Chapter 13

Protein Structure Visualisation, Comparison and Classification

Written examination:

Friday 7th February

10h00 – 11h30

L.01.220

Lecture slides:

All slides will be uploaded to OLAT

Evaluation:

A link is sent by email,
please do participate!

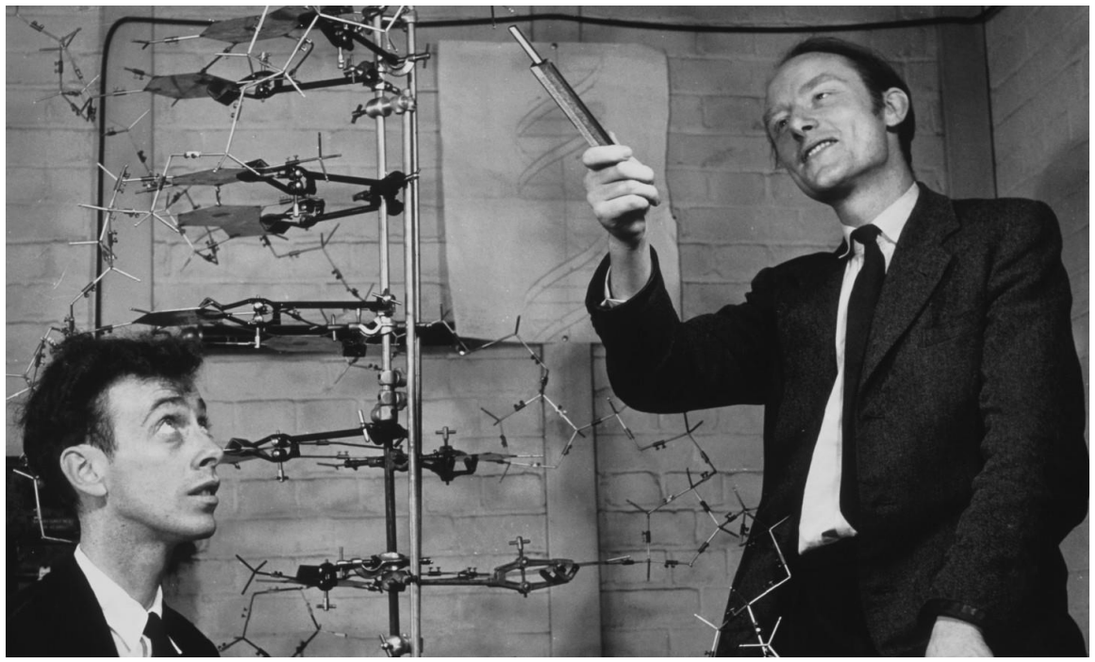
Overview

1. Introduction
2. Introduction to Biological Databases
3. Pairwise Sequence Alignment
4. Database Similarity Searching
5. Multiple Sequence Alignment
6. Profiles and Hidden Markov Models
7. Protein Motifs and Domain Prediction
8. Gene Prediction
9. Promoter and Regulatory Element Prediction
10. Phylogenetics Basics
11. Phylogenetic Tree Construction Methods and Programs
12. Protein Structure Basics
13. Protein Structure Visualization, Comparison and Classification
14. Protein Secondary Structure Prediction
15. Protein Tertiary Structure Prediction
16. RNA Structure Prediction
17. Genome Mapping, Assembly and Comparison
18. Functional Genomics
19. Proteomics

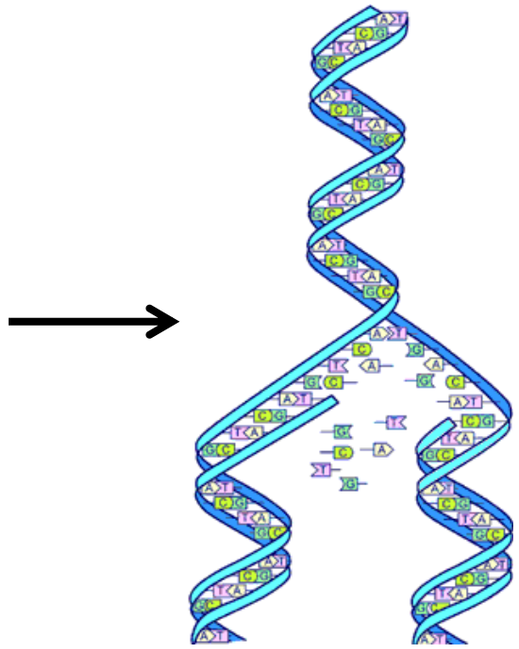
Structural models provide functional insight

Insight from structure: DNA (1953)

