Chapter 13
Protein structure

Bioinformatics and Functional Genomics
3rd edition (Wiley, 2015)
Jonathan Pevsner, Ph.D.

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

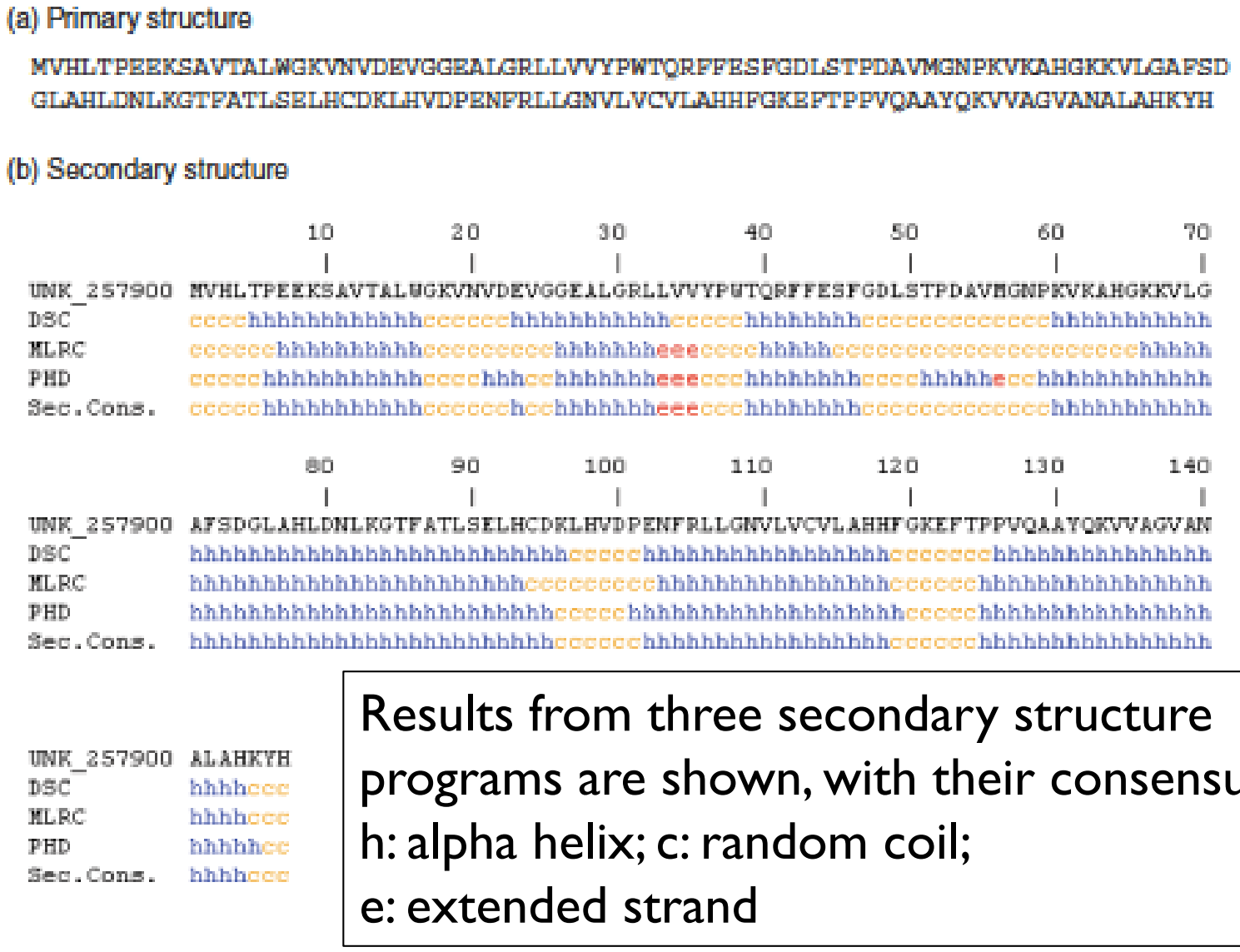
Protein Data Bank

Protein structure prediction

Intrinsically disordered proteins

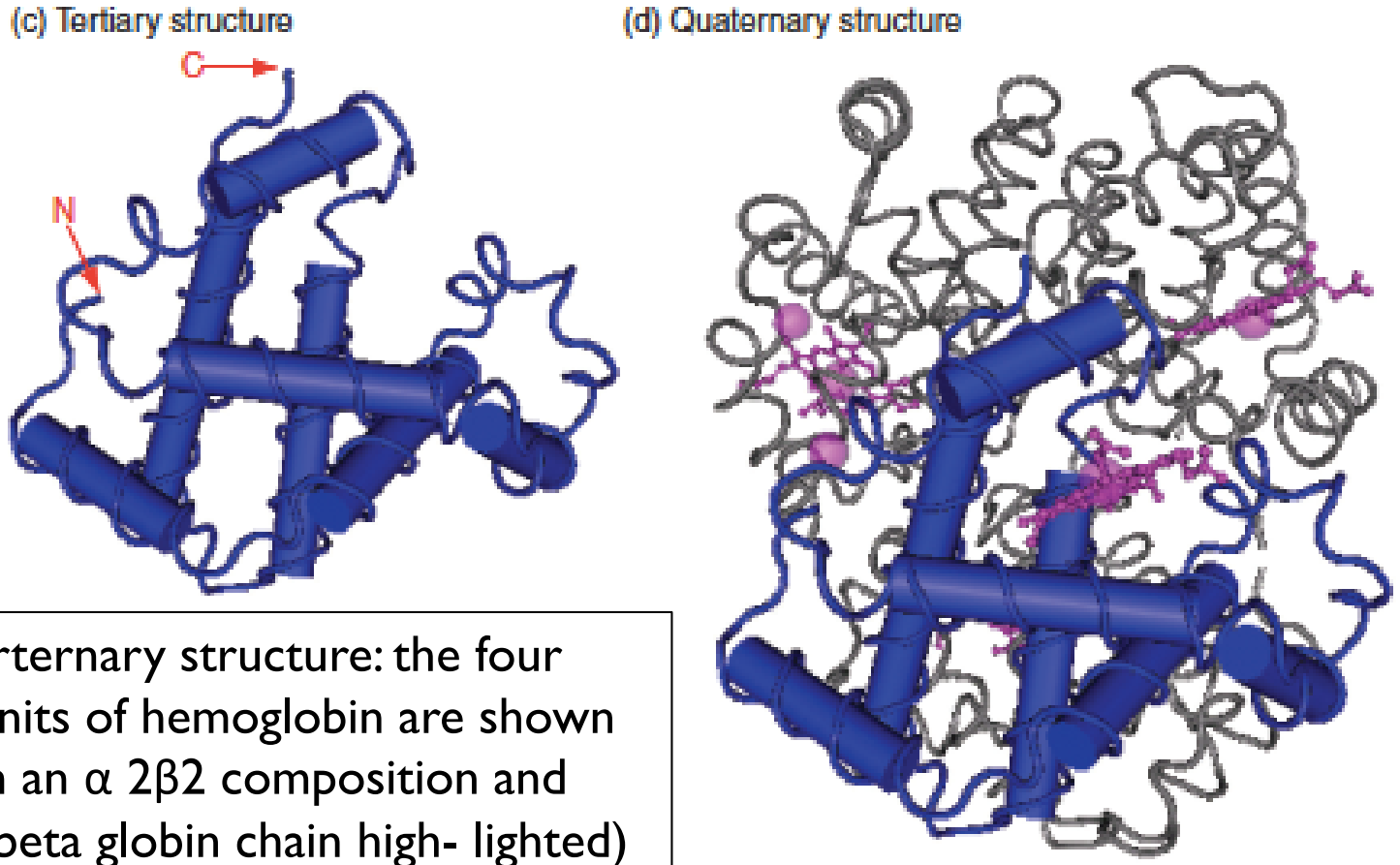
Protein structure and disease

Protein primary and secondary structure



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

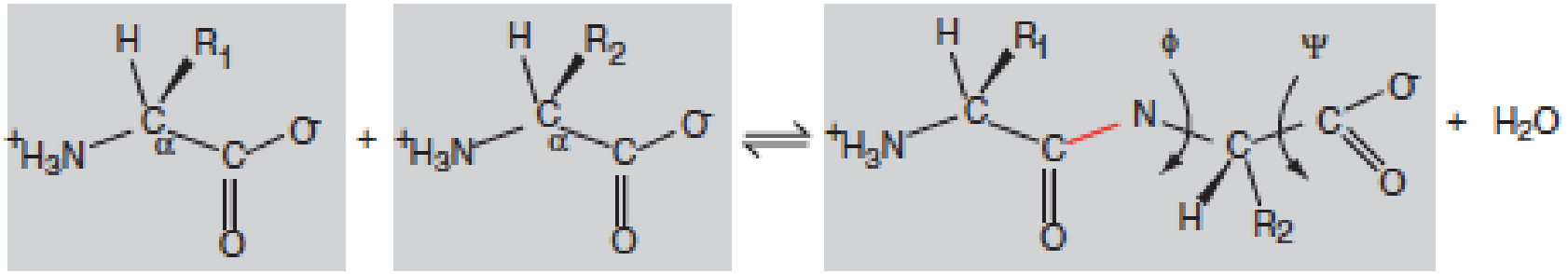


Quarternary structure: the four subunits of hemoglobin are shown (with an α 2β2 composition and one beta globin chain high- lighted) as well as four noncovalently attached heme groups.

