Universitas Esa Unggul

Smart, Creative and Entrepreneurial

# IBT 432 Aplikasi Bioinformatika

## Protein modelling I: Protein structure and Protein Data Bank

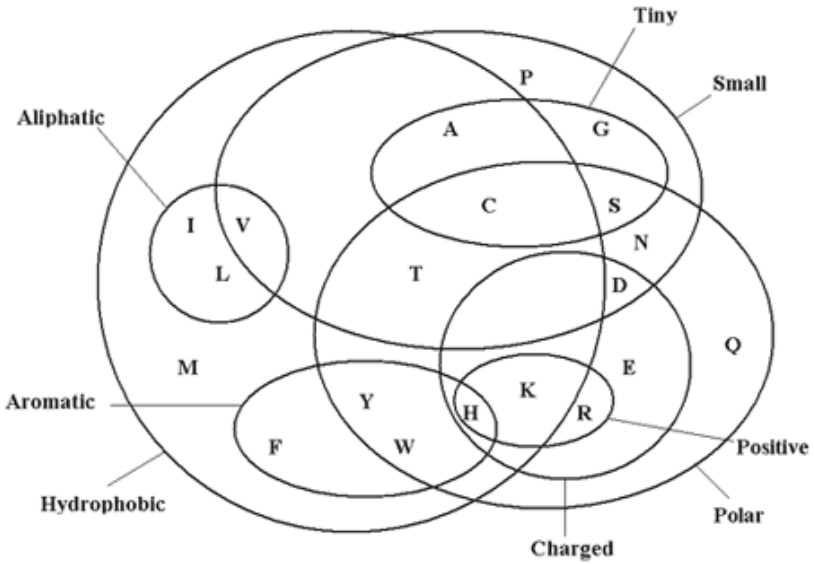
**Riza Arief Putranto**

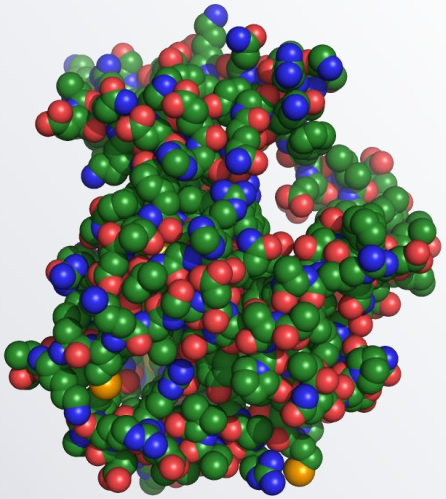
# Rencana Perkuliahan

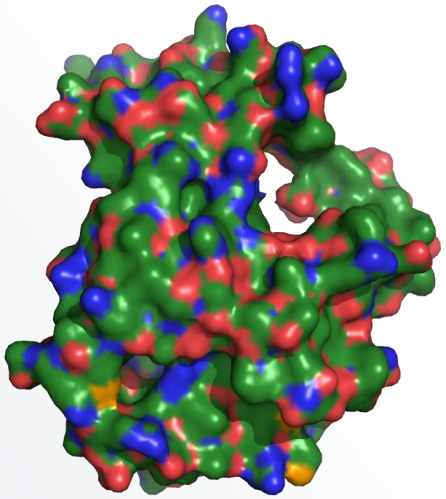
1. ~~Kontrak belajar dan pengenalan bioinformatika aplikatif~~
2. ~~Database sekuen dan analisis genomika~~
3. ~~Anotasi sekuen ke genom - Praktik~~
4. ~~Analisis komparasi genomika I~~
5. ~~Analisis komparasi genomika II~~
6. ~~Analisis komparasi genomika III~~
7. ~~Analisis komparasi genomika – Praktik~~
8. Protein modelling I
9. Protein modelling II
10. Protein modelling III
11. Protein modelling - Praktik
12. Visualisasi protein modelling
13. Visualisasi protein modelling - Praktik
14. Presentasi mahasiswa

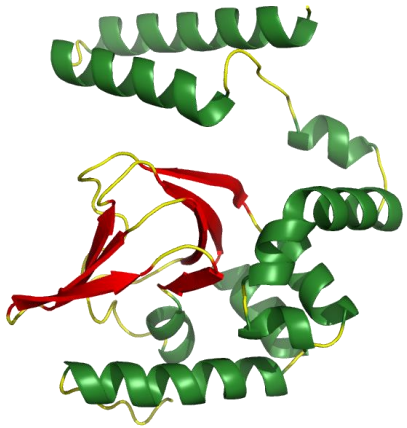
# Amino acid residues

- Proteins are made up of amino acids, which are interconnected by peptide bonds
- There are 20 naturally occurring amino acids
- Amino acids may be subdivided by their individual properties



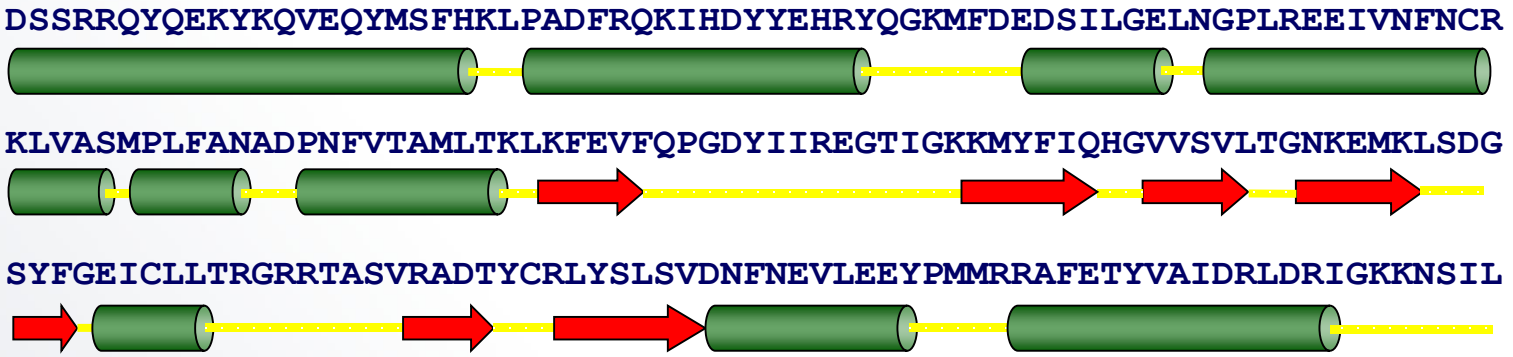



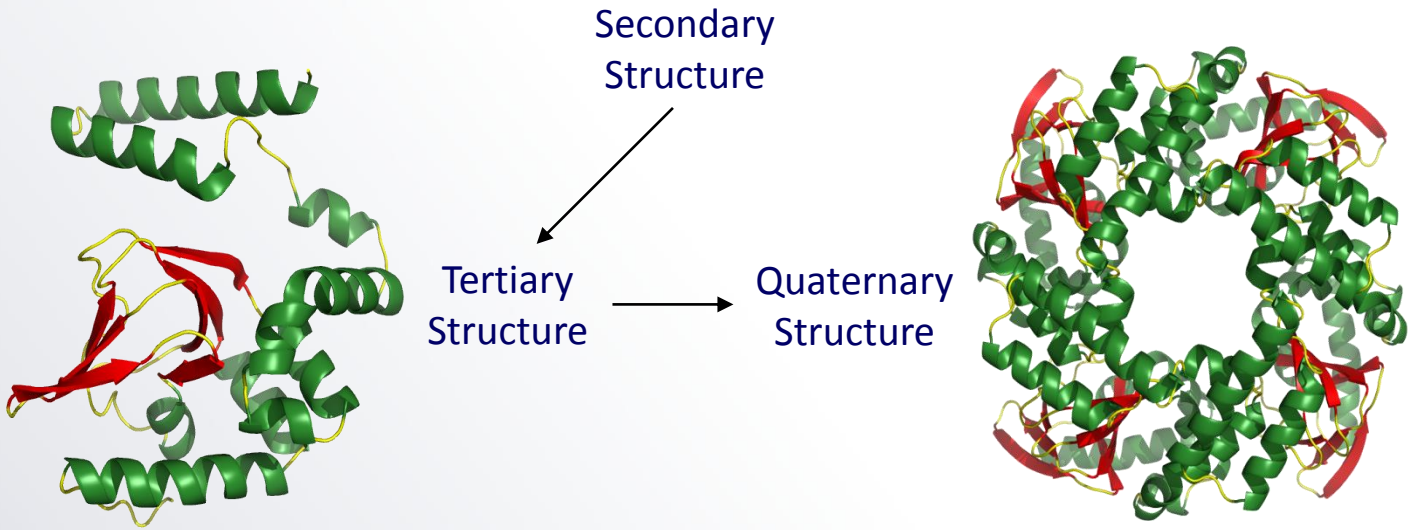




# From sequence to structure

## Primary Structure – Amino Acid Sequence

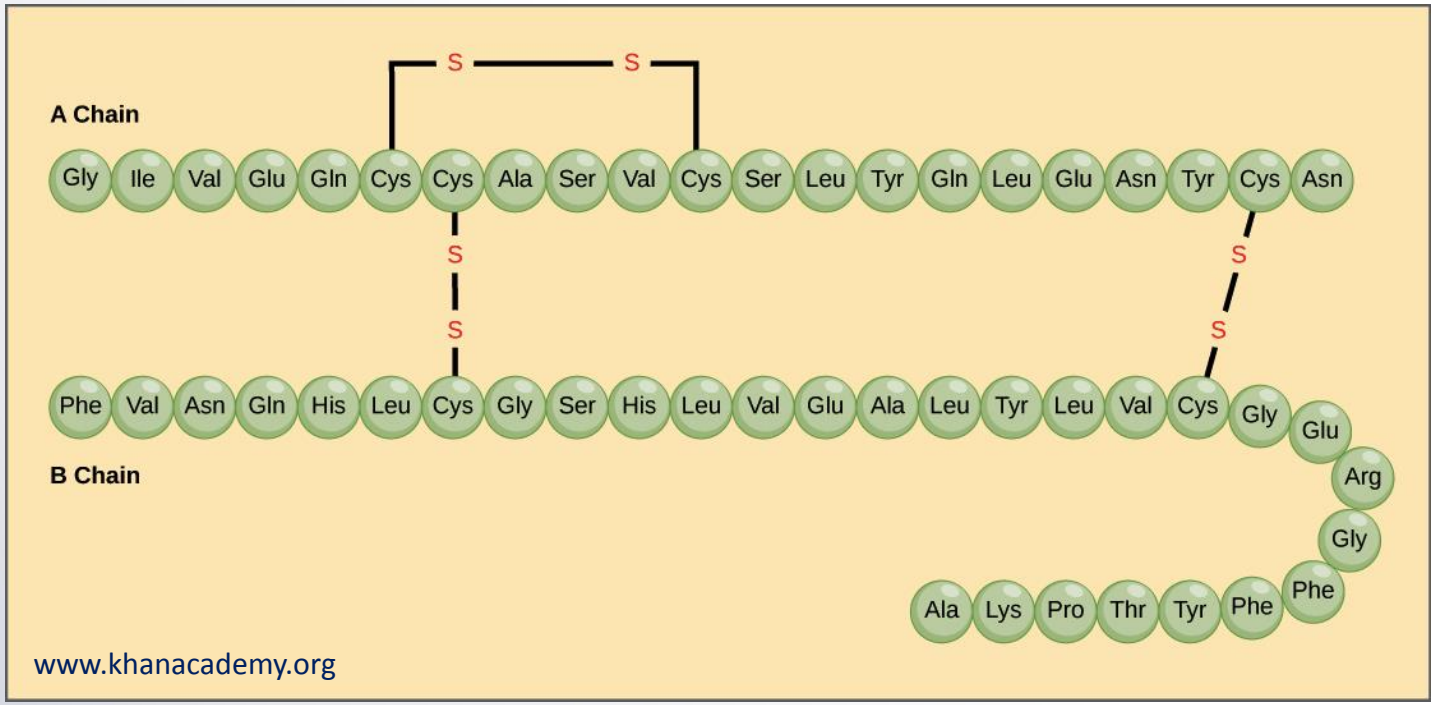



**What information can we get from a Sequence of amino acids?**

# Primary structure

The simplest level of protein structure, **primary structure**, is simply **the sequence of amino acids in a polypeptide chain**. For example, the hormone insulin has two polypeptide chains, A and B, shown in diagram below. (The insulin molecule shown here is cow insulin, although its structure is similar to that of human insulin.) Each chain has its own set of amino acids, assembled in a particular order. For instance, the sequence of the A chain starts with glycine at the N-terminus and ends with asparagine at the C-terminus, and is different from the sequence of the B chain.



