

LOSCHMIDT  
LABORATORIES



# Structural databases & Models of structures



# 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

	HEADER	LYASE (CARBON-CARBON)				03-JUL-95		1DNP				
structure annotation	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
amino acid field	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
	....											
cofactor filed	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
residue name
residue number
x, y, z coordinates
occupancy
temperature factor
atom type

atom name
polypeptide chain identifier

# PDB format



- ❑ atomic coordinates
- ❑ chemical and biological features
- ❑ experimental details of the structure determination
- ❑ structural features
  - secondary structure assignments
  - hydrogen bonding
  - biological assemblies
  - 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



- ❑ **macromolecular Crystallographic Information File (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

- ❑ Protein Data Bank Markup Language (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
        VPSRFGSGSGTQYSLKINSLQPEDFGSYQCQHFWSLTPRTFGGGTKLEIK
      </PDBx:pdbx_seq_one_letter_code>
      <PDBx:pdbx_seq_one_letter_code_can>
        DIVLTQSPASLSASVGETVTITCRASGNIHNYLAWYQQKQGKSPQLLVYYTTTLADG
        VPSRFGSGSGTQYSLKINSLQPEDFGSYQCQHFWSLTPRTFGGGTKLEIK
      </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

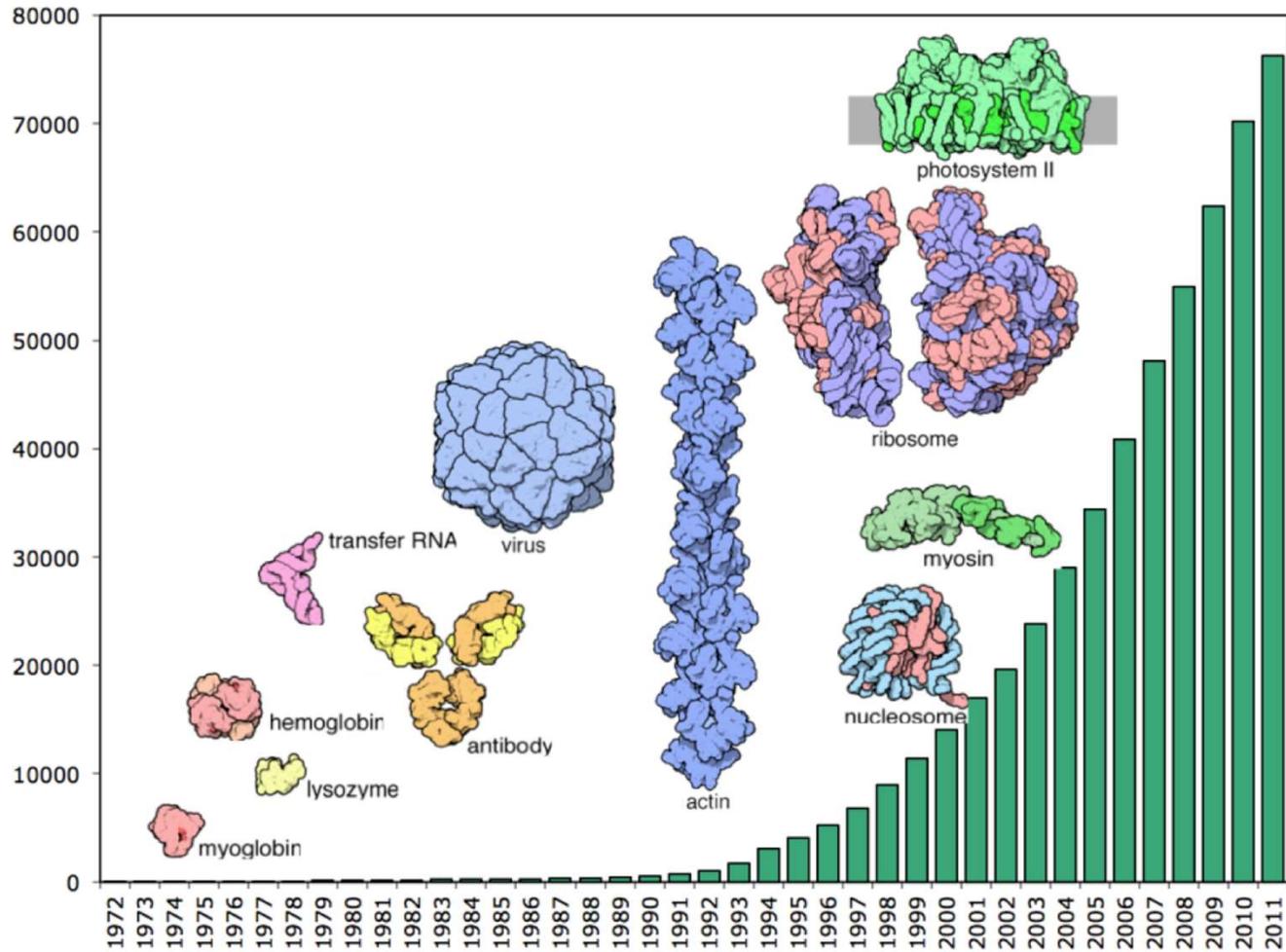
- PDBsum, SCOP, Protopedia, Structural Biology KnowledgeBase



- 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)



## □ database growth





## □ worldwide Protein Data Bank (wwPDB)

- <http://www.wwpdb.org/>
- central repository of **experimental macromolecular structures**
- more than 170,000 structures (October 2020), updated every week
- mostly **protein structures** (93 %), structures of protein/nucleic acids complexes (5 %) and nucleic acid structures (3 %)
- majority of structures from **X-ray** crystallography (88 % ), **NMR** (8 %), or **EM** (4%)
- 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>

The screenshot shows the top navigation bar of the RCSB PDB website. It includes a dark blue header with the following menu items: RCSB PDB, Deposit, Search, Visualize, Analyze, Download, Learn, and More. On the right side of the header is a 'MyPDB Login' button. Below the header is a white search bar with the text 'Search by PDB ID, author, macromolecule, sequence, or ligands' and a 'Go' button. To the left of the search bar is the RCSB PDB logo and the text '134251 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education'. Below the search bar are several logos for partner databases: PDB-101, Worldwide PDB Protein Data Bank, EMDatabank, Nucleic Acid Database, and Worldwide Protein Data Bank Foundation. On the right side of the search bar area, there are links for 'Advanced Search', 'Browse by Annotations', 'Search History (2)', and 'Previous Results (110)'. At the bottom right of the header area, there is a 'Take the RCSB PDB User Survey' button and social media icons for Facebook, Twitter, YouTube, and LinkedIn.

A vertical navigation sidebar with a dark blue background and white text. The items are: Welcome (with a blue bookmark icon), Deposit (with a blue upload icon), Search (with a magnifying glass icon), Visualize (with a blue camera icon), Analyze (with a blue grid icon), Download (with a blue download icon), and Learn (with a blue book icon).

## A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

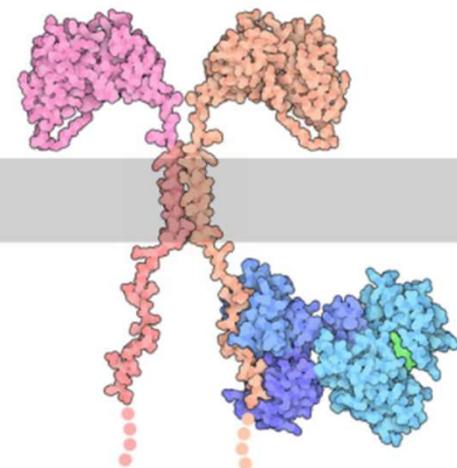
As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

### 2017 RCSB PDB User Survey

A graphic for the 2017 RCSB PDB User Survey. It features a central orange speech bubble with the text 'RCSB PDB User Survey'. Surrounding it are several grey speech bubbles with survey questions: 'I visit rcsb.org every ...', 'I'm a depositor', 'I'm a student', 'My research interests are ...', 'I use PDB-101 every ...', 'I wish rcsb.org offered ...', 'My favorite rcsb.org feature is ...', and 'I work at ...'.

## October Molecule of the Month



Chimeric Antigen Receptors

# PDBe

□ <http://www.ebi.ac.uk/pdbe/>

The screenshot shows the PDBe website homepage. At the top, there is a navigation bar with links for Services, Research, Training, and About us. Below this is the PDBe logo and the text "Protein Data Bank in Europe Bringing Structure to Biology". A search bar is located on the right side of the header, with a search button and a dropdown menu showing examples like "hemoglobin" and "BRCA1\_HUMAN". Below the header is a secondary navigation bar with links for PDBe home, Deposition, PDBe services, PDBe training, Documentation, and About PDBe. The main content area features a "Featured structure" section for Solanezumab, an anti-Alzheimer's antibody, with a 3D ribbon diagram and a description. Below this are sections for "News" and "Events", each with a list of recent updates. On the right side, there is a "Popular" section with links to various tools and resources, a "Latest archive statistics" section, and a "Connect with us" section with social media links for Facebook, YouTube, Twitter, and RSS. At the bottom right, there is a "Tweets" section with a "Follow" button.

EMBL-EBI Services Research Training About us

**Protein Data Bank in Europe**  
Bringing Structure to Biology

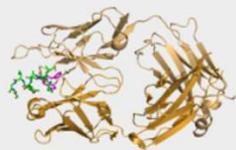
Search  Search  
Examples: hemoglobin, BRCA1\_HUMAN EMsearch

**PDBe home** Deposition PDBe services PDBe training Documentation About PDBe Share Feedback

PDBe is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures.  
[Read more about PDBe.](#)

### Featured structure

Solanezumab. An anti-Alzheimer's antibody 23rd July 2015



Solanezumab, an antibody which may slow the progression of Alzheimer's disease, recognizes a central portion of amyloid beta. Its structure reveals how.

[Read more...](#)

[Previous featured structures](#)

### News

PDBe webinar available online  
27 July, 2015

EMDataBank announces 2015 Map Challenge  
23 July, 2015

### Events

[EMBL-EBI course: Structural Bioinformatics](#)  
**Hinxton, UK**  
12 Oct 2015 to 16 Oct 2015

[PDBe Roadshow](#)

### Popular

- EMsearch
- PDBeFold
- PDBePISA
- Sequence search
- PDBe REST API
- EM resources
- NMR resources
- EMPIAR
- News
- Events
- Training
- Contact us

### Latest archive statistics

As of 30 September 2015 the PDB contains 112561 entries ([latest PDB entries](#), [chemistry](#), [biology](#)) and EMDB contains 3200 entries ([latest map releases](#), [latest header releases](#), [latest updates](#)).

### Connect with us

- Facebook
- Twitter
- YouTube
- RSS

### Tweets

[Follow](#)

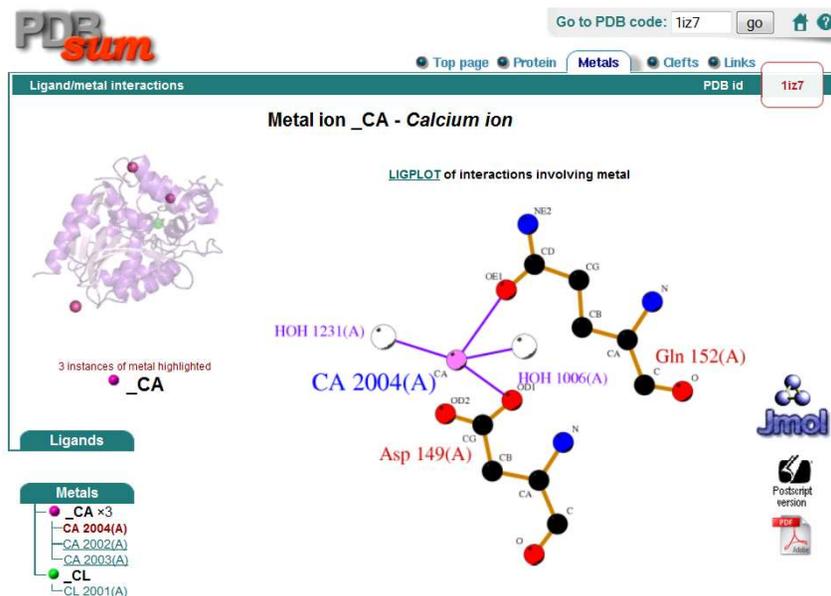
□ <http://www.pdbj.org/>

The screenshot displays the PDBj website interface. At the top left, it shows '112561 entries available on 2015-09-29 17:00 UTC / 09:00 JST'. The main header includes the PDBj logo and navigation links for English, Japanese, Chinese, and Korean. A search bar is prominently featured. The left sidebar contains several menu categories: Home, Data deposition, Download, New format, and Search. The main content area includes a 'Guide for first time visitors', a 'Find the service you need' section with a grid of radio buttons for various services (PDB, BMRB, EMDB, search, deposition, viewer, education/dictionary, NMR, electron microscopy, secondary structure, sequence, similarity, function prediction, chemical component, structure prediction, binding site, surface structure, 3D structure), and a 'Latest news' section. On the right, there is a 'Molecule of the Month' section featuring a 3D model of amyloids and a 'Latest new entries' link.

# 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

- ❑ **Structural Classification of Proteins (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:**

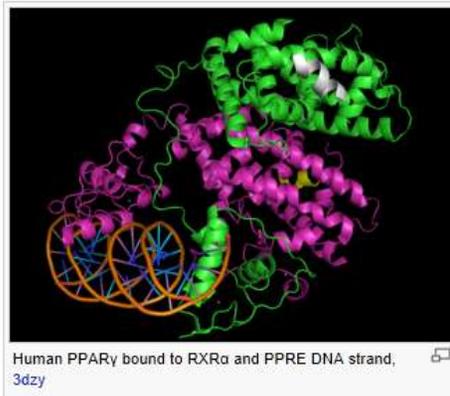
1. Root: [scop](#)
2. Class: [Alpha and beta proteins \(a/b\)](#) [51349]  
*Mainly parallel beta sheets (beta-alpha-beta units)*
3. Fold: [alpha/beta-Hydrolases](#) [53473]  
*core: 3 layers, a/b/a; mixed beta-sheet of 8 strands, order 12435678, strand 2 is antiparallel to the rest*
4. Superfamily: [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

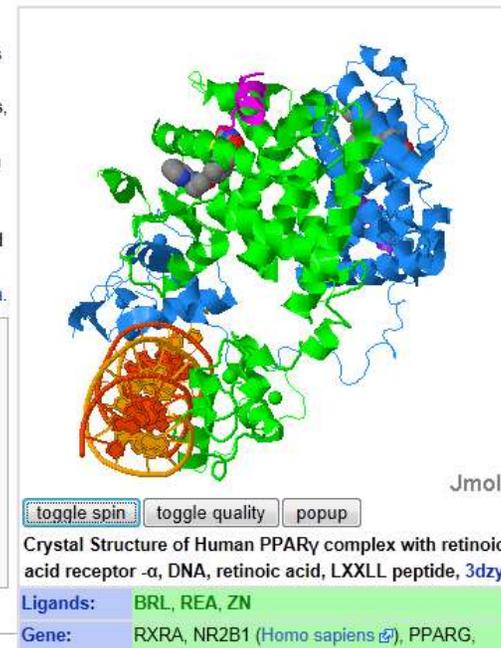


The Peroxisome Proliferator-Activated Receptors (PPAR)  $\alpha$ ,  $\gamma$ , and  $\delta$  are members of the nuclear receptor family. Since their discovery in the early 90s, it has become clear that the PPARs are essential modulators of external stimuli, acting as transcription factors to regulate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

#### Contents [hide]

- 1 Biological Role
- 2 Natural Ligands
- 3 PPAR Structure
- 4 Binding of Synthetic Agonists and Medical Implications
- 5 Additional 3D Structures of PPAR
- 6 Additional Resources
- 7 References

#### Biological Role

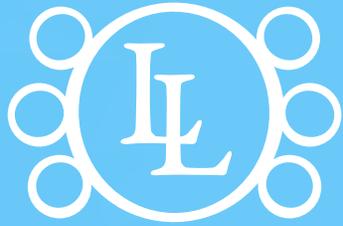


# Other structure-based resources

## ❑ Structural Biology Knowledgebase

- <http://sbkb.org/>
- provides up-to-date information about advances in structural biology and structural genomics

The screenshot shows the homepage of the Structural Biology Knowledgebase (SBKB). The header includes the site name "PSI | StructuralBiologyKnowledgebase" and navigation links: Home, Protein Resources, Homology Models, Methods & Technologies, E-Collection, and About. A search bar is prominently displayed with the text "Search for proteins, models, methods, and more...". Below the search bar are radio buttons for search criteria: "by sequence", "by text", "by pdb id", and "by uniprot ac". A search input field contains the example sequence "MKLTLKNLISMAIMMSTIVMGSSAMAADSNEKIVAHRGASGYLPEHTLPKAMAAYA" and a "Go" button. The main content area is divided into several sections: "About this Site" with a protein structure image and text describing the knowledgebase; "Protein Resources" with links to Sequence Data Repositories, Structural Biology Resources, Function Resources, KB-Rank Structure Search Tool, KB-Role Function Prediction Tool, Functional Sleuth, and Sequence Comparison Tool; "Homology Models" with links to Protein Model Portal (PMP), Interactive Modelling, and Model Archive; "Methods & Technologies" with links to TargetTrack, Technology Reports, Order PSI Clones, and Synchrotron Information (BioSync); "Latest PSI Results" with a table showing statistics: New structures last month: 21, Total structures to date: 6920, Total distinct structures: 5472, Total community structures: 599, and a link to View PSI Metrics; "Latest Structures" with a 3D protein structure image and text: Centers: MCGS NatPro, PDBID: 5CQF, Crystal structure of L-lysine 6-monooxygenase from Pseudomonas syringae, and a link to View all latest structures; and "E-Collection" with a link to E-Collection and a featured article titled "Nuclear Pore Complex: Integrative Approach to Probe Nup133" with a 3D protein structure image and percentages (50.6%, 24.2%, 5.0%).



