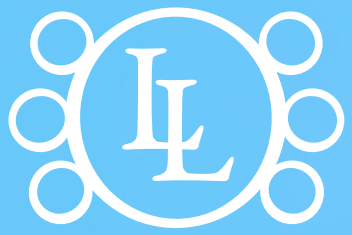


LOSCHMIDT
LABORATORIES

Structural databases

Outline

- Structural databases
- 3D data validation
- 3D protein modelling
- Models validation and databases

Outline

- Structural databases
  - Data formats (PDB, mmCIF, PDBML)
  - wwPDB
  - Other resources
- 3D data validation
- 3D protein modelling
- Models validation and databases

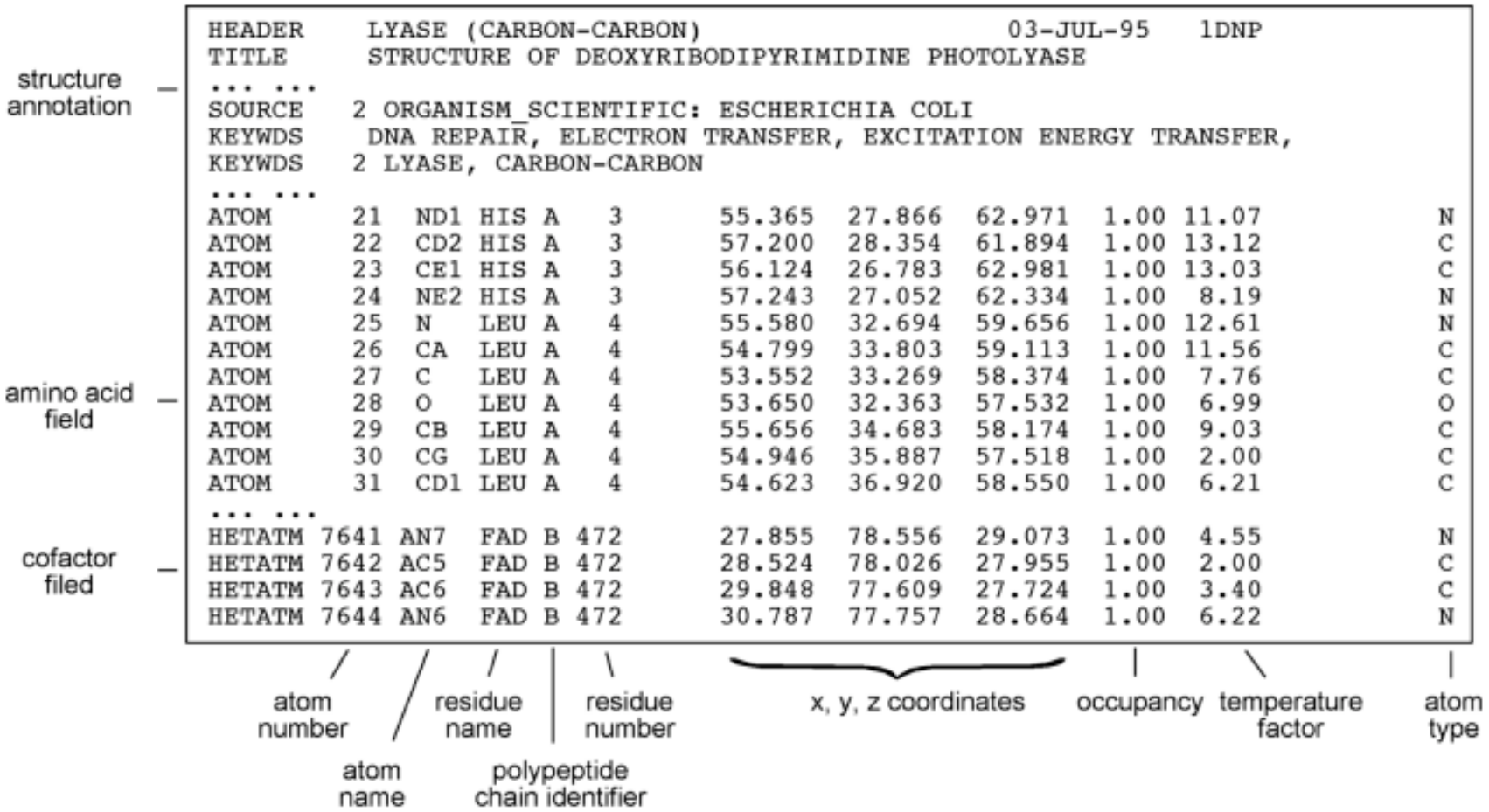
Data formats

- different formats are used to represent primary macromolecular 3D structure data
  - PDB
  - mmCIF
  - PDBML
  - …
- The spatial 3D coordinates for each atom are recorded

PDB format

- designed in the early 1970s - first entries of PDB database
- rigid structure of 80 characters per line, including spaces
- still the most widely supported format

PDB format


structure annotation
amino acid field
cofactor filed
HEADER LYASE (CARBON-CARBON) 03-JUL-95 1DNP
TITLE STRUCTURE OF DEOXYRIBODIPYRIMIDINE PHOTOLYASE
... ...
SOURCE 2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI
KEYWDS DNA REPAIR, ELECTRON TRANSFER, EXCITATION ENERGY TRANSFER,
KEYWDS 2 LYASE, CARBON-CARBON
... ...
ATOM 21 ND1 HIS A 3 55.365 27.866 62.971 1.00 11.07 N
ATOM 22 CD2 HIS A 3 57.200 28.354 61.894 1.00 13.12 C
ATOM 23 CE1 HIS A 3 56.124 26.783 62.981 1.00 13.03 C
ATOM 24 NE2 HIS A 3 57.243 27.052 62.334 1.00 8.19 N
ATOM 25 N LEU A 4 55.580 32.694 59.656 1.00 12.61 N
ATOM 26 CA LEU A 4 54.799 33.803 59.113 1.00 11.56 C
ATOM 27 C LEU A 4 53.552 33.269 58.374 1.00 7.76 C
ATOM 28 O LEU A 4 53.650 32.363 57.532 1.00 6.99 O
ATOM 29 CB LEU A 4 55.656 34.683 58.174 1.00 9.03 C
ATOM 30 CG LEU A 4 54.946 35.887 57.518 1.00 2.00 C
ATOM 31 CD1 LEU A 4 54.623 36.920 58.550 1.00 6.21 C
... ...
HETATM 7641 AN7 FAD B 472 27.855 78.556 29.073 1.00 4.55 N
HETATM 7642 AC5 FAD B 472 28.524 78.026 27.955 1.00 2.00 C
HETATM 7643 AC6 FAD B 472 29.848 77.609 27.724 1.00 3.40 C
HETATM 7644 AN6 FAD B 472 30.787 77.757 28.664 1.00 6.22 N
atom number
atom name
residue name
residue number
polypeptide chain identifier
x, y, z coordinates
occupancy
temperature factor
atom type

PDB format

- atomic coordinates
- chemical and biological features
- experimental details of the structure determination
- structural features
  - secondary structure assignments
  - hydrogen bonding
  - biological assemblies REMARK 350
  - active sites
  - ...

PDB format

- advantages
  - widely used → supported by majority of tools
  - easy to read and easy to use

→ suitable for accessing individual entries

PDB format

- disadvantages
  - inconsistency between individual PDB entries as well as PDB records within one entry (e.g., different residue numbering in SEQRES and ATOM sections) → not suitable for computer extraction of information

```
SEQRES   1    396  MET ASP GLU ASN ILE THR ALA ALA PRO ALA ASP PRO ILE
SEQRES   2    396  LEU GLY LEU ALA ASP LEU PHE ARG ALA ASP GLU ARG PRO
. . .
. . .
ATOM      1  N   MET     5      41.402  11.897  15.262  1.00 48.61
ATOM      2  CA  MET     5      40.919  13.262  15.600  1.00 47.70
ATOM      9  N   PHE     6      39.627  14.840  14.228  1.00 48.66
ATOM     10  CA  PHE     6      39.199  15.440  12.964  1.00 45.33
. . .
```

PDB format

- disadvantages
  - inconsistency between individual PDB entries as well as PDB records within one entry → not suitable for computer extraction of information
  - absolute limits on the size of certain items of data, e.g.: max. number of atom records limited to 99,999; max. number of chains limited to 26 → large systems such as the ribosomal subunit must be divided into multiple PDB files

→ not suitable for analysis and comparison of experimental and structure data across the entire database

mmCIF format

- **m**acro**m**olecular **C**rystallographic **I**nformation **F**ile (mmCIF)
- developed to handle increasingly complicated structure data
- each field of information is explicitly assigned by a tag and linked to other fields through a special syntax

PDB HEADER PLANT SEED PROTEIN 11-OCT-91 1CBN

mmCIF

```
_struct.entry_id '1CBN'
_struct.title 'PLANT SEED PROTEIN'
_struct_keywords.entry_id '1CBN'
_struct_keywords.text 'plant seed protein'
_database_2.database_id 'PDB'
_database_2.database_code '1CBN'
_database_PDB_rev.rev_num 1
_database_PDB_rev.date_original '1991-10-11'
```

mmCIF format

- advantages
  - easily parsable by computer software
  - consistency of data across the database
- disadvantages
  - difficult to read
  - rarely supported by visualization and computational tools

→ suitable for analysis and comparison of experimental and structure data across the entire database