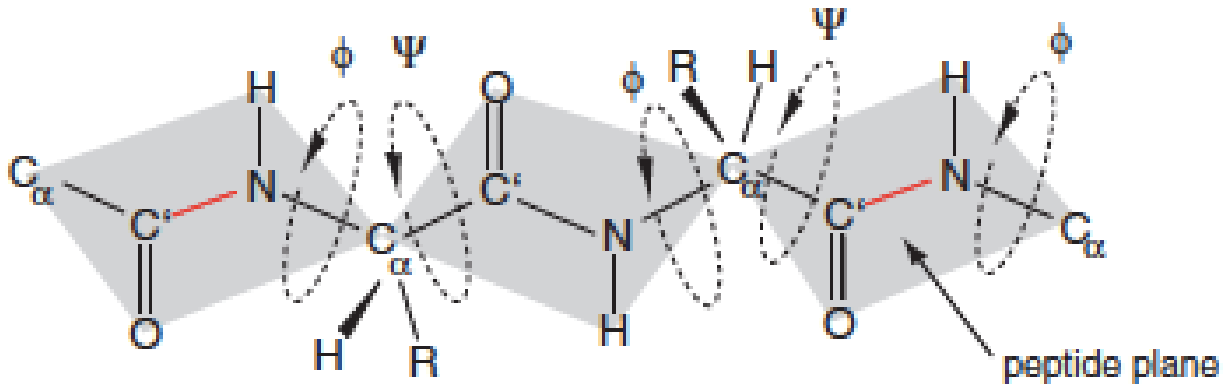
B&FG 3e
Fig. 13.1
Page 592

The peptide bond; phi and psi angles

(a) peptide bond



(b) phi and psi angles of polypeptide



B&FG 3e
Fig. 13.2
Page 593

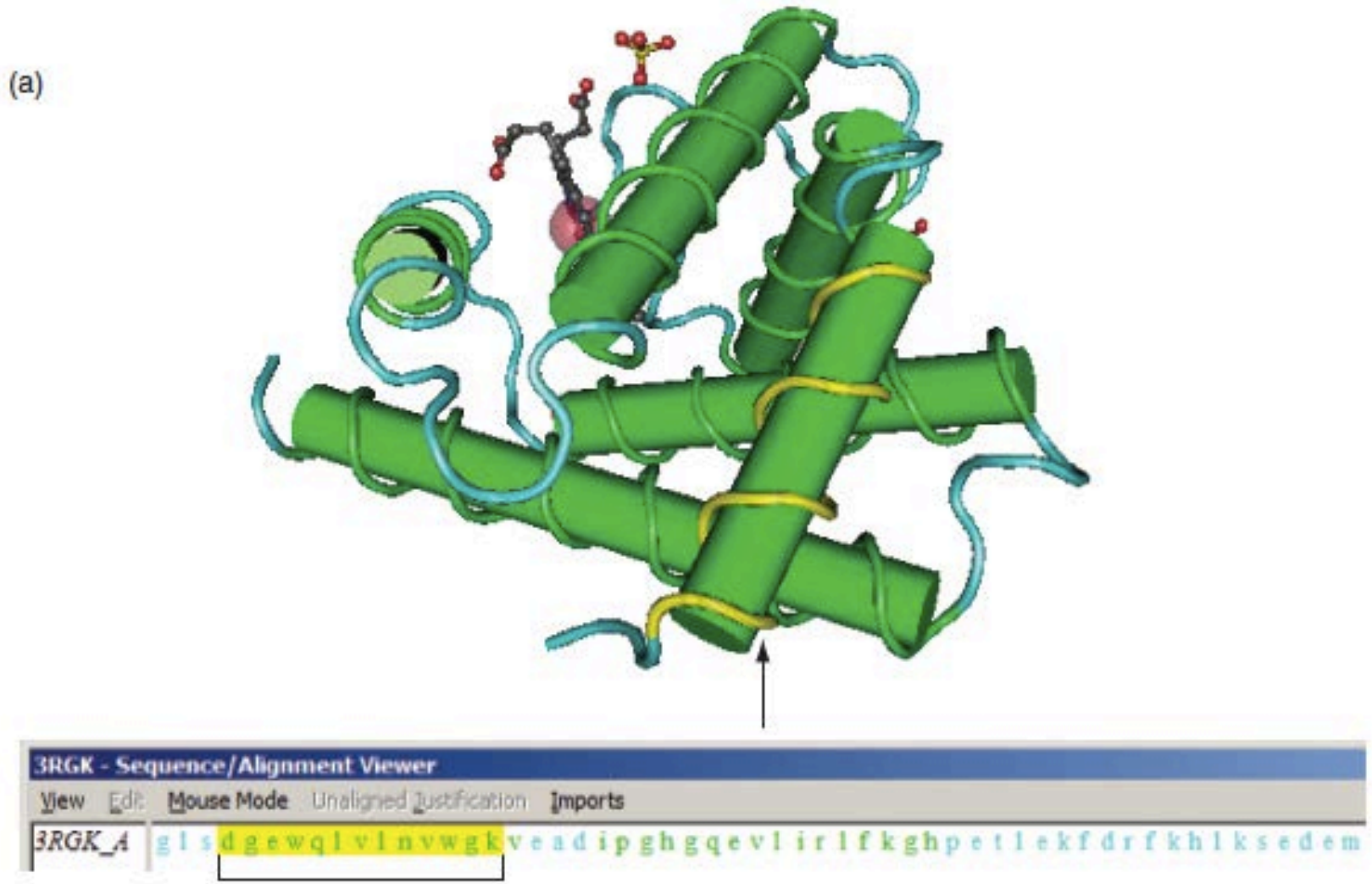
Protein secondary structure

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.

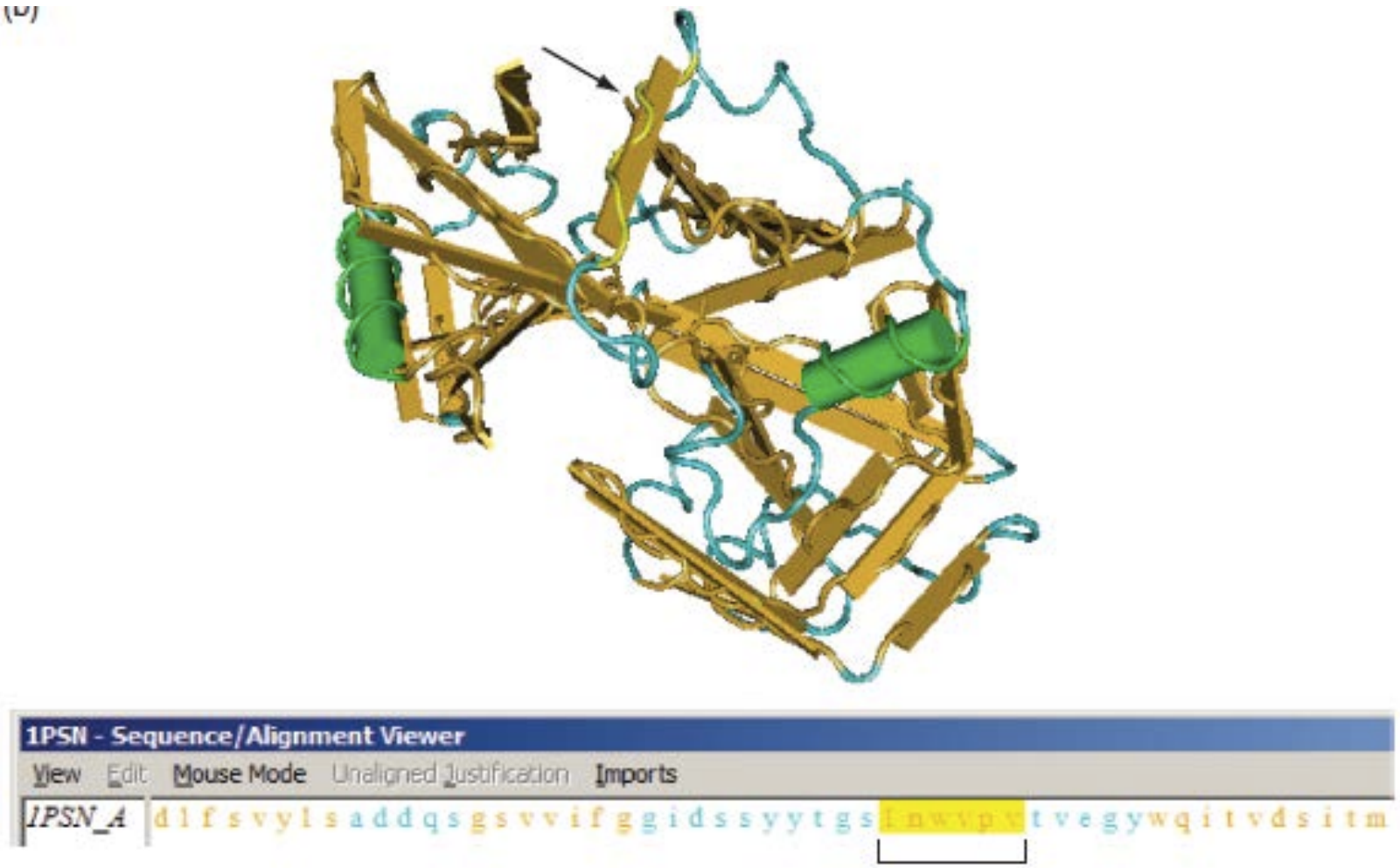
Protein secondary structure: myoglobin (alpha helical)



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

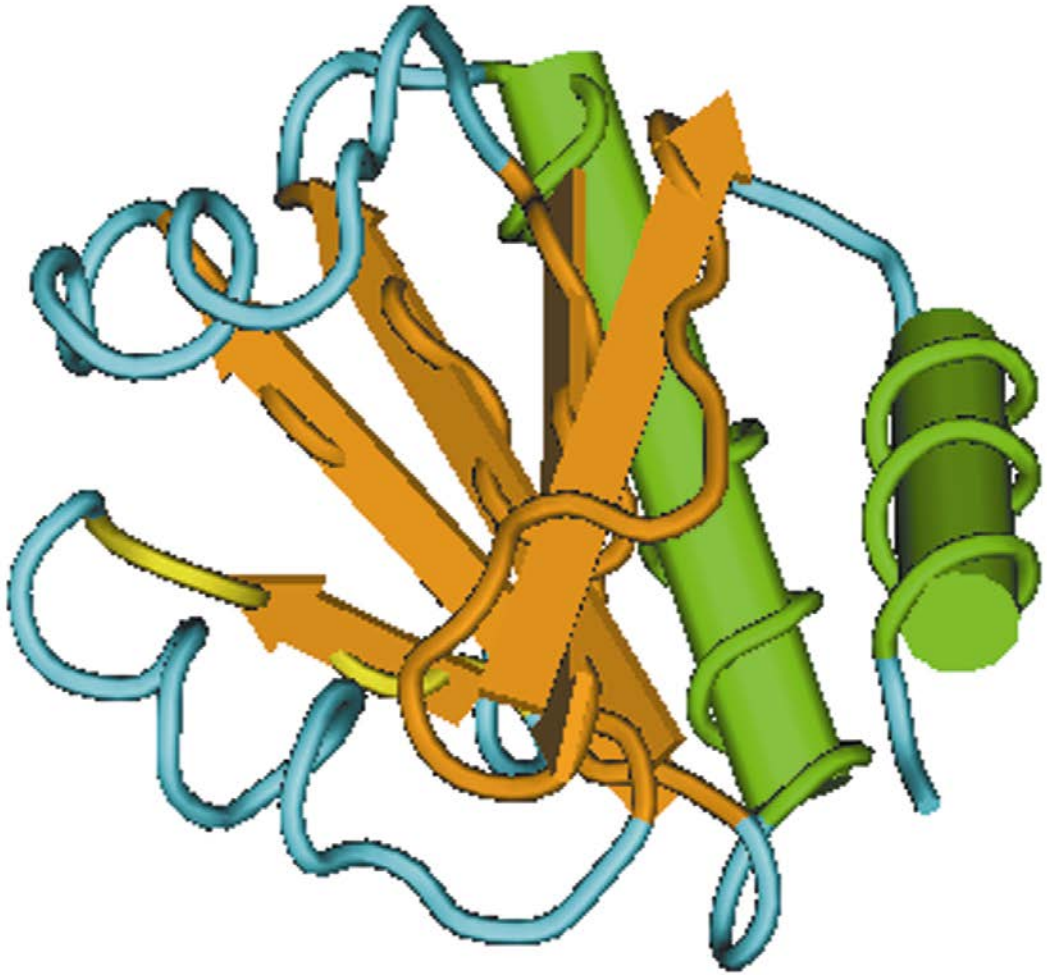
B&FG 3e
Fig. 13.3
Page 595

Protein secondary structure: pepsin (beta sheets)



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.

B&FG 3e
Fig. 13.3
Page 595



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

Secondary structure prediction

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

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

GOR (Garnier, Osguthorpe, Robson): related algorithm

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

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**	

B&FG 3e
Tab. 13.1
Page 597

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.