LOSCHMIDT  
LABORATORIES



## Structural quality assurance

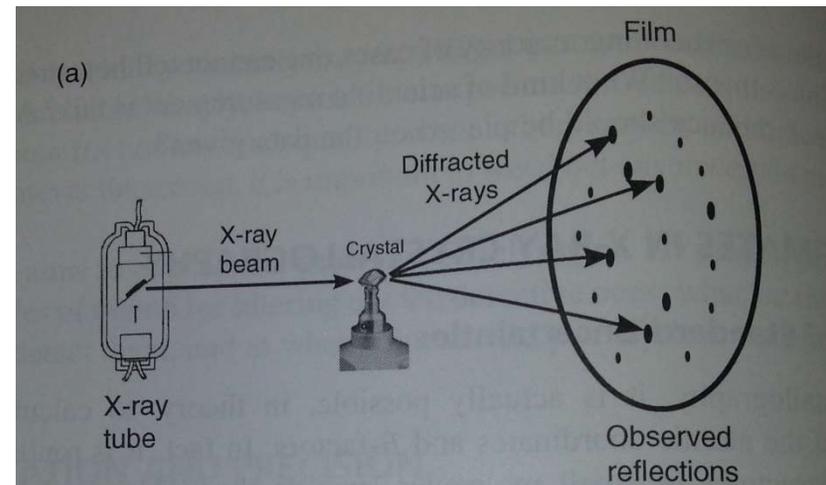
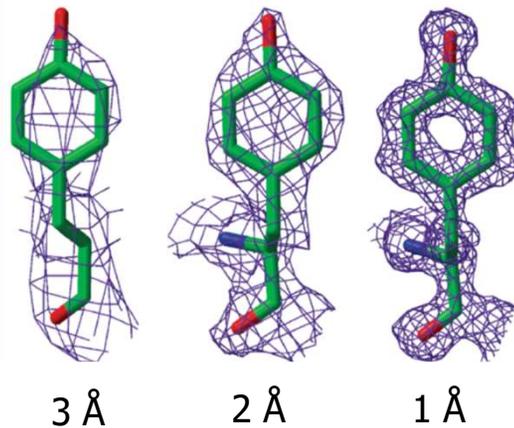
# Outline

- ❑ 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

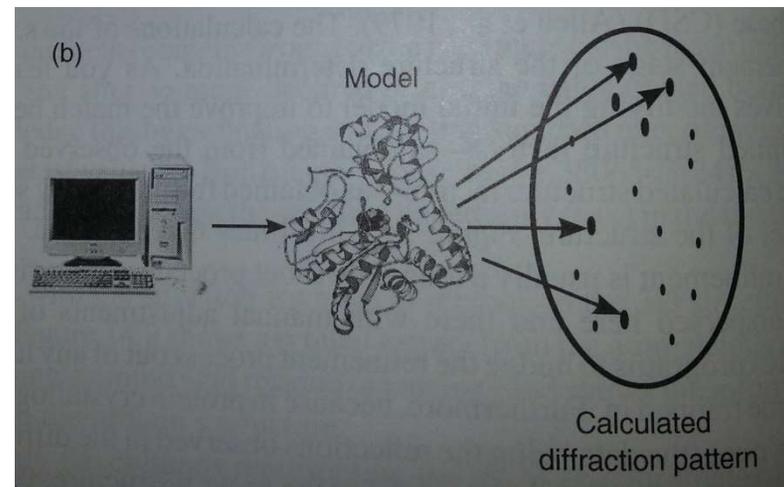
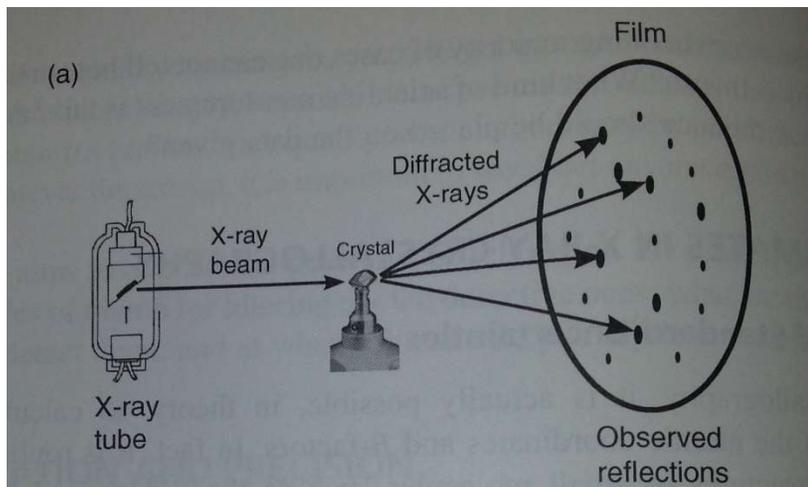
- measure of the level of detail present in the diffraction pattern



# Concepts

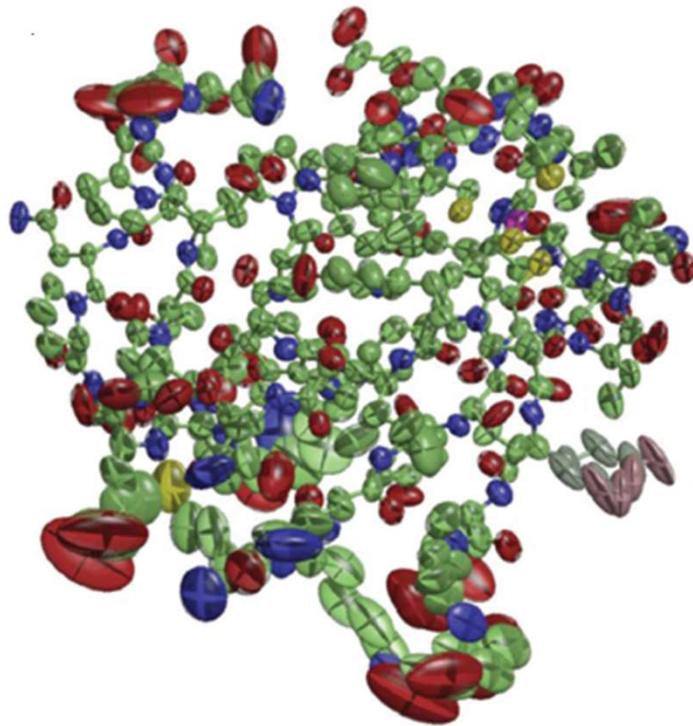
## □ R-factor (R-value)

- 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

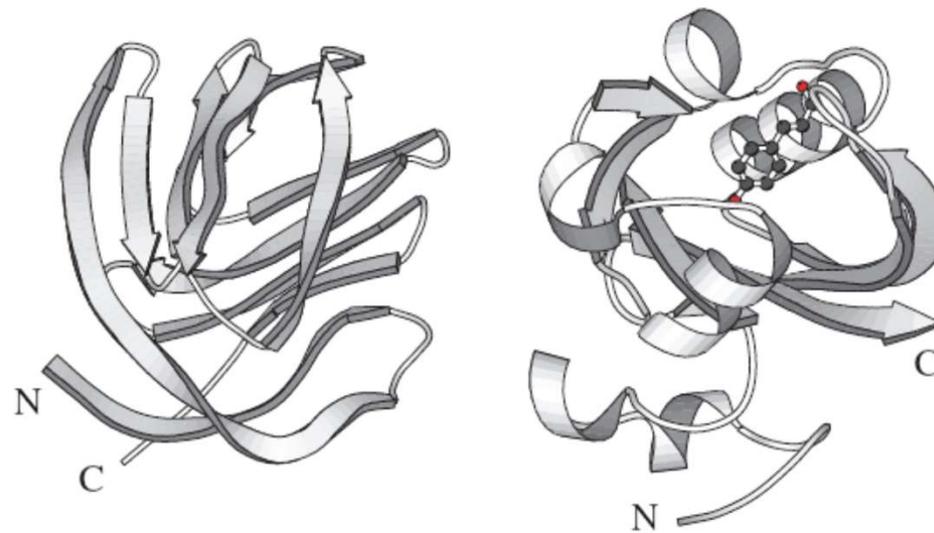


- ❑ relate to the **accuracy** of the model—how well it corresponds to the “true” structure of the molecule in question
- ❑ often include errors of **interpretation**
  - low quality of electron density map → difficult to find the correct tracing of the molecule(s) through it → misstracing and “frame-shift” 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**



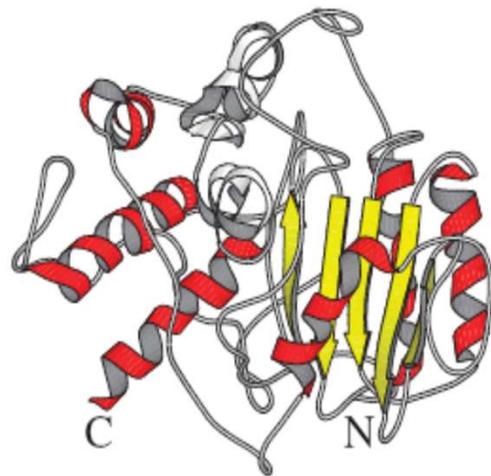
Incorrect model (1PHY)

Corrected model (2PHY)

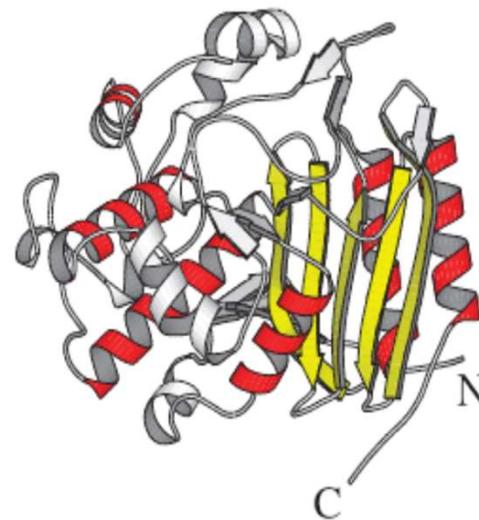
# Examples of systematic errors



- ❑ 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** ( $> 3.0 \text{ \AA}$ ), 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

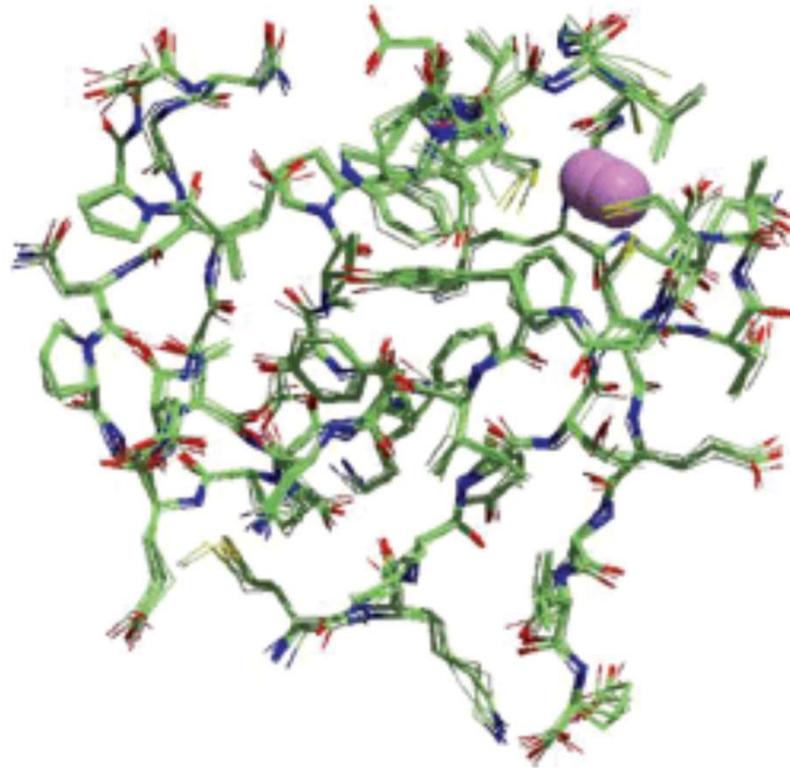


- ❑ depend on how **precisely** a given measurement can be made
- ❑ all measurements contain errors at some degree of precision
- uncertainties in atomic positions
- ❑ **less serious** than systematic errors
- ❑ if a structure is essentially correct, the sizes of the random errors determine how precise the structure is

# Examples of random errors



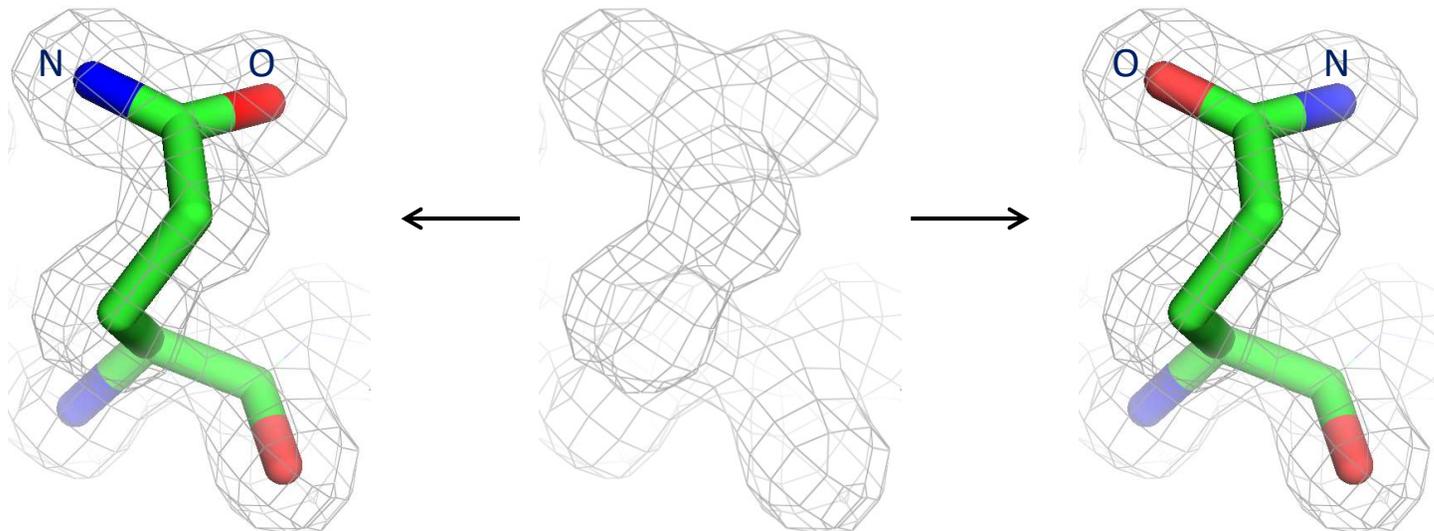
- ❑ uncertainties in atomic positions
- ❑ typically in range of 0.01 - 1.27 Å, median 0.28 Å



# 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  $\leq 2.0 \text{ \AA}$**  and  **$R$ -factor  $\leq 0.2$**
- selection criteria always **depend on the type of analysis** required (e.g., comparison of folds –  $3.0 \text{ \AA}$  resolution is sufficient vs. analysis of side chain torsional conformers – resolution  $\leq 1.2 \text{ \AA}$  is required)
- $R$ -factor can easily be fooled  $\rightarrow$  a better indicator of model reliability is  **$R_{\text{free}}$**  – calculated in the same way as  $R$ -factor but using only a small fraction of the experimental data;  $R_{\text{free}}$  should be  **$\leq 0.4$**
- 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

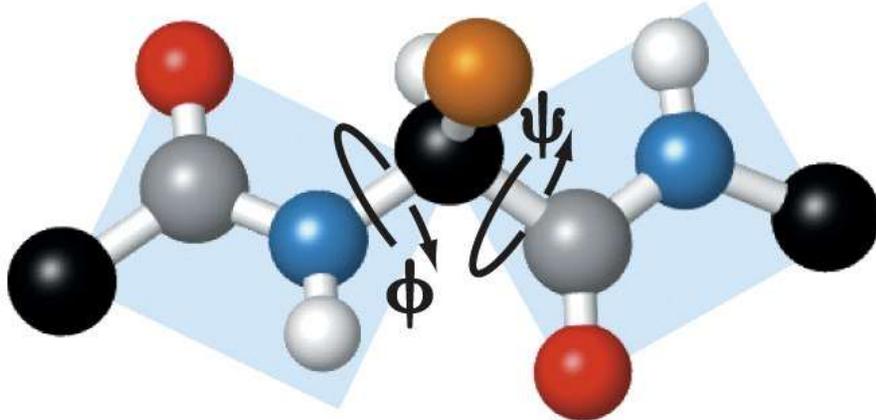
- check of stereochemical quality of protein structures
- plot of the  $\Psi$  versus the  $\Phi$  main chain torsion angles for every amino acid residue in the protein (except the two terminal residues)
- **favorable** and **“disallowed”** regions of the plot determined from analyses of existing structures
- 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

(a)



(b)

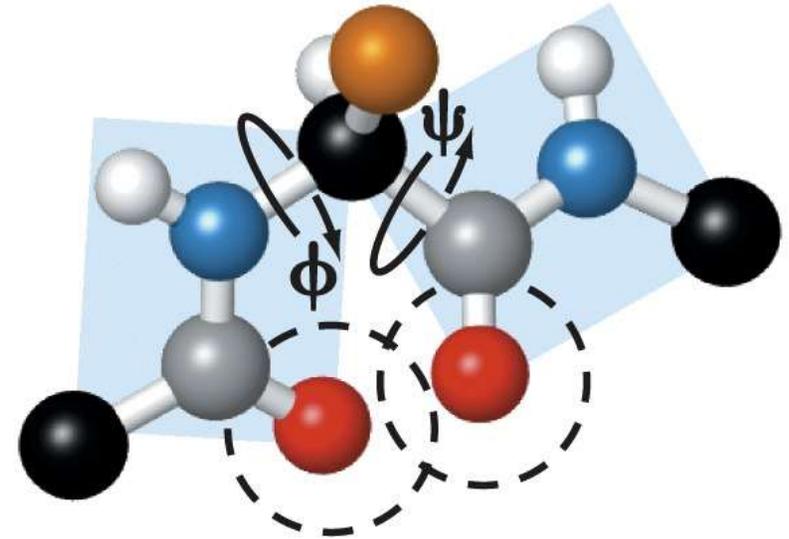
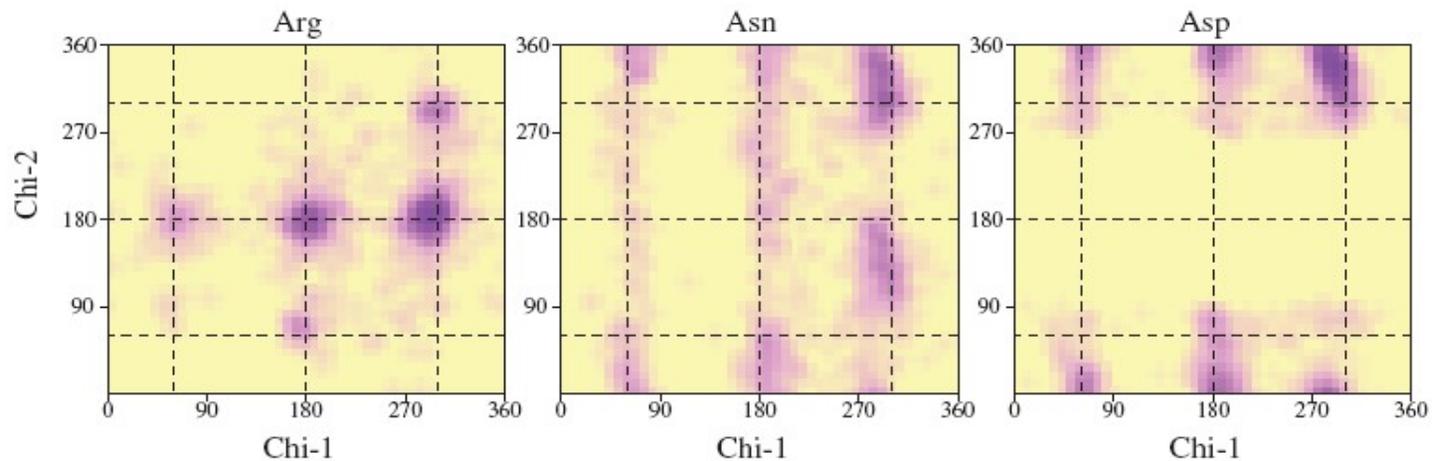
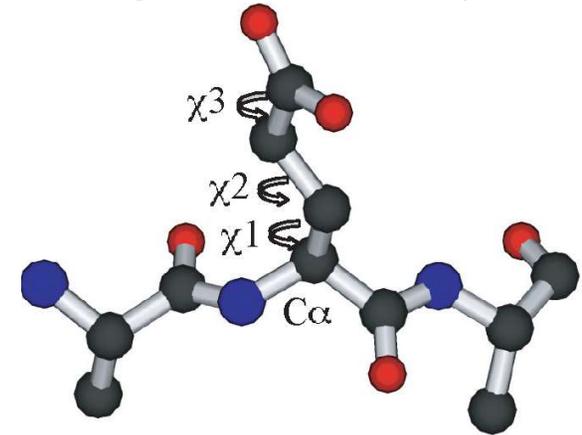


Figure 4-8 Principles of Biochemistry, 4/e  
© 2006 Pearson Prentice Hall, Inc.

# Validation of protein structures

## □ side chain torsion angles

- preferred **conformations of side chain torsion angles** obtained by analyses of existing structures
- $\chi_1$  – torsion angle about N-C $^\alpha$ -C $^\beta$ -A $^\gamma$
- $\chi_2$  – torsion angle about C $^\alpha$ -C $^\beta$ -A $^\gamma$ -A $^\delta$ , ...