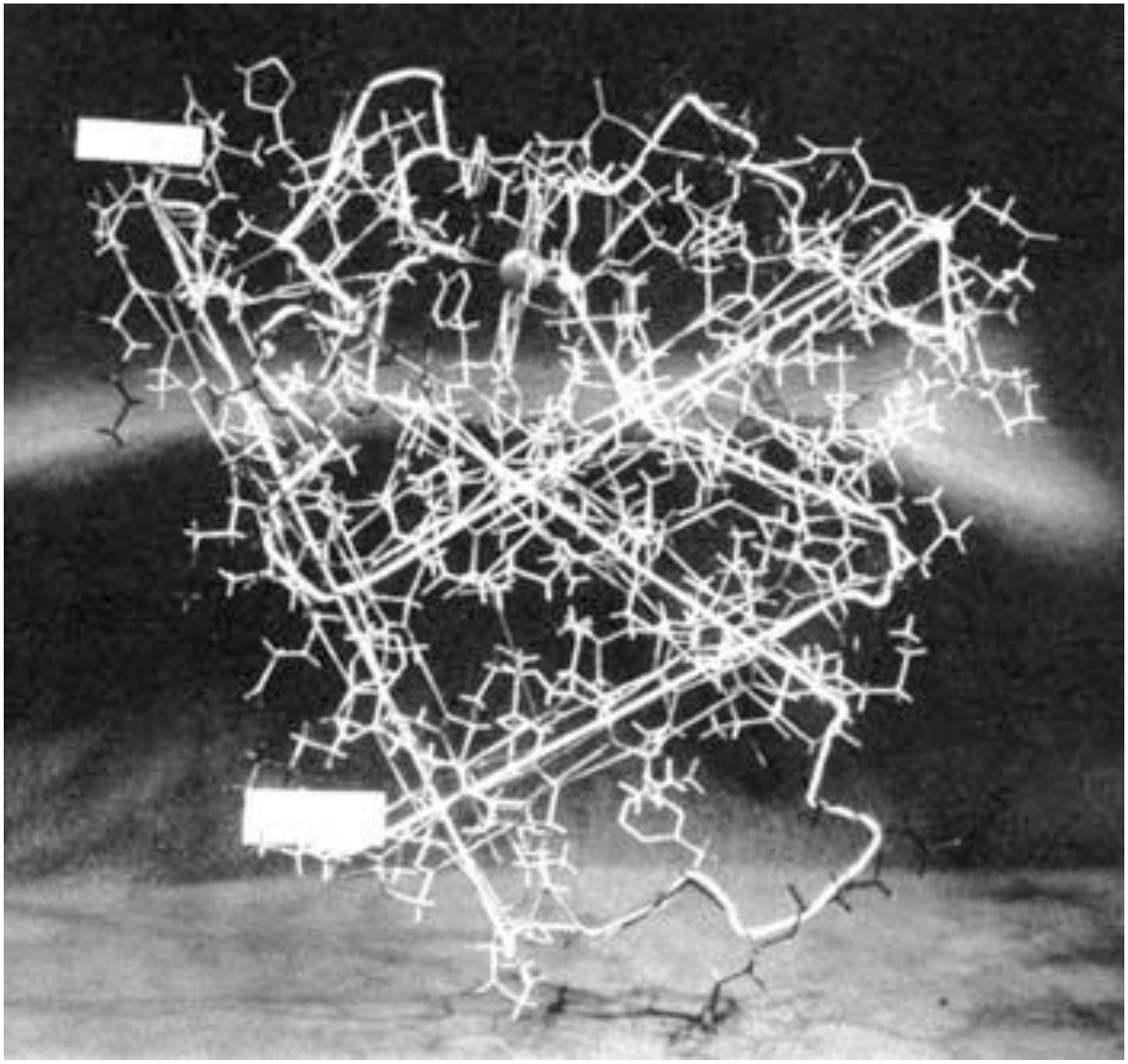
Structure



Replication mechanism



The first atomic protein model: myoglobin (1958)

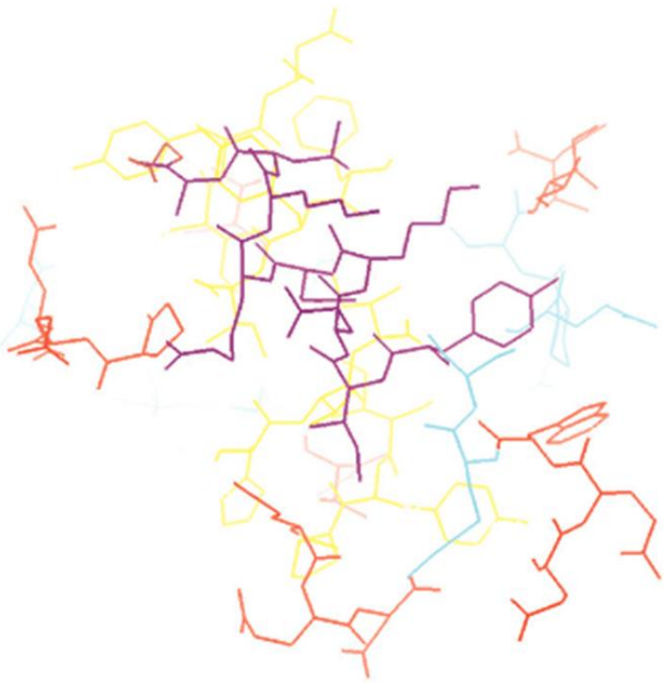


Computer visualisation

- PDB data files only contain atomic coordinates (x, y, z) without any chemical bonding information, so connectivity has to be inferred by the program
- Visualisation software allows interactive manipulation of structural images: the atomic model can be moved, rotated and specific areas zoomed in on
- Structures can be represented in a variety of styles, with different levels of abstraction, in order to emphasise particular aspects

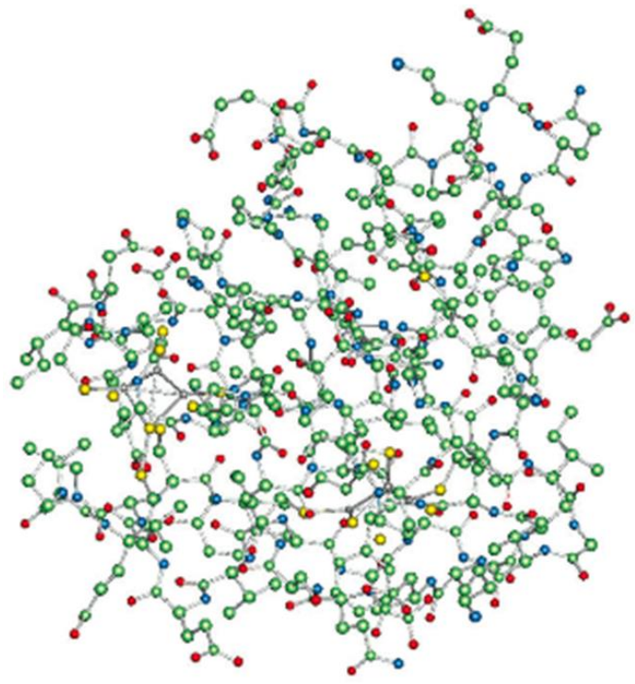
Atomic-level visualisation

- Useful to visualise specific details, *e.g.* ligand interaction
- Rather confusing if an entire protein is represented

