Lecture 1 - Introduction to Structural Bioinformatics

Motivation and Basics of Protein Structure

Objectives of the course

- Understanding protein function.
- Applications to Computer Aided Drug Design.
- Development of efficient algorithms to evaluate the above “in silico”.
- Emphasis on the “structure” related problems – Geometric Computing in Molecular Biology.
- Show relevance to other spatial “pattern discovery” tasks.

Most of the Protein Structure slides – courtesy of Hadar Benyaminy.

Textbook

There is no single, double or triple textbook for this course.

Most of the material is based on journal articles and research done by the Wolfson-Nussinov Structural Bioinformatics group at TAU.

Nevertheless :

Recommended Literature (1):

- Setubal and Meidanis, Introduction to Computational Biology, (1997).
- A. Lesk, *Introduction to* Protein Architecture, 2’nd edition (2001).
- S.L. Salzberg, D.B.Searls, S. Kasif (editors), Computational Methods in Molecular Biology, (1998).

Recommended Literature (2):

- Branden and Tooze, Introduction to Protein Structure (2’nd edition).
- D. Gusfield, Algorithms on Strings, Trees and Sequences, (1997).
- Voet and Voet, Biochemistry (or, any other Biochemistry book in the Library).
- M. Waterman, Introduction to Computational Biology.

Strongly Recommended Literature (currently not in the library):

- Protein Bioinformatics.
- Structural Bioinformatics.

Recommended Web Sites:

- Enormous number of sites.
- Search using “google”.
- PDB site http://www.rcsb.org/pdb/
- Birbeck course on protein structure.

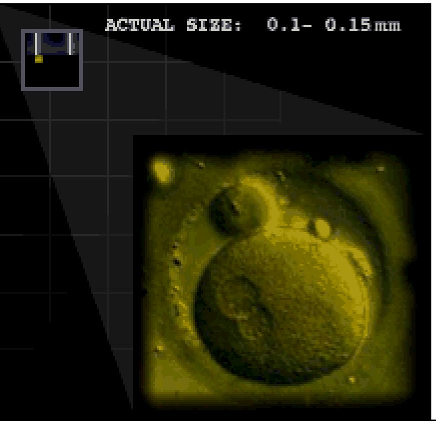
Journals :

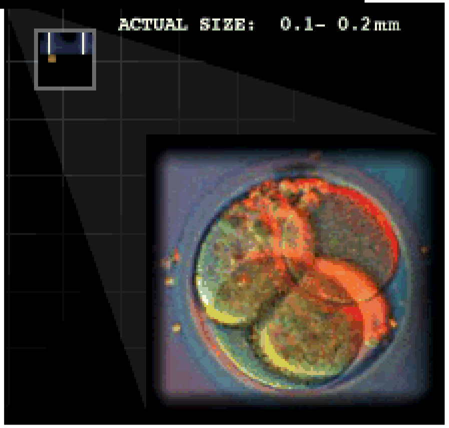
- Proteins : Structure, Function, bioinformatics.
- Journal of Computational Biology.
- Bioinformatics (former CABIOS).
- Journal of Molecular Biology.
- Journal of Computer Aided Molecular Design.
- Journal of Molecular Graphics and Modelling.
- Protein Engineering.

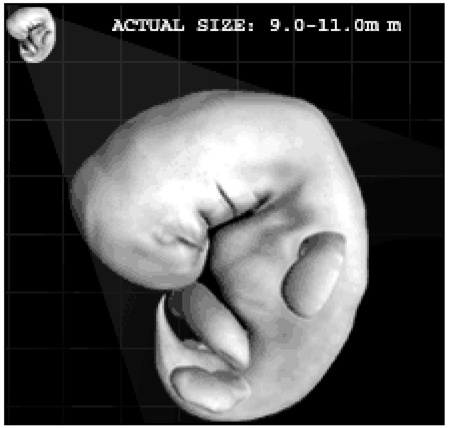
Computational Biology Conferences:

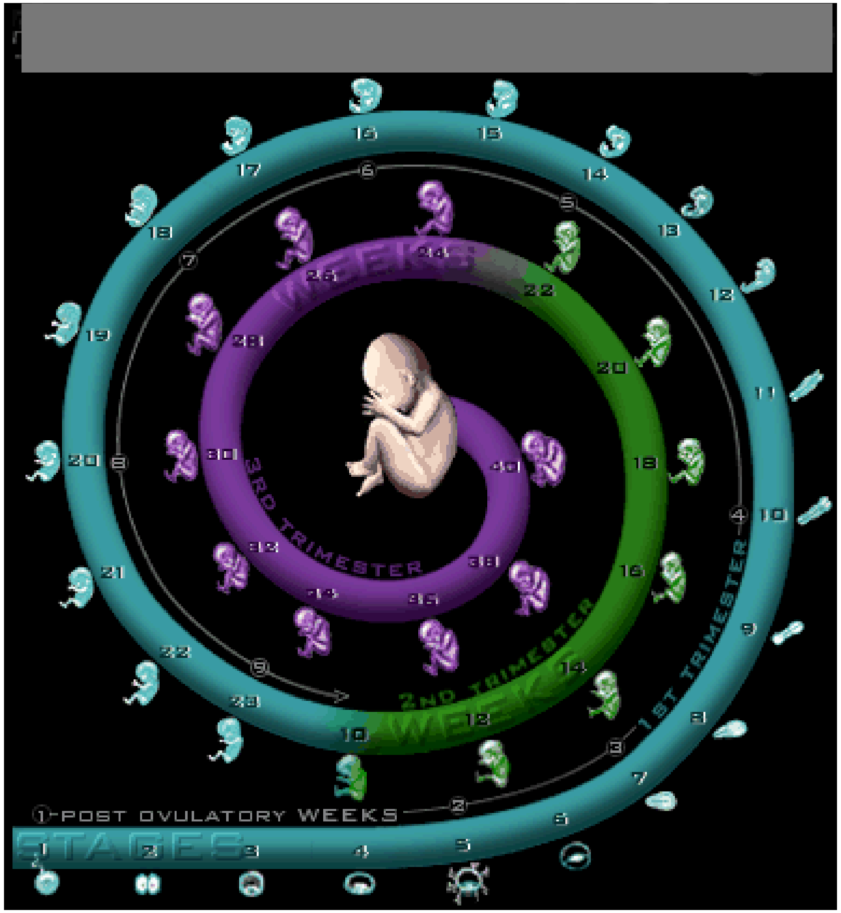
- **ISMB** - International Conference on Intelligent Systems in Molecular Biology.
- **RECOMB** - Int. Conference of Computational Molecular Biology.
- **ECCB** - European Conference on Computational Bio.
- **WABI** - Workshop of Algorithms in Bioinformatics .

Cell- the basic life unit

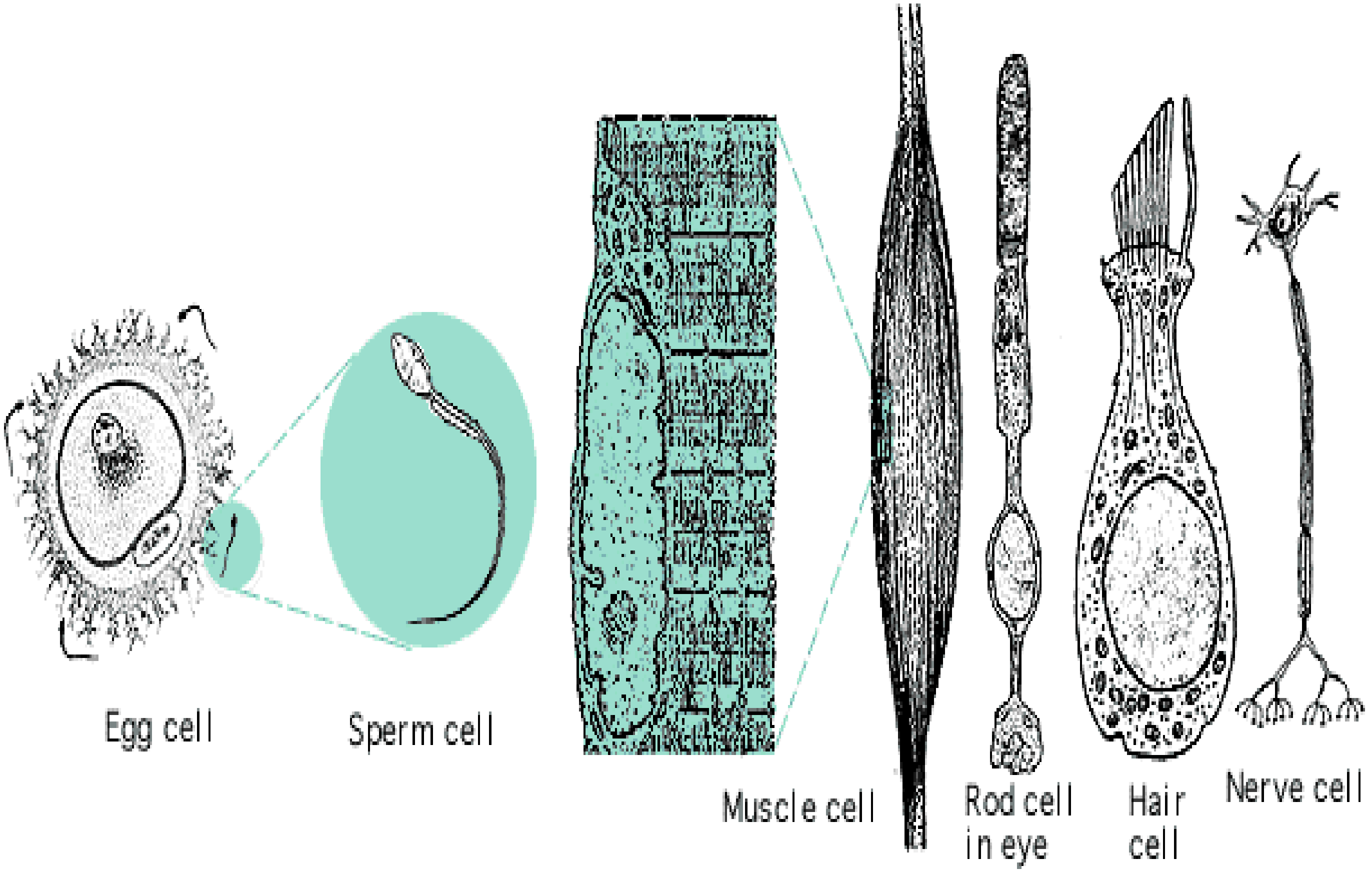








Different cell types

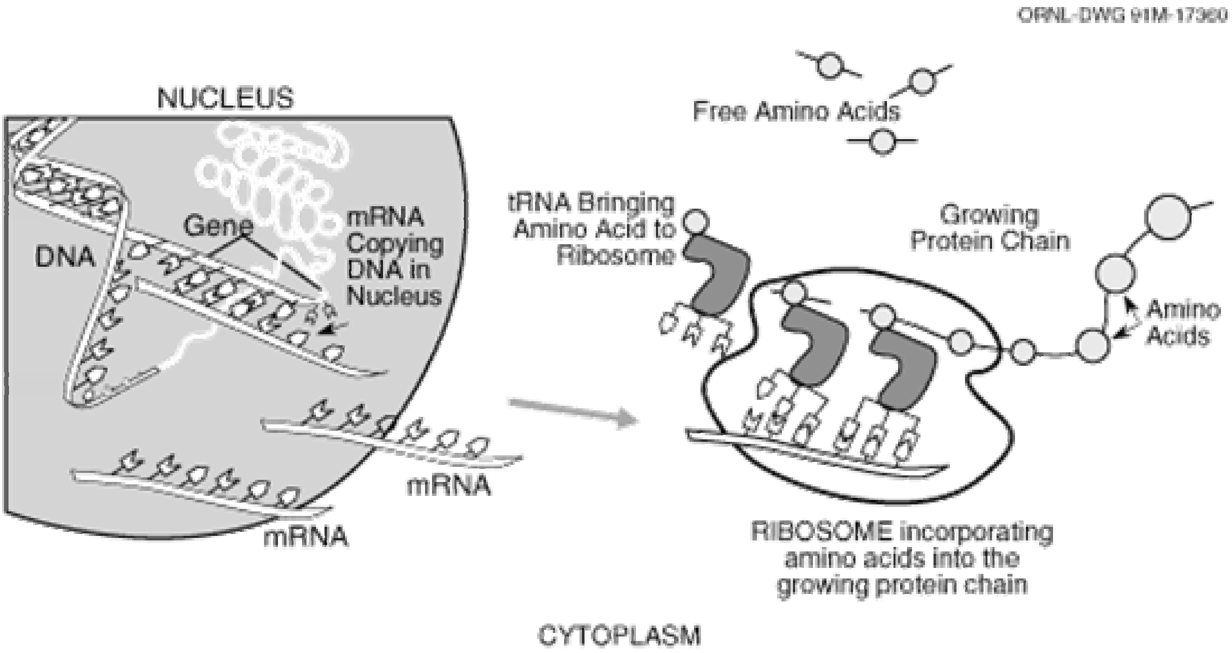


Size of protein molecules (diameter)

- cell (1x10^{-6} **m**) μ microns
- ribosome (1x10^{-9} **m**) nanometers
- protein (1x10^{-10} **m**) angstroms

The central dogma

- DNA ---> RNA ---> Protein
- {A,C,G,T} {A,C,G,U} {A,D,..Y}
- *4 letter alphabets* *20 letter alphabet*
- Sequence of nucleic acids seq of amino acids


ORNL-DWG 91M-17360
NUCLEUS
Free Amino Acids
tRNA Bringing Amino Acid to Ribosome
Growing Protein Chain
Gene
mRNA Copying DNA in Nucleus
DNA
Amino Acids
mRNA
mRNA
RIBOSOME incorporating amino acids into the growing protein chain
CYTOPLASM

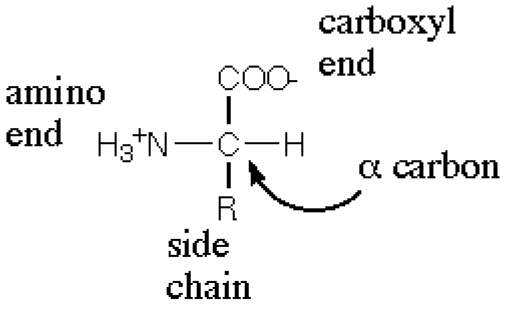
When genes are expressed, the genetic information (base sequence) on DNA is first transcribed (copied) to a molecule of messenger RNA in a process similar to DNA replicatior The mRNA molecules then leave the cell nucleus and enter the cytoplasm, where triplets of (codons) forming the genetic code specify the particular amino acids that make up an) bases individual protein.