→ not suitable for accessing individual entries

PDBML format

- **P**rotein **D**ata **B**ank **M**arkup **L**anguage (PDBML)
- XML version of PDB format

```
<?xml version="1.0" encoding="UTF-8" ?>
<PDBx:datablock  datablockName="EXAMPLE"
   xmlns:PDBx="http://deposit.pdb.org/pdbML/pdbx-v1.000.xsd"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://deposit.pdb.org/pdbML/pdbx-v1.000.xsd
            pdbx-v1.000.xsd">
   <PDBx:entity_polyCategory>
      <PDBx:entity_poly entity_id="1">
         <PDBx:type>polypeptide(L)</PDBx:type>
         <PDBx:nstd_linkage>no</PDBx:nstd_linkage>
         <PDBx:nstd_monomer>no</PDBx:nstd_monomer>
         <PDBx:pdbx_seq_one_letter_code>
          DIVLTQSPASLSASVGETVTITCRASGNIHNYLAWYQQKQGKSPQLLVYYTTTLADG
          VPSRFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWSTPRTFGGGTKLEIK
         </PDBx:pdbx_seq_one_letter_code>
         <PDBx:pdbx_seq_one_letter_code_can>
          DIVLTQSPASLSASVGETVTITCRASGNIHNYLAWYQQKQGKSPQLLVYYTTTLADG
          VPSRFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWSTPRTFGGGTKLEIK
         </PDBx:pdbx_seq_one_letter_code_can>
      </PDBx:entity_poly>
   </PDBx:entity_polyCategory>
</PDBx:datablock>
```

Structural databases

- Primary
  - **wwPDB: 3D structure of biopolymers**
    - BMRB: Nuclear Magnetic Resonance specific
    - EMDB: Electron-Microscopy specific
  - NDB: 3D structure of nucleic acids: http://ndbserver.rutgers.edu/
  - CSD: 3D structure of small molecules (commercial)

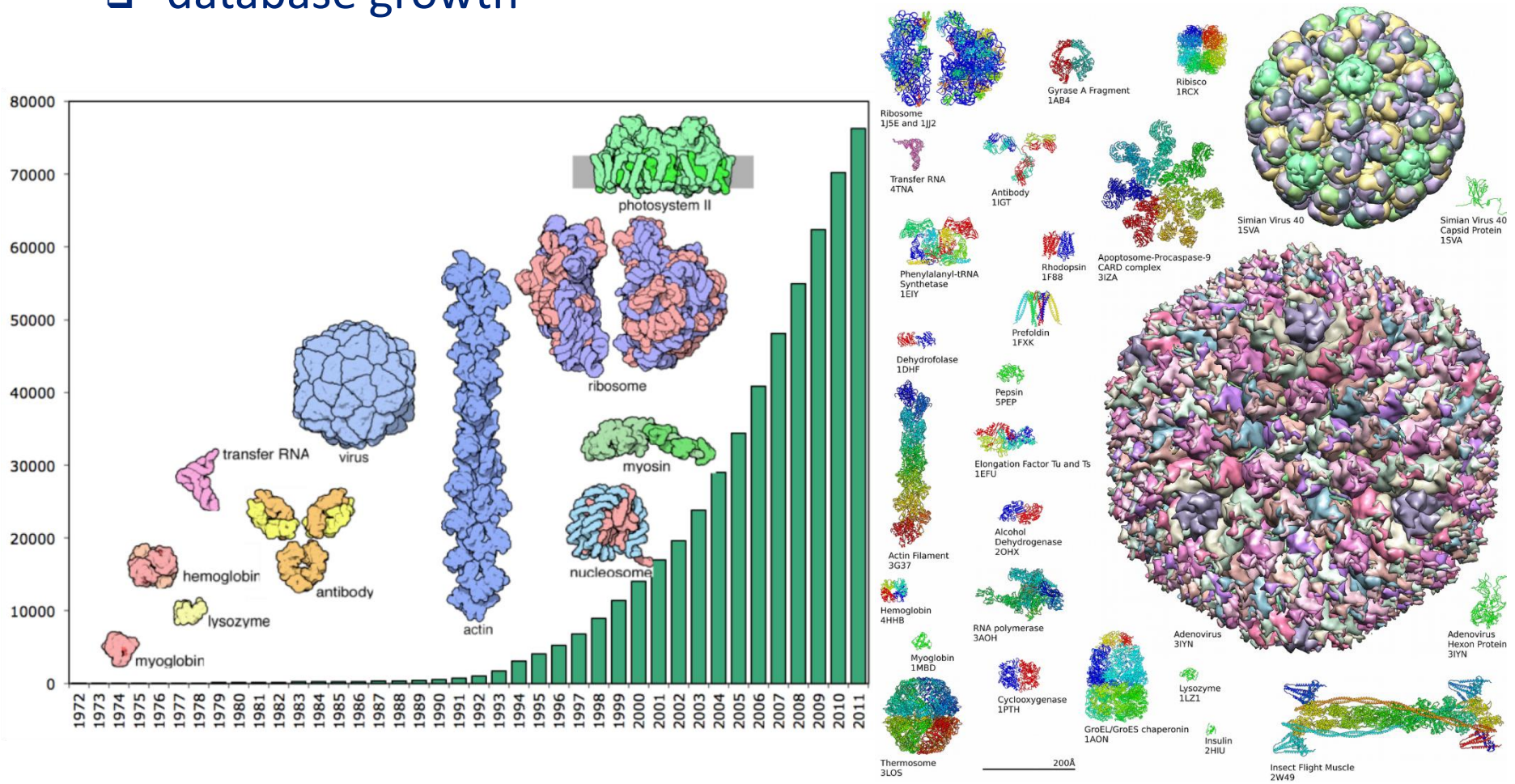
    http://www.ccdc.cam.ac.uk/products/csd/
- Other sources
  - PDBsum, SCOP, Protopedia, Structural Biology KnowledgeBase

wwPDB

- joint initiative of four organizations
  - Research Collaboratory for Structural Bioinformatics (RCSB PDB)
  - Protein Data Bank in Europe (PDBe)
  - Protein Data Bank Japan (PDBj)
  - Biological Magnetic Resonance Data Bank (BMRB)

wwPDB

- database growth



wwPDB

- **w**orld**w**ide **P**rotein **D**ata **B**ank (wwPDB)
  - http://www.wwpdb.org/
  - central repository of experimental macromolecular structures
  - more than 225,000 structures (October 2024), updated every week
  - mostly protein structures (87 %), structures of protein/nucleic acids or oligosaccharides complexes (11 %) and nucleic acid structures (2 %)
  - majority of structures from X-ray crystallography (84 %), NMR (6 %), or EM (10%)
  - deposition of the structure into wwPDB is a requirement for its publication

wwPDB – data deposition

- All data can be deposited at RCSBPDB, PDBe or PDBj site
  - Same requirements content and format of the final files:
    - structures of biopolymers
    - structures determined by experimental techniques
    - structures containing required information
  - Same validation methods

  → uniformity of the final archive
- PDB-ID
  - assigned to each deposition
  - unique identifier of each structure
  - four-character code

wwPDB – data validation

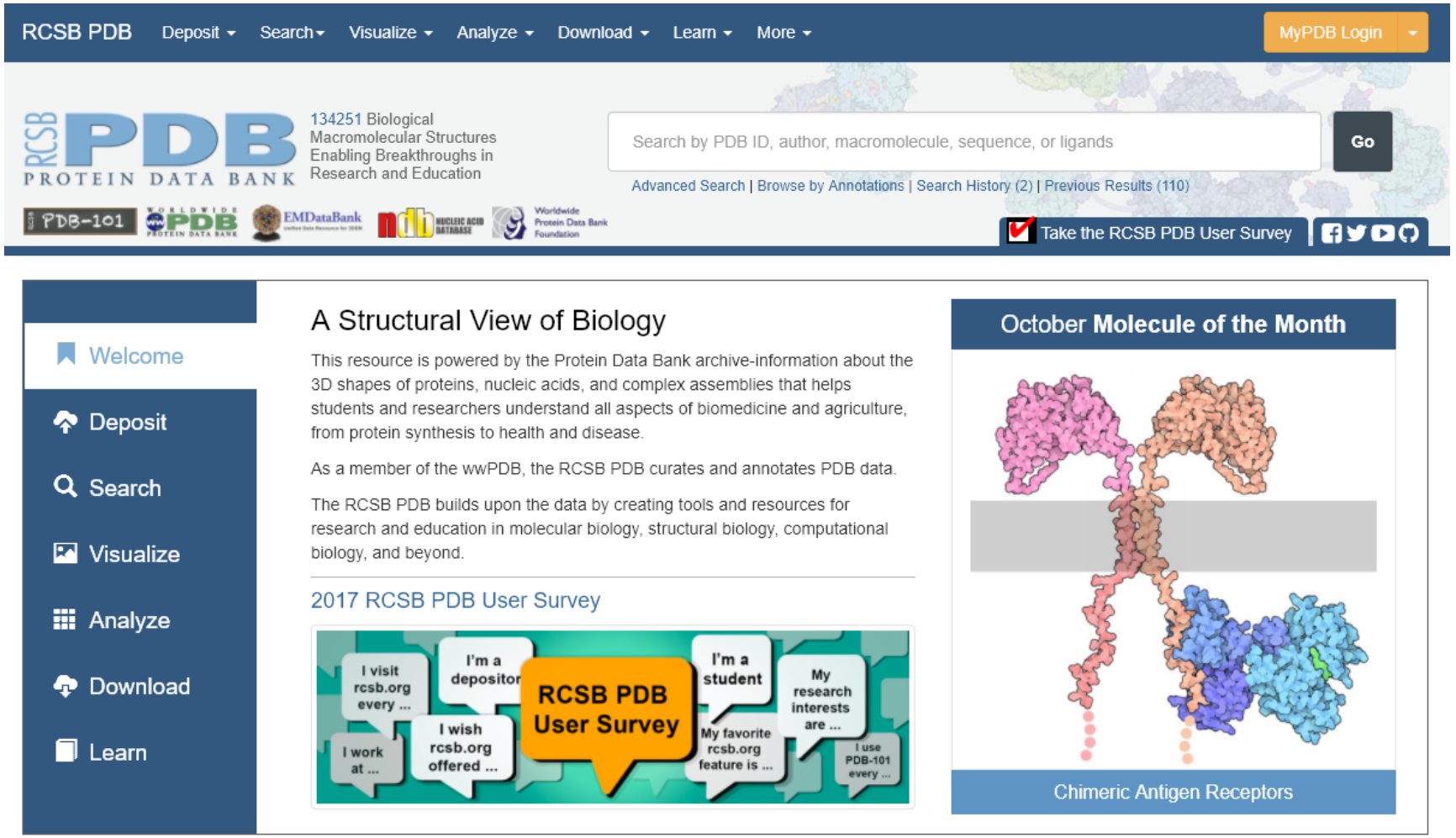
- assessment of the quality of deposited atomic models (structure validation) and how well these models fit experimental data (experimental validation)
- validation using accepted community standards
  - covalent bond distances and angles
  - stereochemical validation
  - atom and ligand nomenclature
  - geometry
  - NMR data specific checks
  - …

