SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOINFORMATICS

UNIT – I Structural Bioinformatics – SBI1403

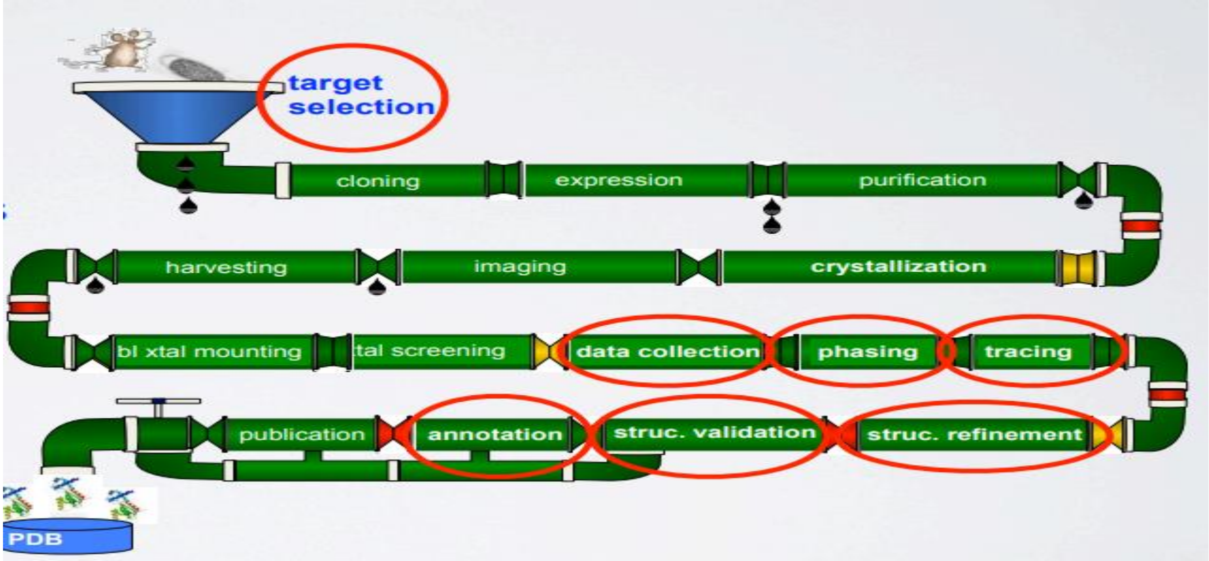
INTRODUCTION TO STRUCTURAL BIOINFORMATICS

Introduction

Structural bioinformatics is the branch of bioinformatics that is related to the analysis and prediction of the three-dimensional structure of biological macromolecules such as proteins, RNA, and DNA.
Structural Bioinformatics was the first major effort to show the application of the principles and basic knowledge of the larger field of bioinformatics to questions focusing on macromolecular structure, such as the prediction of protein structure and how proteins carry out cellular functions, and how the application of bioinformatics to these life science issues can improve healthcare by accelerating drug discovery and development. Designed primarily as a reference, the first edition nevertheless saw widespread use as a textbook in graduate and undergraduate university courses dealing with the theories and associated algorithms, resources, and tools used in the analysis, prediction, and theoretical underpinnings of DNA, RNA, and proteins.

Structural biology, determining the three-dimensional shapes of biomacromolecules and their complexes, can tell us a lot about how these molecules function and the roles they play within a cell. Bioinformatics data derived from structure determination experiments enables life-science researchers to address a wide variety of questions. For example, it aids the understanding of how mutations in a gene might alter a protein’s shape, disrupt a catalytic site, or alter the binding affinity of a pharmaceutical compound.

THE HOLY TRINITY OF STRUCTURAL BIOINFORMATICS
Sequence > Structure > Function
Motivation 1:
Detailed understanding of molecular interactions Provides an invaluable structural context for conservation and mechanistic analysis leading to functional insight
Motivation 2: Lots of structural data is becoming available
Data from: http://www.rcsb.org/pdb/statistics/ Structural Genomics has contributed to driving down the cost and time required for structural determination



Sequence > Structure > Function

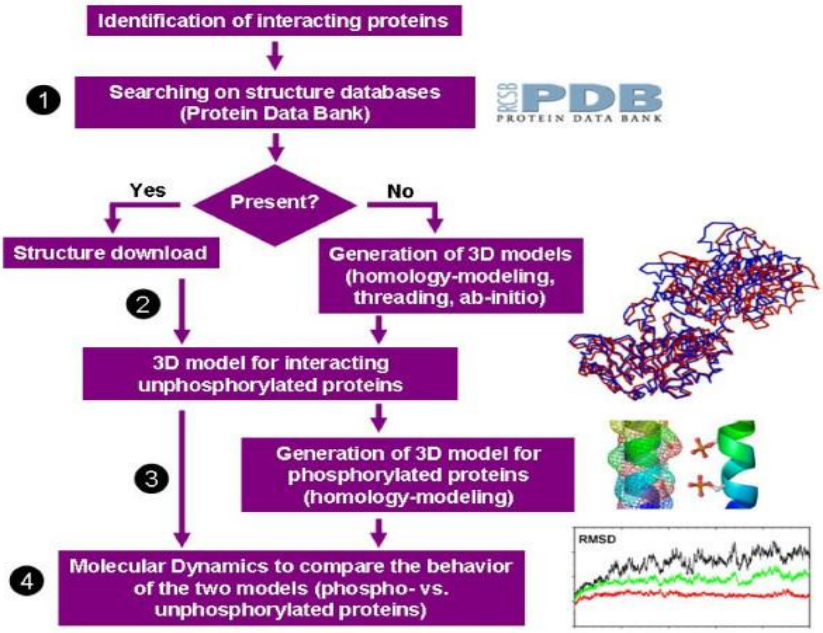
• Structure determines function, so understanding structure helps our understanding of function

Structure is more conserved than sequence

• Structure allows identification of more distant evolutionary relationships

Structure is encoded in sequence

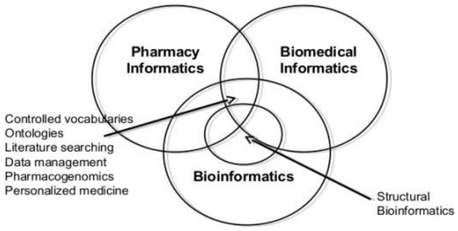
• Understanding the determinants of structure allows design and manipulation of proteins for industrial and medical advantage


Identification of interacting proteins
1
Searching on structure databases (Protein Data Bank)
RCSB PDB PROTEIN DATA BANK
Yes
Present?
No
Structure download
Generation of 3D models (homology-modeling, threading, ab-initio)
2
3D model for interacting unphosphorylated proteins
3
Generation of 3D model for phosphorylated proteins (homology-modeling)
4
Molecular Dynamics to compare the behavior of the two models (phospho- vs. unphosphorylated proteins)
RMSD

Structural Bioinformatics

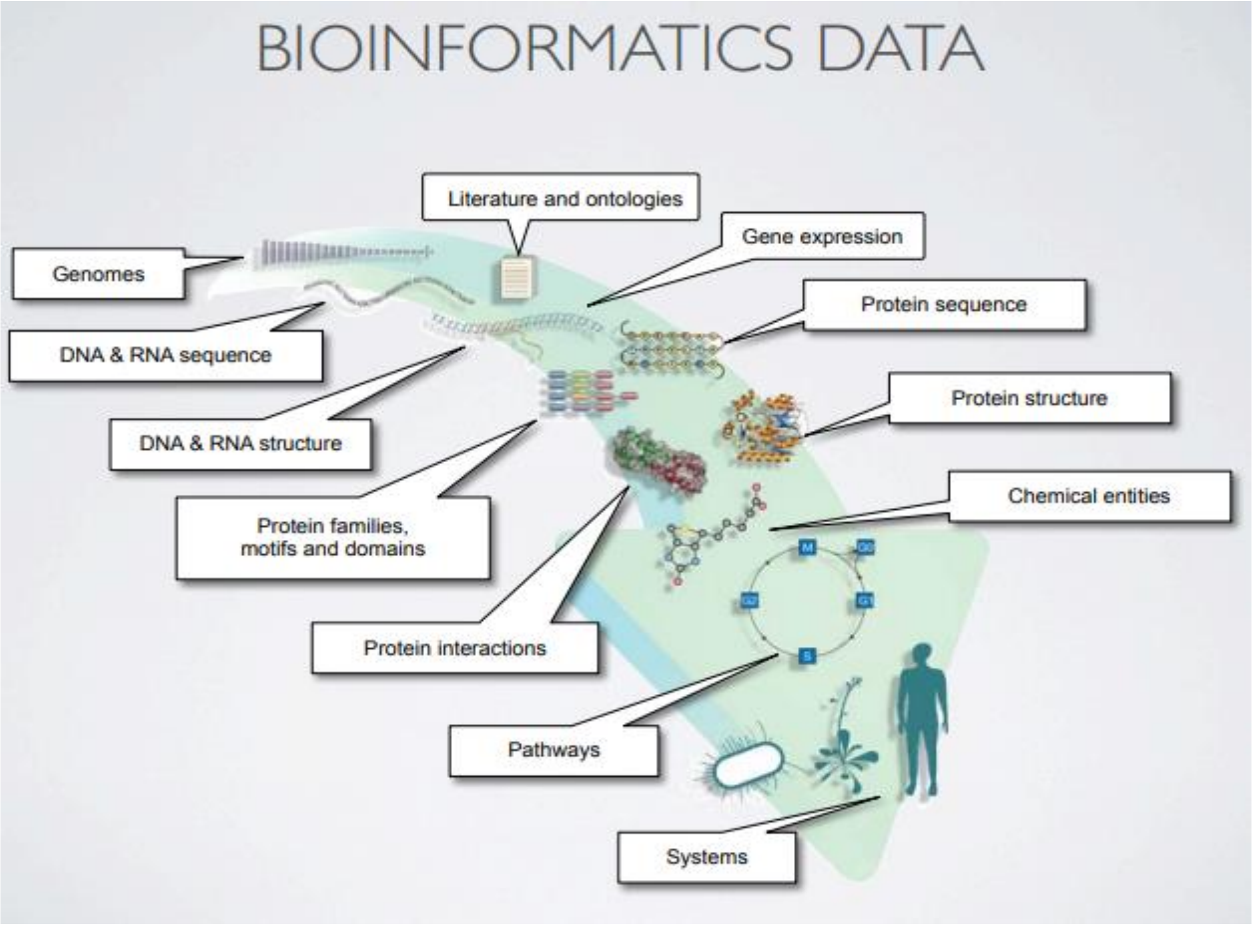
Structural Bioinformatics is an interdisciplinary field that deals with the three dimensional structures of biomolecules. It attempts to model and discover the basic principles underlying biological machinery at the molecular level. It is based on the assumption that 3D structural information of a biological system is the core to understanding its mechanism of action and function. Structural bioinformatics combines applications of physical and chemical principles with algorithms from computational science.

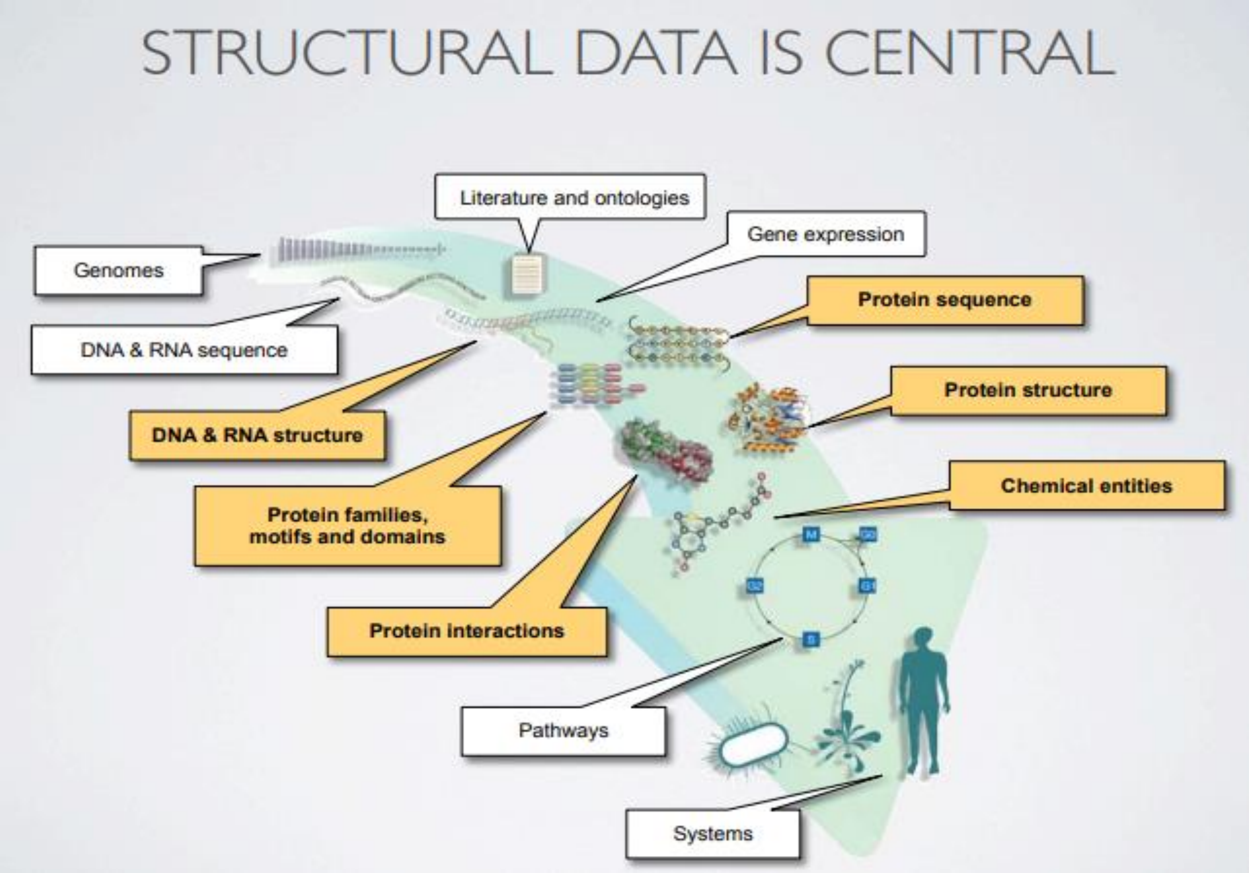
Putting Structural Bioinformatics in Perspective



Major areas:

- protein and nucleic acid 3D structure determination
- prediction of protein 3D structure from sequence
- protein structure validation
- protein structure comparison and alignment
- protein and nucleic acid structure classification
- inferring protein function from structure
- prediction of protein-ligand interaction
- prediction of protein-protein interactions
- development of databases


BIOINFORMATICS DATA



STRUCTURAL DATA IS CENTRAL

Literature and ontologies

Gene expression

Genomes

Protein sequence

DNA & RNA sequence

Protein structure

DNA & RNA structure

Chemical entities

Protein families, motifs and domains

Protein interactions

Pathways

Systems

Given the structural information created by efforts in X-ray crystallography and NMR, there are a wide range of analytic and scientific challenges to informatics. It is not possible to cover the full scope of activities, but they can be reviewed briefly to show the richness of opportunities in the analysis of structural data.

Visualization:

- ❑ The creation of images of molecular structure remains a primary activity within structural biology .
- ❑ The complexity of these molecules seems to demand novel display methods that are able to combine structural information with other information sources (such as electrostatic fields, the location of functional sites, and areas of structural or genetic variability).
- ❑ The issues for informatics include the creation of flexible software infrastructures for extending display capabilities, and the use of novel methods for rapidly rendering complex molecular structures (Huang et al., 1996; Sanner et al., 1999).

Classification:

- ❑ The database of known structures is already sufficiently large that it is necessary to cluster similar structures together, in order to form families of proteins.
- ❑ These families are often aggregated into superfamilies, and indeed entire structural hierarchies have been created.
- ❑ The Structural Classification of Proteins (SCOP) is an example of a semiautomated classification of all protein structures (Murzin et al., 1995), and there have been numerous efforts to create automated classification—usually based on the pairwise comparison of all structures to create a matrix of distances (Holm and Sander, 1996; Orengo et al., 1997).
- ❑

Prediction:

- Despite the growth of the structural databases, the number of known 3D structures has lagged far behind the availability of sequence information. Thus, the prediction of 3D structure remains an area of keen interest. The Critical Assessment for Structure Prediction (CASP) meetings have provided a biennial forum for the comparison of methods for structure prediction.

- The main categories of prediction have been homology modeling (based on high sequence homology to a known structure (Sanchez and Sali, 1997), threading (based on remote sequence homology) ´ (Bryant and Altschul, 1995), and ab initio prediction (based on no detectable homology (Osguthorpe, 2000).

- The diversity of methods invented and evaluated is quite inspiring, and the resulting lessons about how proteins are put together have been significant.