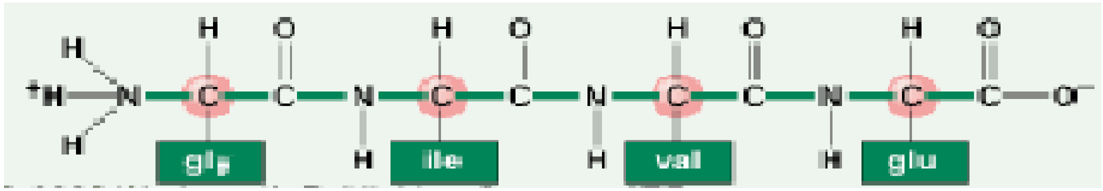
This process, called translation, is accomplished by ribosomes (cellular components compose of proteins and another class of RNA) that read the genetic code from the mRNA, and transfer RNAs (tRNAs) that transport amino acids to the ribosomes for attachment to the growing protein. (From www.ornl.gov/hgmis/publicat/primer/)

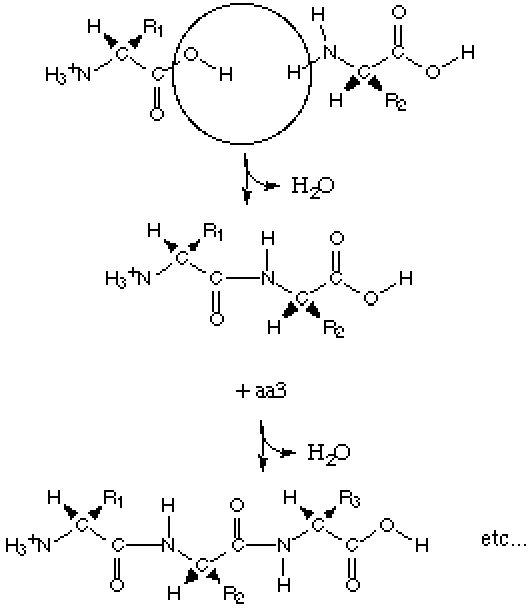
Proteins – our molecular machines (samples of protein tasks)

- Catalysis (enzymes).
- Signal propagation.
- Transport.
- Storage.
- Receptors (e.g. antibodies – immune system).
- Structural proteins (hair, skin, nails).

Amino acids and the peptide bond

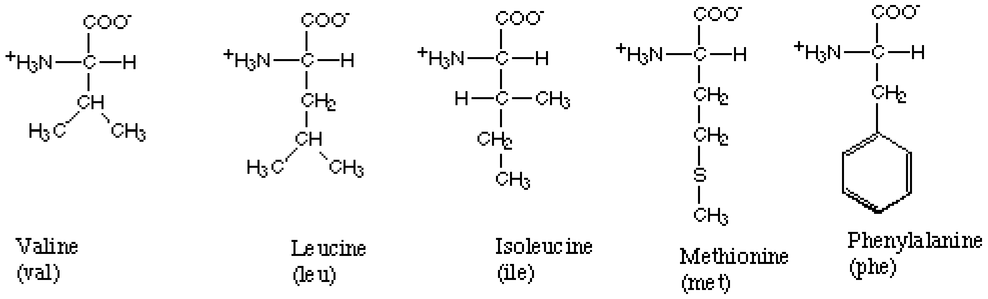






C_β – first side chain carbon (except for glycine).

Amino acids with hydrophobic side groups

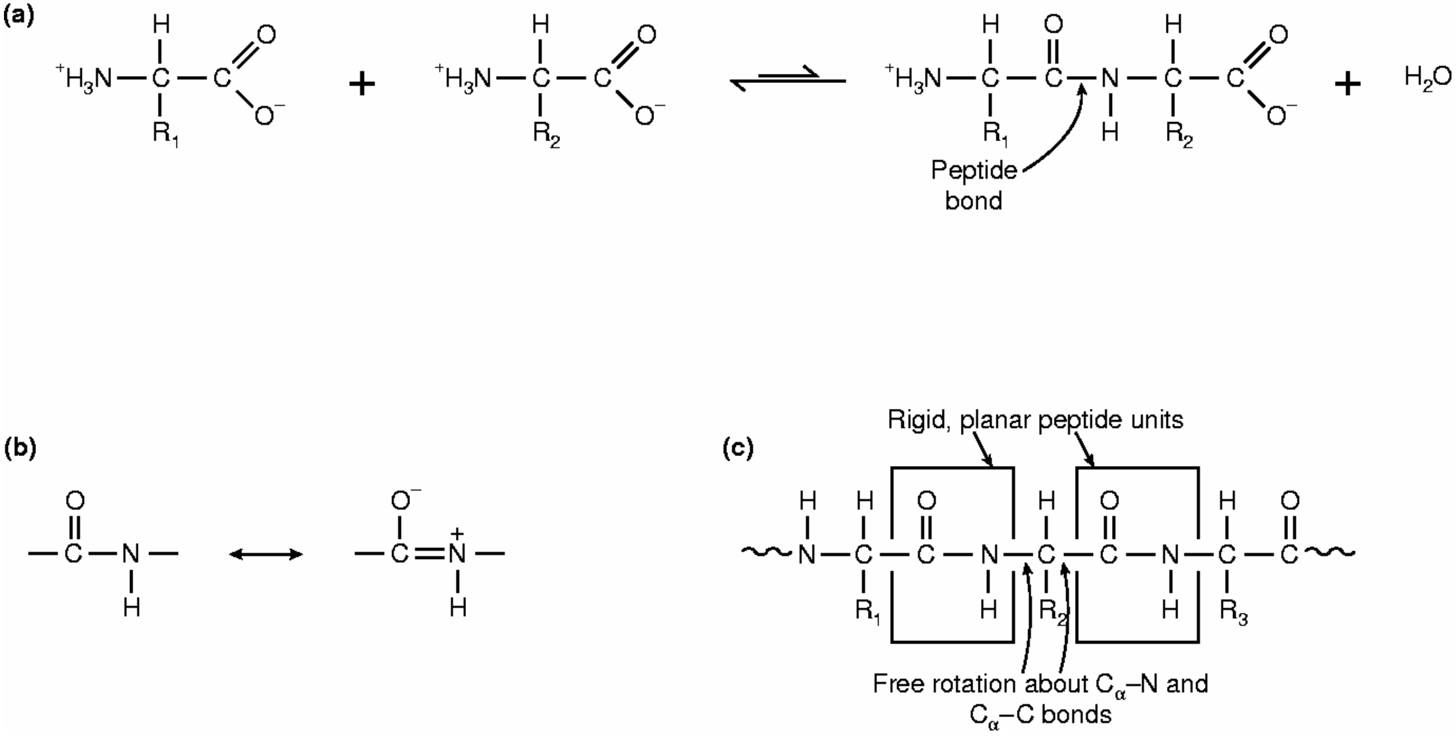


Primary through Quaternary structure

- Primary structure: The order of the amino acids composing the protein.
- AASGDXSLVEVHXXVFIVPPXIL…..

Folding of the Protein Backbone

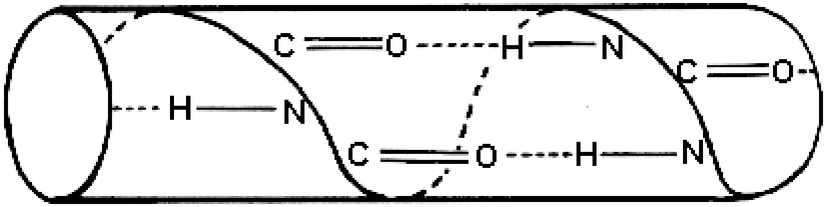
B2.1


(a)
^+H_3N
R_1
R_2
+
H_2O
Peptide bond
(b)
(c)
Rigid, planar peptide units
R_3
Free rotation about C_α–N and C_α–C bonds

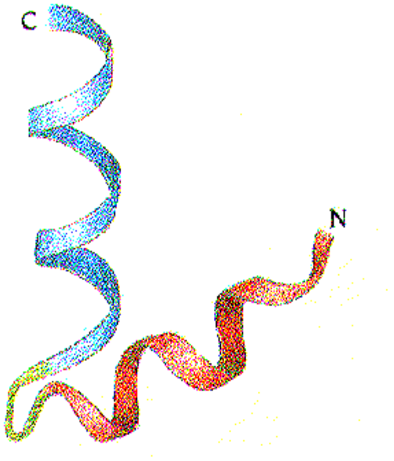
The Holy Grail - Protein Folding

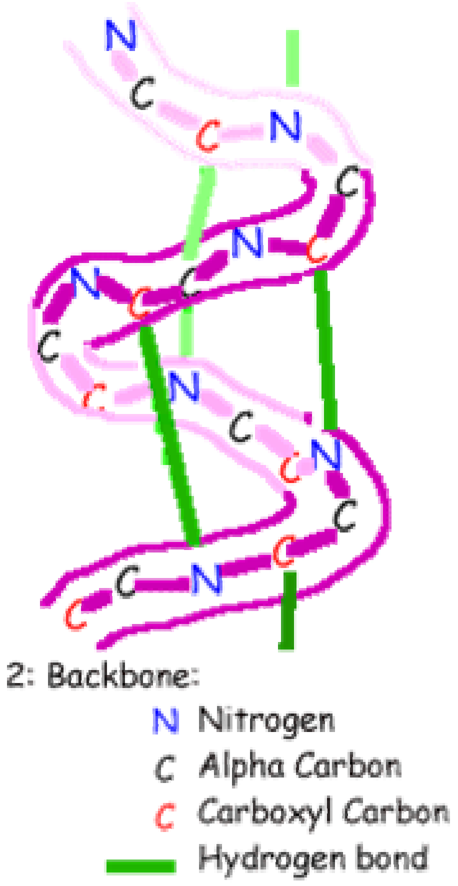
- How does a protein “know” its 3-D structure ?
- How does it compute it so fast ?
- Relatively primitive computational folding models have proved to be NP complete even in the 2-D case.

Secondary structure

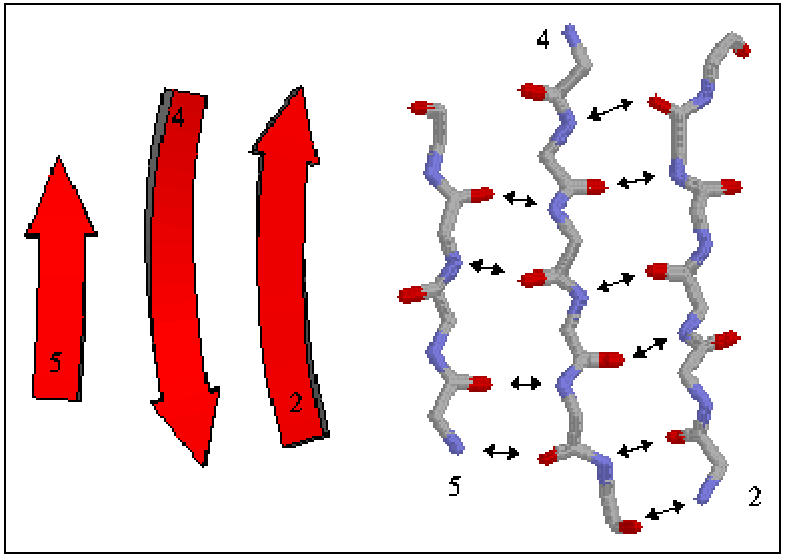


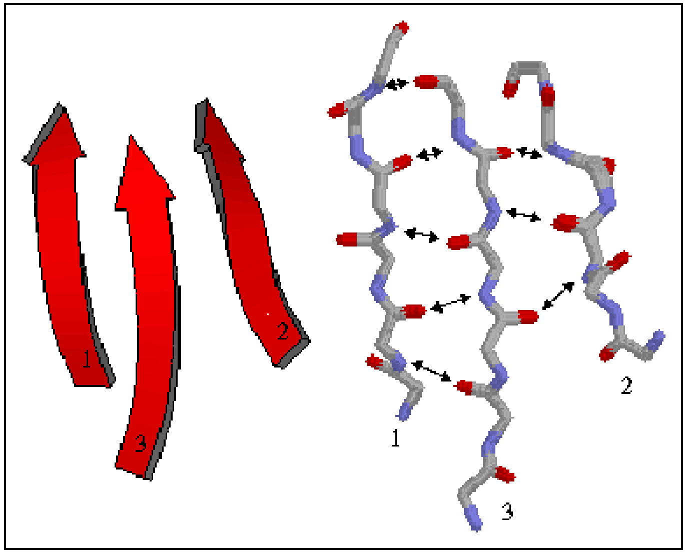
3.6 residues/turn (5.4 A dist.)



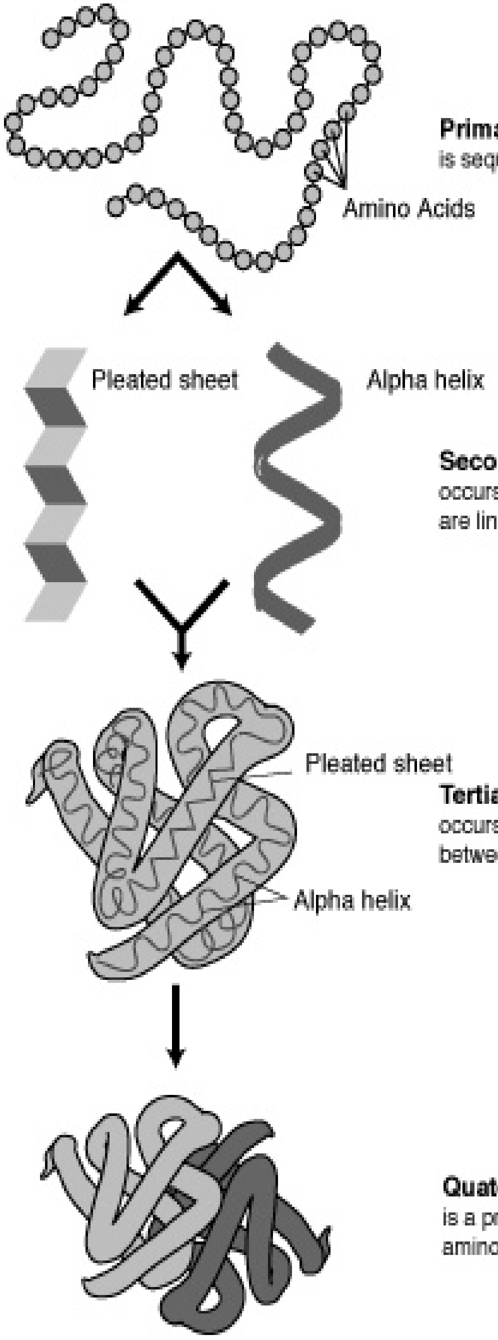


β strands and sheets





Bond.   Hydrogen bond.