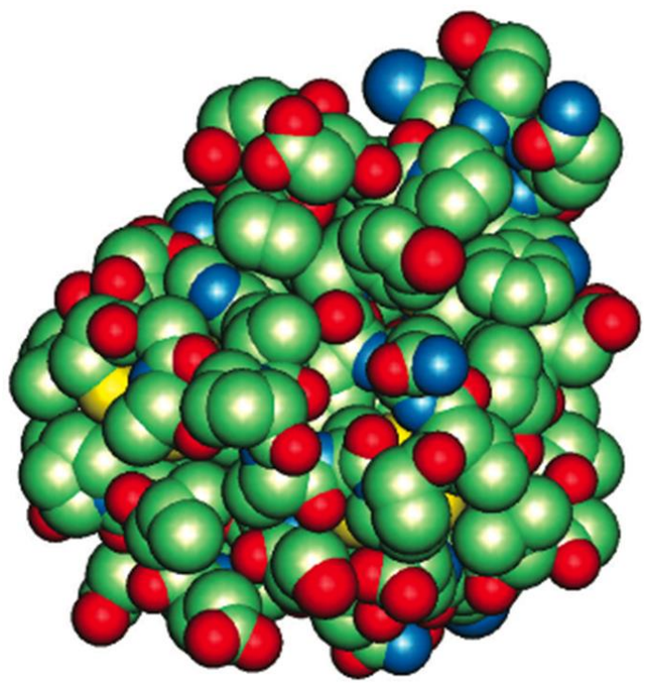


Wire-frame

Chemical bonds are drawn as lines or 3D sticks



Ball-and-stick

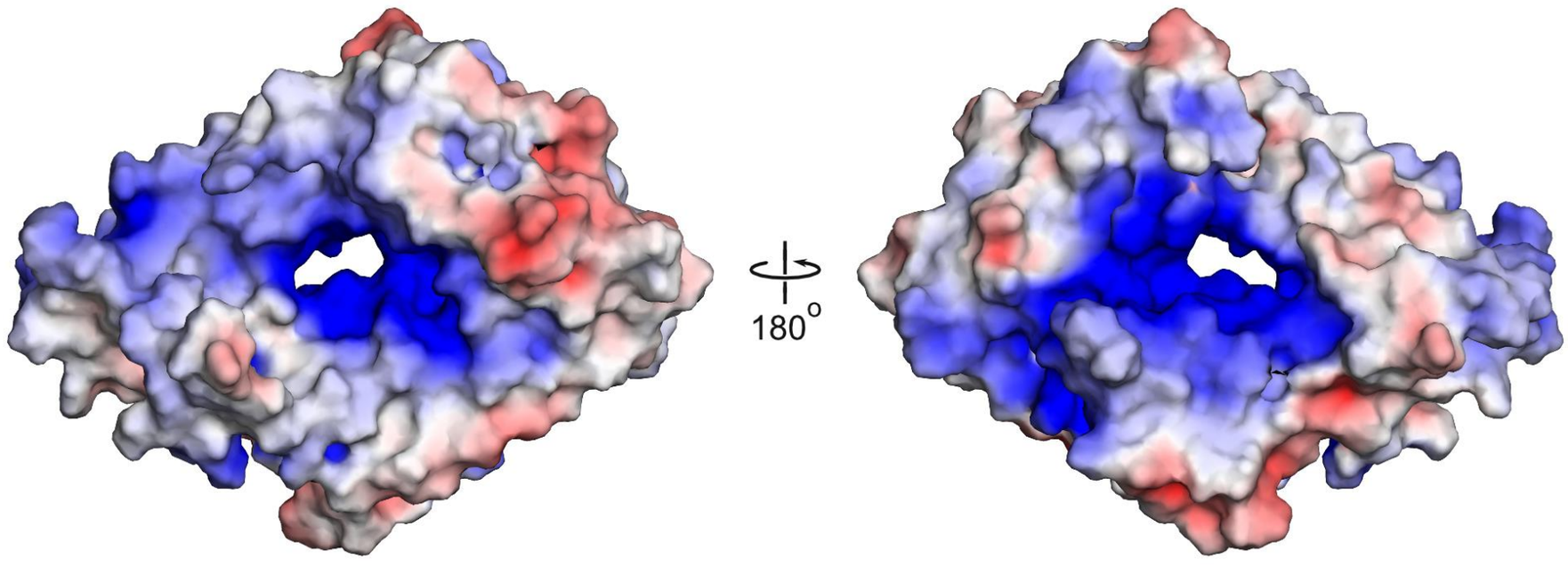


Space-filling

Each atom is shown as a sphere of a size corresponding to its Van der Waals radius

Surface representation

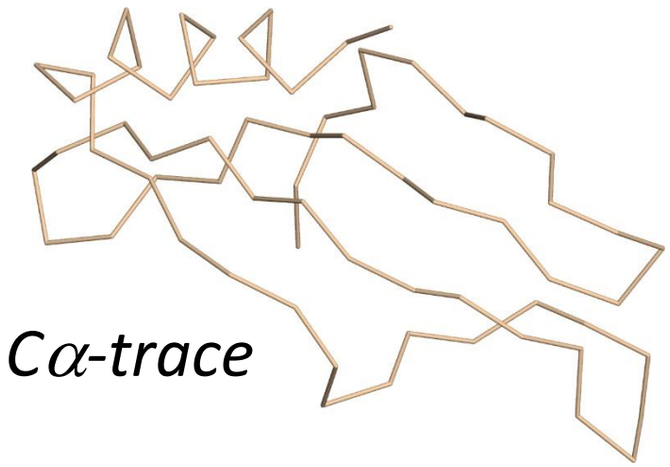
- Especially useful to reveal pockets and tunnels that act as binding sites
- Surface properties (such as charge distribution) can be calculated and visualised using colours



6RN5

Backbone representation

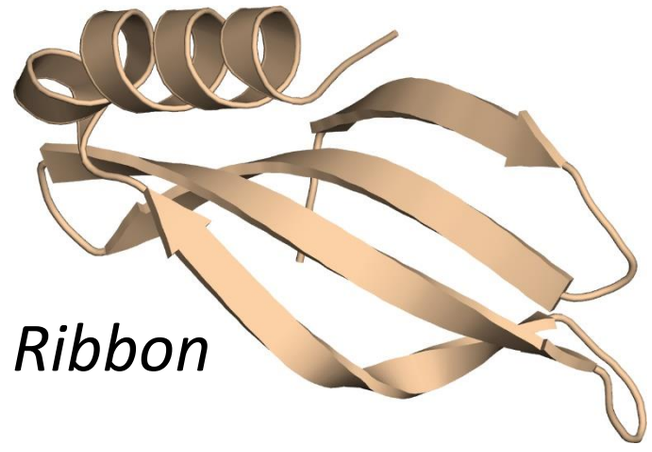
Emphasis on 3D fold rather than on atomic detail:



Cα-trace

Cα-trace

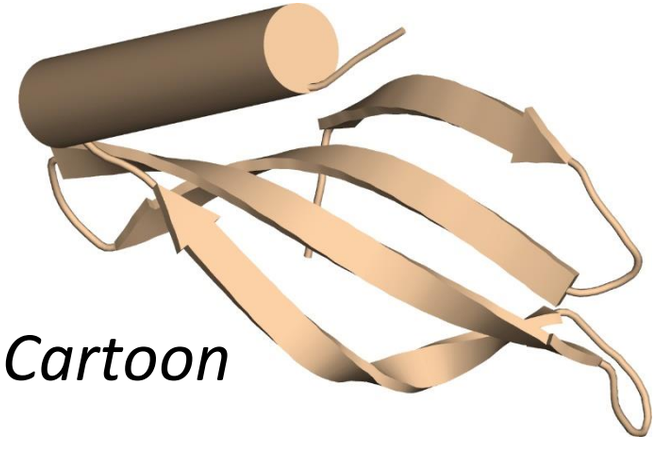
- Like wireframe or stick representation, but only showing “connections” between consecutive Cα-atoms