# Validation of protein structures

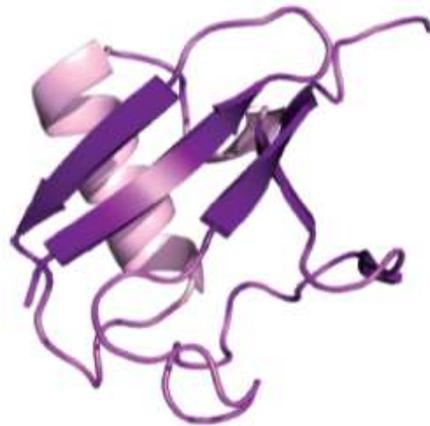


- ❑ bad and unfavorable atom-atom contacts
  - “simple” **count of bad contacts**, e.g., two nonbonded atoms with a center-to-center distance  $<$  sum of their van der Waals radii
  - evaluation of the **environment of individual atoms** or residue fragments with respect to the environments found in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

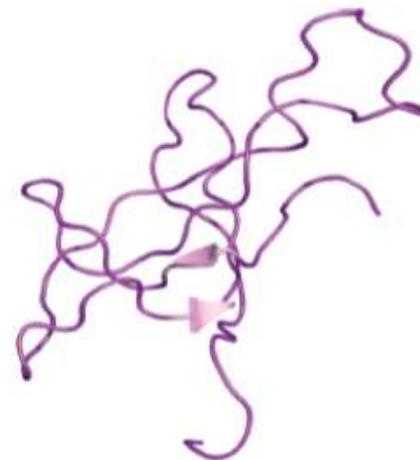
# Validation of protein structures

## □ 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  $\alpha$ -helices and  $\beta$ -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,  $\chi_1$ - $\chi_2$  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

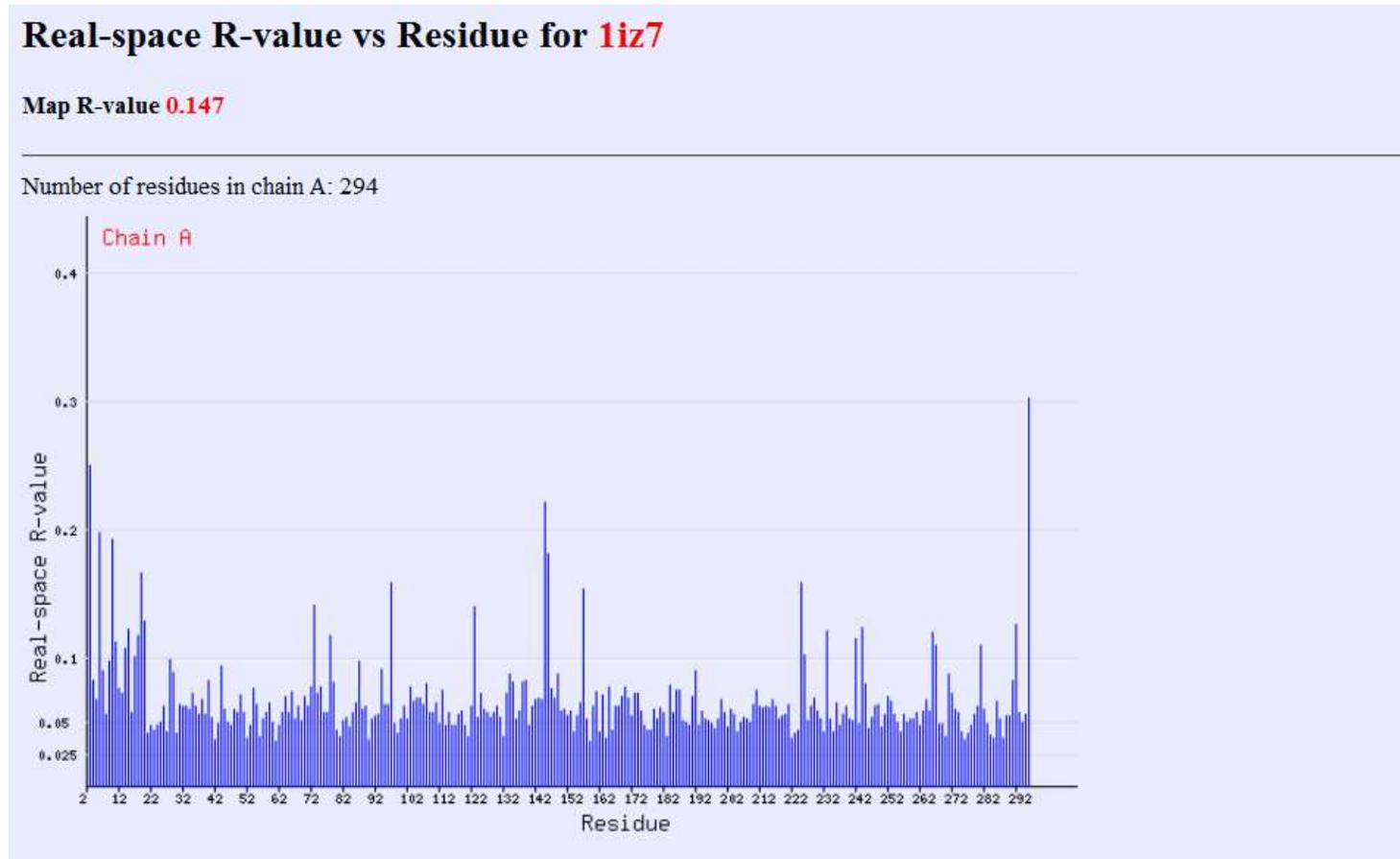
# Quality information on the web



- 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)



# 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)

**PDBsum**

Go to PDB code:

[Top page](#) [Protein](#) [Metals](#) [Glefs](#) [Links](#)

Hydrolase PDB id **1iz7**

**PDB id: 1iz7** [Links](#)

**Name: Hydrolase**

**Title:** Re-refinement of the structure of hydrolytic haloalkane deha linb from sphingomonas paucimobilis ut26 at 1.6 a resolution

**Structure:** Haloalkane dehalogenase, linb. Chain: a. Synonym: 1,3,4,6-tetrachloro-1,4-cyclohexadiene hydrolase. Engineered: yes

**Source:** Sphingomonas paucimobilis. Organism\_taxid: 13689. Strain: ut26. Expressed in: escherichia coli. Expression\_system\_taxid: 562.

**Resolution:** 1.58Å **R-factor:** 0.140 **R-free:** 0.178

**Authors:** V.A.Streltsov

**Key ref:** V.A.Streltsov et al. (2003). Haloalkane dehalogenase LinB from Sphingomonas paucimobilis UT26: X-ray crystallographic studies of dehalogenation of brominated substrates. *Biochemistry*, **42**, 10104-10112. PubMed id: [12939138](#) DOI: [10.1021/bi027280a](#)

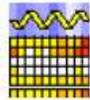
**PROCHECK**

[Headers](#)

[References](#)

# Quality information on the web

## □ PDBsum



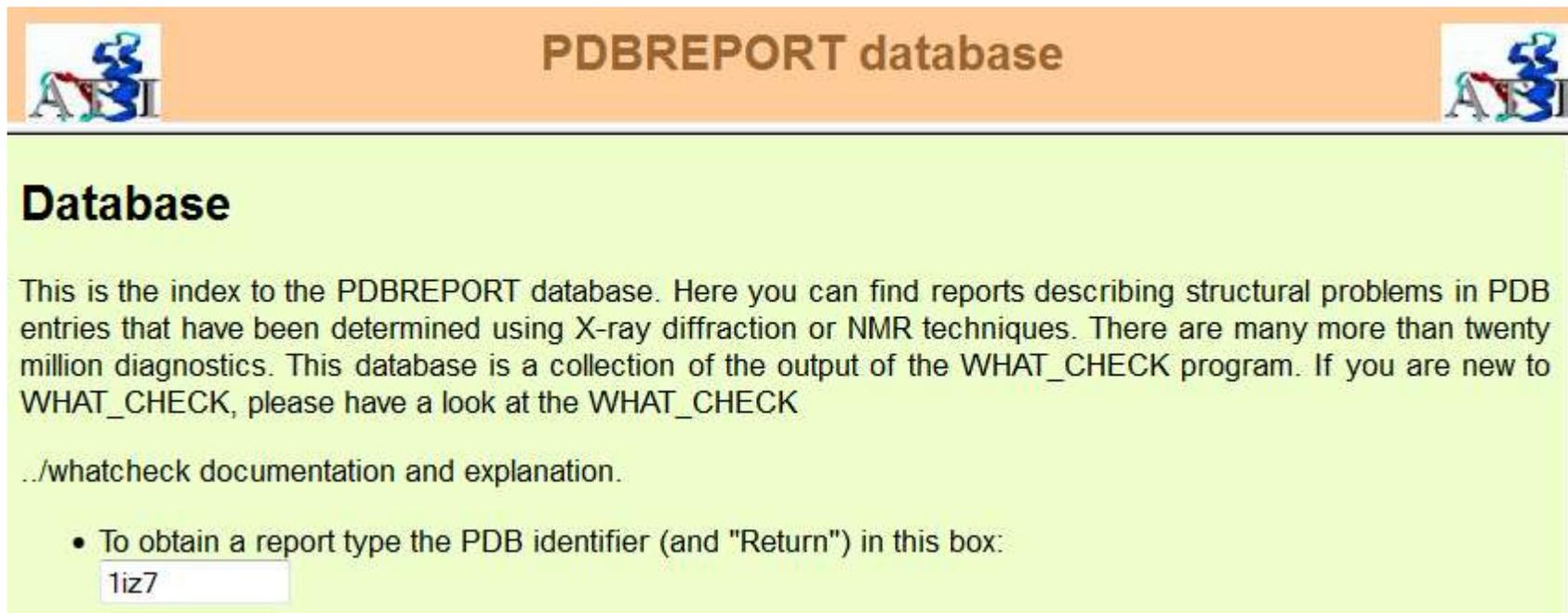
PROCHECK analyses for 1iz7

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



**PDBREPORT database**

### Database

This is the index to the PDBREPORT database. Here you can find reports describing structural problems in PDB entries that have been determined using X-ray diffraction or NMR techniques. There are many more than twenty million diagnostics. This database is a collection of the output of the WHAT\_CHECK program. If you are new to WHAT\_CHECK, please have a look at the WHAT\_CHECK

../whatcheck documentation and explanation.

- To obtain a report type the PDB identifier (and "Return") in this box:

# 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

Summary Sequence Annotations Seq. Similarity 3D Similarity Literature Biol. & Chem. Methods **Geometry** Links

**Re-refinement of the structure of hydrolytic haloalkane dehalogenase linb from sphingomonas paucimobilis UT26 AT 1.6 A resolution**

**1IZ7**

Display Files ▾  
Download Files ▾  
Share this Page ▾

Geometry: Structure Variance Analysis Results

**RCSB Graphics**

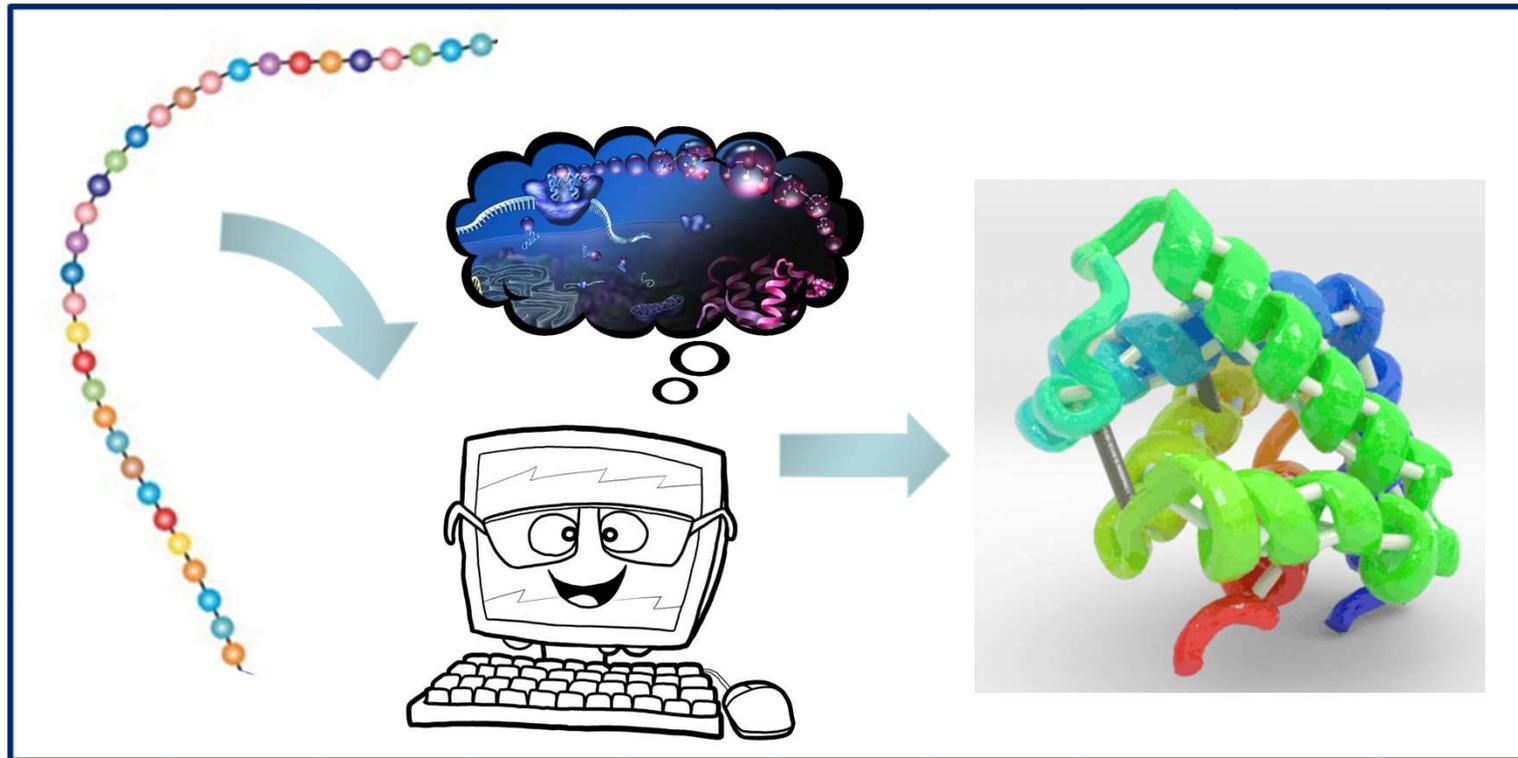
Chain Id	B factor	Omega	FDS Summary
A	Plot    Summary    3D	Plot    3D	Plot    Summary    3D

\*Note: FDS (fold deviation score) is defined as a multiple of the standard deviation for a specific reference value.

**MolProbity Ramachandran Plot**

Click here to download the MolProbity Ramachandran Plot.

# 3D structure prediction



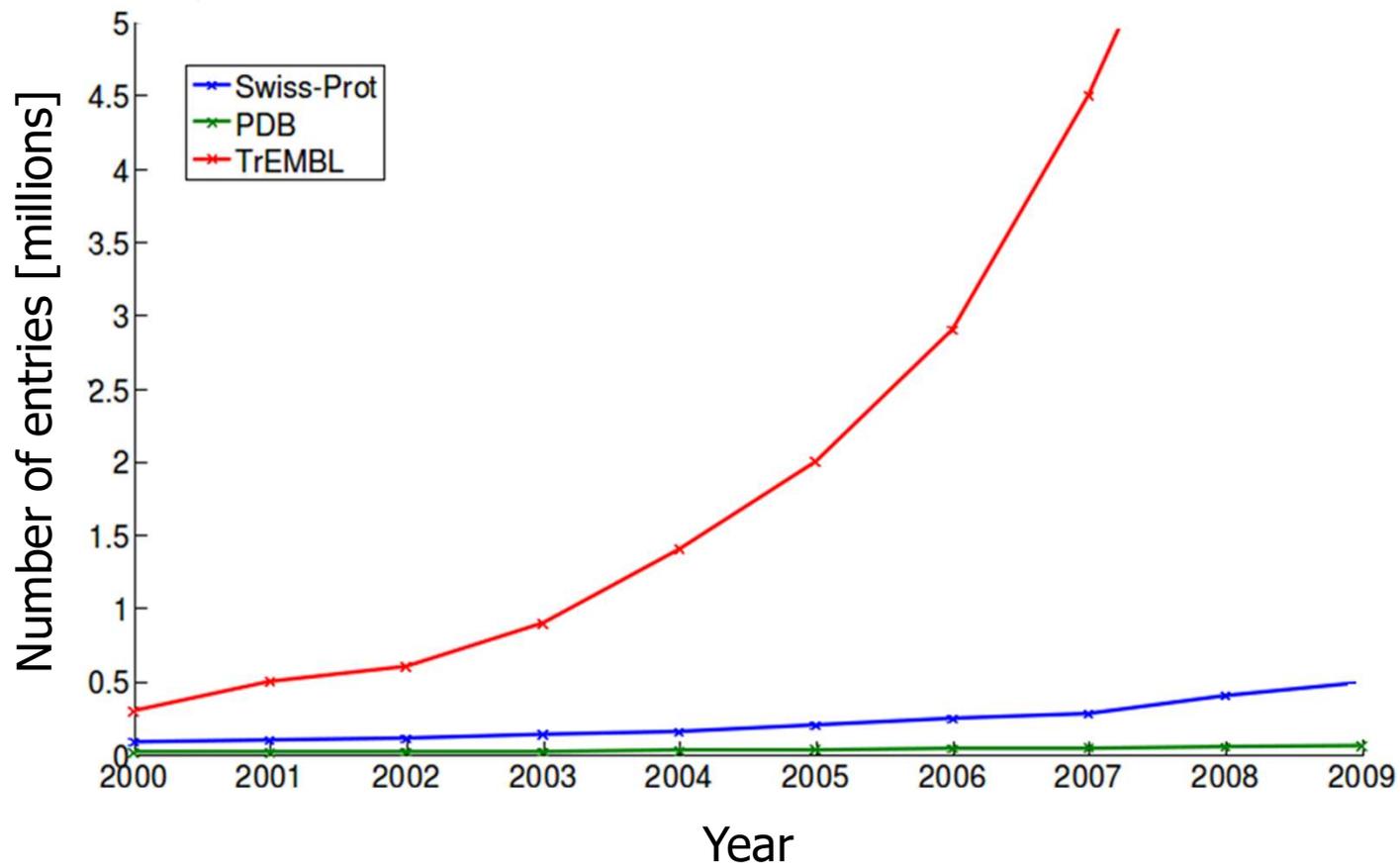
# 3D structure prediction

- ❑ homology modeling
- ❑ fold recognition
- ❑ *ab initio* prediction
- ❑ “hybrid” approaches
- ❑ Assessment
- ❑ databases of protein models

