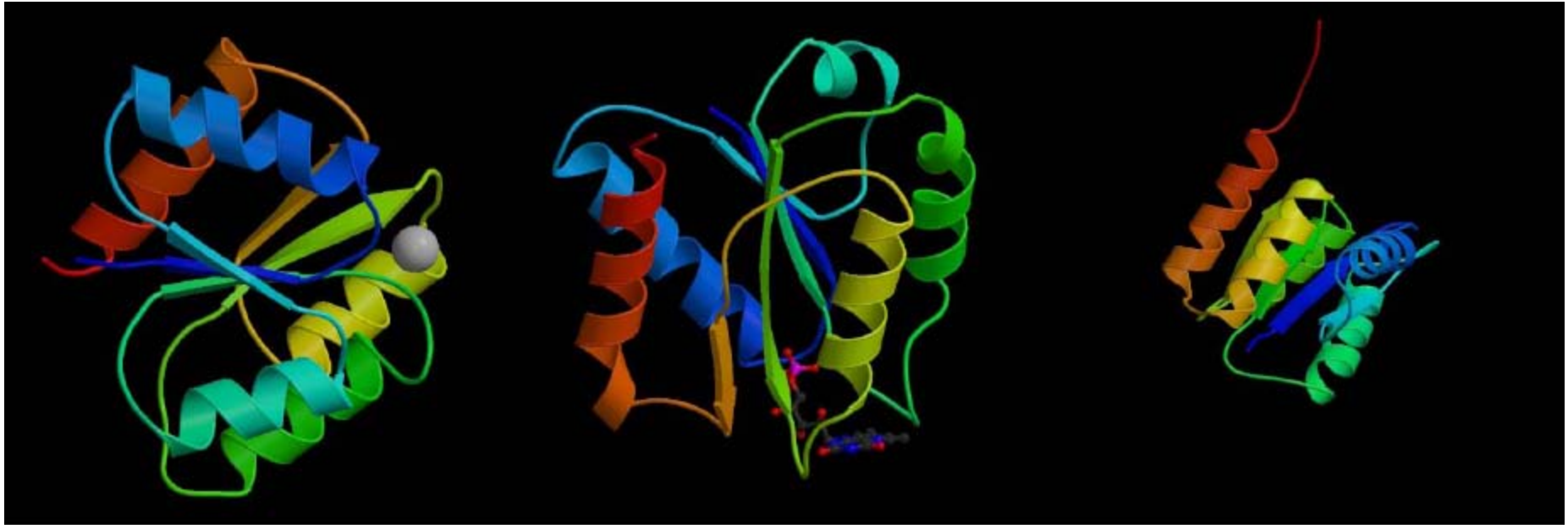


80 Residue Stretch (Yellow) with Over 40% Sequence Identity



INDIANA UNIVERSITY

Similar Structure – Different Function



1ymv

CheY

Signal Transduction

1fla

Flavodoxin

Electron Transport

1pdo

Mannose Transporter



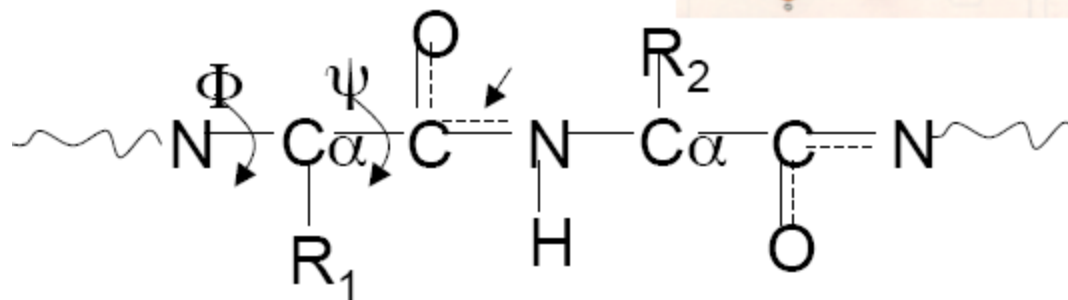
Structural Alignment

- Given two shapes (or structures) A and B
 - **Goal:** Find optimal match between shapes
 - NP-hard problem
- **Main Steps**
 - Define a measure of protein structure similarity
 - Usually involves a representation of structures in coordinate-independent space
 - Align structures based on such similarities
 - How to sample space of possible alignment solutions?
 - Define a statistical measure of similarity
 - How to score a given alignment and determine its statistical significance?