- **Simulation**
- The results of crystallographic studies (and to some extent, NMR studies) are primarily static structural models. However, the properties of these molecules that are of the greatest interest are often the result of their dynamic motions.
- The definition of energy functions that govern the folding of proteins and their subsequent stable dynamics has been an area of great interest since the first structure was determined.
- Unfortunately, the time scales on which macromolecular dynamics must be sampled (fractions of picoseconds) are much shorter than the time scale on which biologically important phenomena occur (microseconds to seconds).
- Nevertheless, the **availability of increasingly powerful computers and the clever approximation and search methods are enabling molecular simulations of sufficient length and accuracy to emerge, and are making contributions to our understanding of protein function**.
- The associated computation of electrostatic fields of macromolecular structures has emerged as an important component of understanding molecular function.
- **Toward a High-Resolution Understanding of Biology**
- The great **promise of structural bioinformatics is predicated on the belief that the availability of high resolution structural information about biological systems** will allow us to reason precisely about the function of these systems and the affects of modifications or perturbations.
- **Whereas genetic analyses can only associate genetic sequences with their functional consequences, structural biological analyses offer the additional promise of ultimate insight into the mechanisms of these consequences**, and therefore a more profound understanding of how biological function follows from structure.
- **The promise for structural bioinformatics lies in four areas:**
- **(1) creating an infrastructure for building up structural models from component parts;**
- **(2) gaining the ability to understand the design principles of proteins so that new functionalities can be created;**
- **(3) learning how to design drugs efficiently based on structural knowledge of their target; and**

- **(4) catalyzing the development of simulation models that can give insight into function based on structural simulations**.
- Each of these four areas has already seen success, and the structural genomics projects promise to create data sets sufficient to catalyze accelerated progress in all these areas.
-
- **Special Challenges in Computing with Structural Data**
- Structural bioinformatics **must overcome some special challenges that are either not present or not dominant in other types of bioinformatics domains (such as the analysis of sequence or microarray data).**
- It is important to remember these challenges when assessing the opportunities in the field.
- **They include:**
- • **Structural data is not linear and therefore is not easily amenable to algorithms based on strings**. In addition to **this obvious nonlinearity, there are also nonlinear relationships between atoms** (the forces are not linear), which means that most **computations on structure need either to make approximations or to be very expensive.**
- • The **search space for most structural problems is continuous**. Structures are represented generally by **atomic Cartesian coordinates (or internal angular coordinates) that are continuous variables**. Thus, there are infinite search spaces for algorithms attempting to assign atomic coordinate values. Many simplifications can be applied, such as **lattice models for 3D structure (Hinds and Levitt, 1994),** but these are attempts to manage the inherent continuous nature of these problems
- There is a fundamental connection **between molecular structure and physics**.
- Although this statement seems obvious and trivial, it means that when reduced representations, such as pseudo atoms (Wuthrich et al., 1983) or lattice models are applied, they become more difficult to relate to the underlying physics that govern the interactions.
- The need to keep structural calculations physically reasonable is an important constraint.
- Reasoning **about structure requires visualization**.Creation of computer graphics was driven, in part, by the need of structural biologists to look at molecules .
- This visualization is both a benefit and a detriment: Structure is well defined and well-designed visualizations can provide **insight into structural problems**. However, **graphic displays have a human user as a target and are not easily parsed or understood by computers**, and thus represent something of a computational "dead end."
- The need to have expressive data structures underlying these visualizations allows the information to be understood and analyzed by computer programs, and thus opens the possibility of further downstream analysis
- Structural data, like all biological data, can be noisy and imperfect.
- Despite some amazing successes in the elucidation of very high-resolution structures, the precision of our knowledge about many structures is likely to be limited by their flexibility, dynamics, or experimental noise.

- Understanding the **protein structural disorder may be critical for understanding the protein's function**.
- Thus, we must be comfortable reasoning about structures about which we have only partial knowledge.
- Protein and nucleic acid structures are generally conserved more than their associated sequence.
- Thus, sequences will accumulate mutations over time that may make identification of their similarities more difficult, while their structures may remain essentially identical.
- However, sequence information is still much more abundant than structural information, and so for **many molecules it is the sequence information that is readily available.**
- Thus, the **need to identify distant sequential similarities in order to gain structural insights can be a major challenge**
- Finally, we must recognize that there is a major gap in our knowledge of a large fraction of proteins that are not globular and water soluble.
- In particular, membrane-bound and fibrous proteins are simply not well understood and structures are not available in the numbers required to allow routine statistical and informatics approaches to their study.
- The importance of this shortcoming cannot be overemphasized, since these classes of proteins are **among the most important for understanding a large number of cellular processes of great interest, including signal transduction, cytoskeletal dynamics, and cellular localizations and compartmentalization.**
- **Target Selection.**
- Structural genomics efforts with finite resources must carefully select proteins to study.
- Informatics methods are used to compare the database of existing structures and known sequences with potential targets in order to identify those that are most likely to add to our structural knowledge base.
- This **selection can be informed by the expected novelty of the structure, and even its importance as reflected in the published literature.**
- A critical **part of target selection is the identification of domains within large proteins**.
- Domains are often **easier to study initially in isolation, and then to study in complexes**. The definition of domains from sequence data alone is a challenging problem.

- ❑ **Tracking Experimental Crystallization Trials**.
- ❑ One of the major bottlenecks in structural genomics is the discovery of crystallization conditions that work for proteins of interest.
- ❑ In addition to the obvious need for storing and tracking information on the proteins, the conditions attempted, and the results, there is also an opportunity to apply machine-learning methods to these data in order to extract rules that may help increase the yield of crystals based on previous experience
- ❑ Until recently the results of failed crystallization experiments were not generally available, making it **difficult to apply automated machine-learning methods to these data sets.**
- ❑ **Analysis of Crystallographic Data.**
- ❑ A **long-standing area of computation within structural biology are the algorithms for deconvoluting the X-ray diffraction pattern**, which involves

computing an inverse Fourier transform with partial information (i.e., with missing phase information).

- ❑ There is **interest in ab initio methods for automating these computations**, and success in this area reduces the number of heavy atom derivatives that must be created for structures of interest.
- ❑ **Multiwavelength Anomalous Diffraction (MAD)** is now the **preferred method for solving the crystallographic phase problem.**
- ❑ Recent progress has been made **on automated electron density map fitting and refinement**.
- ❑ **Analysis of NMR Data.**
- ❑ **NMR experiments provide complementary data to the crystallographic analyses.**
- ❑ NMR experiments produce two-dimensional (or higher) spectra for which each individual peak must be assigned to an atomic interaction.
- ❑ The automated analysis and assignment of atoms in these spectra is a difficult search problem, but one in which progress has been made to accelerate the analysis of structure.

- ❑ Given a set of atomic proximities from NMR, we need methods to **"embed" these distance measures into 3D structures that satisfy these constraints**.
- ❑ Distance geometry, restrained molecular dynamics and other nonlinear optimization methods have been developed for this purpose.

Assessment and Evaluation of Structures.

- Given the results of a crystallographic or NMR structure determination effort, we must check the structures to be sure that they meet certain quality standards.
- Algorithms have been developed for assessing the basic chemistry of structural models, and also for identifying active sites and binding sites in these structures (Laskowski et al., 1993; Feng, Westbrook, and Berman, 1998; Vaguine, Richelle, and Wodak, 1999).
- Computational methods are still needed for automatically annotating 3D structures with functional information, based on an understanding of how molecular properties aggregate in three dimensions to produce function (such as binding, catalysis, motion, and signal transduction) (Wei, Huang, and Altman, 1999).

Storing Molecular Structures in Databases.

- The storage of the results of structural genomics efforts is an important task, requiring data structures and organizations that facilitate the most common queries.
- Ideally, databases of structure will store not only the resulting model, but also the raw data on which it is based.
- The PDB is the major repository for 3D structural information on proteins; the Nucleic Acids Database (NDB) serves this function for nucleic acids.
- There is also an effort to store the raw data associated with crystallography in the PDB/NDB and the raw data associated with NMR in the BioMagResBank (BMRB).

Correlating Molecular Structural Information with Structural and Functional Information Gained from Other Types of Experimentation.

- ❑ In the end, we perform structural studies in order to get an insight into how the molecules work.
- ❑ Structural studies with crystallography and NMR are but two methods that can be used to probe structure–function relationships.
- ❑ The integration of the results of these methods with other structural and functional data allows us to build comprehensive models of mechanism, specificity, and dynamics.

- ❑ A major bottleneck for using informatics methods for this integration is the lack of repositories of structural and functional data that can be accessed by computer programs doing systematic analyses.
- ❑ **One exception is the noncrystallographic structural data about the 30S and 50S ribosomal subunits stored in the RiboWEB knowledge base (http://riboweb.stanford.edu/).**
- ❑ RiboWEB is a knowledge base of ribosomal structural components that stores more than 8000 noncrystallographic structural and functional observations about the bacterial ribosome.
- ❑ It stores its information in structured "information templates" that are easily parsed by computer programs, thus making possible automated comparison and evaluation of structural models.
- ❑ For example, RiboWEB has been used to assess the compatibility of the published ribosomal crystal structures with over 1000 proximity measurements from cross-linking, chemical protection, and labeling experiments (collected during the last 25 years). Incompatibilities between these data and the crystal structures may suggest artifactual data or (more usefully) may suggest areas of important dynamic motion for the ribosome

INTEGRATING STRUCTURAL DATA WITH OTHER DATA SOURCES

- ❑ **Structural bioinformatics** has existed in some form or other ever since the determination of **the first myoglobin structure.**
- ❑ One could **argue that the roots go back further to the time when small molecular structure determination was introduced**.
- ❑ In any case, **the challenges for the field are clearly abundant and significant**.
- ❑ As we look to coming decades, it **appears that a primary challenge in structural bioinformatics will be the integration of structural information with other biological information to yield a higher resolution understanding of biological function**.
- ❑ The **success of genome-sequencing projects has created information about all the structures that are present in individual organisms, as well as both the shared and unique features of these organisms**.
- ❑ Even with the success **of structural genomics projects, bioinformatics techniques will probably be used to create homology models of most of these genomic components**.
- ❑ The resulting **structures will be studied with respect to how they interact and perform their functions**.
- ❑ Similarly, the emergence **of microarray expression measurements provides an ability to consider how the expression of macromolecular structures is regulated at a structural level** (including the key structural machinery associated with transcription, translation, and degradation).
- ❑ **Mass spectroscopic methods that allow the identification of structural modifications and variations** (such as genetic mutation or post-translational modifications) will need to be integrated with structural models to understand how they alter functional characteristics.
- ❑ **Finally, cellular localization data will allow us to place 3D molecular structures into compartments within the cell as we build more complex models of how cells are organized structurally in order to optimize their function**. This exciting activity will mark the next phase of structural bioinformatics—when the organization and physical structure of entire cells is understood and represented in computational models that provide insight into how thousands of structures within a cell work together to create the functions associated with life.

UNIT – II- Structural Bioinformatics – SBI1403

Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description

1. Introduction

The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of biological macromolecules that serves a global community of researchers, educators, and students. The data contained in the archive include atomic coordinates, crystallographic structure factors and NMR experimental data. Aside from coordinates, each deposition also includes the names of molecules, primary and secondary structure information, sequence database references, where appropriate, and ligand and biological assembly information, details about data collection and structure solution, and bibliographic citations.

Basic Notions of the Format Description

Character Set

Only non-control ASCII characters, as well as the space and end-of-line indicator, appear in a PDB coordinate entry file. Namely:

abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ 1234567890
` - = [ ] \ ; ' , . / ~ ! @ # $ % ^ & * ( ) _ + { } | : " < > ?

The use of punctuation characters in the place of alphanumeric characters is discouraged.

The space, and end-of-line:. The end-of-line indicator is system-specific character; some systems may use a carriage return followed by a line feed, others only a line-feed character.

Special Characters

Greek letters are spelled out, i.e., alpha, beta, gamma, etc. Bullets are represented as (DOT).
Right arrow is represented as -->. Left arrow is represented as <--.

If "=" is surrounded by at least one space on each side, then it is assumed to be an equal sign, e.g., 2 + 4 = 6.

Commas, colons, and semi-colons are used as list delimiters in records that have one of the following data types:

List SList
Specification List Specification
If a comma, colon, or semi-colon is used in any context other than as a delimiting character, then the character must be escaped, i.e., immediately preceded by a backslash, "\".

Example - Use of “\” character:

```
COMPND      MOL_ID:   1;
```

COMPND 2 MOLECULE: GLUTATHIONE SYNTHETASE; COMPND 3 CHAIN: A;
COMPND 4 SYNONYM: GAMMA-L-GLUTAMYL-L-CYSTEINE\:GLYCINE LIGASE COMPND 5 (ADP-FORMING);
COMPND 6 EC: 6.3.2.3;
COMPND 7 ENGINEERED: YES

COMPND MOL_ID: 1;
COMPND 2 MOLECULE: S-ADENOSYLMETHIONINE SYNTHETASE;
COMPND 3 CHAIN: A, B;
COMPND 4 SYNONYM: MAT, ATP\:L-METHIONINE S-ADENOSYLTRANSFERASE; COMPND 5 EC: 2.5.1.6;
COMPND 6 ENGINEERED: YES;
COMPND 7 BIOLOGICAL_UNIT: TETRAMER;
COMPND 8 OTHER_DETAILS: TETRAGONAL MODIFICATION

Record Format

Every PDB file is presented in a number of lines. Each line in the PDB entry file consists of 80 columns. The last character in each PDB entry should be an end-of- line indicator.

Each line in the PDB file is self-identifying. The first six columns of every line contains a record name, that is left-justified and separated by a blank. The record name must be an exact match to one of the stated record names in this format guide.

The PDB file may also be viewed as a collection of record types. Each record type consists of one or more lines.

Each record type is further divided into fields.

Each record type is detailed in this document. The description of each record type includes the following sections:

- Overview
- Record Format
- Details
- Verification/Validation/Value Authority Control
- Relationship to Other Record Types
- Examples
- Known Problems

For records that are fully described in fixed column format, columns not assigned to fields must be left blank.

Types of Records

It is possible to group records into categories based upon how often the record type appears in an entry.

One time, single line: There are records that may only appear one time and without continuations in a file. Listed alphabetically, these are:

RECORD TYPE	DESCRIPTION
CRYST1	Unit cell parameters, space group, and Z.
END	Last record in the file.
HEADER	First line of the entry, contains PDB ID code, classification, and date of deposition.
NUMMDL	Number of models.
MASTER	Control record for bookkeeping.
ORIGXn	Transformation from orthogonal coordinates to the submitted coordinates (n = 1, 2, or 3).
SCALEn	Transformation from orthogonal coordinates to fractional crystallographic coordinates (n = 1, 2, or 3).

It is an error for a duplicate of any of these records to appear in an entry.

One time, multiple lines: There are records that conceptually exist only once in an entry, but the information content may exceed the number of columns available. These records are therefore continued on subsequent lines. Listed alphabetically, these are:

RECORD TYPE	DESCRIPTION
AUTHOR	List of contributors.
CAVEAT	Severe error indicator.
COMPND	Description of macromolecular contents of the entry.
EXPDTA	Experimental technique used for the structure determination.
MDLTYP	Contains additional annotation pertinent to the coordinates presented in the entry.
KEYWDS	List of keywords describing the macromolecule.
OBSLTE	Statement that the entry has been removed from distribution and list of the ID code(s) which replaced it.
SOURCE	Biological source of macromolecules in the entry.

SPLIT List of PDB entries that compose a larger macromolecular complexes.

SPRSDE List of entries obsoleted from public release and replaced by current entry.

TITLE Description of the experiment represented in the entry.

The second and subsequent lines contain a continuation field, which is a right-justified integer. This number increments by one for each additional line of the record, and is followed by a blank character.

Multiple times, one line: Most record types appear multiple times, often in groups where the information is not logically concatenated but is presented in the form of a list. Many of these record types have a custom serialization that may be used not only to order the records, but also to connect to other record types. Listed alphabetically, these are:

RECORD TYPE	DESCRIPTION
ANISOU	Anisotropic temperature factors.
ATOM	Atomic coordinate records for standard groups.
CISPEP	Identification of peptide residues in cis conformation.
CONECT	Connectivity records.
DBREF	Reference to the entry in the sequence database(s).
HELIX	Identification of helical substructures.
HET	Identification of non-standard groups heterogens).
HETATM	Atomic coordinate records for heterogens.
LINK	Identification of inter-residue bonds.
MODRES	Identification of modifications to standard residues.
MTRIXn	Transformations expressing non-crystallographic symmetry (n = 1, 2, or 3). There may be multiple sets of these records.
REVDAT	Revision date and related information.
SEQADV	Identification of conflicts between PDB and the named sequence database.
SHEET	Identification of sheet substructures.
SSBOND	Identification of disulfide bonds.

Multiple times, multiple lines: There are records that conceptually exist multiple times in an entry, but the information content may exceed the number of columns available. These records are therefore continued on subsequent lines. Listed alphabetically, these are:

RECORD TYPE	DESCRIPTION
FORMUL	Chemical formula of non-standard groups.
HETNAM	Compound name of the heterogens.
HETSYN	Synonymous compound names for heterogens.
SEQRES	Primary sequence of backbone residues.
SITE	Identification of groups comprising important entity sites.

The second and subsequent lines contain a continuation field which is a right-justified integer. This number increments by one for each additional line of the record, and is followed by a blank character.

Grouping: There are three record types used to group other records. Listed alphabetically, these are:

RECORD TYPE	DESCRIPTION
ENDMDL	End-of-model record for multiple structures in a single coordinate entry.
MODEL	Specification of model number for multiple structures in a single coordinate entry.
TER	Chain terminator.

The MODEL/ENDMDL records surround groups of ATOM, HETATM, ANISOU, and TER records. TER records indicate the end of a chain.

Other: The remaining record types have a detailed inner structure. Listed alphabetically, these are:

RECORD TYPE	DESCRIPTION
JRNL	Literature citation that defines the coordinate set.
REMARK	General remarks; they can be structured or free form.

PDB Format Change Policy

The wwPDB will use the following protocol in making changes to the way PDB coordinate entries are represented and archived. The purpose of the policy is to allow ample time for everyone to understand these changes and to assess their impact on existing programs. PDB format modifications are necessary to address the changing needs of PDB users as well as the changing nature of the data that is archived.

1. Comments and suggestions will be solicited from the community on specific problems and data representation issues as they arise.

2. Proposed format changes will be disseminated through pdb-l@rcsb.org and wwpdb.org.

3. A 60-day discussion period will follow the announcement of proposed changes. Comments and suggestions must be received within this time period. Major changes that are not upwardly compatible will be allotted up to twice the standard amount of discussion time.