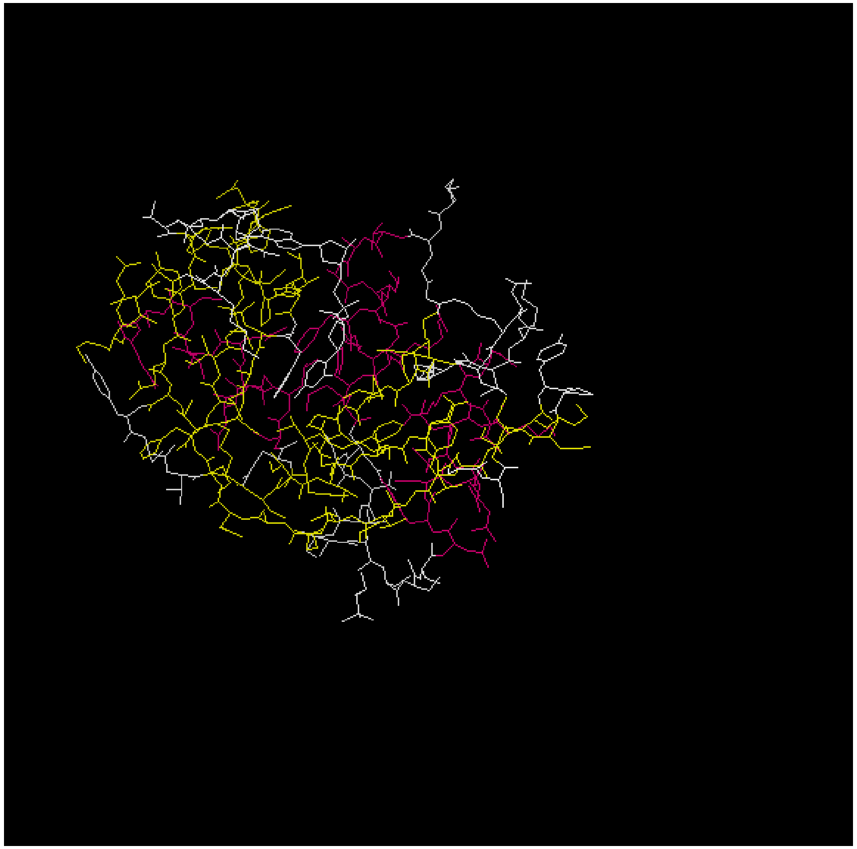
Primary protein structure
is sequence of a chain of amino acids

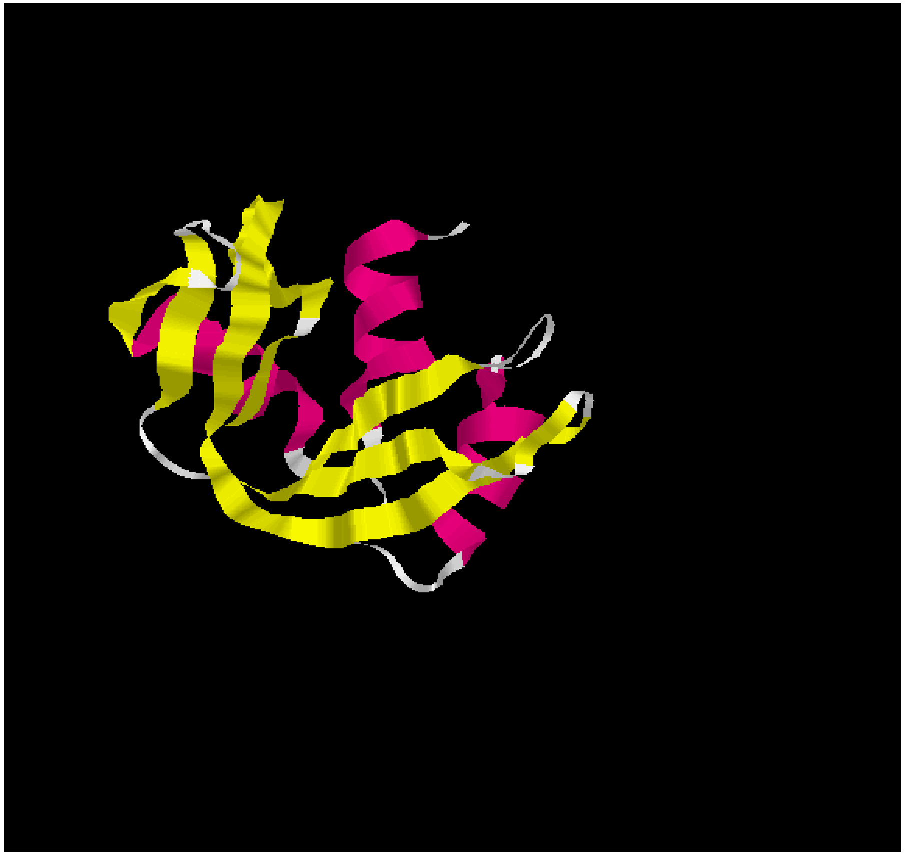
Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds

Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.

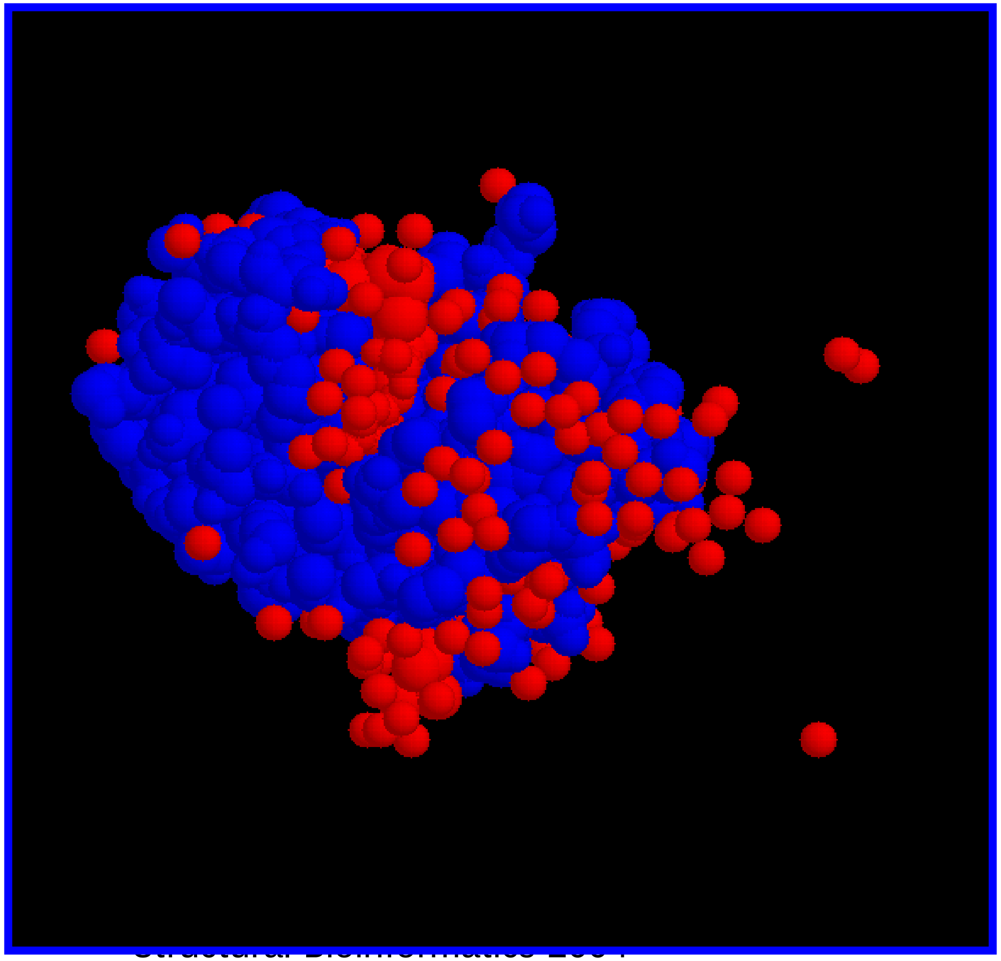
Quaternary protein structure
is a protein consisting of more than one amino acid chain.

Wire-frame or ribbons display



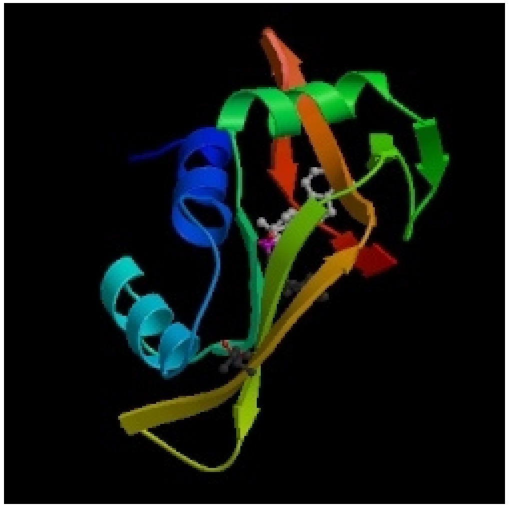


Space-fill display



Tertiary structure: full 3D folded structure of the polypeptide chain

Ribonuclease - PDB code 1rpg

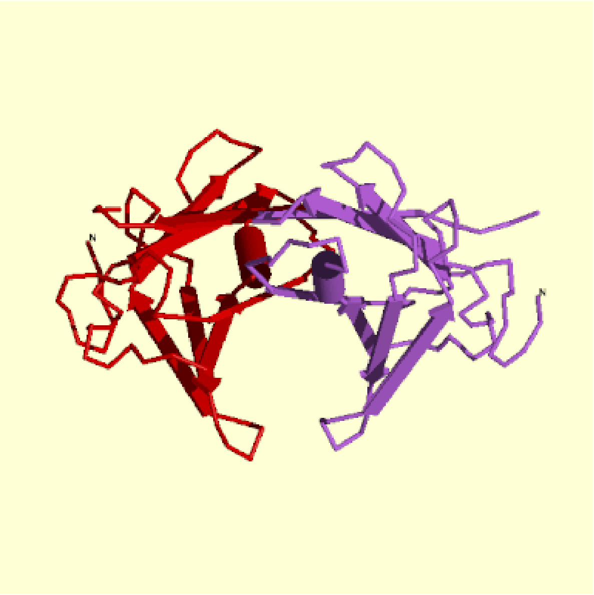


Quaternary structure

The interconnections and organization of more than one polypeptide chain.

Example :Transthyretin

dimer **(1tta)**



Determination of protein structures

- X-ray Crystallography
- NMR (Nuclear Magnetic Resonance)
- EM (Electron microscopy)
- Nano – sensors (?)

X-ray Crystallography

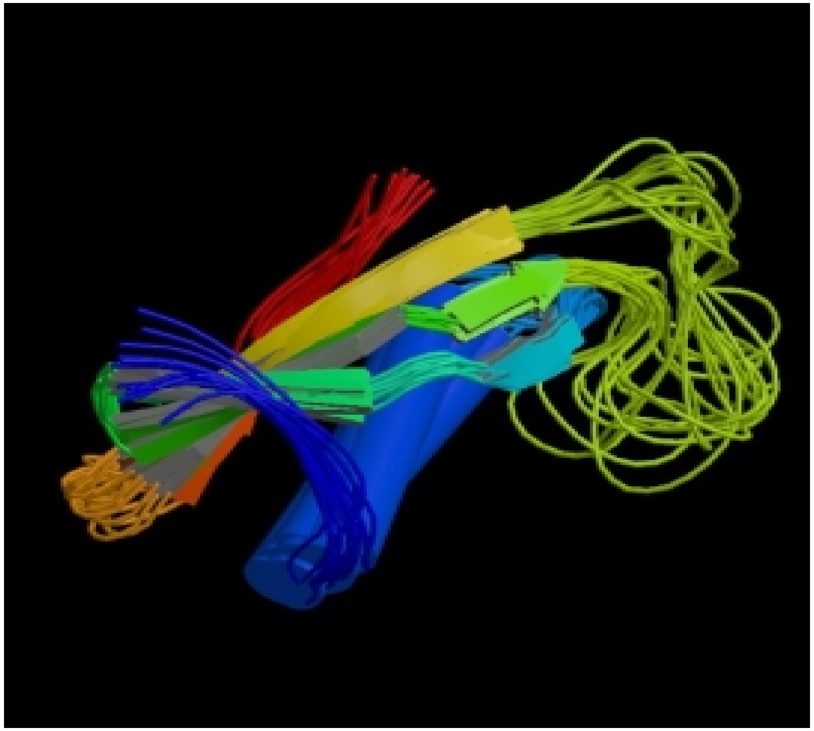
- Crystallization
- Each protein has a unique **X-ray pattern diffraction.**
- The **electron density map** is used to build a model of the protein.

Nuclear Magnetic Resonance

- Performed in an **aqueous solution**.
- NMR analysis gives **a set of estimates of distances between specific pairs of protons (H – atoms).**
- Solved by Distance Geometry methods.
- The result is an **ensemble of models** rather than a single structure.

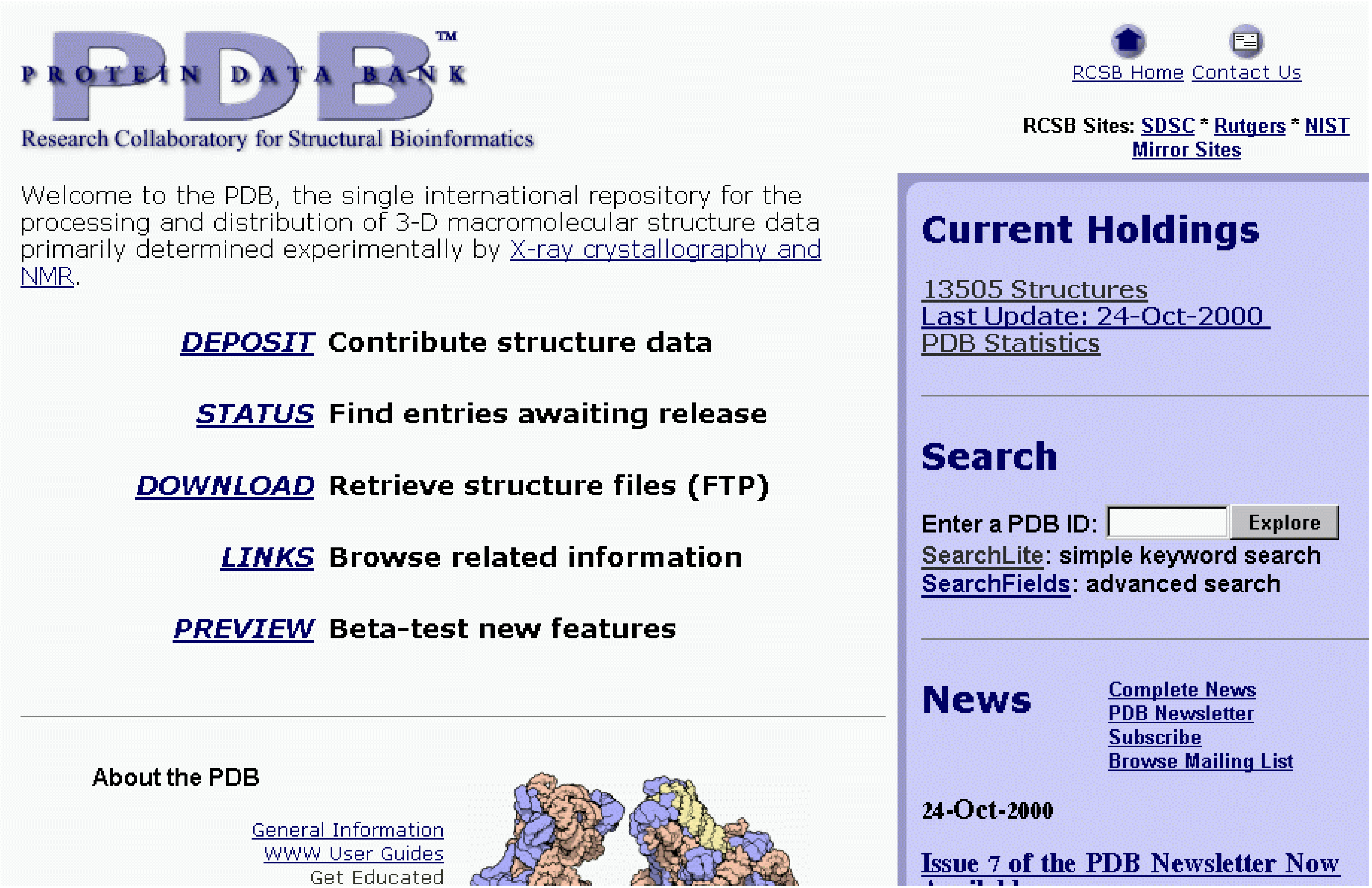
An NMR result is an ensemble of models

Cystatin (**1a67**)



The Protein Data Bank (PDB)

- International repository of 3D molecular data.
- Contains x-y-z coordinates of all atoms of the molecule and additional data.