# Importance of structure



- no experimental structure for most of the sequences



# Homology modelling

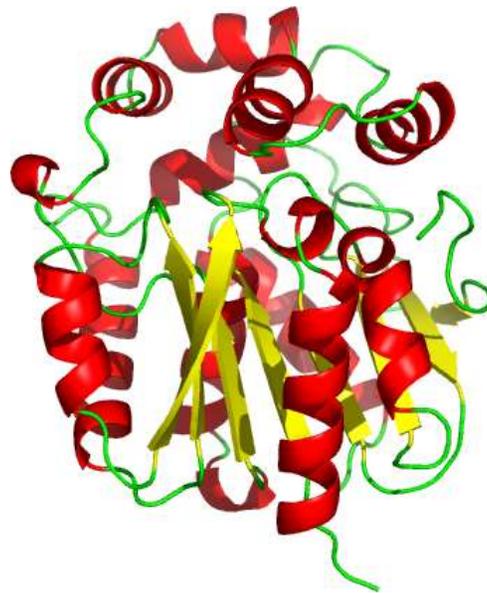


- ❑ basic principle – structure is more conserved than sequence

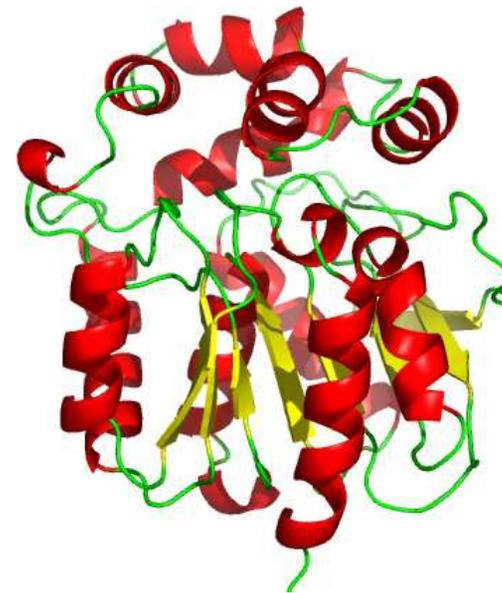
# Homology modeling



- basic principle – structure is more conserved than sequence
  - similar sequences adopt practically identical structures



haloalkane dehalogenase  
LinB (PDB-ID 1iz7)



haloalkane dehalogenase  
DhaA (PDB-ID 1cqW)

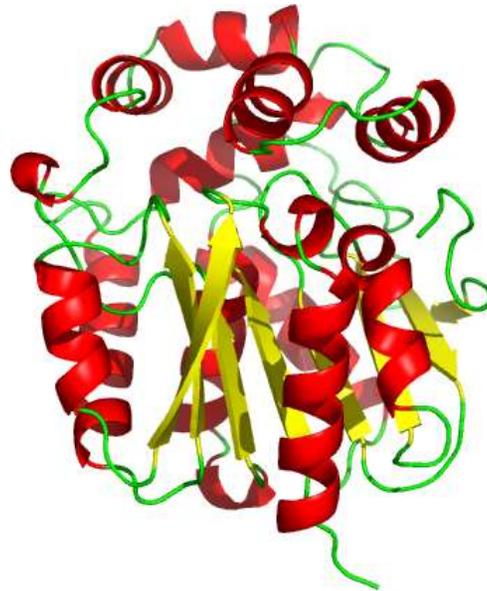
sequence identity: ~ 50 %

Structure prediction → protein structure prediction → 3D structure prediction → homology modeling

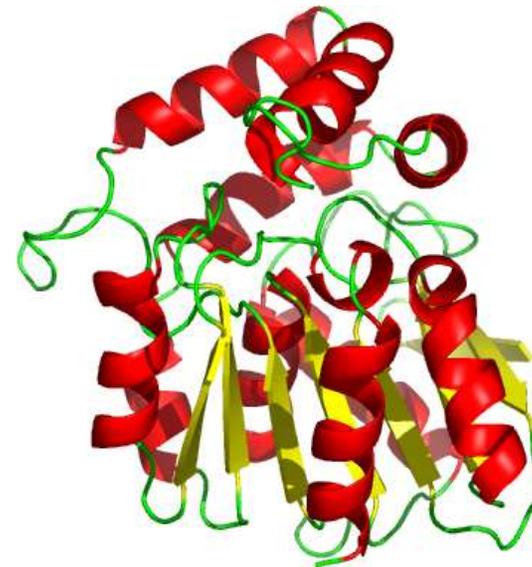
# Homology modeling



- basic principle – structure is more conserved than sequence
  - distantly related sequences still fold into similar structures



haloalkane dehalogenase  
LinB (PDB-ID 1iz7)



chloroperoxidase L  
(PDB-ID 1a88)

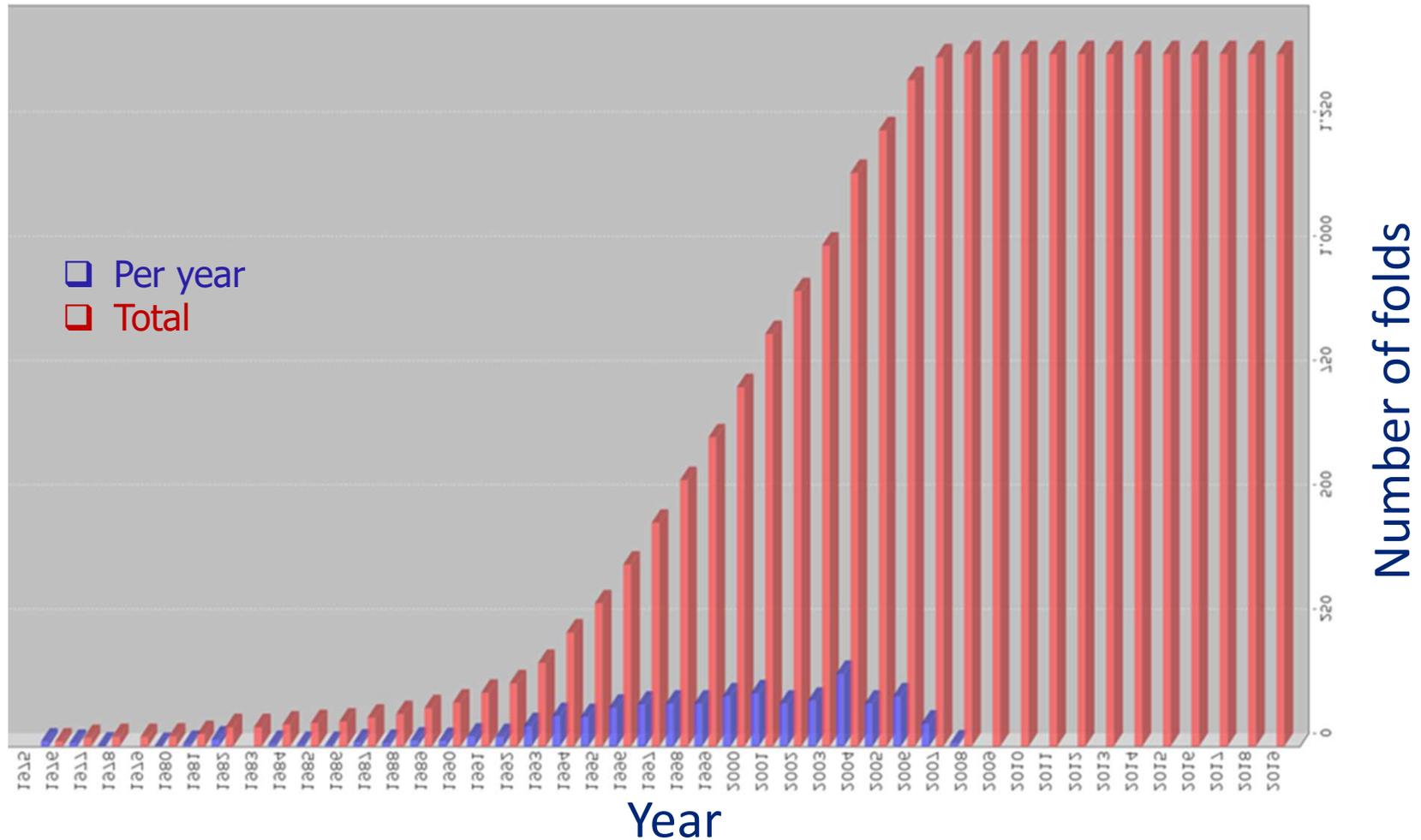
sequence identity: ~ 15 %

Structure prediction → protein structure prediction → 3D structure prediction → homology modeling

# Homology modeling



- number of folds in SCOP database



Structure prediction → protein structure prediction → 3D structure prediction → homology modeling

# Homology modeling



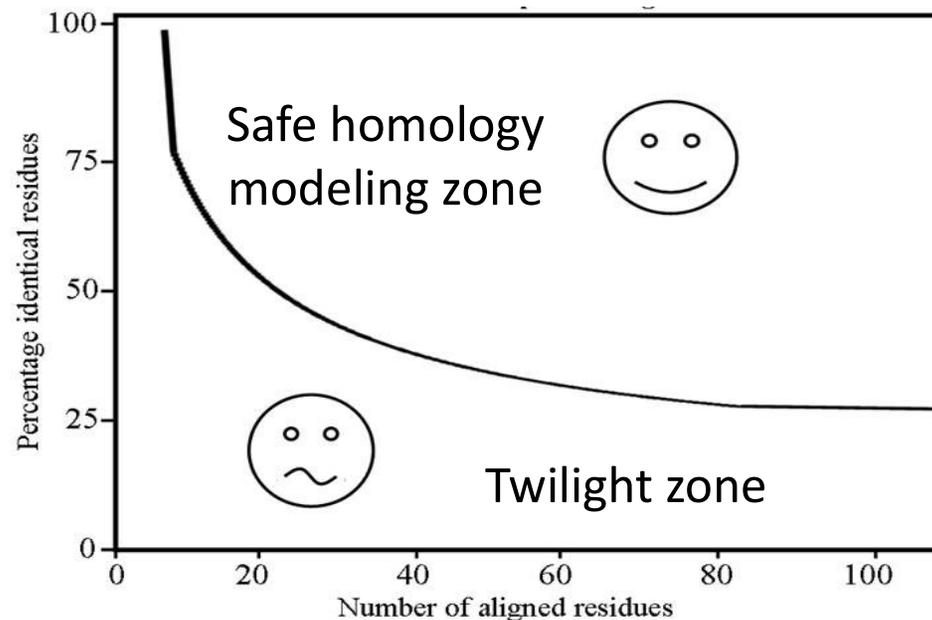
- ❑ basic principle – structure is more conserved than sequence
  - similar sequences adopt practically identical structures
  - distantly related sequences still fold into similar structures
- ❑ builds an atomic-resolution model of the target protein  
**based on the experimental 3D structure** (template) of a homologous protein
- ❑ the **most accurate** 3D prediction approach
- ❑ if no reliable template is available → fold recognition or *ab initio* prediction

Structure prediction → protein structure prediction → 3D structure prediction → homology modeling

# Homology modeling



- the quality of the model depends on the **sequence identity** / **similarity** between the **target and template** proteins
- For a **standard length protein** it should be **> 25%** / **> 40%**



Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999 Feb;12(2):85-94. doi: 10.1093/protein/12.2.85. PMID: 10195279.

Structure prediction → protein structure prediction → 3D structure prediction → homology modeling

# Homology modelling – steps



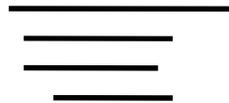
...MSLGAKPFGE...

**target  
sequence**

# Homology modelling – steps



...MSLGAKPFGE...



target  
sequence

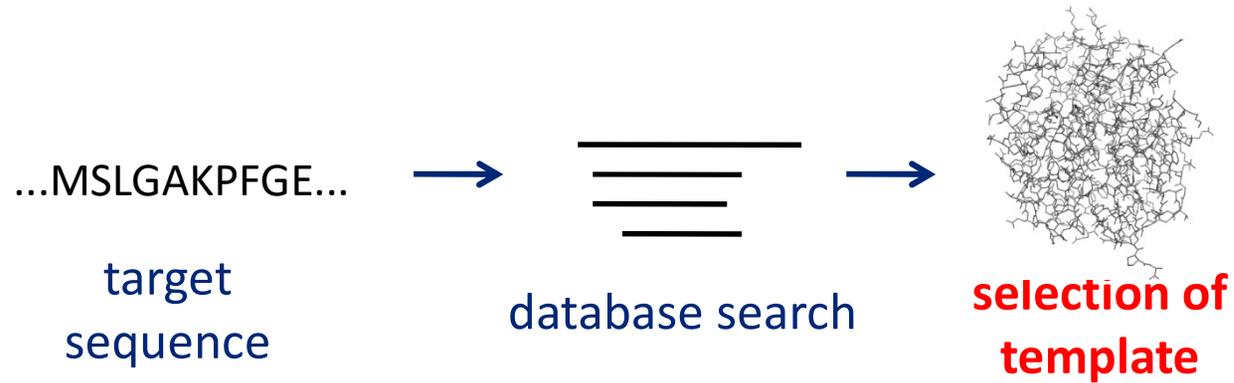
**database search**

# Database search



- ❑ standard **sequence-similarity** searches
  - comparison of the target sequence to all sequences with known 3D structures in the wwPDB database
  - BLAST, FASTA,...
- ❑ **profile-based** searches
  - more sensitive than standard sequence-similarity searches
  - PSI-BLAST, HHMER, HHblits, ...
- ❑ **fold recognition** methods
  - applied if no template can reliably be identified by the sequence or profile based methods (sequence identity < recommended 25 %)
  - FUGUE, GenTHREADER, pro-sp3-TASSER..

# Homology modelling – steps

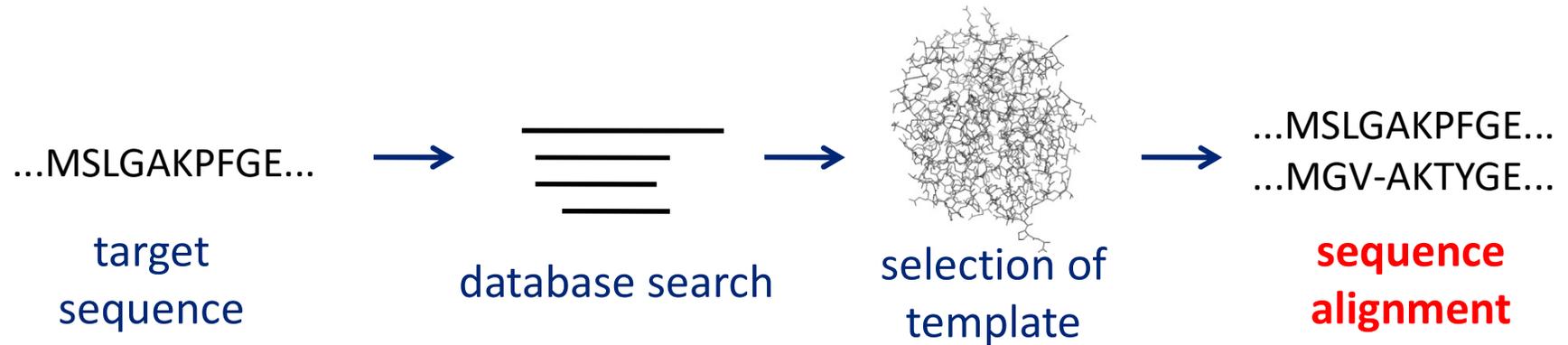


# Selection of template



- ❑ wrong template = wrong model
- ❑ more than one possible template may be identified → a combination of different criteria to select the final template:
  - sequence identity between the template and target protein
  - coverage between the template and query sequences
  - the resolution of the template structure, number of errors
  - a portion of conserved residues in the region of interest (e.g., binding site residues)
  - ...
- ❑ multiple templates can be used to create a combined model

# Homology modelling – steps



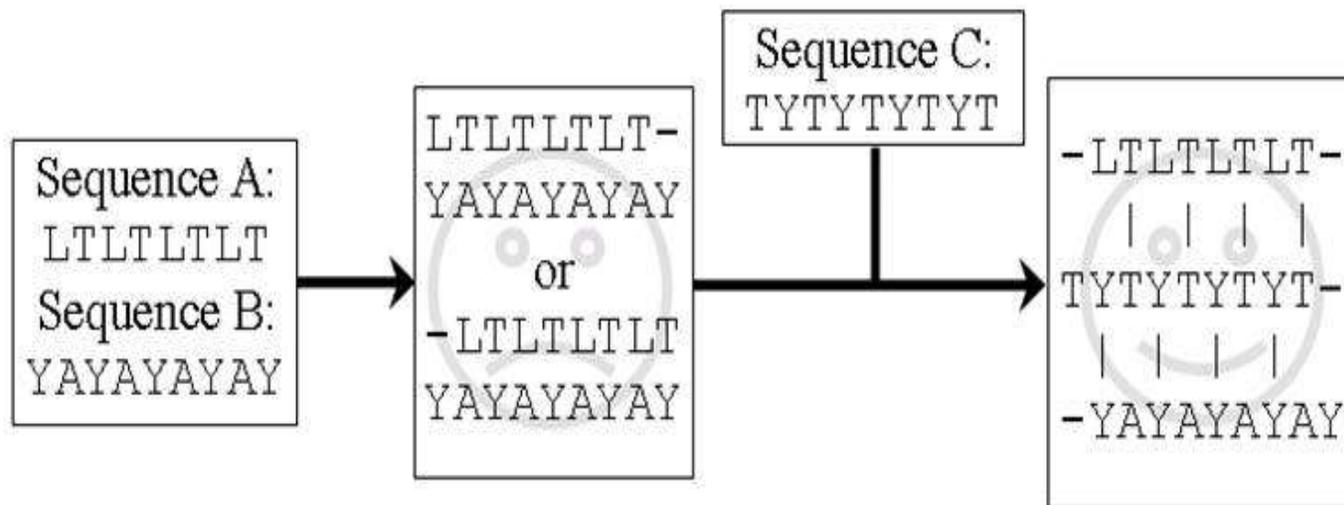
# Sequence alignments



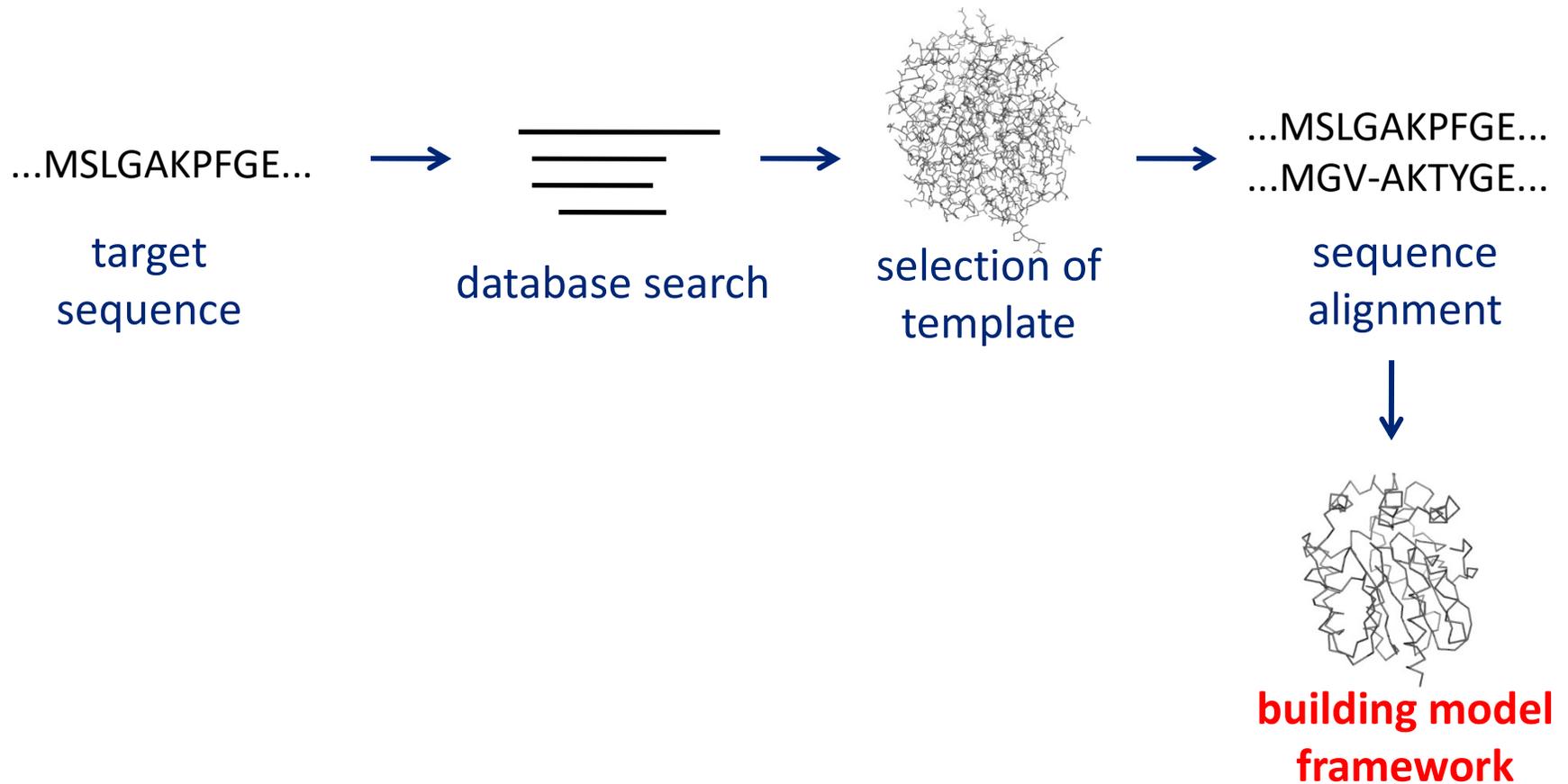
- ❑ reliability of alignment decreases with decreasing similarity of the target and template sequences
- ❑ quality of **alignment is crucial** – it determines the quality of the final model
- ❑ the pairwise target-template alignment provided by the database search methods is almost guaranteed to contain errors → more sophisticated methods needed
  - **multiple sequence alignment**
  - **Profile-driven alignments**
  - correction of alignment based on the template structure