4. The wwPDB will then work in consultation with the wwPDB Advisory Committee and the equivalent partner Scientific Advisory Committees to evaluate and reconcile all suggestions. The final decision will be officially announced via pdb-l@rcsb.org and wwpdb.org.

5. Implementation will follow official announcement of the format change. Major changes will not appear in PDB files earlier than 60 days after the announcement, allowing sufficient time to modify files and programs.

Order of Records

All records in a PDB coordinate entry must appear in a defined order. Mandatory record types are present in all entries. When mandatory data are not provided, the record name must appear in the entry with a NULL indicator. Optional items become mandatory when certain conditions exist. Old records that are not described here are deprecated. Record order and existence are described in the following table:

RECORD TYPE	EXISTENCE	CONDITIONS IF OPTIONAL
HEADER	Mandatory	
OBSLTE	Optional	Mandatory in entries that have been replaced by a newer entry.
TITLE	Mandatory	
SPLIT	Optional	Mandatory complexes entries. when large macromolecular are split into multiple PDB
CAVEAT	Optional	Mandatory when there are outstanding errors such as chirality.
COMPND	Mandatory	
SOURCE	Mandatory	
KEYWDS	Mandatory	

EXPDTA Mandatory

NUMMDL Optional Mandatory for NMR ensemble entries.

MDLTYP Optional Mandatory for NMR minimized average Structures or when the entire polymer chain contains C alpha or P atoms only.

AUTHOR REVDAT SPRSDE
JRNL Mandatory Mandatory Optional
Optional

Mandatory for a Mandatory for a replacement entry. publication describes
the experiment.
REMARK 0 Optional Mandatory for a re-refined structure
REMARK 1 Optional
REMARK 2 Mandatory
REMARK 3 Mandatory
REMARK N Optional Mandatory under certain conditions.

DBREF Optional Mandatory for all polymers.
DBREF1/DBREF2 Optional Mandatory when certain sequence database

accession and/or sequence numbering does not fit preceding DBREF format.

SEQADV Optional Mandatory if sequence conflict exists.

SEQRES Mandatory Mandatory if ATOM records exist.

MODRES Optional Mandatory if modified group exists in the coordinates.

HET Optional Mandatory if a non-standard group other than water appears in the coordinates.

HETNAM Optional Mandatory if a non-standard group other than water appears in the coordinates.

HETSYN Optional

FORMUL Optional Mandatory if a non-standard group or water appears in the coordinates.

HELIXOptional

SHEET Optional

SSBOND Optional Mandatory if a disulfide bond is present.

LINK Optional Mandatory if non-standard residues appear in a polymer

CISPEP Optional

SITE Optional

CRYST1 Mandatory

ORIGX1 ORIGX2 ORIGX3 Mandatory SCALE1 SCALE2 SCALE3 Mandatory
MTRIX1 MTRIX2 MTRIX3 Optional Mandatory if the complete asymmetric unit
must be generated from the given coordinates using non-crystallographic symmetry.

MODEL Optional Mandatory if more than one model is present in the entry.

ATOMOptional Mandatory if standard residues exist.

ANISOU Optional

TER Optional Mandatory if ATOM records exist.

HETATM Optional Mandatory if non-standard group exists.

ENDMDL Optional Mandatory if MODEL appears.

CONECT Optional Mandatory if non-standard group appears and if LINK or SSBOND records exist.

MASTER Mandatory

END Mandatory

Sections of an Entry

The following table lists the various sections of a PDB entry and the records within it:

SECTION DESCRIPTION RECORD TYPE

Title Summary descriptive remarks HEADER, OBSLTE, TITLE, SPLIT,
CAVEAT, COMPND, SOURCE,
KEYWDS,EXPDTA, NUMMDL, MDLTYP,
AUTHOR, REVDAT, SPRSDE, JRNL
Remark Various comments about entry REMARKs 0-999 annotations in more depth than
standard records

Primary structure Peptide and/or nucleotide DBREF, SEQADV, SEQRES MODRES sequence and the
relationship between the PDB sequence and that found in the sequence database(s)

Heterogen	Description of non-standard groups	HET, HETNAM, HETSYN, FORMUL
Secondary structure	Description of secondary structure	HELIX, SHEET
Connectivity annotation	Chemical connectivity	SSBOND, LINK, CISPEP
Miscellaneous features	Features within the macromolecule	SITE
Crystallographic	Description of the crystallographic cell	CRYST1
Coordinate transformation	Coordinate transformation operators	ORIGXn, SCALEn, MTRIXn,
Coordinate	Atomic coordinate data	MODEL, ATOM, ANISOU, TER, HETATM, ENDMDL
Connectivity	Chemical connectivity	CONECT
Bookkeeping	Summary information, end-of-file marker	MASTER, END

Field Formats and Data Types

Each record type is presented in a table which contains the division of the records into fields by column number, defined data type, field name or a quoted string which must appear in the field, and field definition. Any column not specified must be left blank.

Each field contains an identified data type that can be validated by a program. These are:

DATA TYPE DESCRIPTION

AChar An alphabetic character (A-Z, a-z).

Atom Atom name.

Character Any non-control character in the ASCII character set or a space.

Continuation A two-character field that is either blank (for the first record of a set) or contains a two digit number
right-justified and blank-filled which counts continuation records starting with 2. The continuation number must be followed by a blank.

Date A 9 character string in the form DD-MMM-YY where DD is the day of the month, zero-filled on the left (e.g., 04); MMM is the common English 3-letter abbreviation of the month; and YY is the last two digits of the year. This must represent a valid date.

IDcode A PDB identification code which consists of 4 characters, the first of which is a digit in the range 0 - 9; the remaining 3 are alpha-numeric, and letters are upper case only. Entries with a 0 as the first character do not contain coordinate data.

Integer Right-justified blank-filled integer value.

Token A sequence of non-space characters followed by a colon and a space.

List A String that is composed of text separated with commas.

LString A literal string of characters. All spacing is significant and must be preserved.

LString(n) An LString with exactly n characters.

Real(n,m) Real (floating point) number in the FORTRAN format Fn.m.

Record name The name of the record: 6 characters, left-justified and blank-filled.

Residue name One of the standard amino acid or nucleic acids, as listed below, or the non-standard group designation as defined in the HET dictionary. Field is right-justified.

SList A String that is composed of text separated with semi-colons. Specification A String composed of a token and its associated value

separated by a colon.

Specification List A sequence of Specifications, separated by semi-colons.

String A sequence of characters. These characters may have arbitrary spacing, but should be interpreted as directed below.

String(n) A String with exactly n characters.

SymOP An integer field of from 4 to 6 digits, right-justified, of the form nnnMMM where nnn is the symmetry operator number and MMM is the translation vector.

To interpret a String, concatenate the contents of all continued fields together, collapse all sequences of multiple blanks to a single blank, and remove any leading and trailing blanks. This permits very long strings to be properly reconstructed.

2. Title Section

This section contains records used to describe the experiment and the biological macromolecules present in the entry: HEADER, OBSLTE, TITLE, SPLIT, CAVEAT, COMPND, SOURCE, KEYWDS, EXPDTA, AUTHOR, REVDAT, SPRSDE, JRNL, and REMARK records.

HEADER

Overview

The HEADER record uniquely identifies a PDB entry through the idCode field. This record also provides a classification for the entry. Finally, it contains the date when the coordinates were deposited to the PDB archive.

Record Format

COLUMNS DATA TYPE FIELD DEFINITION
----

1 - 6 Record name "HEADER"
11 - 50 String(40) classification Classifies the molecule(s).
51 - 59 Date depDate Deposition date. This is the date the coordinates were received at the PDB.
63 - 66 IDcode idCode This identifier is unique within the
PDB.

Details

* The classification string is left-justified and exactly matches one of a collection of strings.
A class list is available from the current wwPDB Annotation Documentation Appendices (http://www.wwpdb.org/docs.html). In the case of macromolecular complexes, the classification field must present a class for each macromolecule present. Due to the limited length of the classification field, strings must sometimes be abbreviated. In these cases, the full terms are given in KEYWDS.

* Classification may be based on function, metabolic role, molecule type, cellular location, etc. This record can describe dual functions of a molecules, and when applicable, separated by a comma “,”. Entries with multiple molecules in a complex will list the classifications of each macromolecule separated by slash “/”.

Verification/Validation/Value Authority Control

The verification program checks that the deposition date is a legitimate date and that the ID code is well-formed.

PDB coordinate entry ID codes do not begin with 0. “No coordinates”, or NOC files, given as 0xxx codes, contained no structural information and were bibliographic only. These entries were subsequently removed from PDB archive.

Relationships to Other Record Types

The classification found in HEADER also appears in KEYWDS, unabbreviated and in no strict order.

Example

1 2 3 4 5 6 7 8

12345678901234567890123456789012345678901234567890123456789012345678901234567890

HEADER PHOTOSYNTHESIS 28-MAR-07 2UXK
HEADER TRANSFERASE/TRANSFERASE INHIBITOR 17-SEP-04 1XH6
HEADER MEMBRANE PROTEIN, TRANSPORT PROTEIN 20-JUL-06 2HRT

OBSLTE

Overview

OBSLTE appears in entries that have been removed from public distribution.

This record acts as a flag in an entry that has been removed (“obsoleted”) from the PDB's full release. It indicates which, if any, new entries have replaced the entry that was obsoleted. The format allows for the case of multiple new entries replacing one existing entry.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"OBSLTE"	
9 - 10	Continuation	continuation	Allows concatenation of multiple records
12 - 20	Date	repDate	Date that this entry was replaced.
22 - 25	IDcode	idCode	ID code of this entry.
32 - 35	IDcode	rIdCode	ID code of entry that replaced this one.
37 - 40	IDcode	rIdCode	ID code of entry that replaced this one.
42 - 45	IDcode	rIdCode	ID code of entry that replaced this one.
47 - 50	IDcode	rIdCode	ID code of entry that replaced this one.
52 - 55	IDcode	rIdCode	ID code of entry that replaced this one.
57 - 60	IDcode	rIdCode	ID code of entry that replaced this one.
62 - 65	IDcode	rIdCode	ID code of entry that replaced this one.
67 - 70	IDcode	rIdCode	ID code of entry that replaced this one.
72 - 75	IDcode	rIdCode	ID code of entry that replaced this one.

Details

* Major revisions to coordinates that change the structure's geometry or chemical composition (such as a change in the sequence of the polymers or ligand identity) require the entry to be obsoleted and superseded by a new deposition. Further information can be

found at wwPDB policies (http://www.wwpdb.org/policy.html) . All OBSLTE entries are available from the PDB archive (ftp://ftp.wwpdb.org/pub/pdb/data/structures/obsolete).

* Though the obsolete entry is removed from the public archive, the initial citation that reported the structure is carried over to the superseding entry.

Verification/Validation/Value Authority Control

wwPDB staff adds this record at the time an entry is removed from release.

Relationships to Other Record Types

None.

Example

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890


OBSLTE     31-JAN-94 1MBP    2MBP
```

TITLE

Overview

The TITLE record contains a title for the experiment or analysis that is represented in the entry. It should identify an entry in the same way that a citation title identifies a publication.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"TITLE "	
9 - 10	Continuation	continuation	Allows concatenation of multiple records.
11 - 80	String	title	Title of the experiment.

Details

* The title of the entry is free text and should describe the contents of the entry and any procedures or conditions that distinguish this entry from similar entries. It presents an opportunity for the depositor to emphasize the underlying purpose of this particular experiment.

* Some items that may be included in TITLE are:

- Experiment type.
- Description of the mutation.
- The fact that only alpha carbon coordinates have been provided in the entry.

Verification/Validation/Value Authority Control

This record is free text so no verification of format is required. The title is supplied by the depositor, but staff may exercise editorial judgment in consultation with depositors in assigning the title.

Relationships to Other Record Types

COMPND, SOURCE, EXPDTA, and REMARKs provide information that may also be found in TITLE. You may think of the title as describing the experiment, and the compound record as describing the molecule(s).

Examples

1 2 3 4 5 6 7 8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
TITLE RHIZOPUSPEPSIN COMPLEXED WITH REDUCED PE**PTIDE INHIBITOR**

TITLE STRUCTURE OF THE TRANSFORMED MONOCLINIC LYSOZYME BY TITLE 2 CONTROLLED DEHYDRATION

TITLE NMR STUDY OF OXIDIZED THIOREDOXIN MUTANT (C62A,C69A,C73A) TITLE 2 MINIMIZED AVERAGE STRUCTURE

SPLIT (added)

Overview

The SPLIT record is used in instances where a specific entry composes part of a large macromolecular complex. It will identify the PDB entries that are required to reconstitute a complete complex.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"SPLIT "	
9 - 10	Continuation	continuation	Allows concatenation of multiple records.
12 - 15	IDcode	idCode	ID code of related entry.
17 - 20	IDcode	idCode	ID code of related entry.
22 - 25	IDcode	idCode	ID code of related entry.
27 – 30	IDcode	idCode	ID code of related entry.

32 - 35	IDcode	idCode	ID code of related entry.
37 - 40	IDcode	idCode	ID code of related entry.
42 - 45	IDcode	idCode	ID code of related entry.
47 - 50	IDcode	idCode	ID code of related entry.
52 - 55	IDcode	idCode	ID code of related entry.
57 - 60	IDcode	idCode	ID code of related entry.
62 - 65	IDcode	idCode	ID code of related entry.
67 - 70	IDcode	idCode	ID code of related entry.
72 - 75	IDcode	idCode	ID code of related entry.
77 - 80	IDcode	idCode	ID code of related entry.

Details

* The SPLIT record can be continued on multiple lines, so that all related PDB entries are cataloged.

Verification/Validation/Value Authority Control

This record will be generated at the time of processing the component PDB files of the large macromolecular complex when all complex constituents are deposited.

Relationships to Other Record Types

REMARK 350 will contain an amended statement to reflect the entire complex.

Examples

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890

SPLIT  1VOQ 1VOR 1VOS 1VOU 1VOV 1VOW 1VOX 1VOY 1VP0 1VOZ
```

CAVEAT

Overview

CAVEAT warns of errors and unresolved issues in the entry. Use caution when using an entry containing this record.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"CAVEAT"	
9 - 10	Continuation	continuation	Allows concatenation of multiple records.

12 - 15	IDcode	idCode	PDB ID code of this entry.
20 - 79	String	comment	Free text giving the reason for the CAVEAT.

Details

*	The CAVEAT will also be included in cases where the wwPDB is unable to verify the transformation of the coordinates back to the crystallographic cell. In these cases, the molecular structure may still be correct.

Verification/Validation/Value Authority Control

CAVEAT will be added to entries known to be incorrect.

COMPND (updated)

Overview

The COMPND record describes the macromolecular contents of an entry. Some cases where the entry contains a standalone drug or inhibitor, the name of the non-polymeric molecule will appear in this record. Each macromolecule found in the entry is described by a set of token: value pairs, and is referred to as a COMPND record component. Since the concept of a molecule is difficult to specify exactly, staff may exercise editorial judgment in consultation with depositors in assigning these names.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"COMPND"	
8 - 10	Continuation	continuation	Allows concatenation of multiple records.
11 - 80	Specification list	compound	Description of the molecular components.

Details

* The compound record is a Specification list. The specifications, or tokens, that may be used are listed below:

TOKEN	VALUE DEFINITION
MOL_ID	Numbers each component; also used in SOURCE to associate the information.
MOLECULE	Name of the macromolecule.
CHAIN	Comma-separated list of chain identifier(s).

FRAGMENT Specifies a domain or region of the molecule.

SYNONYM Comma-separated list of synonyms for the MOLECULE.

EC The Enzyme Commission number associated with the molecule.
If there is more than one EC number, they are presented as a comma-separated list.

ENGINEERED Indicates that the molecule was produced using
recombinant technology or by purely chemical synthesis.

MUTATION Indicates if there is a mutation.

OTHER_DETAILS Additional comments.

* In the case of synthetic molecules, the depositor will provide the description.

* For chimeric proteins, the protein name is comma-separated and may refer to the presence of a linker (protein_1, linker, protein_2).

* Asterisks in nucleic acid names (in MOLECULE) are for ease of reading.

* No specific rules apply to the ordering of the tokens, except that the occurrence of MOL_ID or FRAGMENT indicates that the subsequent tokens are related to that specific molecule or fragment of the molecule.

* When insertion codes are given as part of the residue name, they must be given within square brackets, i.e., H57[A]N. This might occur when listing residues in FRAGMENT or OTHER_DETAILS.

* For multi-chain molecules, e.g., the hemoglobin tetramer, a comma-separated list of CHAIN identifiers is used.

Verification/Validation/Value Authority Control

CHAIN must match the chain identifiers(s) of the molecule(s). EC numbers are also checked.

Relationships to Other Record Types

In the case of mutations, the SEQADV records will present differences from the reference molecule. REMARK records may further describe the contents of the entry. Also see verification above.