Ribbon

Ribbon/cartoon

- Ribbons or cylinders represent α-helices
- Arrows (N -> C) represent β-strands



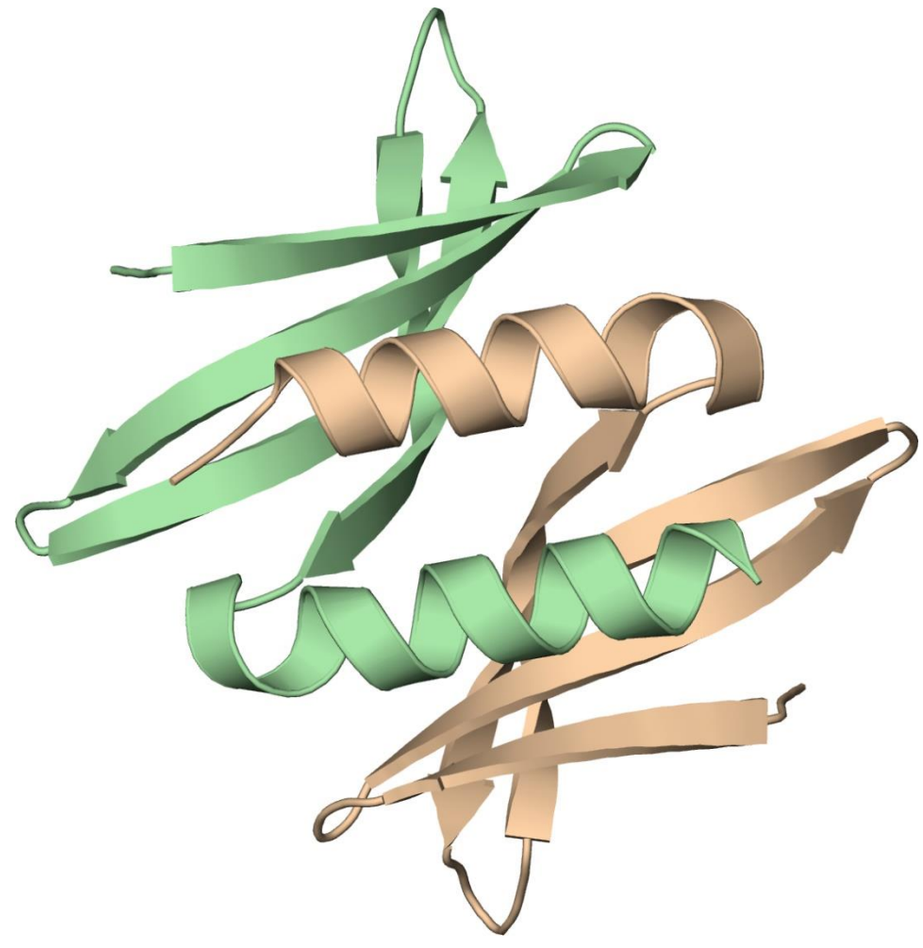
Cartoon

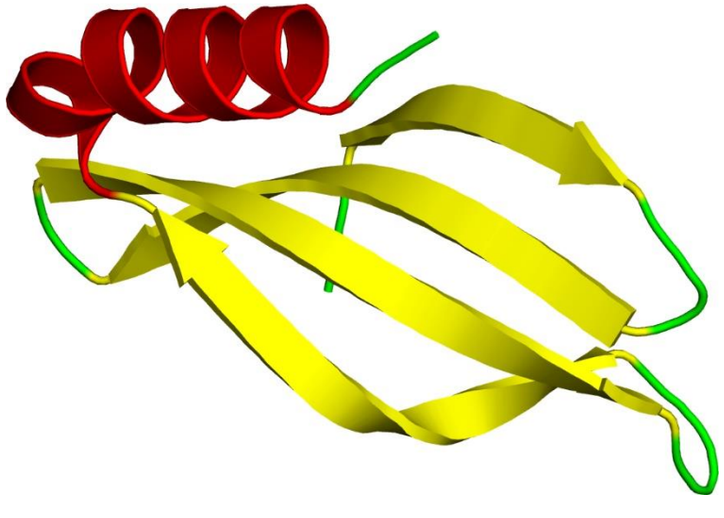
5A4N

Colours

Colour are often used to:

- Indicate various atom types
- Indicate surface charge
- Show different proteins in a complex
- Emphasise secondary structure elements

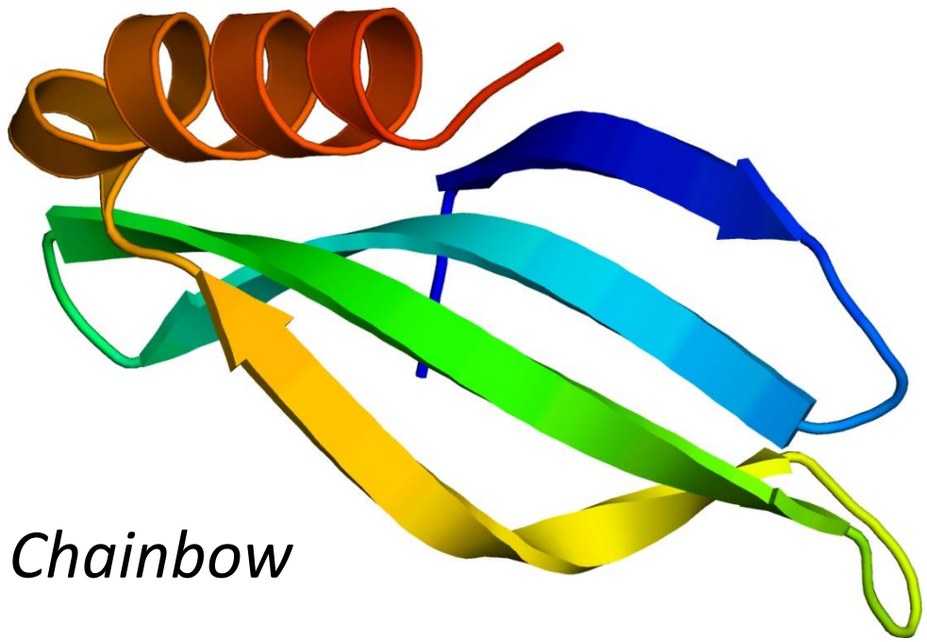




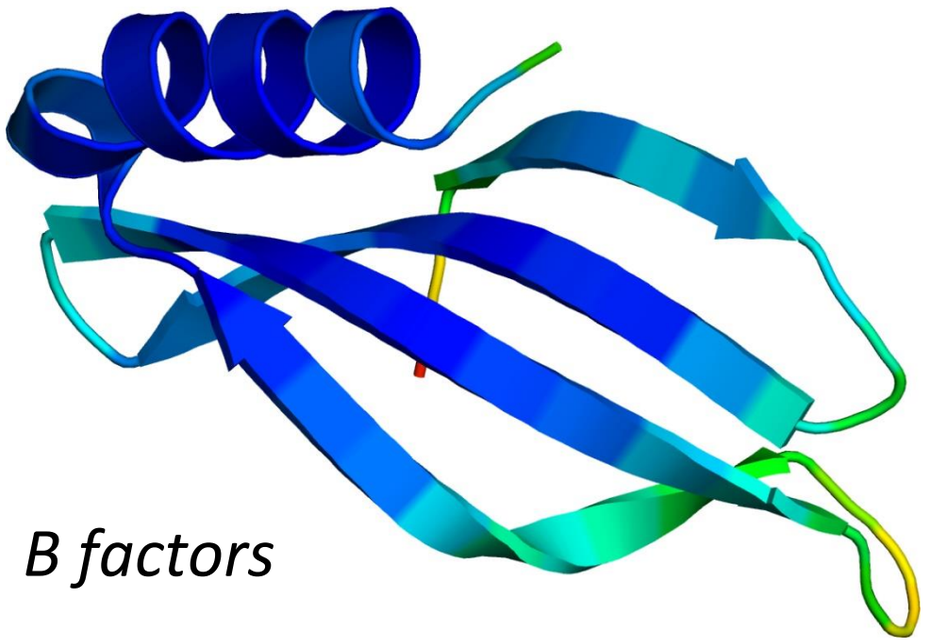
5A4N

Colour gradients

- According to position in the protein sequence (*"Chainbow"*)
- Temperature factors (B factors)