# Sequence alignments

- multiple sequence alignment
  - works with **more information than pairwise alignment** → more reliable
  - MUSCLE, CLUSTAL Omega, T-Coffee



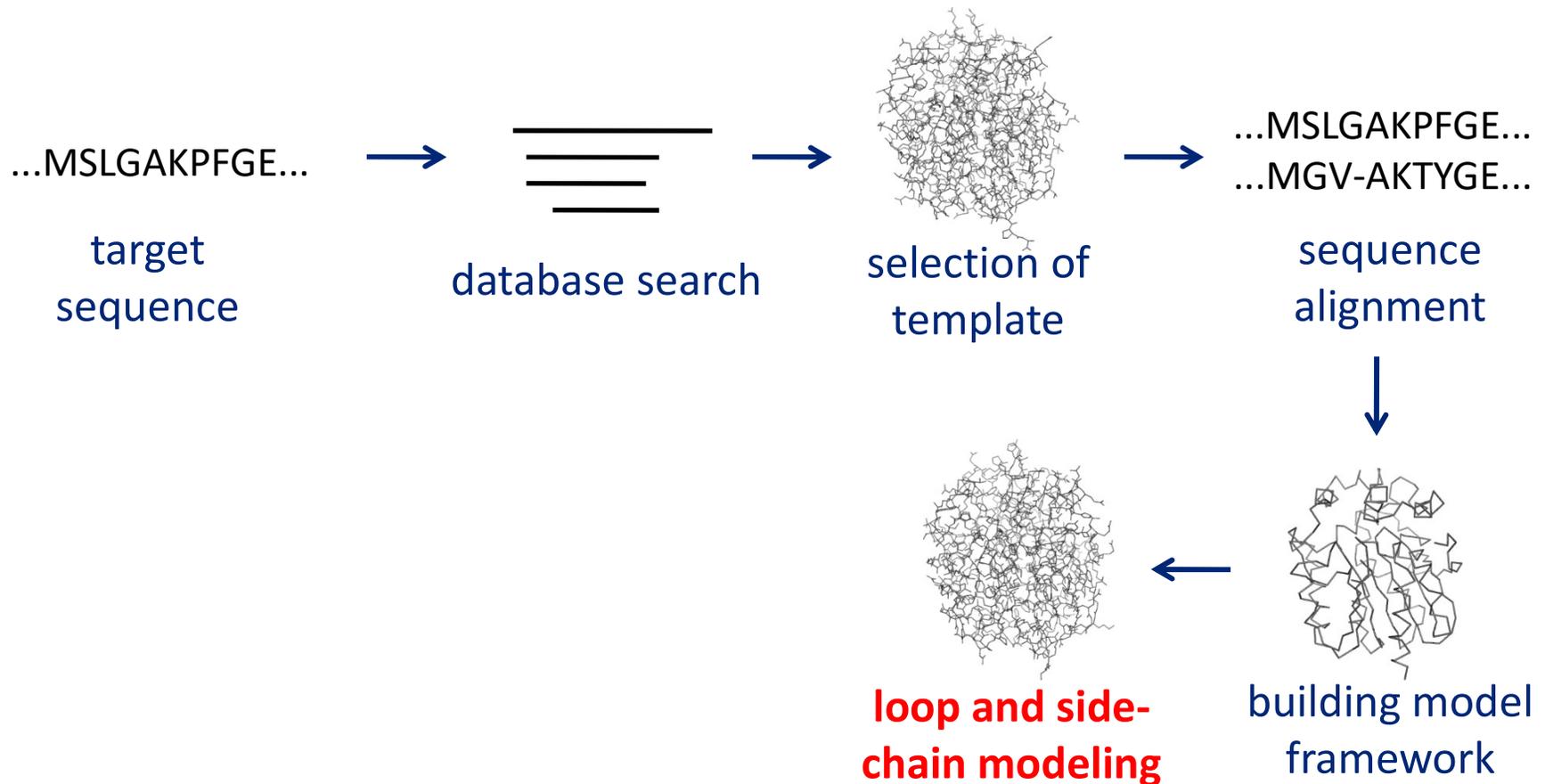
# Homology modelling – steps



# Building model framework

- **copying the basic shape** of the template to the model
  - if the two aligned residues differ, the backbone coordinates for N, C $\alpha$ , C and O, and often also C $\beta$  can be copied
  - conserved residues can be copied completely to provide an initial guess
  - residues that are not present in the target (because the target can have less residues than the template) are not copied

# Homology modelling – steps



# Loop modelling

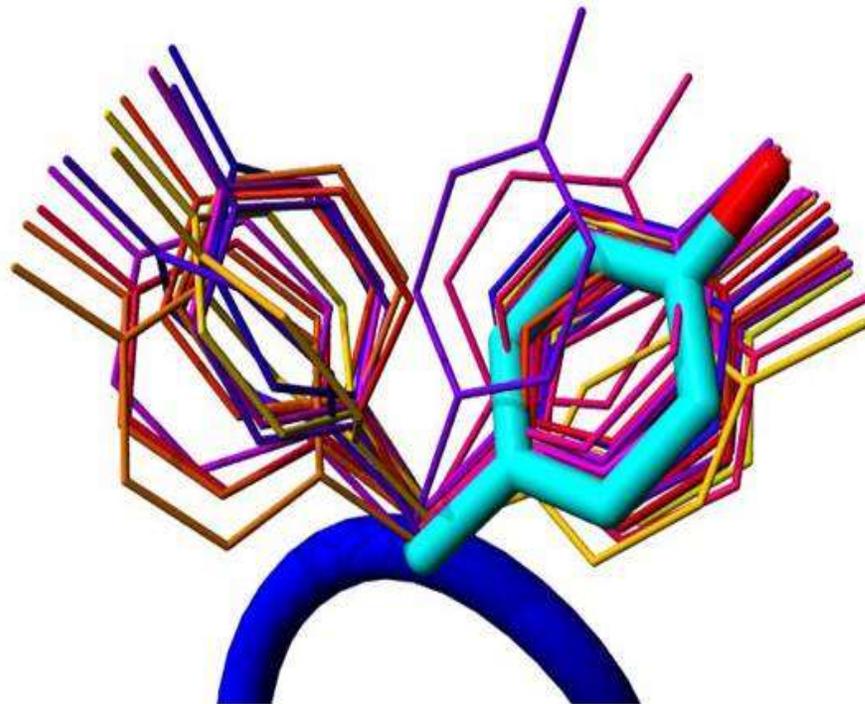
- ❑ inserting missing residues into the continuous backbone
- ❑ prediction of loop conformation is a **difficult task** (especially for loops > 5-8 residues long)
  - **knowledge based** prediction – use of libraries of possible loop conformations known from experimentally determined structures with the same local sequence
  - **ab initio** prediction – use of energy functions to find the most optimal conformation, followed by minimization of the structure
  - **hybrid** approach – the loop is divided into small fragments that are all separately compared to known structures

# Side-chain modelling

- ❑ adding side-chains of amino acids to the model backbone
  - ❑ **rotamer libraries** – common side-chain conformations (**rotamers**) extracted from high-resolution X-ray structures → possible rotamers explored and scored based on energy function
  - ❑ **backbone-dependent rotamer libraries** – the optimal conformation of the side chain depends on the local backbone conformation (5 - 9 neighboring residues) → explored only possible rotamers corresponding to the best backbone matches – greatly reduces conformational search space

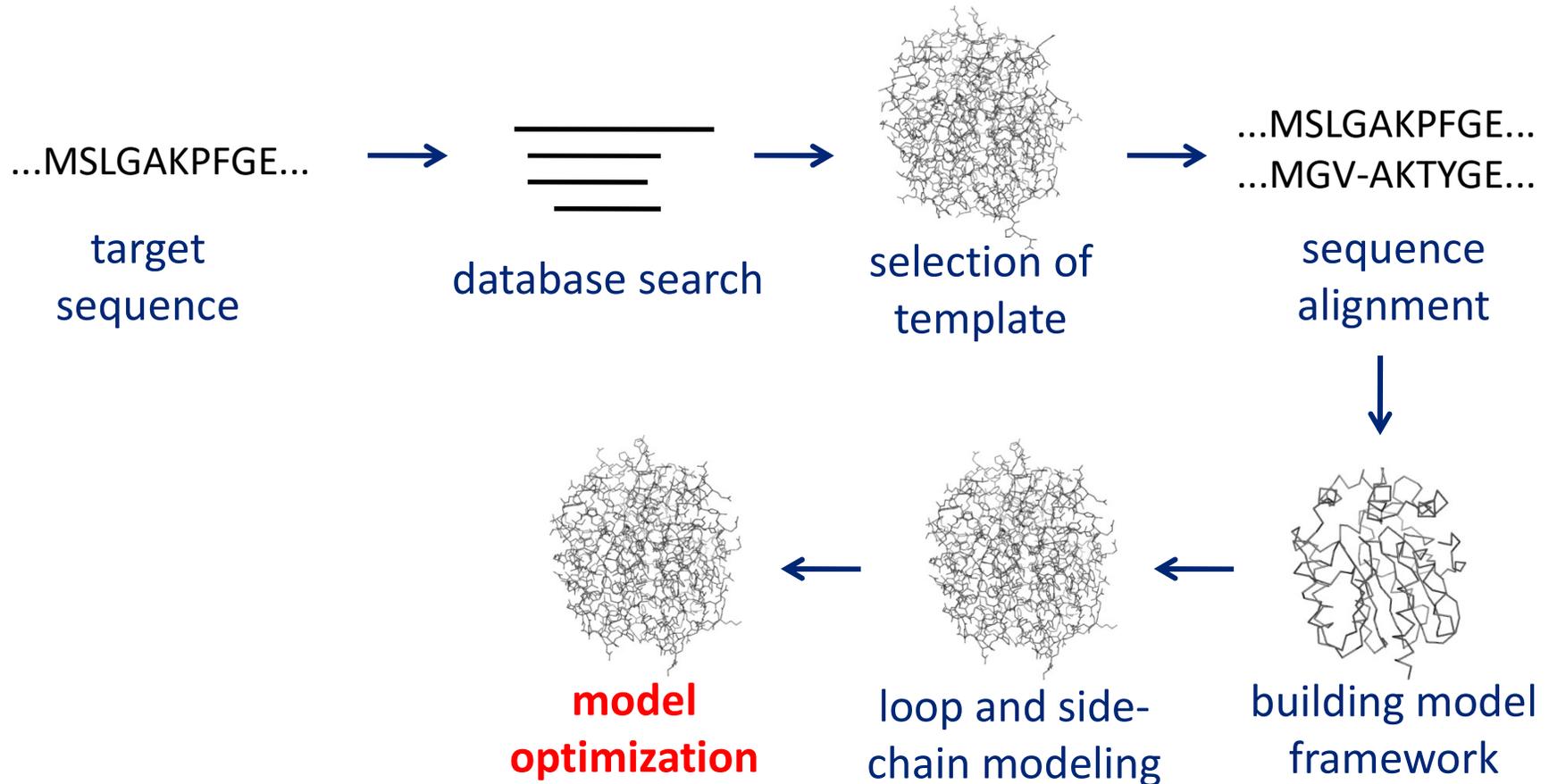
# Side-chain modelling

- backbone-dependent rotamer library



According to the backbone-dependent rotamer library, the backbone favors two different conformations for Tyrosine which appear about equally often in the database

# Homology modelling – steps

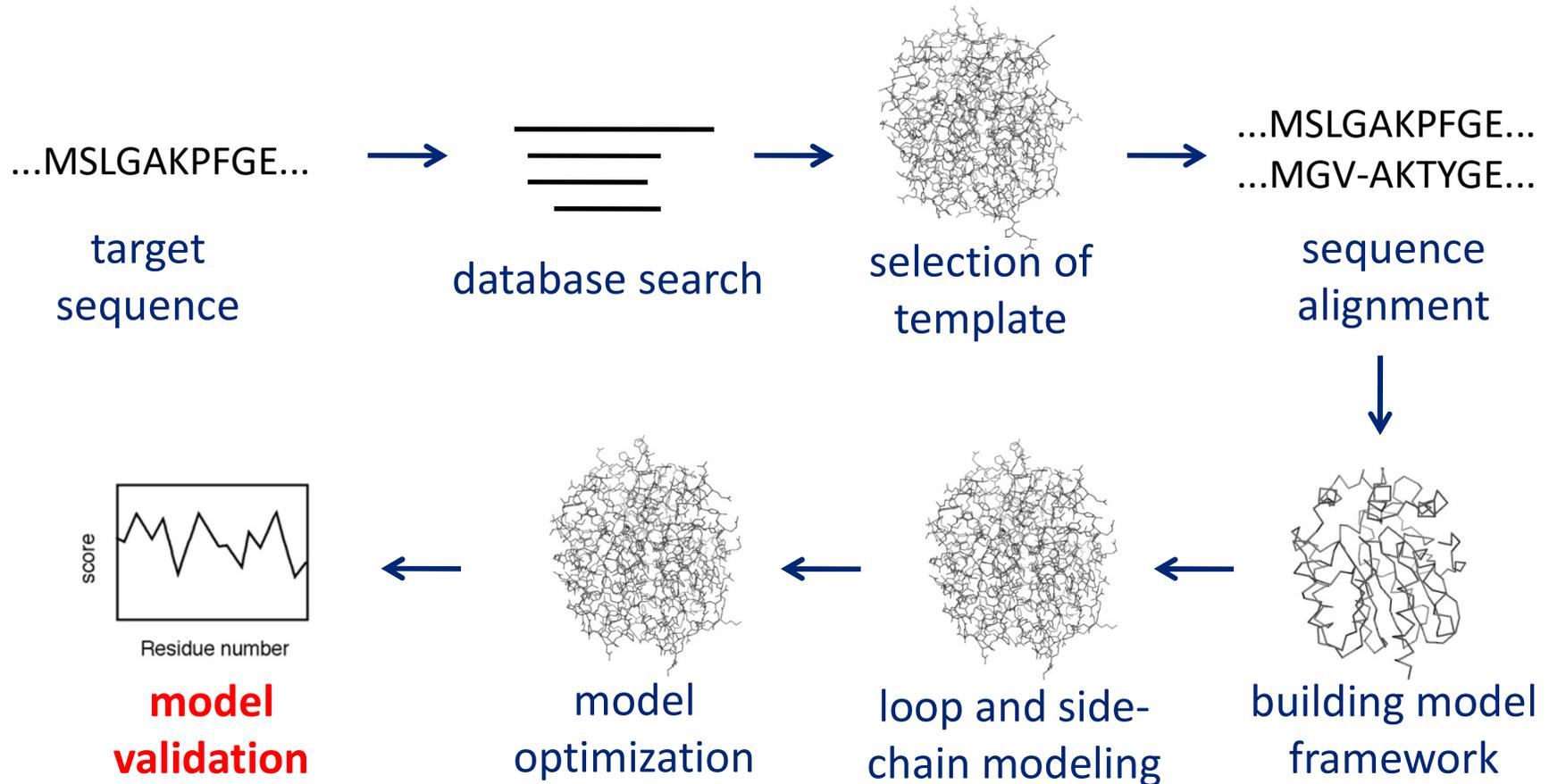


# Model optimization



- ❑ energy minimization – **may introduce many errors** moving the model away from its correct structure → must be used carefully
- ❑ **molecular dynamics** simulation – follows the motions of the protein and mimics the folding process