Examples

1 2 3 4 5 6 7 8
1234567890123456789012345678901234567890123456789012345678901234567890123
4567890

COMPND

```
COMPND COMPND COMPND COMPND COMPND COMPND COMPND COMPND
COMPND COMPND COMPND      2
3
4
5
6
7
8
9
10
11
12        MOL_ID: 1;
MOLECULE: HEMOGLOBIN ALPHA CHAIN; CHAIN:   A,   C;
SYNONYM: DEOXYHEMOGLOBIN ALPHA CHAIN; ENGINEERED:    YES;
MUTATION: YES; MOL_ID:  2;
MOLECULE: HEMOGLOBIN BETA CHAIN; CHAIN:   B,   D;
SYNONYM: DEOXYHEMOGLOBIN BETA CHAIN; ENGINEERED:    YES;
MUTATION:  YES
COMPND COMPND
2         MOL_ID: 1;
MOLECULE: COWPEA CHLOROTIC MOTTLE VIRUS;
COMPND      3         CHAIN: A, B, C;
COMPND      4         SYNONYM:  CCMV;
COMPND      5         MOL_ID: 2;
COMPND      6         MOLECULE:    RNA    (5'-(*AP*UP*AP*U)-3');
COMPND      7         CHAIN: D, F;
COMPND      8         ENGINEERED: YES;
COMPND      9         MOL_ID: 3;
COMPND      10        MOLECULE:  RNA  (5'-(*AP*U)-3');
COMPND      11        CHAIN: E;
COMPND      12        ENGINEERED:   YES
COMPND COMPND
2         MOL_ID: 1;
MOLECULE: HEVAMINE A;
COMPND      3         CHAIN: A;


COMPND      4   EC:   3.2.1.14,  3.2.1.17;
COMPND      5  OTHER_DETAILS:  PLANT  ENDOCHITINASE/LYSOZYME
```

SOURCE (updated)

Overview

The SOURCE record specifies the biological and/or chemical source of each biological molecule in the entry. Some cases where the entry contains a standalone drug or inhibitor, the source information of this molecule will appear in this record. Sources are described by both the common name and the scientific name, e.g., genus and species. Strain and/or cell-line for immortalized cells are given when they help to uniquely identify the biological entity studied.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"SOURCE"	
8 - 10	Continuation	continuation	Allows concatenation of multiple records.
11 - 79	Specification List	srcName	Identifies the source of the macromolecule in a token: value format.

Details

TOKEN VALUE DEFINITION

MOL_ID Numbers each molecule. Same as appears in COMPND.

SYNTHETIC Indicates a chemically-synthesized source.

FRAGMENT A domain or fragment of the molecule may be specified.

ORGANISM_SCIENTIFIC Scientific name of the organism.

ORGANISM_COMMON Common name of the organism.

ORGANISM_TAXIDNCBI Taxonomy ID number of the organism.

STRAIN Identifies the strain.

VARIANT Identifies the variant.

CELL_LINE The specific line of cells used in the experiment.

ATCC American Type Culture Collection tissue culture number.

ORGAN Organized group of tissues that carries on a specialized function.

TISSUE Organized group of cells with a common function and structure.

CELL Identifies the particular cell type.

ORGANELLEOrganized structure within a cell.

SECRETION Identifies the secretion, such as saliva, urine, or venom, from which the molecule was isolated.

CELLULAR_LOCATION Identifies the location inside/outside the cell.

PLASMID Identifies the plasmid containing the gene.

GENE Identifies the gene.

EXPRESSION_SYSTEM Scientific name of the organism in which the molecule was expressed.

EXPRESSION_SYSTEM_COMMON Common name of the organism in which the molecule
was expressed.

EXPRESSION_SYSTEM_TAXID NCBI Taxonomy ID of the organism used as the expression system.

EXPRESSION_SYSTEM_STRAIN Strain of the organism in which the molecule
was expressed.

EXPRESSION_SYSTEM_VARIANT Variant of the organism used as the
expression system.

EXPRESSION_SYSTEM_CELL_LINE The specific line of cells used as the
expression system.

EXPRESSION_SYSTEM_ATCC_NUMBER Identifies the ATCC number of the expression System.

EXPRESSION_SYSTEM_ORGAN Specific organ which expressed the molecule.
EXPRESSION_SYSTEM_TISSUE Specific tissue which expressed the molecule.
EXPRESSION_SYSTEM_CELL Specific cell type which expressed the molecule.
EXPRESSION_SYSTEM_ORGANELLE Specific organelle which expressed the molecule.
EXPRESSION_SYSTEM_CELLULAR_LOCATION Identifies the location inside or outside
the cell which expressed the molecule.

EXPRESSION_SYSTEM_VECTOR_TYPE Identifies the type of vector used, i.e.,
plasmid, virus, or cosmid.

EXPRESSION_SYSTEM_VECTOR Identifies the vector used.
EXPRESSION_SYSTEM_PLASMID Plasmid used in the recombinant experiment.
EXPRESSION_SYSTEM_GENE Name of the gene used in recombinant experiment.
OTHER_DETAILS Used to present information on the source which is not given elsewhere.

* The srcName is a list of tokens: value pairs describing each biological component of the entry.

* As in COMPND, the order is not specified except that MOL_ID or FRAGMENT indicates subsequent specifications are related to that molecule or fragment of the molecule.

* Only the relevant tokens need to appear in an entry.

* Molecules prepared by purely chemical synthetic methods are described by the specification SYNTHETIC followed by "YES" or an optional value, such as NON-BIOLOGICAL SOURCE or BASED ON THE NATURAL SEQUENCE. ENGINEERED must appear in the COMPND record.

* In the case of a chemically synthesized molecule using a biologically functional sequence (nucleic or amino acid), SOURCE reflects the biological origin of the sequence and COMPND reflects its synthetic nature by inclusion of the token ENGINEERED. The token SYNTHETIC appears in SOURCE.

* If made from a synthetic gene, ENGINEERED appears in COMPND and the expression system is described in SOURCE (SYNTHETIC does NOT appear in SOURCE).

* If the molecule was made using recombinant techniques, ENGINEERED appears in COMPND and the system is described in SOURCE.

* When multiple macromolecules appear in the entry, each MOL_ID, as given in the COMPND record, must be repeated in the SOURCE record along with the source information for the corresponding molecule.

* Hybrid molecules prepared by fusion of genes are treated as multi-molecular systems for the purpose of specifying the source. The token FRAGMENT is used to associate the source with its corresponding fragment.

• When necessary to fully describe hybrid molecules, tokens may appear more than once for a given MOL_ID.

• All relevant token: value pairs that taken together fully describe each fragment are grouped following the appropriate FRAGMENT.

• Descriptors relative to the full system appear before the FRAGMENT (see third example below).

* ORGANISM_SCIENTIFIC provides the Latin genus and species. Virus names are listed as the scientific name.

* Cellular origin is described by giving cellular compartment, organelle, cell, tissue, organ, or body part from which the molecule was isolated.

* CELLULAR_LOCATION may be used to indicate where in the organism the compound was found. Examples are: extracellular, periplasmic, cytosol.

* Entries containing molecules prepared by recombinant techniques are described as follows:

- The expression system is described.

- The organism and cell location given are for the source of the gene used in the cloning experiment.

* Transgenic organisms, such as mouse producing human proteins, are treated as expression systems.

* New tokens may be added by the wwPDB.

Verification/Validation/Value Authority Control

The biological source is compared to that found in the sequence databases. The Tax ID is identified and the corresponding scientific and common names for the organism is matched to a standard taxonomy database (such as NCBI).

Relationships to Other Record Types

Each macromolecule listed in COMPND must have a corresponding source.

Examples

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
SOURCE    MOL_ID:  1;
SOURCE    2  ORGANISM_SCIENTIFIC:  AVIAN  SARCOMA  VIRUS; SOURCE 3
ORGANISM_TAXID:  11876
SOURCE    4 STRAIN: SCHMIDT-RUPPIN B;
SOURCE    5  EXPRESSION_SYSTEM:  ESCHERICHIA  COLI; SOURCE   6
EXPRESSION_SYSTEM_TAXID:  562
SOURCE    7  EXPRESSION_SYSTEM_PLASMID:  PRC23IN

SOURCE    MOL_ID:  1;
SOURCE    2  ORGANISM_SCIENTIFIC:  GALLUS  GALLUS; SOURCE    3
ORGANISM_COMMON: CHICKEN;
SOURCE    3  ORGANISM_TAXID:  9031 SOURCE    4 ORGAN: HEART;
SOURCE    5  TISSUE:  MUSCLE
```

For a Chimera protein:

```
SOURCE    MOL_ID:  1;
SOURCE    2    ORGANISM_SCIENTIFIC:    MUS    MUSCULUS,    HOMO
SAPIENS; SOURCE  3 ORGANISM_COMMON: MOUSE, HUMAN;
SOURCE    3 ORGANISM_TAXID: 10090, 9606
SOURCE    5 EXPRESSION_SYSTEM: ESCHERICHIA COLI; SOURCE     6
EXPRESSION_SYSTEM_TAXID: 344601 SOURCE       6
EXPRESSION_SYSTEM_STRAIN: B171;
```

SOURCE 7 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID; SOURCE 8 EXPRESSION_SYSTEM_PLASMID: P4XH-M13;

KEYWDS

Overview

The KEYWDS record contains a set of terms relevant to the entry. Terms in the KEYWDS record provide a simple means of categorizing entries and may be used to generate index files. This record addresses some of the limitations found in the classification field of the HEADER record. It provides the opportunity to add further annotation to the entry in a concise and computer- searchable fashion.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"KEYWDS"	
9 - 10	Continuation	continuation	Allows concatenation of records if necessary.
11 - 79	List	keywds	Comma-separated list of keywords relevant to the entry.

Details

* The KEYWDS record contains a list of terms relevant to the entry, similar to that found in journal articles. A phrase may be used if it presents a single concept (e.g., reaction center). Terms provided in this record may include those that describe the following:

- Functional classification.
- Metabolic role.
- Known biological or chemical activity.
- Structural classification.

*Other classifying terms may be used. No particular ordering is required. A number of PDB entries contain complexes of macromolecules. In these cases, all terms applicable to each molecule should be provided separated by a comma.

*Note that the terms in the KEYWDS record duplicate those found in the classification field of the HEADER record. Terms abbreviated in the HEADER record are unabbreviated in KEYWDS.

Verification/Validation/Value Authority Control

Terms used in the KEYWDS record are subject to scientific and editorial review. A list of terms, definitions, and synonyms will be maintained by the wwPDB. Every attempt will be made to provide some level of consistency with keywords used in other biological databases.

Relationships to Other Record Types

HEADER records contain a classification term which must also appear in KEYWDS. Scientific judgment will dictate when terms used in one entry to describe a molecule should be included in other entries with the same or similar molecules.

Example

1   2   3   4   5   6   7   8
123456789012345678901234567890123456789012345678901234567890123
4567890
KEYWDS LYASE, TRICARBOXYLIC ACID CYCLE, MITOCHONDRION, OXIDATIVE KEYWDS 2 METABOLISM

EXPDTA (updated)

Overview

The EXPDTA record presents information about the experiment.

The EXPDTA record identifies the experimental technique used. This may refer to the type of radiation and sample, or include the spectroscopic or modeling technique. Permitted values include:

X-RAY DIFFRACTION FIBER DIFFRACTION NEUTRON DIFFRACTION
ELECTRON CRYSTALLOGRAPHY ELECTRON MICROSCOPY
SOLID-STATE NMR SOLUTION NMR SOLUTION SCATTERING

*Note:Since October 15, 2006, theoretical models are no longer accepted for deposition. Any theoretical models deposited prior to this date are archived at ftp://ftp.wwpdb.org/pub/pdb/data/structures/models.
Please see the documentation from previous versions for the related file format description.

Record Format

COLUMNS DATA TYPE FIELD DEFINITION
----
1 - 6 Record name "EXPDTA"

9 - 10 Continuation continuation Allows concatenation of multiple records.

11 - 79 SList technique The experimental technique(s) with
optional comment describing the sample or experiment.

Details

* EXPDTA is mandatory and appears in all entries. The technique must match one of the permitted values. See above.

* If more than one technique was used for the structure determination and is being represented in the entry, EXPDTA presents the techniques as a semi-colon separated list.

Verification/Validation/Value Authority Control

The verification program checks that the EXPDTA record appears in the entry and that the technique matches one of the allowed values. It also checks that the relevant standard REMARK is added, as in the cases of NMR or electron microscopy studies, that the appropriate CRYST1 and SCALE values are used.

Relationships to Other Record Types

If the experiment is an NMR or electron microscopy study, this may be stated in the TITLE, and the appropriate EXPDTA and REMARK records should appear. Specific details of the data collection and experiment appear in the REMARKs.

In the case of a polycrystalline fiber diffraction study, CRYST1 and SCALE contain the normal unit cell data.

Examples

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890123
4567890
EXPDTA    X-RAY   DIFFRACTION

EXPDTA    NEUTRON    DIFFRACTION;    X-RAY    DIFFRACTION   EXPDTA
          SOLUTION  NMR
EXPDTA    ELECTRON  MICROSCOPY
```

NUMMDL (added)

Overview

The NUMMDL record indicates total number of models in a PDB entry.

Details

* The modelNumber field lists total number of models in a PDB entry and is left justified.

* If more than one model appears in the entry, the number of models included must be stated.

* NUMMDL is mandatory if a PDB entry contains more than one models.

Verification/Validation/Value Authority Control

The verification program checks that the modelNumber field is correctly formatted.

Example

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234
567890
NUMMDL    20
```

MDLTYP (added)

Overview

The MDLTYP record contains additional annotation pertinent to the coordinates presented in the entry.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"MDLTYP"	
9 - 10	Continuation	continuation	Allows concatenation of multiple records.
11 - 80	SList	comment	Free Text providing additional structural annotation.

Details

* The MDLTYP record will be used by the wwPDB to highlight certain features of the deposited coordinates as described below.

* For entries that are determined by NMR methods and the coordinates deposited are either a minimized average or regularized mean structure, this record will contain the tag "MINIMIZED AVERAGE" to highlight the nature of the deposited coordinates in the entry.

* Where the entry contains entire polymer chains that have only either C-alpha (for proteins) or P atoms (for nucleotides), the MDLTYP record will be used to describe the contents of such chains along with the chain identifier. For these polymeric chains, REMARK 470 (Missing Atoms) will be omitted.

* If multiple features need to be described in this record, they will be separated by a ";" delineator.

*	Where an entry has multiple features requiring description in this record including MINIMIZED AVERAGE, the MINIMIZED AVERAGE value will precede all other annotation.

*	New descriptors may be added by the wwPDB.

Verification/Validation/Value Authority Control

The chain_identifiers described in this record must be present in the COMPND, SEQRES and the coordinate section of the entry.

Example

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
MDLTYP    MINIMIZED AVERAGE

MDLTYP    CA  ATOMS  ONLY,  CHAIN  A, B, C, D, E, F, G, H, I, J, K  ; P  ATOMS
ONLY, MDLTYP     2 CHAIN  X, Y, Z



MDLTYP    MINIMIZED  AVERAGE;  CA  ATOMS  ONLY,  CHAIN  A,  B
```

AUTHOR

Overview

The AUTHOR record contains the names of the people responsible for the contents of the entry.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"AUTHOR"	
9 - 10	Continuation	continuation	Allows concatenation of multiple records.
11 - 79	List	authorList	List of the author names, separated by commas.

Details

*	The authorList field lists author names separated by commas with no subsequent spaces.

*	Representation of personal names:

- First and middle names are indicated by initials, each followed by a period, and precede the surname.

- Only the surname (family or last name) of the author is given in full.

- Hyphens can be used if they are part of the author's name.

- Apostrophes are allowed in surnames.

- Umlauts and other character modifiers are not given.

*	Structure of personal names:

- There is no space after any initial and its following period.

- Blank spaces are used in a name only if properly part of the surname (e.g., J.VAN DORN), or between surname and Jr., II, or III

Abbreviations that are part of a surname, such as Jr., St. or Ste., are followed by a period and a space before the next part of the surname.

*	Representation of corporate, organization or university names:

- Group names used for one or all of the authors should be spelled out in full.

- The name of the larger group comes before the name of a subdivision, e.g., University of Somewhere, Department of Chemistry.

*	Structure of list:

- Line breaks between multiple lines in the authorList occur only after a comma.

- Personal names are not split across two lines.

*	Special cases:

- Names are given in English if there is an accepted English version; otherwise in the native language, transliterated if necessary.

Verification/Validation/Value Authority Control

The verification program checks that the authorList field is correctly formatted. It does not perform any spelling checks or name verification.

Relationships to Other Record Types

The format of the names in the AUTHOR record is the same as in JRNL and REMARK 1 references.

Example

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890123
4567890
AUTHOR
          M.B.BERRY,B.MEADOR,T.BILDERBACK,P.LIANG,M.GLASER, AUTHOR  2
G.N.PHILLIPS    JR.,T.L.ST.    STEVENS
```

REVDAT (updated)

Overview

REVDAT records contain a history of the modifications made to an entry since its release.

Record Format

COLUMNS DATA TYPE FIELD DEFINITION

----

1 - 6 Record name "REVDAT"

8 - 10 Integer modNum Modification number.

11 - 12 Continuation continuation Allows concatenation of multiple records.

14 - 22 Date modDate Date of modification (or release for new entries) in DD-MMM-YY format. This is not repeated on continued lines.

24 - 27 IDCode modId ID code of this entry. This is not repeated on continuation lines.

32 Integer modType An integer identifying the type of modification. For all revisions, the

modification type is listed as 1

40 - 45 LString(6) record Modification detail.

47 - 52 LString(6) record Modification detail.

54 - 59 LString(6) record Modification detail.

61 - 66 LString(6) record Modification detail.

Details

* Each time revisions are made to the entry, a modification number is assigned in increasing (by 1) numerical order. REVDAT records appear in descending order (most recent modification appears first). New entries have a REVDAT record with modNum equal to 1 and modType equal to 0.

Allowed modTypes are:

0 Initial released entry.

1 Other modification.

* Each revision may have more than one REVDAT record, and each revision has a separate continuation field.

* Modification details are typically PDB record names such as JRNL, SOURCE, TITLE, or COMPND. A special modification detail VERSN indicates that the file has undergone a change in version. The current version will be specified in REMARK 4.

Verification/Validation/Value Authority Control

The modType must be one of the defined types, and the given record type must be valid. If modType is 0, the modId must match the entry's ID code in the HEADER record.

Relationships to Other Record Types

In the case of a version revision, the current will be specified in REMARK 4.

Template

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890

REVDAT    2         15-OCT-99     1ABC  1         REMARK
REVDAT    1         09-JAN-89     1ABC  0
```

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890

REVDAT    2         11-MAR-08 2ABC    1         JRNL   VERSN
REVDAT    1         09-DEC-03  2ABC    0
```

SPRSDE

Overview

The SPRSDE records contain a list of the ID codes of entries that were made obsolete by the given coordinate entry and removed from the PDB release set. One entry may replace many.

It is wwPDB policy that only the principal investigator of a structure has the authority to obsolete it.

Record Format

```
COLUMNS     DATA TYPE  FIELD DEFINITION
--------------------------------------------------------------------------------
```

1 - 6	Record name	"SPRSDE"	
9 - 10	Continuation	continuation	Allows for multiple ID codes.
12 - 20	Date	sprsdeDate	Date this entry superseded the listed entries. This field is not copied on continuations.
22 - 25	IDcode	idCode	ID code of this entry. This field is not copied on continuations.
32 - 35	IDcode	sIdCode	ID code of a superseded entry.
37 - 40	IDcode	sIdCode	ID code of a superseded entry.
42 - 45	IDcode	sIdCode	ID code of a superseded entry.
47 - 50	IDcode	sIdCode	ID code of a superseded entry.
52 - 55	IDcode	sIdCode	ID code of a superseded entry.
57 - 60	IDcode	sIdCode	ID code of a superseded entry.
62 - 65	IDcode	sIdCode	ID code of a superseded entry.
67 - 70	IDcode	sIdCode	ID code of a superseded entry.
72 - 75	IDcode	sIdCode	ID code of a superseded entry.

Details

- The ID code list is terminated by the first blank sIdCode field.

Verification/Validation/Value Authority Control

wwPDB checks that the superseded entries have actually been removed from release.

Relationships to Other Record Types

The sprsdeDate is usually the date the entry is released, and therefore matches the date in the REVDAT 1 record. The ID code found in the idCode field must be the same as one found in the idCode field of the HEADER record.

Example

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890

SPRSDE     17-JUL-84  4HHB      1HHB
SPRSDE     27-FEB-95 1GDJ       1LH4  2LH4
```

JRNL (updated)

Overview

The JRNL record contains the primary literature citation that describes the experiment which resulted in the deposited coordinate set. There is at most one JRNL reference per entry. If there is no primary reference, then there is no JRNL reference. Other references are given in REMARK 1.