PROTEIN DATA BANK PDB™

Research Collaboratory for Structural Bioinformatics

RCSB Home Contact Us

RCSB Sites: SDSC * Rutgers * NIST
Mirror Sites

Welcome to the PDB, the single international repository for the processing and distribution of 3-D macromolecular structure data primarily determined experimentally by X-ray crystallography and NMR.

DEPOSIT Contribute structure data

STATUS Find entries awaiting release

DOWNLOAD Retrieve structure files (FTP)

LINKS Browse related information

PREVIEW Beta-test new features

About the PDB

General Information
WWW User Guides
Get Educated

Current Holdings

13505 Structures
Last Update: 24-Oct-2000
PDB Statistics

Search

Enter a PDB ID: ▭ Explore

SearchLite: simple keyword search
SearchFields: advanced search

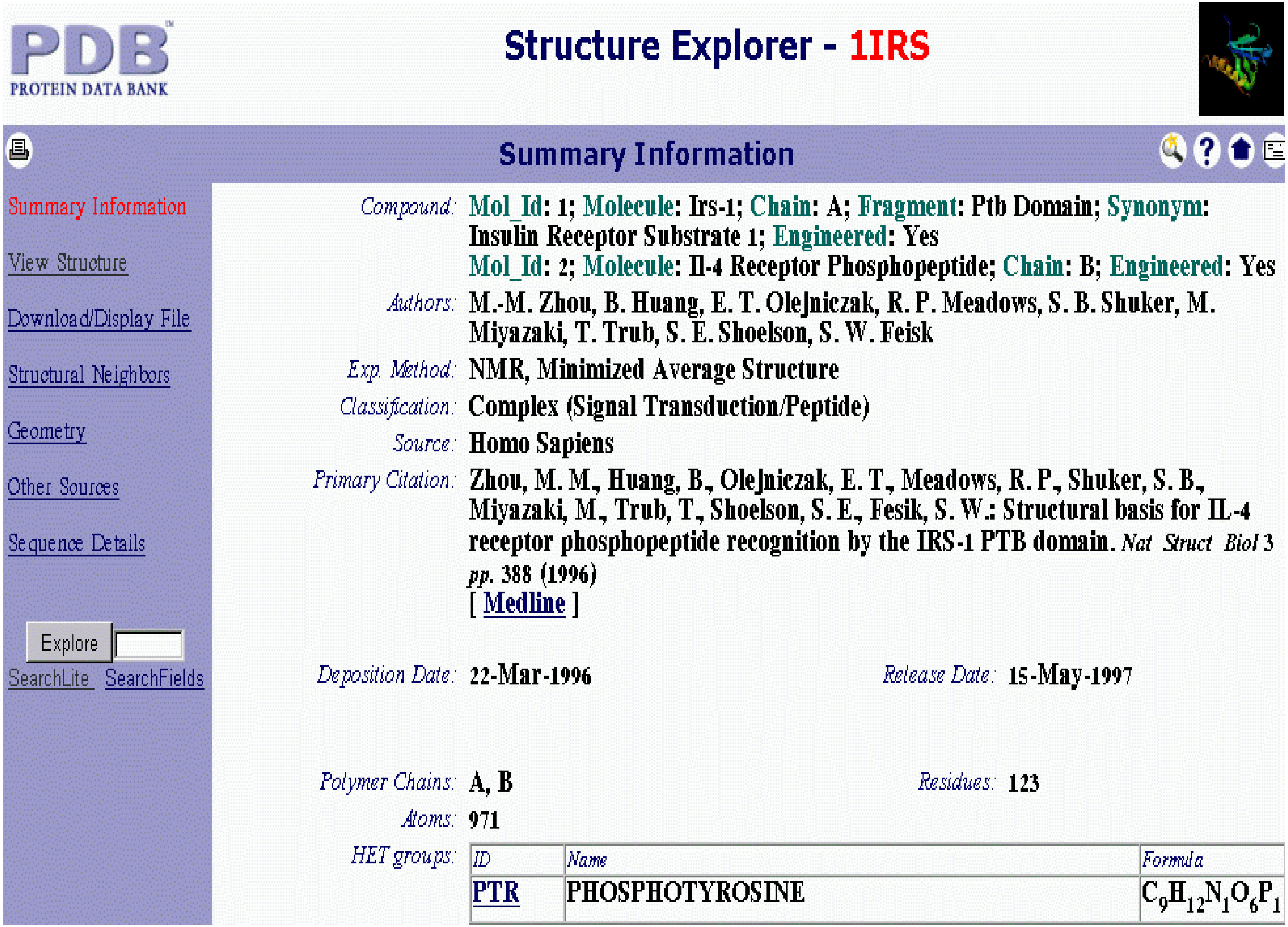
News

Complete News
PDB Newsletter
Subscribe
Browse Mailing List

24-Oct-2000

Issue 7 of the PDB Newsletter Now

Feb. 2003 – about 20,000 structures.



Structure Explorer - 1IRS

Summary Information

Compound: Mol_Id: 1; Molecule: Irs-1; Chain: A; Fragment: Ptb Domain; Synonym: Insulin Receptor Substrate 1; Engineered: Yes
Mol_Id: 2; Molecule: Il-4 Receptor Phosphopeptide; Chain: B; Engineered: Yes

Authors: M.-M. Zhou, B. Huang, E. T. Olejniczak, R. P. Meadows, S. B. Shuker, M. Miyazaki, T. Trub, S. E. Shoelson, S. W. Feisk

Exp. Method: NMR, Minimized Average Structure

Classification: Complex (Signal Transduction/Peptide)

Source: Homo Sapiens

Primary Citation: Zhou, M. M., Huang, B., Olejniczak, E. T., Meadows, R. P., Shuker, S. B., Miyazaki, M., Trub, T., Shoelson, S. E., Fesik, S. W.: Structural basis for IL-4 receptor phosphopeptide recognition by the IRS-1 PTB domain. *Nat Struct Biol* 3 *pp.* 388 (1996)
[Medline]

Deposition Date: 22-Mar-1996 *Release Date:* 15-May-1997

Polymer Chains: A, B *Residues:* 123

Atoms: 971

HET groups:

ID	*Name*	*Formula*
PTR	PHOSPHOTYROSINE	$C_9H_{12}N_1O_6P_1$

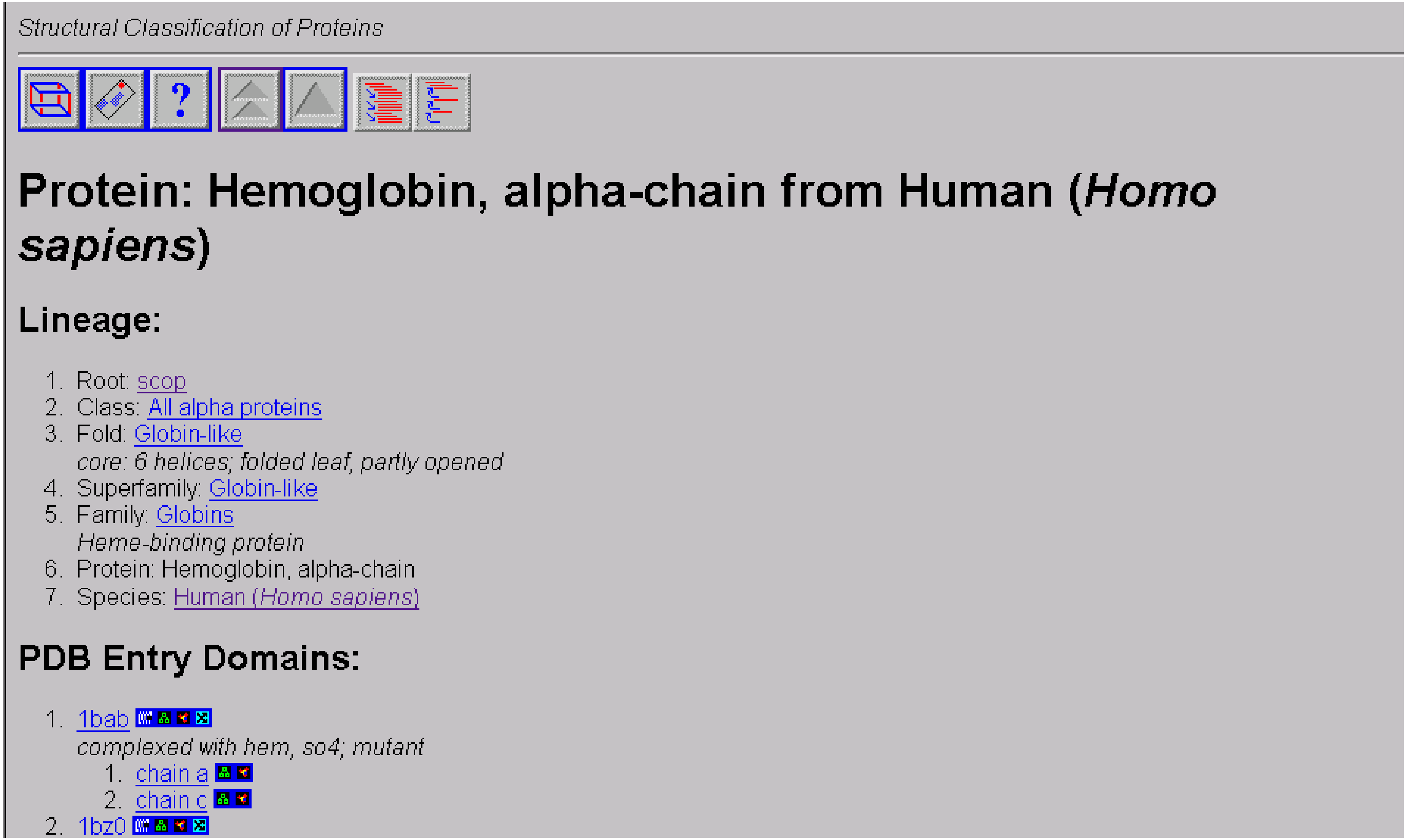
Classification of 3D structures

SCOP

- Provides a description of the structural and evolutionary relationships between all proteins whose structure is known.
- Created largely by manual inspection.

- J. Mol. Biol. 247, 536-540, 1995

SCOP



Structural Classification of Proteins

Protein: Hemoglobin, alpha-chain from Human (*Homo sapiens*)

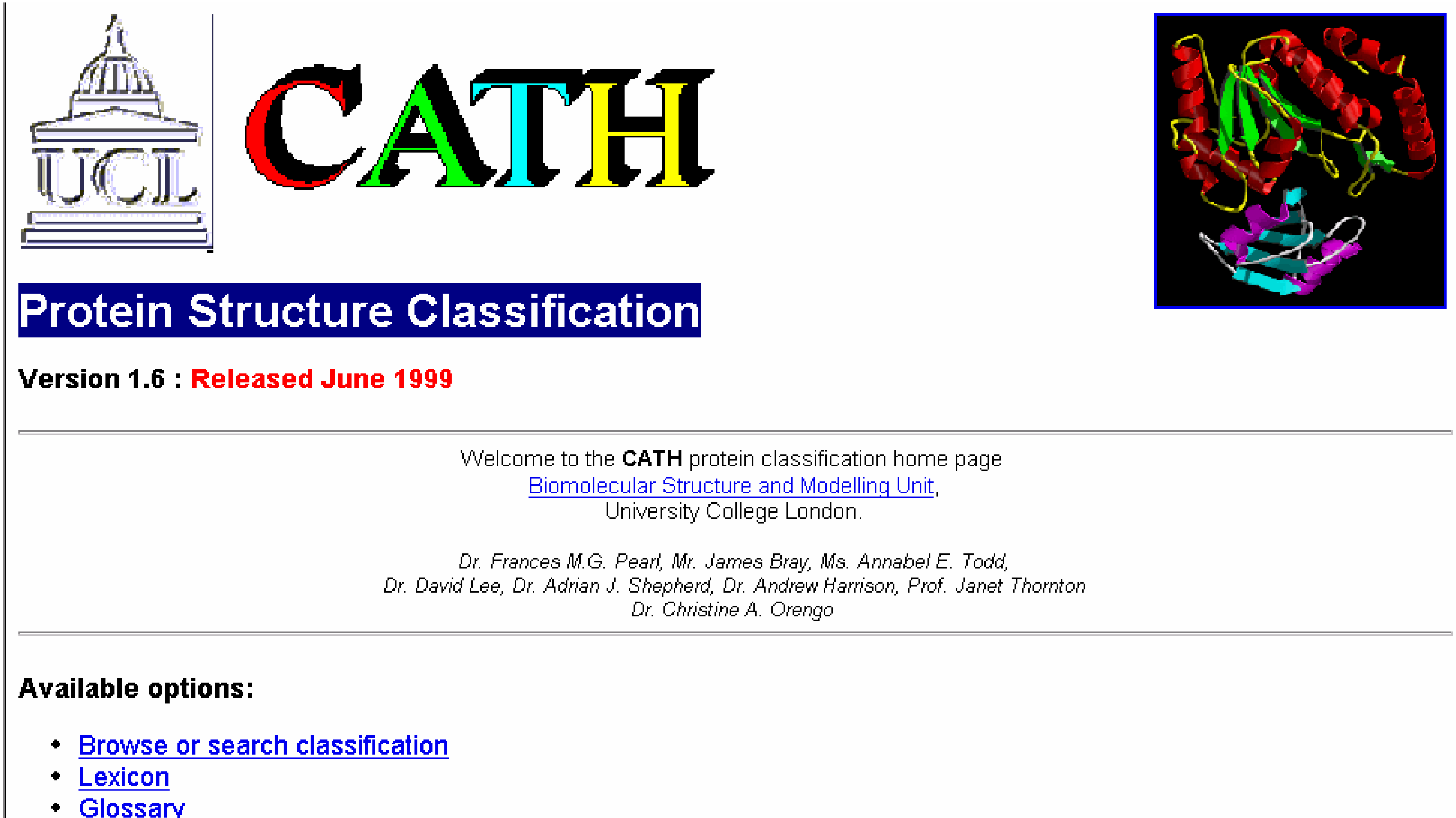
Lineage:

1. Root: scop
2. Class: All alpha proteins
3. Fold: Globin-like
   core: 6 helices; folded leaf, partly opened
4. Superfamily: Globin-like
5. Family: Globins
   Heme-binding protein
6. Protein: Hemoglobin, alpha-chain
7. Species: Human (*Homo sapiens*)

PDB Entry Domains:

1. 1bab
   complexed with hem, so4; mutant
   1. chain a
   2. chain c
2. 1bz0

CATH - Protein Structure Classification
http://www.biochem.ucl.ac.uk/bsm/cath/



Protein Structure Classification

Version 1.6 : Released June 1999

Welcome to the **CATH** protein classification home page
Biomolecular Structure and Modelling Unit,
University College London.