B&FG 3e
Tab. 13.3
Page 598

Tertiary protein structure: protein folding

Main approaches:

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

[2] Prediction

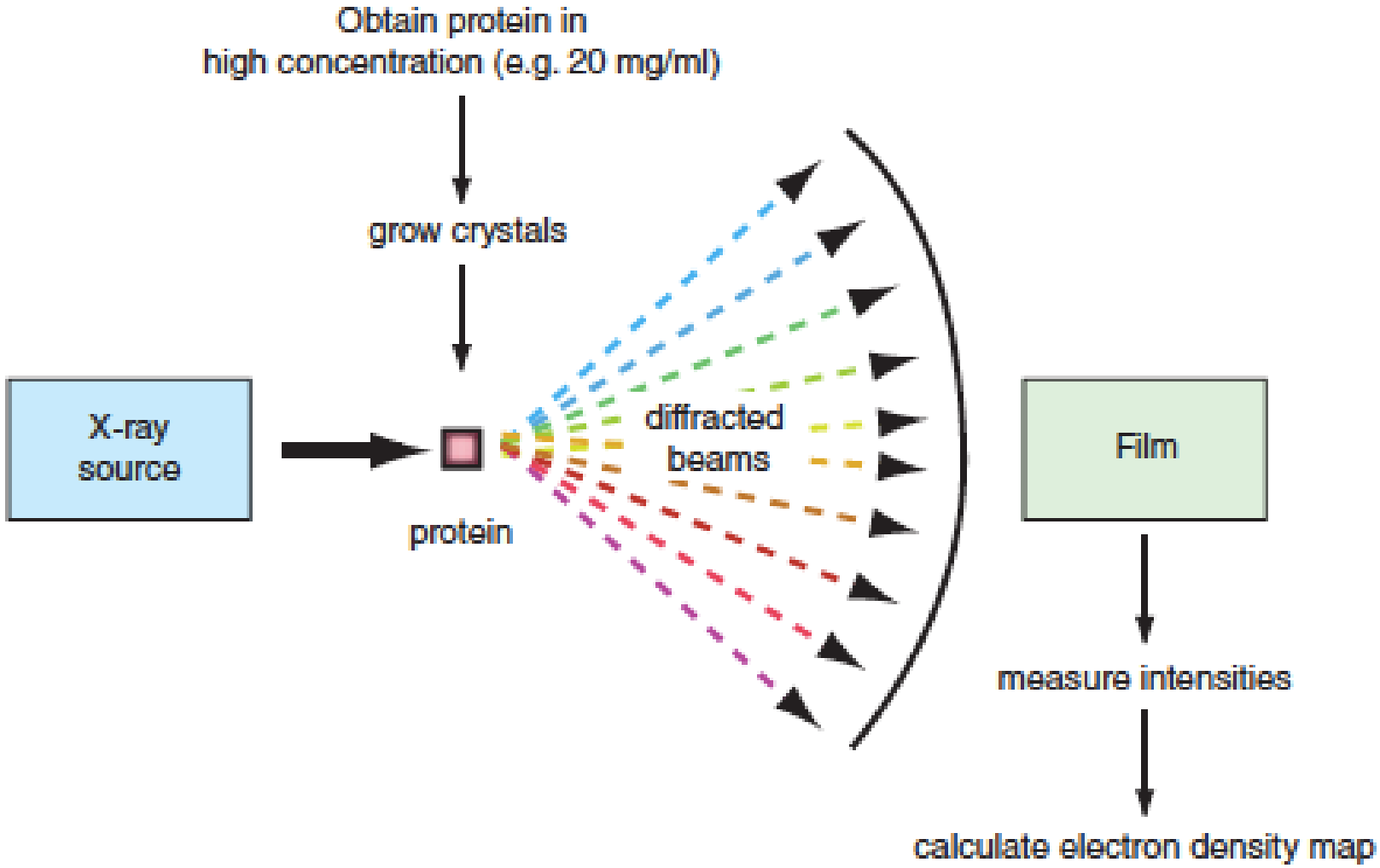
- Comparative modeling (based on homology)
- Threading
- *Ab initio* (de novo) prediction

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

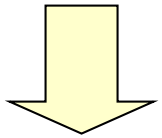
X-ray crystallography



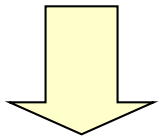
B&FG 3e
Box 13.1
Page 599

Steps in obtaining a protein structure

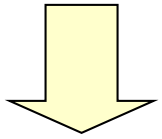
Target selection



Obtain, characterize protein

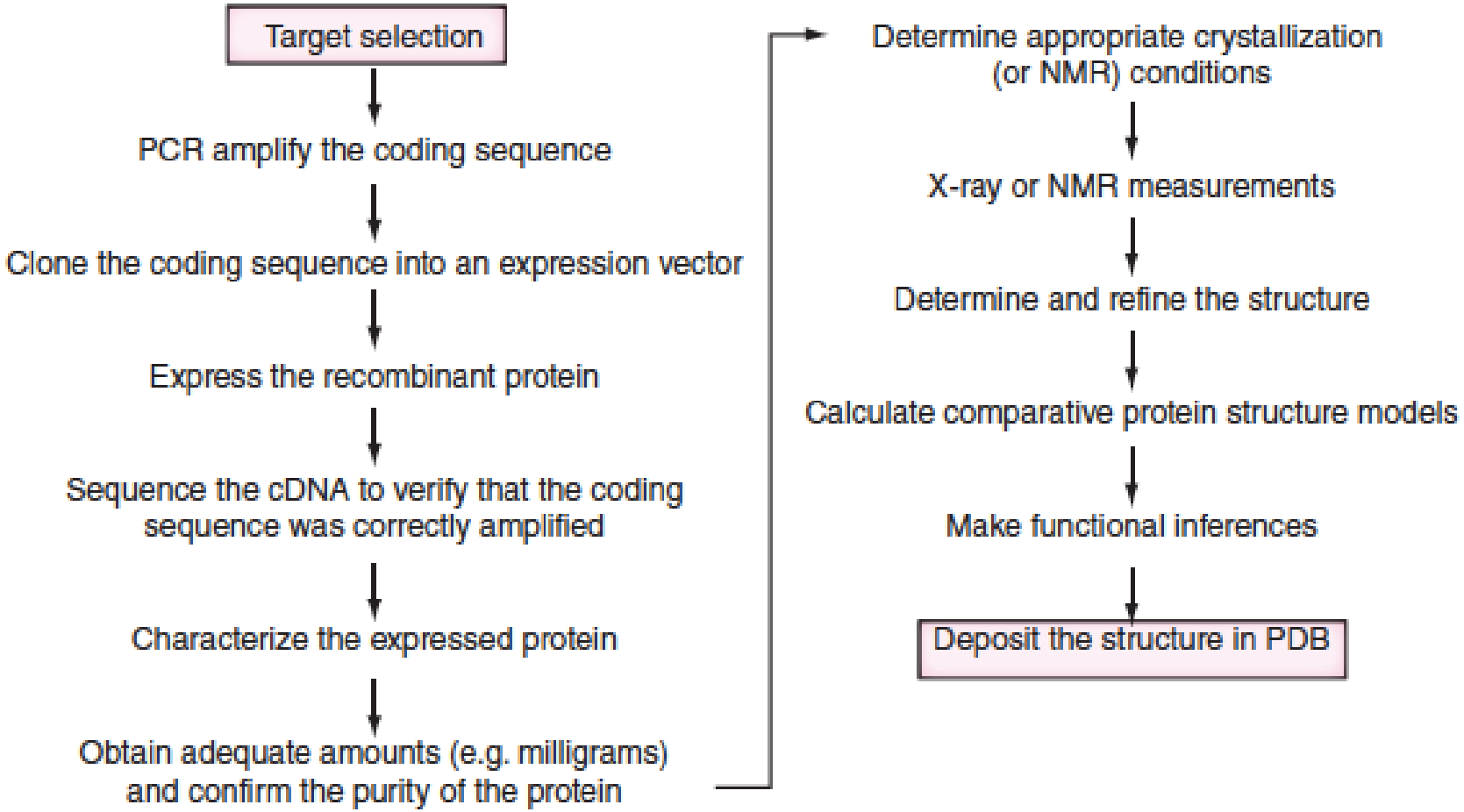


Determine, refine, model the structure



Deposit in repository

Target selection for protein structure determination



B&FG 3e
Fig. 13.5
Page 599

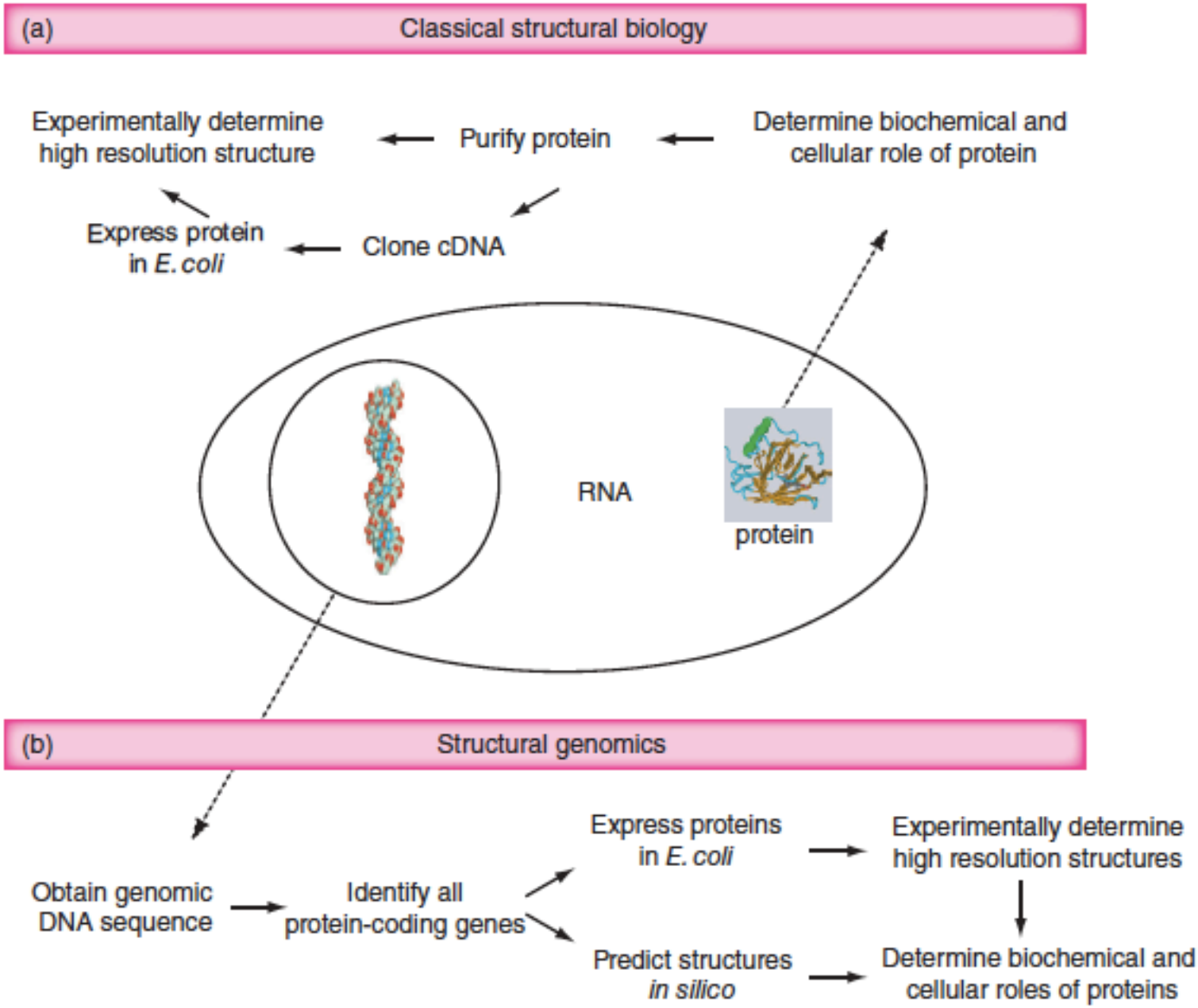
Priorities for target selection for protein structures

