

Chapter 13

Protein structure

Learning objectives

Upon completing this material you should be able to:

- understand the principles of protein primary, secondary, tertiary, and quaternary structure;
- use the NCBI tool CN3D to view a protein structure;
- use the NCBI tool VAST to align two structures;
- explain the role of PDB including its purpose, contents, and tools;
- explain the role of structure annotation databases such as SCOP and CATH; and
- describe approaches to modeling the three-dimensional structure of proteins.

Outline

Overview of protein structure

Principles of protein structure

Protein Data Bank

Protein structure prediction

Intrinsically disordered proteins

Protein structure and disease

Overview: protein structure

The three-dimensional structure of a protein determines its capacity to function. Christian Anfinsen and others denatured ribonuclease, observed rapid refolding, and demonstrated that the primary amino acid sequence determines its three-dimensional structure.

We can study protein structure to understand problems such as the consequence of disease-causing mutations; the properties of ligand-binding sites; and the functions of homologs.

Outline

Overview of protein structure

Principles of protein structure

Protein Data Bank

Protein structure prediction

Intrinsically disordered proteins

Protein structure and disease

Protein primary and secondary structure

(a) Primary structure

MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSD
GLAHLNDLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAAYQKVAVANALAHKYH

(b) Secondary structure

	10	20	30	40	50	60	70
UNK_257900	MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG						
DSC	cccc	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh
MLRC	cccc	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh
PHD	cccc	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh
Sec.Cons.	cccc	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh

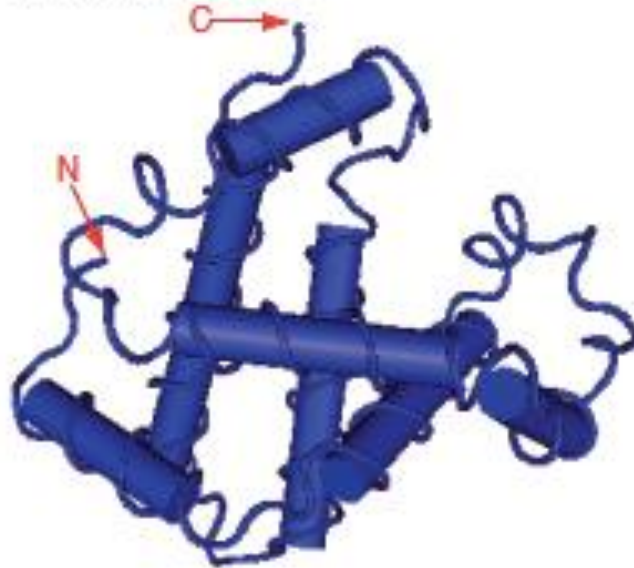
	80	90	100	110	120	130	140
UNK_257900	AFSDGLAHLNDLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAAYQKVAVAN						
DSC	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh
MLRC	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh
PHD	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh
Sec.Cons.	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh	hhhh

UNK_257900	ALAHKYH
DSC	hhhhccc
MLRC	hhhhccc
PHD	hhhhccc
Sec.Cons.	hhhhccc

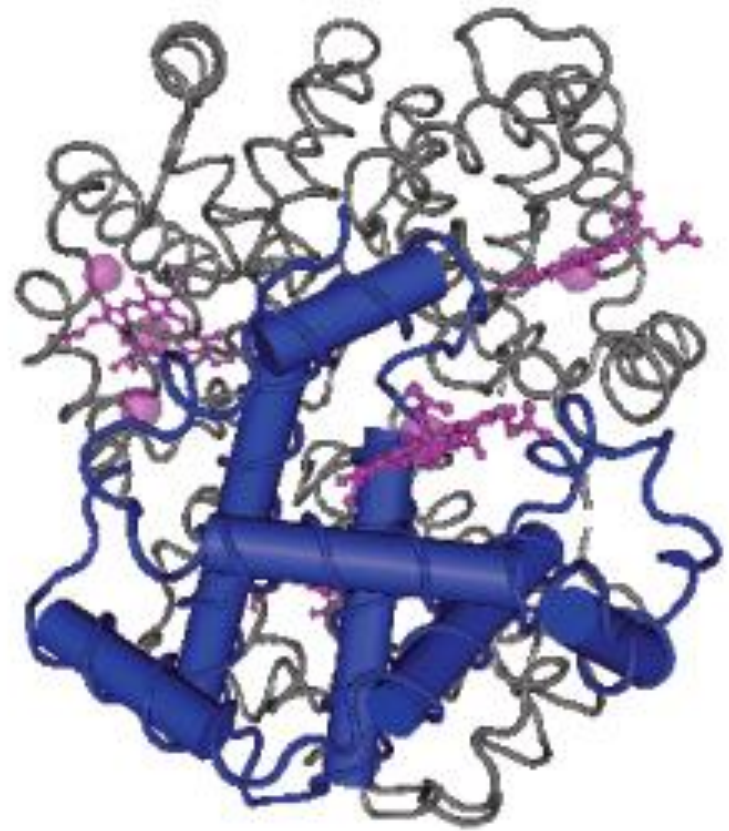
Results from three secondary structure programs are shown, with their consensus.
h: alpha helix; c: random coil;
e: extended strand

Protein tertiary and quaternary structure

(c) Tertiary structure



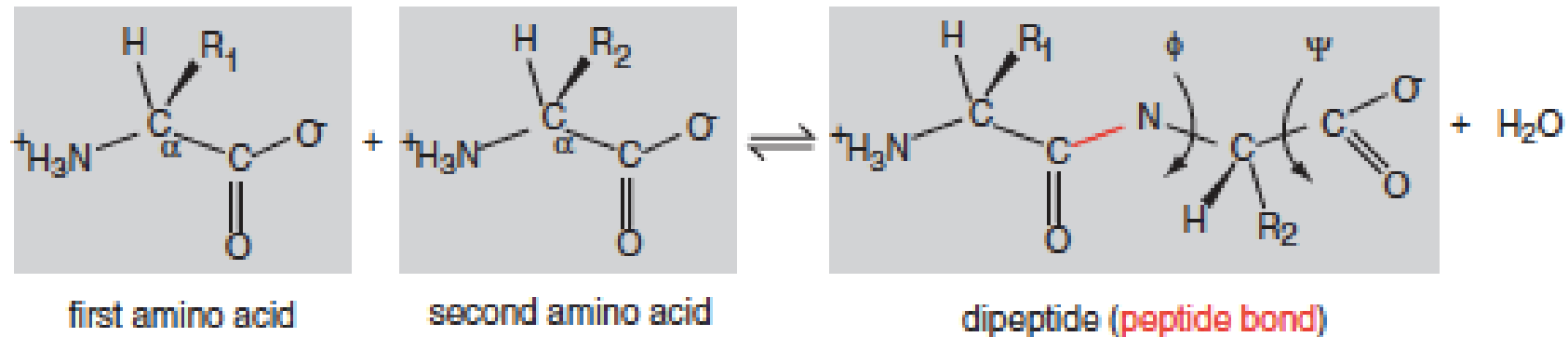
(d) Quaternary structure



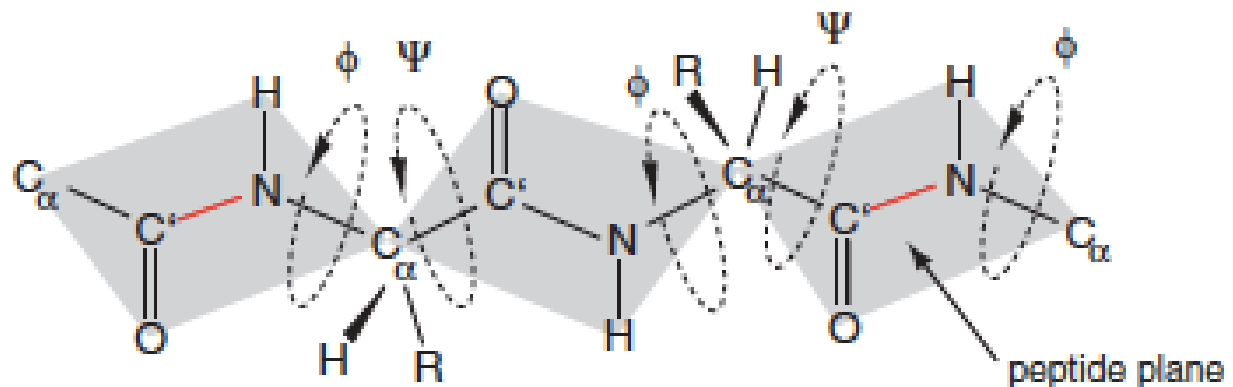
Quarternary structure: the four subunits of hemoglobin are shown (with an $\alpha 2\beta 2$ composition and one beta globin chain high- lighted) as well as four noncovalently attached heme groups.

The peptide bond; phi and psi angles

(a) peptide bond



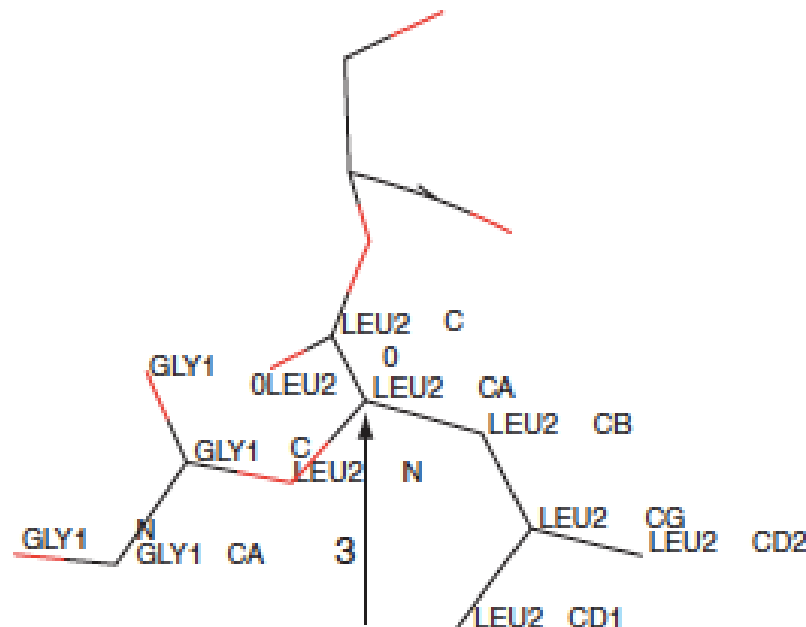
(b) phi and psi angles of polypeptide



(c) DeepView control bar



(d) DeepView viewer



Protein secondary structure

Protein secondary structure is determined by the amino acid side chains.

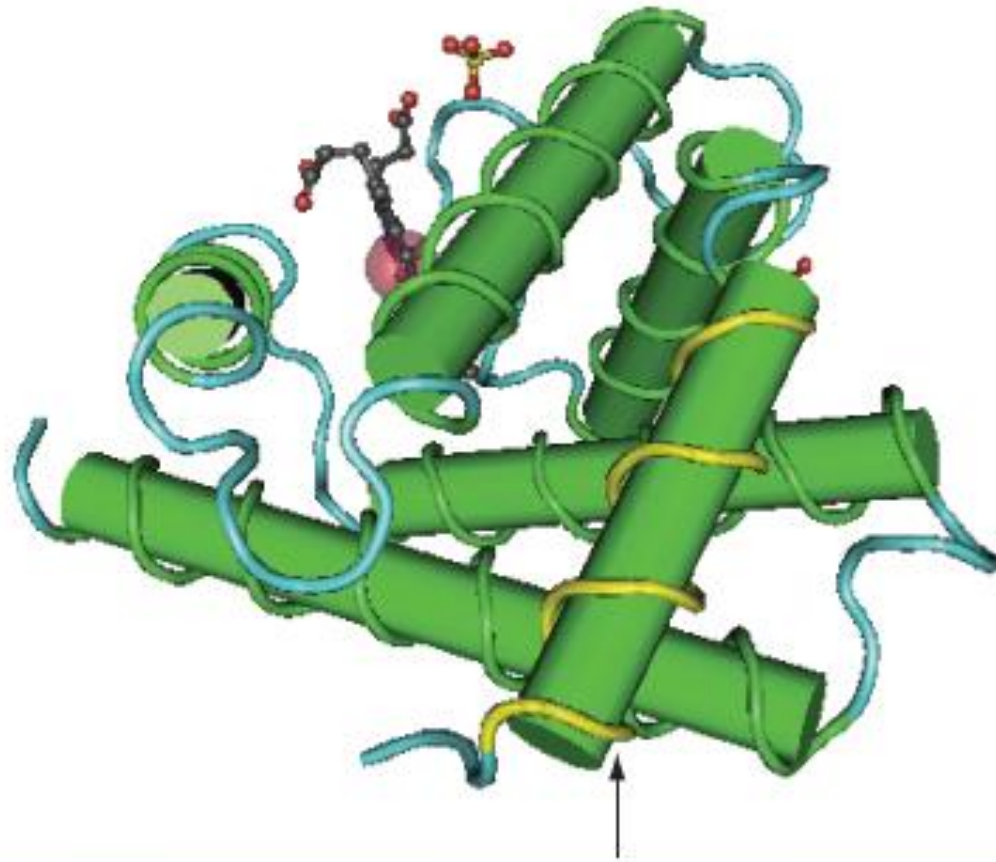
Myoglobin is an example of a protein having many α -helices. These are formed by amino acid stretches 4-40 residues in length.

Thioredoxin from *E. coli* is an example of a protein with many β sheets, formed from β strands composed of 5-10 residues. They are arranged in parallel or antiparallel orientations.