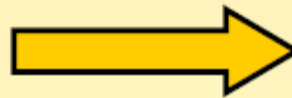
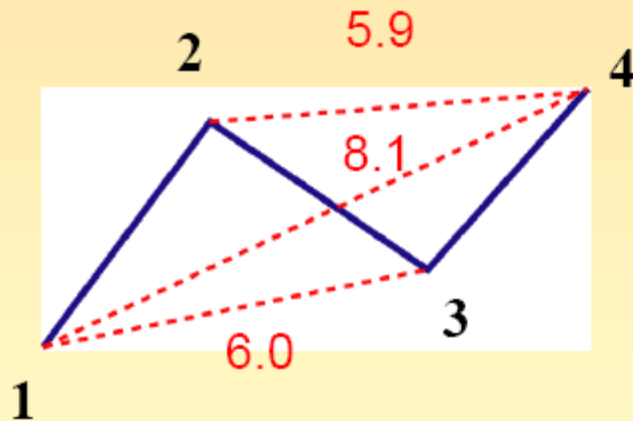
Measuring Protein Structure Similarity

- Visual comparison
- Dihedral angle comparison
 - Torsion angles (Φ, ψ)
- Distance matrix
- Root mean square distance (a.k.a. RMSD)





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

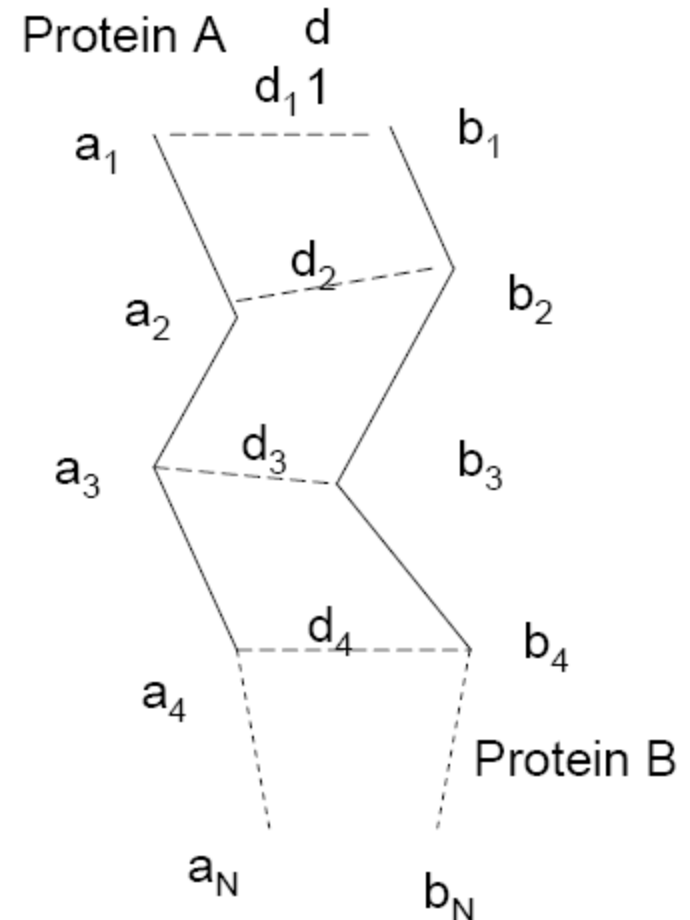
Root Mean Squared Distance (RMSD)

$$\text{RMSD} = \text{Sqrt} \left(\frac{1}{N} \sum_{i=1}^{i=N} |d_i|^2 \right)$$

1-2Å - the proteins are close

<6Å - they are likely related

Structure superposition –
placement of two or more 3D
protein structures (as rigid bodies)
in space, so that certain scoring
function (e.g. RMSD) is optimized





How to score a given alignment and determine its statistical significance?

- **Parameters/factors to consider:**
 - Baseline: **what is random structure?**
 - Distance: RMSD vs. intra-protein distances or contacts/interactions?
 - Gaps vs. unaccounted unaligned areas?
 - Similar sequence vs. similar structure?
 - Known functionally important residues vs. similarity in structure?
 - How to handle structural movements and flexible areas?



Levitt and Gerstein Approach (1998)

1. Use a resource where structure similarity is defined – SCOP.
2. Find a scoring function which allows good separation of true and false positives:

$$S_{str} = M \left(\sum \frac{1}{1 + \left(\frac{d_{ij}}{d_0} \right)^2} - \frac{N_{gap}}{2} \right)$$

d_{ij} - distance between residues i and j in alignment;

$M = 25$;