Chainbow



B factors

5A4N

Basic visualisation software

Web-based viewers:

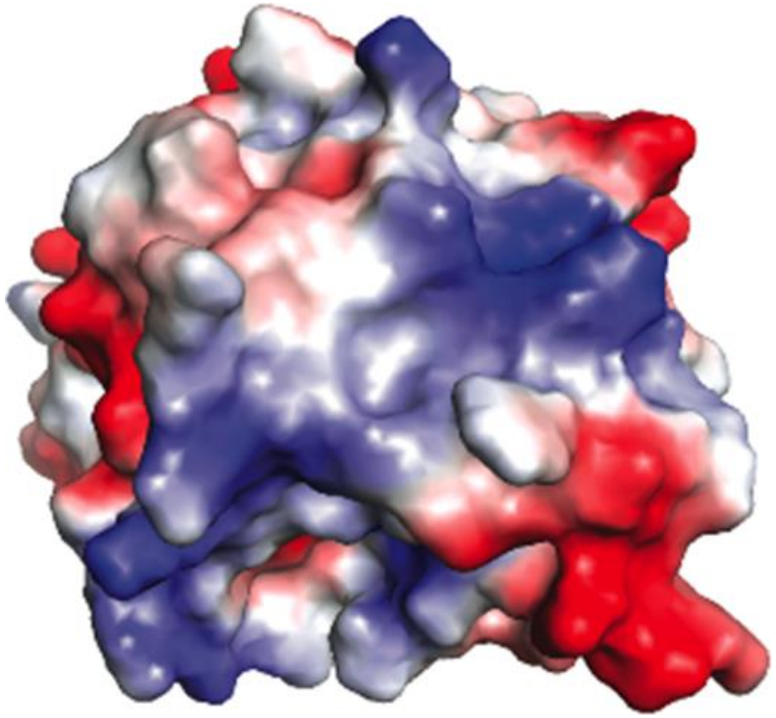
- PDB (simple viewer integrated in the PDB entry page of each protein)
- iCn3D (part of the Entrez database): https://www.ncbi.nlm.nih.gov/Structure/icn3d/docs/icn3d_about.html

Downloadable:

- Rasmol: http://www.openrasmol.org
- Swiss-PDBViewer: https://spdbv.unil.ch

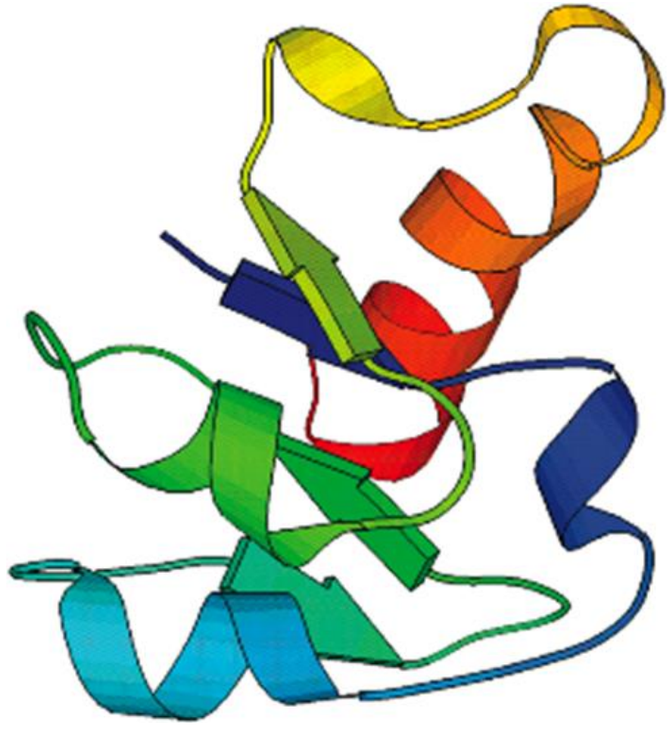
Surface rendering with charge distribution

- GRASP/GRASP2: https://honig.c2b2.columbia.edu/grasp



Script-based visualisation software

- Molscript: https://pekrau.github.io/MolScript



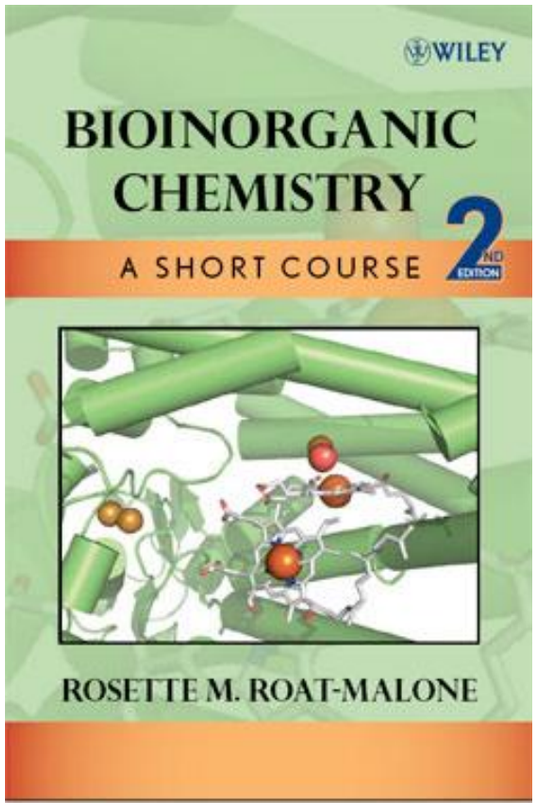
“State-of-the-art” structure visualisation software

- ChimeraX: https://www.cgl.ucsf.edu/chimerax
- PyMOL: https://pymolwiki.org/index.php/Windows_Install