<https://proteinstrutures.com/Structure/Structure/secondary-sructure.html>

Protein secondary structure: myoglobin (alpha helical)

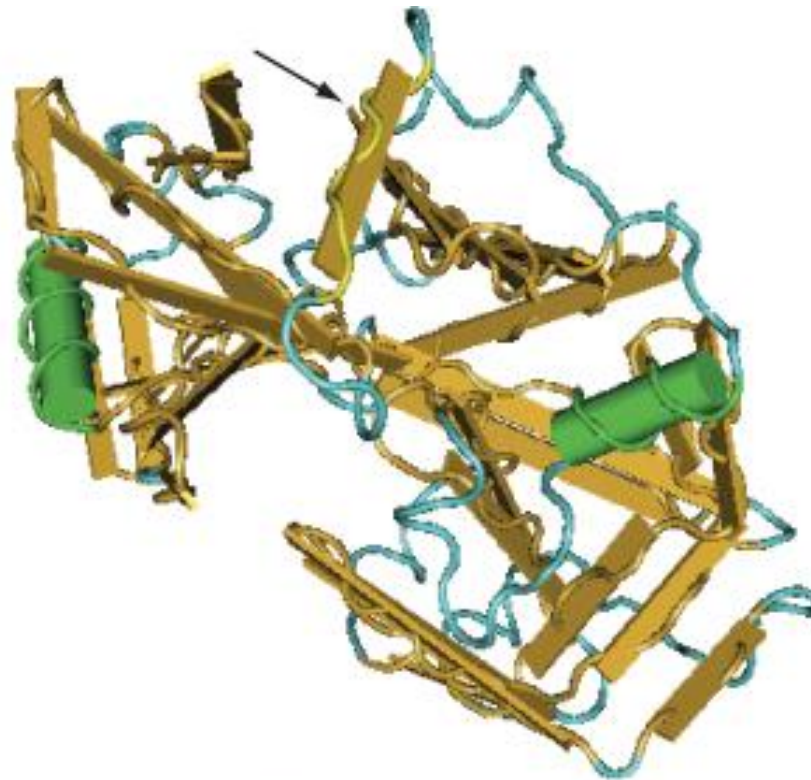
(a)



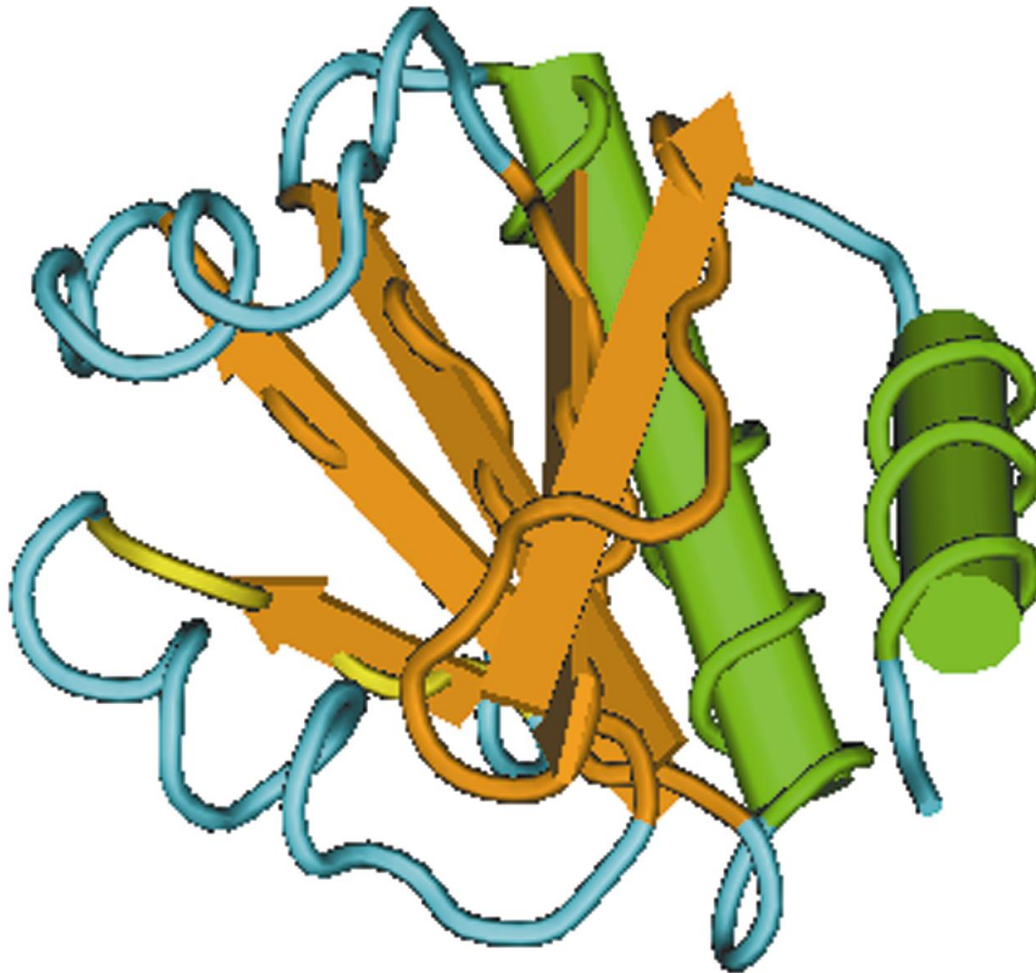
Myoglobin (John Kendrew, 1958) in Cn3D software (NCBI)

Protein secondary structure: pepsin (beta sheets)

(5)



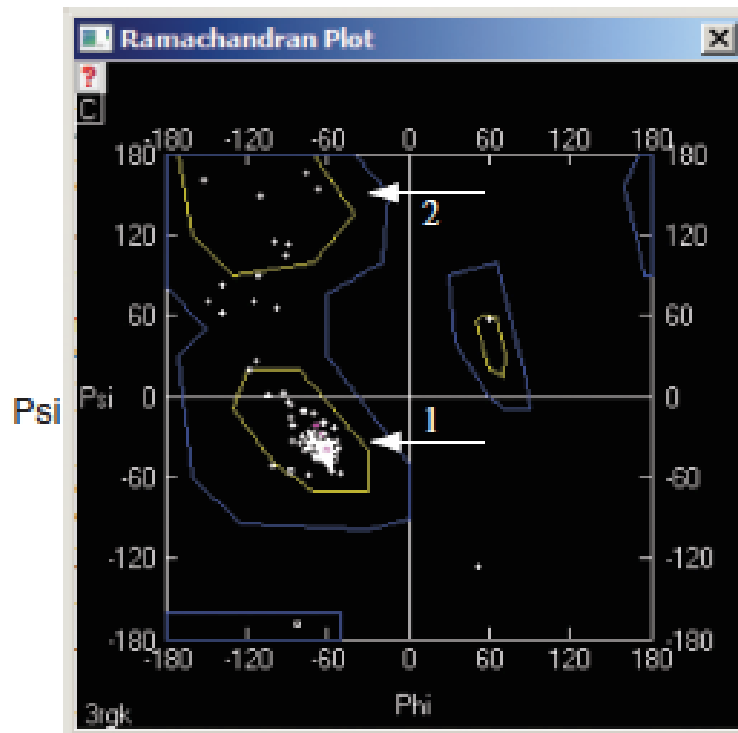
Click residues in the sequence viewer (highlighted in yellow) to see the corresponding residues (here a beta strand; arrow at top) highlighted in the structure image.



Thioredoxin: structure having beta sheets (brown arrows) and alpha helices (green cylinders).

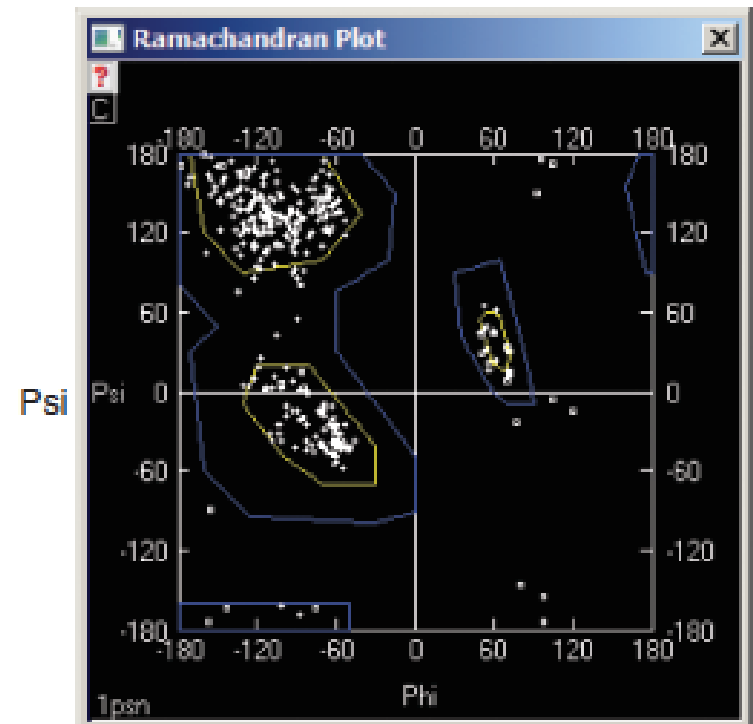
Protein secondary structure: Ramachandran plot

(a) Ramachandran plot: myoglobin (3RGK)



Phi

(b) Ramachandran plot: pepsin (1PSN)



Phi

Myoglobin (left) is mainly alpha helical (see arrow 1); pepsin (right) has beta sheets (see region of arrow 2)

Secondary structure prediction

Chou and Fasman (1974) developed an algorithm based on the frequencies of amino acids found in α helices, β -sheets, and turns.

Proline: occurs at turns, but not in α helices.

GOR (Garnier, Osguthorpe, Robson): related algorithm

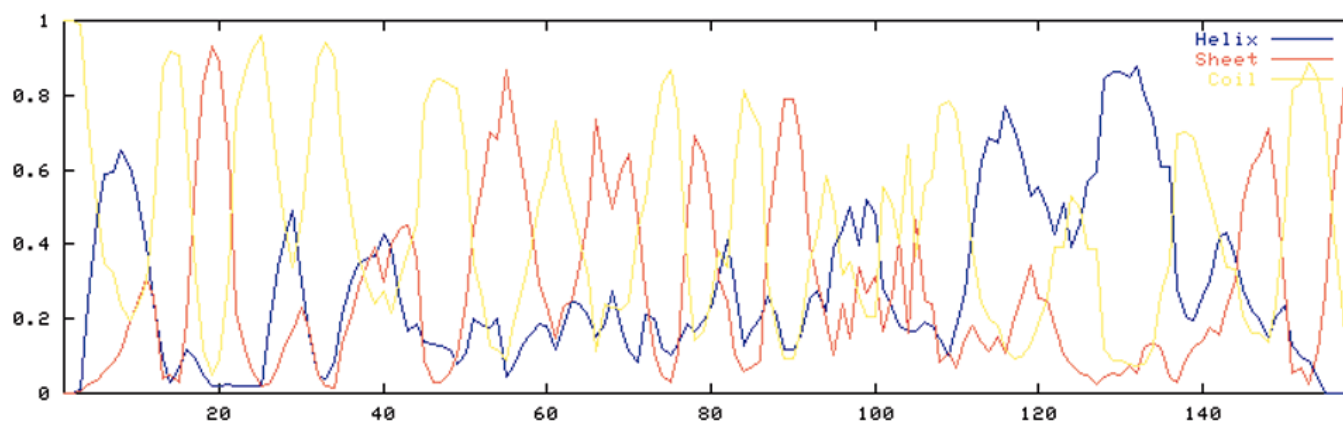
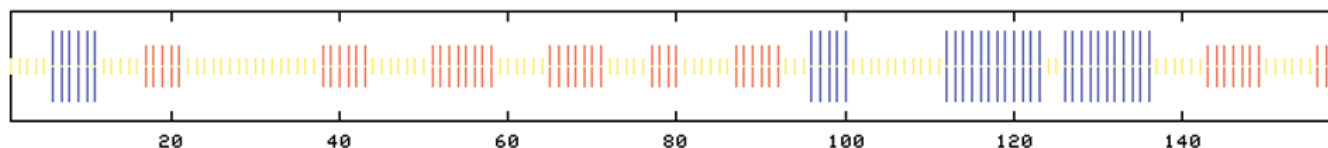
Modern algorithms: use multiple sequence alignments and achieve higher success rate (about 70-75%)

10 20 30 40 50 60 70
 | | | | | | |
 AQEEEEAEQNLSSELSGPWRTVYIGSTNPEKIQENGPFRTYFRELVFDDEKGTVDFFYSVKRDGKWKNVHVK
 cccccc hhhhhh ccc
 ATKQDDGTYVADYEGQNVFKIVSLSRTHLVAHNINVDKHGTTELTELFVKLNVEDEDLEKFWKLTEDKG
 ecccccccccccccccccccccccc hhhhhh ccccccccc hhhhhhhhhhhhh c hhhhhhhhhhhhh ccc
 IDKKNVNFLENEDHPHPE
 ccccccccccccccccccc

Sequence length : 159

GOR4 :

Alpha helix	(Hh) :	34 is	21.38%
3 ₁₀ helix	(Gg) :	0 is	0.00%
Pi helix	(Ii) :	0 is	0.00%
Beta bridge	(Bb) :	0 is	0.00%
Extended strand	(Ee) :	45 is	28.30%
Beta turn	(Tt) :	0 is	0.00%
Bend region	(Ss) :	0 is	0.00%
Random coil	(Cc) :	80 is	50.31%
Ambiguous states (?)	:	0 is	0.00%
Other states	:	0 is	0.00%



	10	20	30	40	50	60	70
3DSEQ pdb1pboA pdb1pboA	AQEEEEAEQNLSGSPWRTVYIGSTNPEKIQENGPFRTYFRELVFDDDEKGTVDFFYFSVKRDGKWKNVHVK						
DPM	cc <h>hhhhhh</h> hcctctccceeeettctccc <h>tcctccccc</h> ee <h>hhhe</h> hht <h>cccc</h> eeeeeeettctctccceeh						
DSC	cc						
GOR4	cccccc <h>hhhhhh</h> cc						
HNNC	cc <h>hhhhhhhhhh</h> cccccccccccccccccccccccccccccccc <h>hee</h> hecccccccccccccccccccccccccccc						
PHD	cccccccccccccccccccccccccccccccccccc <h>hh</h> hecccccccccccccccccccccccccccccccccccc						
Predator	ccc <h>hhhhhh</h> cccccccccccccccccccccccccccc <h>hhhhhh</h> cccccccccccccccccccccccccccccccc						
SIMPA96	c <h>hhhhhhhhhh</h> cccccccccccccccccccccccc <h>hhhhhh</h> hecccccccccccccccccccccccccccccccc						
SOPM	h <h>hhhhhhhhhh</h> htcccccccccccttctcttcccc <h>hhhh</h> heectttccccccccccctttcccccccccc						
Sec.Cons.	cc?h <h>hhhhhhhh</h> cccccccccccccccccccccccc <h>h</h> ?e?hecccccccccccccccccccc?eeee						

	80	90	100	110	120	130	140
3DSEQ pdb1pboA pdb1pboA	ATKQDDGTYVADYEGQNVFKIVSLSRTHLVAHNINVDKHGQKTEL TGLFVKLNVEDEDLEKFWKLTEDKG						
DPM	hhttttccce <h>hct</h> ctccceeeeeeeee <h>ee</h> hhcetctcccccc <h>he</h> ceee <h>hhhhhhhhhhhhhhhh</h> htcc						
DSC	eecc <h>hhhhhh</h> eecccc						
GOR4	eecccccccccccccccccccccccc <h>hhhh</h> cccccccccccc <h>hhhh</h> heeecccc <h>hhhhhhhhhhhh</h> ccc						
HNNC	eecccccccccccccccccccccccc <h>he</h> cccccccccccccccccccccccccccc <h>hhhhhhhhhhhh</h> ccc						
PHD	eecc <h>hhhhhhhhhhhhhh</h> ccc						
Predator	eecc <h>hhhhhh</h> cccc <h>hhhhhhhhhh</h> ccc						
SIMPA96	cc <h>hhhhhh</h> heeecccc <h>hhhhhhhhhhhh</h> ccc						
SOPM	ecctttccccceettcccccccccccccccccccccccttccc <h>hh</h> heeecccc <h>hhhhhhhhhh</h> ccctt						
Sec.Cons.	eecccccccccccccccccccc?cccccccc?c?cccccc?hh?eeeecccc <h>hhhhhhhhhh</h> ccc						

	150
3DSEQ pdb1pboA pdb1pboA	IDKKNVVNFLENEDHPH
DPM	ctccccce <h>hh</h> htcccc
DSC	cccccccccccccccc
GOR4	ccccccccccccceec
HNNC	cccccccccccccccc
PHD	cccc <h>hh</h> cccccccc
Predator	ccccceeecccccccc
SIMPA96	cc <h>hhhhhhhh</h> cccccc
SOPM	cc <h>hhhhhhhh</h> htcccc
Sec.Cons.	cccc?eeee?cccccc

Secondary structure prediction: conformational preferences of the amino acids

Amino acid	Preference			Properties
	Helix	Strand	Turn	
Glu	1.59	0.52	1.01	Helical preference; extended flexible side chain
Ala	1.41	0.72	0.82	
Leu	1.34	1.22	0.57	
Met	1.30	1.14	0.52	
Gln	1.27	0.98	0.84	
Lys	1.23	0.69	1.07	
Arg	1.21	0.84	0.90	
His	1.05	0.80	0.81	
Val	0.90	1.87	0.41	Strand preference; bulky side chains, beta-branched
Ile	1.09	1.67	0.47	
Tyr	0.74	1.45	0.76	
Cys	0.66	1.40	0.54	
Trp	1.02	1.35	0.65	
Phe	1.16	1.33	0.59	
Thr	0.76	1.17	0.90	
Gly	0.43	0.58	1.77	Turn preference; restricted conformations, side-main chain interactions
Asn	0.76	0.48	1.34	
Pro	0.34	0.31	1.32	
Ser	0.57	0.96	1.22	
Asp	0.99	0.39	1.24	

Secondary structure prediction

Web servers include:

GOR4

Jpred

NNPREDICT

PHD

Predator

PredictProtein

PSIPRED

SAM-T99sec

Secondary structure prediction: codes from the DSSP database

DSSP code	Secondary structure assignment
H	Alpha helix
B	Residue in isolated beta-bridge
E	Extended strand, participates in beta ladder
G	3-helix (3/10 helix)
I	5 helix (pi helix)
T	Hydrogen bonded turn
S	Bend
Blank or C	Loop or irregular element, incorrectly called "random coil" or "coil."