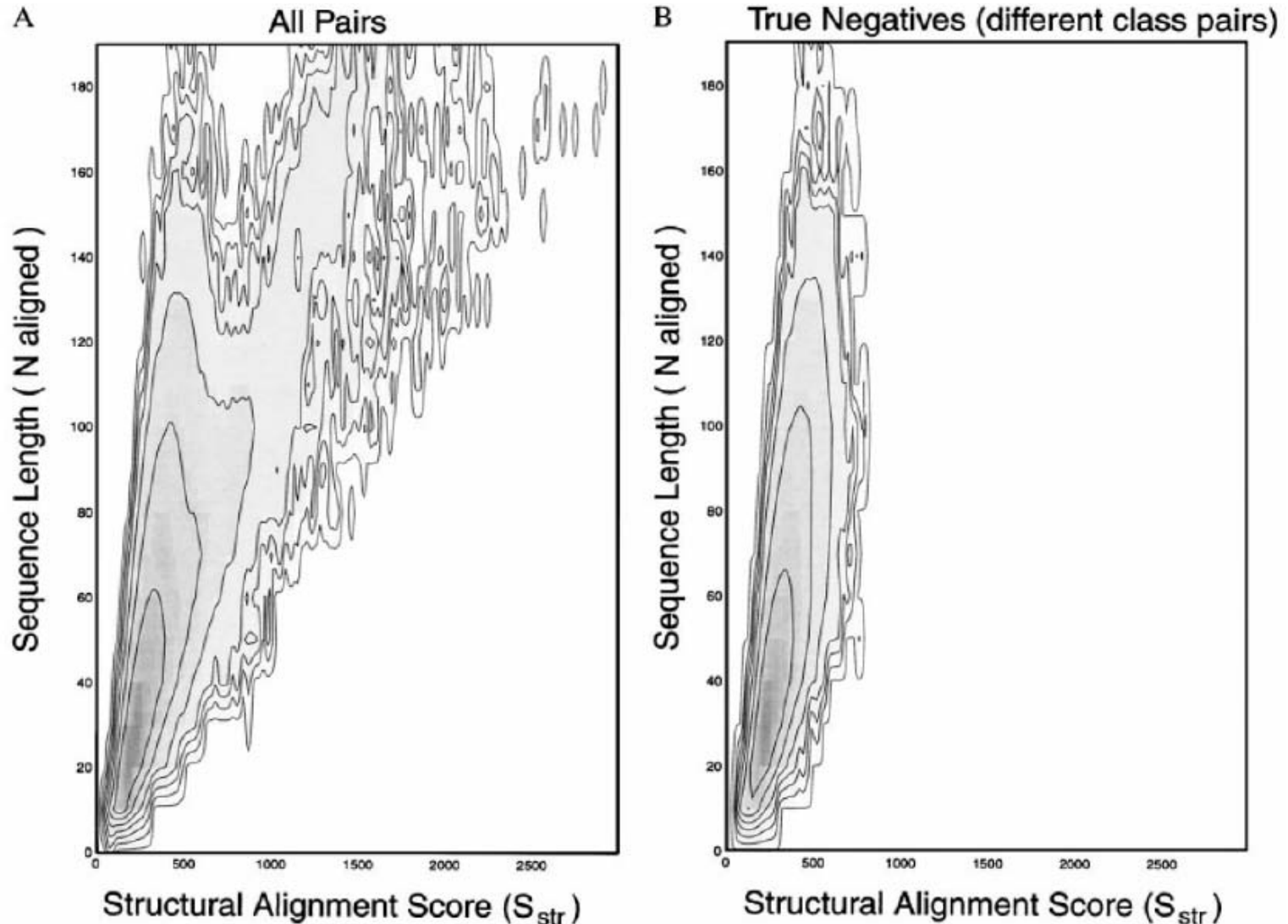
$d_0 = 5 \text{ \AA}$;

N_{gap} – total number of gaps in local alignment (i.e. not counting terminal gaps);



INDIANA UNIVERSITY

Levitt and Gerstein Approach (1998)





Some Methods for Structure Comparison and Alignment

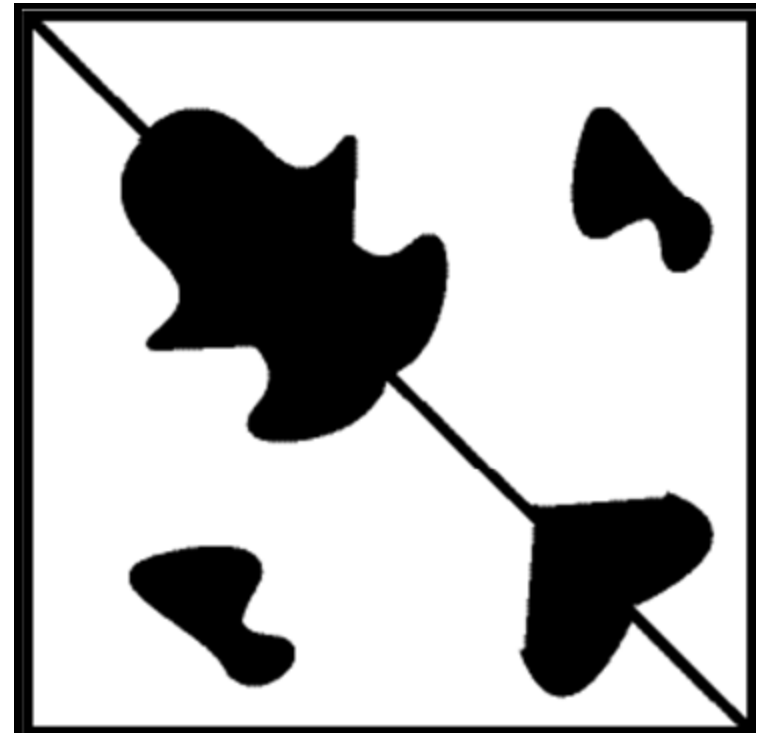
TABLE 16.1. Web Resources Associated with Methods for Structure Comparison and Alignment