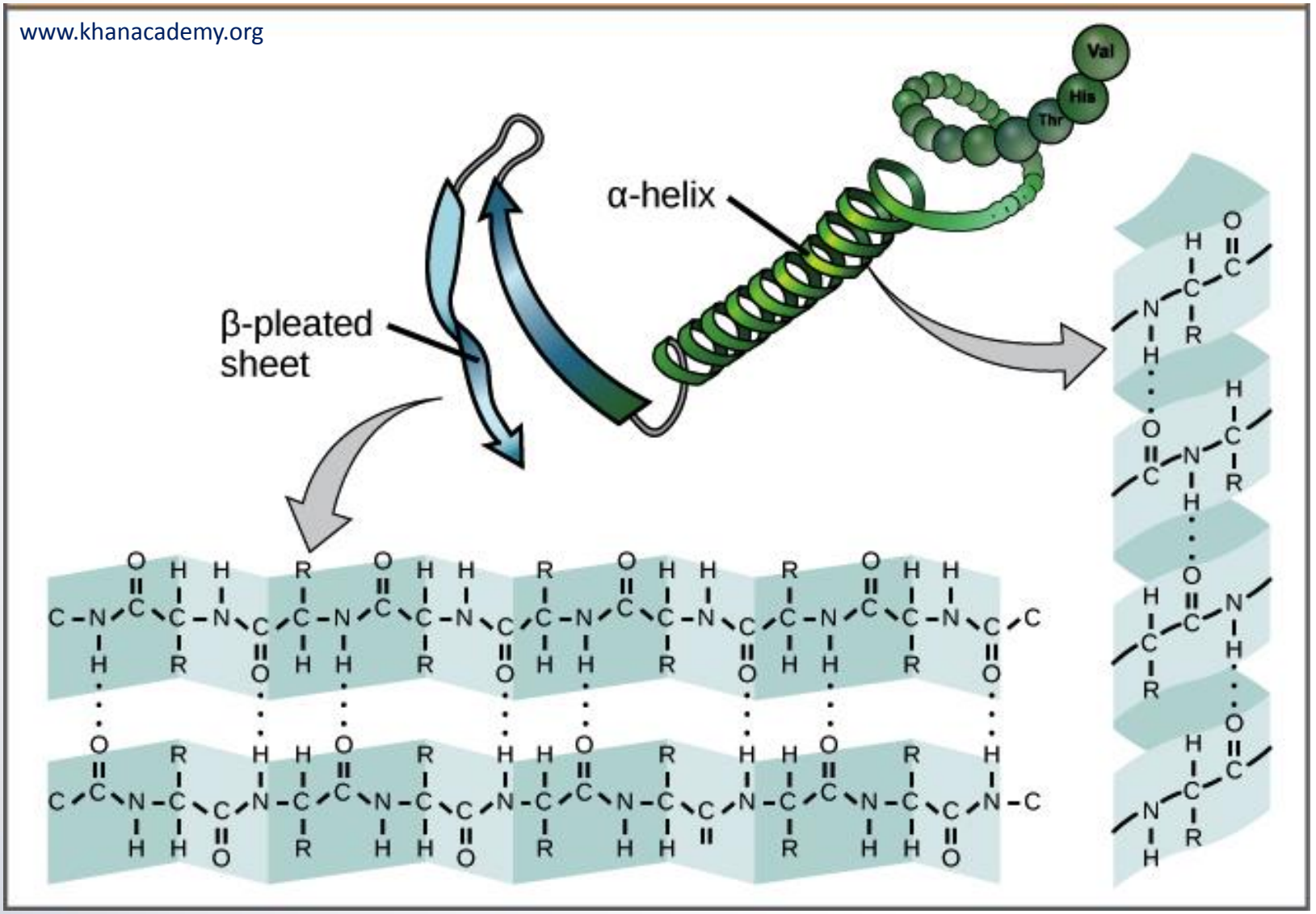
# Secondary structure

The next level of protein structure, **secondary structure**, refers to local folded structures that form within a polypeptide due to interactions between atoms of the backbone. (The backbone just refers to the polypeptide chain apart from the R groups – so all we mean here is that secondary structure does not involve R group atoms.) The most common types of secondary structures are the α helix and the β pleated sheet. Both structures are held in shape by hydrogen bonds, which form between the carbonyl O of one amino acid and the amino H of another.

In an **α helix**, the carbonyl (C=O) of one amino acid is hydrogen bonded to the amino H (N-H) of an amino acid that is four down the chain. This pattern of bonding pulls the polypeptide chain into a helical structure that resembles a curled ribbon, with each turn of the helix containing 3.6 amino acids.

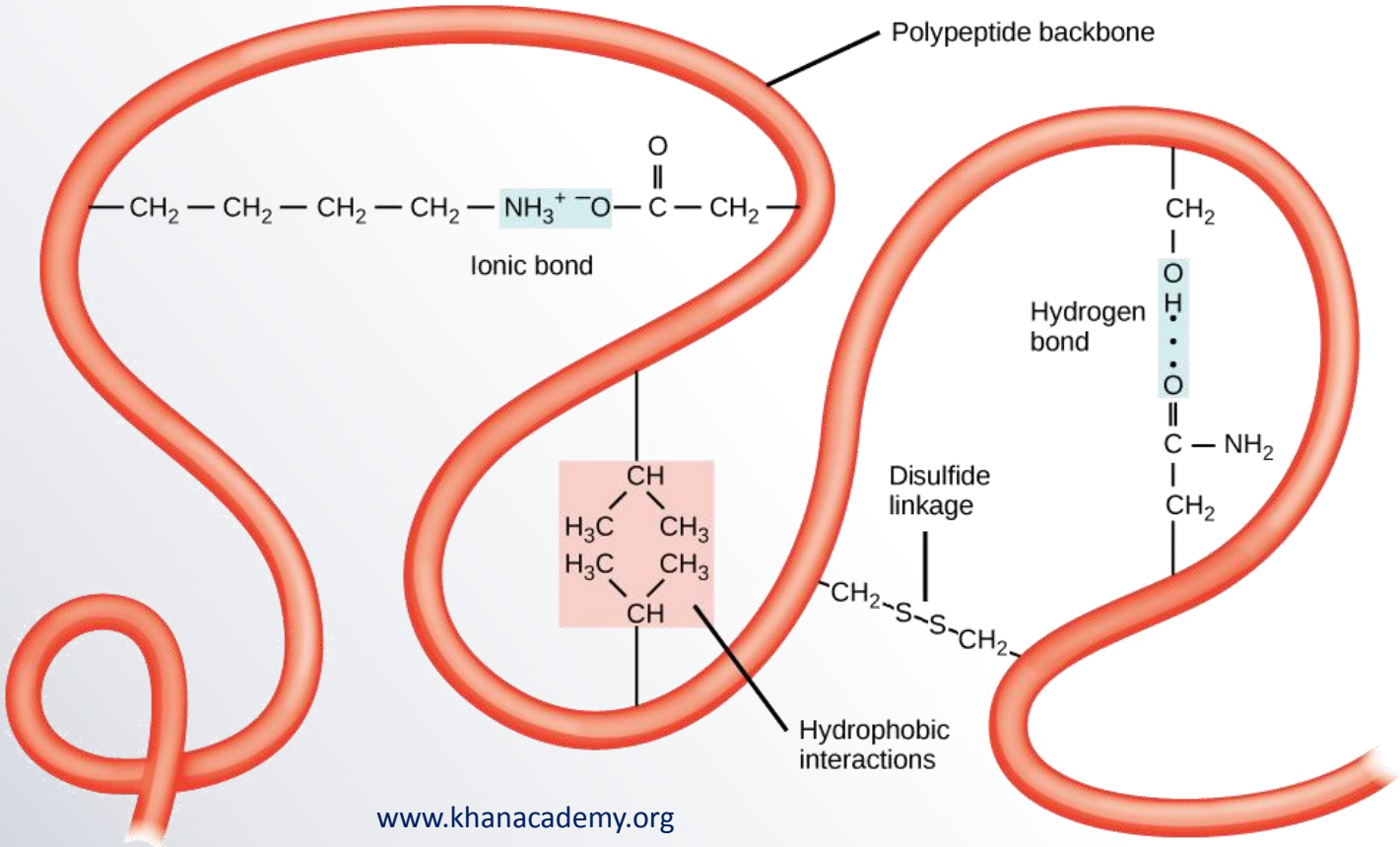
In a **β pleated sheet**, two or more segments of a polypeptide chain line up next to each other, forming a sheet-like structure held together by hydrogen bonds. The strands of a β pleated sheet may be **parallel**, pointing in the same direction (meaning that their N- and C-termini match up), or **antiparallel**, pointing in opposite directions (meaning that the N-terminus of one strand is positioned next to the C-terminus of the other).

# Secondary structure



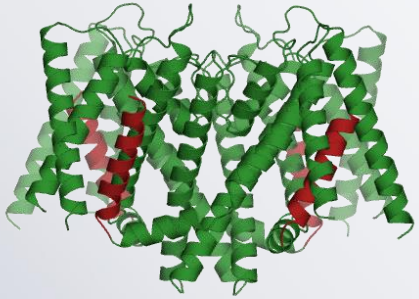
# Tertiary structure

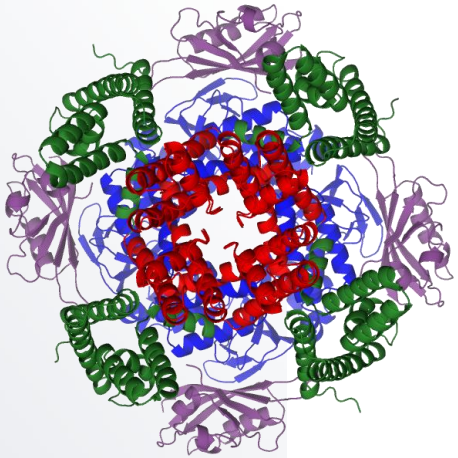
The overall three-dimensional structure of a polypeptide is called its **tertiary structure**. The tertiary structure is primarily due to interactions between the R groups of the amino acids that make up the protein.