DSSP is a dictionary of secondary structure, including a standardized code for secondary structure assignment.

Tertiary protein structure: protein folding

Main approaches:

[1] Experimental determination
(X-ray crystallography, NMR)

https://en.wikipedia.org/wiki/Nuclear_magnetic_resonance

[2] Prediction

- ▶ Comparative modeling (based on homology)
- ▶ Threading
- ▶ *Ab initio* (de novo) prediction

Experimental approaches to protein structure

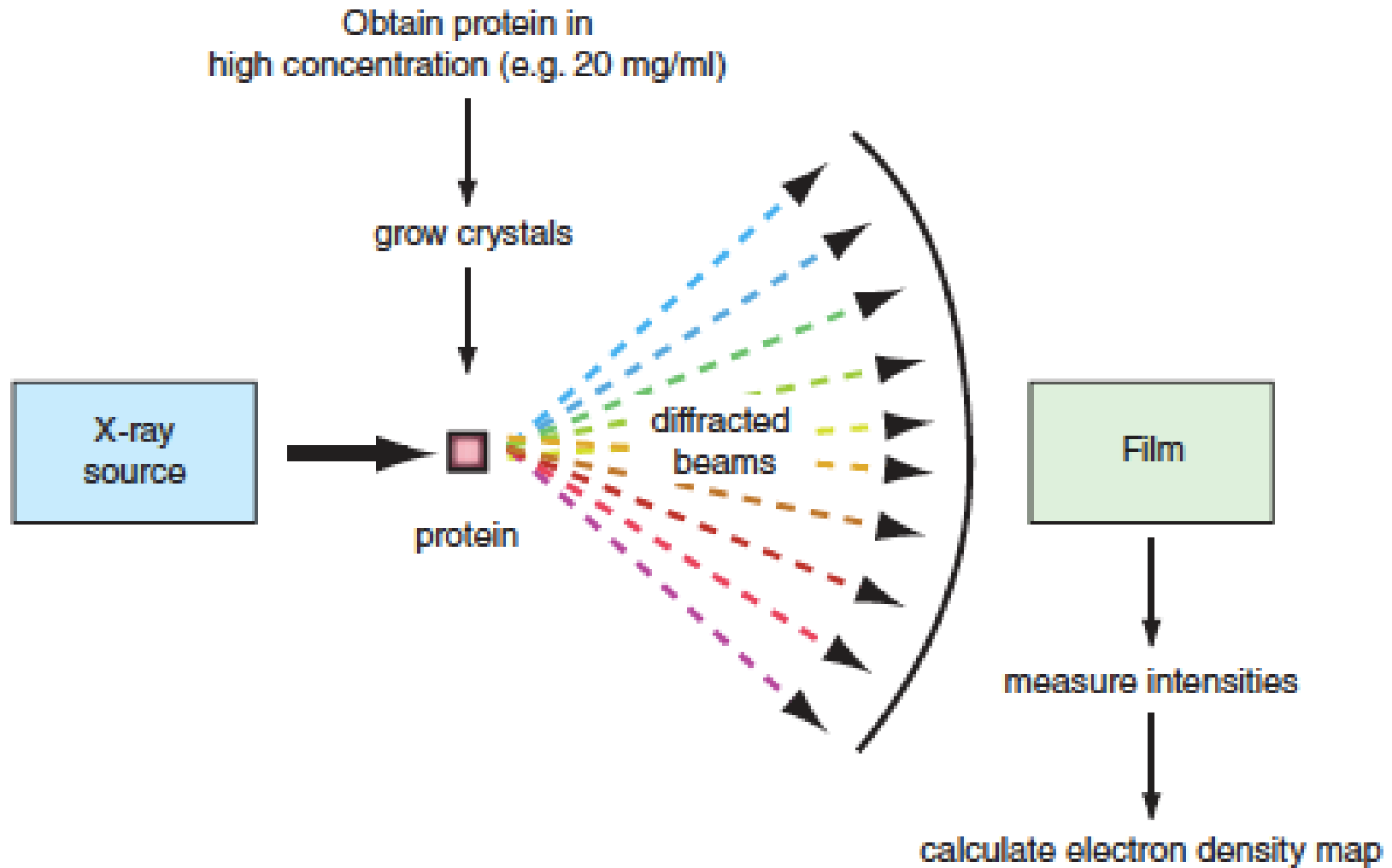
[1] X-ray crystallography

- Used to determine 80% of structures
- Requires high protein concentration
- Requires crystals
- Able to trace amino acid side chains
- Earliest structure solved was myoglobin

[2] NMR

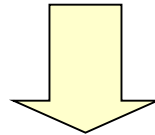
- Magnetic field applied to proteins in solution
- Largest structures: 350 amino acids (40 kD)
- Does not require crystallization

X-ray crystallography

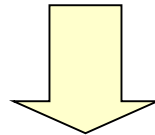


Steps in obtaining a protein structure

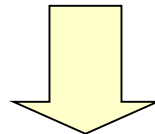
Target selection



Obtain, characterize protein

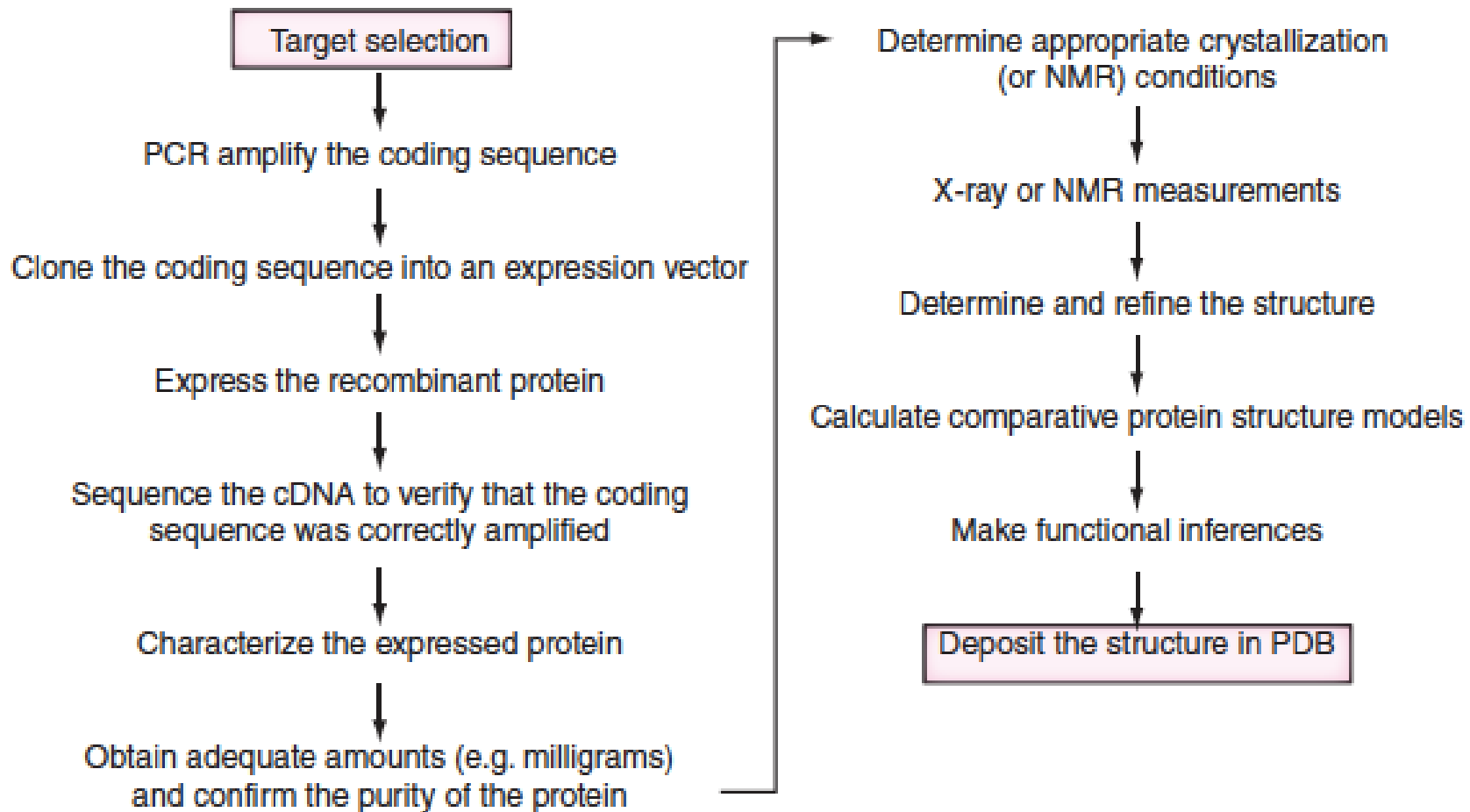


Determine, refine, model the structure



Deposit in repository

Target selection for protein structure determination



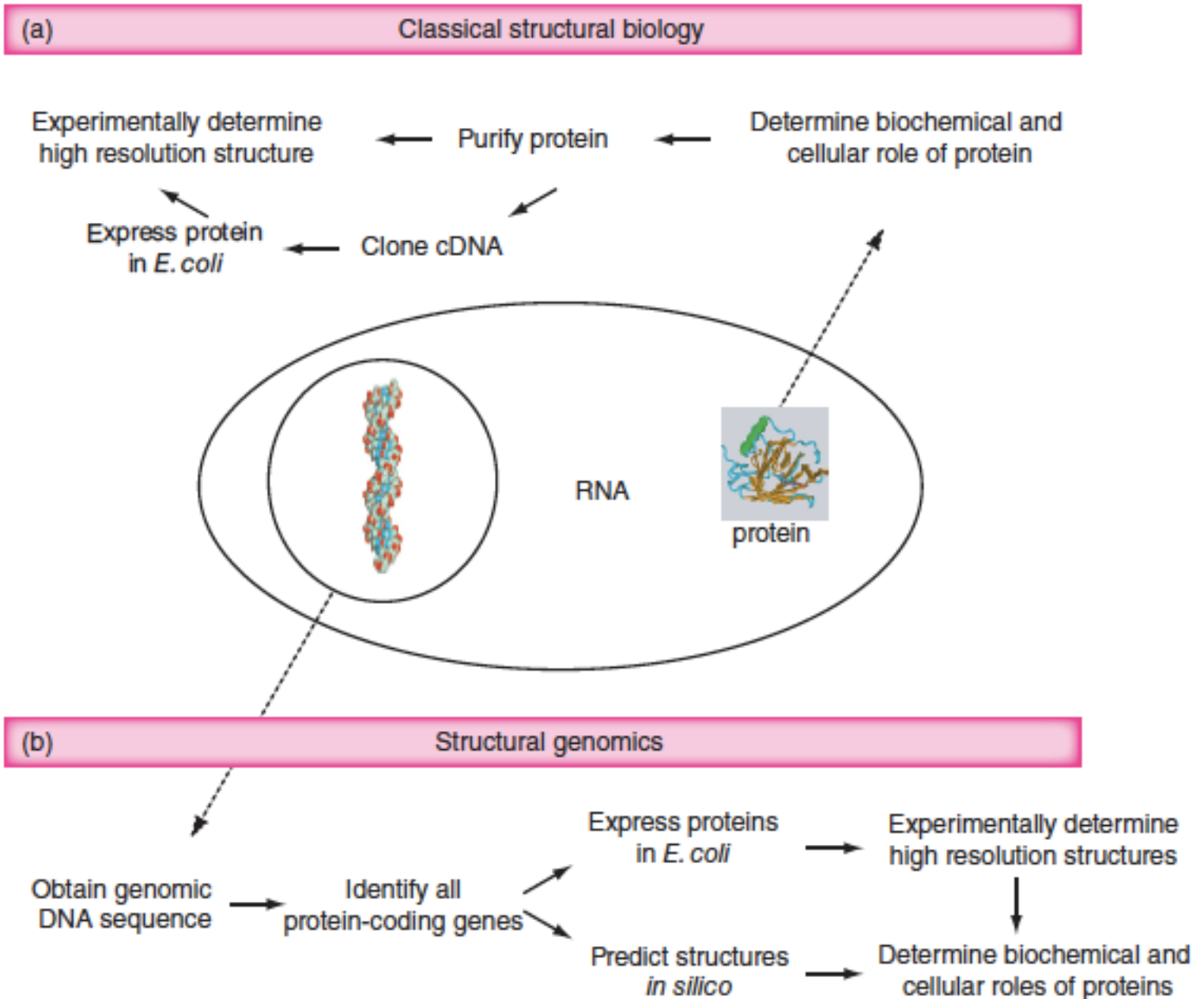
Priorities for target selection for protein structures

Historically, small, soluble, abundant proteins were studied (e.g. hemoglobin, cytochromes c, insulin).