Name	Description	Citations ^a	URL and Web Resource Reference
CE	<u>C</u> ombinatorial <u>E</u> xtension of the Optimum Path (Shindyalov and Bourne, 1998)	76	http://cl.sdsc.edu/ce.html Shindyalov and Bourne (2001)
→ DALI	<u>D</u> istance Matrix <u>A</u> lignment (Holm and Sander, 1993a)	890	http://www.ebi.ac.uk/dali/ Deitmann et al., (2001)
HOMSTRAD	<u>H</u> omologous <u>S</u> tructure <u>A</u> lignment <u>D</u> atabase (Mizuguchi et al., 1998)	47	http://www-cryst.bioc.cam.ac.uk/~homstrad/ Sowdhamini et al., (1998)
SARF2	<u>S</u> patial <u>A</u> rrangement of Backbone <u>F</u> ragments (Alexandrov, 1996)	66	http://123d.ncifcrf.gov/sarf2.html
SSAP	<u>S</u> equential <u>S</u> tructure <u>A</u> lignment <u>P</u> rogram (Taylor and Orengo, 1989)	248	http://www.biochem.ucl.ac.uk/~orengo/-ssap.html
VAST	<u>V</u> ector <u>A</u> lignment <u>S</u> earch <u>T</u> ool (Gibrat et al., 1996)	122	http://www.ncbi.nlm.nih.gov/Structure/-VAST/vast.shtml Wang et al., (2001)

^aCitations of the original paper reporting the method as of May 20, 2002.

DALI (Holm and Sander, 1996)

- Uses distance matrix to represent 3D structure
 - Usually distance between all their C α atoms
- Plot against itself
- Identify pairs of residues close in space to each other
 - Closeness of C α atoms (dark regions)





DALI: Idea + Algorithm

- Not useful to compare entire matrices
- Create a set of sub-matrices of fixed size (6x6 aa) by collapsing original distance matrices into regions of overlap

Algorithm

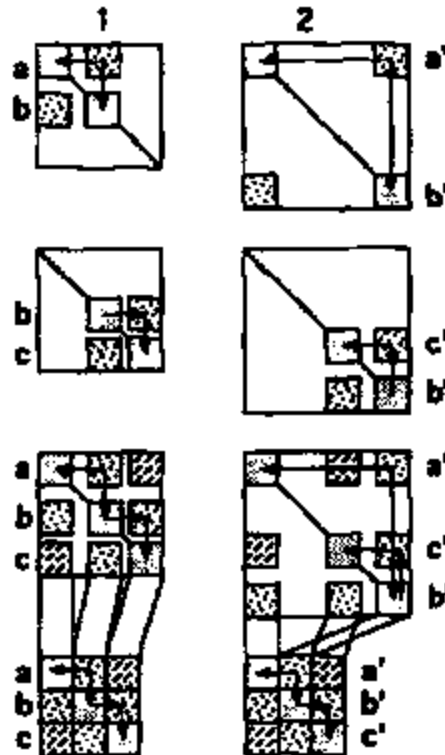
1. Identify a pair of matching sub-matrices within two matrices (adjacent fragments).
2. Identify a second pair which overlaps the first (contains one common structural element)
3. Since there is overlap between adjacent fragments, joined sub-matrices
4. Rearrange and “collapse” the matrix according to the aligned regions of the sequence