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



B&FG 3e
Fig. 13.6
Page 601

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

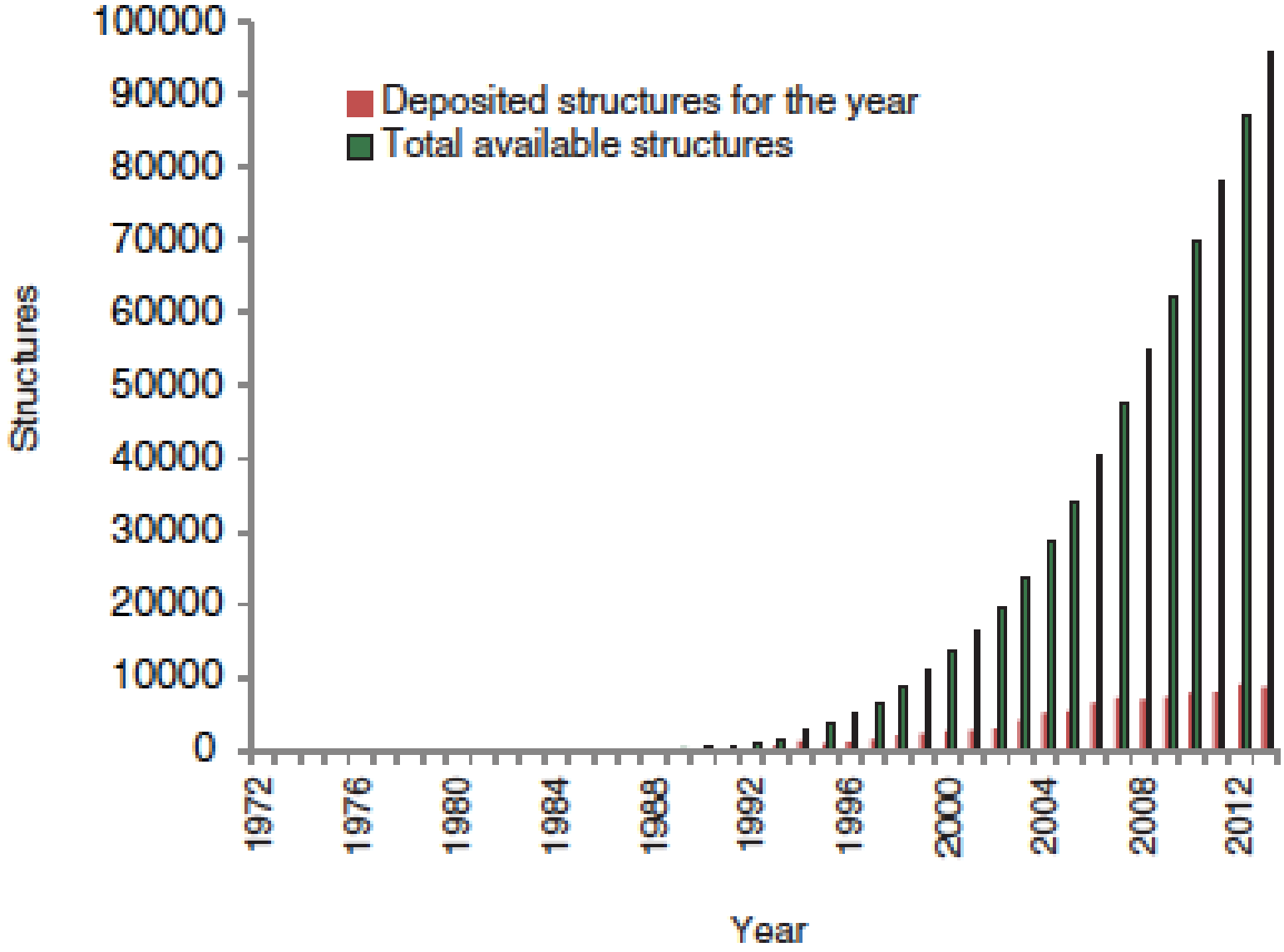
Protein Data Bank (PDB) holdings

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

B&FG 3e
Tab. 13.4
Page 604

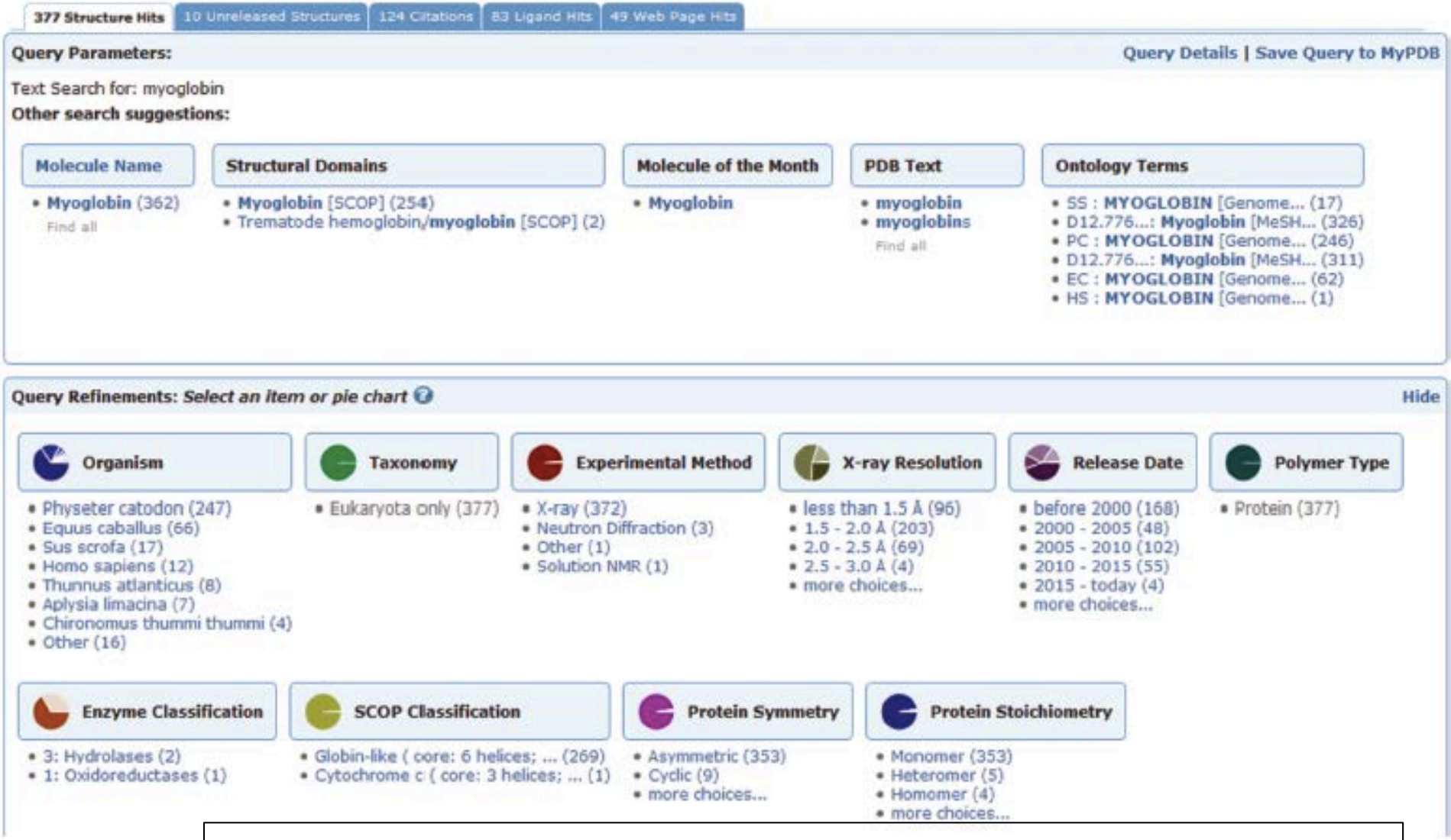
Accessed 2015

PDB: number of searchable structures per year



B&FG 3e
Fig. 13.8
Page 604

PDB query for myoglobin



B&FG 3e
Fig. 13.9
Page 605

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.

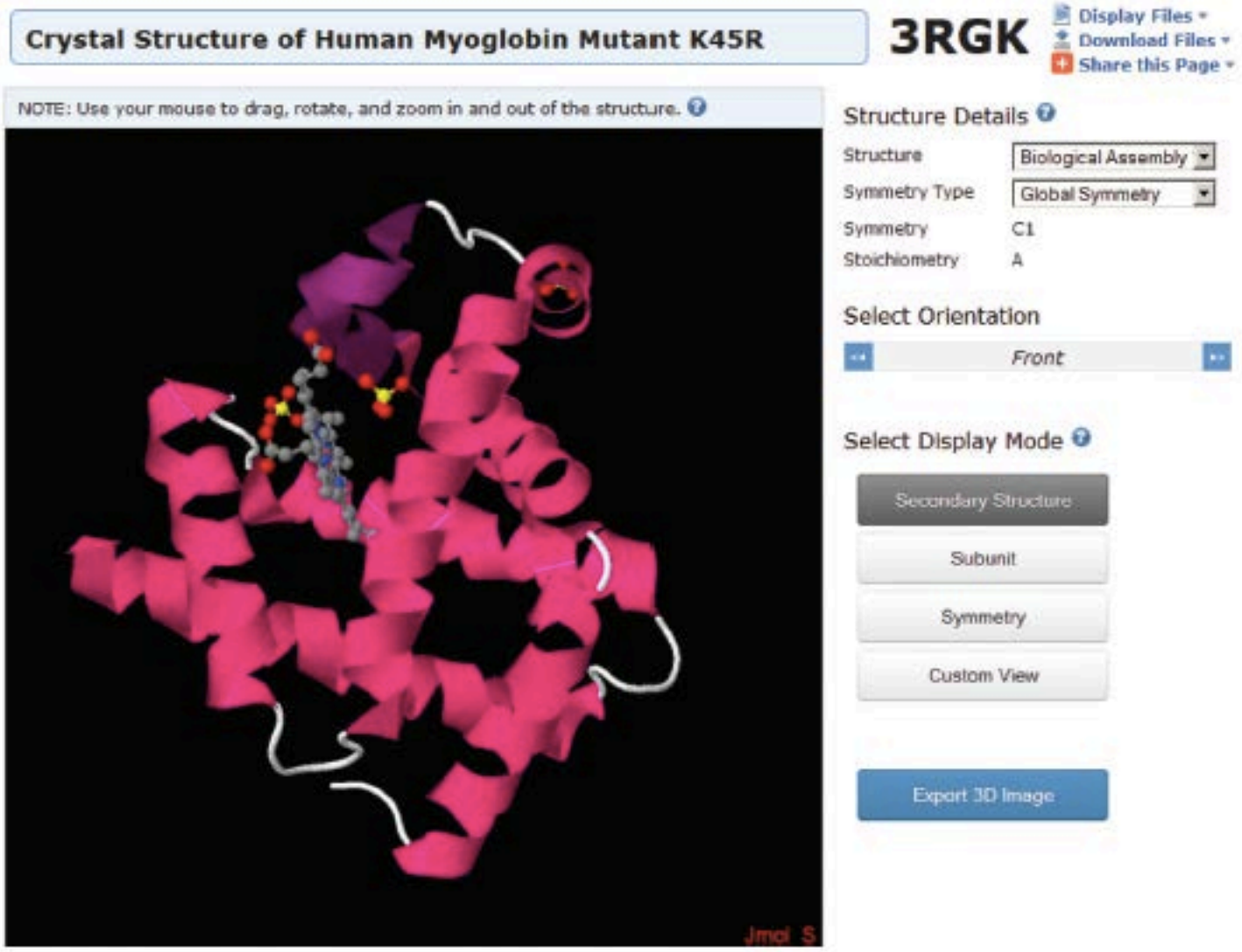
Interactive visualization tools for PDB protein structures

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/

B&FG 3e
Tab. 13.5
Page 606

Visualization tools are available within PDB and elsewhere.