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"JRNL "	
13 - 79	LString	text	See Details below.

Details

* The following tables are used to describe the sub-record types of the JRNL record.

* The AUTH sub-record is mandatory in JRNL. This is followed by TITL, EDIT, REF, PUBL, REFN, PMID and DOI sub- record types. REF and REFN are also mandatory in JRNL. EDIT and PUBL may appear only if the reference is to a non-journal.

1. AUTH

* AUTH contains the list of authors associated with the cited article or contribution to a larger work (i.e., AUTH is not used for the editor of a book).

* The author list is formatted similarly to the AUTHOR record. It is a comma-separated list of names. Spaces at the end of a sub-record are not significant; all other spaces are significant. See the AUTHOR record for full details.

* The authorList field of continuation sub-records in JRNL differs from that in AUTHOR by leaving no leading blank in column 20 of any continuation lines.

* One author's name, consisting of the initials and family name, cannot be split across two lines. If there are continuation sub-records, then all but the last sub-record must end in a comma.

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"REMARK"	
10	LString(1)	"1"	
13 - 16	LString(4)	"AUTH"	Appears on all continuation records.
17 - 18	Continuation	continuation	Allows a long list of authors.
20 - 79	List	authorList	List of the authors.

2. TITL

* TITL specifies the title of the reference. This is used for the title of a journal article, chapter, or part of a book. The TITL line is omitted if the author(s) listed in authorList

wrote the entire book (or other work) listed in REF and no section of the book is being cited.

* If an article is in a language other than English and is printed with an alternate title in English, the English language title is given, followed by a space and then the name of the language (in its English form, in square brackets) in which the article is written.

* If the title of an article is in a non-Roman alphabet the title is transliterated.

* The actual title cited is reconstructed in a manner identical to other continued records, i.e., trailing blanks are discarded and the continuation line is concatenated with a space inserted.

* A line cannot end with a hyphen. A compound term (two elements connected by a hyphen) or chemical names which include a hyphen must appear on a single line, unless they are too long to fit on one line, in which case the split is made at a normally-occurring hyphen. An individual word cannot be hyphenated at the end of a line and put on two lines. An exception is when there is a repeating compound term where the second element is omitted, e.g., "DOUBLE- AND TRIPLE- RESONANCE". In such a case the non-completed word "DOUBLE-" could end a line and not alter reconstruction of the title.

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"REMARK"	
10	LString(1)	"1"	
13 - 16	LString(4)	"TITL"	Appears on all continuation records.
17 - 18	Continuation	continuation	Permits long titles.
20 - 79	LString	title	Title of the article.

3. EDIT

* EDIT appears if editors are associated with a non-journal reference. The editor list is formatted and concatenated in the same way that author lists are.

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"REMARK"	
10	LString(1)	"1"	
13 - 16	LString(4)	"TITL"	Appears on all continuation records.
17 - 18	Continuation	continuation	Permits long titles.
20 - 79	LString	title	Title of the article.

4. REF

* REF is a group of fields that contain either the publication status or the name of the publication (and any supplement and/or report information), volume, page, and year. There are two forms of

this sub-record group, depending upon the citation's publication status.

4a. If the reference has not been published yet, the sub-record type group has the form:

COLUMNS	DATA TYPE	FIELD DEFINITION
1 - 6	Record name	"JRNL "
13 - 16	LString(3)	"REF"
20 - 34	LString(15)	"TO BE PUBLISHED"

* Publication name (first item in pubName field):

If the publication is a serial (i.e., a journal, an annual, or other non-book or non-monographic item issued in parts and intended to be continued indefinitely), use the abbreviated name of the publication as listed in PubMed with periods.

If the publication is a book, monograph, or other non-serial item, use its full name according to the Anglo-American Cataloguing Rules, 2nd Revised Edition; (AACR2R). (Non-serial items include theses, videos, computer programs, and anything that is complete in one or a finite number of parts.) If there is a sub-title, verifiable in an online catalog, it will be included using the same punctuation as in the source of verification. Preference will be given to verification using cataloging of the Library of Congress, the National Library of Medicine, and the British Library, in that order.

If a book is part of a monographic series: the full name of the book (according to the AACR2R) is listed first, followed by the name of the series in which it was published. The series information is given within parentheses and the series name is preceded by "IN:" and a space. The series name should be listed in full unless the series has an accepted ISO abbreviation. If applicable, the series name should be followed, after a comma and a space, by a volume (V.) and/or number (NO.) and/or part (PT.) indicator and its number and/or letter in the series.

* Supplement (follows publication name in pubName field):

If a reference is in a supplement to the volume listed, or if information about a "part" is needed to distinguish multiple parts with the same page numbering, such information should be put in the REF sub-record.

A supplement indication should follow the name of the publication and should be preceded by a comma and a space. Supplement should be abbreviated as "SUPPL." If there is a supplement number or letter, it should follow "SUPPL." without an intervening space. A part indication should also follow the name of the publication and be preceded by a comma and a space. A part should be abbreviated as "PT.", and the number or letter should follow without an intervening space.

If there is both a supplement and a part, their order should reflect the order printed on the work itself.

* Report (follows publication name and any supplement or part information in pubName field):

If a book has a report designation, the report information should follow the title and precede series information. The name and number of the report is given in parentheses, and the name is

preceded by "REPORT:" and a space.

* Reconstruction of publication name:

The name of the publication is reconstructed by removing any trailing blanks in the pubName field, and concatenating all of the pubName fields from the continuation lines with an intervening space. There are two conditions where no intervening space is added between lines: when the pubName field on a line ends with a hyphen or a period, or when the line ends with a hyphen (-). When the line ends with a period (.), add a space if this is the only period in the entire pubName field; do not add a space if there are two or more periods throughout the pubName field, excluding any periods after the designations "SUPPL", "V", "NO", or "PT".

* Volume, page, and year (volume, first page, year fields respectively):

The REF sub-record type group also contains information about volume, page, and year when applicable.

In the case of a monograph with multiple volumes which is also in a numbered series, the number in the volume field represents the number of the book, not the series. (The volume number of the series is in parentheses with the name of the series, as described above under publication name.)

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"JRNL "	
13 - 16	LString(3)	"REF "	
17 - 18	Continuation	continuation	Allows long publication names.
20 - 47	LString	pubName	Name of the publication including section or series designation. This is the only field of this sub-record which may be continued on successive sub-records.
50 - 51	LString(2)	"V."	Appears in the first sub-record only, and only if column 55 is non-blank.
52 - 55	String	volume	Right-justified blank-filled volume information; appears in the first sub-record only.
57 - 61	String	page	First page of the article; appears in the first sub-record only.
63 - 66	Integer	year	Year of publication; first sub-record only.

5. PUBL

* PUBL contains the name of the publisher and place of publication if the reference is to a book or other non-journal publication. If the non-journal has not yet been published or released, this sub- record is absent.

* The place of publication is listed first, followed by a space, a colon, another space, and then the name of the publisher/issuer. This arrangement is based on the ISBD(M) International Standard Bibliographic Description for Monographic Publications (Rev.Ed., 1987) and the AACR2R, and is

used in public online catalogs in libraries. Details on the contents of PUBL are given below.

* Place of publication:

Give the place of publication. If the name of the country, state, province, etc. is considered necessary to distinguish the place of publication from others of the same name, or for identification, then follow the city with a comma, a space, and the name of the larger geographic area.

If there is more than one place of publication, only the first listed will be used. If an online catalog record is used to verify the item, the first place listed there will be used, omitting any brackets.
Preference will be given to the cataloging done by the Library of Congress, the National Library of Medicine, and the British Library, in that order.

* Publisher's name (or name of other issuing entity):

Give the name of the publisher in the shortest form in which it can be understood and identified internationally, according to AACR2R rule 1.4D.

If there is more than one publisher listed in the publication, only the first will be used in the PDB file. If an online catalog record is used to verify the item, the first place listed there will be used for the name of the publisher. Preference will be given to the cataloging of the Library of Congress, the National Library of Medicine, and the British Library, in that order.

* Ph.D. and other theses:

Theses are presented in the PUBL record if the degree has been granted and the thesis made available for public consultation by the degree-granting institution.
The name of the degree-granting institution (the issuing agency) is followed by a space and "(THESIS)".

* Reconstruction of place and publisher:

The PUBL sub-record type can be reconstructed by removing all trailing blanks in the pub field and concatenating all of the pub fields from the continuation lines with an intervening space.
Continued lines do not begin with a space.

COLUMNS DATA TYPE FIELD DEFINITION
----

1 - 6 Record name "JRNL "
13 - 16 LString(4) "PUBL"
17 - 18 Continuation continuation Allows long publisher and place names.
20 - 70 LString pub City of publication and name of the publisher/institution.

6. REFN (changed)

* REFN is a group of fields that contain encoded references to the citation. No continuation lines are possible. Each piece of coded information has a designated field.

* There are two forms of this sub-record type group, depending upon the publication status. 6a. This form of the REFN sub-record type group is used if the citation has not been published.
COLUMNS DATA TYPE FIELD DEFINITION
----

1 - 6 Record name "JRNL "

13 - 16 LString(4) "REFN"

6b. This form of the REFN sub-record type group is used if the citation has been published.

COLUMNS DATA TYPE FIELD DEFINITION
----

1 - 6 Record name "JRNL "
13 - 16 LString(4) "REFN"
36 - 39 LString(4) "ISSN"
"ESSN" or International Standard Serial Number or Electronic Standard Serial Number.
41 - 65 LString issn ISSN number (final digit may be a letter and may contain one or more dashes).

7. PMID (added)

* PMID lists the PubMed unique accession number of the publication related to the entry.

COLUMNS DATA TYPE FIELD DEFINITION
----
1 - 6 Record name "JRNL "

13 - 16 LString(4) "PMID"

20 – 79 Integer continuation unique PubMed identifier number assigned to the publication describing the experiment. Allows for a long PubMed ID number.

8. DOI (added)

* DOI is the Digital Object Identifier for the related electronic publication (“e-pub”), if applicable.

* Every DOI consists of a publisher prefix, a fore-slash (“/”), and then a suffix which can be any length and may include a combination of numbers and alphabets.
For example: 10.1073/PNAS.0712393105

COLUMNS DATA TYPE FIELD DEFINITION
----
1 - 6 Record name "JRNL "

13 - 16 LString(4) "DOI "

20 – 79 LString continuation Unique DOI assigned to the publication describing the experiment. Allows for a long DOI string.

Verification/Validation/Value Authority Control

wwPDB verifies that this record is correctly formatted.

Citations appearing in JRNL may not also appear in REMARK 1.

Relationships to Other Record Types

The publication cited as the JRNL record may not be repeated in REMARK 1.

Example

```
1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
JRNL   AUTH G.FERMI,M.F.PERUTZ,B.SHAANAN,R.FOURME
JRNL   TITL  THE CRYSTAL STRUCTURE OF HUMAN DEOXYHAEMOGLOBIN AT JRNL   TITL 2 1.74 A RESOLUTION
JRNL   REF   J.MOL.BIOL.           V. 175       159  1984 JRNL   REFN
       ISSN  0022-2836
JRNL   PMID  6726807
JRNL   DOI   10.1016/0022-2836(84)90472-8
```

Known Problems

- Interchange of bibliographic information and linking with other databases is hampered by the lack of labels or specific locations for certain types of information or by more than one type of information being in a particular location. This is most likely to occur with books, series, and reports. Some of the points below provide details about the variations and/or blending of information.

- Titles of the publications that require more than 28 characters on the REF line must be continued on subsequent lines. There is some awkwardness due to volume, page, and year appearing on the first REF line, thereby splitting up the title.

- Information about a supplement and its number/letter is presented in the publication's title field (on the REF lines in columns 20 - 47).

- When series information for a book is presented, it is added to the REF line. The number of REF lines can become large in some cases because of the 28-column limit for title information in REF.

- Books that are issued in more than one series are not accommodated.

- Pagination is limited to the beginning page.

QUALITY INFORMATION ON THE WEB

Rather than having to install and run one of the above packages, it is possible to obtain much of the information it provides from the Web. Several sites provide precomputed quality criteria for all existing structures in the PDB. Other sites allow you upload your own PDB file, via your Web browser, and will run their validation programs on it and provide you with the results of their checks.

PDBsum—PROCHECK Summaries

The first site that provides precomputed quality criteria is the PDBsum Web site (Laskowski, 2001) at http://www.biochem.ucl.ac.uk/bsm/pdbsum. This Web site spe- cializes in structural analyses and pictorial representations of all PDB structures. Each structure containing one or more protein chains has a PROCHECK and a WHAT CHECK button. The former gives a Ramachandran plot for all protein chains in the structure, together with summary statistics calculated by the PROCHECK program. These results can provide a quick guide to the likely quality of the structure, in addition to the structure's resolution, R-factor and, where available, Rfree .

The WHATCHECK button links to the PDBREPORT for the structure, described

below.

Occasionally the model of a protein structure is so bad that one can tell immediately from merely looking at the secondary structure plot on the PDBsum page. Most proteins have around $50-60\%$ of their residues in regions of regular secondary structure, that is, in α-helices and $\beta-$strands. However, if a model is really poor, the main-chain oxygen and nitrogen atoms responsible for the hydrogen-bonding that maintains the regular secondary structures can lie beyond normal hydrogen-bonding distances; so the algorithms that assign secondary structure (Chapter 17) may fail to detect some of the α-helices and $\beta-$strands that the correct protein structure contains. Figure 14.11 gives an example of the secondary structure contents for a typical protein and for the protein that had the poor Ramachandran plot in Figure 14.9b.

PDBREPORT—WHATCHECK Results

The WHATCHECK button on the PDBsum page leads to the WHAT IF Check report on the given protein's coordinates. This report is a detailed listing (plus an even more detailed one, called the Full report) of the numerous analyses that have been precomputed using the WHATCHECK program. These analyses include space group and symmetry checks, geometrical checks on bond lengths, bond angles, torsion angles, proline puckers, bad contacts, planarity checks, checks on hydrogen-bonds, and more, including an overall summary report intended for users of the model. The PDBREPORT database can be accessed directly at http://www.cmbi.kun.nl/gv/pdbreport.

PDB's Geometry Analyses

The PDB Web site (http://www.rcsb.org/pdb) also has geometrical analyses on each entry, consisting of tables of average, minimum, and maximum values for the protein's bond lengths, bond angles, and dihedral angles. Unusual values are highlighted. It is also possible to view a backbone representation of the structure in RasMol, colored according to the Fold Deviation Score — the redder the coloring the more unusual the residue's conformational parameters.

Validation Servers on the Web

In addition to the sites mentioned above, there are a number of validation servers on the Web that allow you to submit a PDB file for analysis. Table 14.1 lists these servers. They are mostly for protein structures and most use programs that are freely available for in-house use (see Table 14.2). However, the servers can often be easier and more convenient to use, and of course save you having to download and install the programs, particularly the Biotech Validation server that runs the three most commonly used validation programs: PROCHECK, PROVE, and WHATCHECK.

The Structural Classification of Proteins

- The Structural Classification of Proteins (SCOP) database is a classification of protein domains organised according to their evolutionary and structural relationships.
- The SCOP database is a classification that organises proteins of known three-dimensional structure according to their structural and evolutionary relationships .
- It was established in 1994 at MRC LMB and CPE in Cambridge and over the years has attracted a broad range of users, thus becoming a valuable resource in different areas of protein research.
- **Current SCOP classification structure**
- **Two evolutionary levels:** family and superfamily are at the heart of the current SCOP classification.
- Family groups closely related proteins with a clear evidence for their evolutionary origin while superfamily brings together more distantly related protein domains.
- As these relationships can sometimes span structural regions of different size, we provide domain boundaries for both, family and superfamily levels.