Modern criteria:

- Represent all branches of life
- Represent previously uncharacterized families
- Identify medically relevant targets
- Some are attempting to solve all structures within an individual organism (*Methanococcus jannaschii*, *Mycobacterium tuberculosis*)

From classical structural biology to structural genomics



Outline

Overview of protein structure

Principles of protein structure

Protein Data Bank

Protein structure prediction

Intrinsically disordered proteins

Protein structure and disease

The Protein Data Bank (PDB)

- PDB is the principal repository for protein structures
- Established in 1971
- Accessed at <http://www.rcsb.org/pdb> or simply <http://www.pdb.org>
- Currently contains >100,000 structure entities



Organism

- Homo sapiens (24297)
- Escherichia coli (4796)
- Mus musculus (4180)
- Saccharomyces cerevisiae (2287)
- Bos taurus (2241)
- Rattus norvegicus (2036)
- Escherichia coli K-12 (1824)
- Other (52344)



Taxonomy

- Eukaryota (48927)
- Bacteria (35310)
- Viruses (6285)
- Archaea (3589)
- Unassigned (2841)
- Other (743)

Protein Data Bank (PDB)



Experimental Method

- X-ray (84522)
- Solution NMR (10124)
- Electron Microscopy (702)
- Solid-State NMR (62)
- Hybrid (59)
- Neutron Diffraction (43)
- Fiber Diffraction (38)
- Electron Crystallography (38)
- Solution Scattering (32)
- Other (24)



X-ray Resolution

- less than 1.5 Å (6353)
- 1.5 - 2.0 Å (28201)
- 2.0 - 2.5 Å (28405)
- 2.5 - 3.0 Å (15192)
- 3.0 and more Å (6396)
- more choices...



Release Date

- before 2000 (10969)
- 2000 - 2005 (17801)
- 2005 - 2010 (33329)
- 2010 - today (33545)
- this year (8669)
- this month (539)
- more choices...



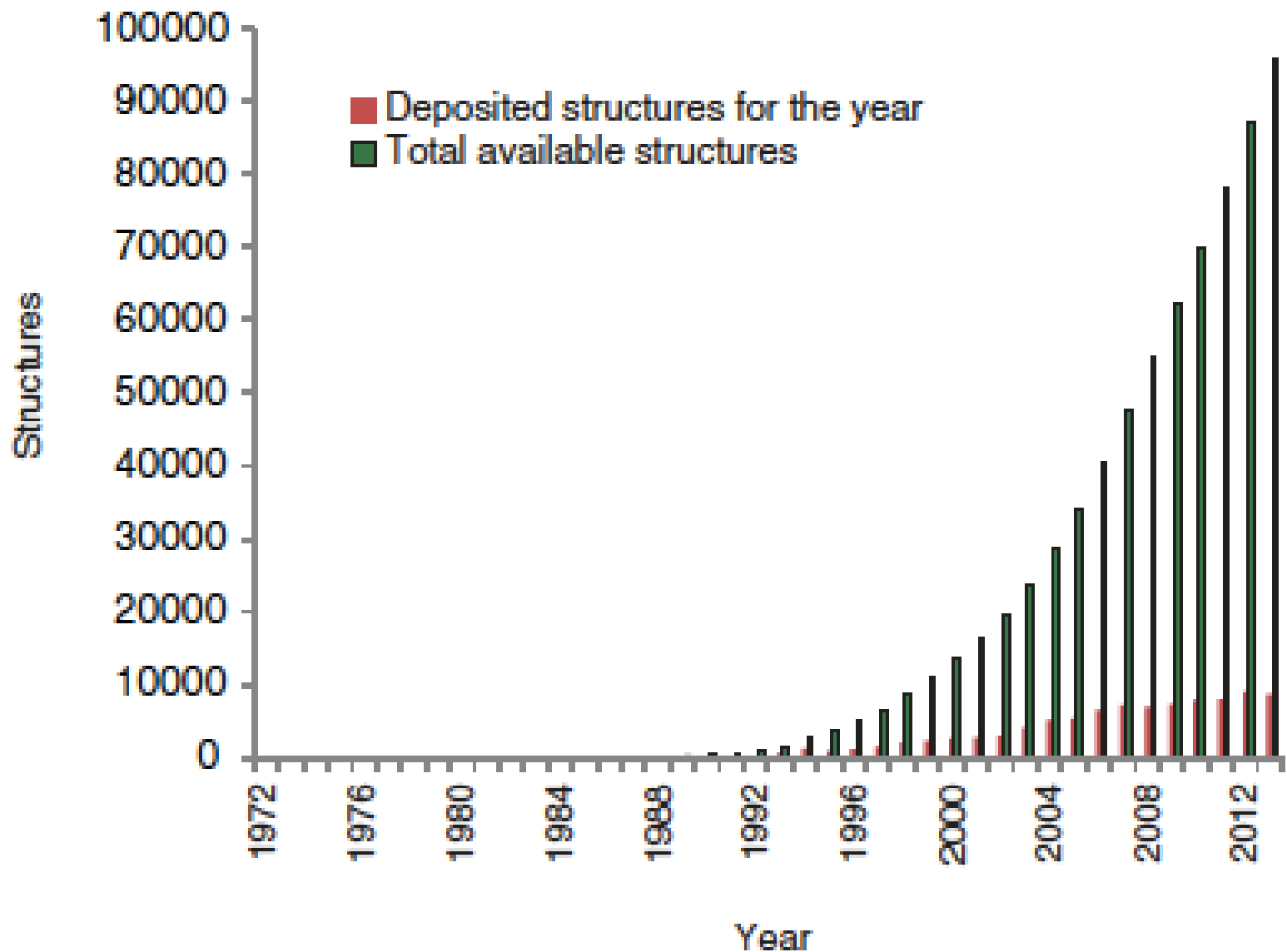
Polymer Type

- Protein (88526)
- Mixed (4614)
- DNA (1472)
- RNA (1007)

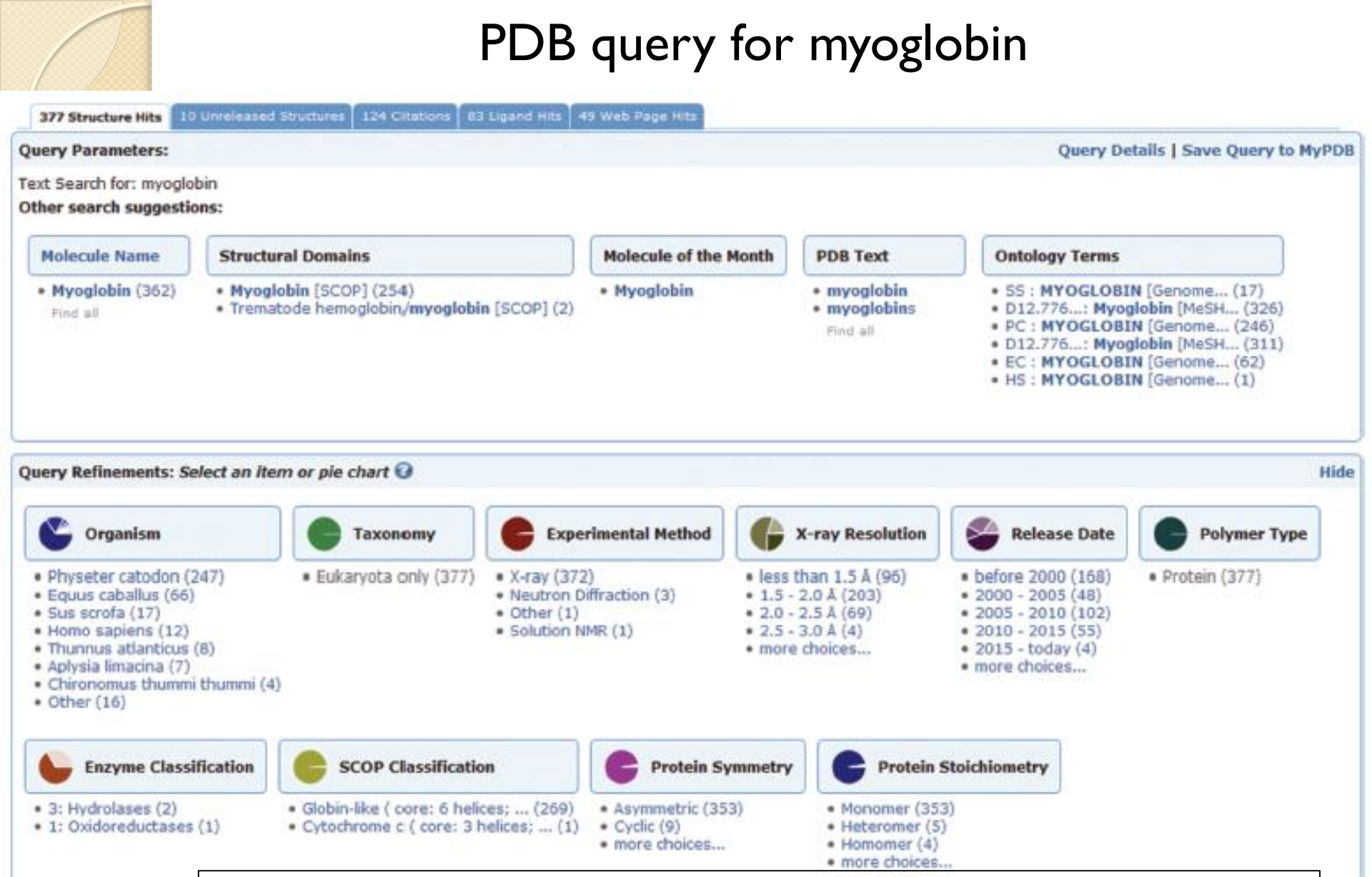
Protein Data Bank (PDB) holdings

Experimental technique	Proteins	Nucleic acids	Protein and nucleic acid complexes	Other	Total
X-ray diffraction	88,991	1,608	4,398	4	95,001
NMR	9,512	1,112	224	8	10,856
Electron microscopy	539	29	172	0	740
Hybrid	68	3	2	1	74
Other	164	4	6	13	187
Total	99,274	2,756	4,802	26	106,858

PDB: number of searchable structures per year



PDB query for myoglobin



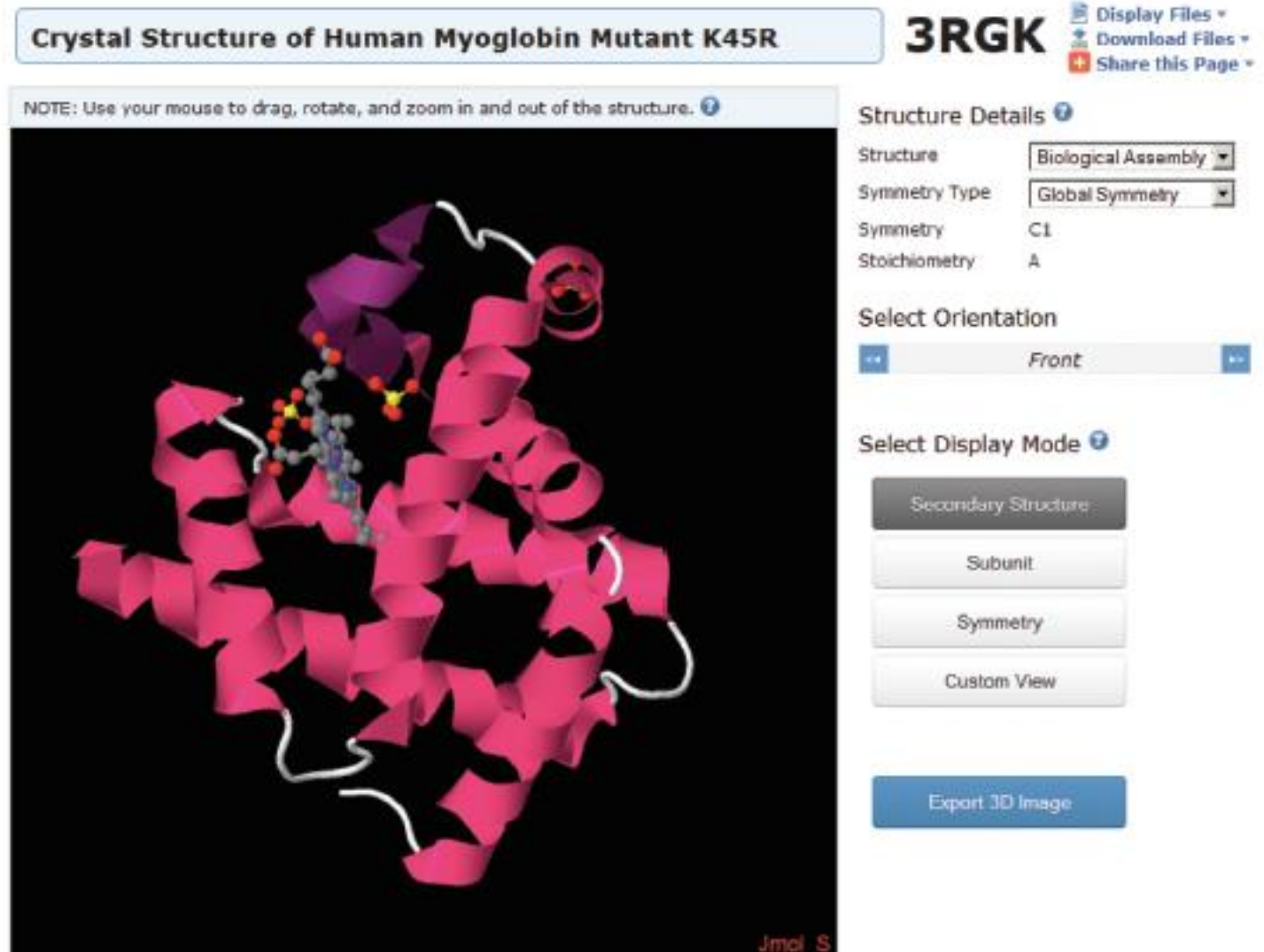
Result of a PDB query for myoglobin. There are several hundred results organized into categories such as UniProt gene names, structural domains, and ontology terms.