# Homology modelling – steps

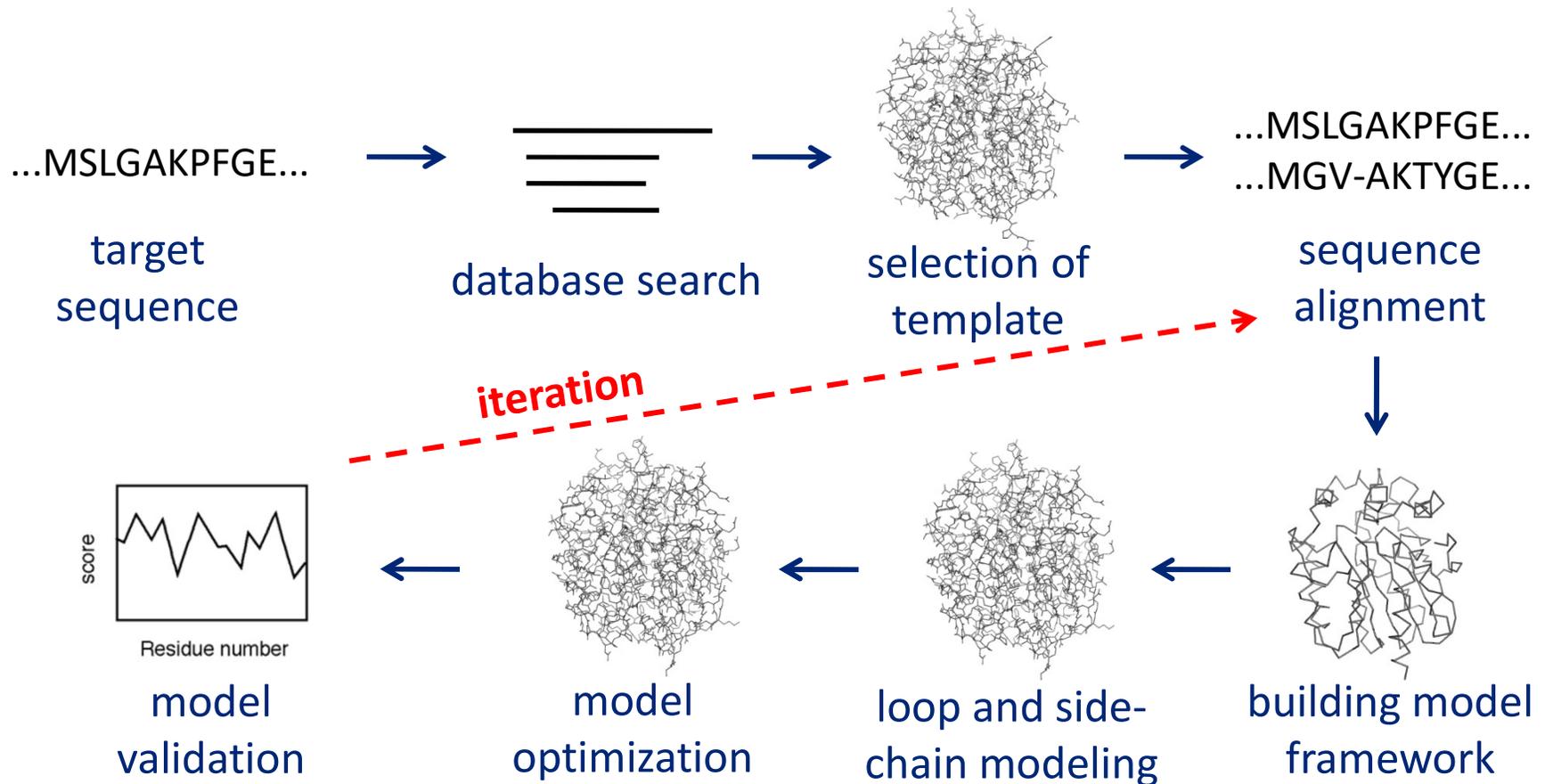


# Model validation

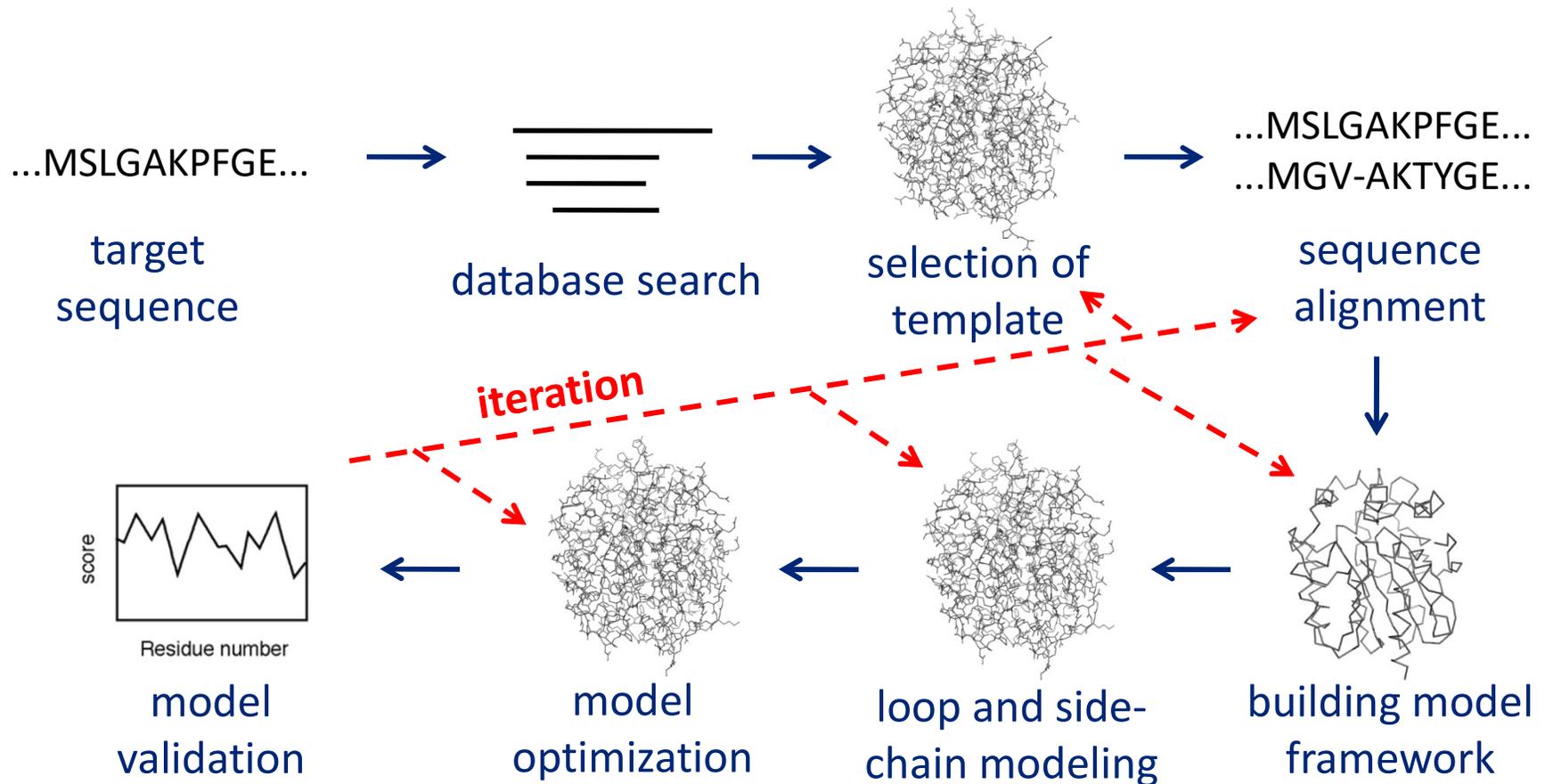


- ❑ finished **model contain errors** (like any other structure) – the number of errors (for a given method) mainly depends on:
  - ❑ the percentage of **sequence identity** between **template and target** sequence, e.g., 90 %: the accuracy of the model comparable to X-ray structures; 50 %-90 %: larger local errors; identity < 25 %: often very large errors
  - ❑ the number of **errors in the template** structure
- ❑ problems that occur far from the site of interest may be ignored, others should be tackled

# Homology modelling – steps



# Homology modelling – steps



# Iteration



- portions of the homology modeling process can be iterated to **correct identified errors**
  - small errors introduced during the optimization → running a shorter molecular dynamics simulation
  - error in a loop → choosing another loop conformation in the loop modeling step
  - large mistakes in the backbone conformation → repeating the whole process with another alignment or even different template
  - ...

# Homology modeling programs



## □ MODELLER

- <http://salilab.org/modeller/>
- models built by **satisfying the spatial restraints** of the C  $\alpha$  - C  $\alpha$  bond lengths and angles, the dihedral angles of the side-chains, and van der Waals interactions
- restraints calculated from the template structures
- available as a web server at different sites, e.g., part of: ModWeb workflow <https://modbase.compbio.ucsf.edu/modweb/>, GeneSilico server <https://genesilico.pl/toolkit/unimod?method=Modeller> or Bioinformatics toolkit <http://toolkit.lmb.uni-muenchen.de/modeller>

# Homology modeling programs



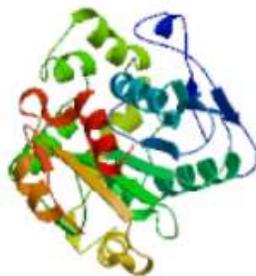
## ❑ SWISS-MODEL

- <http://swissmodel.expasy.org/>
- fully automated protein structure homology modeling server



Print/Save this page

### Model Summary ?



#### Model information:

Modelled residue range: 1 to 297  
Based on template: [2xt0A] (1.90 Å)  
**Remark:** No search for template was performed.  
Only user specified template was used for modelling.  
Sequence Identity [%]: 40.33  
Evalue: 0.00e-1

#### Quality information: [details] ▶

QMEAN Z-Score: -2.61



#### Quaternary structure information: [details] ▶

Template (2xt0): MONOMER  
Model built: SINGLE CHAIN

#### Ligand information: [details] ▶

Ligands in the template: SO4: 2.  
Ligands in the model: none.

**logs:** [Templates] ▶ [Alignment] ▶ [Modelling] ▶

**display model:** as [pdb] ▶ - as [DeepView project] ▶ - in [AstexViewer] ▶

**download model:** as [pdb] ⚡ - as [Deepview project] ⚡ - as [text] ⚡

# Model validation



- ❑ mostly the **same principles** as used for the validation of experimental structures
- ❑ **always check both model and template**
  - The model cannot improve the template if this is “bad” in regions
- ❑ **checks of normality**
  - inside/outside distributions of polar and apolar residues
  - bad contacts
  - evaluation of atom/residue environment
- ❑ **energy-based checks**
  - side-chain clashes
  - bond lengths and angles

# Model validation programs



## QMEAN

- <https://swissmodel.expasy.org/qmean/>
- composite scoring function for the **quality estimation of protein structure models**; evaluates torsion angles, solvation and non-bonded interactions and the agreement between predicted and calculated secondary structure and solvent accessibility

Global scores			Local scores		
Model name_🔗	QMEAN score_🔗	Estimated absolute quality_🔗 <b>NEW</b>	Z-scores of QMEAN terms_🔗 <b>NEW</b>	Residue error_🔗 	Residue error plot_🔗
modbase-model_6d51f947356cc91f0e1be73c6d7e11d2.pdb	0.705	 Z-score=-0.74 [plot 1] [plot 2]	 [png]	 [jpg] [pdb] Jmol	 [png] [table]

# Model validation programs



- ❑ Verify3D
- ❑ ANOLEA
- ❑ PROCHECK
- ❑ WHATCHECK
- ❑ PROSA II
- ❑ ...

# Fold recognition (Threading)

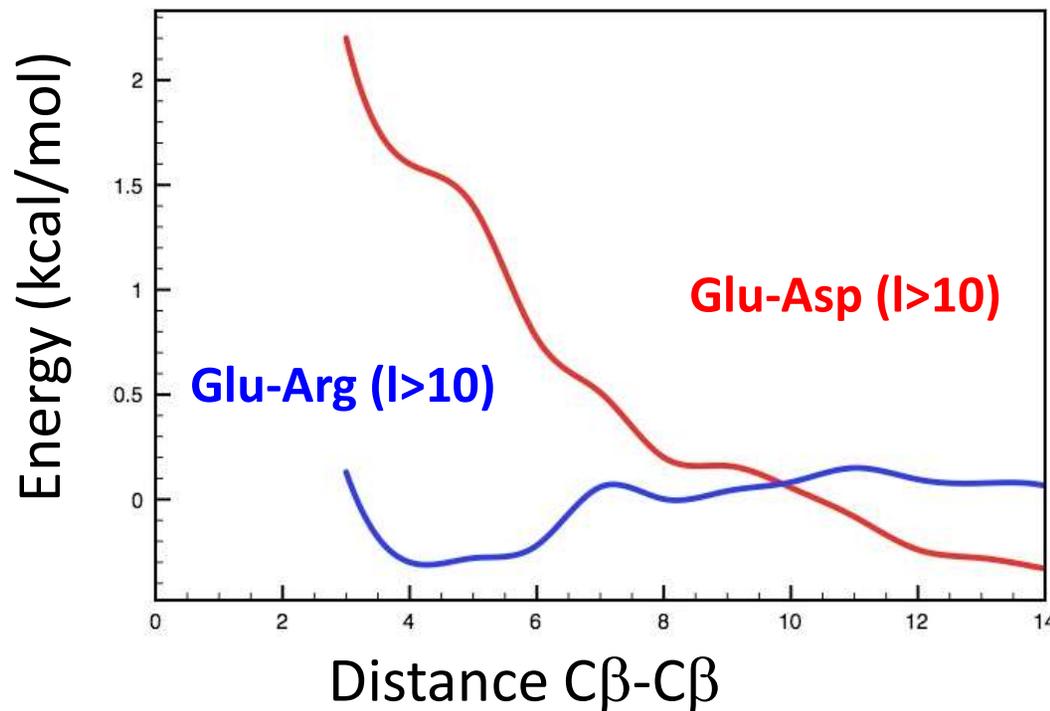


- ❑ predicts the fold of a protein by fitting its sequence into a structural database and selecting the **best fitting fold**
- ❑ provides a rough approximation of the overall topology of the native structure → does **not** generate fully refined **atomic models** for the query sequence
- ❑ can be used when no suitable template structures available for homology modeling
- ❑ **fails** if the correct **protein fold does not exist** in the database
- ❑ high rates of false positives

# Fold recognition (Threading)



- pairwise energy-based methods (threading) – protein sequence is searched for in a structural database to find the best matching structural fold using energy-based criteria



$l$  is distance in sequence (density normalization required)

can be calculated from collections of known structures

# Fold recognition (Threading)

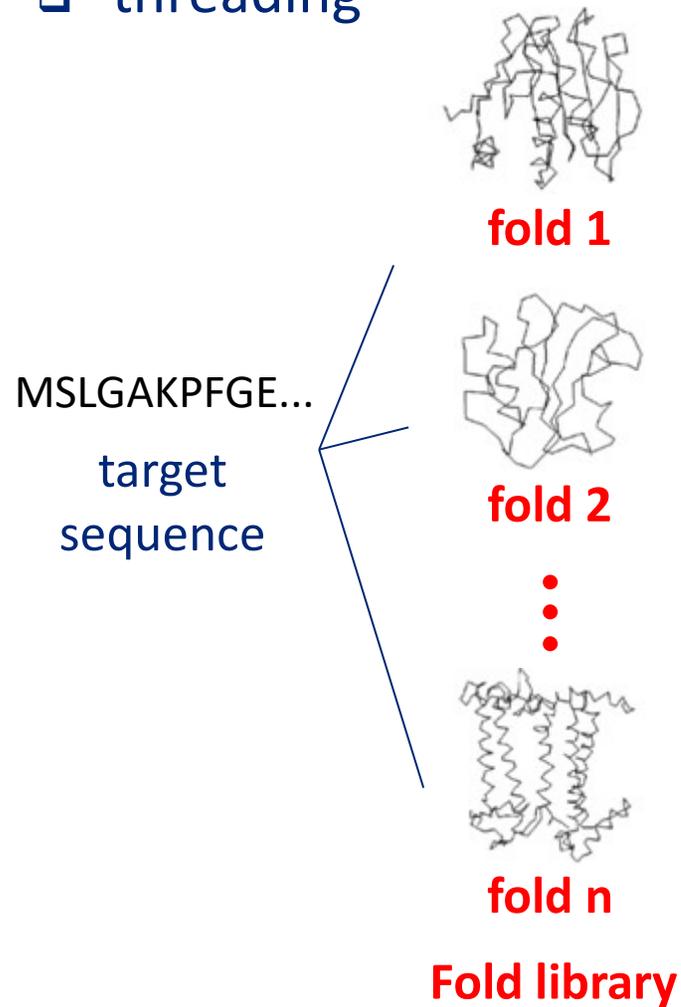
- threading

MSLGAKPFGE...

**target  
sequence**

# Fold recognition (Threading)

- threading



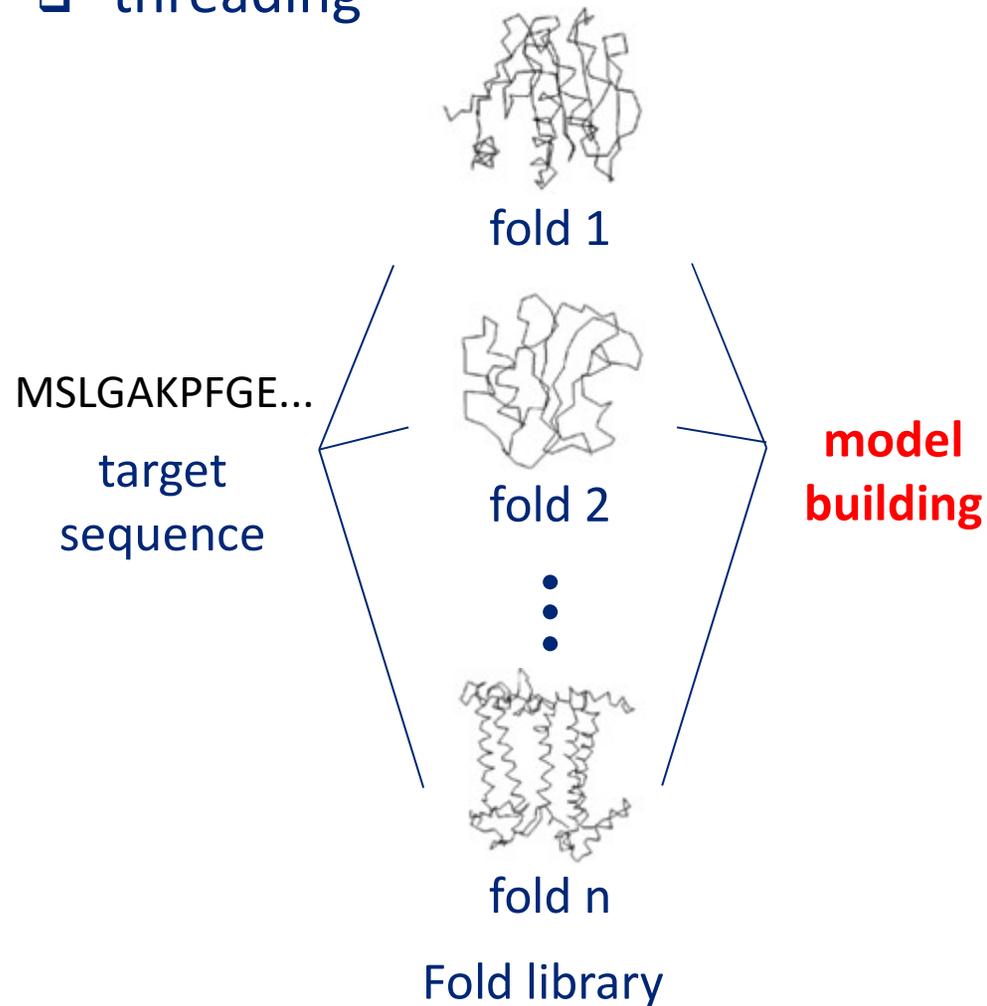
# Fold recognition (Threading)



- ❑ **pairwise energy-based** methods (threading) – protein sequence is searched for in a structural database to find the best matching structural fold using **energy-based criteria**
  1. **alignment** of the query sequence with each structural fold in the fold library (essentially performed at the sequence profile level)

# Fold recognition (Threading)

- threading



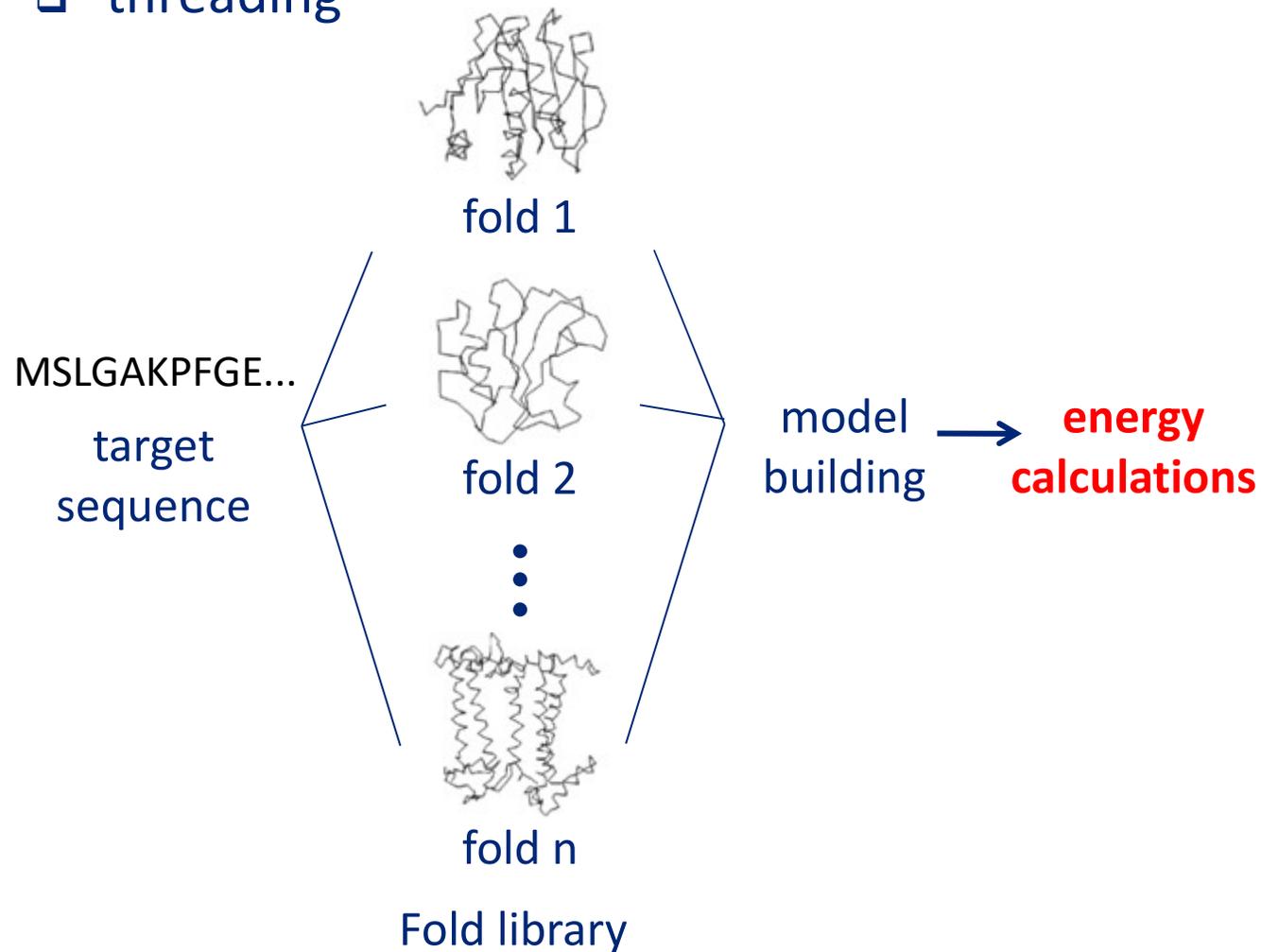
# Fold recognition (Threading)