Dr. Frances M.G. Pearl, Mr. James Bray, Ms. Annabel E. Todd,
Dr. David Lee, Dr. Adrian J. Shepherd, Dr. Andrew Harrison, Prof. Janet Thornton
Dr. Christine A. Orengo

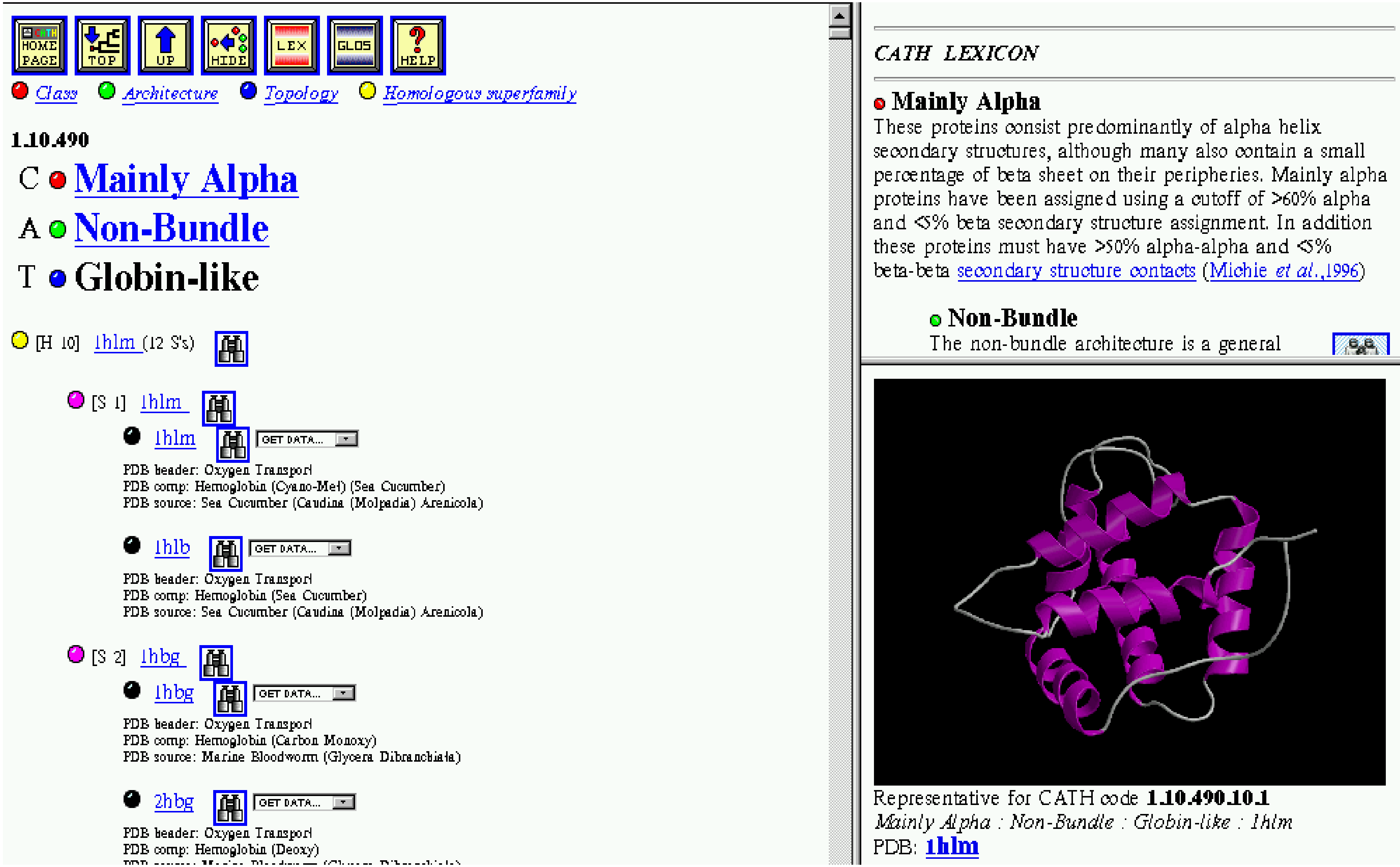
Available options:

- Browse or search classification
- Lexicon
- Glossary

CATH

- Class: derived from secondary structure content.
- Architecture: gross orientation of secondary structures, independent of connectivities.
- Topology: clusters according to topological connections and numbers of secondary structures.

- Homology: clusters according to structure and function.



● Class ● Architecture ● Topology ● Homologous superfamily

1.10.490

C ● Mainly Alpha

A ● Non-Bundle

T ● Globin-like

● [H 10] 1hlm (12 S's)

● [S 1] 1hlm

● 1hlm GET DATA...

PDB header: Oxygen Transport
PDB comp: Hemoglobin (Cyano-Met) (Sea Cucumber)
PDB source: Sea Cucumber (Caudina (Molpadia) Arenicola)

● 1hlb GET DATA...

PDB header: Oxygen Transport
PDB comp: Hemoglobin (Sea Cucumber)
PDB source: Sea Cucumber (Caudina (Molpadia) Arenicola)

● [S 2] 1hbg

● 1hbg GET DATA...

PDB header: Oxygen Transport
PDB comp: Hemoglobin (Carbon Monoxy)
PDB source: Marine Bloodworm (Glycera Dibranchiata)

● 2hbg GET DATA...

PDB header: Oxygen Transport
PDB comp: Hemoglobin (Deoxy)
PDB source: Marine Bloodworm (Glycera Dibranchiata)

CATH LEXICON

● Mainly Alpha

These proteins consist predominantly of alpha helix secondary structures, although many also contain a small percentage of beta sheet on their peripheries. Mainly alpha proteins have been assigned using a cutoff of >60% alpha and <5% beta secondary structure assignment. In addition these proteins must have >50% alpha-alpha and <5% beta-beta secondary structure contacts (Michie *et al.*,1996)

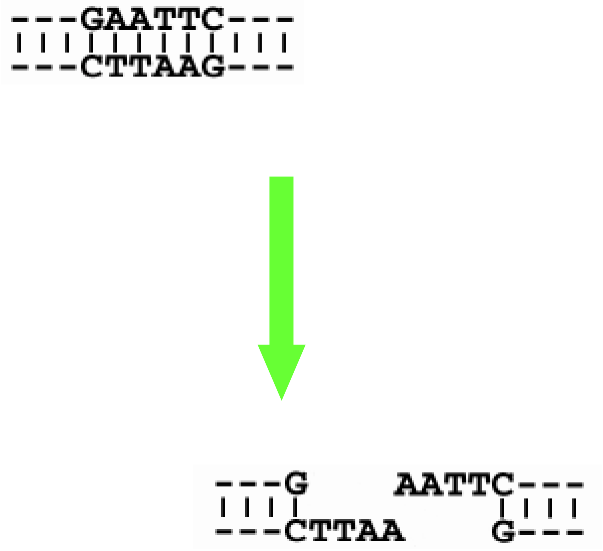
● Non-Bundle

The non-bundle architecture is a general

Representative for CATH code **1.10.490.10.1**
Mainly Alpha : Non-Bundle : Globin-like : 1hlm
PDB: **1hlm**

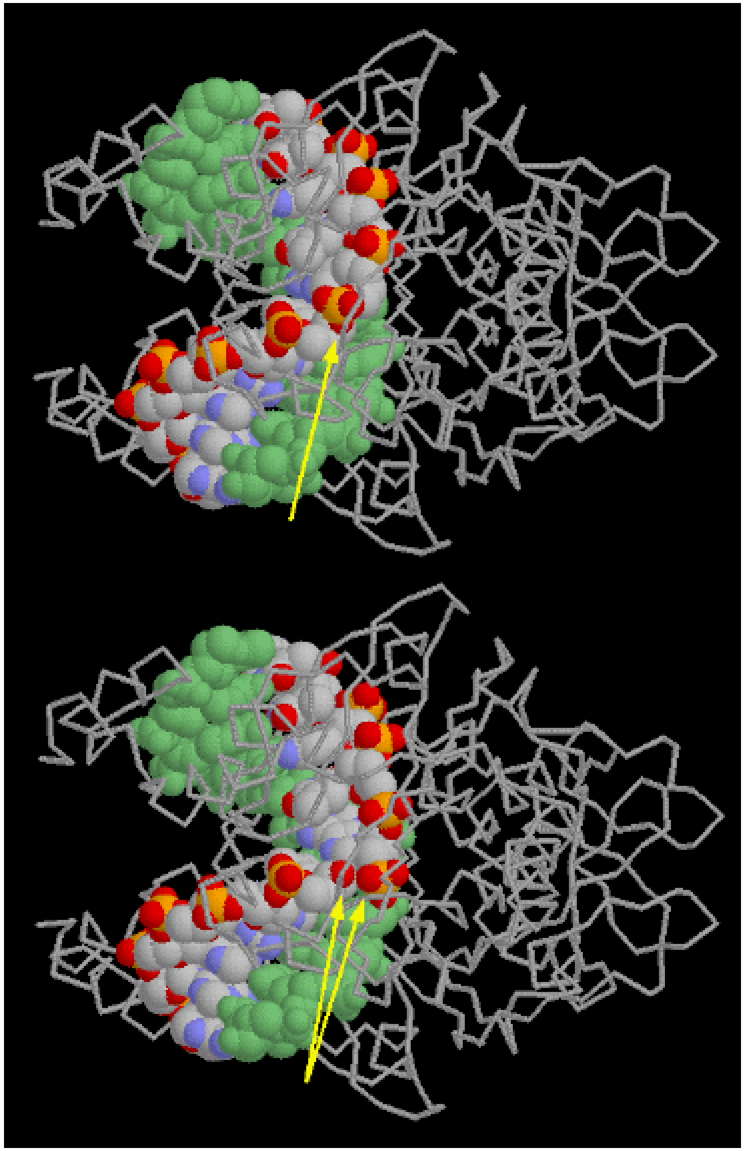
- PDB **http://pdb.tau.ac.il**
- PDB **http://www.rcsb.org/pdb/**
- CATH
  http://www.biochem.ucl.ac.uk/bsm/cath/
- SCOP **http://scop.mrc-lmb.cam.ac.uk/scop/**

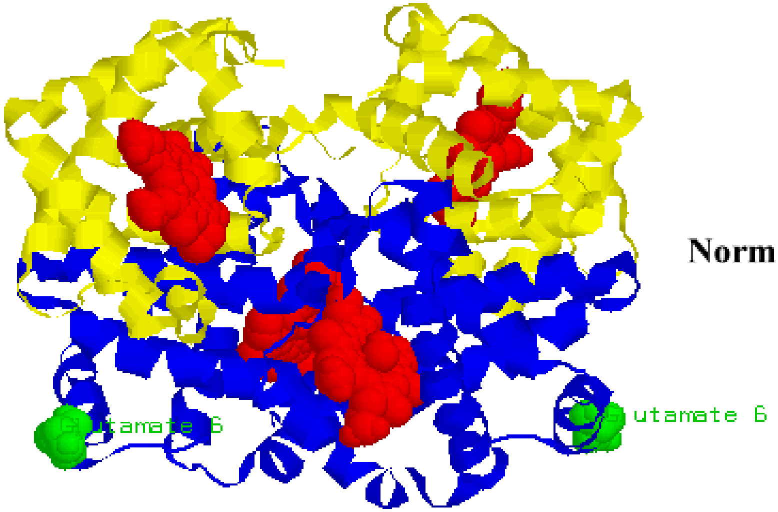
Restriction enzymes



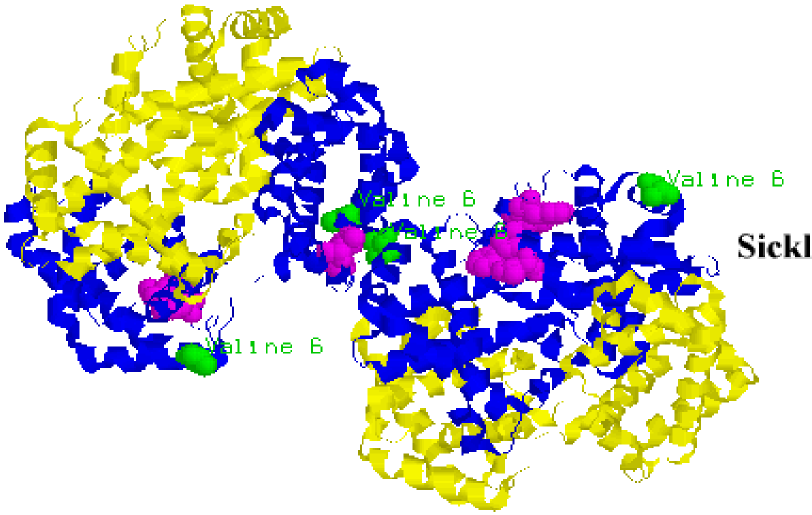
```
---GAATTC---
||||||||||||
---CTTAAG---
```

```
---G     AATTC---
||||         ||||
---CTTAA     G---
```





Normal Hemoglobin



Sickle Hemoglobin

Note: The Sickle hemoglobin image is drawn at 50% of the size of the Normal hemoglobin

The Structural Genomics Pipeline (X-ray Crystallography)