Interactive visualization tools for PDB protein structures

Tool	Comment	URL
Cn3D	From NCBI	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
JMol	Open-source Java viewer for chemical structures in 3D	http://jmol.sourceforge.net/
Kiosk Viewer	Uses Java Web Start	http://pdb.org/
Mage	Reads Kinemages	http://kinemage.biochem.duke.edu
Protein Workshop Viewer	Uses Java Web Start	http://pdb.org/
RasMol	Molecular graphics visualization tool	http://www.rasmol.org/
RasTop	Molecular visualization software adapted from RasMol	http://www.geneinfinity.org/rastop/
Simple Viewer	Uses Java Web Start	http://pdb.org/
SwissPDB viewer	At ExPASy	http://spdbv.vital-it.ch
VMD	Visual Molecular Dynamics; University of Illinois	http://www.ks.uiuc.edu/Research/vmd/

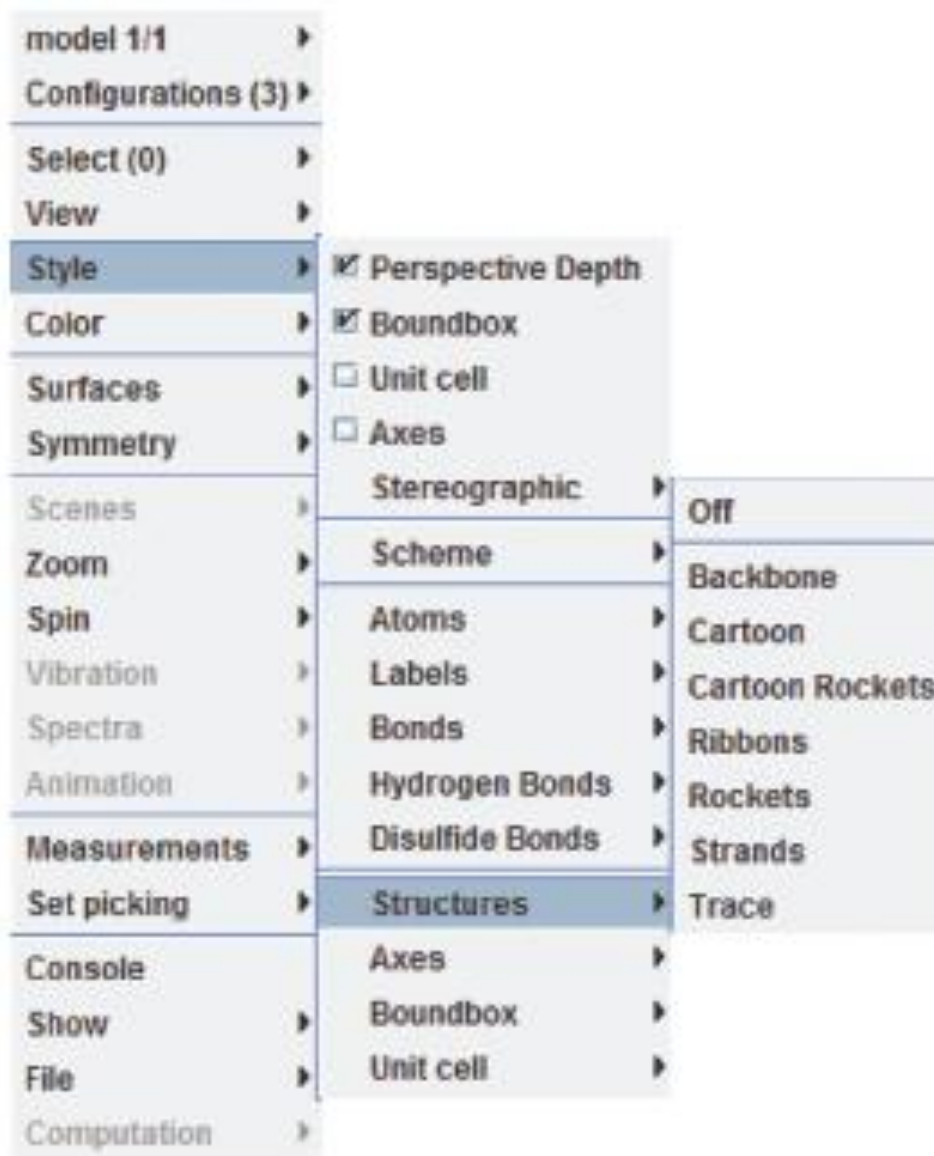
Visualization tools are available within PDB and elsewhere.

Visualizing myoglobin structure 3RGK: Jmol applet

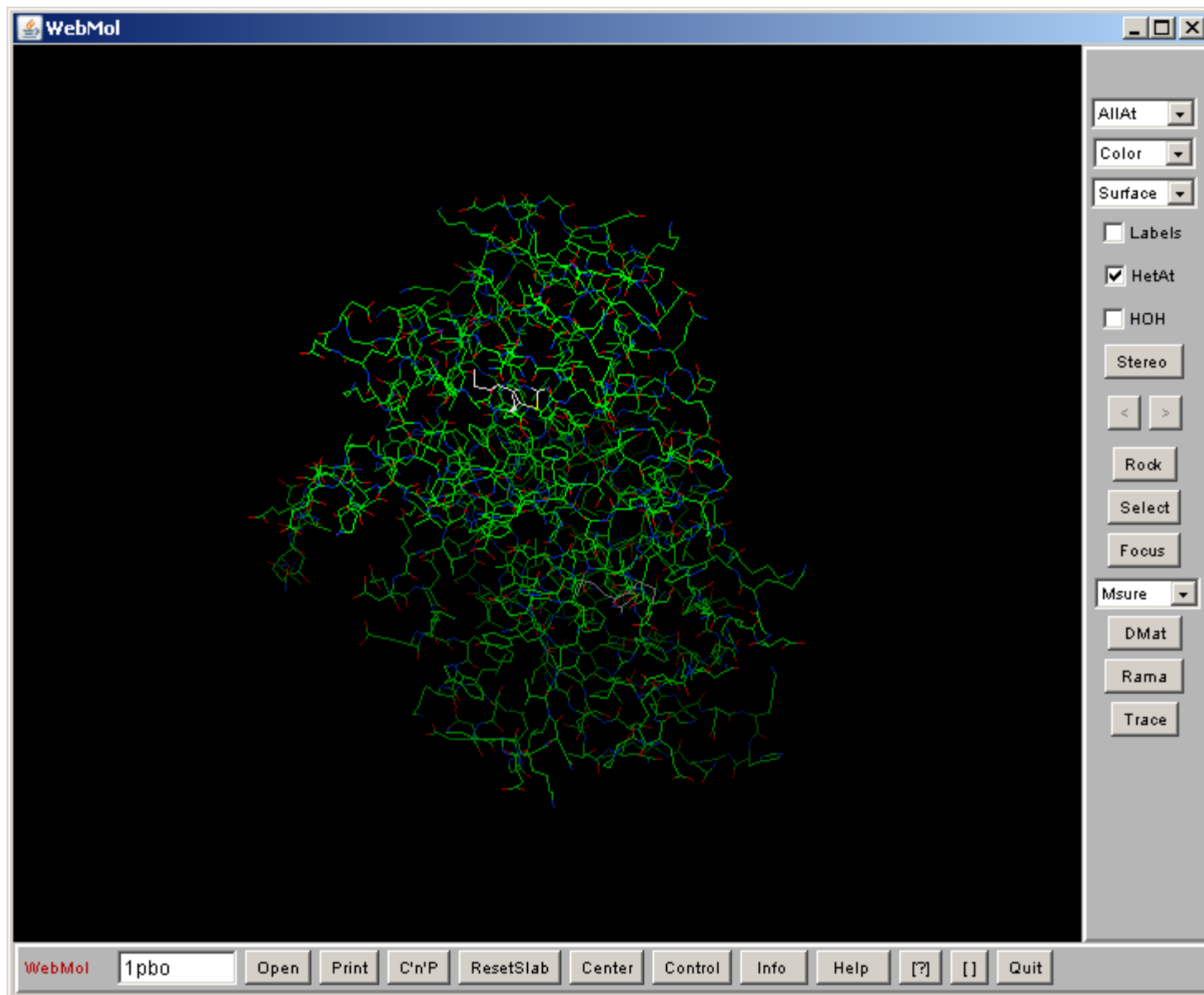


Jmol is available at PDB.

Visualizing structures: Jmol applet options

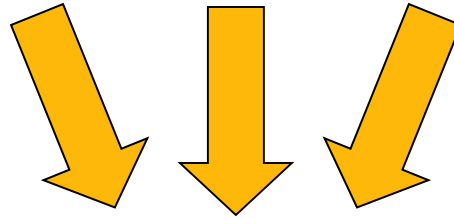


Viewing structures at PDB: WebMol

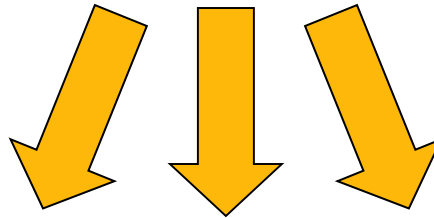


gateways to access PDB files

Swiss-Prot, NCBI, EMBL



Protein Data Bank



CATH, Dali, SCOP, FSSP

databases that interpret PDB files

Access to PDB through NCBI

You can access PDB data at the NCBI several ways.

- Go to the Structure site, from the NCBI homepage
- Perform a DELTA BLAST (or BLASTP) search, restricting the output to the PDB database


Access to PDB structures through NCBI

Molecular Modeling DataBase (MMDB)

Cn3D (“see in 3D” or three dimensions):
structure visualization software

Vector Alignment Search Tool (VAST):
view multiple structures

Access to PDB through NCBI: visit the Structure home page

 NCBI

Resources ▾

How To ▾

pevsner

[My NCBI](#)

[Sign Out](#)

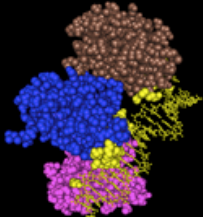
Structure

Structure ▾

Search

Advanced

Help



Structure

Three dimensional structures provide a wealth of information on the biological function and the evolutionary history of macromolecules. They can be used to examine sequence-structure-function relationships, interactions, active sites, and more.

Using Structure

- [Search](#)
- [How to \(Quick Start\) Guides](#)
- [Help](#)
- [News](#)
- [FTP](#)
- [Publications](#)
- [Discover](#)

Structure Tools

- [Macromolecular Resources Overview](#)
- [CBLAST](#)
- [Cn3D](#)
- [IBIS](#)
- [VAST](#)
- [VAST+](#)

More Resources

- [PDB](#)
- [Protein](#)
- [CDD](#)
- [PubChem](#)
- [NCBI Structure Group Resources & Research](#)

Access to PDB through NCBI: query the Structure home page

NCBI Resources ▾ How To ▾

Structure

[Create alert](#) [Advanced](#) [Help](#)

Summary ▾ 20 per page ▾ Sort by Default order ▾

Search results

Items: 1 to 20 of 177


[Send to: ▾](#) **Filter your results:**

[All \(177\)](#)

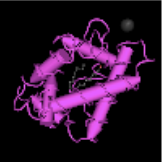
[NMR \(14\)](#)

[X-ray \(159\)](#)


[Manage Filters](#)

☐ 1.  [Complex Of Bovine Odorant Binding Protein \(obp\) With A Selenium Containing Odorant\[Odorant-Binding\]](#)

Taxonomy: Bos taurus
Proteins: 2 Chemicals: 2 modified: 2013-02-06
MMDB ID: 6240 PDB ID: 1PBO
[View in iCn3D](#) [Similar Structures](#) [PubMed](#) [Proteins](#) [Conserved Domains](#) [PubChem Compound](#)

☐ 2.  [Structure Of Bmori Gobp2 \(General Odorant Binding Protein 2\) With \(10e\)-Hexadecen-12-Yn-1-Ol\[Transport Protein\]](#)

Taxonomy: Bombyx mori
Proteins: 1 Chemicals: 3 modified: 2011-05-26
MMDB ID: 75871 PDB ID: 2WCM
[View in iCn3D](#) [Similar Structures](#) [PubMed](#) [Proteins](#) [Conserved Domains](#) [PubChem Compound](#)

☐ 3.  [Structure Of Bmori Gobp2 \(General Odorant Binding Protein 2\) With \(8e,10z\)-Hexadecadien-1-Ol\[Transport Protein\]](#)

Taxonomy: Bombyx mori
Proteins: 1 Chemicals: 4 modified: 2011-05-26
MMDB ID: 75870 PDB ID: 2WCL
[View in iCn3D](#) [Similar Structures](#) [PubMed](#) [Proteins](#) [Conserved Domains](#) [PubChem Compound](#)

Refine your results • What's this? ▴

Protein Domain Families

Families (163)

Superfamilies (173)

Complexes

Protein-Protein (47)

Protein-RNA (3)

Protein-Chemical (142)

Literature

PubMed (157)

PMC (55)

Taxonomy (177)

Find related data ▴

Database:

Molecular Modeling Database (MMDB) at NCBI

NCBI
National Center for
Biotechnology Information

Structure Summary MMDB

HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains Pubchem BioSystems Help

Crystal Structure of Human Myoglobin Mutant K45r

Citation: [?](#)
X-ray crystal structure of a recombinant human myoglobin mutant at 2.8 Å resolution.
Hubbard SR, Lambright SG, Boxer SG, Hendrickson WA
J.Mol.Biol. (1990) 20 p.215

MMDB ID: 90232 **PDB ID:** 3R GK [?](#)

PDB Deposition Date: 2011/4/8 [?](#)
Updated in MMDB: 10/2011 [?](#)
Experimental Method: X-Ray Diffraction [?](#)
Resolution: 1.65 Å [?](#)
Source Organism: Homo sapiens [?](#)
Similar Structures: [VAST](#) [?](#)

☒ **Default Biological Unit** ☐ All Biological Unit (1) ☐ Asymmetric Unit [?](#)

Biological Unit: monomeric; determined by author, and by software (PISA) [?](#)

Interactions [?](#)

Molecular Graphic [?](#)

View or Save 3D Structure [?](#)

File Format: [Cn3D](#)
Display As: [3D structure](#)
Data Set: [Single 3D structure](#)

[View structure](#)

[Download Cn3D](#)

NOTICE
In order to view this biological unit properly, please upgrade to Cn3D 4.3.

You can study PDB structures from NCBI. MMDB offers tools to analyze protein (and other) structures.


[PubMed](#)
[BLAST](#)
[Structure](#)
[Taxonomy](#)
[OMIM](#)
[Help?](#)
[Cn3d](#)

Description: Complex Of Bovine Odorant Binding Protein (Obp) With A Selenium Containing Odorant.