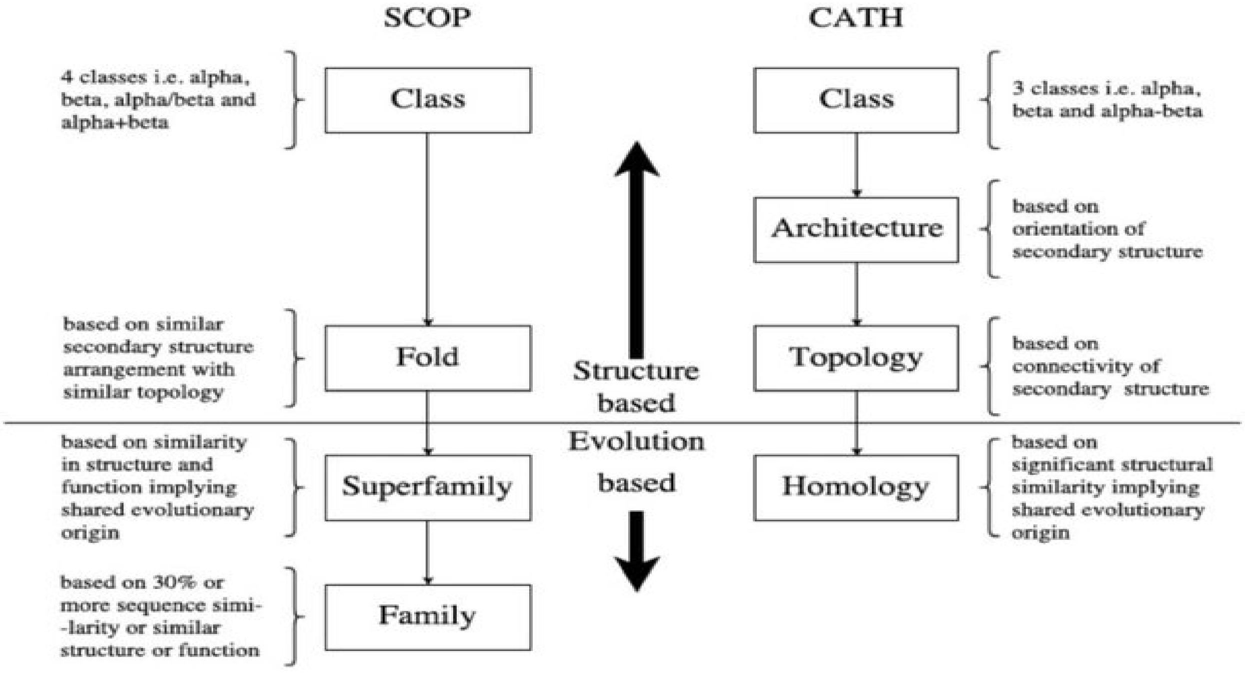
- **Superfamilies** are grouped into distinct folds on the basis of the global structural features shared by the majority of their members. These features are the composition of the secondary structures in the domain core, their architecture and topology.

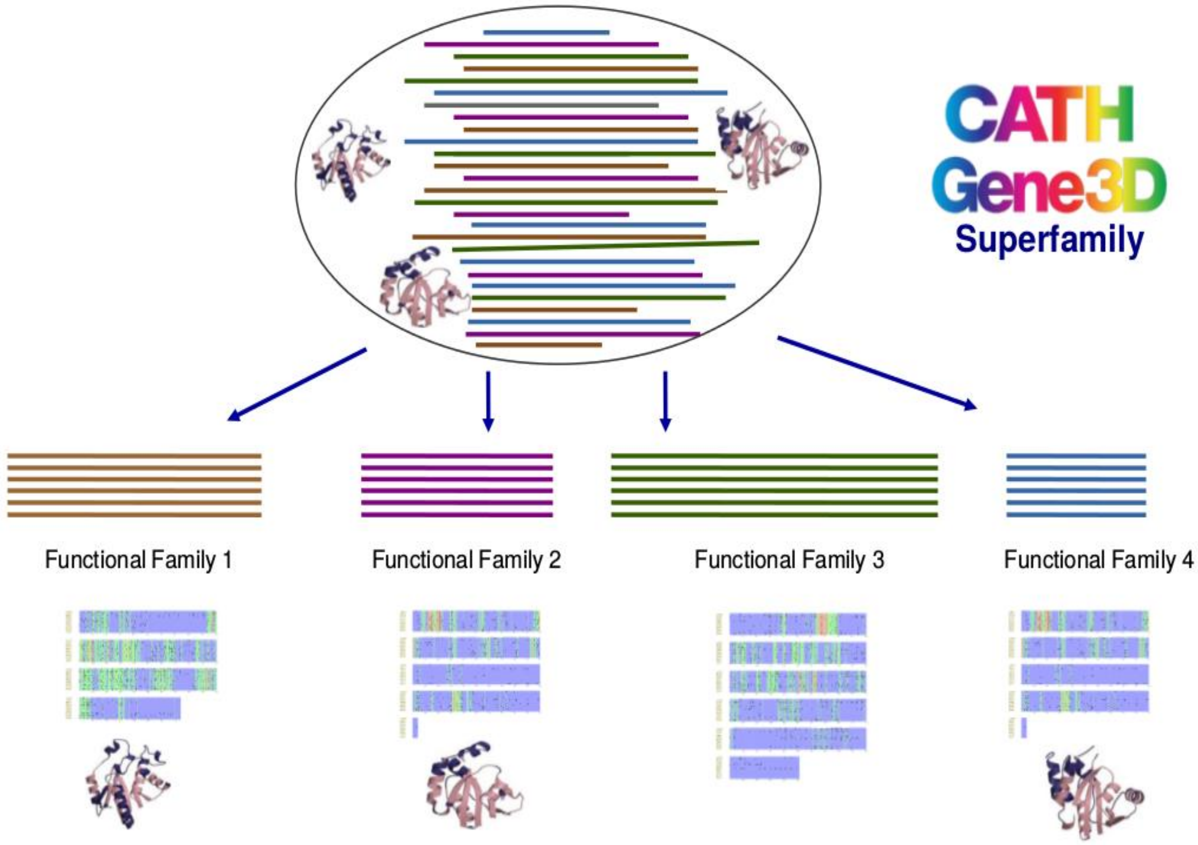
- **Fold is an attribute of a superfamily but the constituent families of some superfamilies** that have evolved distinct structural features can belong to a different fold.

- Superfamilies of proteins or protein regions that do not adopt globular folded structure are grouped in IUPRs (Intrinsically Unstructured Protein Region).

- Some of these proteins exist in an ensemble of different conformations or are unstructured in free state but adopt an ordered conformation upon binding to other macromolecules.

- Folds and IUPRs with different secondary structural content are placed into one of the five different structural classes.
- These include all-alpha and all-beta proteins, containing predominantly alpha-helices and beta-strands, respectively, and 'mixed' alpha and beta classes (a/b) and (a+b) with respectively alternating and segregated alpha-helices and beta-strands, and the fifth class of small proteins with little or no secondary structures.
- Folds and IUPRs are also grouped based on their protein type, into four groups: soluble, membrane, fibrous and intrinsically disordered. Each of these types to a large extent correlates with characteristic sequence and structural features.


SCOP
CATH
4 classes i.e. alpha, beta, alpha/beta and alpha+beta
Class
Class
3 classes i.e. alpha, beta and alpha-beta
Architecture
based on orientation of secondary structure
based on similar secondary structure arrangement with similar topology
Fold
Structure based
Topology
based on connectivity of secondary structure
based on similarity in structure and function implying shared evolutionary origin
Superfamily
Evolution based
Homology
based on significant structural similarity implying shared evolutionary origin
based on 30% or more sequence simi- -larity or similar structure or function
Family



Each colour denotes a unique function/sub-function.

UNIT – III- Structural Bioinformatics – SBI1403

Stabilising forces in protein structure

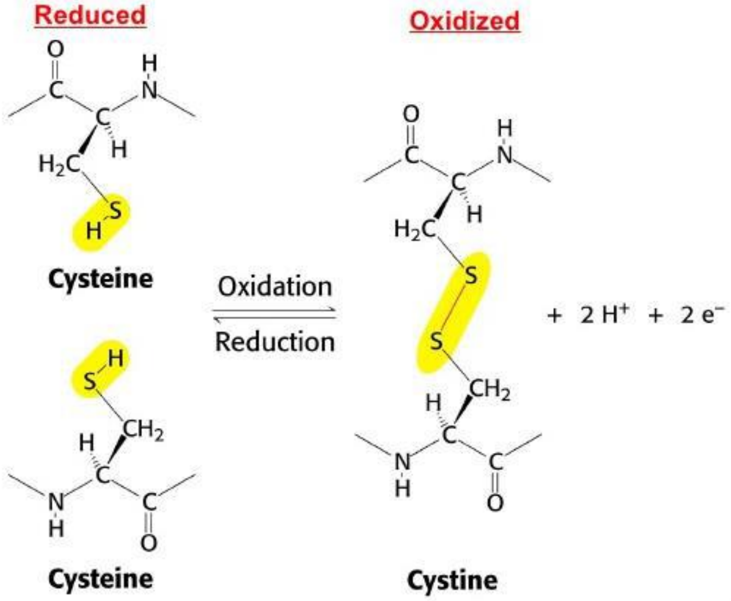
Stabilizing the Shape of Proteins
Proteins are made of amino acid chains, or polypeptides. Amino acids have a basic backbone made of an amino group and a carboxyl group, and differ in their side-chains.
These polypeptide chains of amino acids can be shaped as helixes or sheets, which come together to form a 3-Dimensional structure. The 3-Dimensional structure of proteins is referred to as its "tertiary structure".
The process of folding proteins into their tertiary structures is spontaneous and involves bonds and intermolecular forces to make the structure stable, are broadly categorized into two classes
1. Covalent interactions
2. Non-covalent interactions

Covalent interactions

Covalent bonds are the strongest chemical bonds contributing to protein structure. In addition to the covalent bonds that connect the atoms of a single amino acid and the covalent peptide bond that links amino acids in a protein chain, covalent bonds between cysteine side chains are important determinants of protein structure. Cysteine is the sole amino acid whose side chain can form covalent bonds, yielding disulfide links or bridges.

Disulfide links

Disulfide bonds are formed between two sulfur (SH) atoms, which are found in the side-chain of the amino acid cysteine. When two cysteines are brought into close proximity in the tertiary structure, covalent disulfide bond can be formed as a result of oxidation. The two cysteine residues involved in the bond formation may be far apart in the primary structure but are brought close together as a result of protein folding

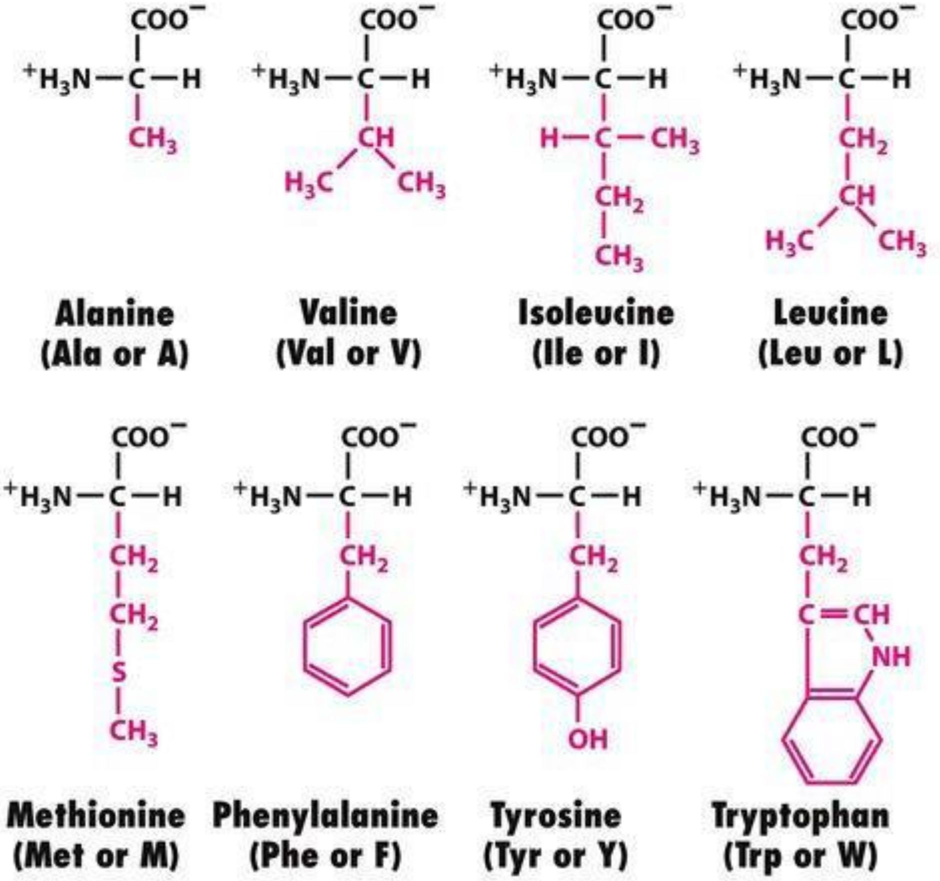


Disulfide links
Disulfide bonds within and between polypeptide chains form as a protein folds to its native conformation.
Some polypeptides whose Cys residues have been derivatized to prevent disulfide bond formation can still assume their fully active conformations, suggesting that disulfide bonds are not essential

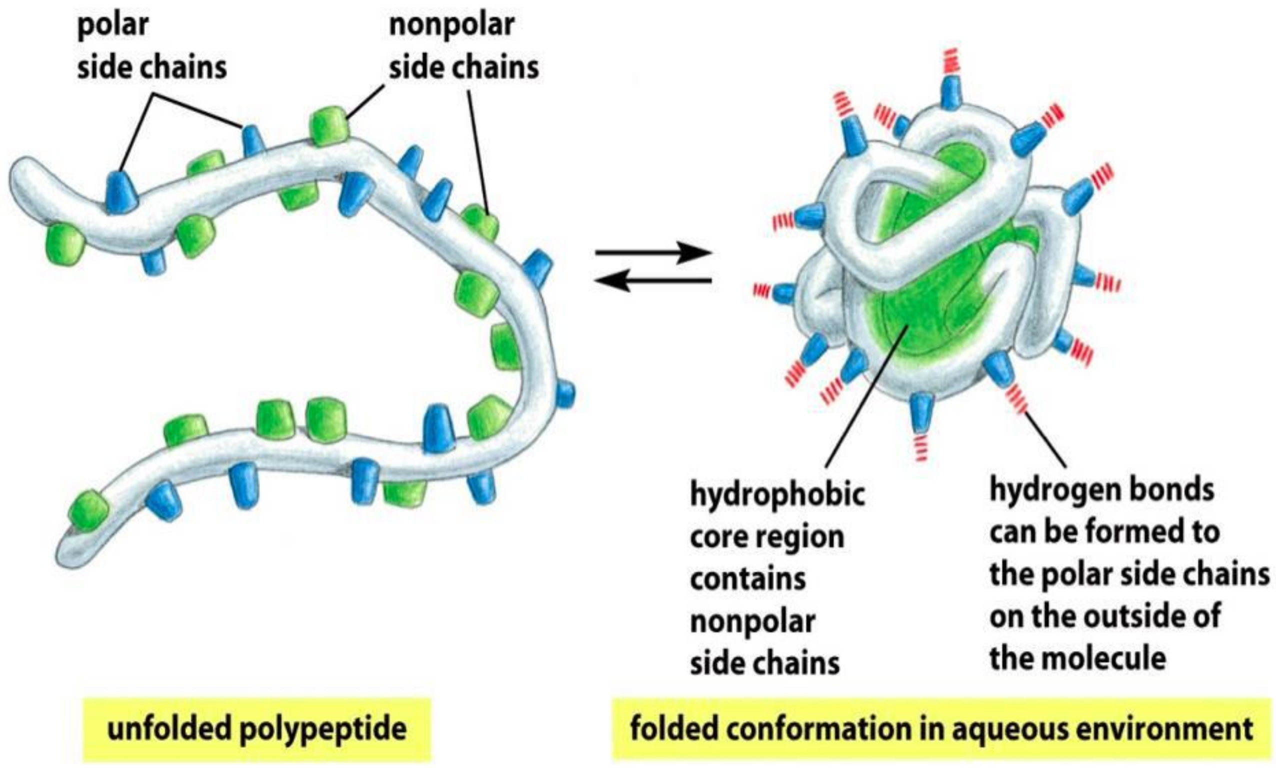
stabilizing forces. They may, however, be important for “locking in” a particular backbone folding pattern as the protein proceeds from its fully extended state to its mature form.