- ❑ **pairwise energy-based** methods (threading) – protein sequence is searched for in a structural database to find the best matching structural fold using **energy-based criteria**
  1. **alignment** of the query sequence with each structural fold in the fold library (essentially performed at the sequence profile level)
  2. building a **crude model** for the target sequence (replacing aligned residues in the template structure with the corresponding residues in the query)

# Fold recognition (Threading)

## □ threading



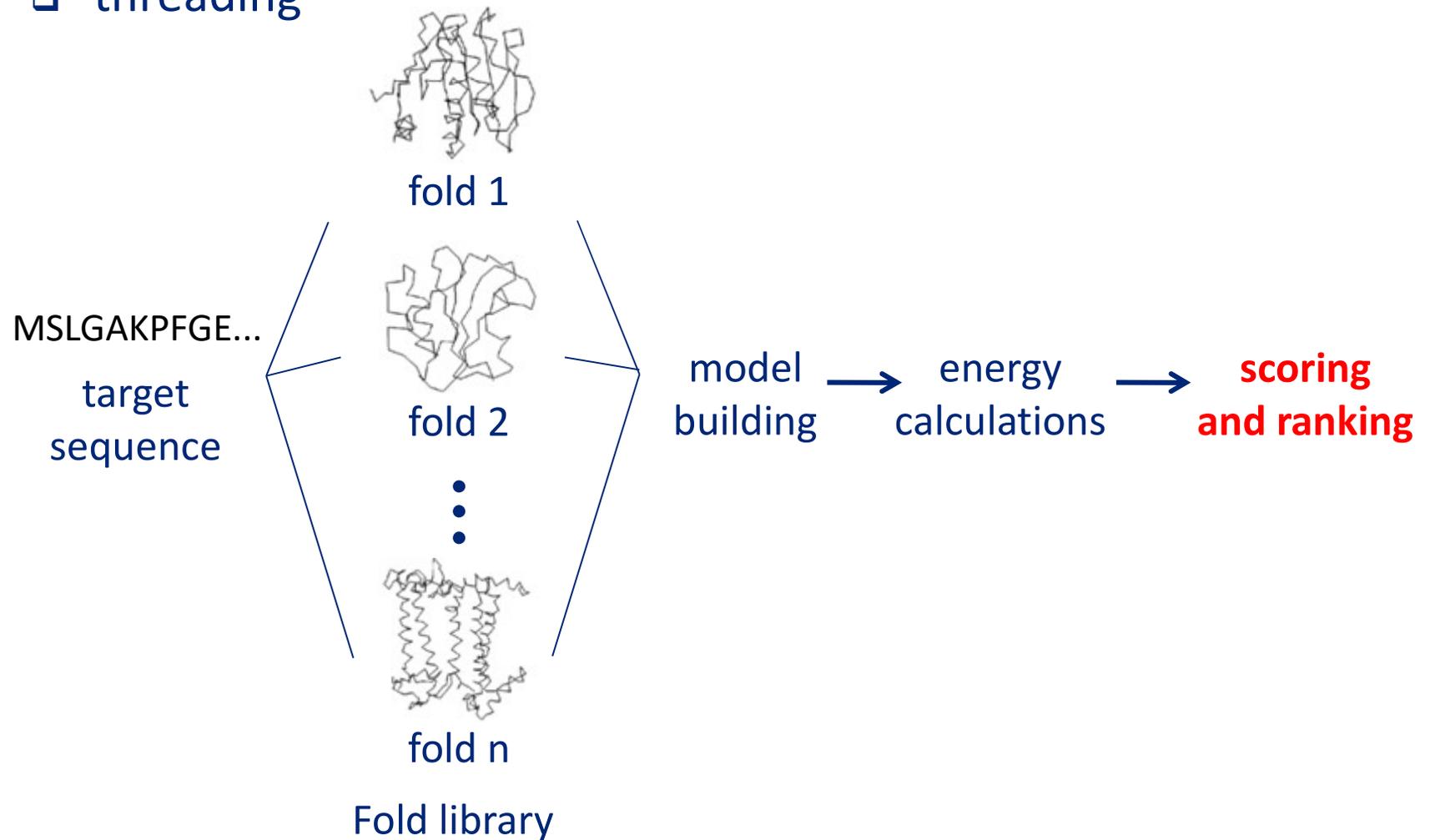
# Fold recognition (Threading)



- ❑ **pairwise energy-based** methods (threading) – protein sequence is searched for in a structural database to find the best matching structural fold using **energy-based criteria**
  1. **alignment** of the query sequence with each structural fold in the fold library (essentially performed at the sequence profile level)
  2. building a **crude model** for the target sequence (replacing aligned residues in the template structure with the corresponding residues in the query)
  3. calculating **energy of the raw model**

# Fold recognition (Threading)

## □ threading



# Fold recognition (Threading)



- ❑ **pairwise energy-based** methods (threading) – protein sequence is searched for in a structural database to find the best matching structural fold using **energy-based criteria**
  1. **alignment** of the query sequence with each structural fold in the fold library (essentially performed at the sequence profile level)
  2. building a **crude model** for the target sequence (replacing aligned residues in the template structure with the corresponding residues in the query)
  3. calculating **energy of the raw model**
  4. **ranking** of the models based on the energetics – the lowest energy fold represents the structurally most compatible fold

# Fold recognition (Profiles)

- profile methods

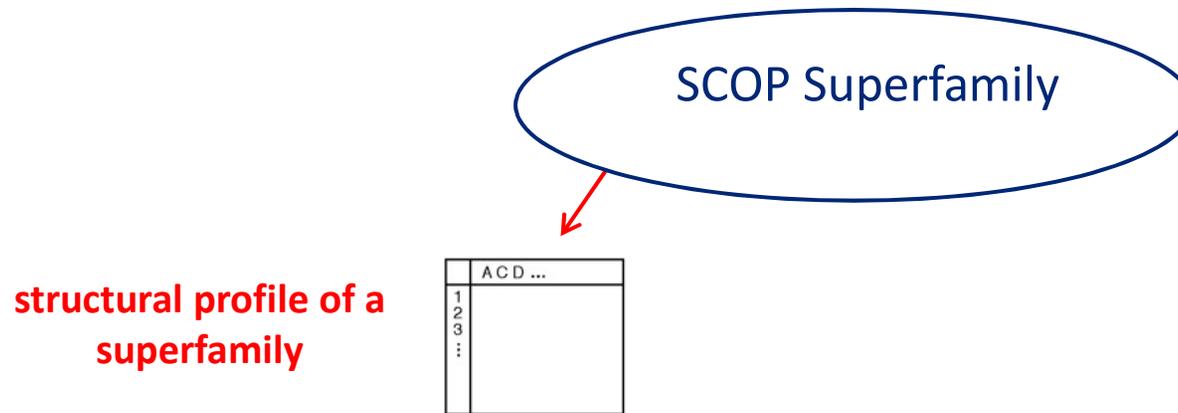
# Fold recognition (Profiles)

- profile methods

**SCOP Superfamily  
(one of many)**

# Fold recognition (Profiles)

- profile methods



# Fold recognition (Profiles)

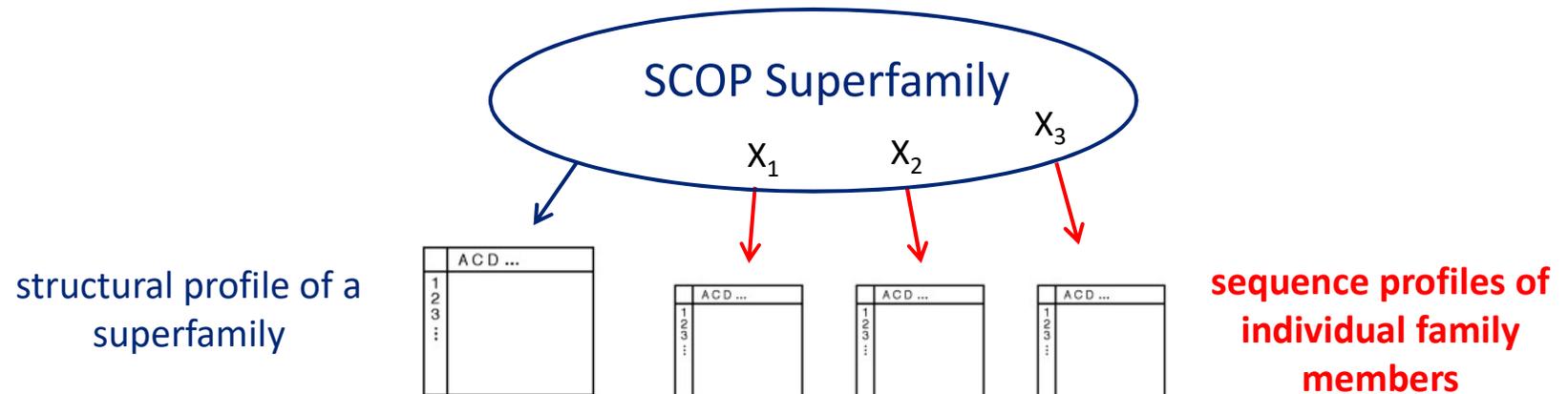


## □ profile methods

1. construction of **profile(s)** for a group of **related protein structures** (e.g., for each SCOP superfamily) – scores describing the propensity of each residue to be at each profile position, information for secondary structural types, solvent accessibility, polarity, sequence-based profiles, ...

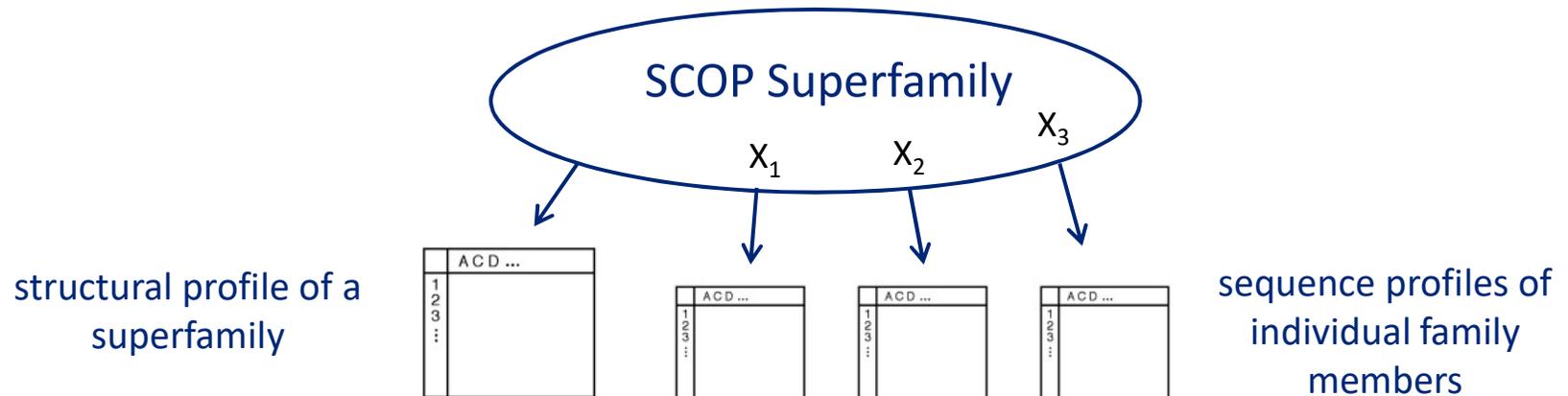
# Fold recognition (Profiles)

- profile methods



# Fold recognition (Profiles)

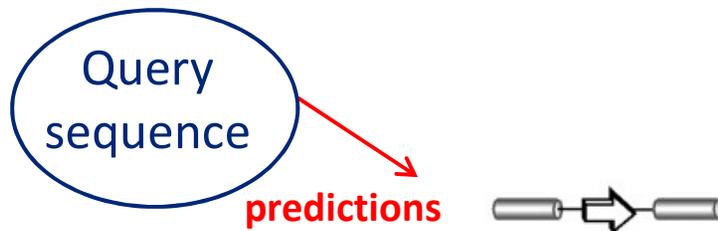
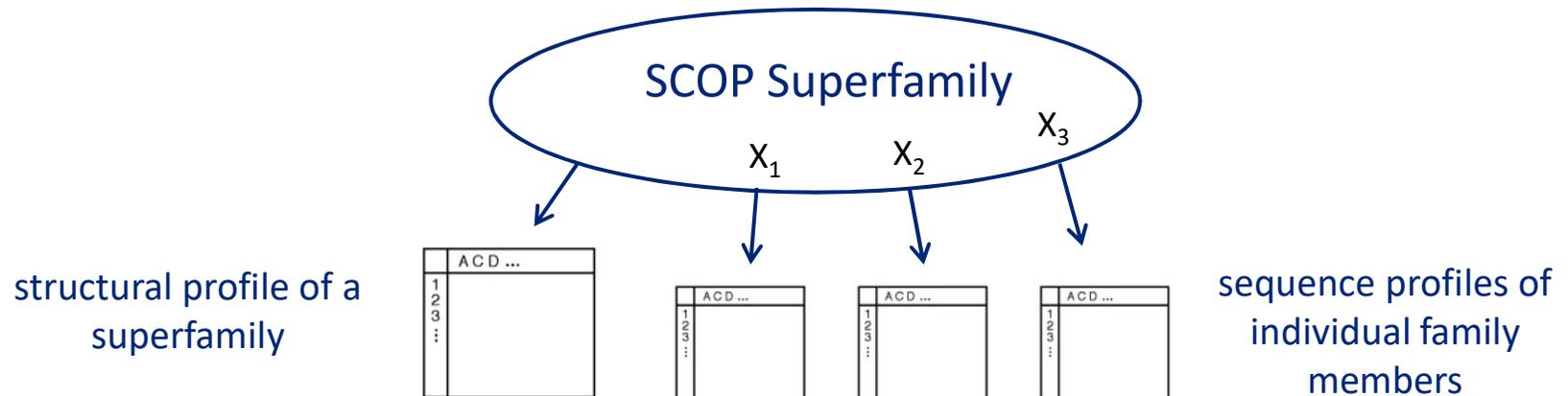
- profile methods



**Query  
sequence**

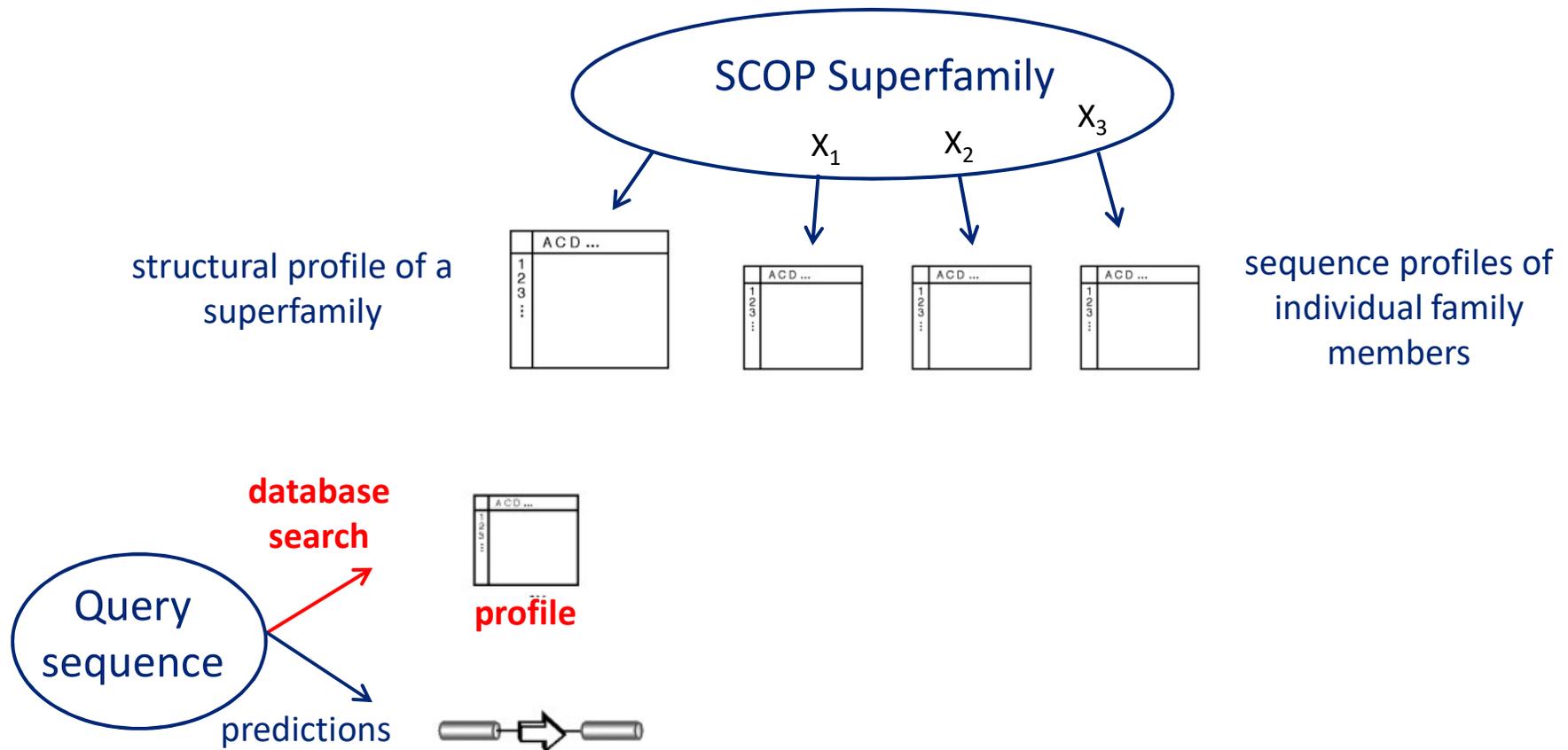
# Fold recognition (Profiles)