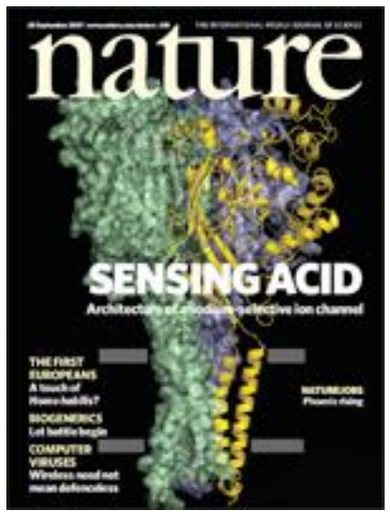
PyMOL

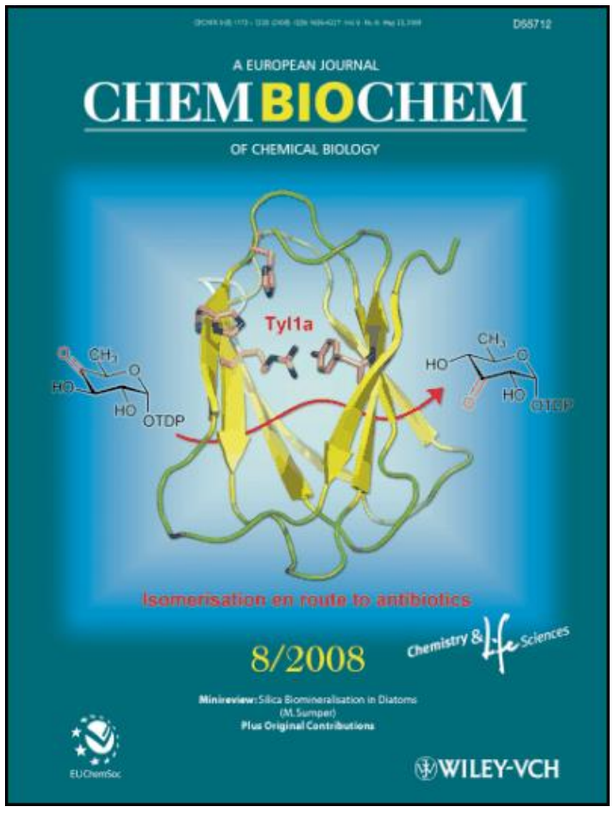

jbc
the journal of biological chemistry

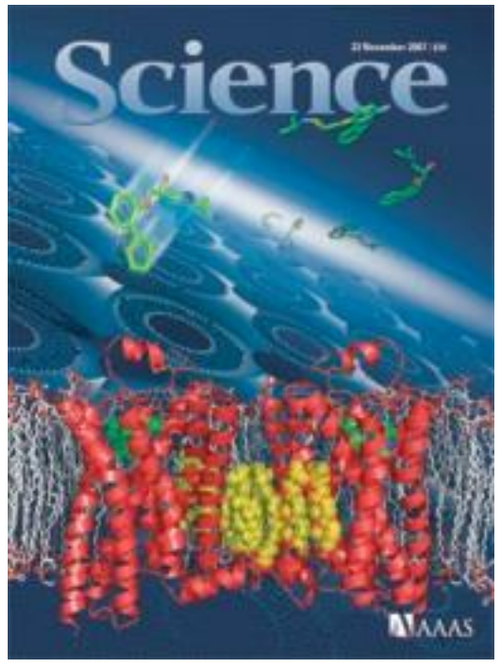

nature
PUMPING IONS
Beating the transmembrane concentration gradient


WILEY
BIOINORGANIC CHEMISTRY
A SHORT COURSE
2ND EDITION
ROSETTE M. ROAT-MALONE


Structure
Aminoacyl-tRNA Synthetases from Human Mitochondria


nature
SENSING ACID


A EUROPEAN JOURNAL
CHEMBIOCHEM
OF CHEMICAL BIOLOGY
Tyl1a
OTDP
Isomerisation en route to antibiotics
8/2008
Chemistry & Life Sciences
WILEY-VCH


Science
AAAS

Usefulness of protein structure comparison

- Identifying conformational changes in a protein and studying allosteric effects
- Revealing which parts of a structure are conserved and which parts are different in evolutionarily related proteins
- Identifying distant homologs: protein structures tend to have a much higher degree of conservation than sequences
- Improving multiple sequence alignments
- Protein structure comparison is a prerequisite for classification of proteins into fold families

Protein structure comparison methods

Main categories:

- Intermolecular methods
- Intramolecular methods
- Combined intra/intermolecular methods

Intermolecular methods

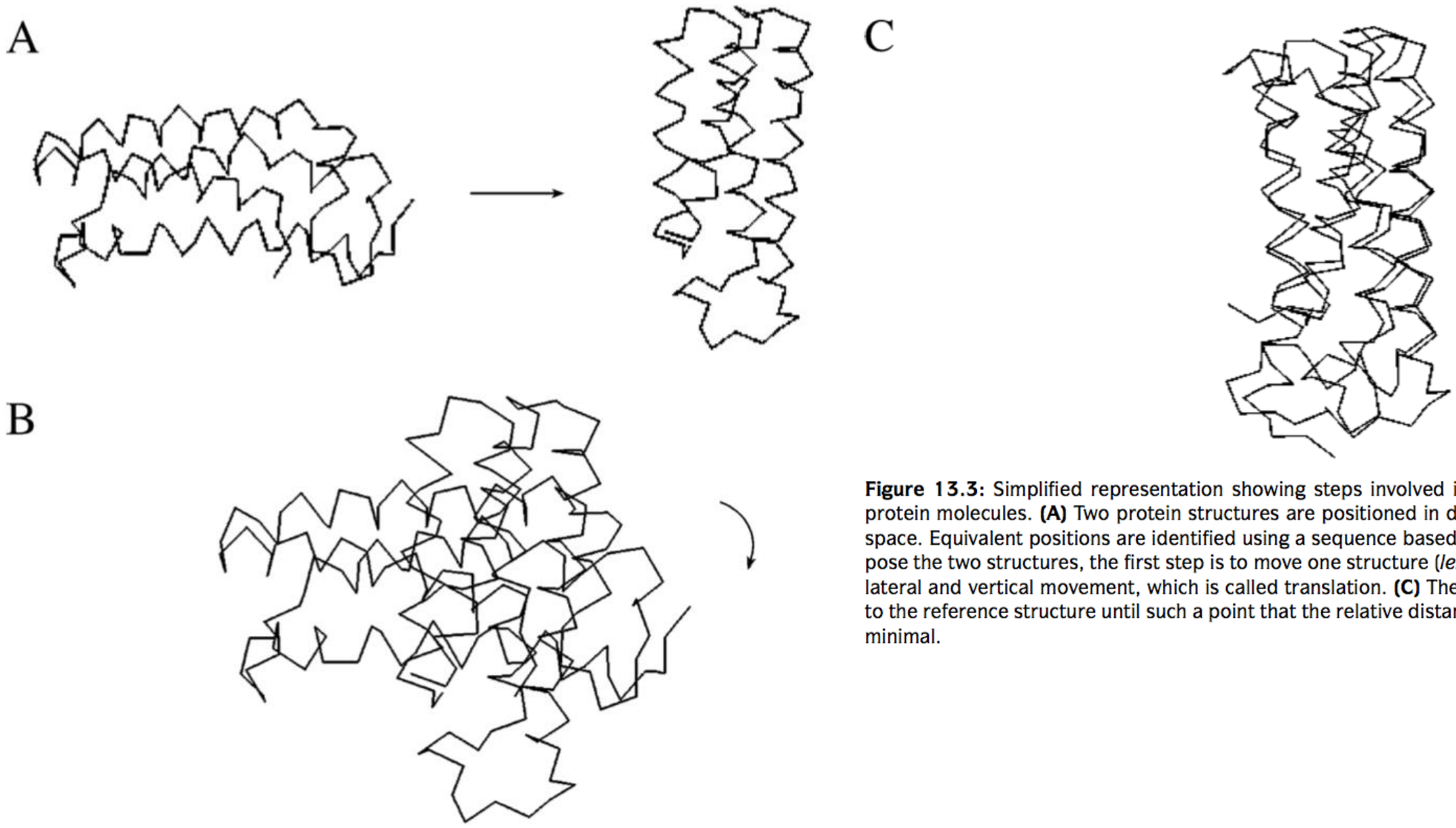


Figure 13.3: Simplified representation showing steps involved in the structure superposition of two protein molecules. **(A)** Two protein structures are positioned in different places in a three dimensional space. Equivalent positions are identified using a sequence based alignment approach. **(B)** To superimpose the two structures, the first step is to move one structure (*left*) relative to the other (*right*) through lateral and vertical movement, which is called translation. **(C)** The left structure is then rotated relative to the reference structure until such a point that the relative distances between equivalent positions are minimal.