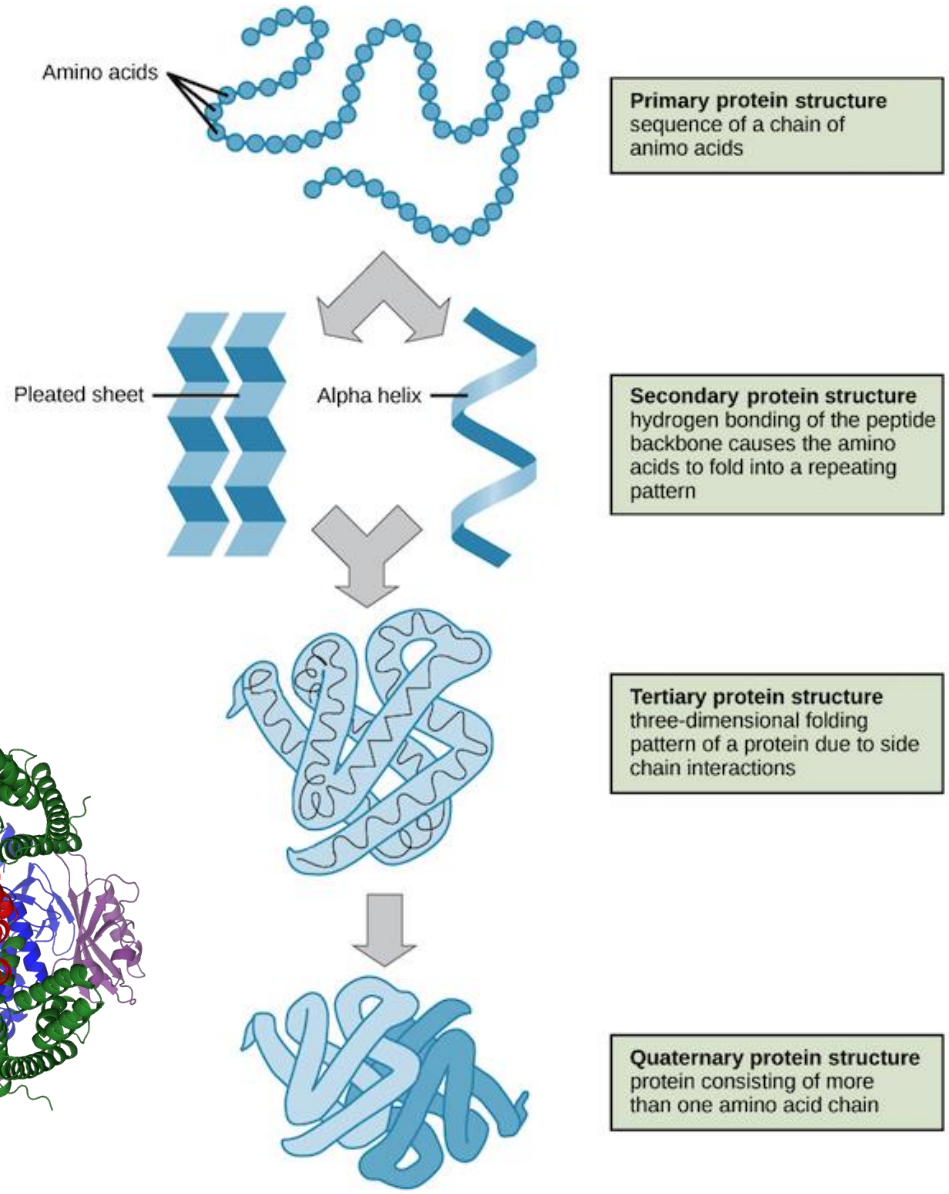


# Quartenary structure

Many proteins are made up of a single polypeptide chain and have only three levels of structure (the ones we've just discussed). However, some proteins are made up of multiple polypeptide chains, also known as subunits. When these subunits come together, they give the protein its **quaternary structure**.





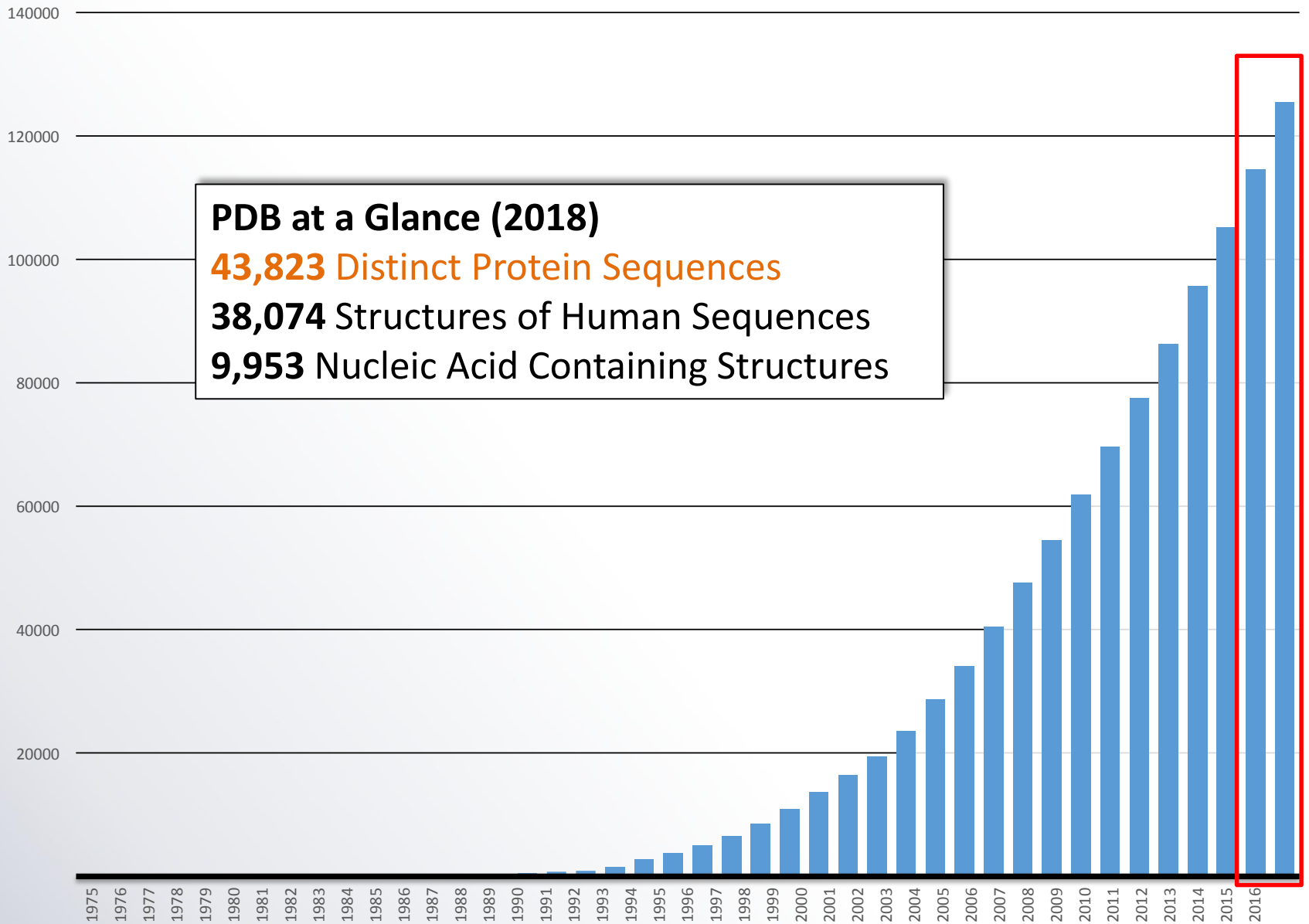


# Protein Data Bank: A Structural View of Biology


- 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.

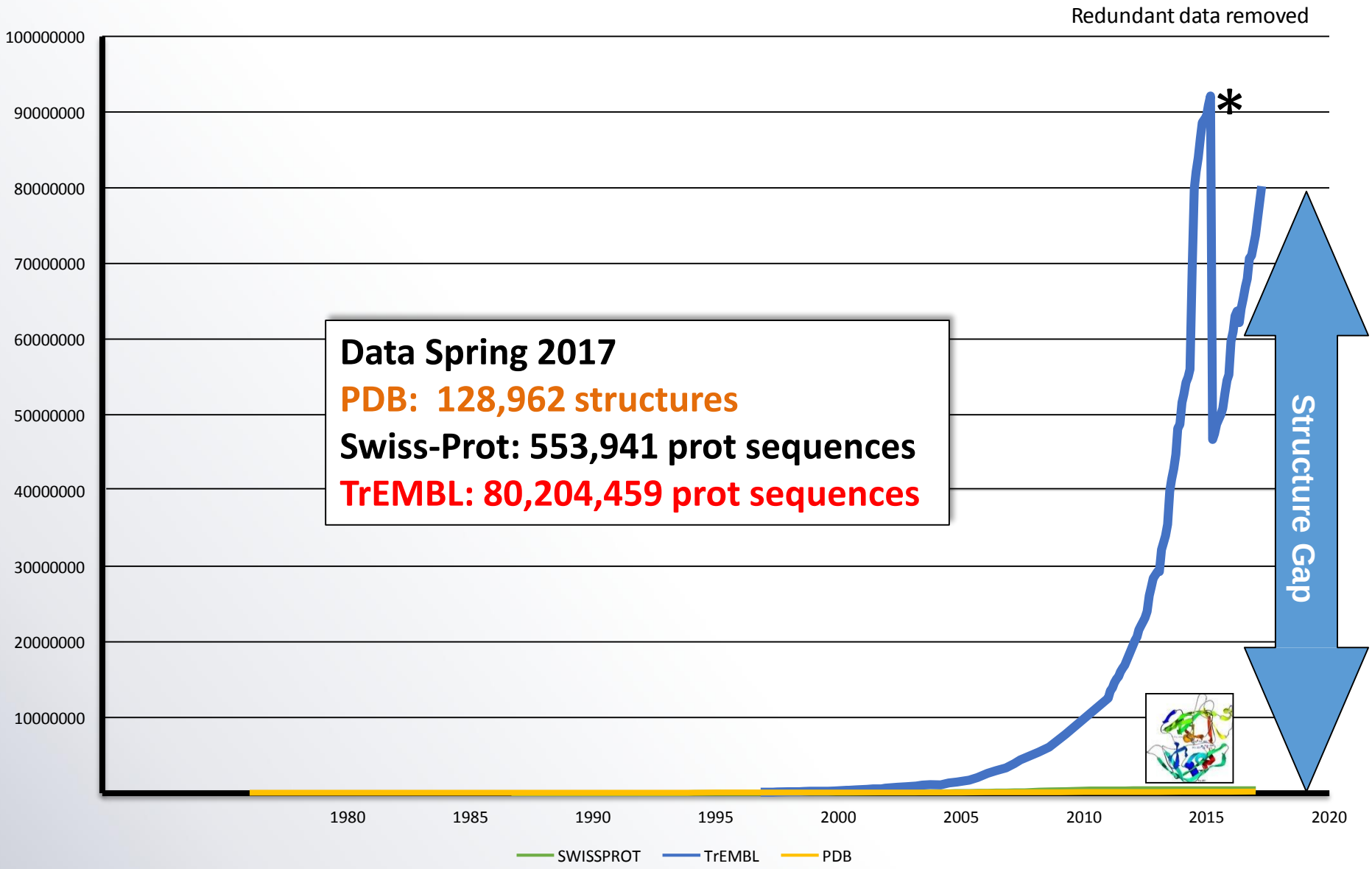
https://www.rcsb.org/

# Protein structure submission in PDB (2018)




https://swissmodel.expasy.org/course

# Structure gap in databases: a real problem




https://swissmodel.expasy.org/course

The RCSB Protein Data Bank - Microsoft Internet Explorer

File Edit View Favorites Tools Help

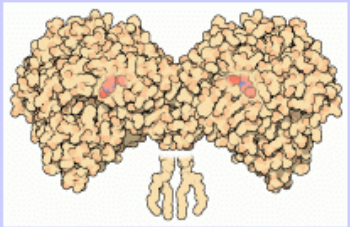
Back Search Favorites Media

Address http://www.rcsb.org/pdb/ Go Links

*DEPOSIT data*
*DOWNLOAD files*
*browse LINKS*
*BETA TEST new features*
*BETA XML files*

## Current Holdings

25760 Structures
Last Update: 01-Jun-2004
PDB Statistics



Molecule of the Month: Acetylcholinesterase

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology/UMBI/NIST -- three members of the Research Collaboratory for Structural Bioinformatics (RCSB).