Deposition: L.M.Amzel, M.A.Bianchet, H.Monaco & G.Bains, 15-Jul-96

Taxonomy: [Bos taurus](#)

Reference: [PubMed](#) **MMDB:** [6240](#) **PDB:** [1PBO](#)

[View 3D Structure](#)

of

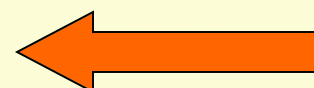
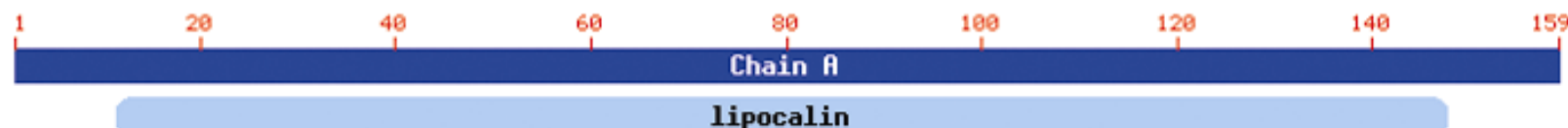
Best Model

with

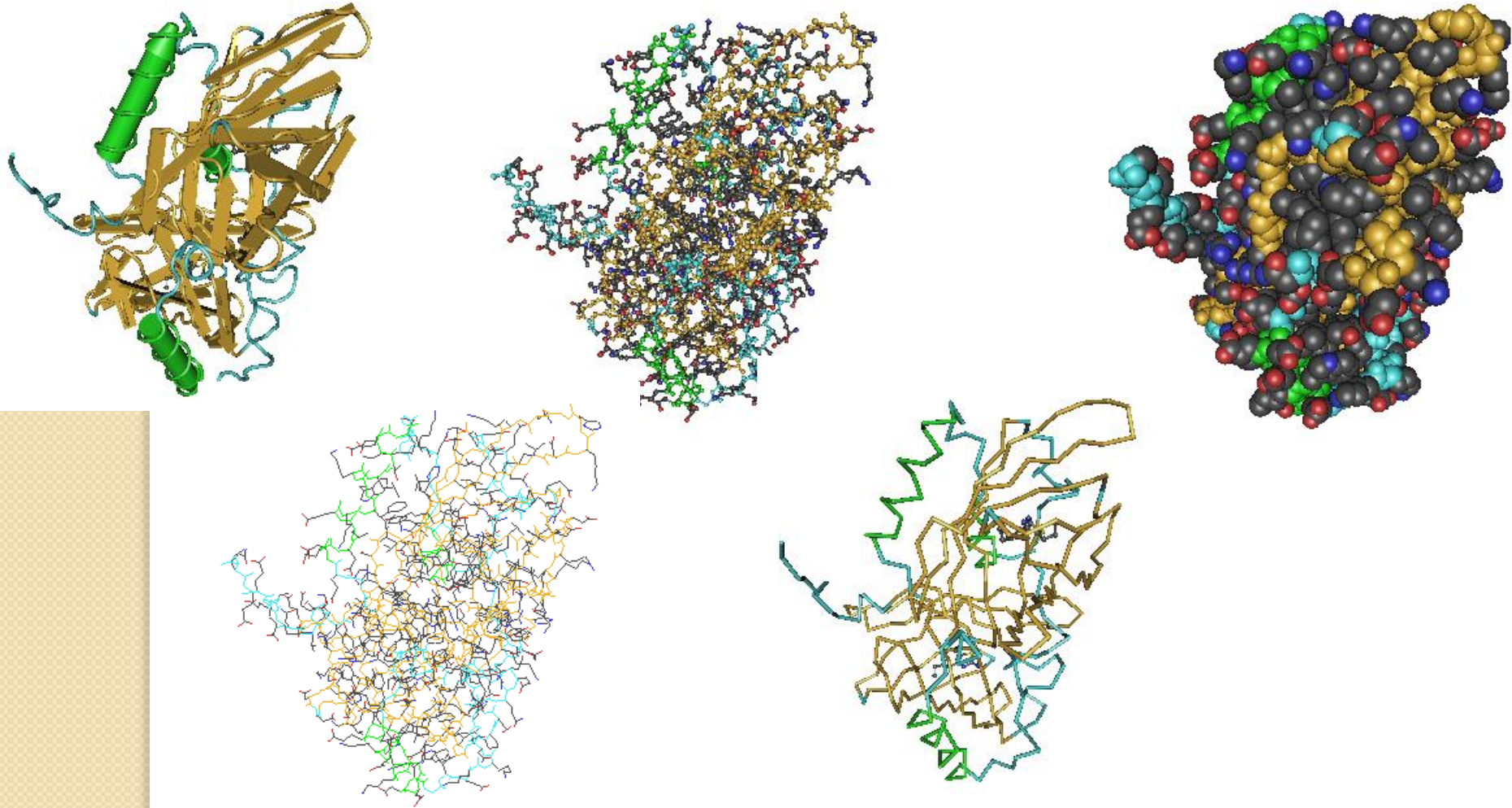
Cn3D

Display

NEW

[Get Cn3D 4.0!](#)

[Protein](#)
[CDs](#)

[Protein](#)
[3d Domains](#)
[CDs](#)


Cn3D: NCBI software for visualizing protein structures



Several display formats are shown

Standard Protein BLAST[blastn](#) **[blastp](#)** [blastx](#) [tblastn](#) [tblastx](#)**Enter Query Sequence**

BLASTP programs search protein databases using a protein query.

Enter accession number(s), gi(s), or FASTA sequence(s) ?

[Clear](#)

Query subrange ?

NP_003270

From

To

Or, upload file

Choose File

no file selected ?

Job Title

NP_003270:troponin C, skeletal muscle [Homo...

Enter a descriptive title for your BLAST search ?

☐ Align two or more sequences ?**Choose Search Set**

Database

Organism

Optional

Exclude

Optional

Non-redundant protein sequences (nr)

✓ Reference proteins (refseq_protein)

Model Organisms (landmark)

UniProtKB/Swiss-Prot (swissprot)

Patented protein sequences (pat)

Protein Data Bank proteins (pdb)

Metagenomic proteins (env_nr)

Transcriptome Shotgun Assembly proteins (tsa_nr)

?

☐ Exclude

+

20 top taxa will be shown. ?

multiple sequences

Do a DELTA BLAST (or BLASTP) search;
set the database to pdb (Protein Data Bank)

Access protein structures by using DELTA-BLAST (or BLASTP) restricting searches to proteins with PDB entries

Sequences producing significant alignments with E-value BETTER than threshold

Select: [All](#) [None](#) Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment						
	Description	Max score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	Chain B, Pigeon Hemoglobin (Oxy Form) >pdb 2R80 D Chain D, Pigeon Hemoglobin (Oxy Form) >pdb 2R80 D	164	99%	1e-53	69%	2R80_B
<input type="checkbox"/>	Chain B, Crystal Structure Of Parrot Hemoglobin (Psittacula Krameri) At Ph 7.5	160	99%	3e-52	69%	2ZFB_B
<input type="checkbox"/>	Chain B, R-State Form Of Chicken Hemoglobin D >pdb 1HBR D Chain D, R-State Form Of Chicken Hemoglobin D >pdb 1HBR D	159	99%	1e-51	69%	1HBR_B
<input type="checkbox"/>	Chain B, Crystal Structure Determination Of Japanese Quail (Coturnix Coturnix Japonica)	159	99%	2e-51	68%	3MJB_B
<input type="checkbox"/>	Chain B, Graylag Goose Hemoglobin (Oxy Form) >pdb 1FAW D Chain D, Graylag Goose Hemoglobin (Oxy Form) >pdb 1FAW D	158	99%	2e-51	68%	1FAW_B
<input type="checkbox"/>	Chain B, Structure Determination Of Haemoglobin From Turkey(meleagris Gallopavo) At 2.2 Angstrom Resolution	158	99%	3e-51	68%	2QMB_B
<input type="checkbox"/>	Chain B, Crystal Structure Determination Of Duck (Anas Platyrhynchos) Hemoglobin At 2.2 Angstrom Resolution	157	99%	4e-51	69%	3EOK_B
<input type="checkbox"/>	Chain B, Bar-Headed Goose Hemoglobin (Oxy Form) >pdb 1C40 B Chain B, Bar-Headed Goose Hemoglobin (Oxy Form) >pdb 1C40 B	157	99%	4e-51	69%	1A4F_B
<input type="checkbox"/>	Chain B, Crystal Structure Determination Of Ostrich Hemoglobin At 2.2 Angstrom Resolution	157	99%	4e-51	69%	3FSA_B
<input type="checkbox"/>	Chain A, R-State Form Of Chicken Hemoglobin D >pdb 1HBR C Chain C, R-State Form Of Chicken Hemoglobin D >pdb 1HBR C	157	99%	4e-51	69%	1HBR_A
<input type="checkbox"/>	Chain A, Crystal Structure Of Parrot Hemoglobin (Psittacula Krameri) At Ph 7.5	157	99%	4e-51	69%	2ZFB_A
<input type="checkbox"/>	Chain A, Crystal Structure Determination Of Ostrich Hemoglobin At 2.2 Angstrom Resolution	125	97%	9e-39	36%	3FSA_A

Structure accession
(e.g. 2JTZ)

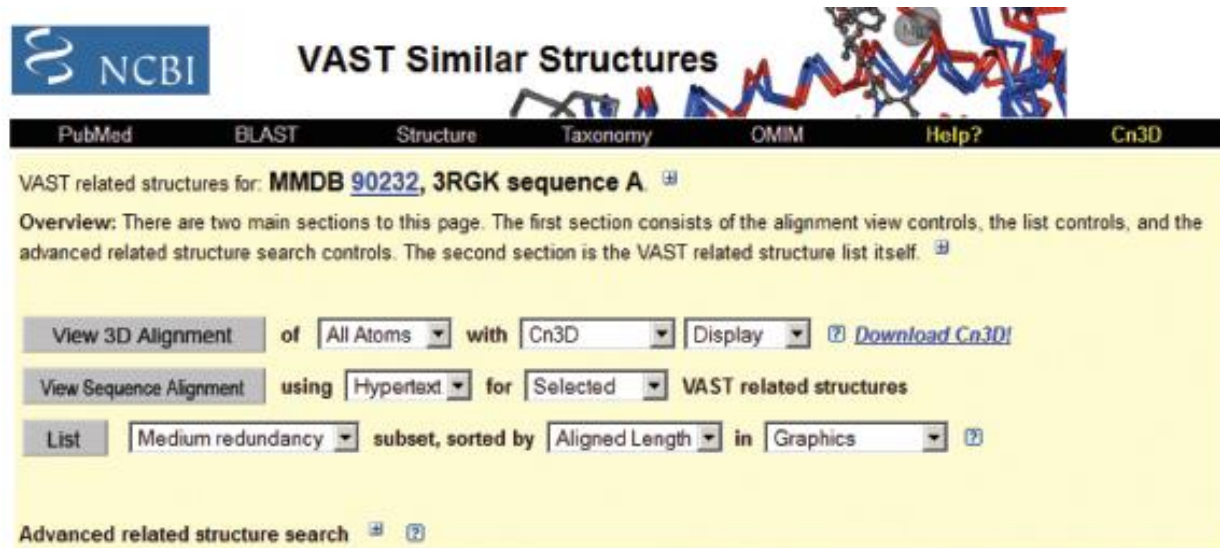
**Structure accession
(e.g. 2JTZ)**

Access to structure data at NCBI:VAST

Vector Alignment Search Tool (VAST) offers a variety of data on protein structures, including

- PDB identifiers
- root-mean-square deviation (RMSD) values to describe structural similarities
- NRES: the number of equivalent pairs of alpha carbon atoms superimposed
- percent identity

Vector Alignment Search Tool (VAST) at NCBI: comparison of two or more structures



NCBI

VAST Similar Structures

PubMed BLAST Structure Taxonomy OMIM Help? Cn3D

VAST related structures for: **MMDB 90232, 3RGK sequence A**

Overview: There are two main sections to this page. The first section consists of the alignment view controls, the list controls, and the advanced related structure search controls. The second section is the VAST related structure list itself.

View 3D Alignment of **All Atoms** with **Cn3D** Display [Download Cn3D!](#)

View Sequence Alignment using **Hypertext** for **Selected** VAST related structures

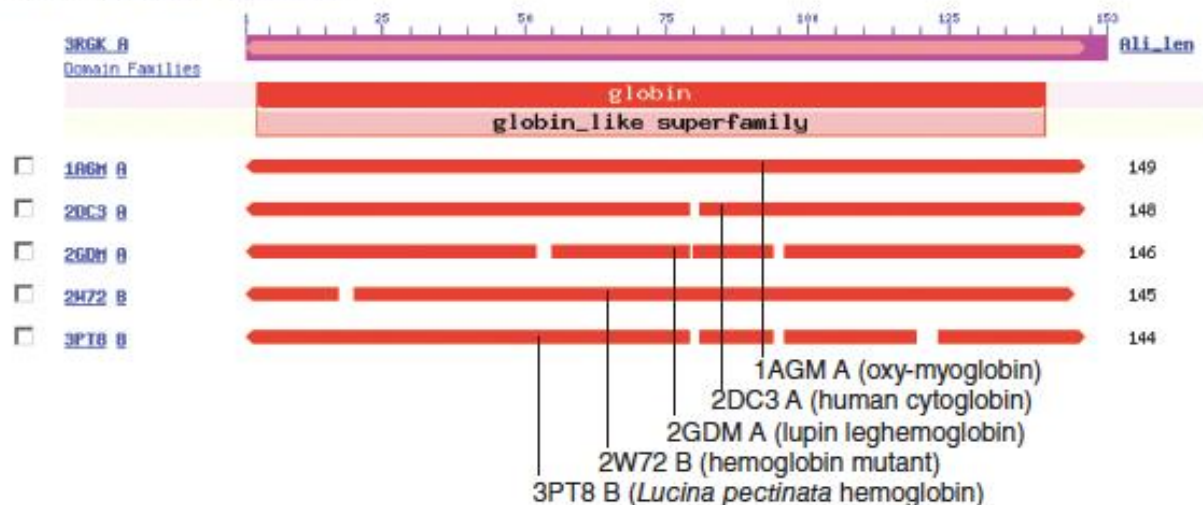
List **Medium redundancy** subset, sorted by **Aligned Length** in **Graphics**

Advanced related structure search

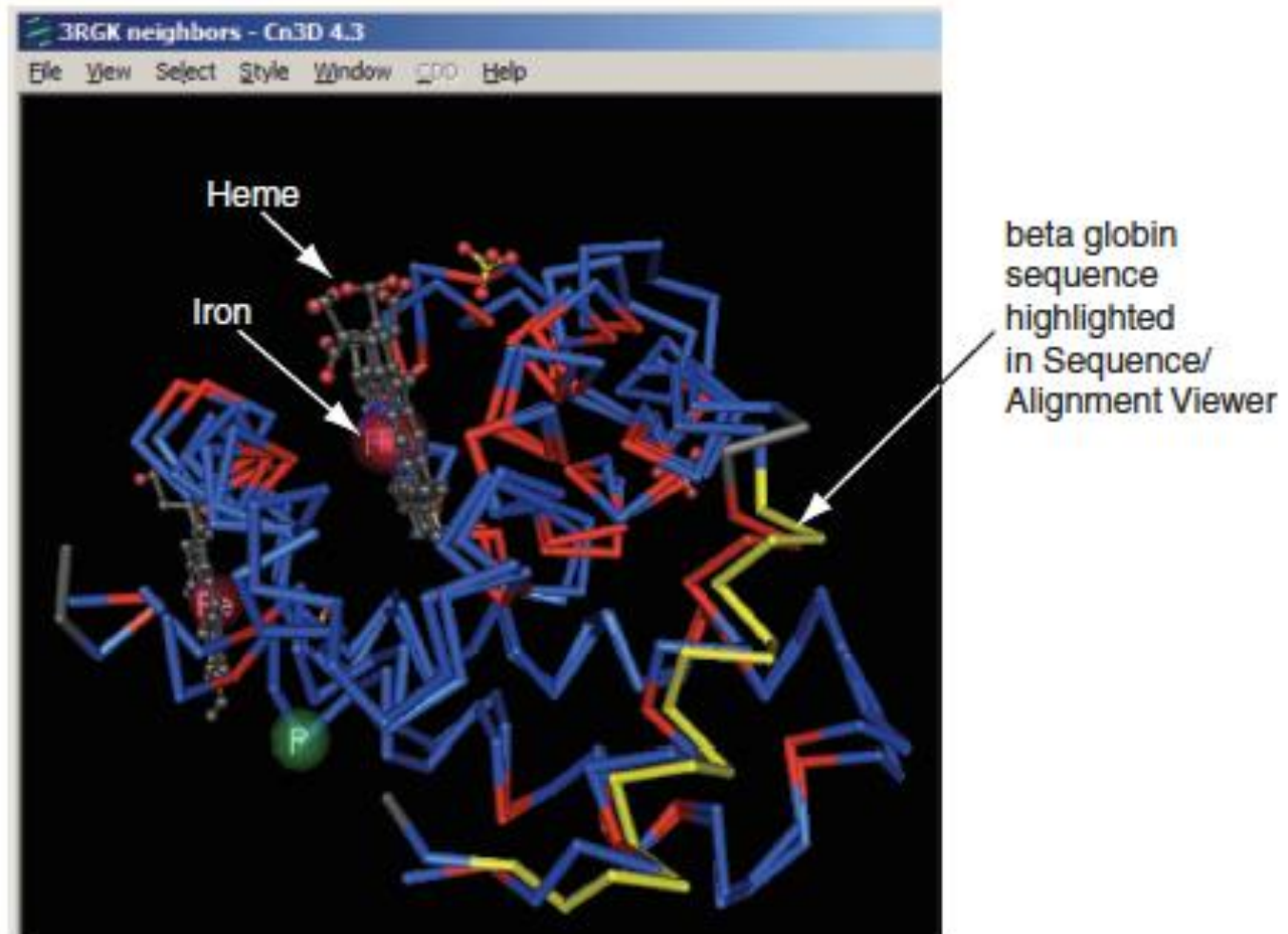
Move the mouse over the **red** alignment footprints in the graphics below and click, you will obtain a structure-based sequence alignment.