Non-covalent interactions The 3-dimensional (3D) structure of proteins results from a delicate balance between various types of non-covalent interactions acting between the amino acid present the polypeptide chain and also with the surrounding environment. Although non-covalent interactions are typically orders of magnitude weaker than covalent bonds but they play important role in the formation and maintenance of 3D structural integrity of protein. Individual amino acids are distinguished by the chemical nature of their side chains. They can be roughly grouped into categories as being hydrophobic , aromatic, hydrophilic, charged, etc. This diversity in the amino acids enables them of forming a wide range of non-covalent interactions. Some of the common non-covalent interactions observed in Proteins are:

1.Hydrophobic bond or interactions
2.Van Der Waals interactions
3.Electrostatic or ionic bond or salt bond or salt bridge
4.Hydrogen Bond

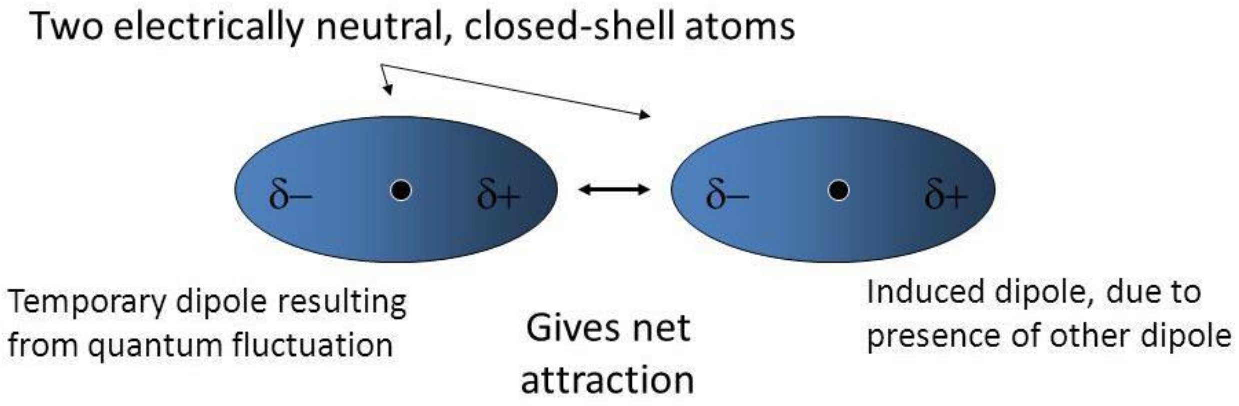


Hydrophobic Interactions



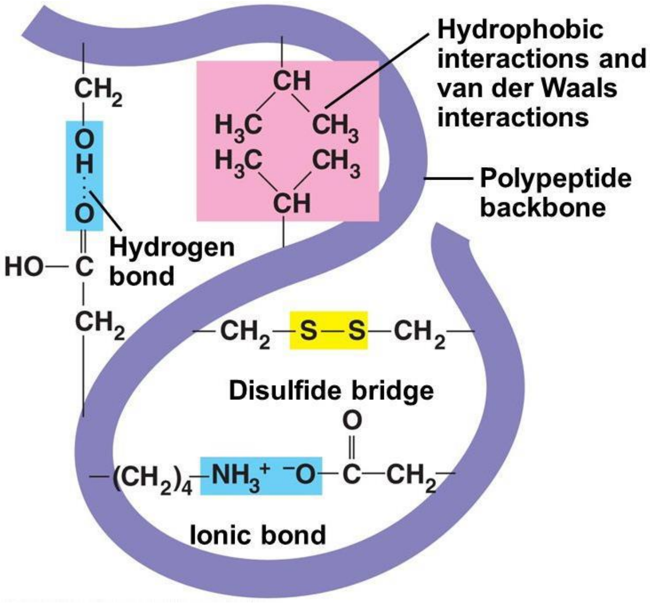
Van Der Waals forces

The Van der Waals force is a transient, weak electrical attraction of one atom for another. Van der Waals attractions exist because every atom has an electron cloud that can fluctuate, yielding a temporary electric dipole. The transient dipole in one atom can induce a complementary dipole in another atom, provided the two atoms are quite close. These short-lived, complementary dipoles provide a weak electrostatic attraction known as the Van der Waals force. The appropriate distance required for Van der Waals attractions depends on the size of each electron cloud of atoms and is referred to as the Van der Waals radius



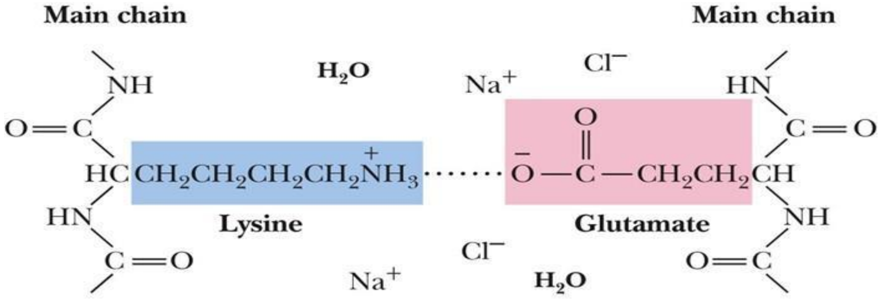
Van der Waals forces

In 3-dimensional structure of proteins , the formation of Van der Waals forces depends on the shape of the side-chain; if the atoms within the side-chains of neighboring amino acids fit well, then Van der Waals force is formed. Well packed hydrophobic cores of proteins represent optimized van der Waals interactions between non-polar residues. Although individually weak, numerous neighbor interactions in such central cores can contribute a significant stabilization to the native structure. Van der Waals forces can play important roles in protein-protein recognition when complementary shapes are involved. This is the case in antibody-antigen recognition, where a "lock and key" fit of the two molecules yields extensive Van der Waals attractions.



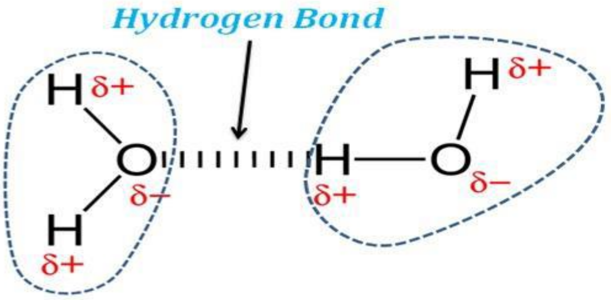
Ionic Bonds- Salt Bridges

Salt bridges in proteins are bonds between oppositely charged residues that are sufficiently close to each other to experience electrostatic attraction. Ionic bonds are formed as amino acids bearing opposite electrical charges are juxtaposed in the hydrophobic core of proteins. Ionic bonding in the interior is rare because most charged amino acids lie on the protein surface. Although rare, ionic bonds can be important to protein structure because they are potent electrostatic attractions that can approach the strength of covalent bonds. An ionic or salt bridge can be formed between the carboxylate ion of an acidic residues such as aspartic acid or glutamic acid and an ammonium ion of the basic residue such as lysine, arginine or histidine



Hydrogen bonds

When two atoms bearing partial negative charges share a partially positively charged hydrogen, the atoms are engaged in a hydrogen bond (H-bond). Hydrogen bonding is a form of weak attractive force between molecules that contain an electric charge. It is caused by electrostatic attraction and can alter the chemical properties of the molecules. The Hydrogen bond attractive force is weaker than full ionic bonding.



Formation of hydrogen bond between two water molecules

Hydrogen bonds in proteins

The correct 3-dimensional structure of a protein is often dependent on an intricate network of H-bonds. These can occur between a variety of atoms, involving:

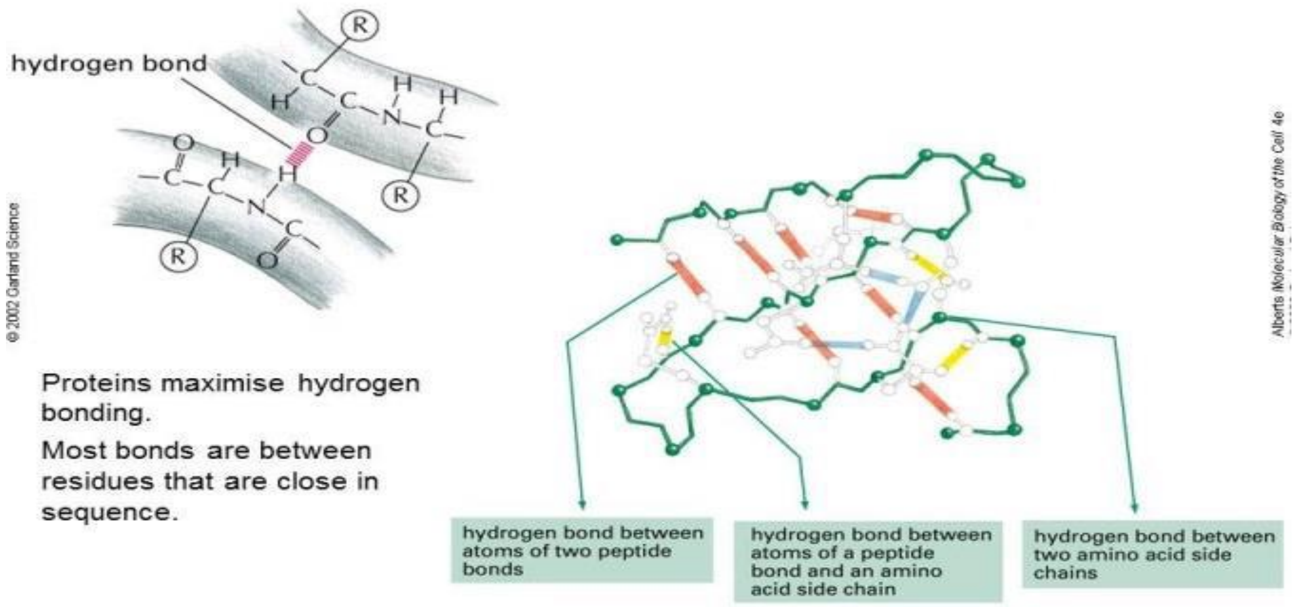
•atoms on two different amino acid sidechains

•atoms on amino acid sidechains and water molecules at the protein surface

•atoms on amino acid sidechains and protein backbone atoms

•backbone atoms and water molecules at the protein surface

•backbone atoms on two different amino acids Polar groups exposed on the surface of proteins often have water as their hydrogen bonding partner. Polar groups within the core region usually form hydrogen bonds with other groups within the protein. H bond can formed between large number of amino acid residues: serine, threonine, aspartic acid, glutamic acid, glutamine, lysine, arginine, histidine, tryptophan, tyrosine and asparagine. Hydrogen bonds are important determinants of native protein structures, because if a protein folded in a way that prevented a hydrogen bond from forming, the stabilizing energy of that hydrogen bond would be lost.

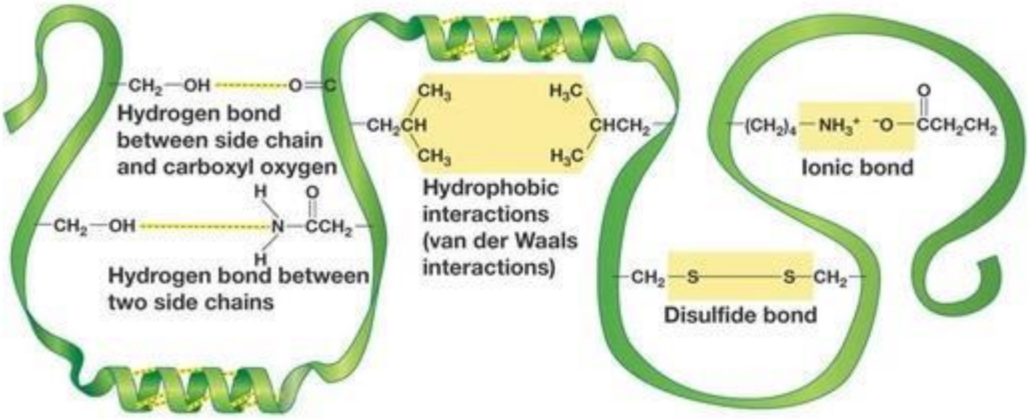

hydrogen bond
R
C
H
C
N
H
H
C
O
O
C
H
C
H
N
C
O
R
R
R
hydrogen bond between atoms of two peptide bonds
hydrogen bond between atoms of a peptide bond and an amino acid side chain
hydrogen bond between two amino acid side chains

Proteins maximise hydrogen bonding.

Most bonds are between residues that are close in sequence.

© 2002 Garland Science

Alberts Molecular Biology of the Cell 4e



$-CH_2-OH \cdots O=C$
Hydrogen bond between side chain and carboxyl oxygen
$-CH_2-OH \cdots N(H)(H)-C(=O)CH_2-$
Hydrogen bond between two side chains
$-CH_2CH(CH_3)_2$ $(H_3C)_2CHCH_2-$
Hydrophobic interactions (van der Waals interactions)
$-CH_2-S-S-CH_2-$
Disulfide bond
$-(CH_2)_4-NH_3^+$ $^-O-C(=O)CH_2CH_2$
Ionic bond

How Transmission Electron Microscopy complements protein X-ray crystallography and NMR

- Protein molecules carry out the majority of functions within cells and most of these activities involve the interaction of multiple proteins and other macromolecules.
- The determination of the structure of a single protein using protein X-ray crystallography or NMR has become easier over the past twenty years.
- However, solving the structure of a protein complex is still very challenging.
- Getting a protein complex to crystallize is not always possible and the large size of many complexes makes them difficult to study with NMR.
- TEM has the ability to determine the structures of these macromolecular complexes.
- TEM works best for complexes that are 250kDa or larger which complements protein X-ray crystallography and NMR studies of individual proteins or domains.
- In recent years, through significant technical advances in electron microscope design and electron detection technology, the resolutions achievable using cryo-TEM have improved and several protein structures have been solved at better than 3.5Å resolution.
- The combination of protein X-ray crystallography, NMR and TEM offer the ability to not only resolve smaller proteins at high resolution but to also examine entire proteins complexes as one large macromolecular structure with structural details of some components and molecular modeling enabling the creation a complete atomic model.
- Cryo-TEM can also study heterogeneous samples and provide structural details about dynamic and complexes that are difficult to examine with other structural biology techniques.

How TEM can help with challenges in structural biology

- TEM is a powerful tool that provides direct images of macromolecules and can help with many structural biology projects.
- TEM can provide a big picture view of larger complexes and provide information on the stability and dynamics of a macromolecular complex.
- When a complex or protein cannot be crystallized, TEM can often provide a structure, sometimes at near atomic resolution.
- Even when a protein complex may be able to be crystallized, TEM can provide an initial 3D model of the structure that may help in determining higher resolution structures.

- TEM provides valuable structural information that complements data from other structural biology techniques to help create a more complete picture of macromolecular complexes and help in the generation or mechanistic models to describe their functions.

Main advantages of TEM:

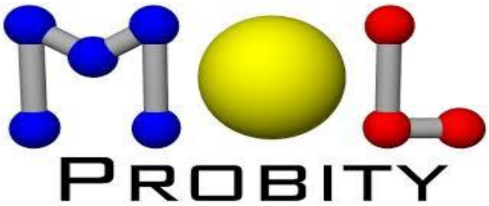
- No need for crystals
- No upper limit to the size or complexity of the macromolecular complex
- Near atomic resolutions are possible with the latest microscopes and detectors
- Heterogeneous samples can be analyzed allowing dynamic and unstable complexes to be studied
- Only micrograms of proteins are required for analysis
- Cryo-TEM visualizes macromolecules in a fully hydrated, close to native state

How TEM is used in Structural Biology

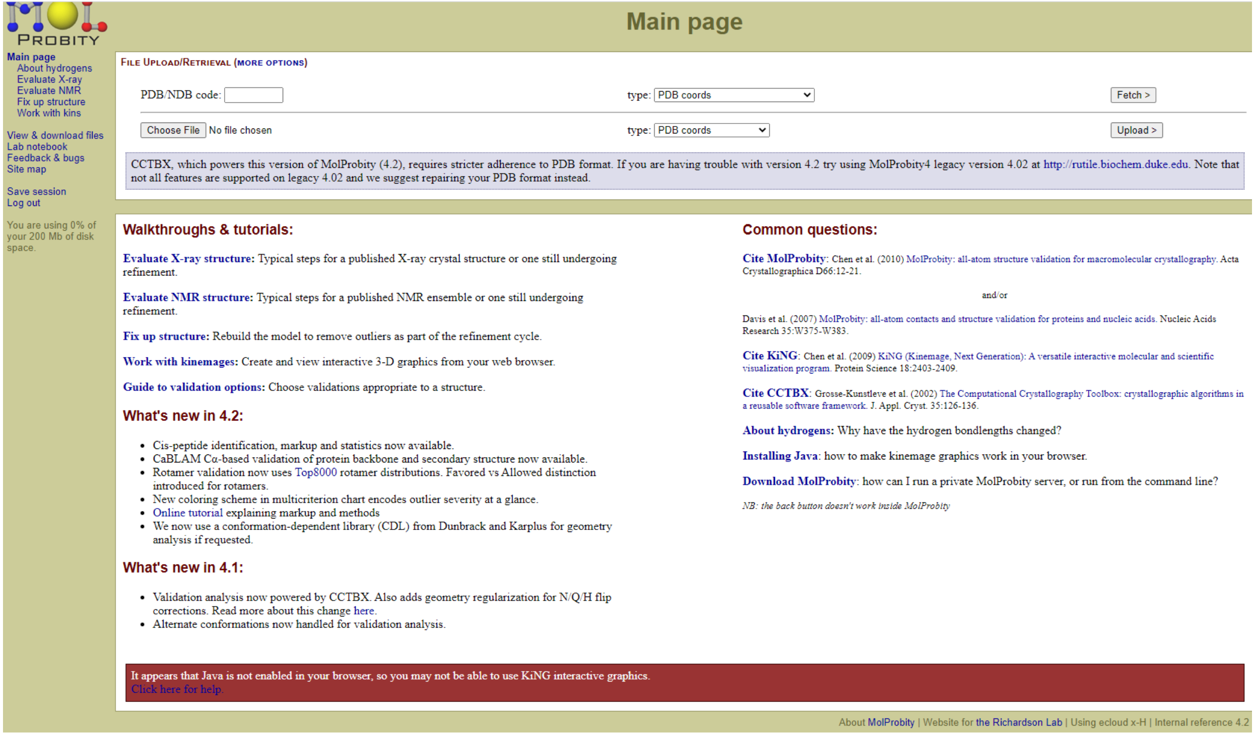
- In structural biology, TEM is a technique where two dimensional images of individual macromolecular complexes are taken with a transmission electron microscope.
- These two dimensional images can be mathematically aligned, through image processing techniques, to generate a 3D volume of the macromolecules.
- The **samples are typically either encased in a heavy metal stain (negative stain), such as uranyl acetate or imaged at cryogenic temperatures with the sample embedded in vitreous ice, free of strain (Cryo-TEM)** .
- In order to prepare proteins for cryo-TEM, the specimen is applied to a carbon coated EM grid with a series of small holes (µm size range).
- The sample is blotted away, leaving a thin film with the specimen residing within the small holes.
- The grid is then rapidly plunged into liquid ethane or propane, cooled to liquid nitrogen temperatures.
- The specimen must then be maintained at cryogenic temperatures in order to prevent a phase transition, which would result in formation of crystalline ice and damage to the specimen.
- The resulting sample is subsequently imaged in an electron microscope, and reconstruction software is used to create the 3D structure from the individual particle images.
- Although individual particles are imaged at low contrast, the resulting structure can be extremely high resolution due to averaging of hundreds or thousands of particles.
- Using high resolution cryo-EM, the authors reveal that the SARS-CoV-2 virus spike proteins are at least 10 times more tightly bound to their host cell receptor than those of the spike protein of the SARS virus (SARS-CoV).