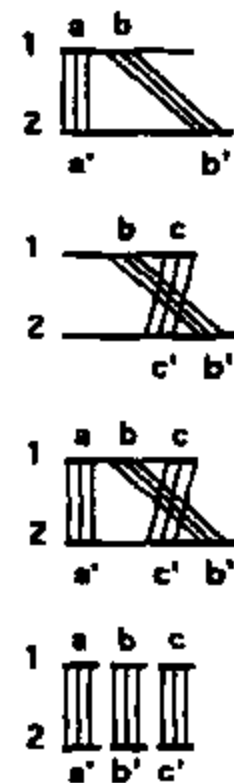
DALI: Example



3D



2D



1D

one pair

an overlapping pair

the two pairs combined

collapse



DALI: Statistical Analysis of Results

- Uses as a background the distribution of scores from an all-against-all comparison in a set of representatives structures (<30% seq. id.)
- **Z-score**: how many standard deviations from the mean of a distribution of alignments derived from the database background distribution
- Interpretation
 - >16 very close,
 - 8-16 pretty close
 - <8 not so close



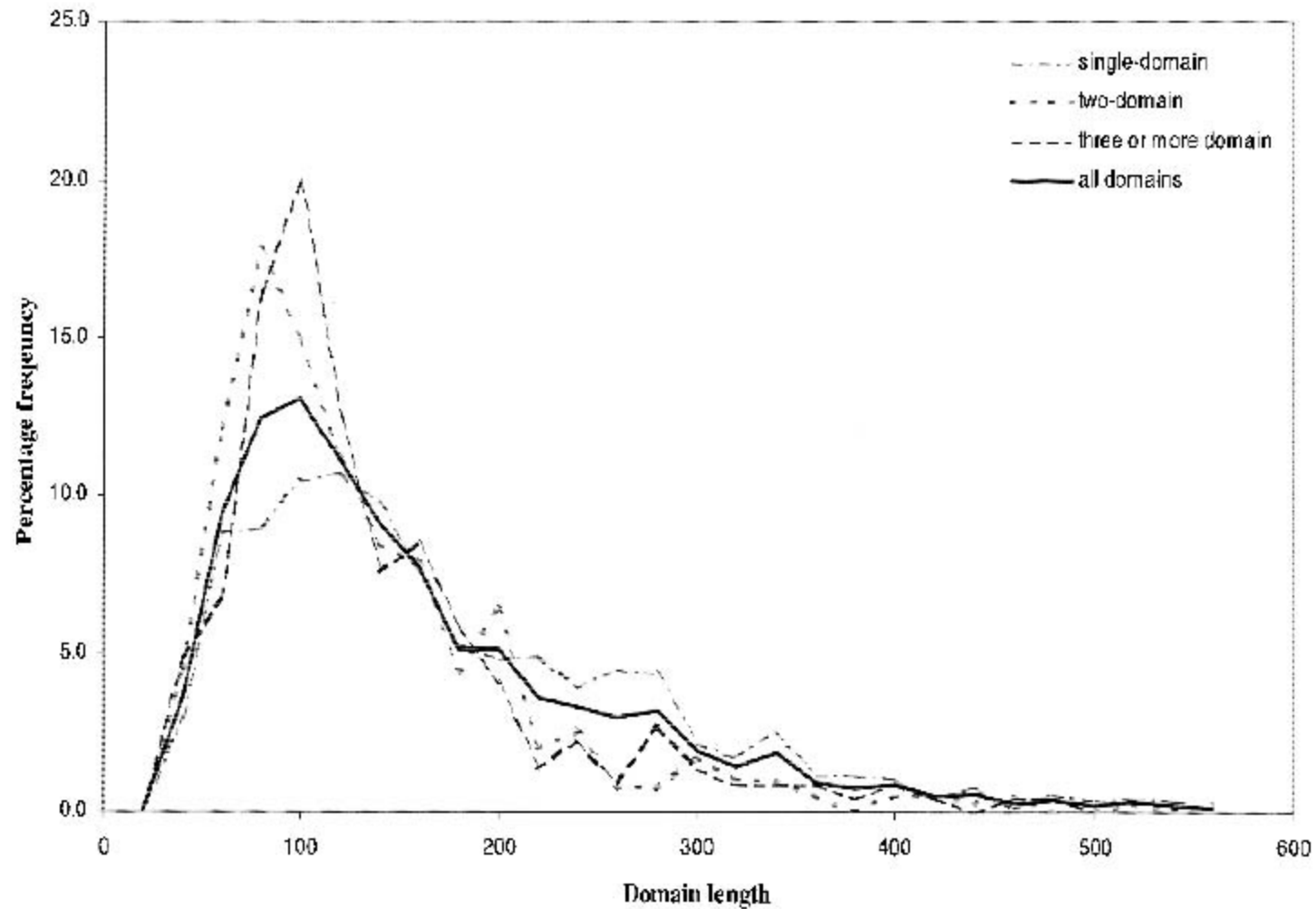
Structural Domains

- At first, domain concept used to describe distinct regions of 3D structures
 - Today, *considered evolutionary units, usually defined as spatially distinct structures with a well-defined hydrophobic core that could fold and function in isolation*
- Proteins with similar folds could be described as proteins sharing similar arrangement of protein domains of fragments (Ye et al. 2003)
 - fold space quite dense and continuous