wwPDB – data access

- the access to the PDB archive is free and publicly available from the RCSB PDB site, PDBe site or PDBj site
- FTP
  - RCSB PDB, PDBe and PDBj sites distribute the same PDB archive
  - updated weekly
- web sites
  - each wwPDB site provides its own services and resources → different views and analyses of the structural data
  - sequence-based and text-based queries

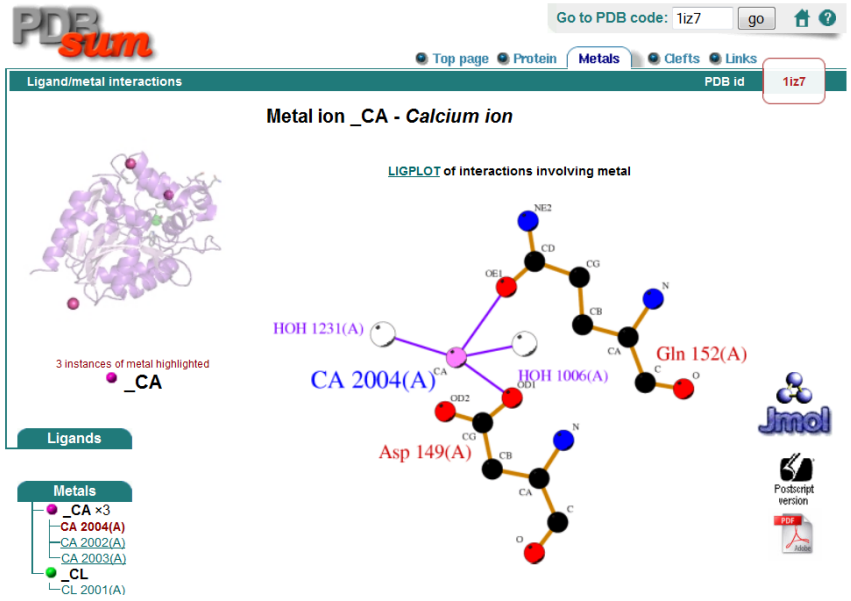
RCSB PDB

- http://pdb.rcsb.org



Other structure-based resources

- **PDBsum**
  - http://www.ebi.ac.uk/pdbsum/
  - provides summaries and pre-computed analyses for structures deposited in the wwPDB



Other structure-based resources

- **S**tructural **C**lassification of **P**roteins (**SCOP**)
  - http://scop.mrc-lmb.cam.ac.uk/scop/
  - provides classifications of proteins with known 3D structure according to their evolutionary and structural relationships

Protein: Haloalkane dehalogenase from Sphingomonas paucimobilis, UT26, LinB [TaxId: 13689]

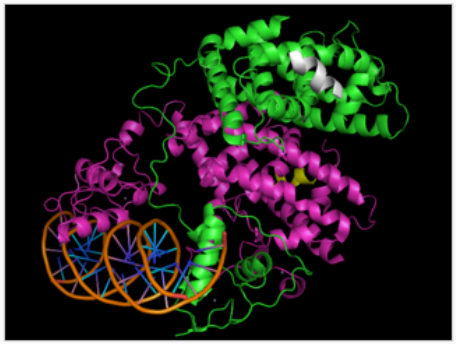
Lineage:

1. Root: scop
2. Class: Alpha and beta proteins (a/b) [51349]
   Mainly parallel beta sheets (beta-alpha-beta units)
3. Fold: alpha/beta-Hydrolases [53473]
   core: 3 layers, a/b/a; mixed beta-sheet of 8 strands, order 12435678, strand 2 is antiparallel to the rest
4. Superfamily: alpha/beta-Hydrolases [53474]
   many members have left-handed crossover connection between strand 8 and additional strand 9
5. Family: Haloalkane dehalogenase [53513]
6. Protein: Haloalkane dehalogenase [53514]
7. Species: Sphingomonas paucimobilis, UT26, LinB [TaxId: 13689] [53517]

Other structure-based resources

- **Proteopedia**
  - http://www.proteopedia.org/wiki/index.php/
  - free, collaborative 3D-encyclopedia of proteins and other molecules

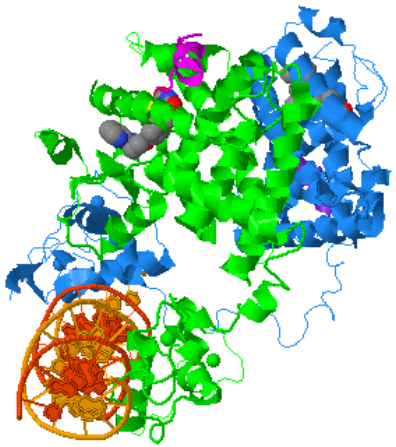
Peroxisome Proliferator-Activated Receptors



Human PPARγ bound to RXRα and PPRE DNA strand, 3dzy

The **Peroxisome Proliferator-Activated Receptors** (PPAR) α, γ, and δ are members of the nuclear receptor family. Since their discovery in the early 90s, it has become clear that the PPARs are essential modulators of external stimuli, acting as transcription factors to regulate mammalian metabolism, cellular differentiation, and tumorigenesis. The PPARs are the targets of numerous pharmaceutical drugs aimed at treating hypolipidemia and diabetes among other diseases.[1] For details on PPARγ see PPAR-gamma.

Contents [hide]



Jmol

toggle spin | toggle quality | popup

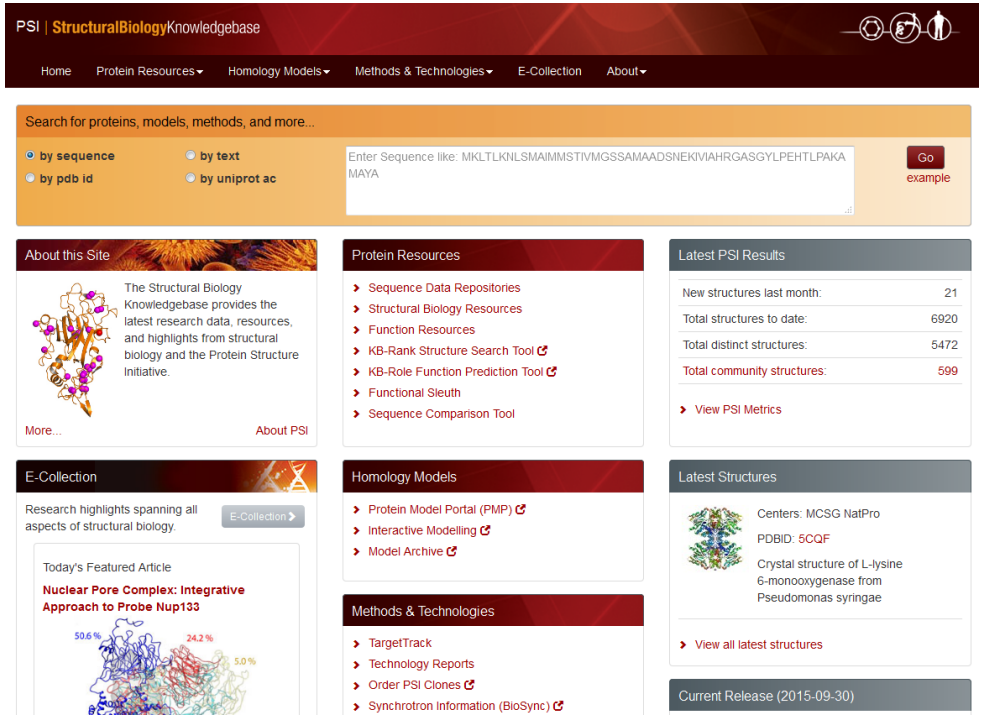
Crystal Structure of Human PPARγ complex with retinoic acid receptor -α, DNA, retinoic acid, LXXLL peptide, 3dzy