Basic Steps

Target Selection

Crystallomics
- Isolation,
- Expression,
- Purification,
- Crystallization

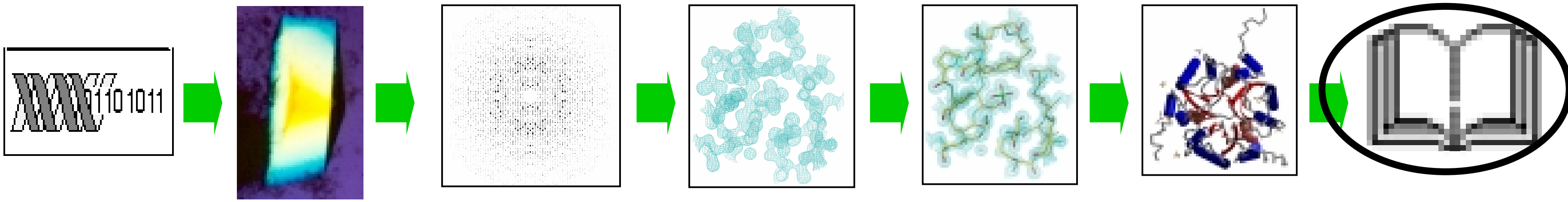
Data Collection

Structure Solution

Structure Refinement

Functional Annotation

Publish



Bioinformatics
- Distant homologs
- Domain recognition

Automation
Bioinformatics
- Empirical rules

Automation
Better sources

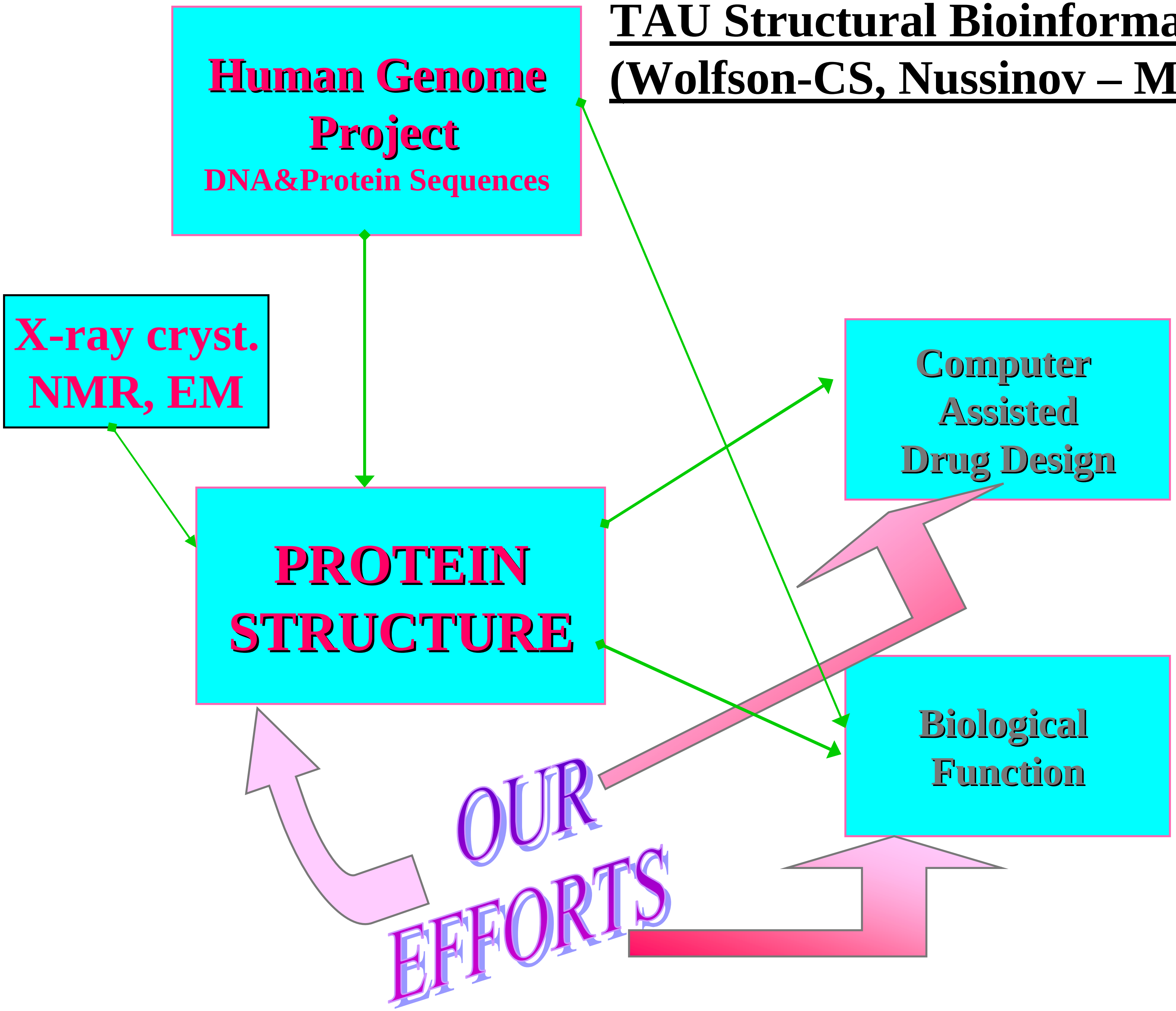
Software integration
Decision Support
MAD Phasing Automated fitting

Bioinformatics
- Alignments
- Protein-protein interactions
- Protein-ligand interactions
- Motif recognition

No?

Borrowed from Bourne’s (UCSD) lecture on CADD

TAU Structural Bioinformatics Lab (Wolfson-CS, Nussinov – MB)



Human Genome Project

DNA&Protein Sequences

X-ray cryst. NMR, EM

PROTEIN STRUCTURE

Computer Assisted Drug Design

Biological Function

OUR EFFORTS

Structural Bioinformatics Lab Goals

Development of *state of the art* algorithmic methods to tackle major computational tasks in protein structure analysis, biomolecular recognition, and *Computer Assisted Drug Design*.

Establish truly *interdisciplinary* collaboration between Life and Computer Sciences.

Bioinformatics and Genomics - Economic Impact

- **Medicine and public health.**
- **Pharmaceutics.**
- **Agriculture.**
- **Food industry.**
- **Biological Computers (?).**

Bioinformatics and Genomics - the Computational Viewpoint

- Molecular Biology is becoming a Computational Science.
- The emergence of large databases of DNA, proteins, small molecules and drugs requires computational techniques to analyze the data.
- Efficient CPU and memory intensive algorithms are being developed.
- Many of the computational tasks have analogs in other well established fields of Computer Science allowing cross-fertilization of ideas.

Bioinformatics - Computational Genomics

- DNA mapping.
- Protein or DNA sequence comparisons , *primary structure*.
- Exploration of huge textual databases.
- In essence one- dimensional methods and intuition.
- Graph - theoretic methods.

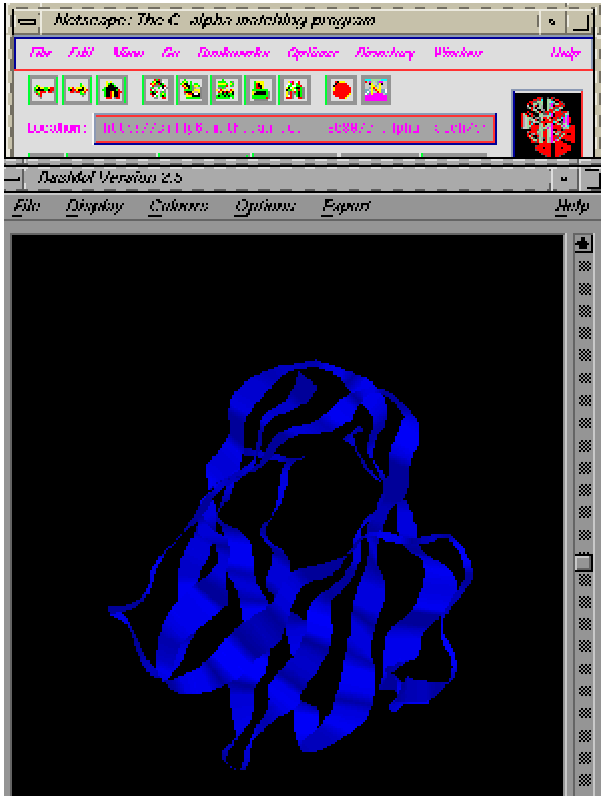
Structural Bioinformatics - Structural Genomics

- **Elucidation of the 3D structures of biomolecules.**
- **Analysis and comparison of biomolecular structures.**
- **Prediction of biomolecular recognition.**
- **Handles three-dimensional (3-D) structures.**
- ***Geometric Computing*.**

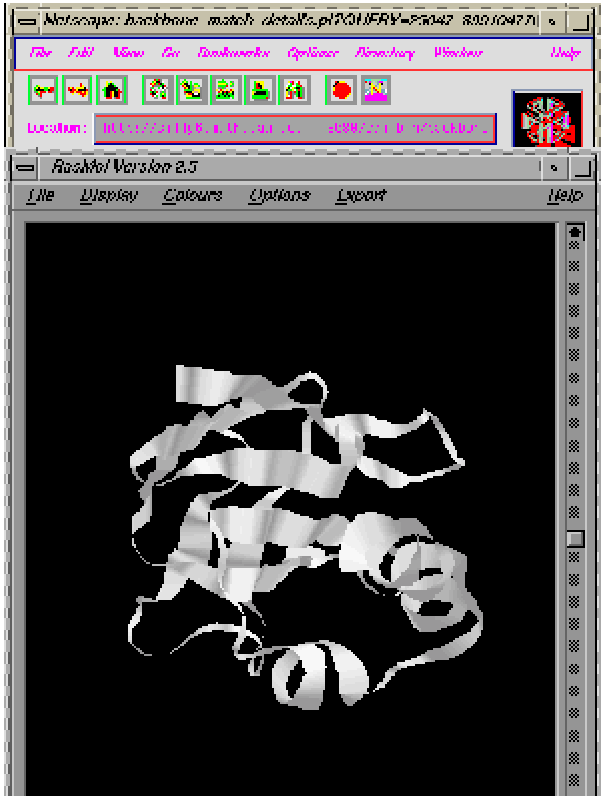
Why bother with structures when we have sequences ?

- In evolutionary related proteins structure is much better preserved than sequence.
- Structural motifs may predict similar biological function .
- Getting insight into protein folding. Recovering the limited (?) number of protein folds.

Case in Point : Protein Structural Comparison


Netscape: The C alpha matching program
File Edit View Go Bookmarks Options Directory Window Help
Location:
RasMol Version 2.5
File Display Colours Options Export Help

ApoAmicyanin - 1aaj


Netscape: backbone_match_details.pl?QUERY=
File Edit View Go Bookmarks Options Directory Window Help
Location:
RasMol Version 2.5
File Display Colours Options Export Help

Pseudoazurin - 1pmy

Geometric Task :

Given two configurations of points in the three dimensional space,

find those rotations and translations of one of the point sets which produce "large" superimpositions of corresponding 3-D points.

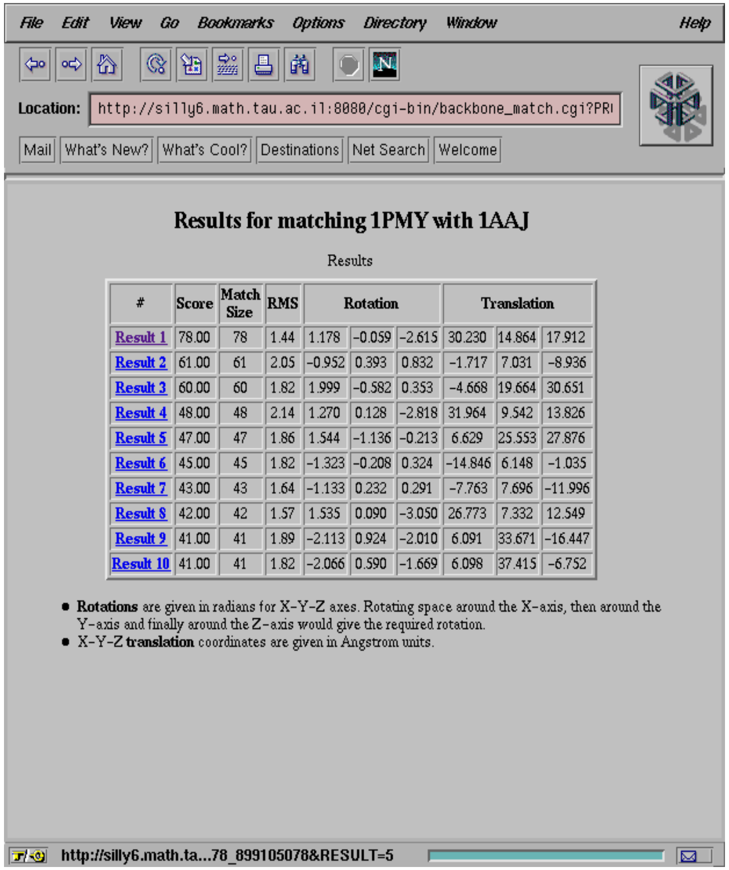
Remarks :

The superimposition pattern is not known a-priori – *pattern discovery* .

The matching recovered can be *inexact*.

We are looking not necessarily for the largest superimposition, since other matchings may have *biological meaning*.

Algorithmic Solution



Location: http://silly6.math.tau.ac.il:8080/cgi-bin/backbone_match.cgi?PR

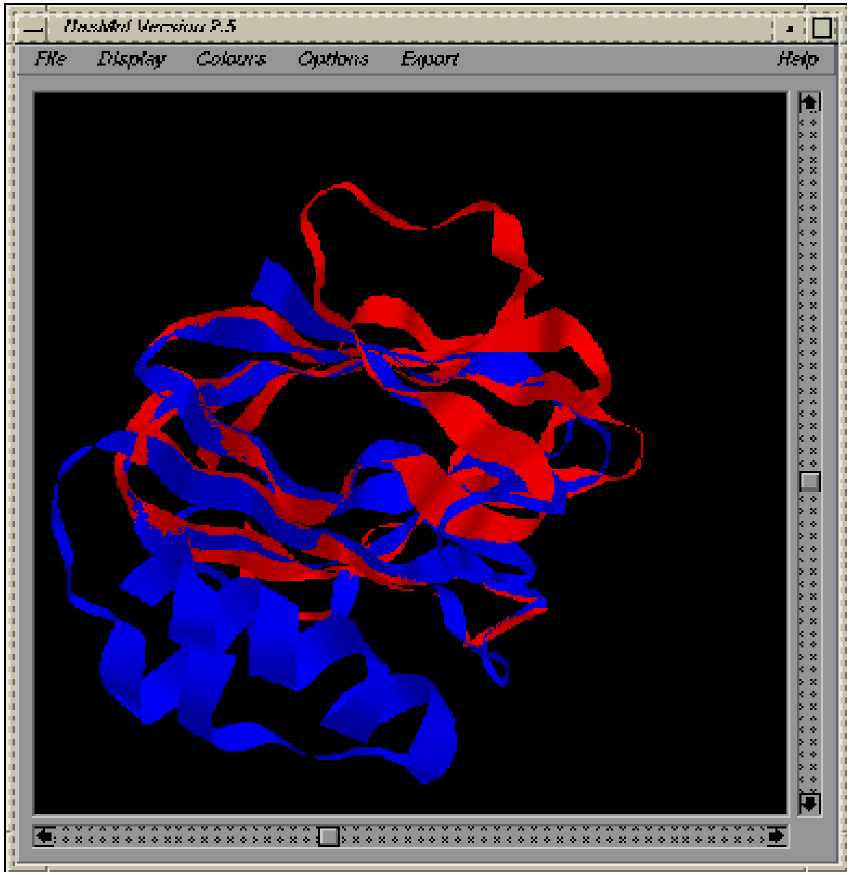
Results for matching 1PMY with 1AAJ

Results

#	Score	Match Size	RMS	Rotation			Translation		
Result 1	78.00	78	1.44	1.178	-0.059	-2.615	30.230	14.864	17.912
Result 2	61.00	61	2.05	-0.952	0.393	0.832	-1.717	7.031	-8.936
Result 3	60.00	60	1.82	1.999	-0.582	0.353	-4.668	19.664	30.651
Result 4	48.00	48	2.14	1.270	0.128	-2.818	31.964	9.542	13.826
Result 5	47.00	47	1.86	1.544	-1.136	-0.213	6.629	25.553	27.876
Result 6	45.00	45	1.82	-1.323	-0.208	0.324	-14.846	6.148	-1.035
Result 7	43.00	43	1.64	-1.133	0.232	0.291	-7.763	7.696	-11.996
Result 8	42.00	42	1.57	1.535	0.090	-3.050	26.773	7.332	12.549
Result 9	41.00	41	1.89	-2.113	0.924	-2.010	6.091	33.671	-16.447
Result 10	41.00	41	1.82	-2.066	0.590	-1.669	6.098	37.415	-6.752