The RCSB PDB is supported by funds from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of

# RCSB PDB
PROTEIN DATA BANK

RCSB Home | wwPDB Home | Contact Us | Help

**Did you find what you wanted?**

Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

ABOUT PDB | NEW FEATURES | USER GUIDES | FILE FORMATS | DATA UNIFORMITY | STRUCTURAL GENOMICS | SOFTWARE | PUBLICATIONS | EDUCATION

## Search the Archive

Enter a PDB ID or keyword

Query Tutorial

Search

PDB ID Authors Full Text Search
match exact word remove similar sequences

**QuickSearch!** search Web pages and structures
**SearchLite** keyword search form with examples
**SearchFields** customizable search form
**Status Search** find entries awaiting release

## News

Complete News Newsletter | pdb-l Archive Subscribe

**1-June-2004**
**TargetDB and Ligand Depot papers published in *Bioinformatics***
Two papers have been published online that describe the TargetDB and Ligand Depot resources. [MORE...]

## PDB Mirrors

***Please bookmark a mirror site***

**San Diego Supercomputer Center, UCSD***
**Rutgers University***
**Center for Advanced Research in Biotechnology, NIST***
**Cambridge Crystallographic Data Centre, UK**
**National University of Singapore**
**Osaka University, Japan**
**Max Delbrück Center for Molecular Medicine, Germany**

**OCA / PDB Lite** MORE...

**RCSB partner*

In citing the PDB please refer to:

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. *Nucleic Acids Research*, **28** pp. 235-242 (2000)

Beta-test new features Internet

Start The ... Bioinf... struc... Docu... Micro... Adob... 02:41

# The Protein Data Bank

- **The Protein Data Bank** was established at Brookhaven National Labs in 1971 as an archive of biological macromolecular crystal structures.

- Since October 1998, the PDB database has been managed by **the Research Collaboratory for Structural Bioinformatics (RCSB)**, which is a consortium consisting of Rutgers, the State University of New Jersey; The San Diego Supercomputer Centre at the University of California, San Diego; and the National Institute of Standards and Technology.

- As of 1st June 2004 25,760 structures have been deposited in the PDB
- As of 16th Nov 2018 **146,266 structures** have been deposited in the PDB

| Experimental Method | Proteins | Nucleic Acids | Protein/NA Complex | Other | Total |
|---|---|---|---|---|---|
| X-Ray | 122565 | 1975 | 6350 | 10 | 130900 |
| NMR | 10907 | 1263 | 253 | 8 | 12431 |
| Electron Microscopy | 1850 | 31 | 660 | 0 | 2541 |
| Other | 244 | 4 | 6 | 13 | 267 |
| Multi Method | 119 | 5 | 2 | 1 | 127 |
| **Total** | **135685** | **3278** | **7271** | **32** | **146266** |

# PDB (https://www.rcsb.org/)

- The PDB archive contains macromolecular structure data on **proteins**, **nucleic acids**, **protein-nucleic acid complexes**, and **viruses**. Files in its holdings are deposited by the international user community and maintained by the RCSB PDB staff. Approximately 50-100 new structures are deposited each week. They are annotated by RCSB and released upon the depositor's specifications. PDB data is freely available worldwide.

- A variety of information associated with each structure is available, including sequence details, **atomic coordinates**, **crystallization conditions**, **3-D structure neighbours** computed using various methods, derived geometric data, structure factors, 3-D images, and a variety of links to other resources.

- Information on structures can be retrieved from the main PDB Web site at **http://www.pdb.org**/, or one of its mirror sites. Structure files can also be obtained through the main FTP site at **ftp://ftp.rcsb.org**/ or one of its mirrors.

RCSB PDB
PROTEIN DATA BANK

# Theoretical Models

- The PDB separated theoretical model coordinate files from the main archive beginning July 1, 2002. Since that date, the main archive has consisted of structures determined using experimental methods only. Theoretical models are only available for download from the PDB FTP site as follows:

  - All theoretical models (current and obsolete) are kept in a separate location in the FTP archive (**pub/pdb/data/structures/models/current, pub/pdb/data/structures/models/obsolete**)

  - Model index files (authors.idx and titles.idx) and a FASTA file (model_seqres.txt) are available at pub/pdb/data/structures/models/index.

  - A simple search interface for theoretical models is available **http://www.rcsb.org/pdb/cgi/models.cgi**. Queries from any other search interface do not return model entries (except for direct lookups by PDB ID).

# Data acquisition and processing

**Public archive**

- Efficient data capture
- Data curation

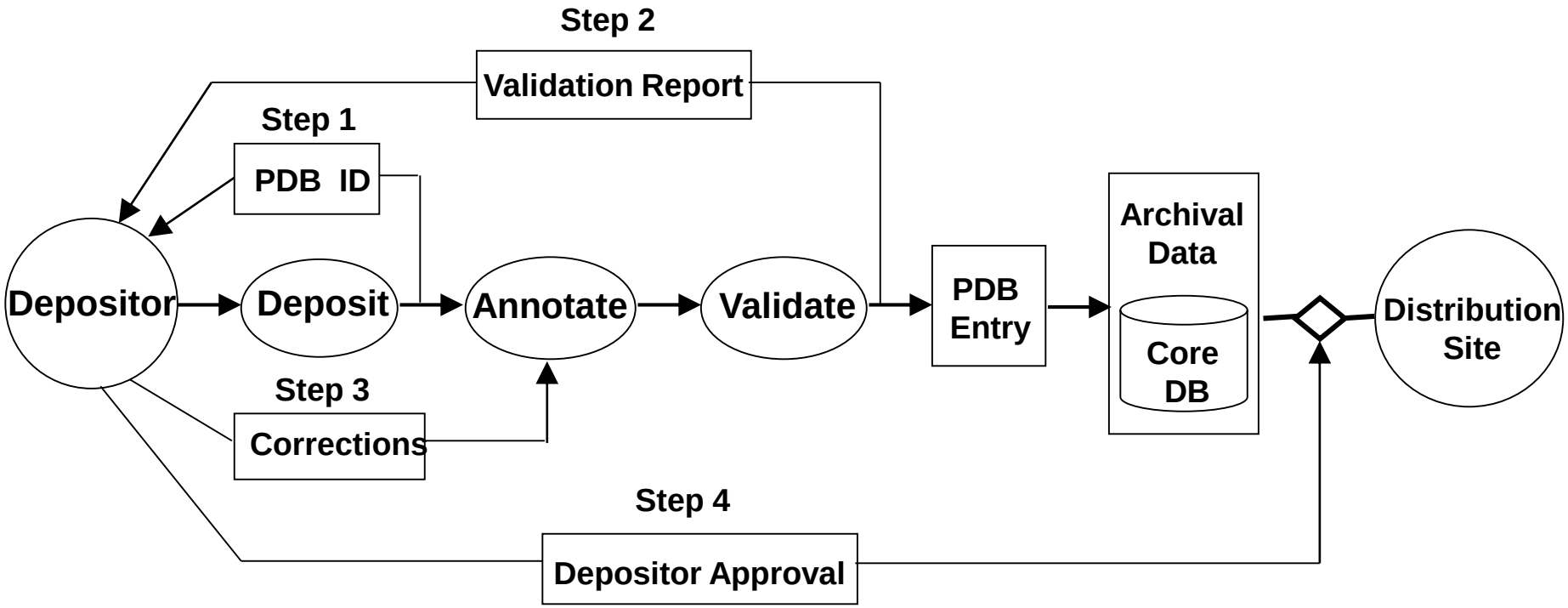
**Data processing**

- Data deposition
- Annotation
- Validation

# Data submission

**Step 1**

After a structure has been deposited using ADIT, a PDB identifier is sent to the author automatically and immediately. This is the first stage in which information about the structure is loaded into the internal core database.

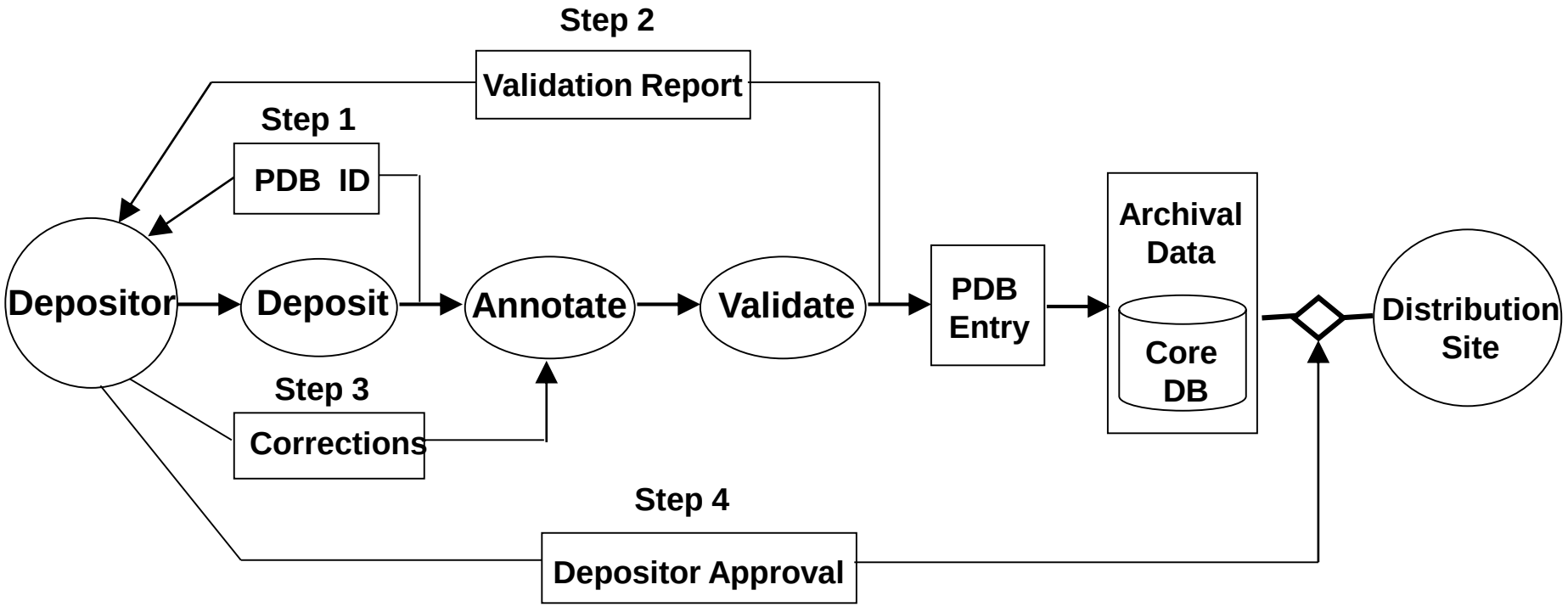



Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). **The Protein Data Bank**. *Nucleic Acids Res.* **28**, 235-242

# Data submission

**Step 2**

The entry is annotated. This process involves using ADIT to help diagnose errors or inconsistencies in the files. The completely annotated entry as it will appear in the PDB resource, together with the validation information, is sent back to the depositor.