Ligands:	BRL, REA, ZN
Gene:	RXRA, NR2B1 (Homo sapiens), PPARG,

Biological Role

Other structure-based resources

- **Structural Biology Knowledgebase**
  - http://sbkb.org/
  - provides up-to-date information about advances in structural biology and structural genomics



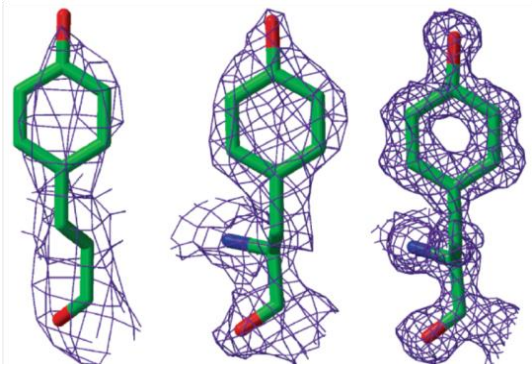
Structural quality assurance

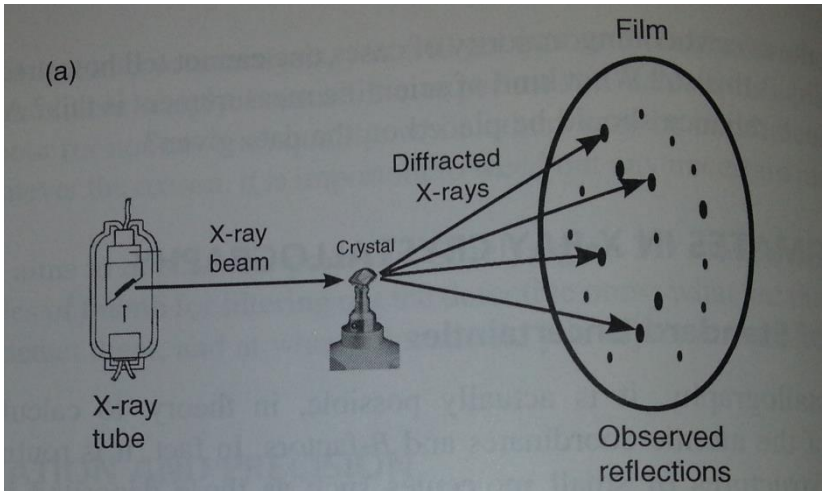
Outline

- Revision of concepts
- Important truths about structures
- Errors in deposited structures
  - systematic errors
  - random errors
- Selecting reliable structure
  - rules of thumbs
  - quality checks
  - programs and databases

Concepts

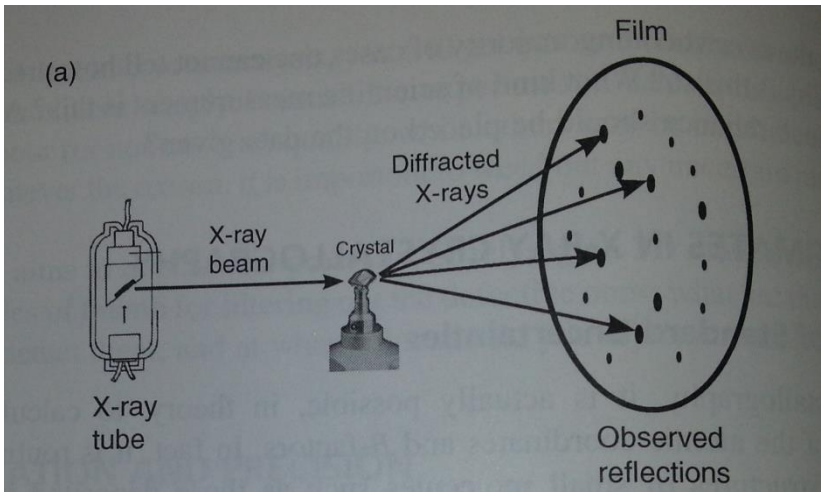
- Resolution
  - measure of the level of detail present in the diffraction pattern

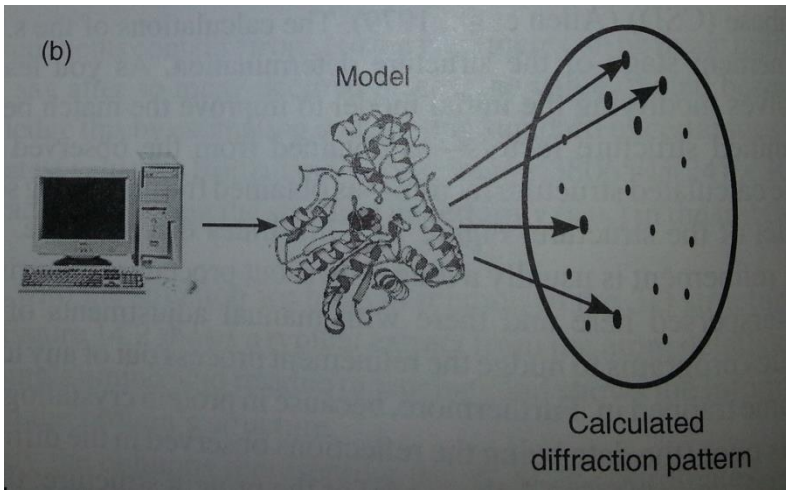




Concepts

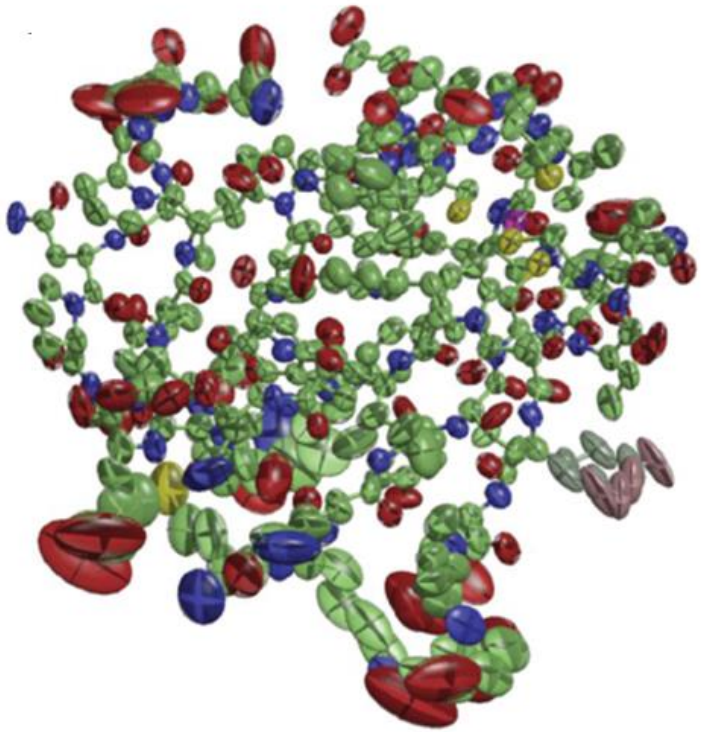
- R-factor (R-value)
  - measure of a model quality - i.e. how well it can reproduce experimental data





Concepts

- Thermal factors (B-factors)
  - measure of how much an atom oscillates or vibrates around the position specified in the model



Important truths about structures

- all structures are just models devised to satisfy experimental data → random and systematic errors
- individual structures differ in the quality
- most structures are reasonably accurate, containing “only” random errors, but some structures are seriously incorrect
- structures should be carefully selected and critically assessed before being used for a specific purpose → quality checks of structures

Errors in deposited structures

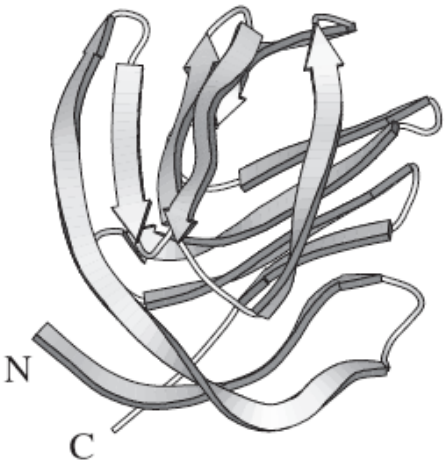
- systematic errors
- random errors

Systematic errors

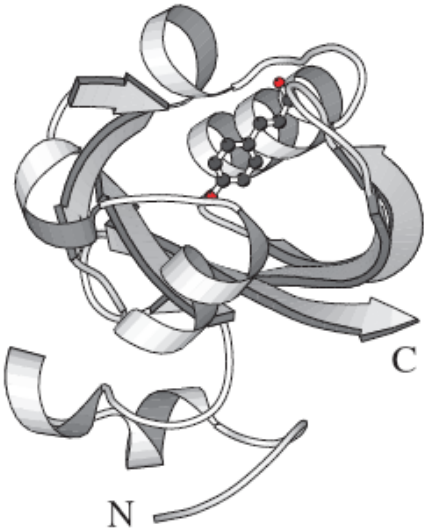
- relate to the accuracy of the model—how well it corresponds to the “true” structure of the molecule in question
- often include errors of interpretation
  - low quality of electron density map → difficult to find the correct tracing of the molecule(s) through it → misstracing and “frame-shift” errors
  - spectral interpretations (assignment of individual NMR signals to individual atoms)
- may lead to completely wrong final structure

Examples of systematic errors

- completely wrong structures
  - trace of the protein chain following the wrong path through the electron density → completely incorrect fold