- **Rotations** are given in radians for X-Y-Z axes. Rotating space around the X-axis, then around the Y-axis and finally around the Z-axis would give the required rotation.
- X-Y-Z **translation** coordinates are given in Angstrom units.

http://silly6.math.ta...78_899105078&RESULT=5

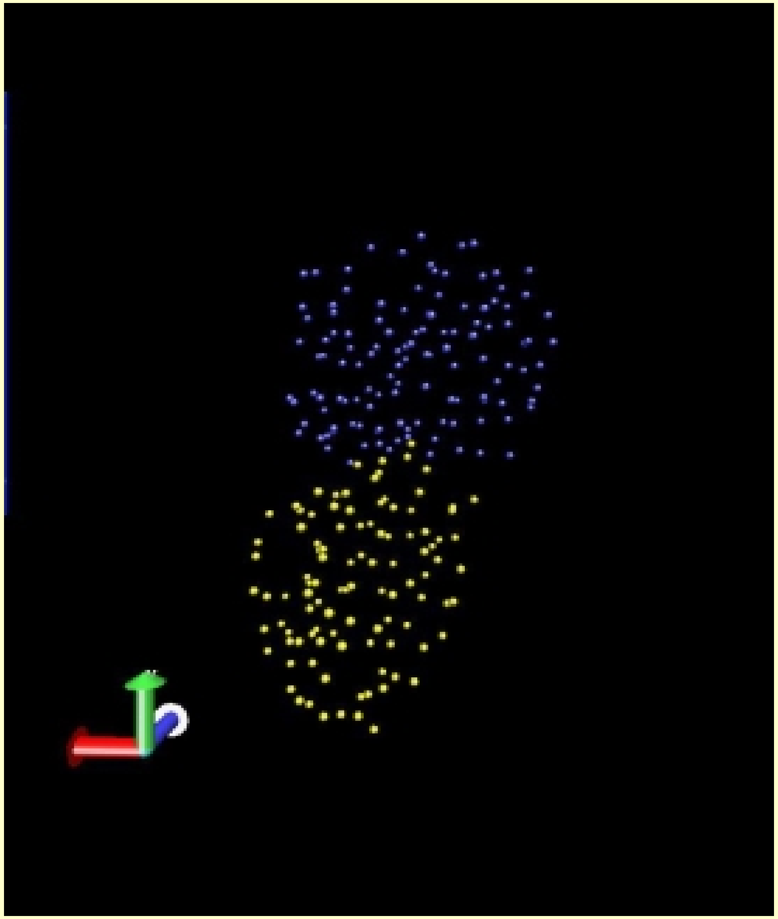


About 1 sec. Fischer, Nussinov, Wolfson ~ 1990.

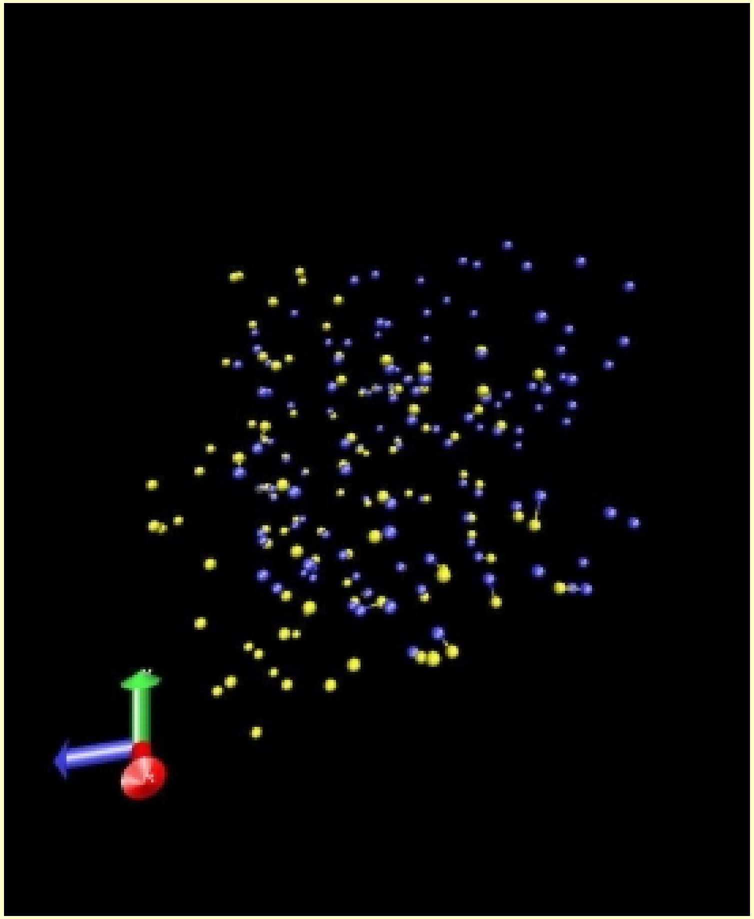
Applications

- **Classification of protein databases by structure.**
- **Search of partial and disconnected structural patterns in large databases.**
- **Detection of structural pharmacophores in an ensemble of drugs.**
- **Comparison and detection of drug receptor *active sites*.**

Geometric Matching task = <u>Geometric Pattern Discovery</u>

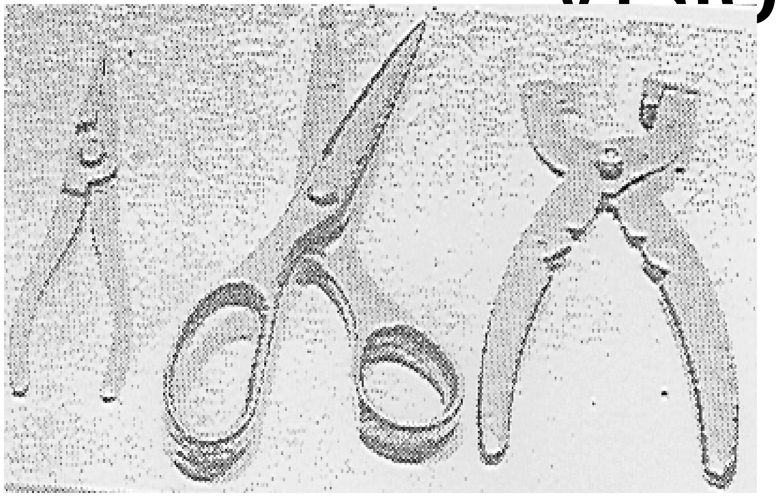


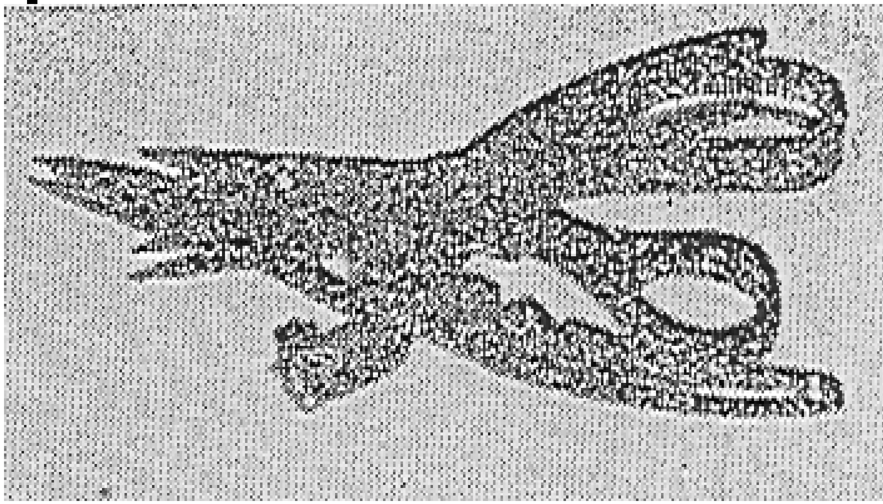
C_{α} constellations - before

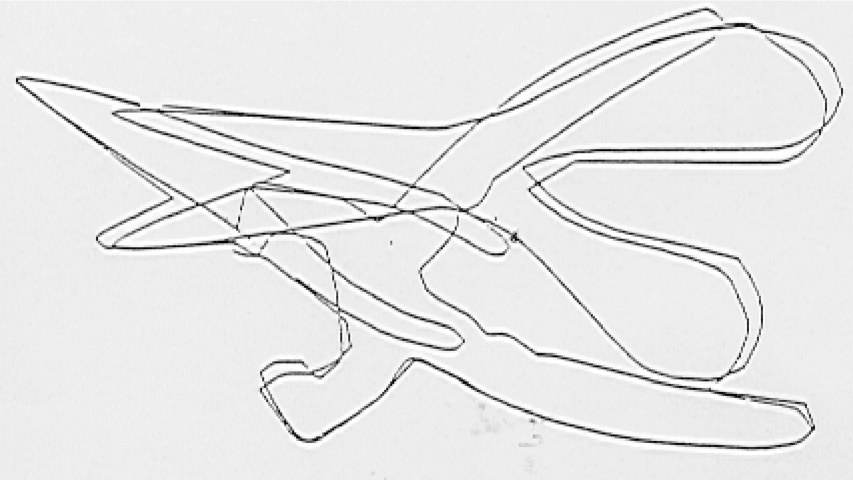


Superimposed constellations

Analogy with Object Recognition in Computer Vision

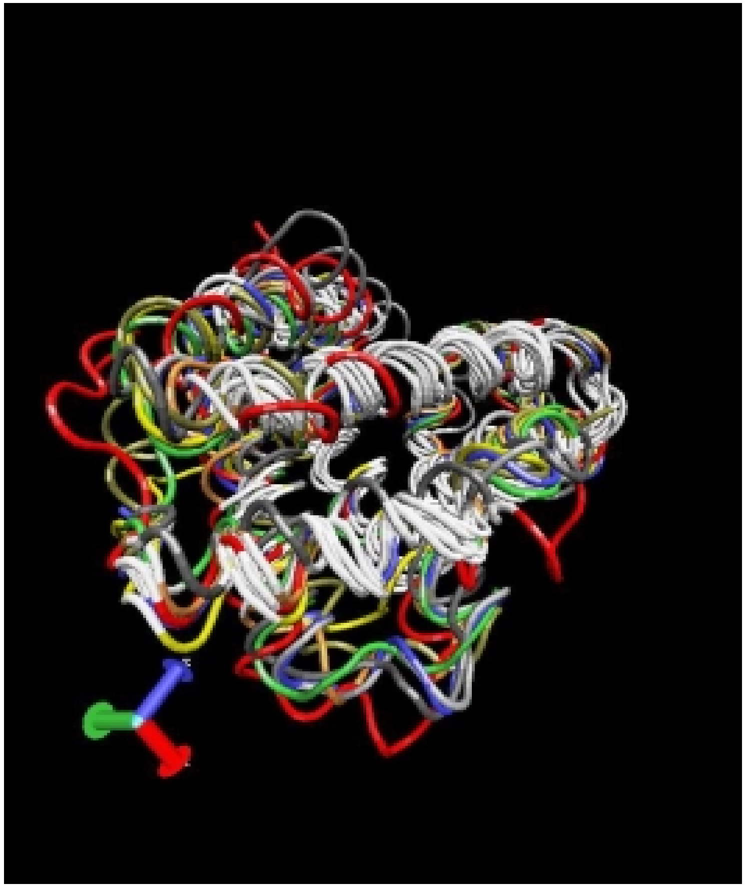


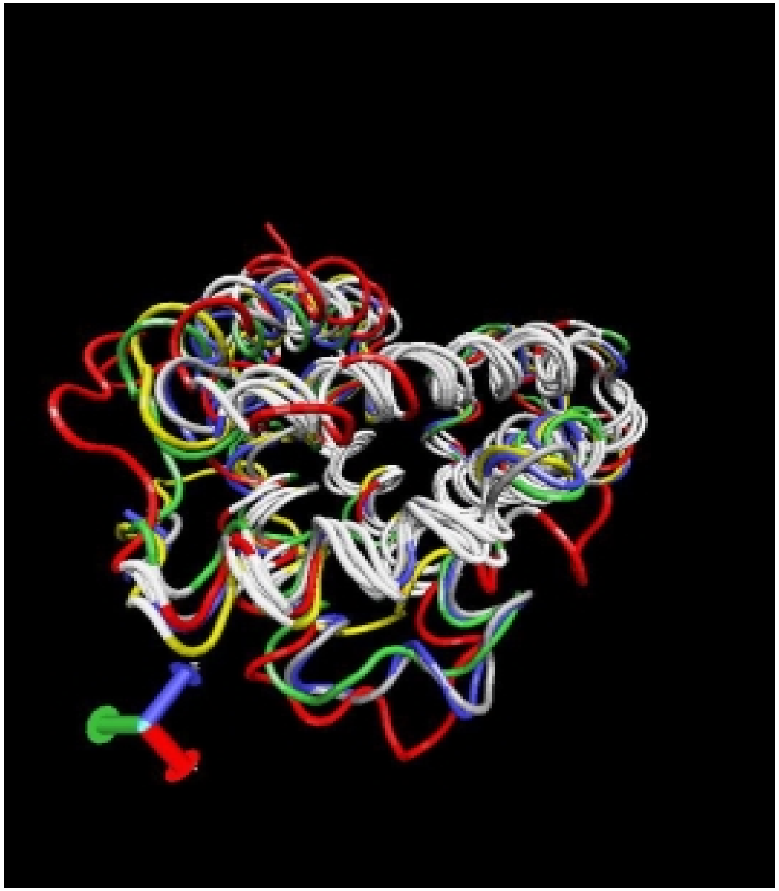




Wolfson, “Curve Matching”,1987.

Multiple Structural Alignment (Globin example)





Leibowitz, Fligelman, Nussinov, Wolfson, - ISMB’99 – Heidelberg.