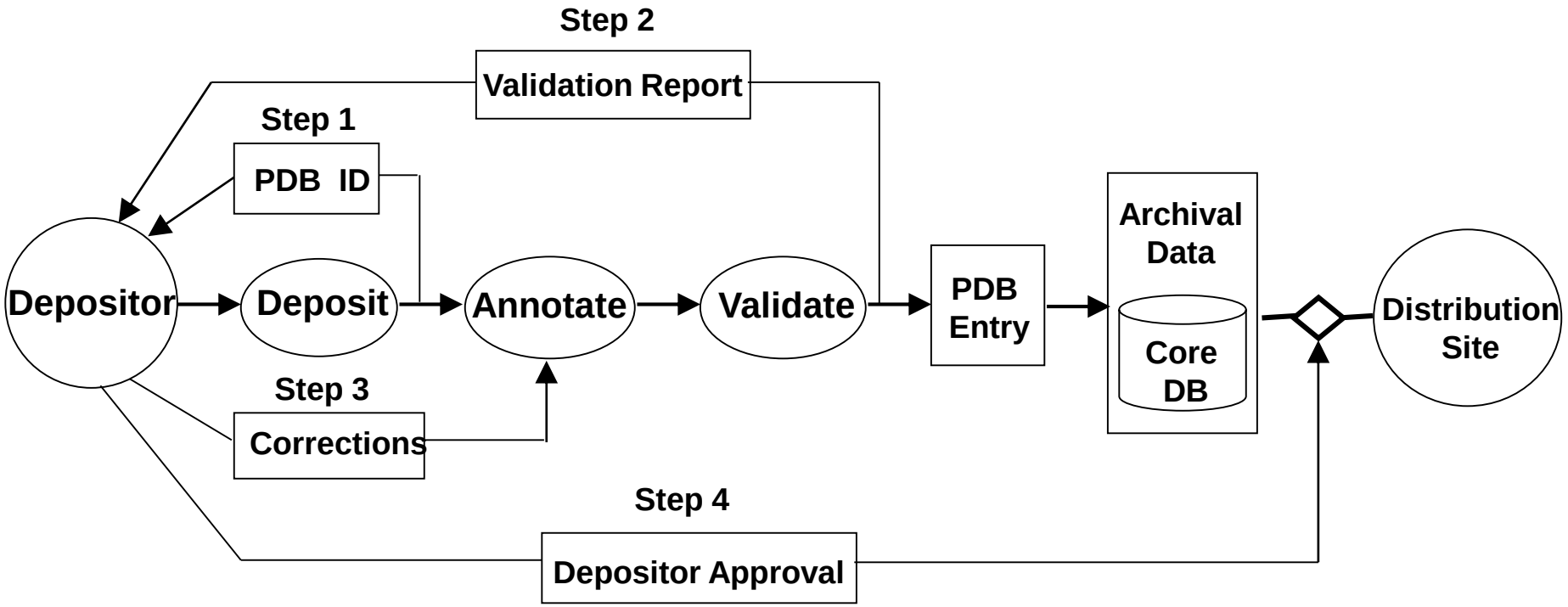

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). **The Protein Data Bank**. *Nucleic Acids Res.* **28**, 235-242

# Data submission

**Step 3**

After reviewing the processed file, the author sends any revisions.

Depending on the nature of these revisions, Steps 2 and 3 may be repeated.

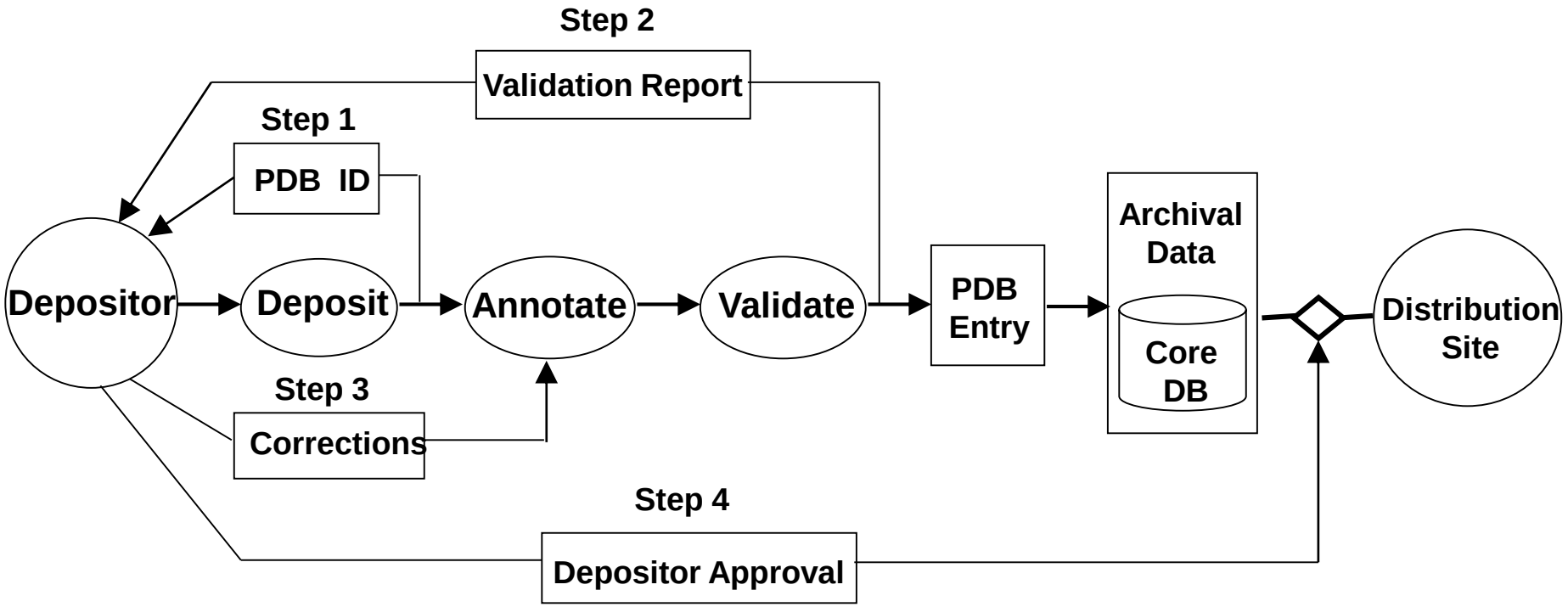



Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). **The Protein Data Bank**. *Nucleic Acids Res.* **28**, 235-242

# Data submission

**Step 4**

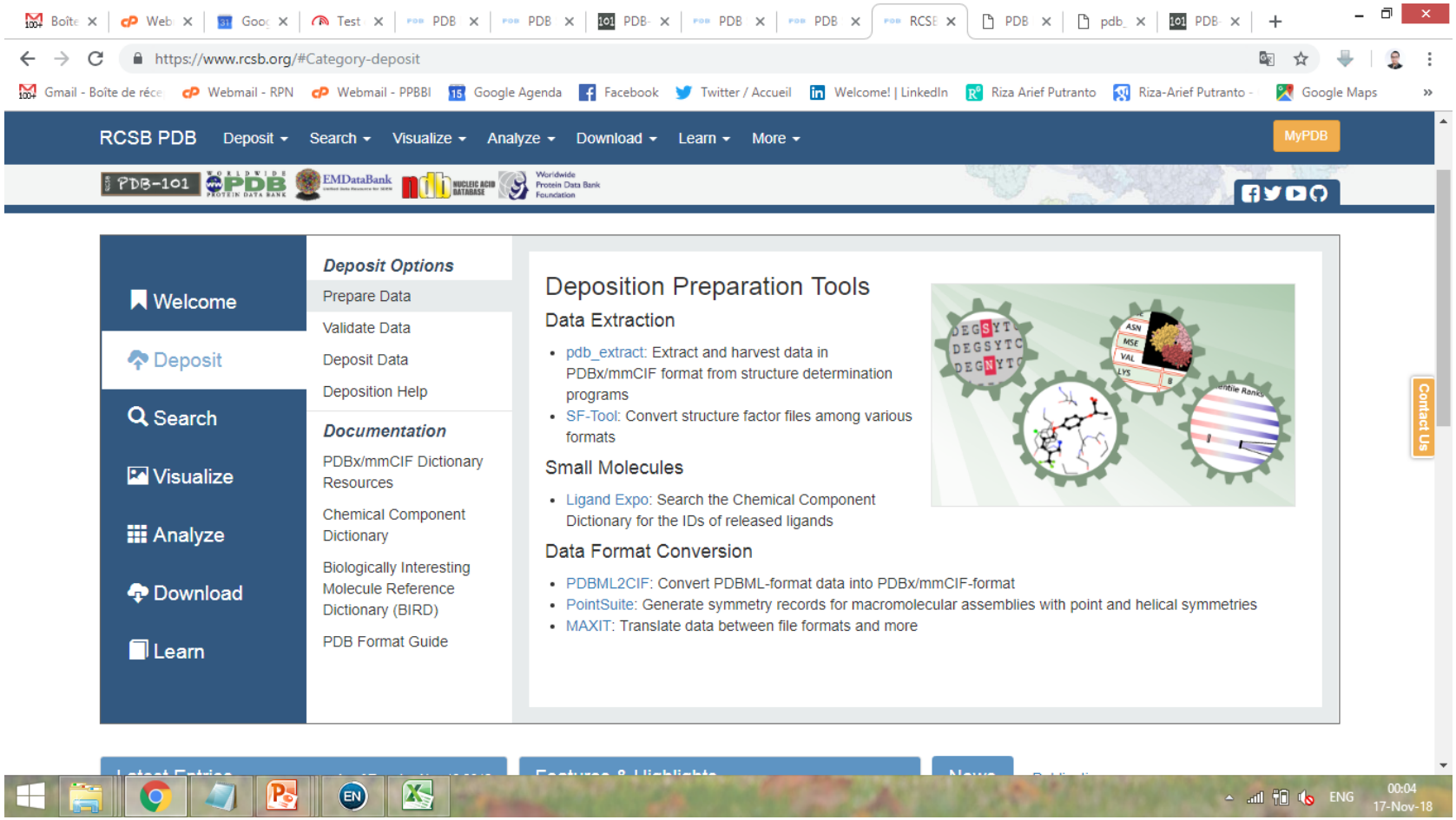
Once approval is received from the author, the entry and the tables in the internal core database are ready for distribution. The schema of this core database is a subset of the conceptual schema specified by the mmCIF dictionary. All aspects of data processing, including communications with the author, are recorded and stored in the correspondence archive. This makes it possible for the PDB staff to retrieve information about any aspect of the deposition process and to closely monitor the efficiency of PDB operations.




Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). **The Protein Data Bank**. *Nucleic Acids Res.* **28**, 235-242

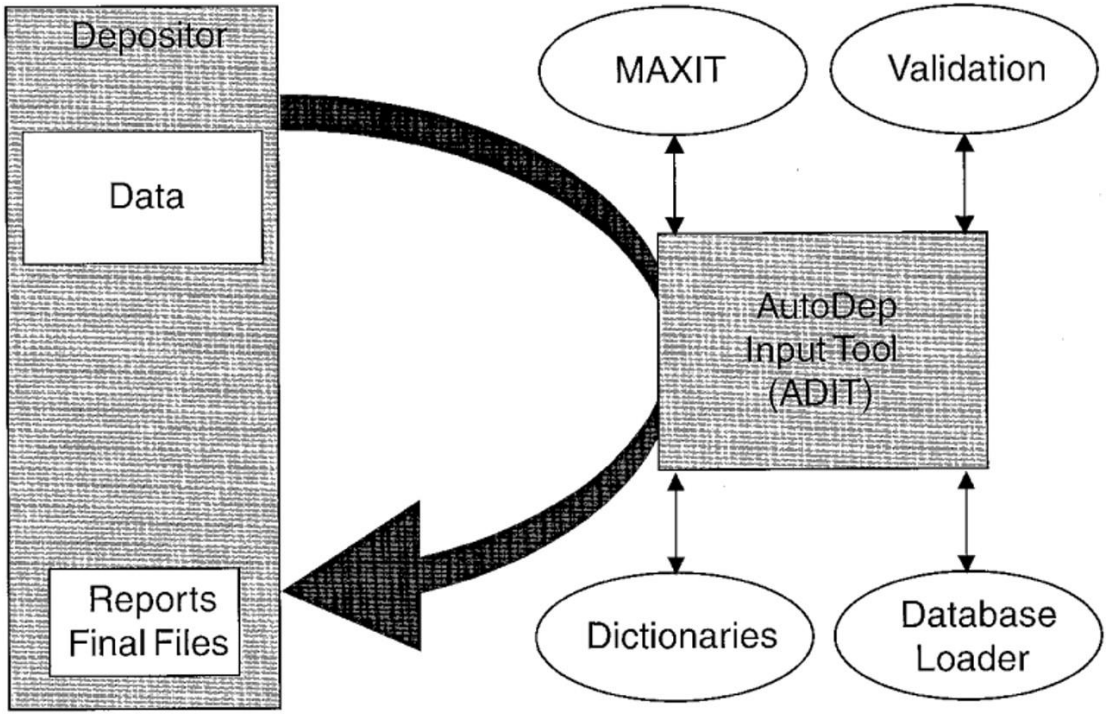
# Data submission

## https://www.rcsb.org/#Category-deposit



# Data submission

ADIT, which is also used to process the entries, is built on top of the mmCIF dictionary which is an ontology of 1700 terms that define the macromolecular structure and the crystallographic experiment, and a data processing program called MAXIT (MAcromolecular EXchange Input Tool). This integrated system helps to ensure that the data submitted are consistent with the mmCIF dictionary which defines data types, enumerates ranges of allowable values where possible and describes allowable relationships between data values.




Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). **The Protein Data Bank**. *Nucleic Acids Res.* **28**, 235-242

# Crystallographic Information File CIF & Self Defining Text Archive and Retrieval (STAR)

**Crystallographic Information File (CIF)** is a data representation used by several disciplines (predominantly crystallography) concerned with **molecular structure**. The basis for this data representation is the Self Defining Text Archive and Retrieval (STAR) definition.

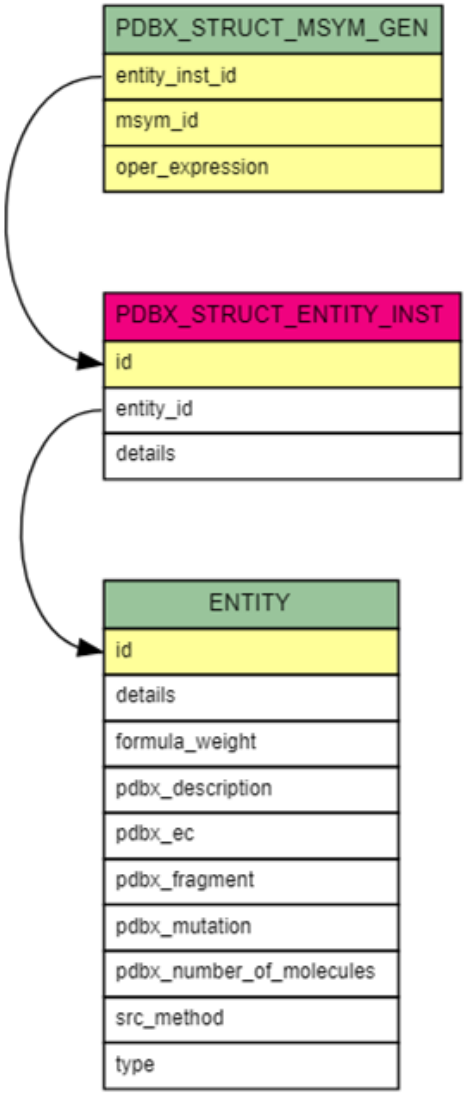
STAR is nothing more than a set of syntax rules. Associated with STAR is a Dictionary Definition Language (DDL) from which STAR compliant dictionaries have been developed by several discipline. From the dictionaries it is possible to define data files which use data items referenced in the dictionaries. The STAR DDL and associated dictionaries is considered as example of metadata - data describing how to represent other data.

•Westbrook, J. D. and Bourne, P. E. (2000). **STAR/mmCIF: an ontology for macromolecular structure**. *Bioinformatics.* **16**, 159-168.

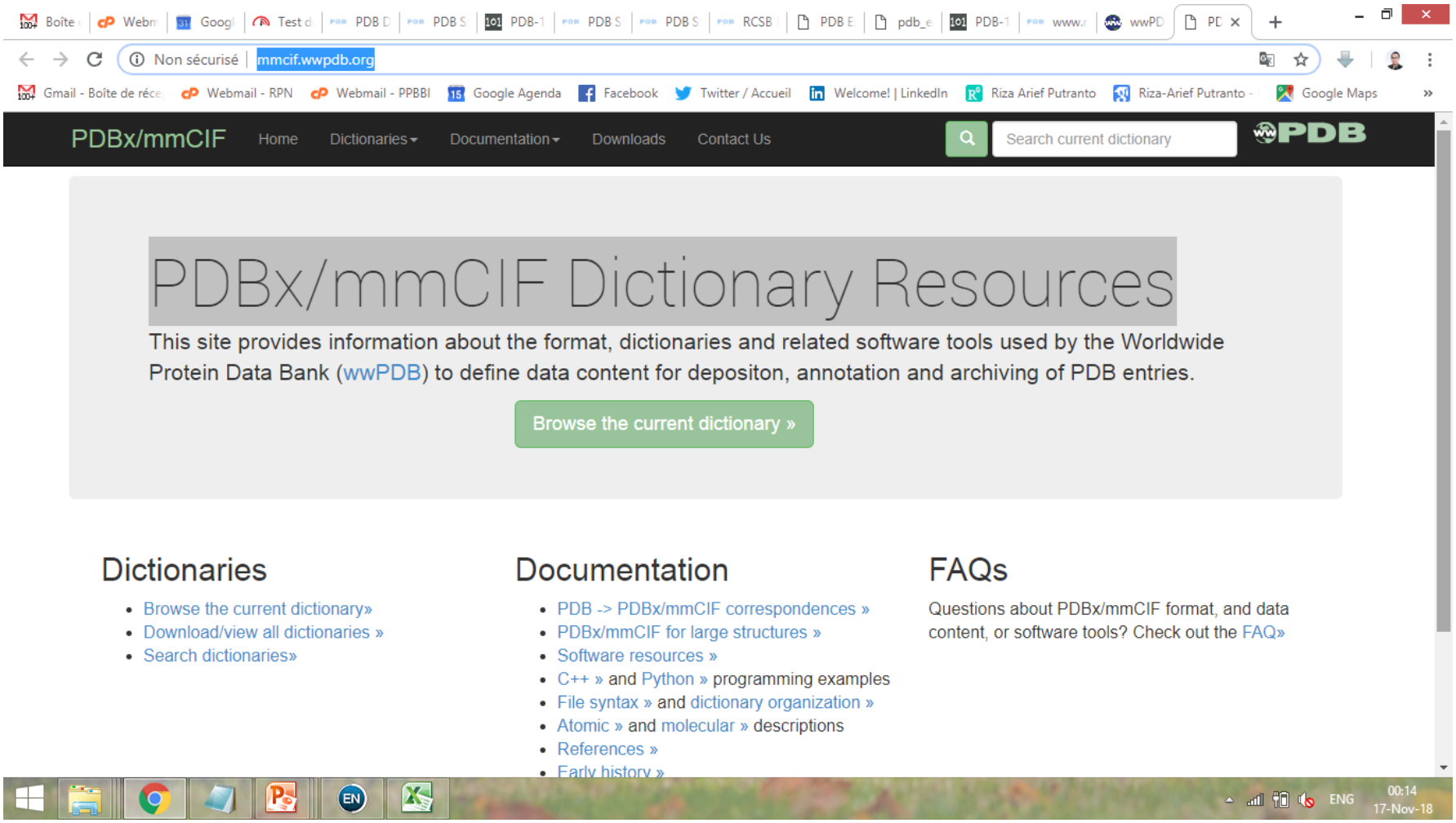
# PDBx/mmCIF Dictionary Resources
# http://mmcif.wwpdb.org/

**The PDBx/mmCIF** is a dictionary of definitions which describes a language for specifying data definitions. DDL defines the data model that provides the foundation for the description of knowledge about an application domain.

The application knowledge is collected in a dictionary of definitions which describes the domain. DDL provides the framework on which this dictionary is organized by defining the levels of abstraction that are available to hold the data description. The DDL defines both the properties that may be associated with each level of abstraction and the relationships that may exist between levels. This DDL defines a relatively simple set of abstractions which include data blocks, categories, category groups, subcategories, and items.



# http://mmcif.wwpdb.org/



# Data validation

**Validation** refers to the procedure for assessing the quality of deposited atomic models (structure validation) and for assessing how well these models fit the experimental data (experimental validation). The PDB validates structures using accepted community standards as part of ADIT's integrated data processing system.

**Covalent bond distances and angles.** Proteins are compared against standard values from Engh and Huber; nucleic acid bases are compared against standard values from Clowney *et al*; sugar and phosphates are compared against standard values from Gelbin *et al.*

**Stereochemical validation.** All chiral centers of proteins and nucleic acids are checked for correct stereochemistry.

**Atom nomenclature.** The nomenclature of all atoms is checked for compliance with IUPAC standards and is adjusted if necessary.

**Close contacts.** The distances between all atoms within the asymmetric unit of crystal structures and the unique molecule of NMR structures are calculated. For crystal structures, contacts between symmetry-related molecules are checked as well.

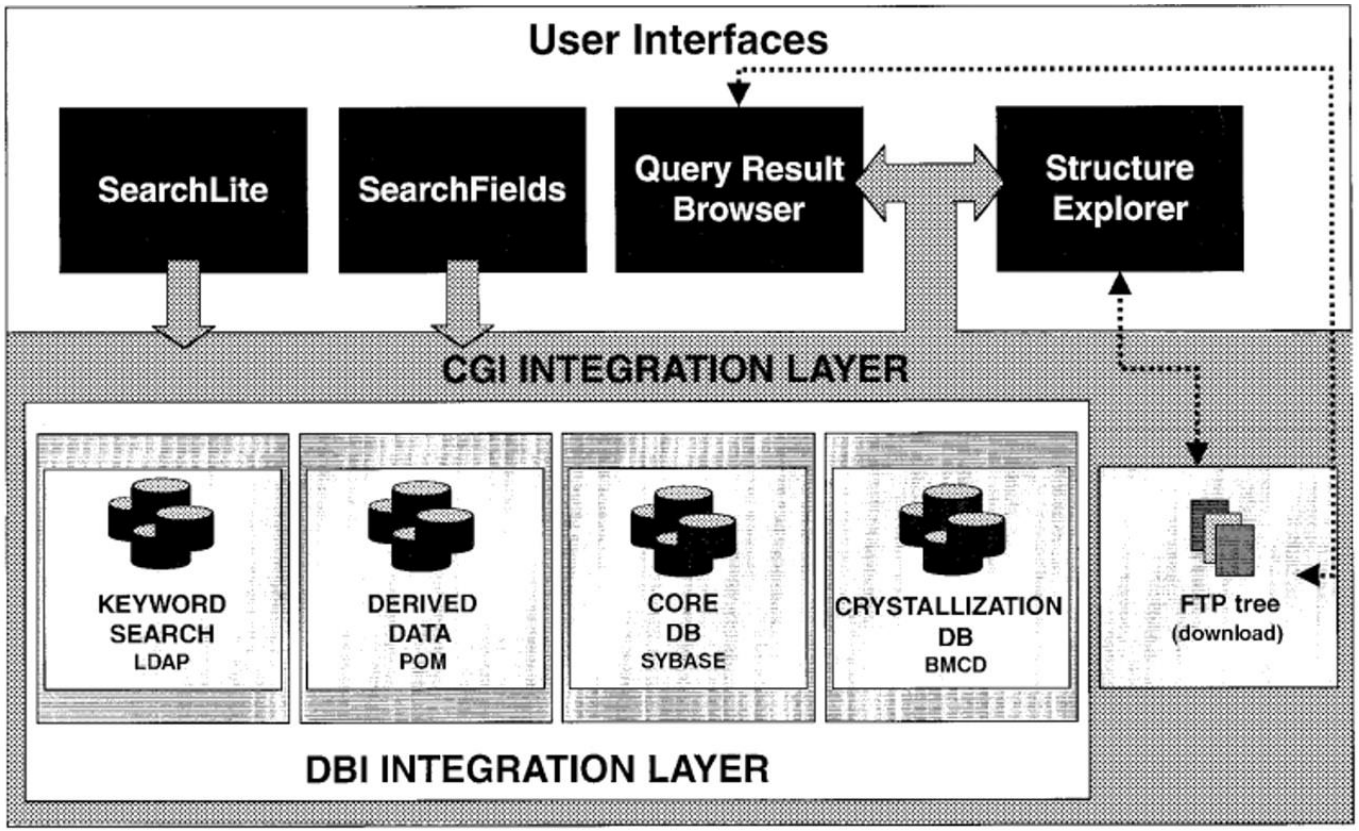
**Ligand and atom nomenclature.** Residue and atom nomenclature is compared against the PDB dictionary (ftp://ftp.rcsb. org/pub/pdb/data/monomers/het_dictionary.txt ) for all ligands as well as standard residues and bases. Unrecognised ligand groups are flagged and any discrepancies in known ligands are listed as extra or missing atoms.