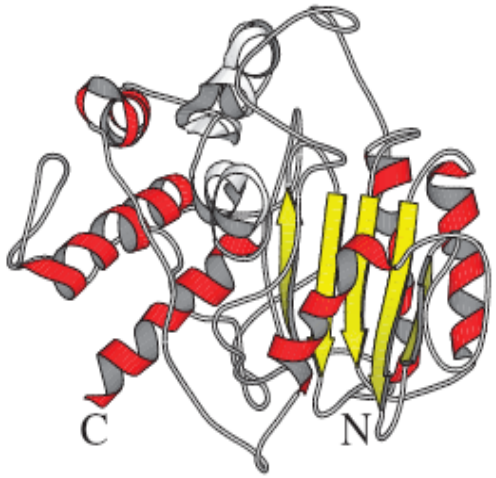
Incorrect model (1PHY)



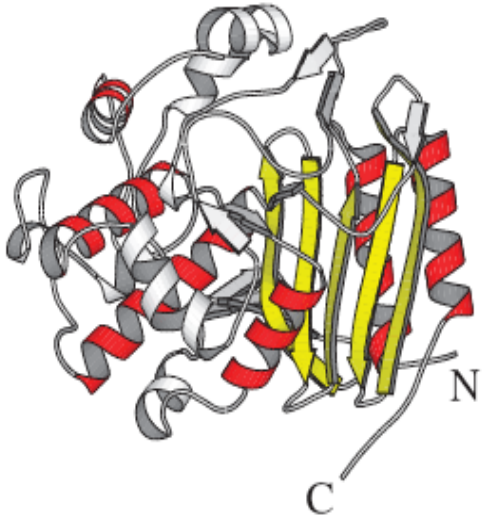
Corrected model (2PHY)

Examples of systematic errors

- wrong connectivity between secondary structure elements
  - incorrect order of secondary structure elements → many protein's residues in the wrong place in the 3D structure



Incorrect model (1PTE)



Corrected model (3PTE)

Examples of systematic errors

- frame-shift errors
  - occur where a residue is fitted into the electron density that belongs to the next residue and persists until compensating error is made (two residues are fitted into the density of a single residue)
  - occur almost exclusively at very low resolution (> 3.0 Å), often in loop regions
- fitting of incorrect main chain or side chain conformations into the density
  - usually the least serious, however still can have effects on biological interpretations

Random errors

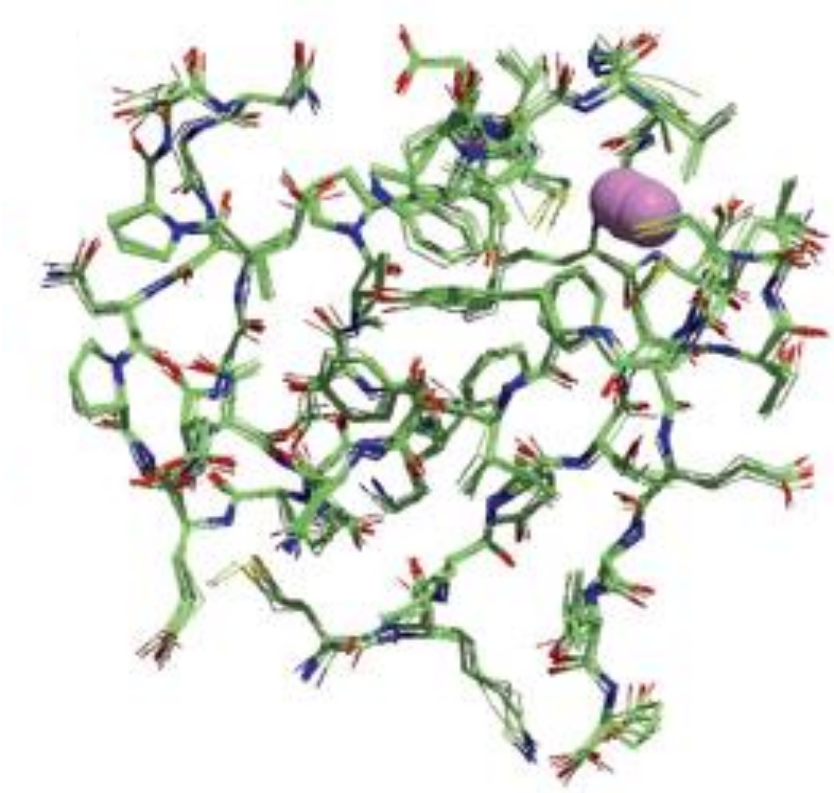
- depend on how precisely a given measurement can be made
- all measurements contain errors at some degree of precision

→ uncertainties in atomic positions

- less serious than systematic errors
- if a structure is essentially correct, the sizes of the random errors determine how precise the structure is

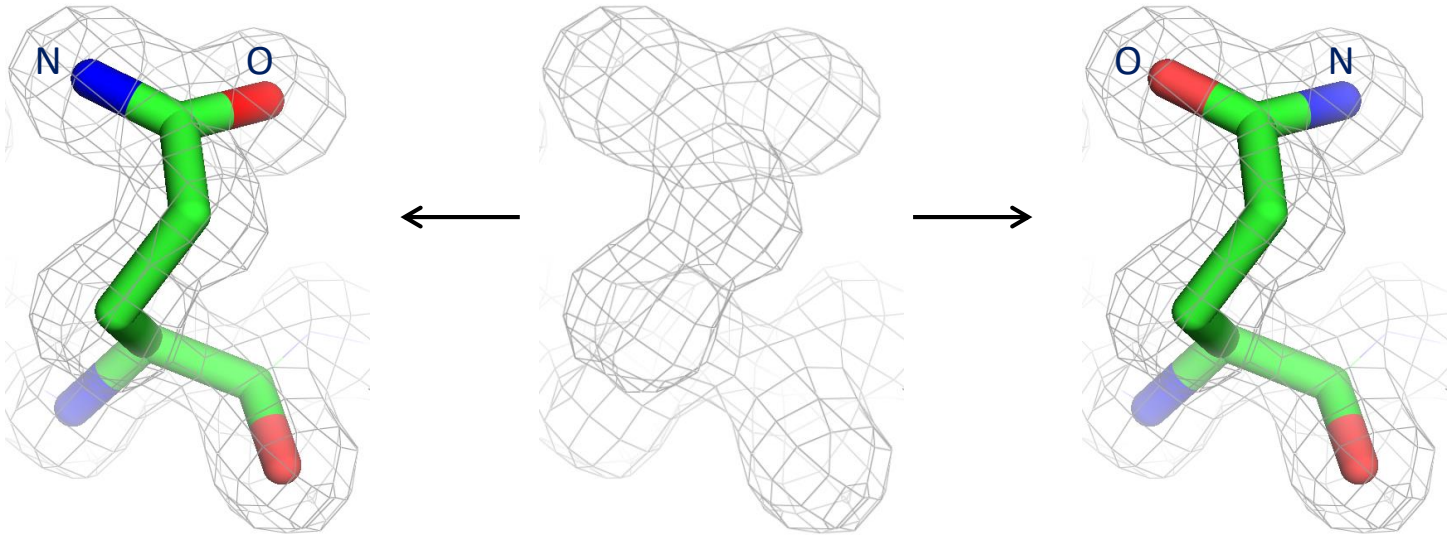
Examples of random errors

- uncertainties in atomic positions
- typically in range of 0.01 - 1.27 Å, median 0.28 Å



Examples of random errors

- side chain flips
  - His/Asn/Gln – symmetrical in terms of shape → fit electron density equally well when rotated by 180°



difficult to distinguish N and O atoms of the side-chain amide from X-ray data

Selecting reliable structure

- rules of thumb for selecting structures
  - X-ray structures
  - NMR structures
- quality checks of structures
  - validation of protein structures
  - programs for quality checks
  - quality information on the web

Rules of thumb for selecting structures

- X-ray structures
  - reasonably accurate structure: resolution ≤ 2.0 Å and *R*-factor ≤ 0.2
  - selection criteria always depend on the type of analysis required (e.g., comparison of folds – 3.0 Å resolution is sufficient vs. analysis of side chain torsional conformers – resolution ≤ 1.2 Å is required)
  - *R*-factor can easily be fooled → a better indicator of model reliability is R_{free} – *c*alculated in the same way as *R*-factor but using only a small fraction of the experimental data; R_{free} should be ≤ 0.4
  - local errors indicated by residue B-factors > 50 but quality checks should always be performed to assess possible local problems in a structure

Rules of thumb for selecting structures

- NMR structures
  - no simple rule of thumb as in the case of X-ray structures
  - information on structure quality can be found in the original paper or obtained by quality checks
  - ResProx (http://www.resprox.ca/) – predicts the atomic resolution of NMR protein structures using machine learning
  - DRESS (http://www.cmbi.ru.nl/dress/) and RECOORD (http://www.ebi.ac.uk/pdbe-apps/nmr/recoord/main.html)web servers – provide improved versions of old NMR models (obtained by re-refinement of the original experimental data using more up-to-date force fields and refinement protocols)