Biomolecular Recognition - docking

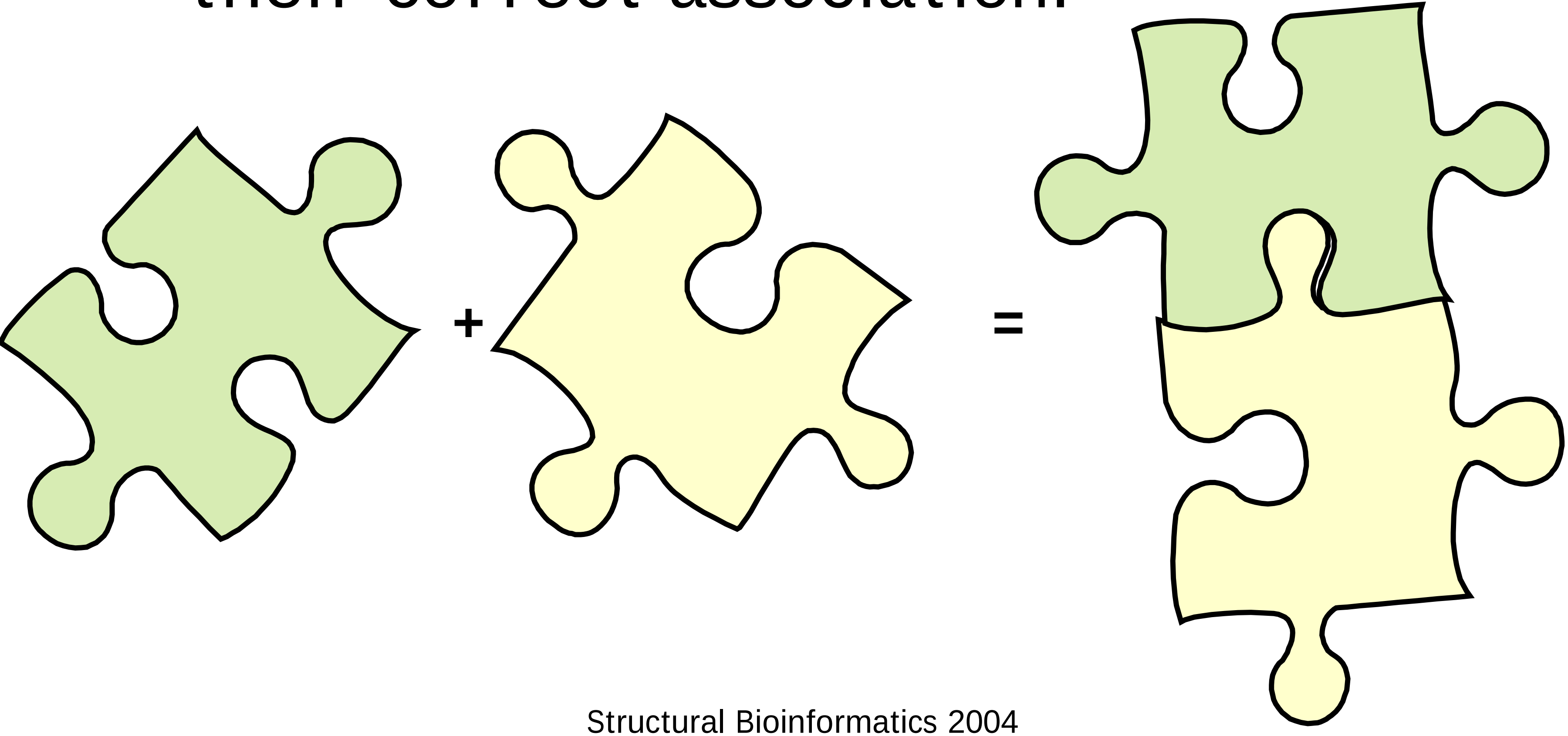
- **Predict association of protein molecules.**
- **Predict binding of a protein molecule with a potential drug.**
- **Scan libraries of drugs to detect a suitable inhibitor for a target molecule.**

Docking Algorithms

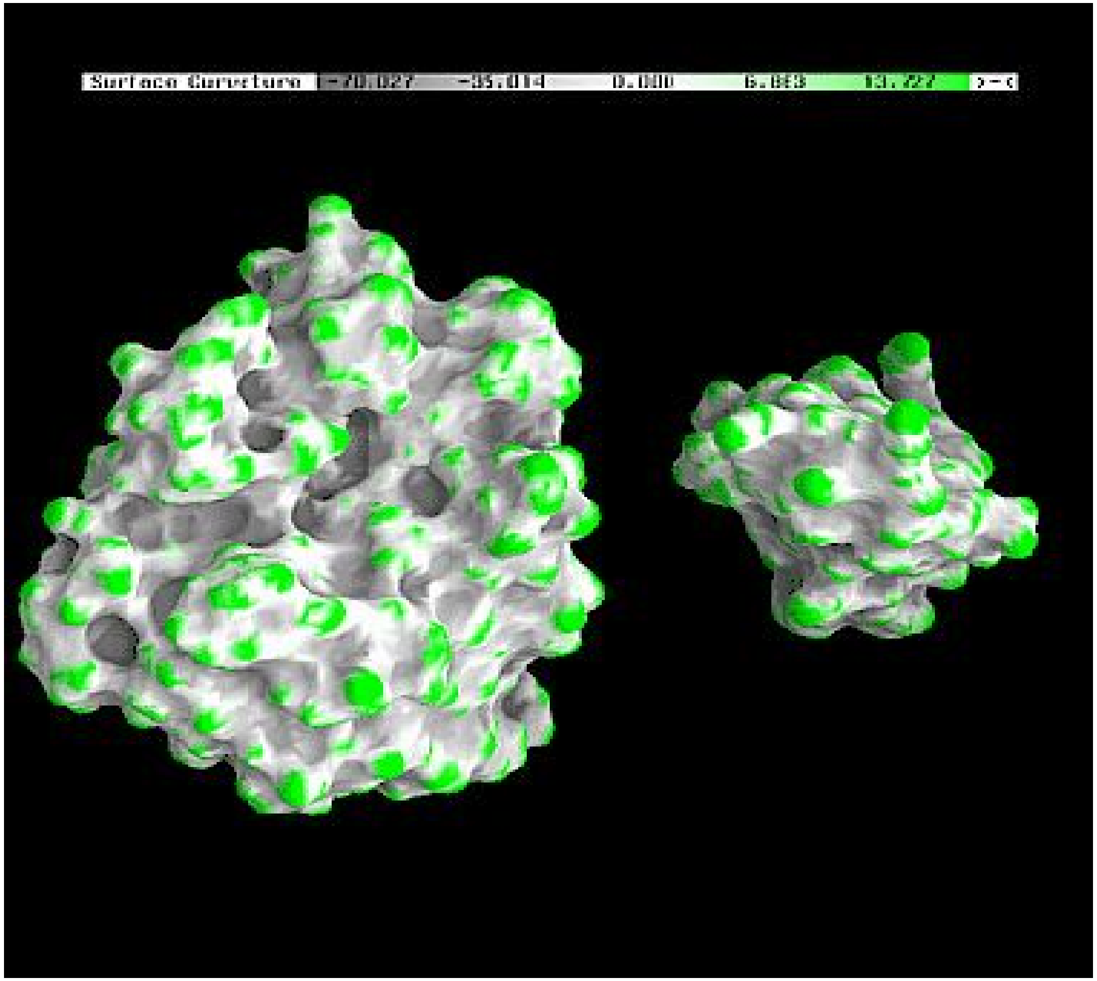
- **Rigid receptor-ligand and protein-protein docking**.
- **Flexible receptor-ligand docking allowing a small number of hinges either in the ligand or the** receptor.

Docking - Problem Definition

- Given a pair of molecules find their correct association:



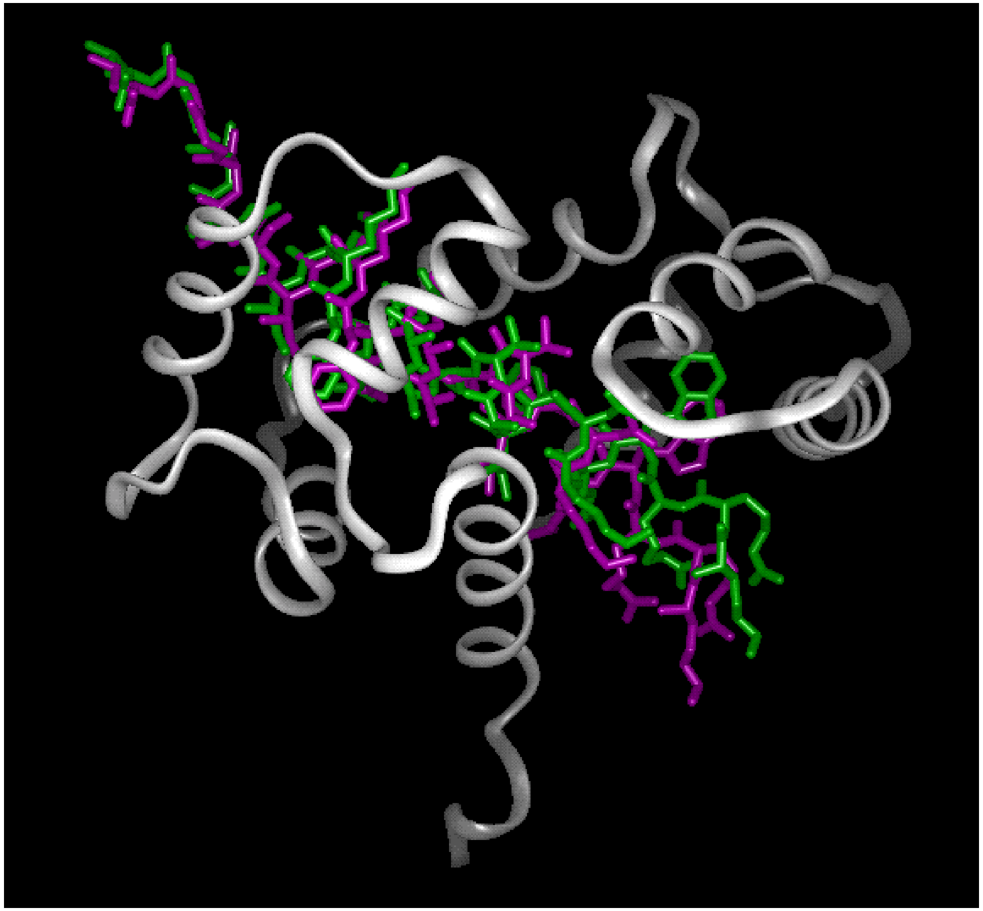
Docking - Trypsin and BPTI



Docking - Relevance

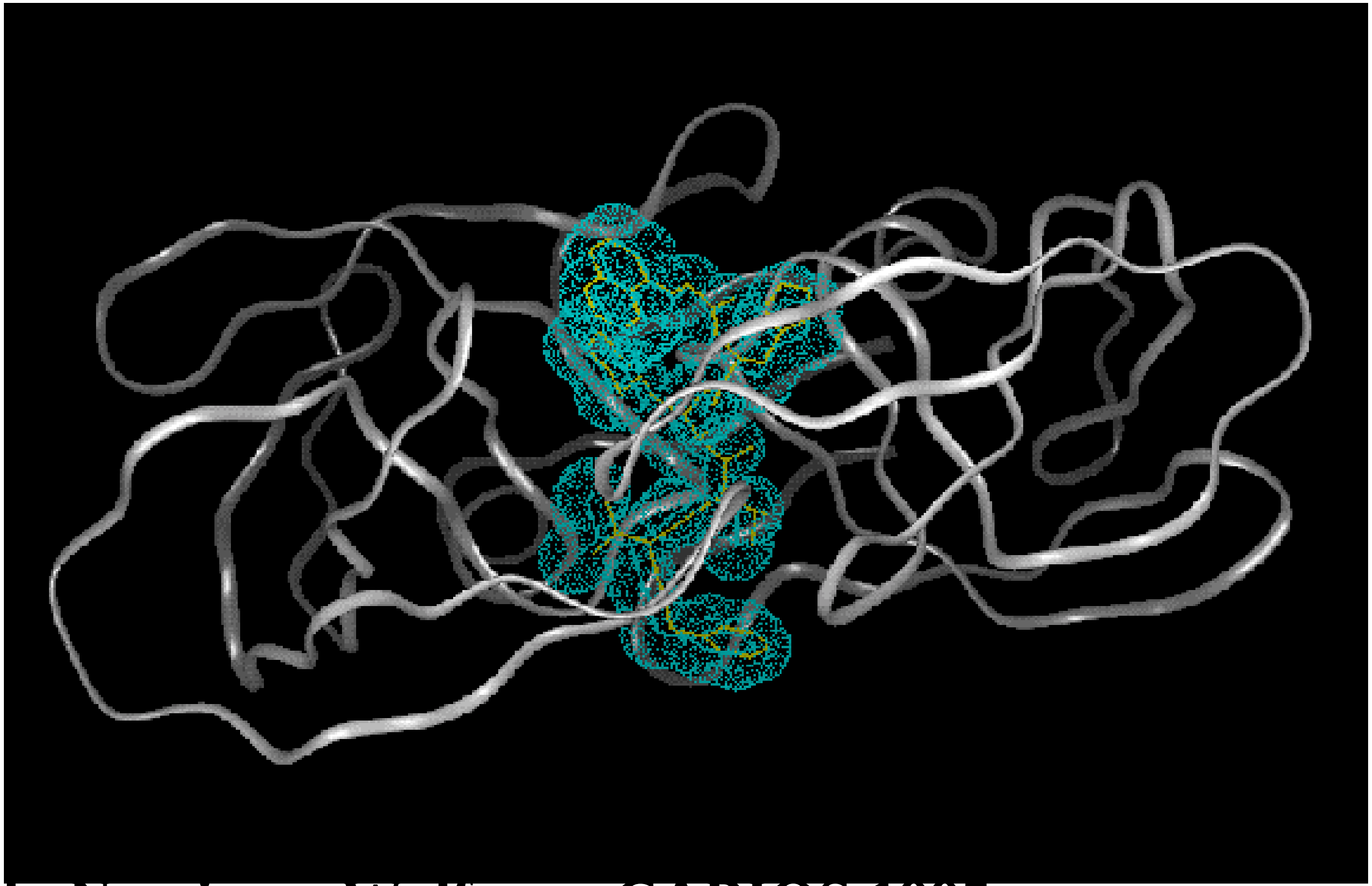
- Computer aided drug design – a new drug should fit the active site of a specific receptor.
- Understanding of the biochemical pathways - many reactions in the cell occur through interactions between the molecules.
- Crystallizing large complexes and finding their structure is difficult.

Flexible Docking
Calmodulin with M13 ligand



Sandak, Nussinov, Wolfson - JCB 1998.

Flexible Docking
HIV Protease Inhibitor



Sandak, Nussinov, Wolfson - CABIOS 1995.

Software Infrastructure

- Development of a software infrastructure for Geometric Computing in Molecular Biology.
- Object oriented, C++ library.
- Speed up development of new and re-usability of old software.
- Development of building blocks for fast testing of new ideas.

Cross - fertilization 1

- **Analogous tasks appear in Computer Vision, Medical Imaging, Structural Bioinformatics, Target Recognition.**
- **Similar software and hardware can handle all of these *Geometric Computing* tasks - *method based cross fertilization*.**

Cross - fertilization 2

- **Bioinformatics brings together Computer Scientists, Molecular Biologists, Chemists etc. to tackle major problems in Computational Biology and Computer Assisted Drug Design - *<u>task based cross-fertilization</u>*.**

Conclusions 1

- **Molecular Biology and Biotechnology have entered a stage in which advanced algorithmic methods make the difference between theory and practice.**
- **Only true interdisciplinary collaboration among Computer and Life scientists can deliver biologically relevant computational techniques.**

Conclusions 2

- **The b.c. (before Computer Science) algorithms in Computational Biology/Biotechnology, which have been mostly developed by chemists and physicists, are analogous to the first generation CS algorithms. The current state-of-the-art of CS (~fifth generation) provides a quantum leap.**

Sample of Topics to be covered

- Protein and DNA sequence alignment.
- Protein structural alignment and classification.
- Biomolecular recognition prediction – docking.
- Folding (homology modelling, threading, ab-initio).
- Distance Geometry for structure calculation from NMR data (?)
- Computer Assisted Structural Drug Design.

GRADING

- Exercises - 50%.
- Final (individual) Project, which involves heavy programming, based on the exercises – 50%.
- Most likely, all the students will get the same project assignment.
- The exact grading details will be supplied by the TA, Maxim Shatsky.