Intermolecular methods

- Starts with identification of equivalent residues, *e.g.* by means of a sequence alignment
- One of the structures is moved towards the other structure (= translation) and rotated, until the shortest possible distances between equivalent postions are reached
- The "over-all distance" to be minimised is calculated as a root mean square deviation (RMSD) of pairwise Cα - Cα distances:

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

Intermolecular methods

RMSD is size-dependent:

- Larger proteins tend to have a higher RMSD value for the same degree of similarity
- Proposed correction formula:

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

Intermolecular methods

Identification of equivalent residues is the most challenging part!

For comparison of distantly related structures:

- Deletion of variable regions outside secondary structure elements
- First divisions of the proteins into small fragments, then fragment based matching, followed by joint superposition for the entire structure

Intermolecular methods: iterative approaches

Comparison of more distantly related structures by *iterative optimisation*:

- First alignment of sequences *e.g.* using dynamic programming
- Equivalent residues are used for first round of superposition
- New equivalent residues are identified by close proximity at three-dimensional level
- Sometimes residues are removed from the analysis if they are further apart than a certain threshold
- New rounds of superposition until RMSD cannot be further improved

Alternative: intramolecular methods

Rely on structural internal statistics

- Does not depend on sequence similarity between the proteins compared
- Does not generate a physical superposition of structures
- Provides quantitative evaluation of the structural similarity between corresponding residue pairs

Intramolecular methods

- Generate a distance matrix between all residues within the same protein
- Distance matrices from the two structures are shifted relative to each other to find areas that overlap
- Similar intramolecular distance patterns imply similar structural regions

Combined intra/intermolecular methods

- Corresponding residues are identified using the intramolecular method
- Structure superposition is performed basing on corresponding residues
- Additional to RMSD, sometimes structural properties such as secondary structure types, torsion angles, accessibility and local hydrogen bonding environment are used

Multiple structure alignment

Progressive approach, *cf.* MSA construction:

- First, all structures are compared in a pairwise fashion
- Distance matrix based on a structural dissimilarity score (*e.g.* RMSD)
- Construction of a phylogenetic tree
- Two most similar structures are aligned to a median structure
- Other structures are progressively added based on the hierarchy of the guide tree

Protein structure comparison servers

DALI:

- http://ekhidna.biocenter.helsinki.fi/dali
- Intramolecular distance method

Various other (but similar) methods are accessible via the **PDB** web site:

- https://www.rcsb.org/alignment

Protein structure classification

- Comparison and classification of protein structures helps finding functional and evolutionary relationships between structures (and proteins)
- Function of a new protein structure can often be identified or better understood by finding out what structural class it belongs to
- This is approach is analogous to finding out what sequence family a newly obtained protein sequence belongs to – but generally structural comparisons are more sensitive/informative

Protein structure classification systems

- **SCOP** (Structural Classification of Proteins)
- **CATH** (Class, Architecture, Topology and Homologous (super)family)

Establishing a structure classification system

Requirements:

1. Removal of redundancy from databases: representatives are selected from groups of redundant structures
2. Structurally distinct domains within a structure are separated
3. Domains of similar structures are clustered

Structural Classification of Proteins (SCOP)

- Since 2014: replaced by SCOP2 (updated version)
- https://www.ebi.ac.uk/pdbe/scop
- Database for comparing and classifying protein structures
- Almost entirely based on manual examination of protein structures
- Grouped into hierarchies of 7 **classes**, 1195 **folds**, 1962 **superfamilies** and 3902 **families** (numbers are for the 23 Feb 2009 release)

Structural Classification of Proteins (SCOP)

Families

- Consist of proteins having high sequence identity (>30%)
- Proteins within a family clearly share a close evolutionary relationship and normally have the same function

Superfamilies

- Consist of families with similar structures, but weak sequence similarity
- Members of superfamilies share a common but distant ancestral origin

Structural Classification of Proteins (SCOP)

Folds

- Consist of superfamilies with a (manually determined) common "core structure":
  - Similar overall secondary structures with similar orientation and connectivity between them
- Members within the same fold do not necessarily have evolutionary relationships, *i.e* the shared core structure may be the result of *analogy*

Structural Classification of Proteins (SCOP)

Classes

- Highest level of the SCOP hierarchy
- Consist of folds with similar (but non-identical) core structures
- Distinguishes groups of proteins by secondary structure composition and general features like "membrane proteins"
- Folds within the same class are most likely unrelated in evolution

Structural Classification of Proteins (SCOP)

Examples of classes:

- All-alpha proteins
- All-beta proteins
- Alpha and beta proteins (a/b): "beta-alpha-beta", mainly parallel beta-sheets
- Alpha and beta proteins (a+b): independent alpha helices, mainly antiparallel beta-sheets
- Multi-domain proteins
- Membrane and cell surface proteins
- Small proteins

CATH

- http://www.cathdb.info
- Classifies proteins based on automatic structural alignment program (SSAP) as well as manual comparison
- Individual domain structures are classified at five major levels: 4 **classes**, 40 **architectures**, 1375 **folds/topologies**, 2738 **homologous superfamilies** and 16933 **homologous families**, 235858 domains (March 26 2013)

CATH

Class

- Based on secondary structure content

Architecture

- Intermediate between fold and class
- Describes overal packing and arrangement of secondary structures independent of connectivity between the elements

CATH

Topology

- Describes overall orientation of secondary structures taking into account connectivity between secondary structure elements

Homologous superfamily and homologous family levels

- Equivalent to superfamily and family level in SCOP

Comparison of SCOP and CATH

SCOP

- Almost entirely based on manual comparison of structures by human experts with no quantitative criteria to group proteins
- Method is subjective, exact boundaries between levels and groups are sometimes arbitrary

CATH

- Combination of manual curation and automated procedure makes process less subjective
- *E.g.* domains are determined by consensus of three different domain recognition algorithms