Quality checks of structures

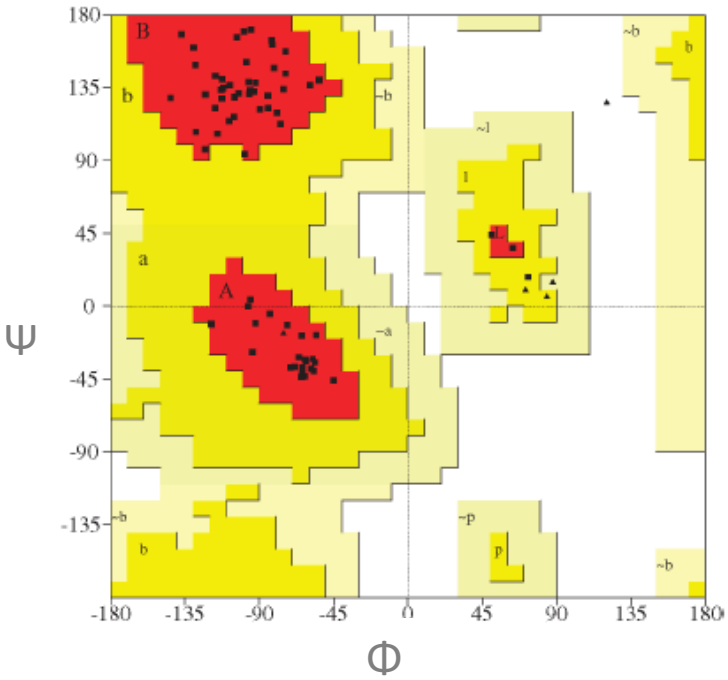
- checks of structure geometry, stereochemistry and other structural properties
- tests of normality
  - comparison of a given protein or nucleic acid structure against what is already known about these molecules
  - knowledge comes from high-resolution structures of small molecules and systematic analyses of existing protein and nucleic acid structures
  - not all outliers from the norm are errors (e.g., an unusual torsion angle of a single residue), however, a structure exhibiting a large number of outliers and oddities is probably problematic

Validation of protein structures

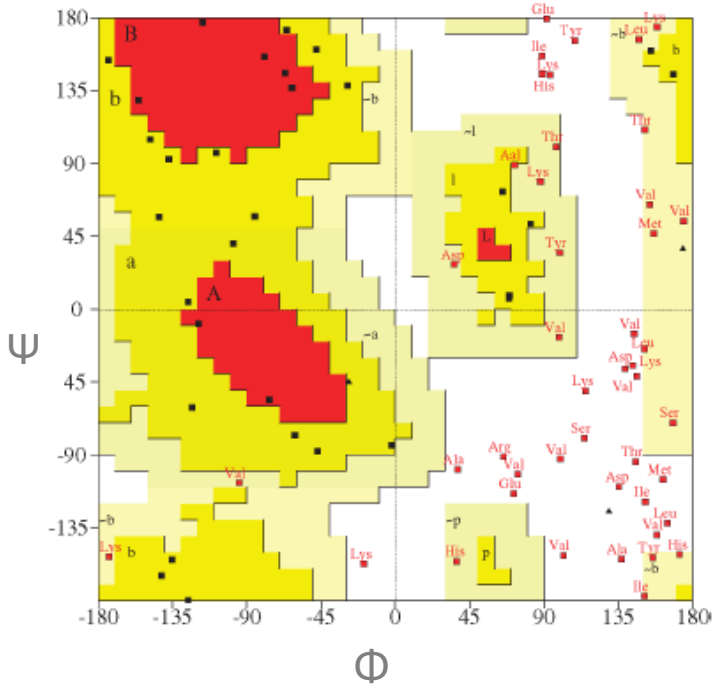
- Ramachandran plot
  - check of stereochemical quality of protein structures
  - plot of the Ψ versus the Φ main chain torsion angles for every amino acid residue in the protein (except the two terminal residues)
  - favorable and "disallowed" regions of the plot determined from analyses of existing structures
  - typical protein structures – residues tightly clustered in the most favored regions, only few or none residues in the "disallowed" regions
  - poorly defined protein structures– residues more dispersed and many of them lie in the "disallowed" regions of the Ramachandran plot

Validation of protein structures

- Ramachandran plot



typical protein structure



poorly defined protein structure

Validation of protein structures

- Ramachandran plot

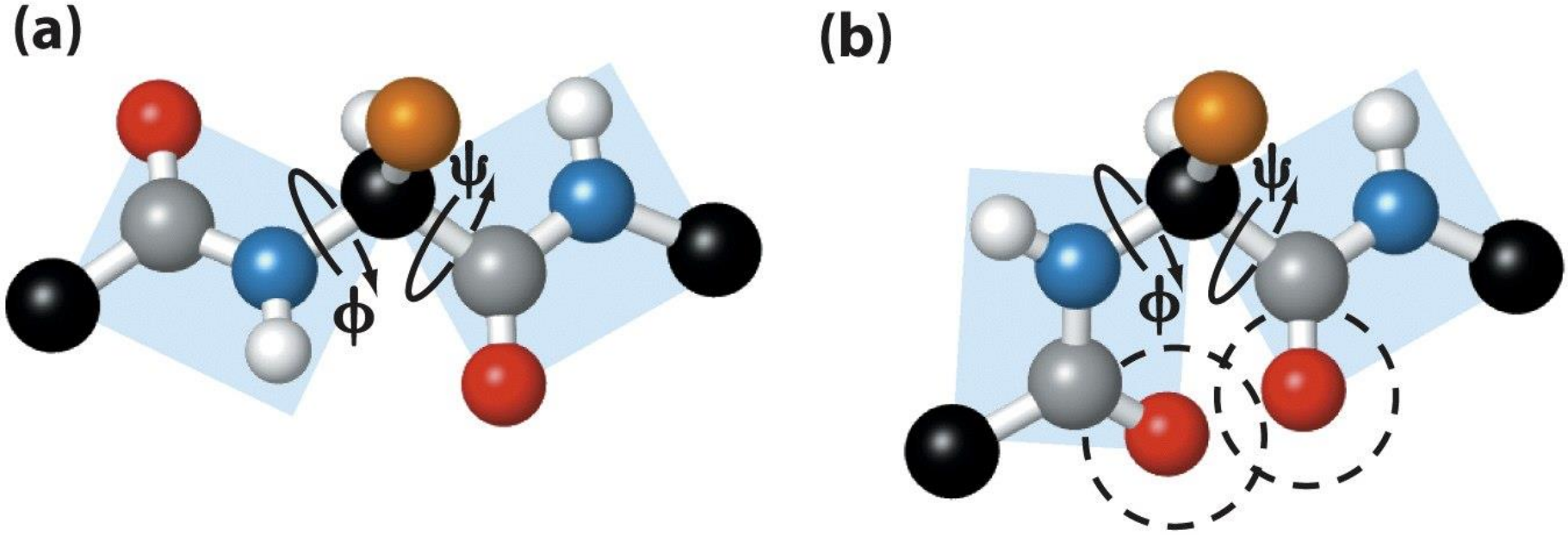
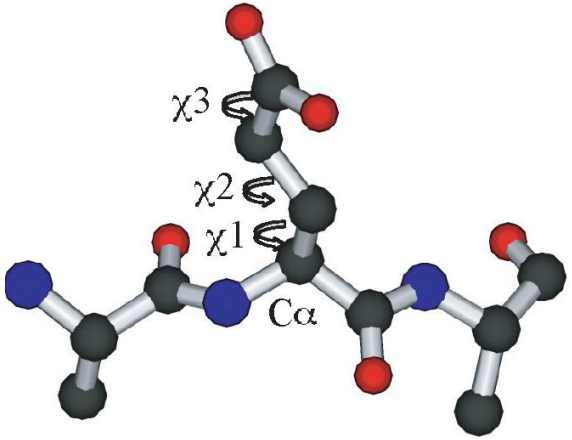
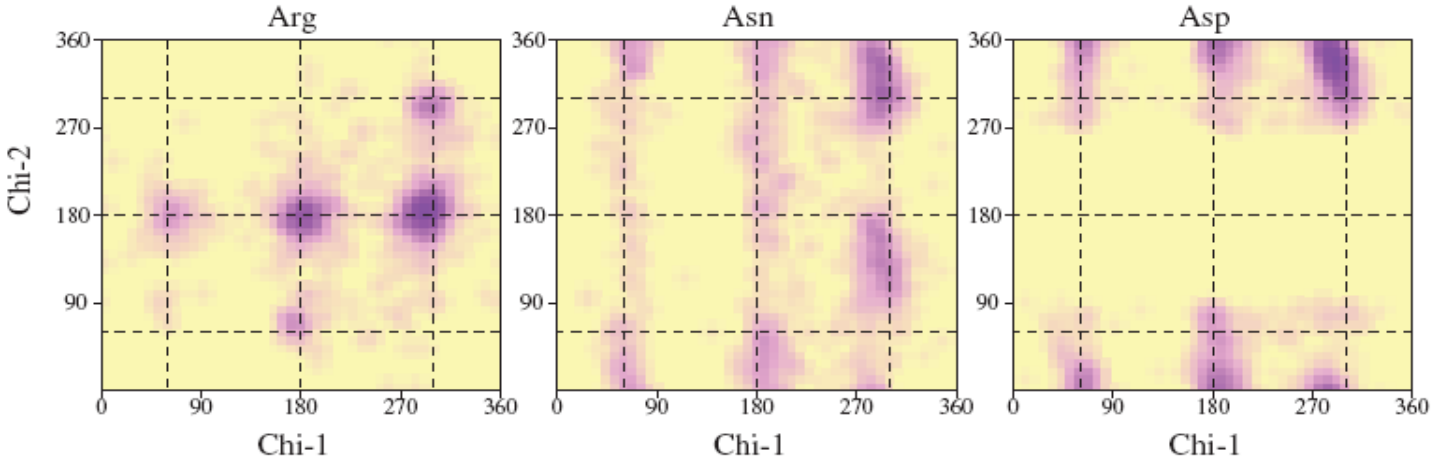


Figure 4-8 Principles of Biochemistry, 4/e
© 2006 Pearson Prentice Hall, Inc.