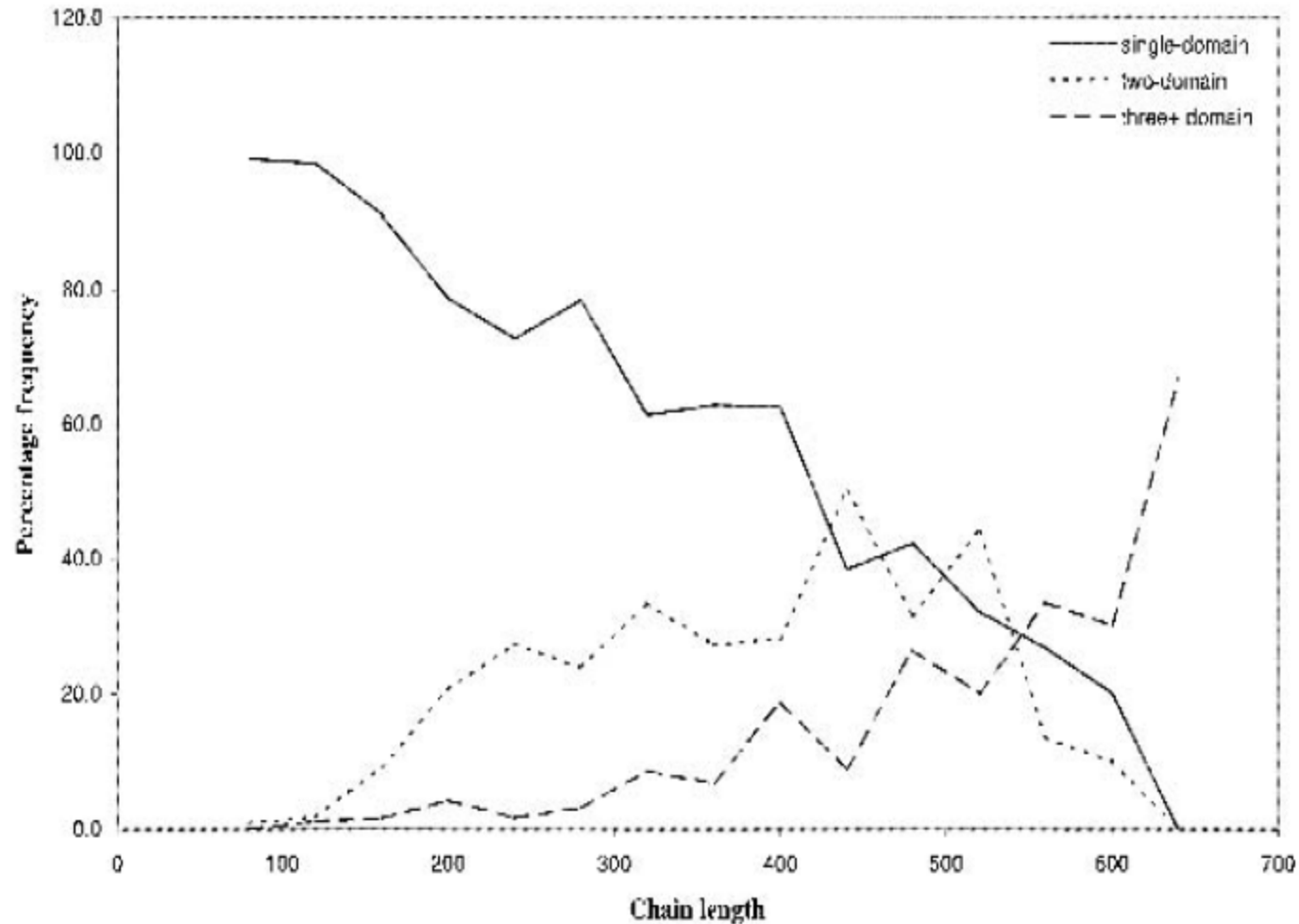
INDIANA UNIVERSITY

Domains Lengths





Domains Lengths





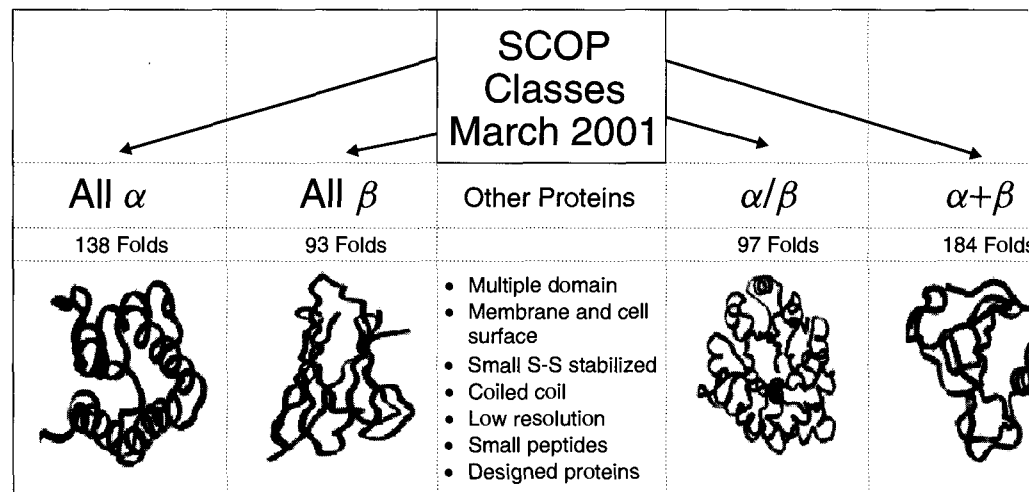
Protein Fold Space and Domains

- Protein structure can elucidate its function and evolutionary history
 - domains are correlated with protein function