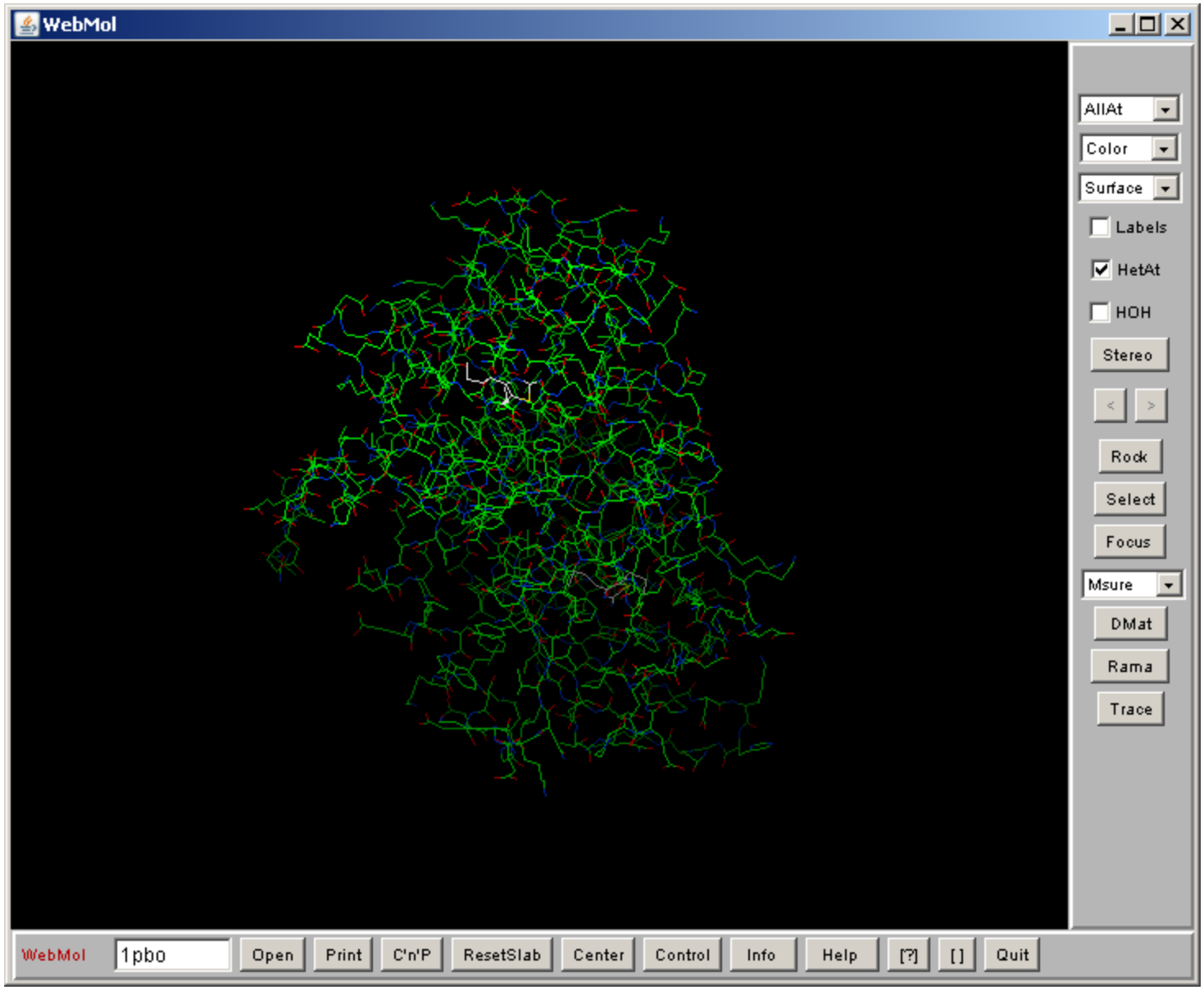
Visualizing myoglobin structure 3RGK: Jmol applet



Jmol is available at PDB.

B&FG 3e
Fig. 13.11
Page 607

Viewing structures at PDB: WebMol



gateways to access PDB files

Swiss-Prot, NCBI, EMBL

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

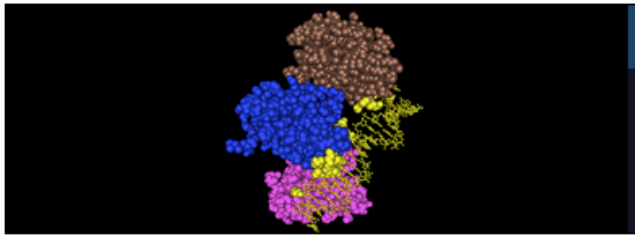
Vector Alignment Search Tool (VAST):
view multiple structures

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 pevsner My NCBI Sign Out

Structure

Structure odorant Search

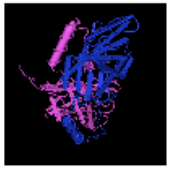
Create alert Advanced Help

Summary 20 per page Sort by Default order Send to:

Search results

Items: 1 to 20 of 177

<< First < Prev Page 1 of 9 Next > Last >>

1. 

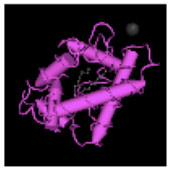
Complex Of Bovine **Odorant** Binding Protein (obp) With A Selenium Containing **Odorant**[**Odorant**-Binding]

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

Filter your results:

All (177)

NMR (14)

X-ray (159)

Manage Filters

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)

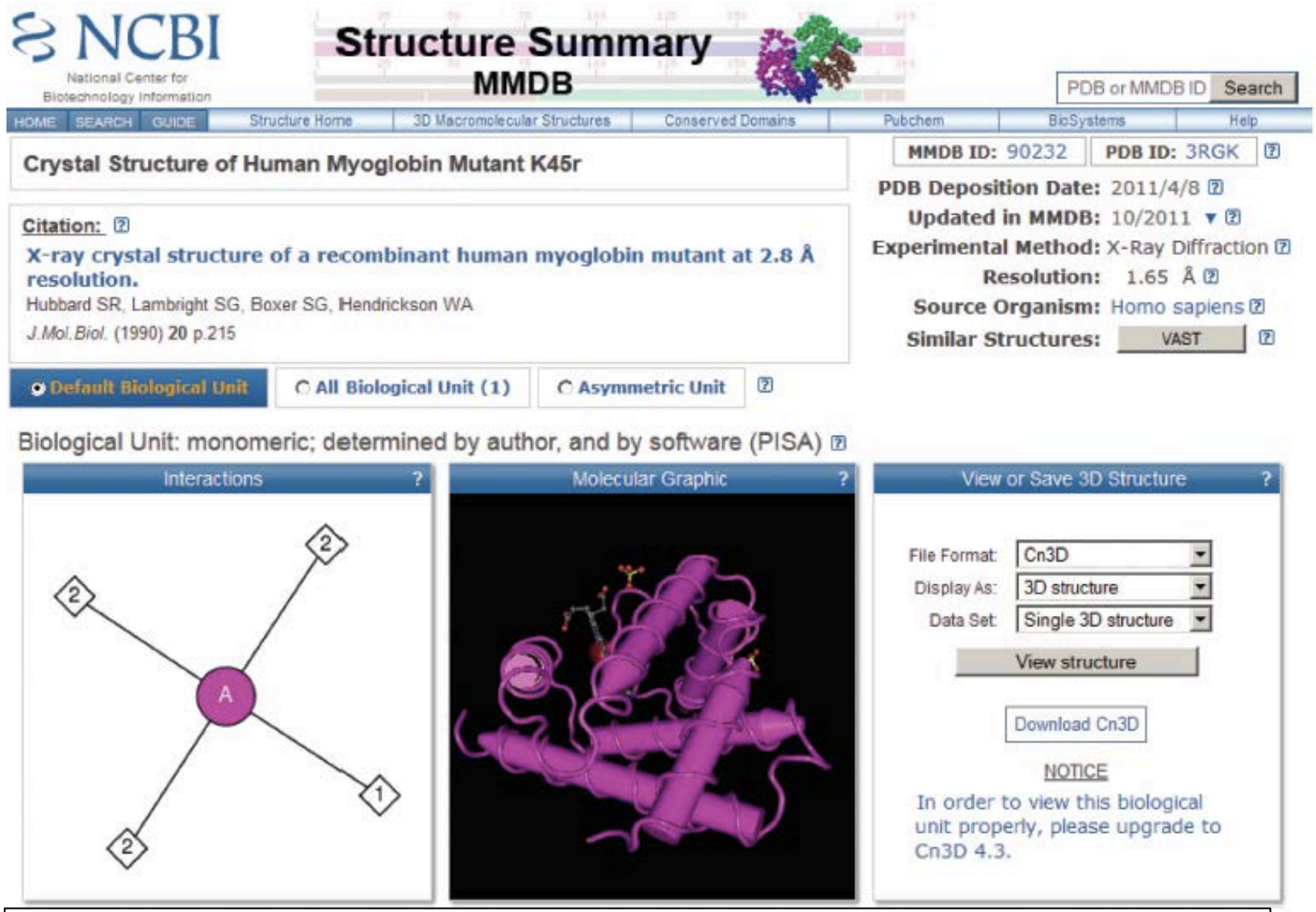
PMC (55)

Taxonomy (177)

Find related data

Database: Select

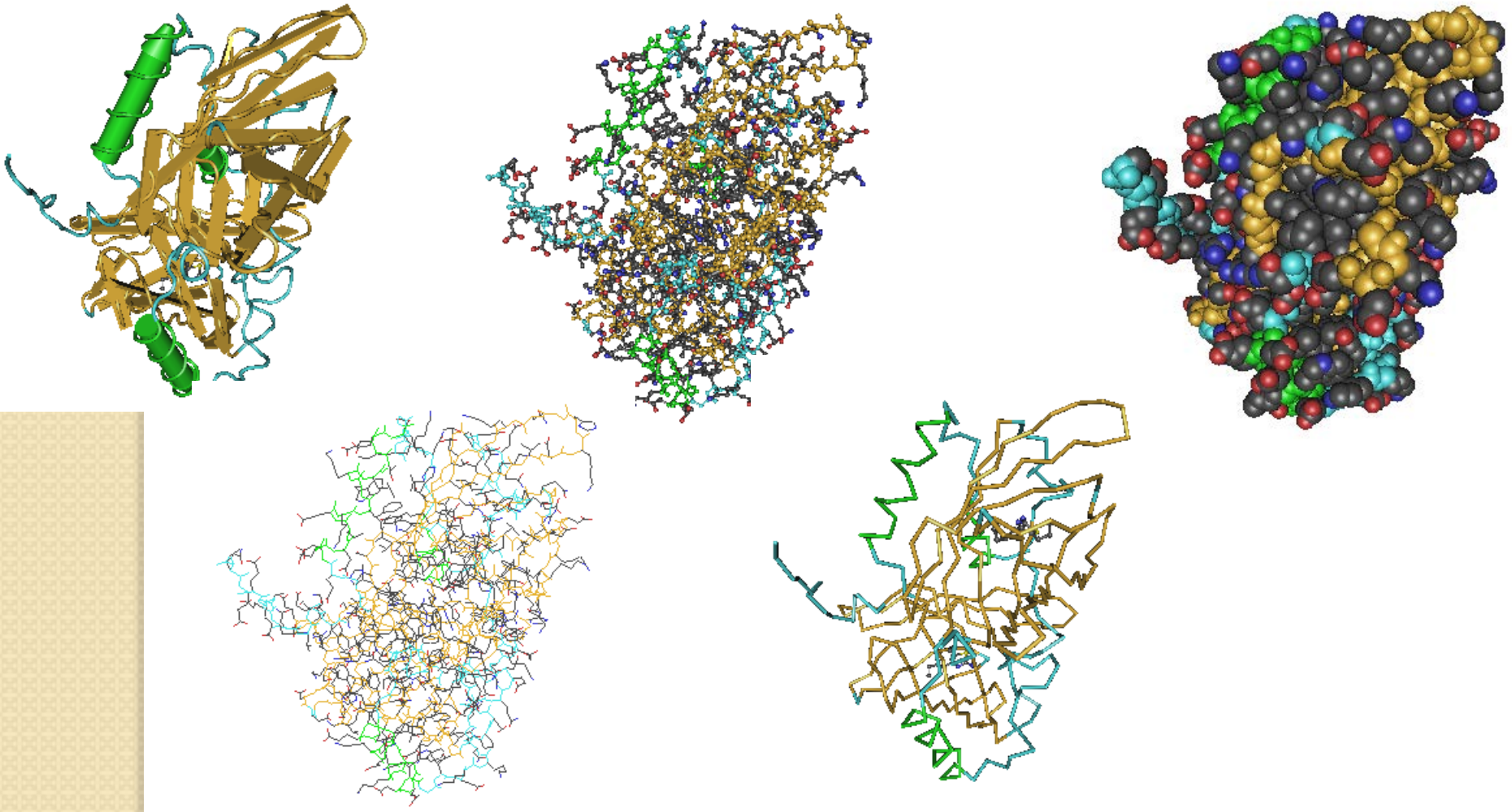
Molecular Modeling Database (MMDB) at NCBI