Validation of protein structures

- side chain torsion angles
  - preferred conformations of side chain torsion angles obtained by analyses of existing structures
  - χ_1 – torsion angle about N-C^{α}-C^{β}-A^{γ}
  - χ_2 – torsion angle about C^{α}-C^{β}-A^{γ}-A^{δ}, ...



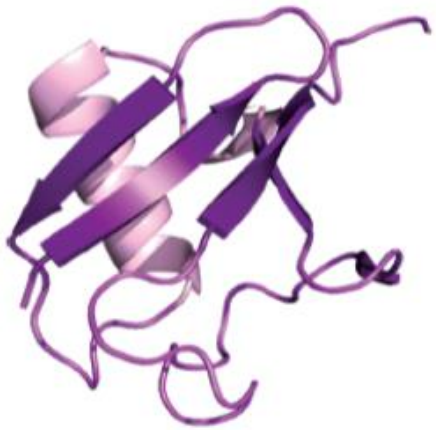


Validation of protein structures

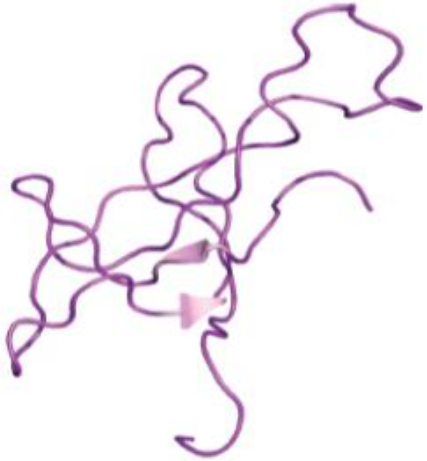
- bad and unfavorable atom-atom contacts
  - "simple" count of bad contacts, e.g., two nonbonded atoms with a center-to-center distance < sum of their van der Waals radii
  - evaluation of the environment of individual atoms or residue fragments with respect to the environments found in the high resolution crystal structures

Validation of protein structures

- secondary structure
  - ~ 50-60% of residues usually in regions of regular secondary structure
  - poorly defined structures – main chain O and N atoms can lie beyond normal hydrogen bonding distances → some of the α-helices and β-strands not detected by the secondary structure assignment programs



typical protein structure



poorly defined protein structure

Validation of protein structures

- other parameters
  - counts of unsatisfied hydrogen bond donors
  - hydrogen bonding energies
  - knowledge-based potentials assessing how “happy” each residue is in its local environment – many unhappy residues → “sad” overall structure
  - real space *R*-factor expressing how well each residue fits its electron density; can also be expressed as a Real-space correlation coefficient

Programs for quality checks

- Proteins
  - PROCHECK
  - WHAT_CHECK
  - Verify 3D
  - MolProbity
  - ANOLEA

Programs for quality checks

- PROCHECK
  - http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/
  - variety of plots for protein structures: Ramachandran plot, χ_1-χ_2 plot for each amino acid type, main chain bond lengths and bond angles, secondary structure plot, ...
  - parameters that deviate from norm are highlighted
  - NMR-PROCHECK – version specific for NMR

Programs for quality checks

- WHAT_CHECK (subset of WHAT IF package)
  - http://swift.cmbi.ru.nl/gv/whatcheck/
  - space group and symmetry
  - bond lengths and angles
  - bad contacts
  - hydrogen bonds
  - ....

  - detailed output of discrepancies of the given protein structure from the norms

Programs for quality checks

- Verify3D
  - https://genesilico.pl/toolkit/unimod?method=Verify3D
  - evaluates residue's environment in terms of secondary structure, buried surface area, and fraction of side chain covered by polar atoms
- MolProbity
  - http://molprobity.biochem.duke.edu/
  - detailed all-atom contact analysis within a given protein structure
- ANOLEA
  - http://melolab.org/anolea/index.html
  - knowledge based evaluation of atom-atom contacts

Quality information on the web

- several databases provide pre-computed quality criteria for all wwPDB structures
  - EDS
  - PDBsum
  - PDBREPORT
  - RCSB PDB

Quality information on the web

- Electron Density Server (EDS)
  - http://eds.bmc.uu.se/eds/, also available via the PDBe site
  - information about local quality of the structure for all structures from wwPDB with deposited experimental data
  - plot of real-space *R*-factor (RSR) – how well each residue fits its electron density
  - plot of Z-score – large positive spike → residue has considerably worse RSR than the average residue of the same type in structures determined at similar resolution.
  - Ramachandran plot
  - …

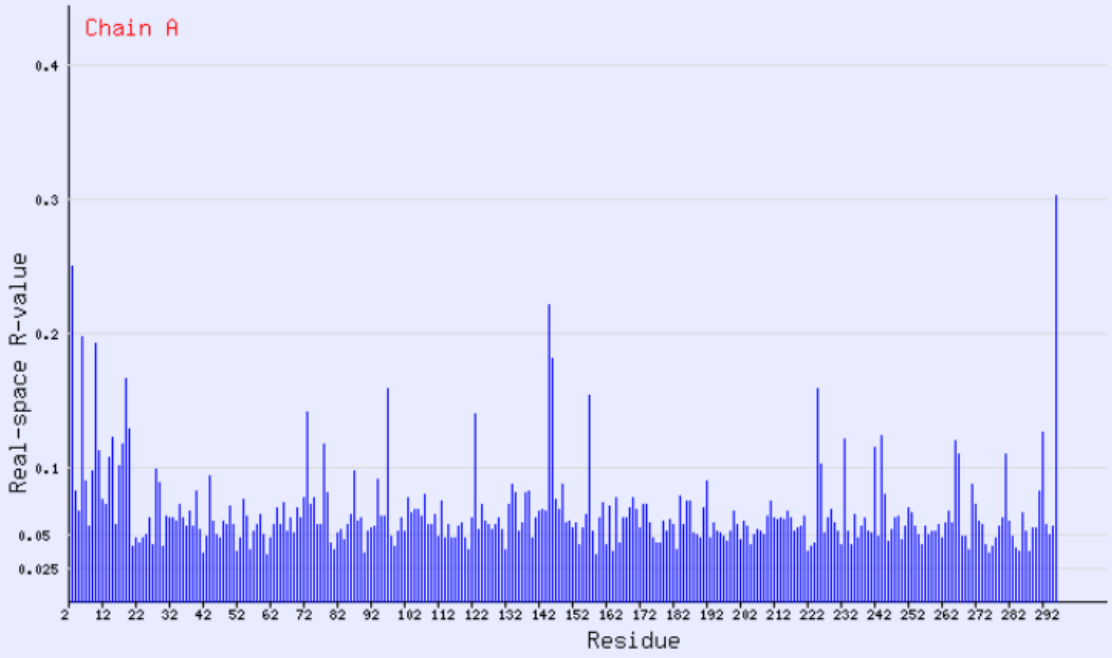
Quality information on the web

- Electron Density Server (EDS)

Real-space R-value vs Residue for 1iz7

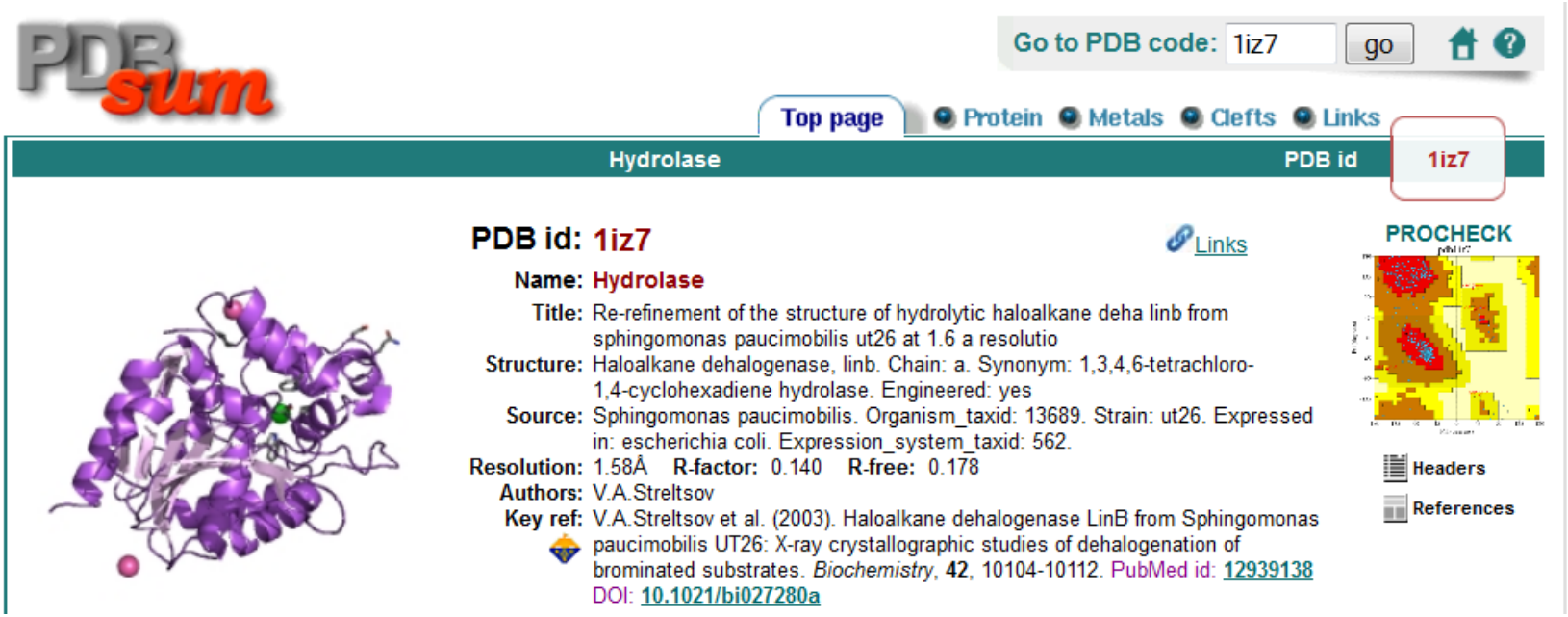
Map R-value 0.147

Number of residues in chain A: 294



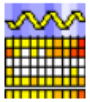
Quality information on the web

- PDBsum
  - http://www.ebi.ac.uk/pdbsum/
  - provides numerous structural analyses of all wwPDB structures, including full PROCHECK output (for all protein-containing entries)