- profile methods



# Fold recognition (Profiles)

- profile methods



# Fold recognition (Profiles)

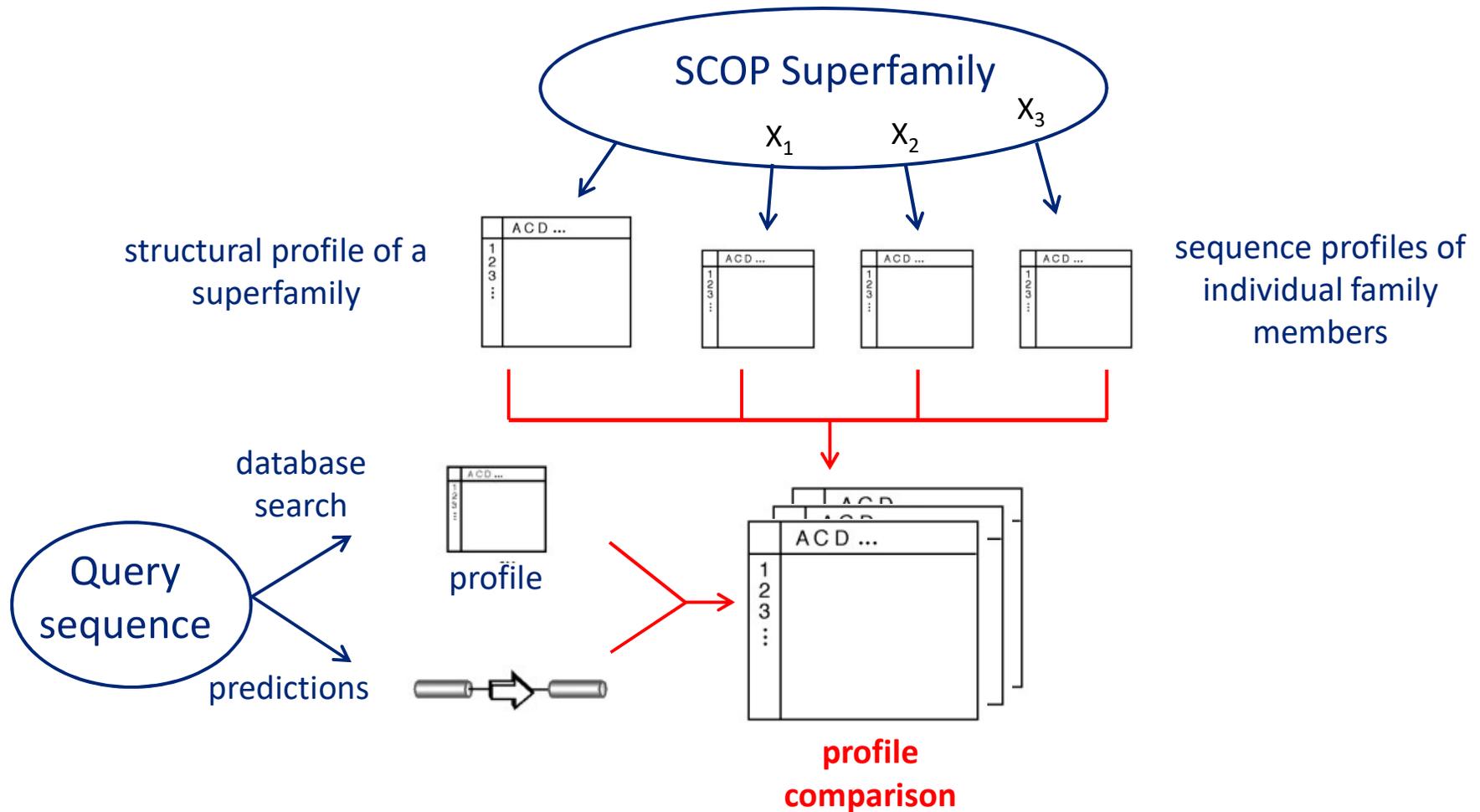


## □ profile methods

1. construction of **profile(s)** for a group of **related protein structures** (e.g., for each SCOP superfamily) – scores describing the propensity of each residue to be at each profile position, information for secondary structural types, solvent accessibility, polarity, sequence-based profiles, ...
2. construction of **profile(s) for the query sequence** – sequence-based profile from the multiple sequence alignment, predicted secondary structure, solvent accessibility, polarity,...

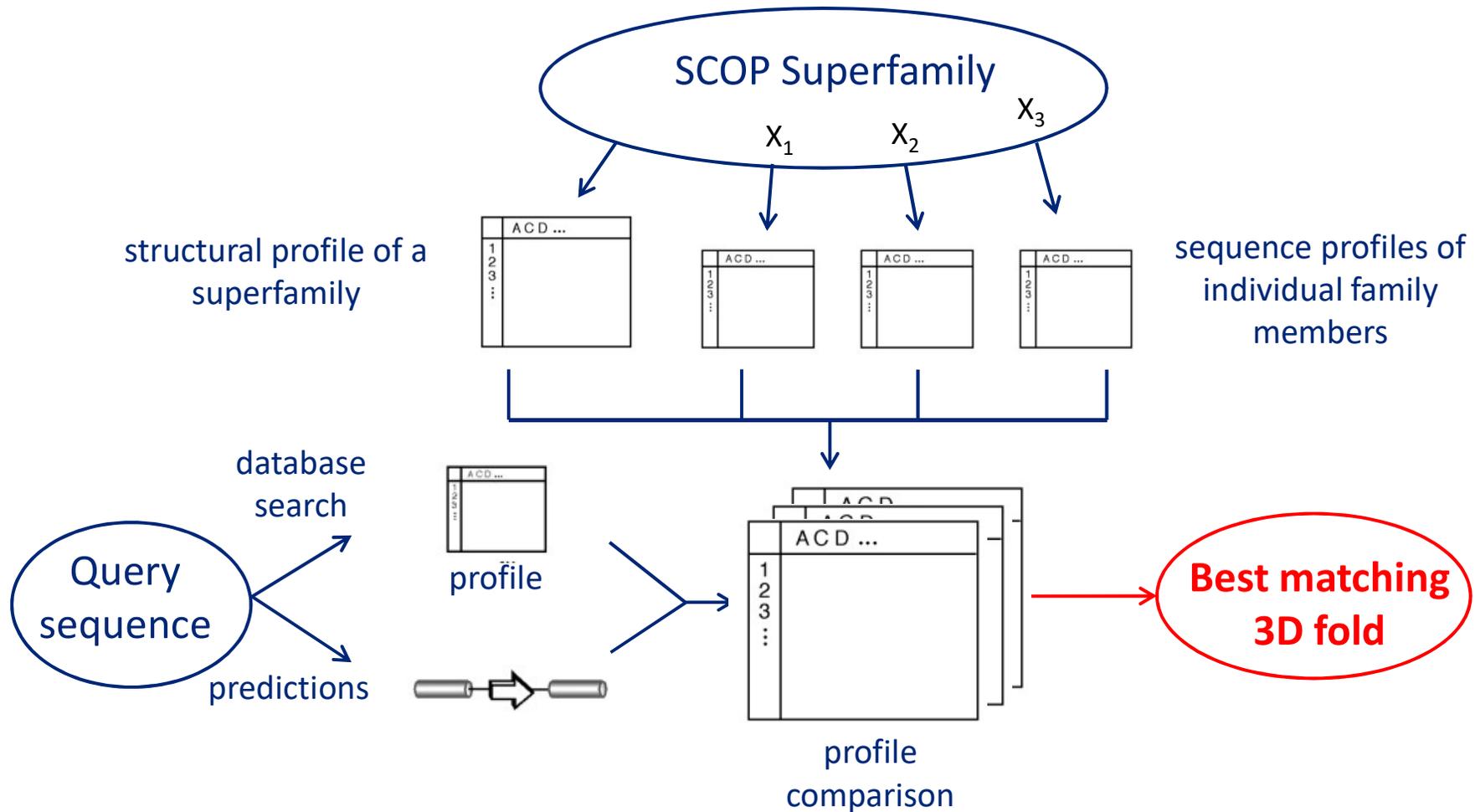
# Fold recognition (Profiles)

- profile methods



# Fold recognition (Profiles)

## □ profile methods



# Fold recognition (Profiles)



## □ profile methods

1. construction of **profile(s)** for a group of **related protein structures** (e.g., for each SCOP superfamily) – scores describing the propensity of each residue to be at each profile position, information for secondary structural types, solvent accessibility, polarity, sequence-based profiles, ...
2. construction of **profile(s) for the query sequence** – sequence-based profile from the multiple sequence alignment, predicted secondary structure, solvent accessibility, polarity,...
3. **comparison** of the query profiles with **profiles of known structural folds** to find the fold that best represents the query sequence

# Fold recognition programs



## □ PHYRE

- <http://www.sbg.bio.ic.ac.uk/phyre2/>
- **profile-based** method
- the highest scoring alignments are used to construct full 3D models of the query – missing or inserted regions are repaired using a loop library and reconstruction procedure, side-chains are placed using a fast graph-based algorithm

# Fold recognition programs

## □ PHYRE

Fold Recognition							
View Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	Fold/PDB descriptor	Superfamily
	<b>c2bk9A</b> (length:145) <b>100% i.d.</b>	 	9.3e-20	100 %	0.90 BioText	Globin-like	Globin-like
	<b>c2bk9A</b> (length:153) <b>23% i.d.</b>	 	7.7e-17	100 %	0.89 BioText	<b>PDB header:</b> oxygen transport	<b>Chain: A: PDB Molecule:</b> cg9734-pa;

# Fold recognition programs



## □ RaptorX

- <http://raptorx.uchicago.edu/>
- provides single-template threading, alignment quality prediction, and multiple-template threading

## □ GenTHREADER

- <http://bioinf.cs.ucl.ac.uk/psipred/>
- uses a hybrid of the profile and pairwise energy methods
- multiple sequence alignment and secondary structure predictions derived for the query are used as input for threading
- threading results are evaluated using neural networks

# *Ab initio* prediction



- ❑ attempts to generate a **structure by using physicochemical principles only**
- ❑ used when neither homology modeling nor fold recognition can be applied
- ❑ search for the structure in the global free-energy minimum
- ❑ so far still limited success in getting correct structures

# *Ab initio* prediction programs



## □ Rosetta

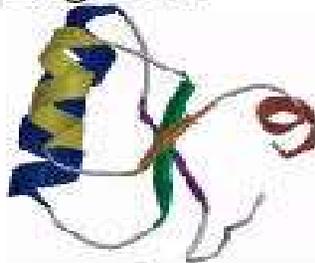
- <http://www.rosettacommons.org/>
- software suite for predicting and designing protein structures, protein folding mechanisms, and protein-protein interactions



# Ab initio prediction programs

- Rosetta

Target 77



native

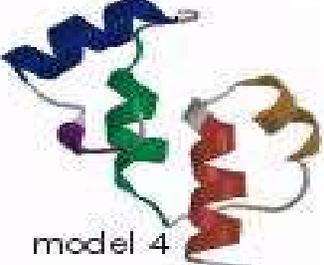


model 4

Target 56

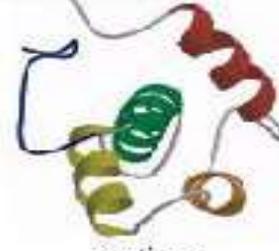


native



model 4

Target 74

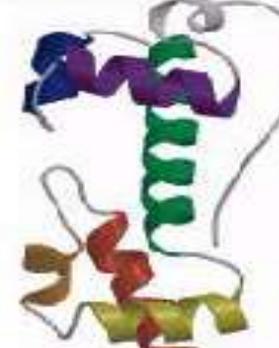


native

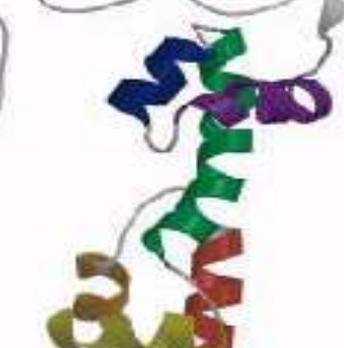


model 4

Target 79



native



model 4

# “Hybrid” 3D structure prediction programs



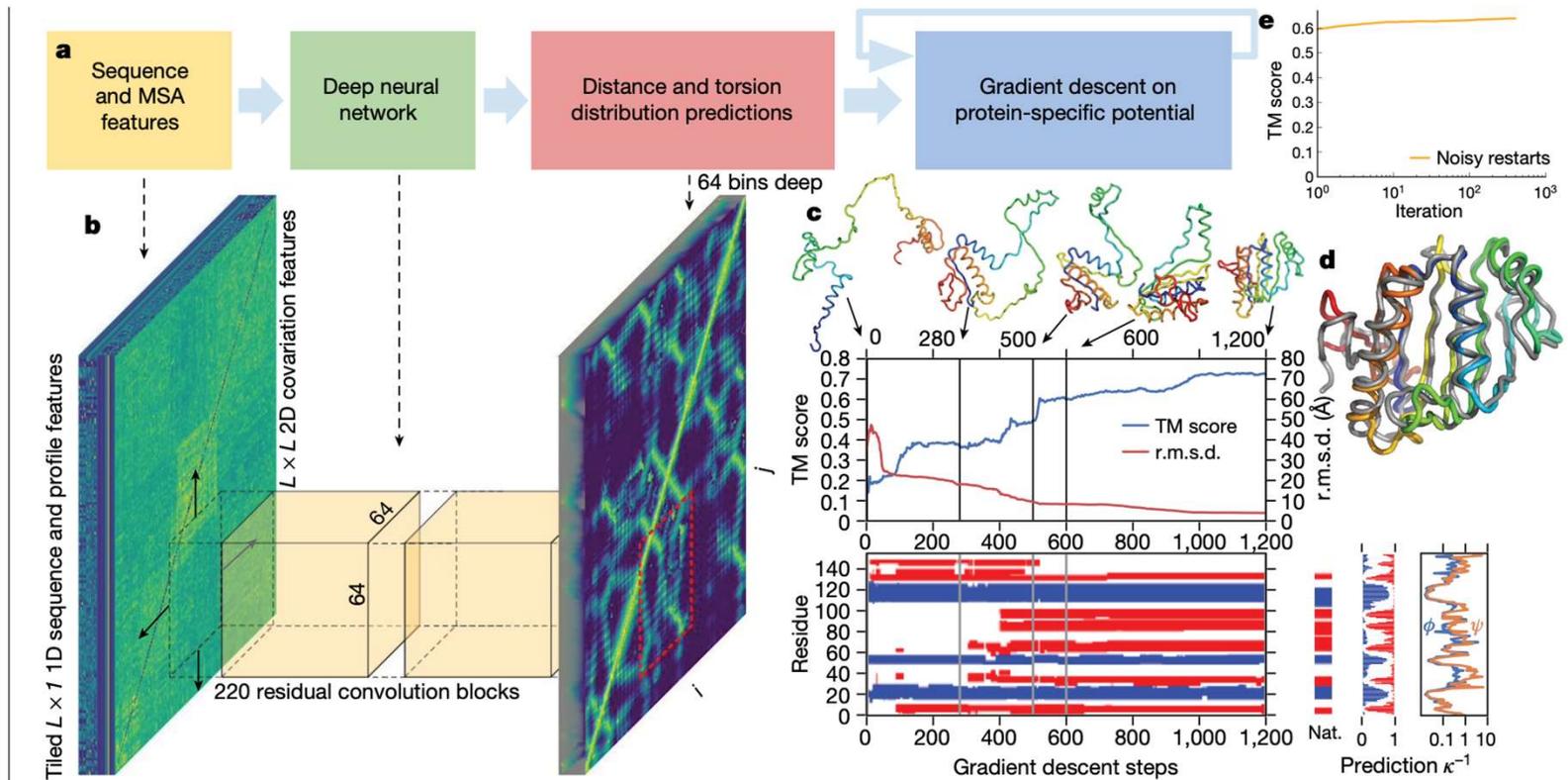
## □ I-TASSER

- <http://zhanglab.ccmb.med.umich.edu/I-TASSER/>
- combines homology modeling, threading and *ab initio* predictions
- **No. 1 server** for protein structure prediction in previous CASP experiments

## □ Robetta

- <http://robetta.bakerlab.org/>
- combines homology modeling and *ab initio* predictions
- implements ROSETTA software

# AlphaFold: ML-powered threading



- Combines threading with ML
- **No. 1 server** for protein structure prediction in the last 2 CASP experiments



# Assessment of prediction methods



- CASP (**C**ritical **A**ssessment of techniques for protein **S**tructure **P**rediction)
  - <http://predictioncenter.org/>
  - biannual international contest providing objective **evaluation of the performance** of individual **prediction methods**
  - evaluation **based on** a large number of **blind predictions** -  
contestants are given protein sequences whose structures have been solved, but not yet published - results of the predictions are compared with the newly determined structure
  - competition in several categories

# Assessment of prediction methods



- CAMEO (**C**ontinuous **A**utomated **M**odel **E**valuati**O**n)
  - <https://www.cameo3d.org/>
  - weekly assessment of new structures in the PDB
  - registered prediction servers are sent weekly requests on not-so-easy new structures in the weekly PDB pre-release.
  - Multiple scores considered, normalized average (IDDT) reported
  - Categories:
    - 3D: Prediction of the 3D coordinates of a protein from sequence
    - QE: Model quality Estimation: Assessment of quality measures reported by participant servers

# Databases of protein models



## ❑ Protein Model Portal

- <http://www.proteinmodelportal.org/>
- **access to pre-computed** (automatically generated) **models** from six structural genomics centers and independent modeling groups, e.g., ModBase and SWISS-MODEL repository
- reliability of model estimated based on the target-template identity

Models:

Model	Rel.	Provider	Type	Templates	%Seq id	from	to	Sel.
[Show]		MODBASE	SC	<b>1b6g</b> ↗	28%	1	296	<input type="checkbox"/>
[Show]		NESG	TC	<b>1y7hA</b> ↗	13%	45	296	<input type="checkbox"/>
[Show]		NESG	TC	<b>1y7iA</b> ↗	13%	45	296	<input type="checkbox"/>
[Show]		MODBASE	SC	<b>1r3dA</b> ↗	12%	34	301	<input type="checkbox"/>
[Show]		NYSGXRC	TC	<b>1r3dA</b> ↗	12%	35	301	<input type="checkbox"/>

# Databases of protein models



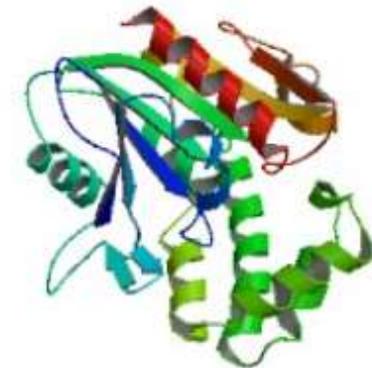
## ❑ ModBase

- <http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi>
- database of annotated protein models generated by the **automated** pipeline including the **MODELLER** program
- contains ~38 millions models for ~6.5 millions unique sequences

Quality criteria indicate whether the model is considered **reliable (green)** or **unreliable (red)**.

Target Region	34-301
Protein Length	301
Template PDB Code	<a href="#">1r3dA</a>
Template Region	4-262
Sequence Identity	12.00%
E-Value	2e-25
GA341	0.18
Dataset	nysgxrc_1r3d_3-06
ModPipe Version	ModPipe1.0
Model Date	2006-04-15

for all Models of this Sequence:



# Databases of protein models



## ❑ SWISS-MODEL repository

- <http://swissmodel.expasy.org/repository/>
- database of annotated protein models generated by the **automated** homology-modeling pipeline **SWISS-MODEL**.
- contains 2.2 millions models for UniProt sequences

## ❑ PMDB (**P**rotein **M**odel **D**ata**B**ase)

- <http://srv00.recas.ba.infn.it/PMDB/>
- contains **manually built** 3D protein models
- users can download as well as submit models along with related supporting evidence

# Databases of protein models

Safari File Edit View History Bookmarks Develop Window Help

alphafold.ebi.ac.uk

AlphaFold Protein Structure Database

Home About FAQs Downloads

# AlphaFold Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism **BETA** **Search**

Examples: Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli Help: AlphaFold DB search help

Feedback on structure: Contact [alphafold@deepmind.com](mailto:alphafold@deepmind.com)

# References



- ❑ Gu, J. & Bourne, P. E. (2009). **Structural Bioinformatics, 2<sup>nd</sup> 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.