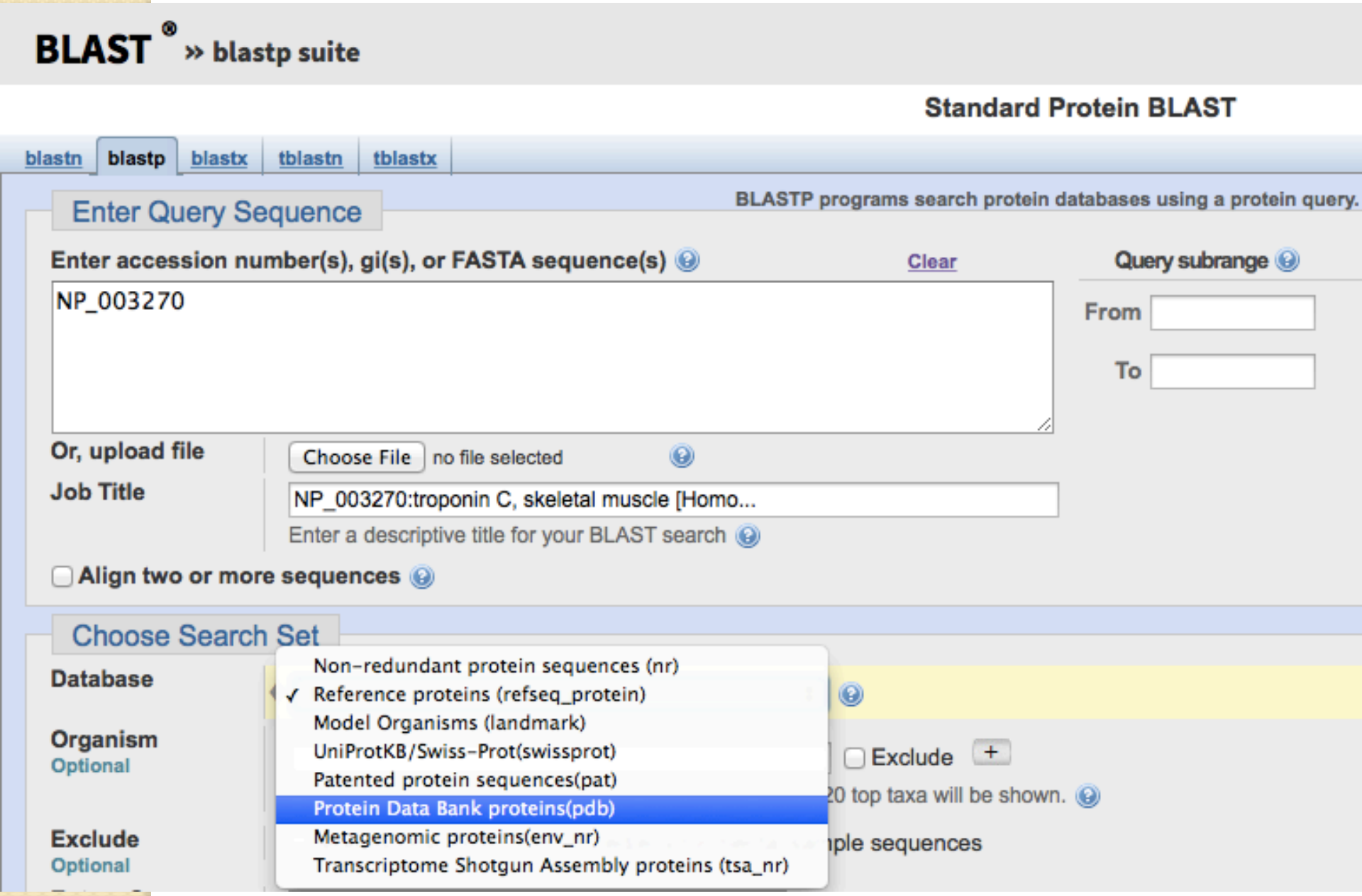
You can study PDB structures from NCBI. MMDB offers tools to analyze protein (and other) structures.

B&FG 3e
Fig. 13.12
Page 608

Cn3D: NCBI software for visualizing protein structures



Several display formats are shown



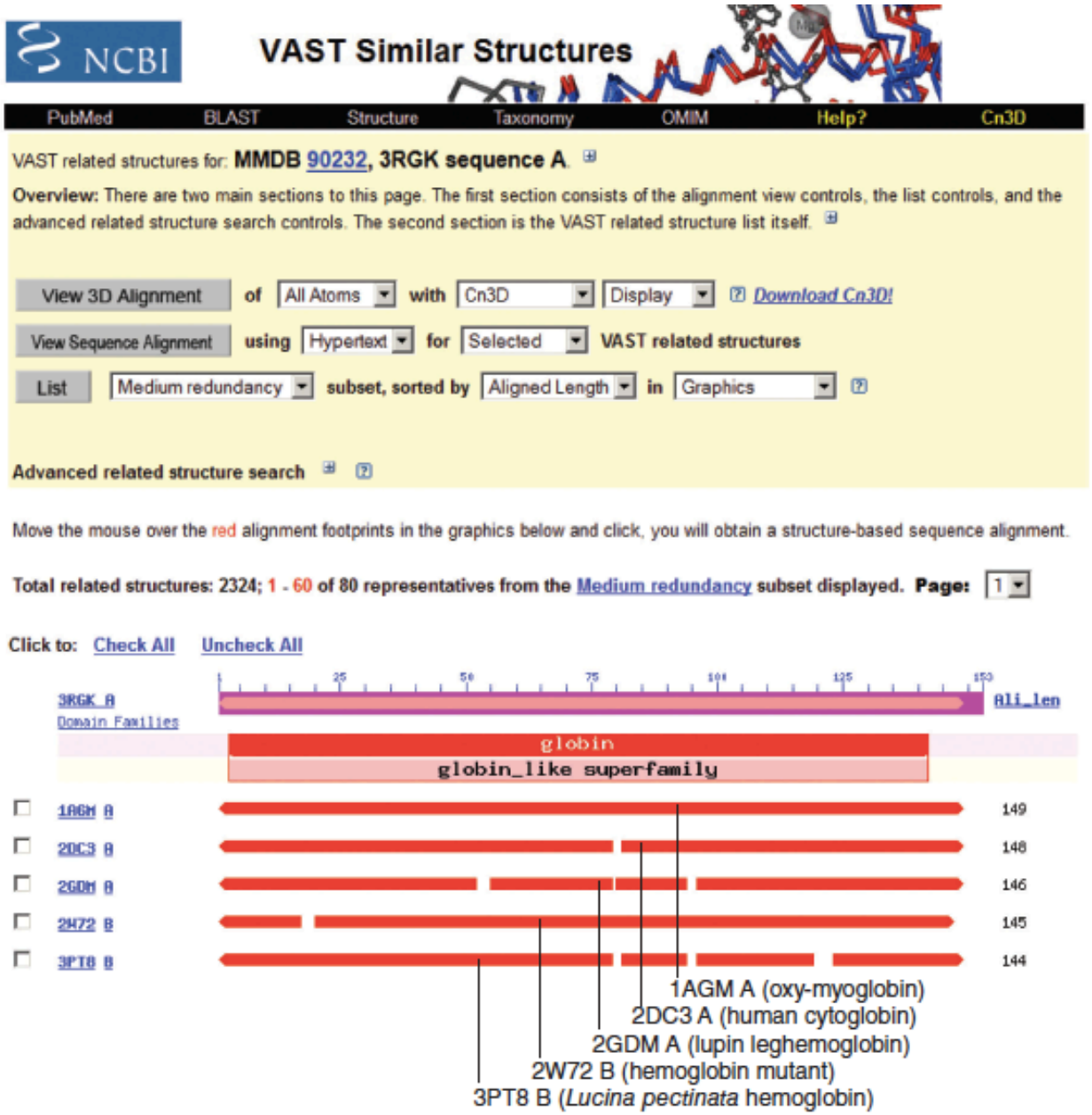
Do a DELTA BLAST (or BLASTP) search;
set the database to pdb (Protein Data Bank)

Access to structure data at NCBI: VAST

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

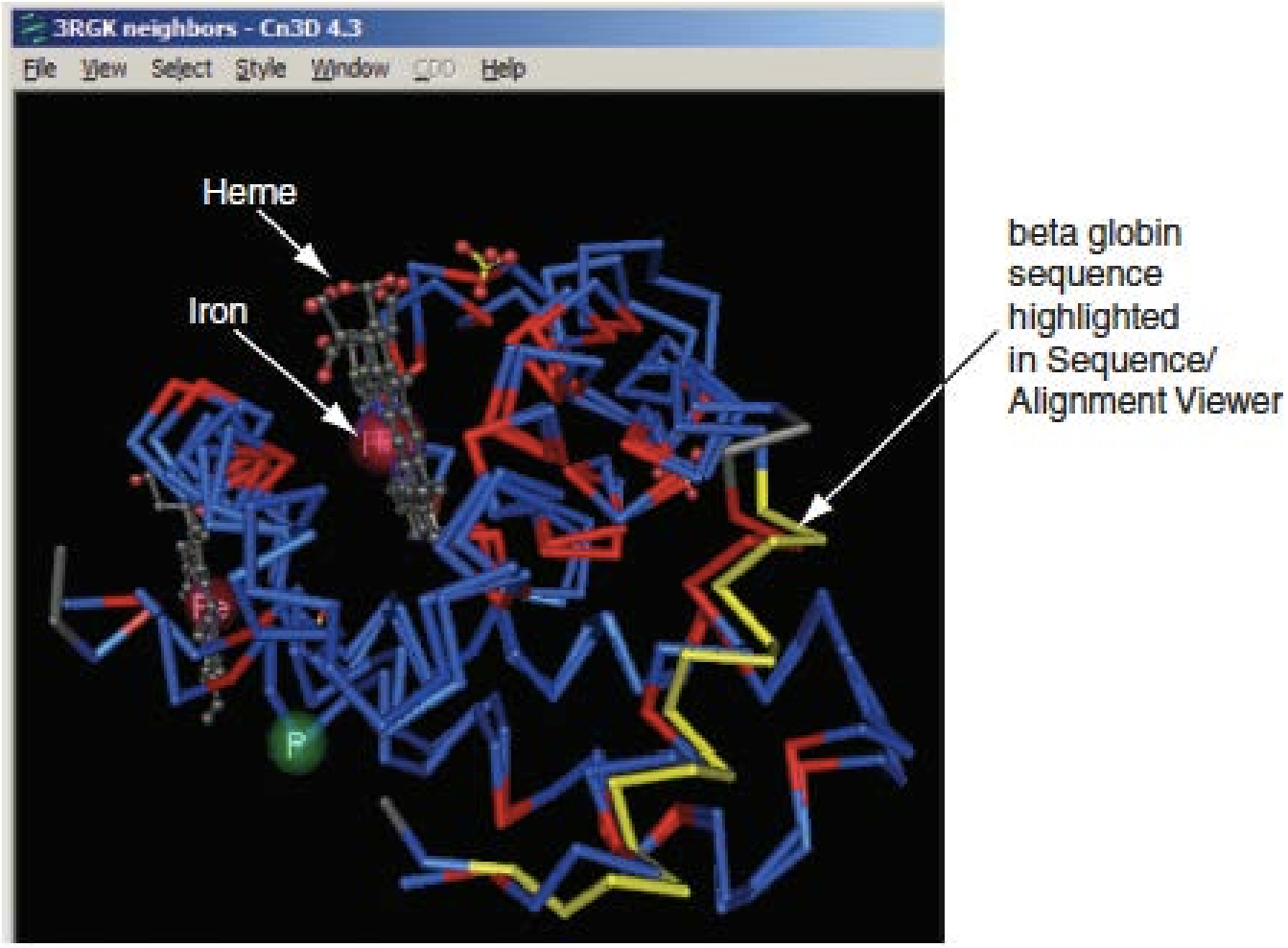
-- PDB identifiers
-- root-mean-square deviation (RMSD) values to describe structural similarities
-- NRES: the number of equivalent pairs of alpha carbon atoms superimposed
-- percent identity