Quality information on the web

- PDBsum

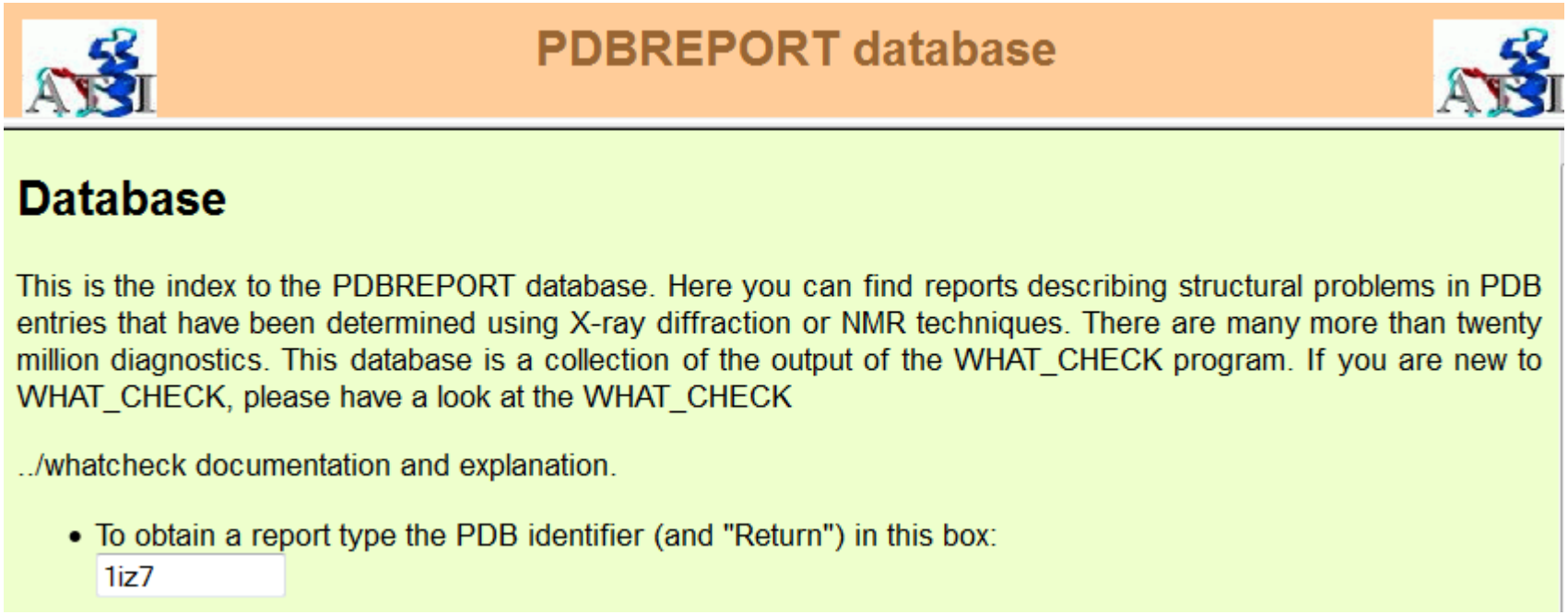


PROCHECK analyses for 1iz7

No.	Plot description	Plot files	Description
1	Main Ramachandran plot	PostScript PDF	
2	All-residue Ramachandran plots	PostScript PDF	
3	All-residue chi1-chi2 plots	PostScript PDF	
4	Main-chain parameters	PostScript PDF	
5	Side-chain parameters	PostScript PDF	
6	Residue properties plot	PostScript PDF	
7	Main-chain bond lengths	PostScript PDF	
8	Main-chain bond angles	PostScript PDF	
9	RMS distances from planarity	PostScript PDF	
10	Distorted geometry	PostScript PDF	

Quality information on the web

- PDBREPORT
  - http://swift.cmbi.ru.nl/gv/pdbreport/
  - provides a pre-computed WHAT_CHECK report for any structure in the wwPDB



Quality information on the web

- PDBREPORT

Warning: Unusual bond angles

The bond angles listed in the table below were found to deviate more than 4 sigma from standard bond angles (both standard values and sigma for protein residues have been taken from Engh and Huber [REF], for DNA/RNA from Parkinson et al [REF]). In the table below for each strange angle the bond angle and the number of standard deviations it differs from the standard values is given. Please note that disulphide bridges are neglected. Atoms starting with "-" belong to the previous residue in the sequence.

17	ARG	(19-)	A	N	CA	C	127.61	5.9
17	ARG	(19-)	A	C	CA	CB	101.78	-4.4
30	ILE	(32-)	A	N	CA	C	97.87	-4.8
132	ILE	(134-)	A	N	CA	C	99.73	-4.1

Error: Nomenclature error(s)

Checking for a hand-check. WHAT IF has over the course of this session already corrected the handedness of atoms in several residues. These were administrative corrections. These residues are listed here.

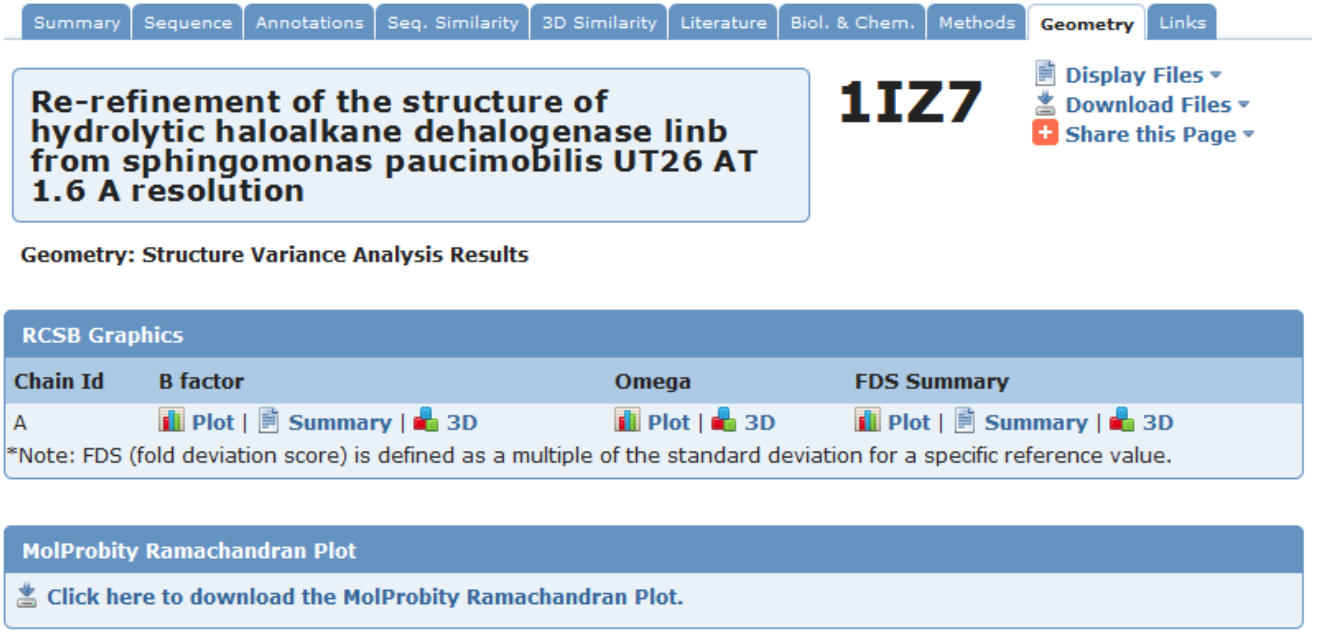
231	GLU	(233-)	A

Error: Tau angle problems

The side chains of the residues listed in the table below contain a tau angle (N-Calpha-C) that was found to deviate

Quality information on the web

- RCSB PDB
  - http://pdb.rcsb.org/
  - provides geometrical analyses for each entry, including information about bond lengths, angles and dihedral angles



References

- Gu, J. & Bourne, P. E. (2009). **Structural Bioinformatics, 2nd Edition**, Wiley-Blackwell, Hoboken, p. 1067.
- Xiong, J. (2006). **Essential Bioinformatics**. Cambridge University Press, New York, p. 352.
- Schwede, T. & Peitsch, M. C. (2008). **Computational Structural Biology: Methods and Applications**, World Scientific Publishing Company, Singapore, p. 700.
- Shapiro, B. A. *et al.* (2007). Bridging the gap in RNA structure prediction. *Current opinion in structural biology* **17**: 157-165.