**Sequence comparison.** The sequence given in the PDB SEQRES records is compared against the sequence derived from the coordinate records. This information is displayed in a table where any differences or missing residues are marked. During structure processing, the sequence database references given by DBREF and SEQADV are checked for accuracy. If no reference is given, a BLAST search is used to find the best match. Any conflict between the PDB SEQRES records and the sequence derived from the coordinate records is resolved by comparison with various sequence databases.

**Distant waters.** The distances between all water oxygen atoms and all polar atoms (oxygen and nitrogen) of the macromolecules, ligands and solvent in the asymmetric unit are calculated. Distant solvent atoms are repositioned using crystallographic symmetry such that they fall within the solvation sphere of the macromolecule.

# Database architecture

In recognition of the fact that no single architecture can fully express and efficiently make available the information content of the PDB, an integrated system of heterogeneous databases has been created that store and organize the structural data. At present there are five major components




Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). **The Protein Data Bank**. *Nucleic Acids Res.* **28**, 235-242

# Database architecture

The core relational database managed by Sybase (Sybase SQL server release 11.0, Emeryville, CA) provides the central physical storage for the primary experimental and coordinate data

The final curated data files (in PDB and mmCIF formats) and data dictionaries are the archival data and are present as ASCII files in the ftp archive.

• **The POM (Property Object Model)-based databases**, which consist of indexed objects containing native (e.g., atomic coordinates) and derived properties (e.g., calculated secondary structure assignments and property profiles). Some properties require no derivation, for example, B factors; others must be derived, for example, exposure of each amino acid residue or C contact maps. Properties requiring significant computation time, such as structure neighbours, are pre-calculated when the database is incremented to save considerable user access time.

• **The Biological Macromolecule Crystallization Database (BMCD;)** is organized as a relational database within Sybase and contains three general categories of literature derived information: macromolecular, crystal and summary data.

• **The Netscape LDAP server** is used to index the textual content of the PDB in a structured format and provides support for keyword searches.

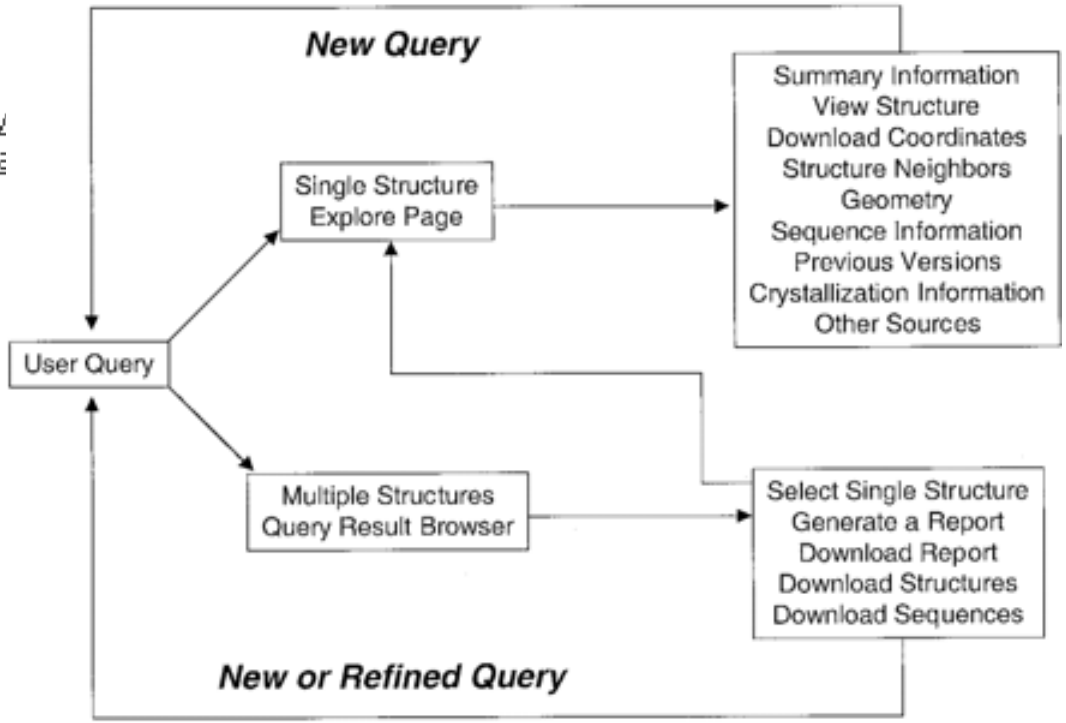
# Database query

Three distinct query interfaces are available for the query of data within PDB:

**Status Query** (**http://www.rcsb.org/pdb/status.html**)

SearchLite (http://www.rcsb.org/pdb/searchlite.html )

SearchFields (http://www.rscb.org/pdb/queryForm.cgi )


Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). **The Protein Data Bank**. *Nucleic Acids Res.* **28**, 235-242

# Search Methods

- The search tools can be accessed from the PDB home page. The types of possible searches are:

**1. By providing a PDB identification code (PDB ID).**

Each structure in the PDB is represented by a 4 character alphanumeric identifier, assigned upon its deposition. For example, 4hhb and 9ins are identification codes for PDB entries for hemoglobin and insulin, respectively. Many of the PDB Web site pages, including the PDB home page, allow you to enter a PDB ID and retrieve information for the corresponding structure

**2. By searching the text of both mmCIF files and the Web pages(QuickSearch).**

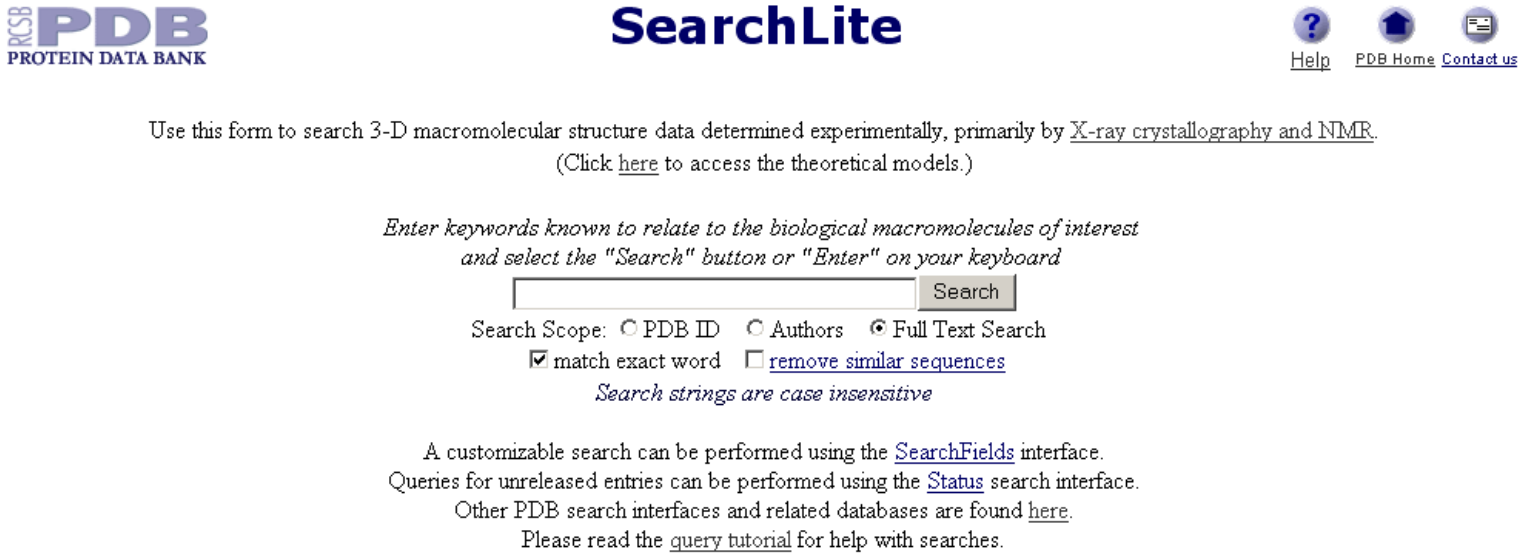
QuickSearch allows to simultaneously search the text of mmCIF files and the Web pages. It supports the same search syntax as the SearchLite search. An 'Exact Word Match' and 'Full Text' search is performed on an index of the mmCIF files and an index of the static PDB Web pages. The structures returned by the search can be browsed, refined and explored using the Query Result Browser and Structure Explorer. The static page results are listed as links and displayed with the keyword highlighted in the context in which it appears.

# Search Methods

**3. By searching the text found in mmCIF files (SearchLite).**

SearchLite searches the text of each mmCIF file as follows:
Queries locate literal text phrases. A search for *protein kinase* will locate the phrase *protein kinase*, NOT *protein* and *kinase* separately.

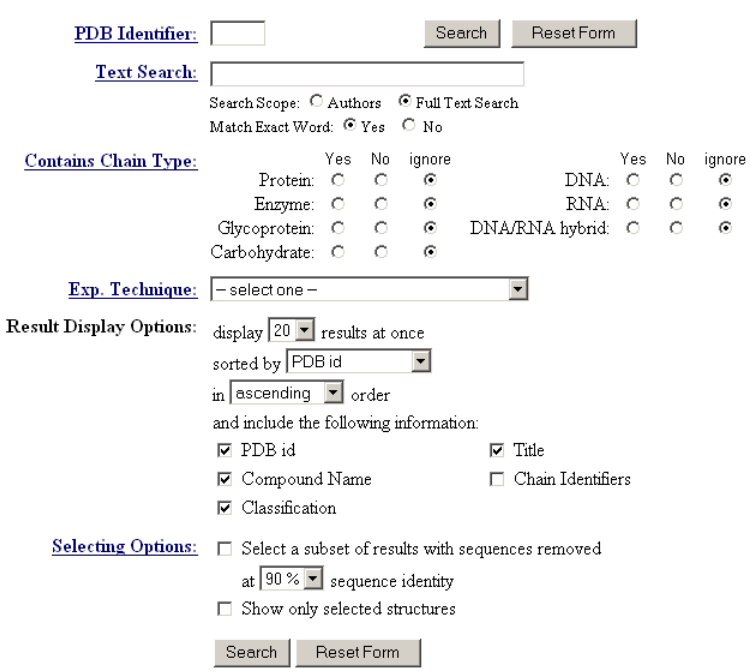
- Partial word searches will retrieve all words they are included in, unless the *match exact word*box is checked. A search for *hend* will locate both *hendrickson* and *henderson* when the box is not checked, but will only retrieve *hend* when the box is checked.
- A second checkbox allows a user to remove sequence homologs from a search.
- Compound searches can be performed using *and, or, not* clauses. A search for *protein and kinase* will locate all structures that contain both *protein* AND *kinase*, not just the structures that contain the phrase *protein kinase*.
- SearchLite will locate entries with an "on hold" status by querying their title records. For queries on unreleased entries specifically, a Status Search is most optimal.