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



B&FG 3e
Fig. 13.14
Page 610

VAST: NCBI tool to compare two structures



VAST: information listed for each structural neighbor

- checkbox: allows for selection of individual neighbors;
- PDB: four-character PDB-identifier of the structural neighbor;
- PDB chain name;
- MMDB domain identifier;
- VAST structure similarity score based on the number of related secondary structure elements and the quality of the superposition;
- RMSD: root-mean-square superposition residual in angstroms (a descriptor of overall structural similarity);
- NRES: number of equivalent pairs of Cα atoms superimposed between the two structures (the alignment length, i.e., how many residues have been used to calculate the three-dimensional superposition);
- %Id: percent identical residues in the aligned sequence region;
- description: string parsed from PDB records;
- metric (Loop Hausdorff Metric): describes how well two structures match in loop regions; and
- gapped score: combines RMSD, the length of the alignment, and the number of gapped regions.

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

B&FG 3e
Table 13.6
Page 612

SCOP-e database: hierarchy of terms

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

Two of the myoglobin structures are shown

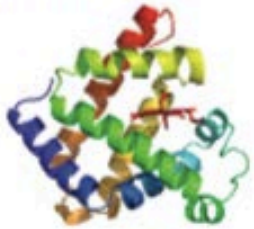
Lineage for Protein: Myoglobin

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)

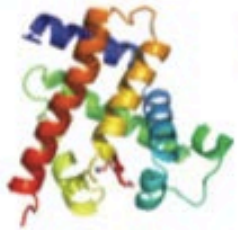
   Domain for 1emy:

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

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

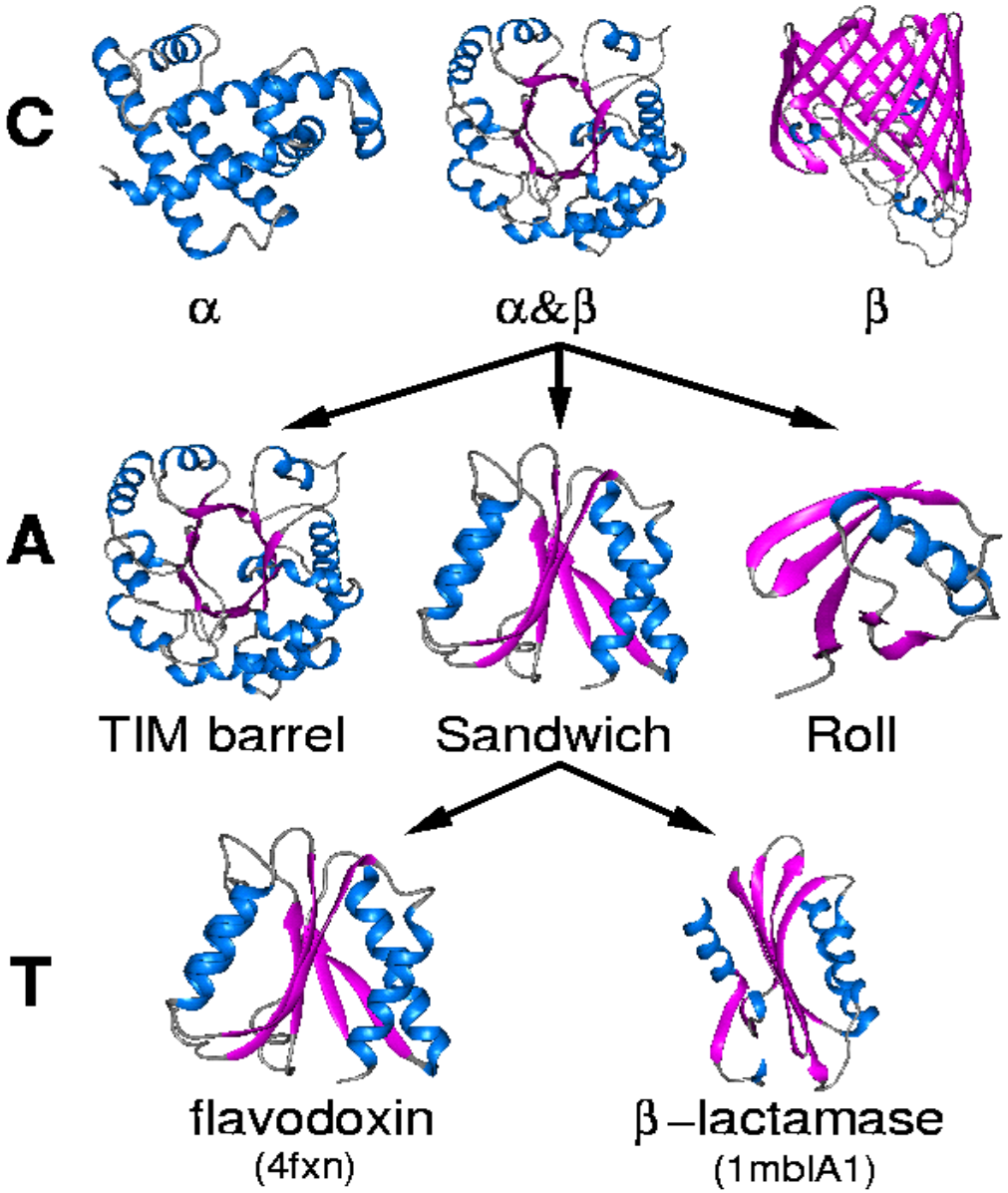
   Domain for 1mbs:

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

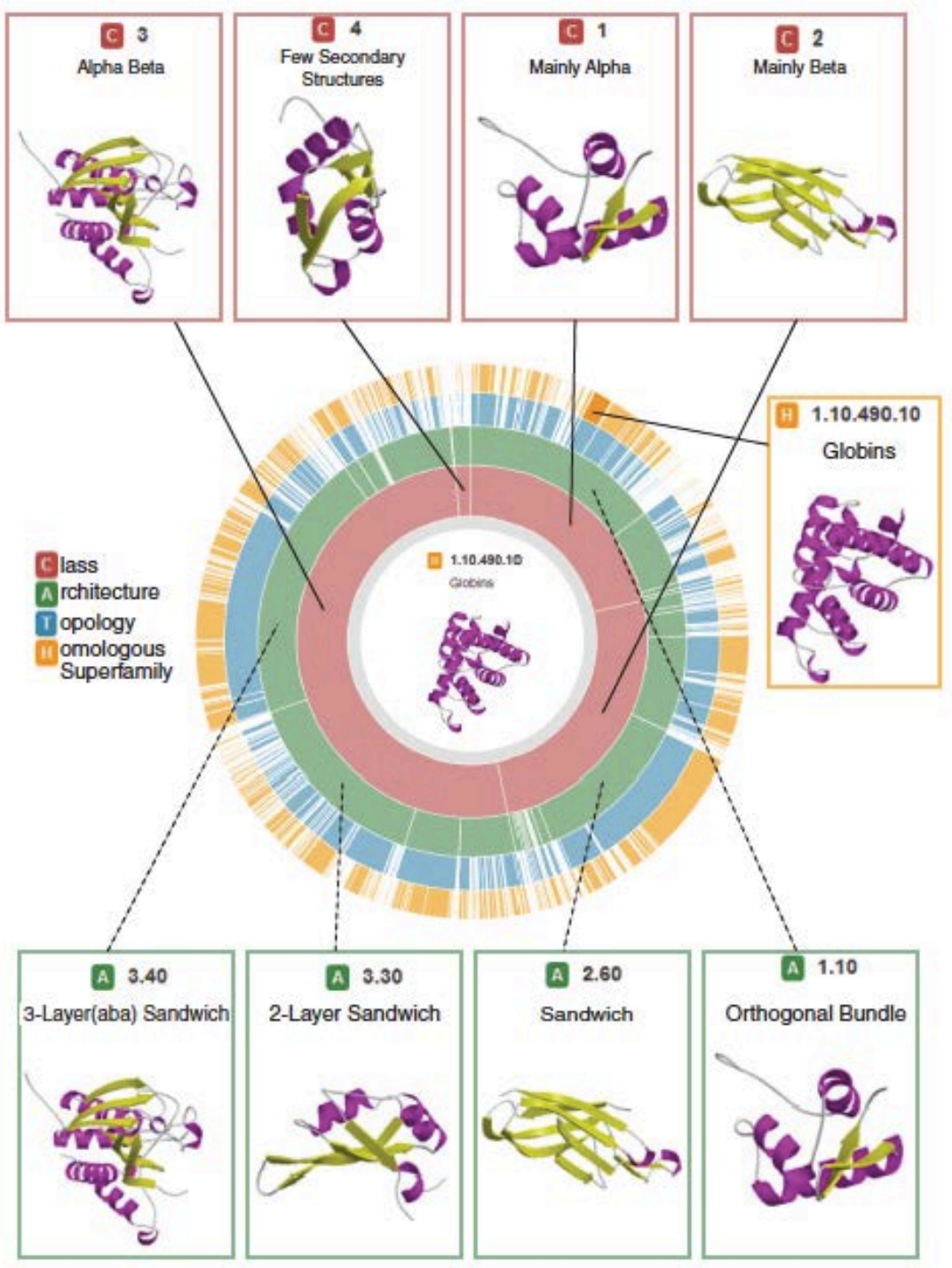
B&FG 3e
Fig. 13.6
Page 613

The CATH Hierarchy



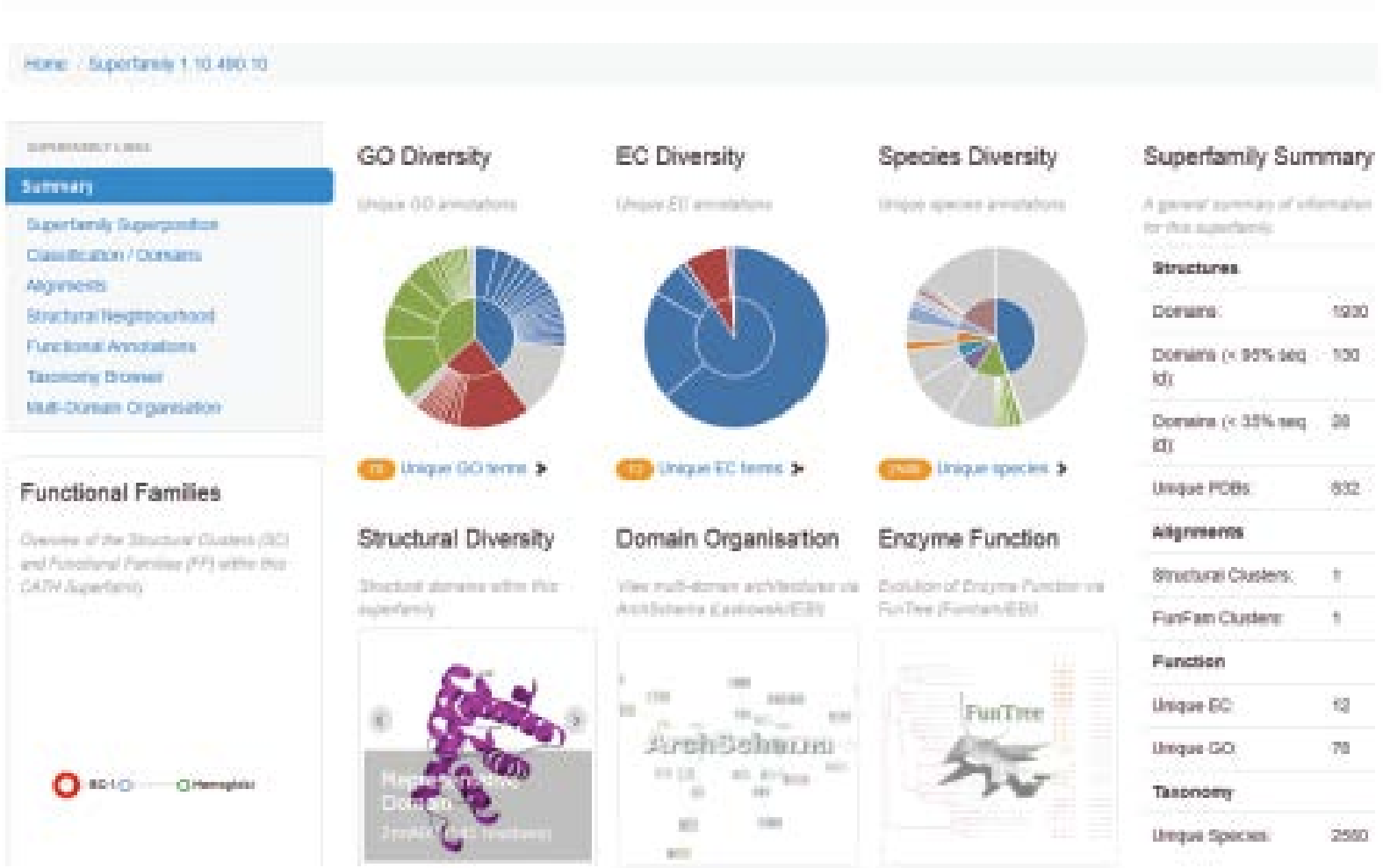
CATH organizes protein structures by a hierarchical scheme of class, architecture, topology (fold family), and homologous superfamily

Globins are highlighted.



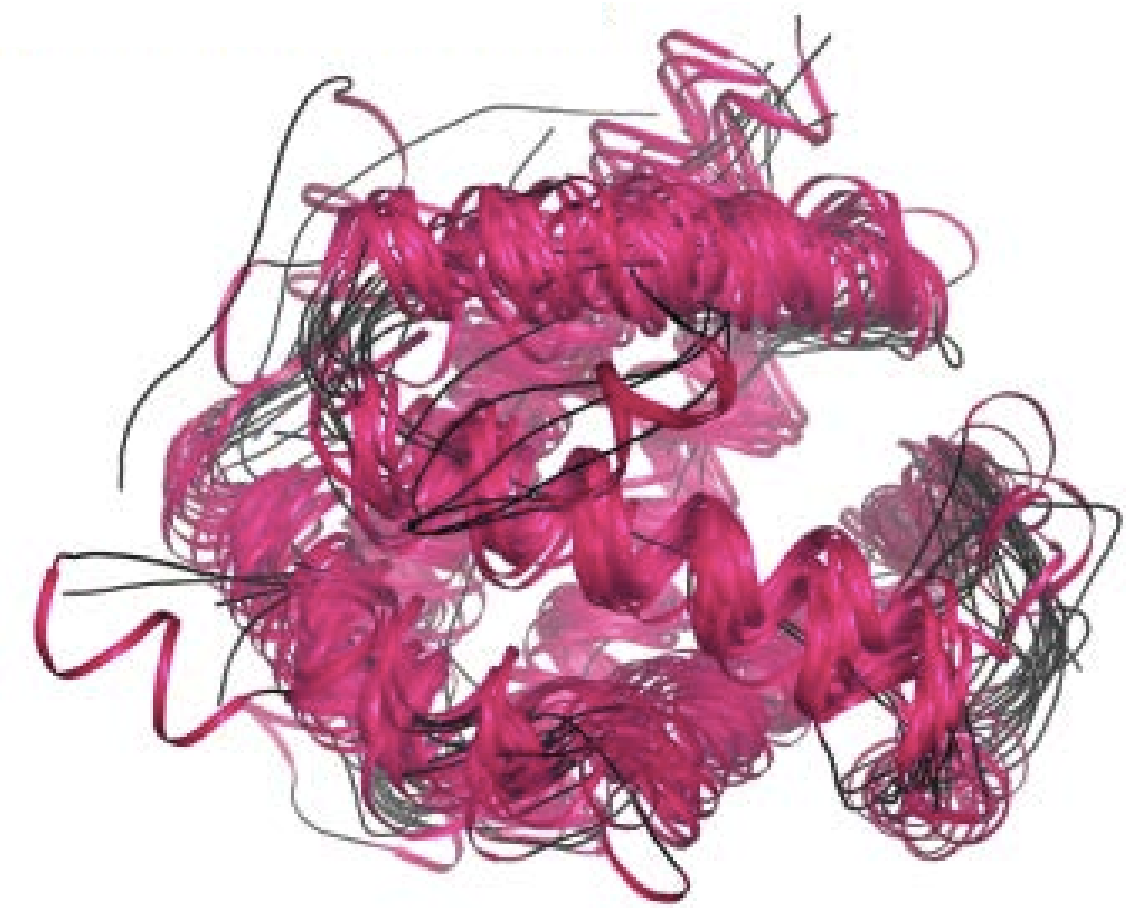
B&FG 3e
Fig. 13.17
Page 614

CATH globin superfamily



B&FG 3e
Fig. 13.18
Page 615

Superposition of globin superfamily members in CATH



B&FG 3e
Fig. 13.18
Page 615

Outline

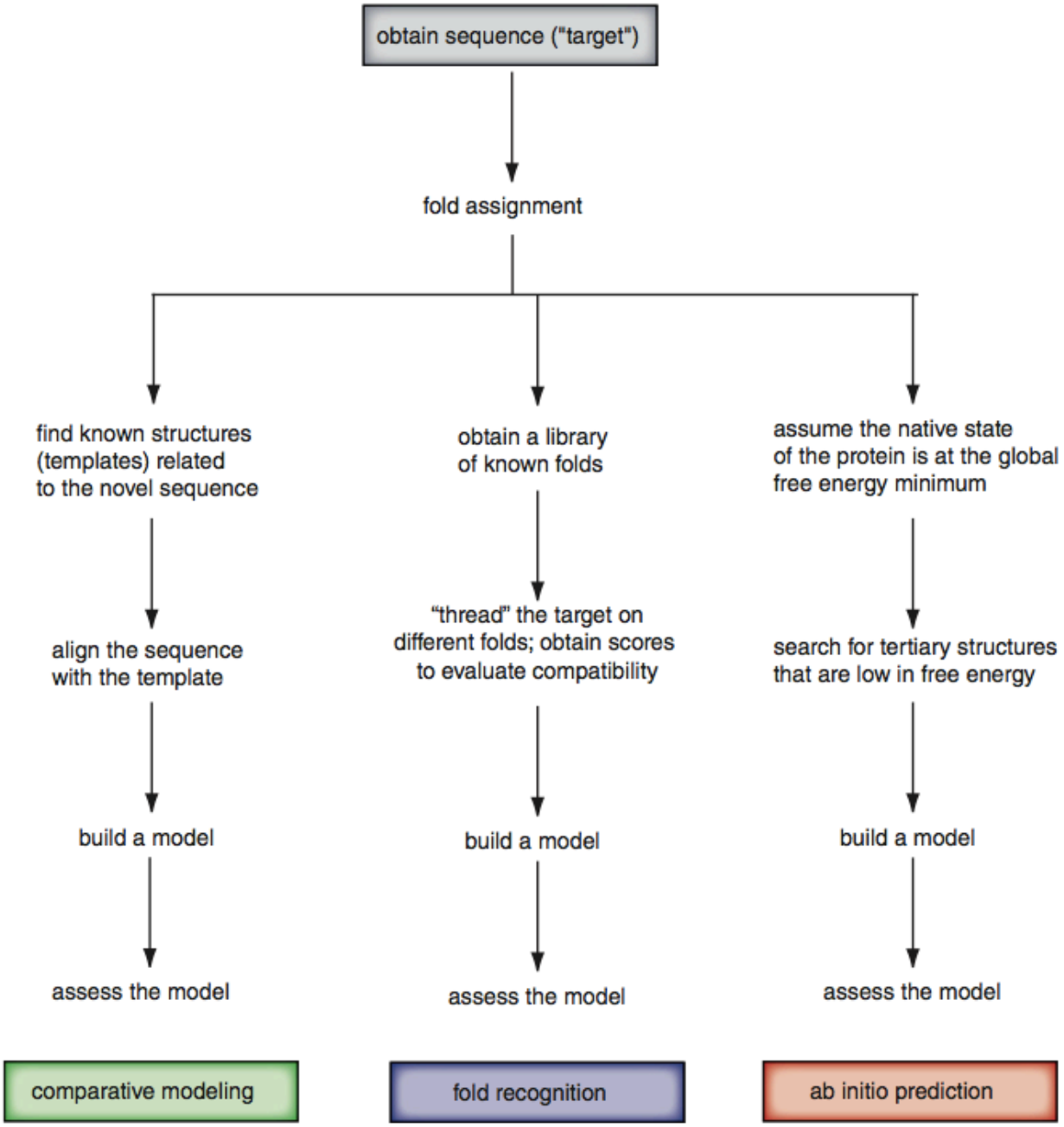
Overview of protein structure
Principles of protein structure
Protein Data Bank
Protein structure prediction
Intrinsically disordered proteins
Protein structure and disease

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



B&FG 3e
Fig. 13.20
Page 618

Structure prediction techniques as a function of sequence identity

sequence identity	model accuracy	resolution	technique	applications
100%				Studying catalytic mechanisms
	100%	1.0 Å	X-ray crystallography, NMR	Designing and improving ligands
				Prediction of protein partners
50%			comparative protein structural modeling	
	95%	1.5 Å		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

B&FG 3e
Fig. 13.21
Page 620

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/

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 a 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.

B&FG 3e
Tab. 13.10
Page 624

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.