Total related structures: 2324; **1 - 60** of 80 representatives from the **Medium redundancy** subset displayed. **Page:** **1**

Click to: [Check All](#) [Uncheck All](#)



VAST: NCBI tool to compare two structures



3RGK neighbors - Sequence/Alignment Viewer

View Edit Mouse Mode Unaligned Justification Imports

3RGK_A	~GLSDGEWQLVLNVWGKV e a DIPGHGQEVLI RLFKGHPETLEK FDRFKHLKSEDEMK.
4HHB_B	v HLTPEEKSAVTALWGKV ~ ~ NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDVAM

Integrated views of universe of protein folds

- Chothia (1992) predicted a total of 1500 protein folds
- It is challenging to map protein fold space because of the varying definitions of domains, folds, and structural elements
- We can consider three resources: CATH, SCOP, and the Dali Domain Dictionary
- Structural Classification of Proteins (SCOP) database provides a comprehensive description of protein structures and evolutionary relationships based upon a hierarchical classification scheme. SCOPe is a SCOP extended database.

Holdings of the SCOP-e database






Class	Number of folds	Number of proteins
All alpha proteins	284	46,456
All beta proteins	174	48,724
Alpha and beta proteins (α/β)	147	51,349
Alpha and beta proteins ($\alpha + \beta$)	376	53,931
Multidomain proteins	66	56,572
Membrane and cell surface proteins	57	56,835
Small proteins	90	56,992
Coiled coil proteins	7	57,942
Low resolution protein structures	25	58,117
Peptides	120	58,231
Designed proteins	44	58,788
Total	1390	603,937

SCOP-e database: hierarchy of terms

<https://scop.berkeley.edu/>

The results of a search for myoglobin are shown, including its membership in a class (all alpha proteins), fold, superfamily, and family.

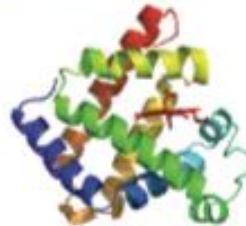
Lineage for Protein: Myoglobin

1. Root: [SCOPe 2.03](#)
2.  Class [a: All alpha proteins](#) [46456] (284 folds)
3.  Fold [a.1: Globin-like](#) [46457] (2 superfamilies)
core: 6 helices; folded leaf, partly opened
4.  Superfamily [a.1.1: Globin-like](#) [46458] (5 families) *S*
5.  Family [a.1.1.2: Globins](#) [46463] (27 protein domains)
Heme-binding protein
6.  Protein Myoglobin [46469] (9 species)

Species:

1.  [Asian elephant \(Elephas maximus\)](#) [TaxId:9783] [46476] (1 PDB entry)

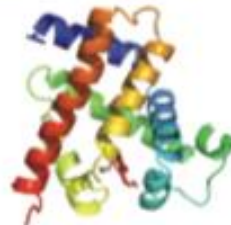
Domain for [1emy](#):



Domain [d1emya : 1emy A:](#) [15204]
complexed with cyn, hem

2.  [Common seal \(Phoca vitulina\)](#) [TaxId:9720] [46472] (1 PDB entry)

Domain for [1mbs](#):

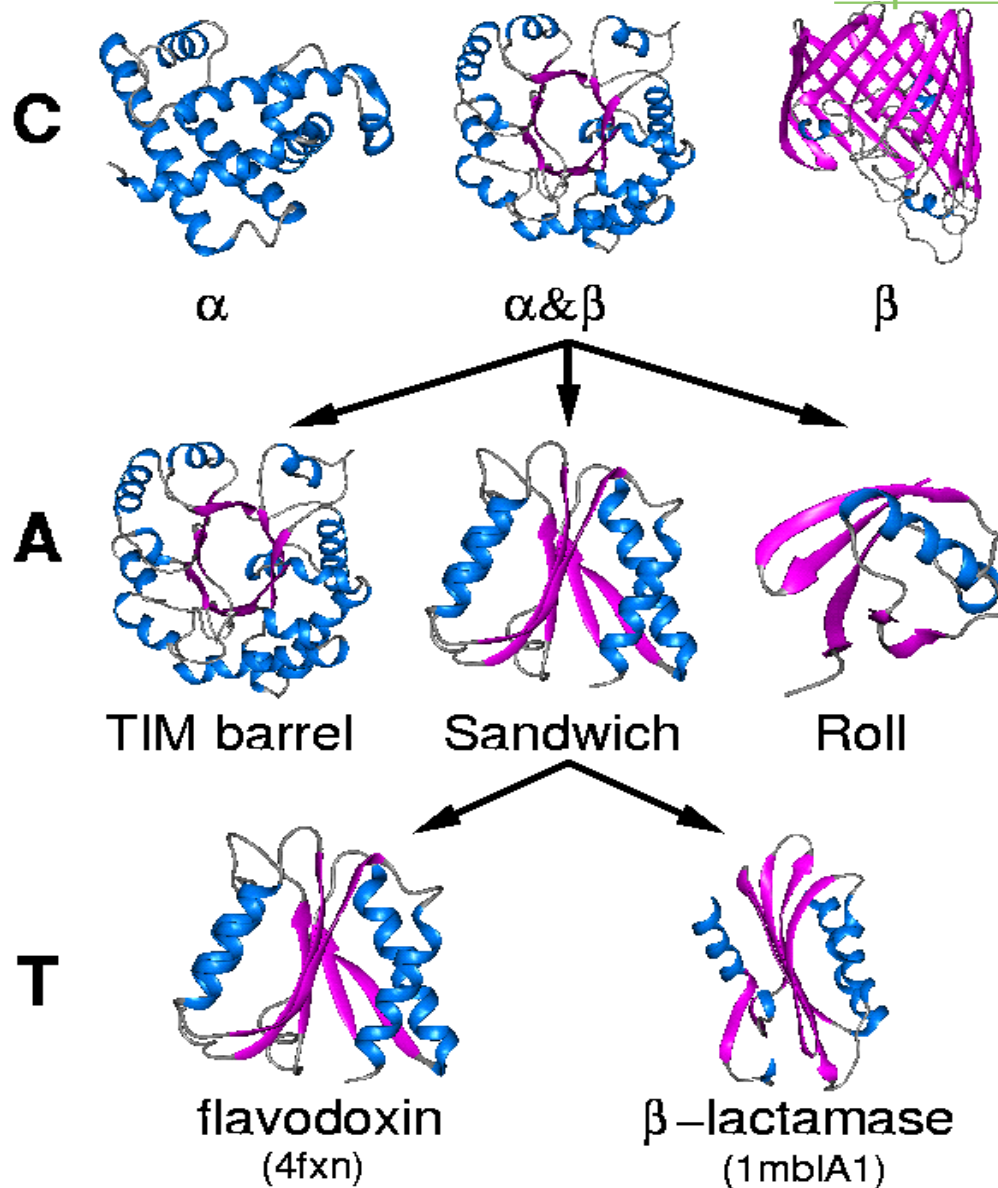


Domain [d1mbsa : 1mbs A:](#) [15156]
complexed with hem

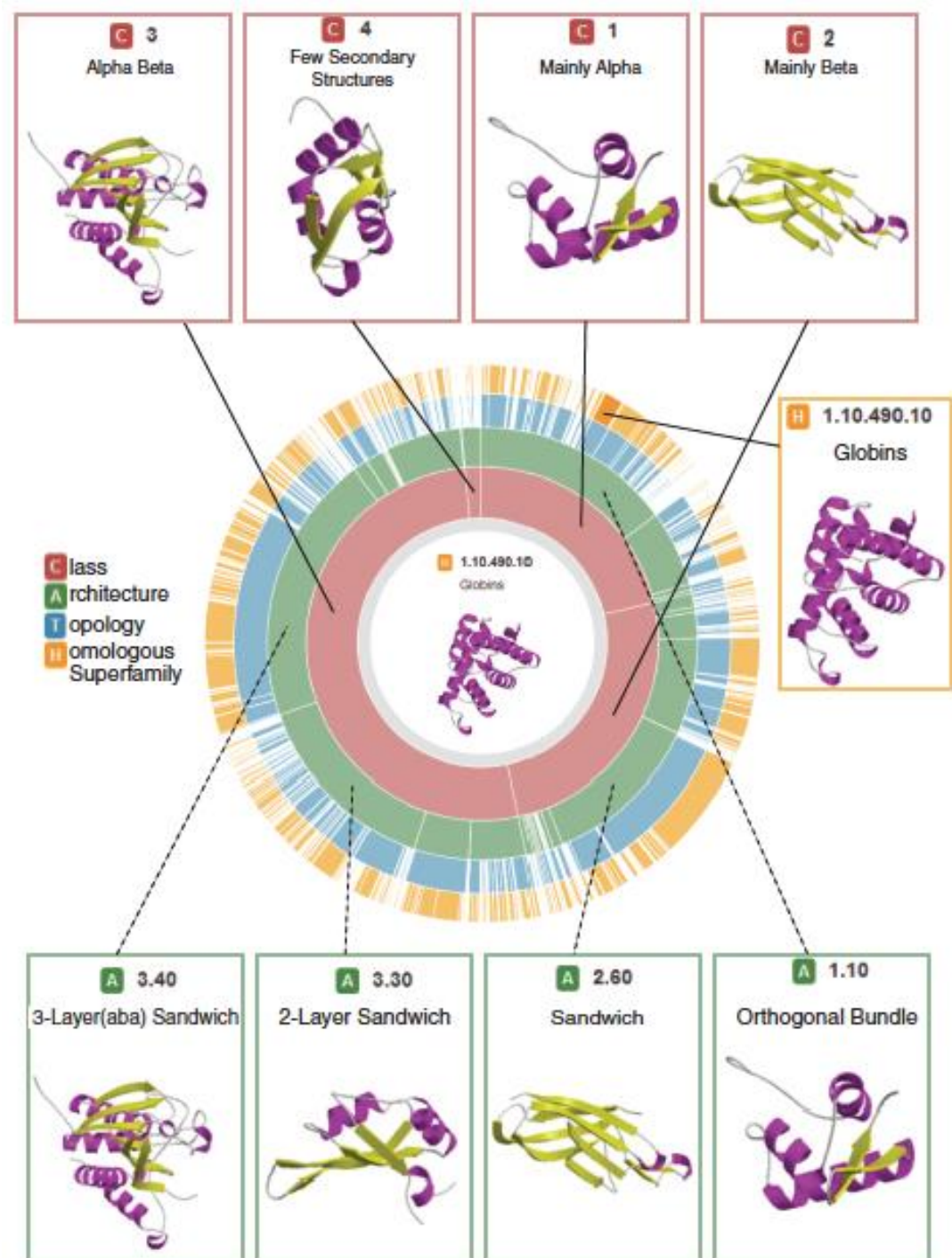
Two of the myoglobin structures are shown

The CATH Hierarchy

<https://www.cathdb.info/>

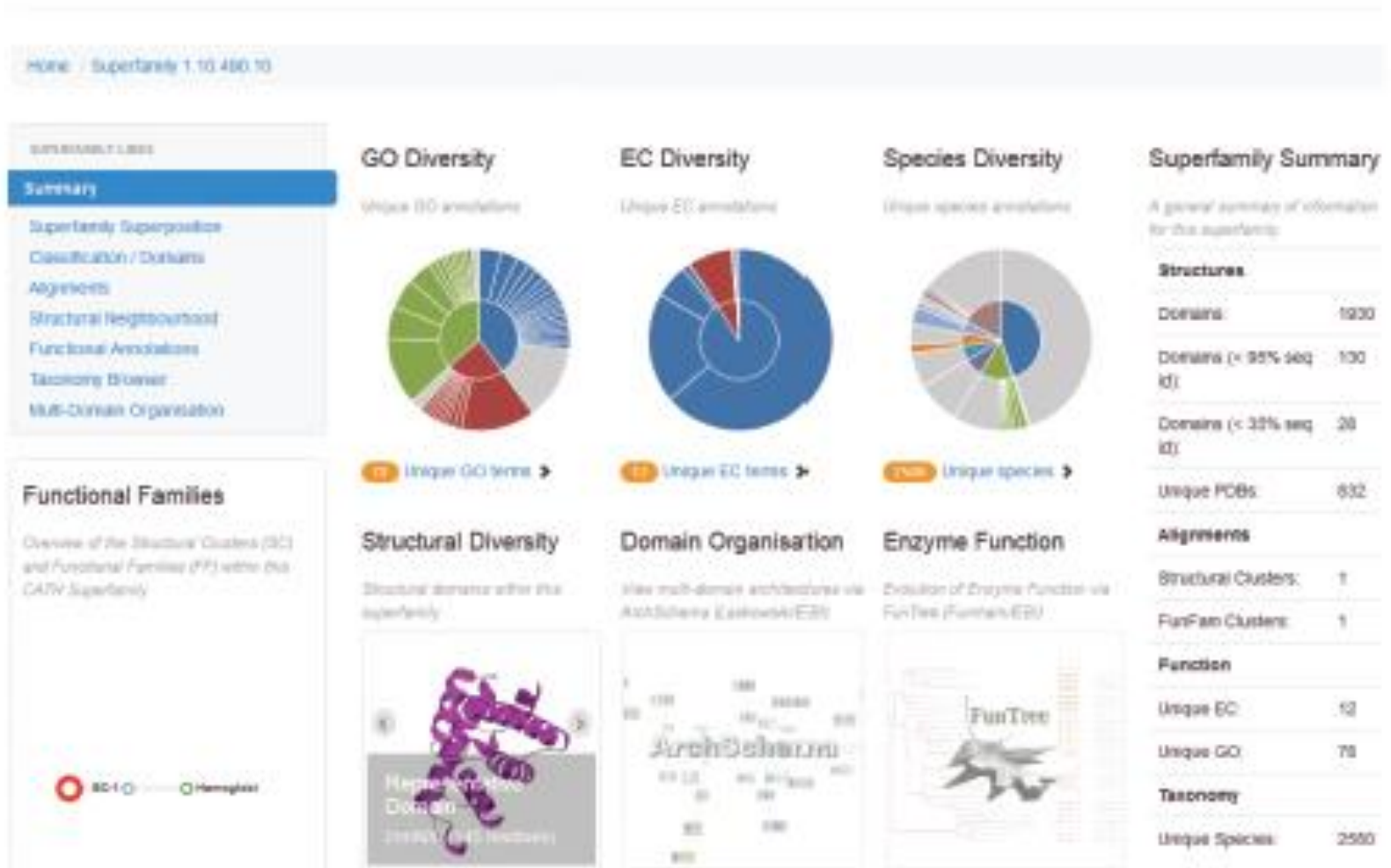


CATH organizes protein structures by a hierarchical scheme of class, architecture, topology (fold family), and homologous superfamily