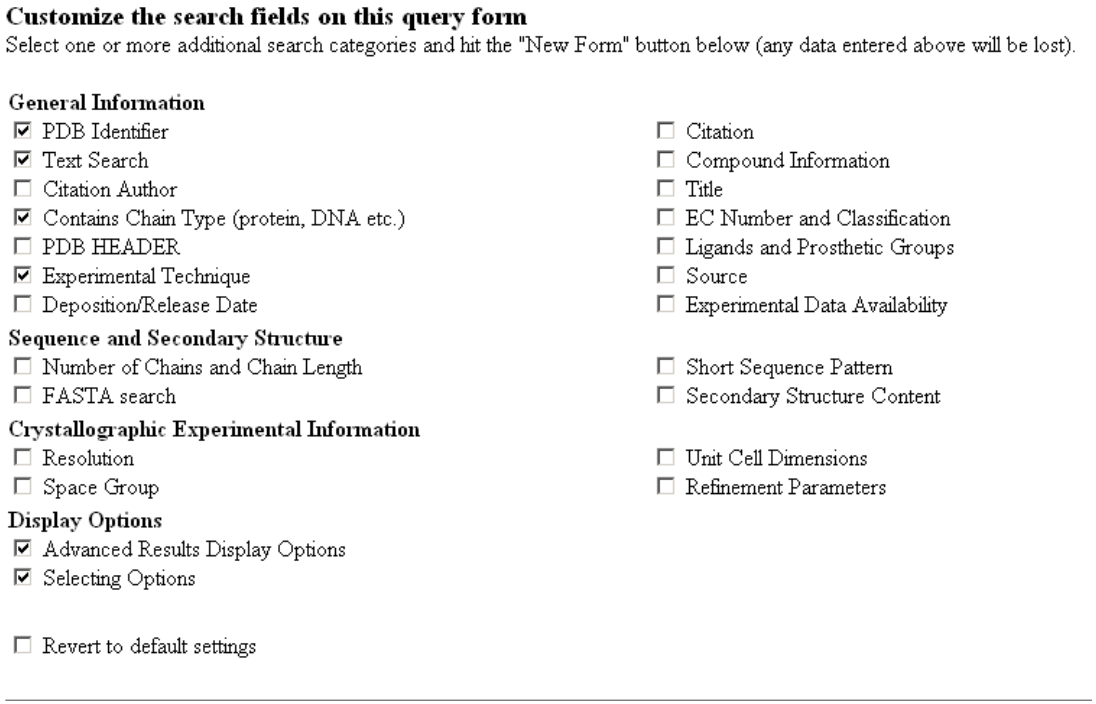


# Search Methods

**4. By searching against specific fields of information - for example, deposition date or author (SearchFields).**

SearchFields supports queries on specific attributes of a structure, such as its author, sequence, or deposition date. Additional search fields can be added or removed from the default form by selecting new fields from choices provided at the bottom of the page, and pressing the New Form button. If multiple fields are used for a search, a list of structures meeting **all** of the specified field requirements is returned. The format of the results can be customized using the options at the bottom of the search interface page.



# Search Methods

**5. By searching on the status of an entry, on hold or released (Status Search).**

To check on the status and obtain summary information on an unreleased entry, use the Status Search link from the PDB home page.
Queries can be performed based on PDB ID, author, title, release date, or deposition date. You may also search based on the holding status of the unpublished entries. Status categories are:

- release on publication - entry will be released when the associated journal article is published (HPUB)
- release on certain date - entry will be released on a date specified by the authors at the time of deposition (HOLD)
- await author input - entry is being processed but requires further interaction between the processor and the depositor (WAIT)
- currently being processed - entry is still being processed (PROC/PROCESSING)
- deposition withdrawn (WDRN)

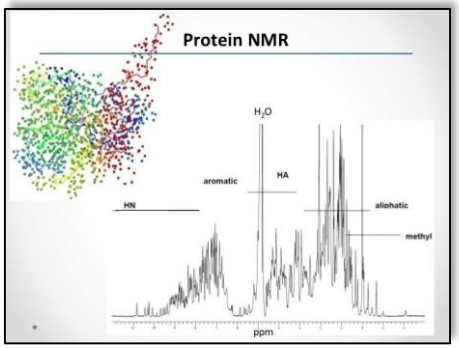
**6. By iterating on a previous search.**

From a list of structures returned from an initial search, the user can select all structures by choosing that option from the pull-down menu, or select a subset of structures by checking the boxes next to them. Additional searches can be performed over the entire or partial result list. Select the Refine Your Query option from the pull down menu at the top of the Query Result Browser, which will return you to the search interface which was used for your initial query.

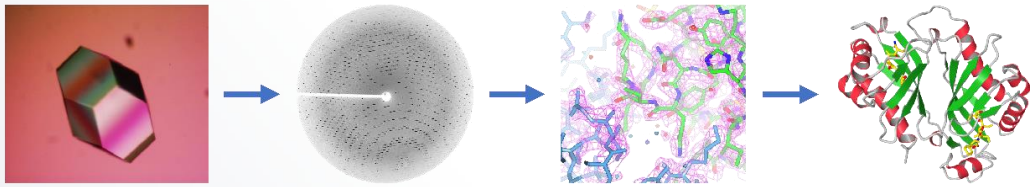
# Protein structure gap

**Experimental protein structure solution** (eg. by **NMR** or **X-Ray crystallography**) is **labor** intensive and **expensive**.

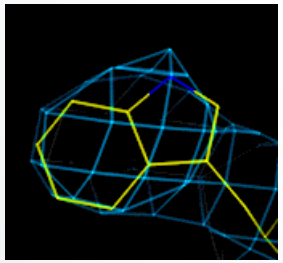
For the majority of proteins in any given proteome, experimental structures are not available.



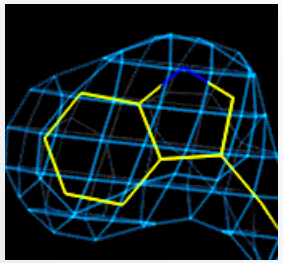

1. Is it possible to **predict** 3-dimensional protein structures **computationally**?
2. Which computational methods are **feasible** and applicable in a life science research context?



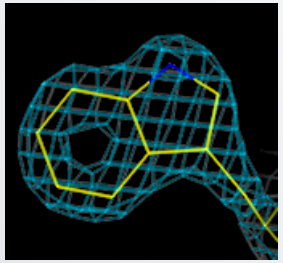
# It is all about resolution



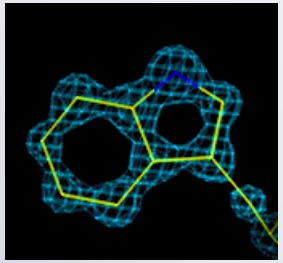
4 Å



3 Å



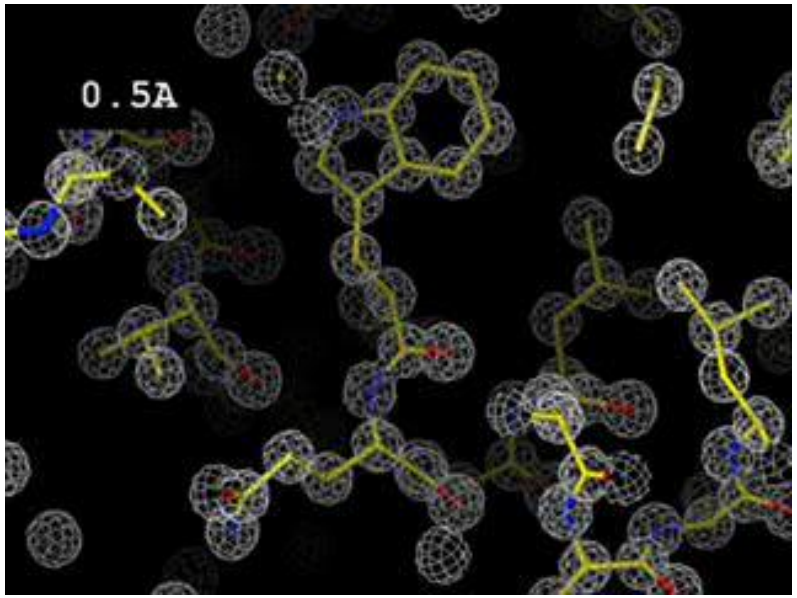
2 Å

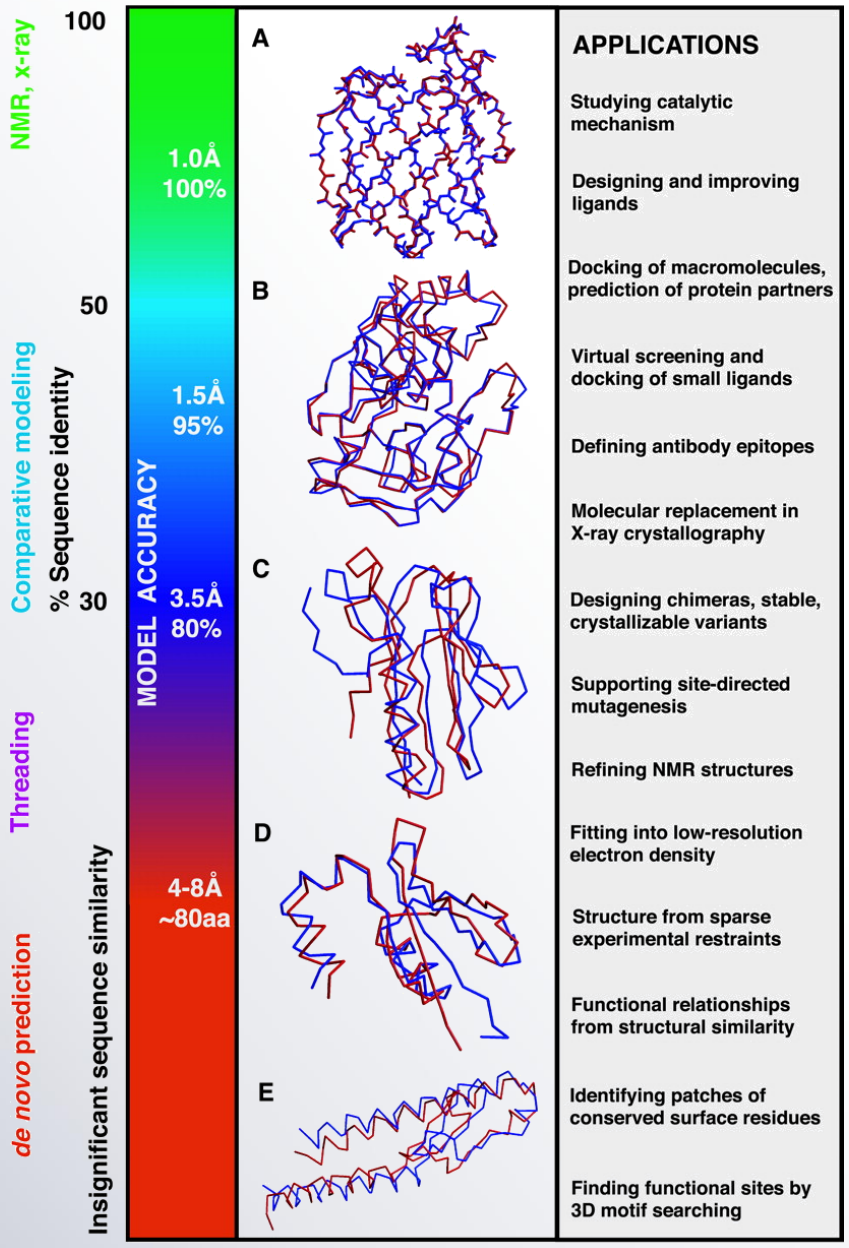


1 Å

low

high




NMR, x-ray
Comparative modeling
Threading
de novo prediction
100
50
30
% Sequence identity
Insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4-8Å
~80aa
A
B
C
D
E
APPLICATIONS
Studying catalytic mechanism
Designing and improving ligands
Docking of macromolecules, prediction of protein partners
Virtual screening and docking of small ligands
Defining antibody epitopes
Molecular replacement in X-ray crystallography
Designing chimeras, stable, crystallizable variants
Supporting site-directed mutagenesis
Refining NMR structures
Fitting into low-resolution electron density
Structure from sparse experimental restraints
Functional relationships from structural similarity
Identifying patches of conserved surface residues
Finding functional sites by 3D motif searching


Jmol

# Quick test: How many monomers?

# It was still the eighth course, don't get dizzy yet