 - domains provide insights into evolution of organisms (e.g. proteins that diverge in evolution maintain their global fold!, Russell et. al., 1997)
- Databases that map protein structure space at domain level
 - SCOP
 - CATH



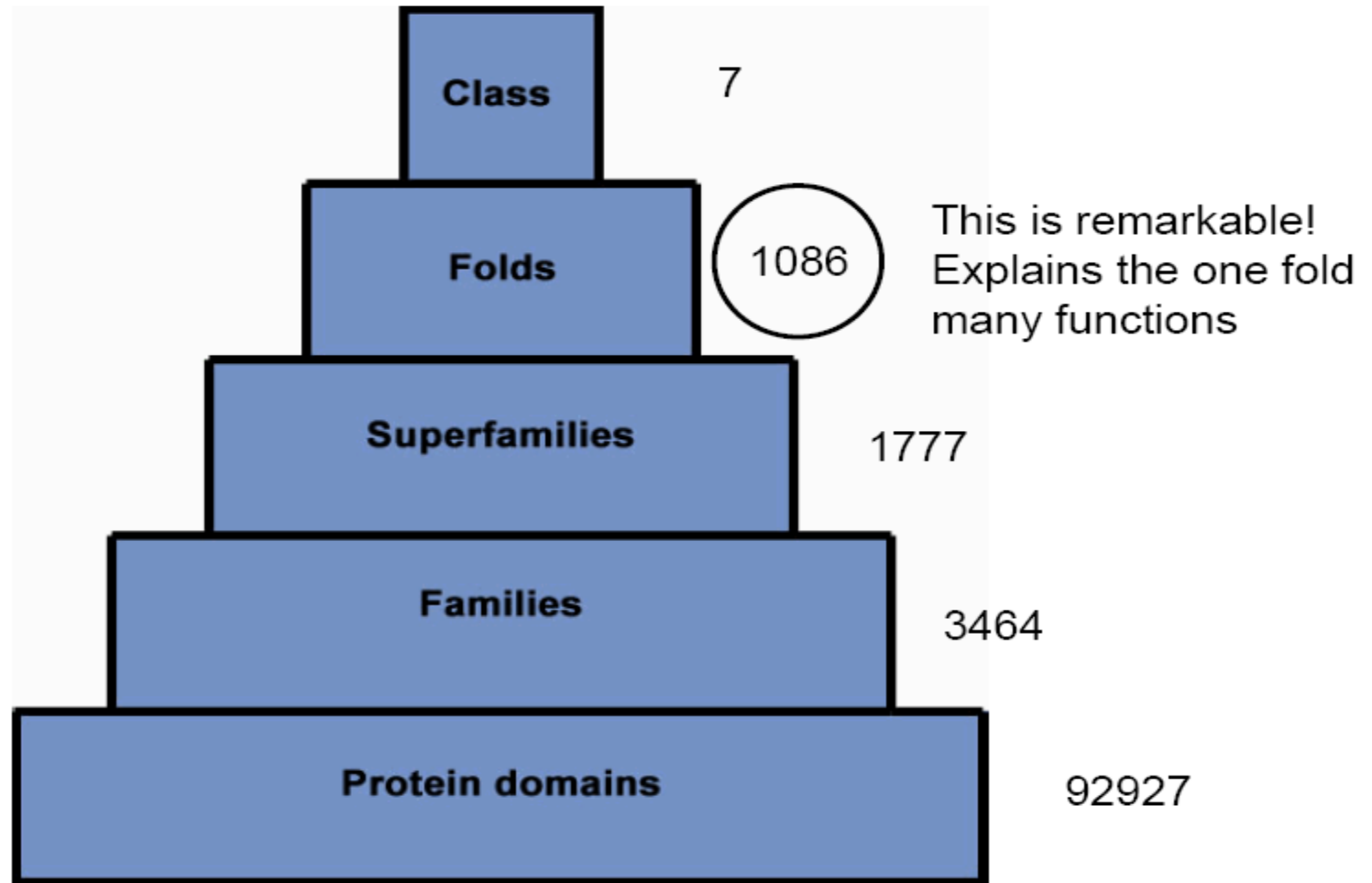
Structure Classification of Proteins, SCOP