Comparison of SCOP and CATH

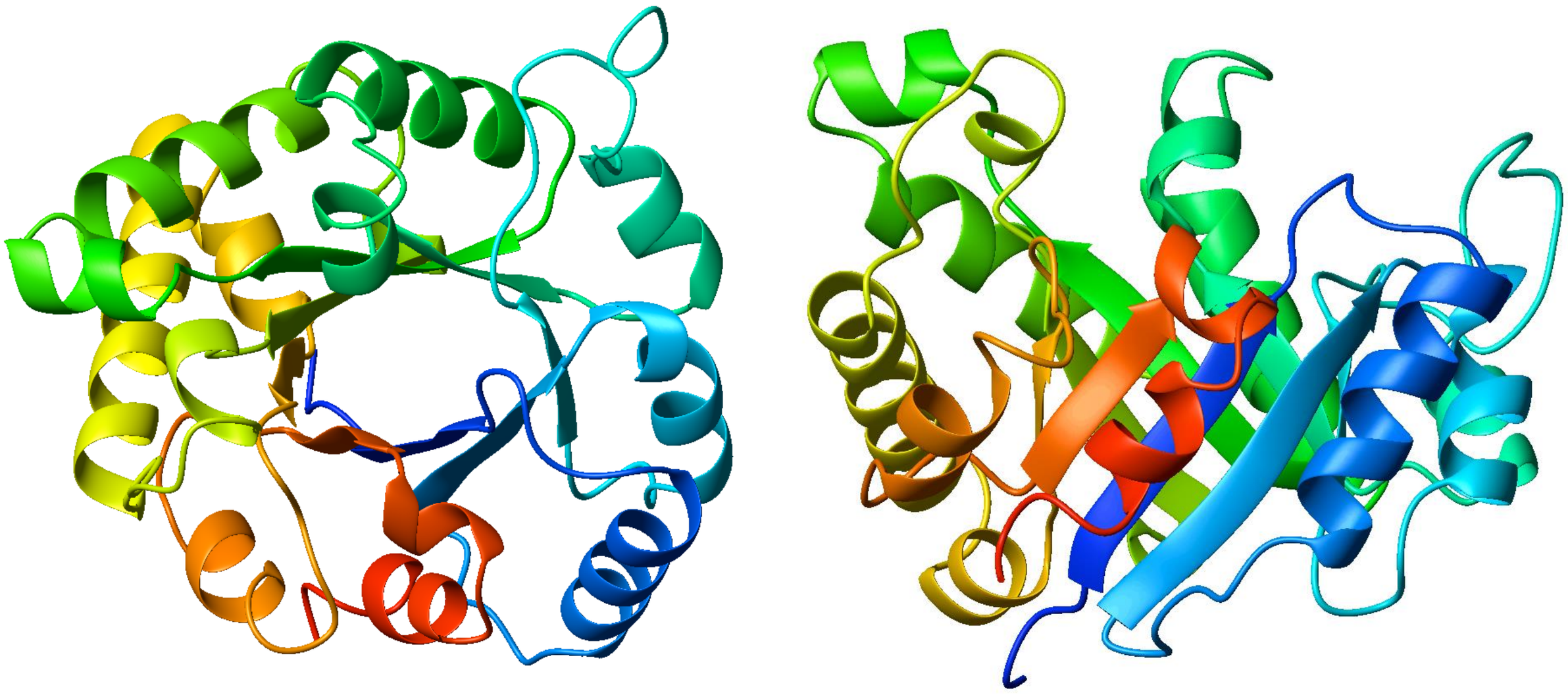
CATH

- Extra "architecture" level makes structure classification more continuous
- Fixed thresholds in structural comparison may make assignment less accurate

Still, classification results of both systems are quite similar:

- Only about 20% of the structure fold assignments are different

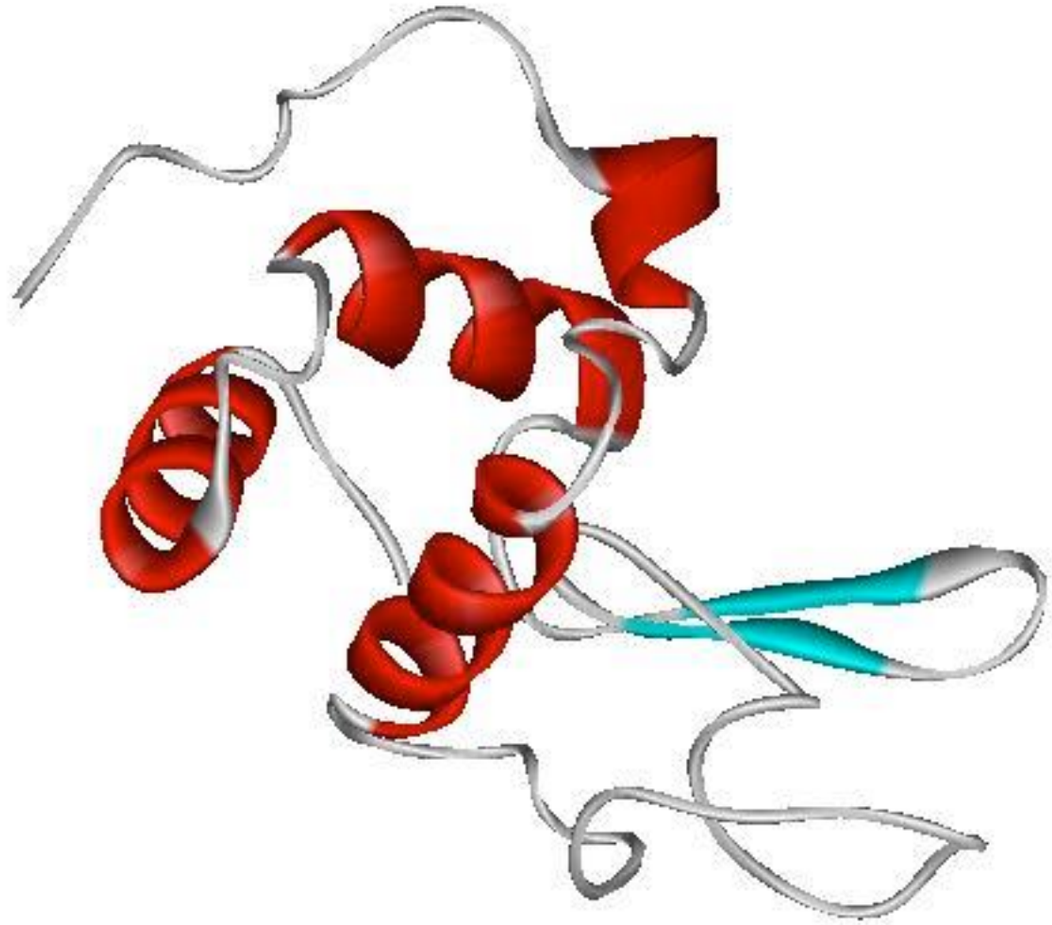
3TIM



PDB code: 4tim

SCOP		CATH	
Class	Alpha and Beta (α/β)	*Class*	Alpha Beta
		Architecture	Barrel
Fold	TIM beta/alpha-barrel	*Topology*	TIM Barrel
Superfamily	Triosephosphate isomerase	*Homologous Superfamily*	Triosephosphate isomerase
Family	Triosephosphate isomerase	*Homologous Family*	Triosephosphate isomerase

1LYS



PDB code: 1lys

SCOP		CATH	
Class	Alpha and Beta (α+β)	*Class*	Mainly Alpha
		Architecture	Orthogonal Bundle
Fold	Lysozyme-like	*Topology*	Lysozyme
Superfamily	Lysozyme-like	*Homologous Superfamily*	Hydrolase (O-glycosyl)
Family	C-type lysozyme	*Homologous Family*	Hydrolase