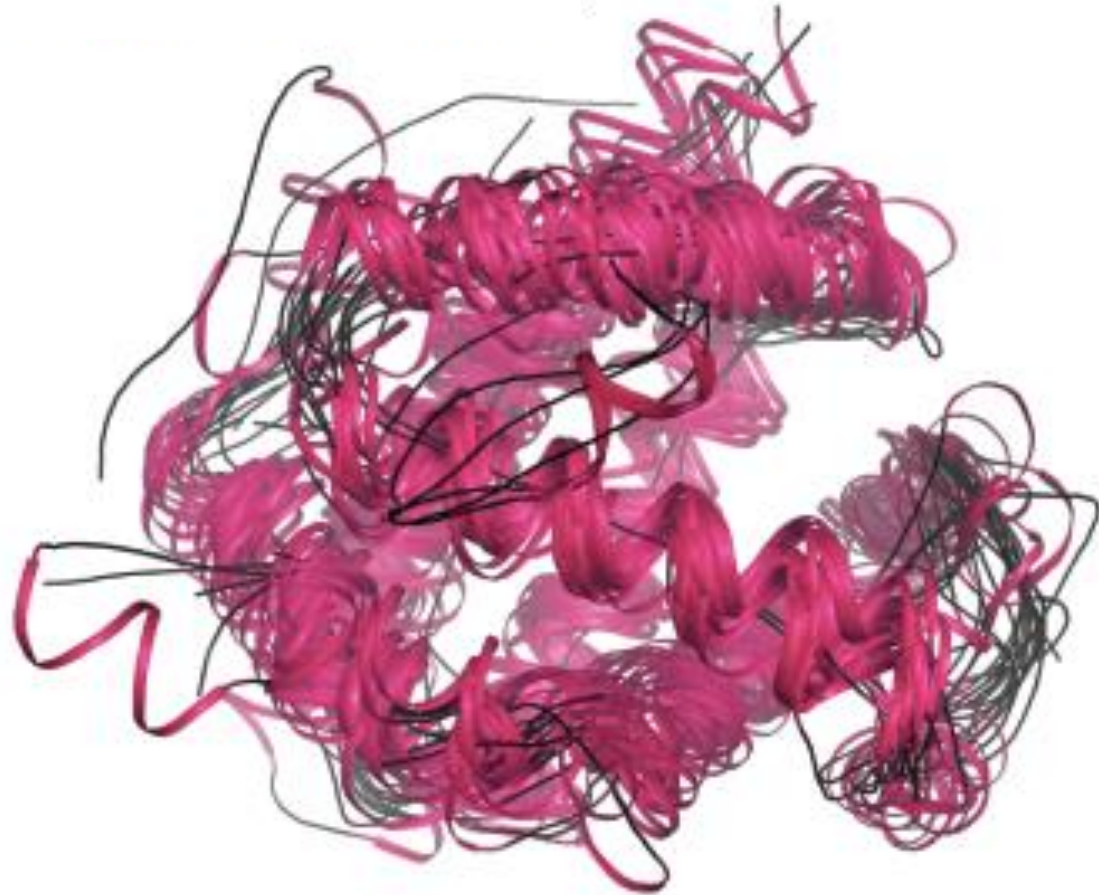


Globins are highlighted.

CATH globin superfamily



Superposition of globin superfamily members in CATH



DaliLite pairwise structural alignment: myoglobin and alpha globin

(a) DaliLite query form
(with myoglobin and alpha globin accessions)

Upload first structure (mol1):

No file selected.

Or enter PDB identifier: chain: (optional)

Upload second structure (mol2):

No file selected.

Or enter PDB identifier: chain: (optional)

(b) DaliLite structure comparison with Jmol



Dali is an acronym for distance matrix alignment. The Dali server allows a comparison of two 3D structures. Left: PDB identifiers for myoglobin and beta globin are entered. Right: output includes a pairwise structural alignment.

DALI server output includes a Z score (here a highly significant value of 21.4) based on quality measures such as: the resolution and amount of shared secondary structure; a root mean squared deviation (RMSD); percent identity; and a sequence alignment indicating secondary structure features.

Comparisons of SCOP, CATH, and Dali

For some proteins (such as those listed here) these three authoritative resources list different numbers of domains:


Name	PDB accession	SCOP	CATH	DALI
Glycogen phosphorylase	1gpb	1	2	3
Annexin V	1avh_A	1	4	4
Submaxillary renin	1smr_A	1	2	1
Fructose-1,6-bisphosphatase	5fbp_A	1	2	2

The field of structural biology provides rigorous measurements of the three-dimensional structure of proteins, and yet classifying domains can be a complex problem requiring expert human judgments. SCOP is especially oriented towards classifying whole proteins, while CATH is oriented towards classifying domains.



Beyond PDB, CATH, SCOP, Dali: partial list of protein structure databases

Database	Comment	URL
3dee	Structural domain definitions	http://www.compbio.dundee.ac.uk/3Dee/
Enzyme Structures Databases	Enzyme classifications and nomenclature	http://www.ebi.ac.uk/thomson-srv/databases/enzymes/
FATCAT	Flexible structure alignment by chaining aligned fragment pairs allowing twists	http://fatcat.burnham.org/
PDBeFold	Secondary-structure matching for fast protein structure alignment in three dimensions	http://www.ebi.ac.uk/msd-srv/ssm/
PDBePISA	Proteins, interfaces, structures and assemblies	http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html
NDB	Database of three-dimensional nucleic acid structures	http://ndbserver.rutgers.edu/
PDBSum	Summary information about protein structures	http://www.ebi.ac.uk/pdbsum/
SWISS-MODEL Repository	Database of annotated three-dimensional comparative protein structure models	http://swissmodel.expasy.org/repository/



Outline

Overview of protein structure

Principles of protein structure

Protein Data Bank

Protein structure prediction

Intrinsically disordered proteins

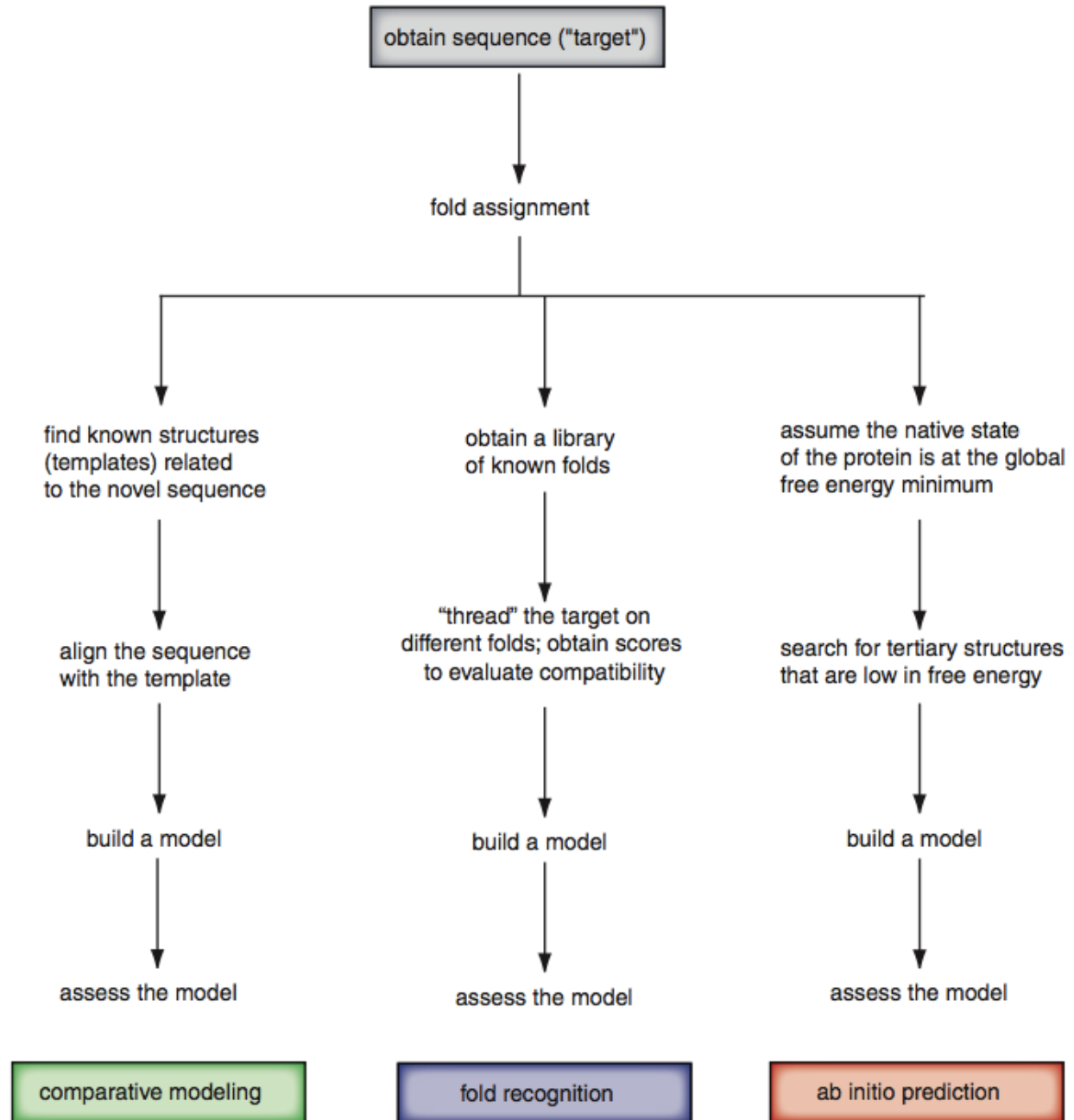
Protein structure and disease

Three main structure prediction strategies

There are three main approaches to protein structure prediction.

1. Homology modeling (comparative modeling). This is most useful when a template (protein of interest) can be matched (e.g. by BLAST) to proteins of known structure.
2. Fold recognition (threading). A target sequence lacks identifiable sequence matches and yet may have folds in common with proteins of known structure.
3. *Ab initio* prediction (template-free modeling). Assumes: (1) all the information about the structure of a protein is contained in its amino acid sequence; and (2) a globular protein folds into the structure with the lowest free energy.

Three main structure prediction strategies



Structure prediction techniques as a function of sequence identity

sequence identity	model accuracy	resolution	technique	applications
100%	100%	1.0 Å	X-ray crystallography, NMR	Studying catalytic mechanisms Designing and improving ligands Prediction of protein partners
50%	95%	1.5 Å	comparative protein structural modeling	Defining antibody epitopes Supporting site-directed mutagenesis
30%	80%	3.5 Å	threading	Refining NMR structures Fitting into low-resolution electron density
<<20%	80 aa	4-8 Å	de novo structure prediction	Identifying regions of conserved surface residues

Websites for structure prediction by comparative modeling, and for quality assessment

Website	Comment	URL
3D-JIGSAW	Laboratory of Paul Bates	http://bmm.cancerresearchuk.org/~3djigsaw/
Geno3D	POLE	http://pbil.ibcp.fr/htm/index.php
MODELLER	From Andrej Sali's group	http://www.salilab.org/modeller/
PredictProtein	Laboratory of Burkhard Rost	http://www.predictprotein.org/
SWISS-MODEL	ExPASy	http://swissmodel.expasy.org/
PROCHECK	Quality assessment	http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/
VERIFY3D	Quality assessment	http://nihserver.mbi.ucla.edu/Verify_3D/
WHATIF	Quality assessment	http://swift.cmbi.ru.nl/whatif/

Predicting protein structure: CASP competition

Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (“CASP”)

<http://predictioncenter.org/>

The competition organizers and selected experts solve a wide range of three dimensional structures (typically by X-ray crystallography) and hold the correct answers. Participants in the competition are given the primary amino acid sequence and a set amount of time to submit predictions. These predictions are then assessed by comparison to the correct structures.

CASP allows the structural biology community to assess which methods perform best (and to identify challenging areas).

Outline

Overview of protein structure

Principles of protein structure

Protein Data Bank

Protein structure prediction

Intrinsically disordered proteins

Protein structure and disease

Intrinsic disorder

- Many proteins do not adopt stable three-dimensional structures, and this may be an essential aspect of their ability to function properly.
- Intrinsically disordered proteins are defined as having unstructured regions of significant size such as at least 30 or 50 amino acids.
- Such regions do not adopt a fixed three-dimensional structure under physiological conditions, but instead exist as dynamic ensembles in which the backbone amino acid positions vary over time without adopting stable equilibrium values.
- The Database of Intrinsic Disorder is available at <http://www.disprot.org>.

Outline

Overview of protein structure
Principles of protein structure
Protein Data Bank
Protein structure prediction
Intrinsically disordered proteins
Protein structure and disease

Protein structure and human disease

In some cases, a single amino acid substitution can induce a dramatic change in protein structure. For example, the DF508 mutation of CFTR alters the helical content of the protein, and disrupts intracellular trafficking.

Other changes are subtle. The E6V mutation in the gene encoding hemoglobin beta causes sickle-cell anemia. The substitution introduces a hydrophobic patch on the protein surface, leading to clumping of hemoglobin molecules.

Protein structure and disease

Disease	OMIM	Gene/Protein	RefSeq	PDB
Alzheimer disease	#104300	Amyloid precursor protein	NP_000475.1	2M4J
Cystic fibrosis	#219700	CFTR	NP_000483.3	2LOB
Huntington disease	#143100	Huntingtin	NP_002102.4	4FED
Creutzfeldt-Jakob disease	#123400	Prion protein	NP_000302.1	2M8T
Parkinson disease	#168600	alpha-synuclein isoform NACP140	NP_000336.1	2M55
Sickle cell anemia	#603903	Hemoglobin beta	NP_000509.1	2M6Z

Examples of proteins associated with diseases for which subtle change in protein sequence leads to change in structure.

Perspective

The aim of structural genomics is to define structures that span the entire space of protein folds. This project has many parallels to the Human Genome Project. Both are ambitious endeavors that require the international cooperation of many laboratories. Both involve central repositories for the deposit of raw data, and in each the growth of the databases is exponential.

It is realistic to expect that the great majority of protein folds will be defined in the near future. Each year, the proportion of novel folds declines rapidly. A number of lessons are emerging:

- proteins assume a limited number of folds;
- a single three-dimensional fold may be used by proteins to perform entirely distinct functions; and
- the same function may be performed by proteins using entirely different folds.