- Hierarchical organization of proteins according to their structures and evolutionary origin
- Proteins classify manually by visual inspection and comparison of structures



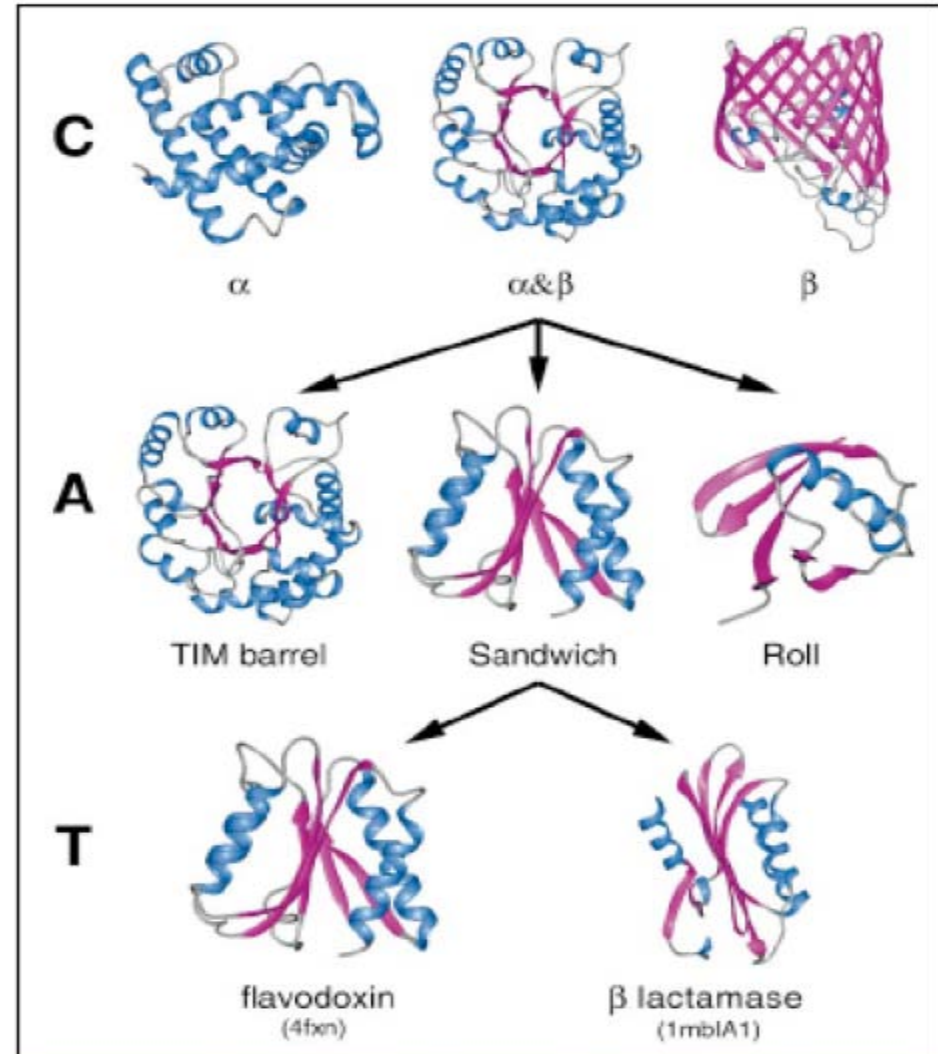


SCOP



Class Architecture Topology Homologous (CATH)

- Proteins structures classified with a combination of manual and automated procedures
- SCOP defines four different classes whereas CATH defines three





What we covered today

- Structure Comparison and Alignment
 - Chapter 16
- Protein Structure Evolution and SCOP DB
 - Chapter 17
- CATH Domain Structure DB
 - Chapter 18
- Identifying Structural Domains
 - Chapter 20



INDIANA UNIVERSITY

Thank You!

QUESTIONS?