- This correlates with the inability of SARS-CoV-2 to effectively bind to the SARS-CoV spike protein and provides insight into how the pathway for vaccine development must differ from previous coronavirus outbreaks.

- Wrapp et al. determined a 3.5-angstrom-resolution structure of the 2019-nCoV **trimeric spike protein by cryo–electron microscopy.**
- Using biophysical assays, the authors show that this protein binds at least 10 times more tightly than the corresponding spike protein of severe acute respiratory syndrome (SARS)–CoV to their common host cell receptor.
- They also tested three antibodies known to bind to the SARS-CoV spike protein but did not detect binding to the 2019-nCoV spike protein.



- MolProbity is a widely used system of model validation for **protein and nucleic acid structures**, accessed at http://molprobity.biochem.duke.edu.
- It builds upon the work of earlier systems such as ProCheck, WhatIf, and Oops, which introduced the use of validation by Ramachandran-plot and sidechain rotamer criteria.
- It complements systems for validating data and model-to-data match such as Rfree or real-space residual
- MolProbity has some features specifically tailored for X-ray crystallography, and is also suitable, and used, for cryoEM, neutron, NMR, and computational models.
- MolProbity's unique feature of **all-atom contact analysis (including hydrogens) was described in 1999** followed by its complementary rotamer, Ramachandran, and Cβ deviation criteria, and the initial MolProbity web service



Main page

You are using 0% of your 200 Mb of disk space.

FILE UPLOAD/RETRIEVAL (MORE OPTIONS)

PDB/NDB code: type: PDB coords

No file chosen type: PDB coords

CCTBX, which powers this version of MolProbity (4.2), requires stricter adherence to PDB format. If you are having trouble with version 4.2 try using MolProbity4 legacy version 4.02 at http://rutile.biochem.duke.edu. Note that not all features are supported on legacy 4.02 and we suggest repairing your PDB format instead.

Walkthroughs & tutorials:

Evaluate X-ray structure: Typical steps for a published X-ray crystal structure or one still undergoing refinement.

Evaluate NMR structure: Typical steps for a published NMR ensemble or one still undergoing refinement.

Fix up structure: Rebuild the model to remove outliers as part of the refinement cycle.

Work with kinemages: Create and view interactive 3-D graphics from your web browser.

Guide to validation options: Choose validations appropriate to a structure.

What's new in 4.2:

- Cis-peptide identification, markup and statistics now available.
- CaBLAM Cα-based validation of protein backbone and secondary structure now available.
- Rotamer validation now uses Top8000 rotamer distributions. Favored vs Allowed distinction introduced for rotamers.
- New coloring scheme in multicriterion chart encodes outlier severity at a glance.
- Online tutorial explaining markup and methods
- We now use a conformation-dependent library (CDL) from Dunbrack and Karplus for geometry analysis if requested.

What's new in 4.1:

- Validation analysis now powered by CCTBX. Also adds geometry regularization for N/Q/H flip corrections. Read more about this change here.
- Alternate conformations now handled for validation analysis.

It appears that Java is not enabled in your browser, so you may not be able to use KiNG interactive graphics.
Click here for help.

Common questions:

Cite MolProbity: Chen et al. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica D66:12-21.

and/or

Davis et al. (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Research 35:W375-W383.

Cite KiNG: Chen et al. (2009) KiNG (Kinemage, Next Generation): A versatile interactive molecular and scientific visualization program. Protein Science 18:2403-2409.

Cite CCTBX: Grosse-Kunstleve et al. (2002) The Computational Crystallography Toolbox: crystallographic algorithms in a reusable software framework. J. Appl. Cryst. 35:126-136.

About hydrogens: Why have the hydrogen bondlengths changed?

Installing Java: how to make kinemage graphics work in your browser.

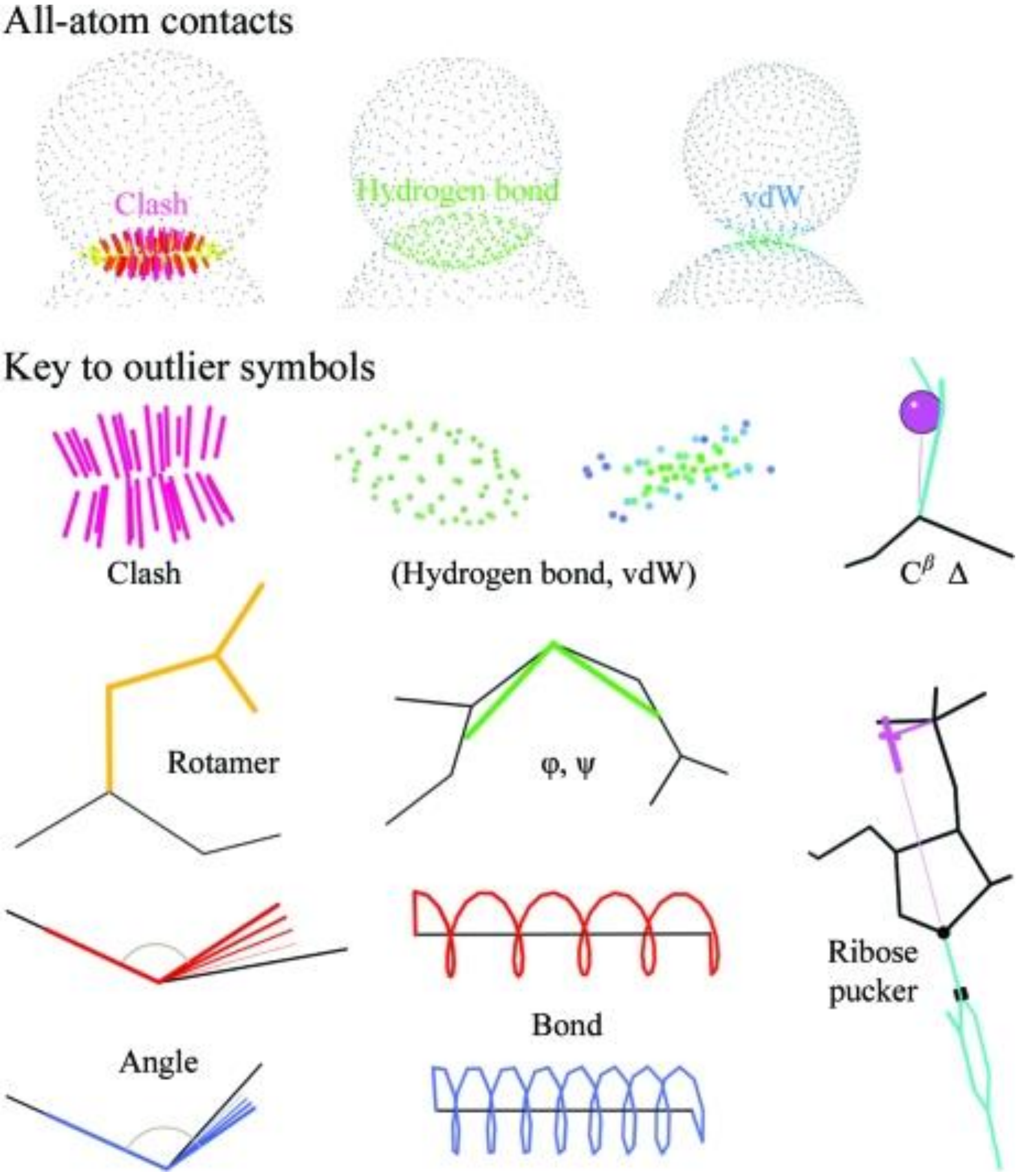
Download MolProbity: how can I run a private MolProbity server, or run from the command line?

NB: the back button doesn't work inside MolProbity

About MolProbity | Website for the Richardson Lab | Using ecloud x-H | Internal reference 4.2

- Addition of H atoms
- The presence of H atoms (both nonpolar and polar) is a critical prerequisite for all-atom contact analysis. Although refinement using H atoms is becoming more common, most crystal structures are still deposited without H atoms.
- Once a PDB structure file has been uploaded, MolProbity detects whether the file contains a suitable number of H atoms; if not, then the ‘Add H atoms’ option is presented to users first.
- *MolProbity* uses the software *REDUCE* to add and optimize hydrogen positions in both protein and nucleic acid structures, including ligands, but does not add explicit H atoms to waters
- A common problem is that the side-chain ends of Asn, Gln and His are easily fitted 180° backwards, since the electron density alone cannot usually distinguish the correct choice of orientation.

- REDUCE can automatically diagnose and correct these types of systematic errors by considering all-atom steric overlaps as well as hydrogen bonding within each local network.
- Automatic correction of Asn/Gln/His flips is the default option in MolProbity during addition of H atoms.
- MolProbity presents each potential flip correction to the user in kinemage view so they have the option of inspecting the before-and-after effects of each flip and approving (or rejecting) each correction



- Users can also choose to add H atoms without Asn/Gln/His flips, which is useful in evaluating the atomic coordinates as they were deposited, but which rejects the easiest and most robustly correct improvement that can be made in a crystallographic model

- All-atom contact analysis

- Once H atoms have been added to (or detected in) a structure, then the complete ‘Analyze all-atom contacts and geometry’ option is enabled.

- A main feature of this option is the all-atom contact analysis, which is performed by the program PROBE
- PROBE operates by, in effect, rolling a 0.5 Å diameter ball around the van der Waals surfaces of atoms to measure the amount of overlap between pairs of nonbonded atoms.
- When non-donor–acceptor atoms overlap by more than 0.4 Å, PROBE denotes the contact as a serious clash, which is included in the reported clashscore and is shown in kinemage format as a cluster of hot-pink spikes in the overlap region
- The ‘clashscore’ is the number of serious clashes per 1000 atoms.
- It is reported in the MolProbity summary with a red/yellow/green color coding for absolute quality.
- The structure’s percentile rank for clashscore value within the relevant resolution range is also given.
- The overall MolProbity score
- In response to user demand, the ‘MolProbity score’ provides a single number that represents the central MolProbity protein quality statistics.
- It is a log-weighted combination of the clashscore, percentage Ramachandran not favored and percentage bad side-chain rotamers, giving one number that reflects the crystallographic resolution at which those values would be expected.
- Therefore, a structure with a numerically lower MolProbity score than its actual crystallographic resolution is, quality-wise, better than the average structure at that resolution.
- There is some distortion in the fit at very high or very low resolutions; for these ranges it is preferable to judge by the resolution-specific percentile score, which is also reported in the summary.
- Percentile scores are currently given for clashscore and for MolProbity score relative to the cohort of PDB structures within 0.25 Å of the file’s resolution

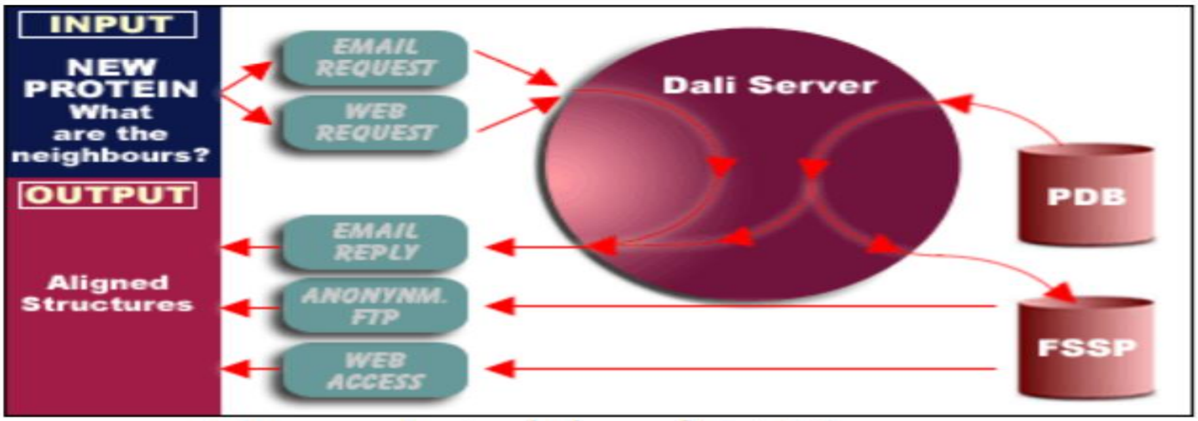
UNIT –IV- Structural Bioinformatics – SBI1403

The FSSP database: Fold Classification based on Structure–Structure alignment of Proteins

- Fold classification based on the structure-structure alignment of proteins and families of structurally similar proteins (FSSP) is a database based on the structural alignment of pair wise combinations of proteins in the Protein Data Bank.
- The Alignments and classification of proteins are done automatically and are updated continuously by the DALI search engine.
- The similarities can be detected by structural comparisons that merge protein families of known 3-D structure into structural classes, the members of which can or might not be evolutionarily related.

Hierarchical clustering in FSSP

- Hierarchical clustering supported structural similarities yields a fold tree that defines 253 folds classes.
- For every representative protein chain, there's a information entry containing structure-structure alignments with its structural neighbours among the PDB.
- The information is accessible online through the World Wide Web browsers and by anonymous ftp (File Transfer Protocol).
- The outline of fold space and therefore the individual datasets offer an upscale supply of data for the study of each divergent and oblique aspects of molecular evolution and for outlining helpful check sets and a regular of truth for assessing the correctness of sequence-sequence or sequence-structure alignments.
- The DALI (Figure 3) database is accessible over the www addressing URL http://ekhidna.biocenter.helsinki.fi/dali/start/.
- The DALI or Distance mAtrix aLIgnment server is a network service used to compare three-dimensional protein structures.
- The query sequence coordinates are compared against those inside the PDB.
- A multiple alignment of structural neighbors is that the output.
- The DALI server is helpful to compare 3D structures wherever similarities don't seem to be detectable by
- comparing sequences directly.
- The comparison uses Max Sprout program to generate backbone and side-chain coordinates if these are not submitted along with the query sequence.
- Secondary structure elements and domains are defined using the DSSP and PUU programs.
- It is additionally attainable to understand the structural neighbours of a protein already within the Protein Data bank from the FSSPdatabase.



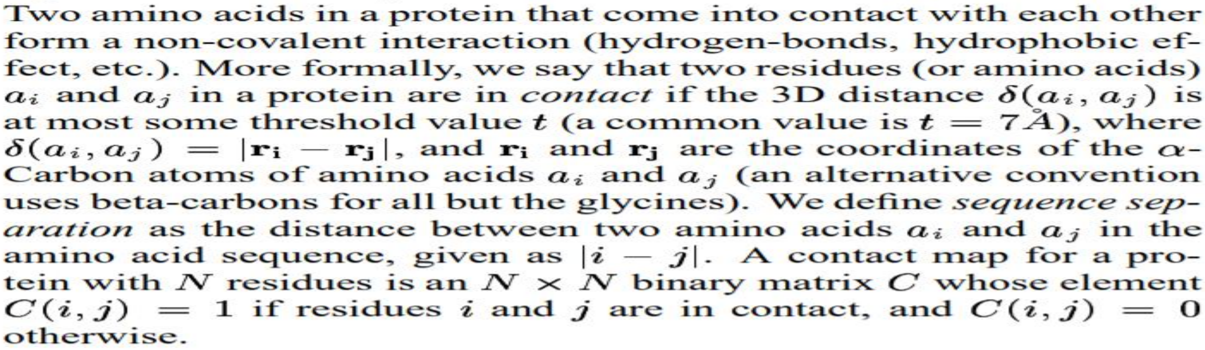
Difference between FSSP and DALI

- The major difference between the two classification schemes, relevant to our work, is their degree of automation.

- FSSP relies on a fully automated structure comparison algorithm, DALI, that calculates a structural similarity measure (represented in terms of Z-score) between pairs of structures of protein chains taken from the PDB.
- A tree is then created by average linkage bunch of the structural similarity score. The tree is cut at DALI Z-score.
- The primary level (Z > 2) can be used as operational definition of folds.

Protein Contact Map

- The 3D conformation of a protein may be compactly represented in a symmetrical, square, boolean matrix of pairwise, inter-residue contacts, or "contact map."
- The contact map of a protein is a particularly useful representation of protein structure.
- The contact map provides useful information about the protein's secondary structure, and it also captures non-local interactions giving clues to its tertiary structure.



Two amino acids in a protein that come into contact with each other form a non-covalent interaction (hydrogen-bonds, hydrophobic effect, etc.). More formally, we say that two residues (or amino acids) a_i and a_j in a protein are in *contact* if the 3D distance $\delta(a_i, a_j)$ is at most some threshold value t (a common value is $t = 7\mathring{A}$), where $\delta(a_i, a_j) = |\mathbf{r_i} - \mathbf{r_j}|$, and $\mathbf{r_i}$ and $\mathbf{r_j}$ are the coordinates of the α-Carbon atoms of amino acids a_i and a_j (an alternative convention uses beta-carbons for all but the glycines). We define *sequence separation* as the distance between two amino acids a_i and a_j in the amino acid sequence, given as $|i - j|$. A contact map for a protein with N residues is an $N \times N$ binary matrix C whose element $C(i, j) = 1$ if residues i and j are in contact, and $C(i, j) = 0$ otherwise.

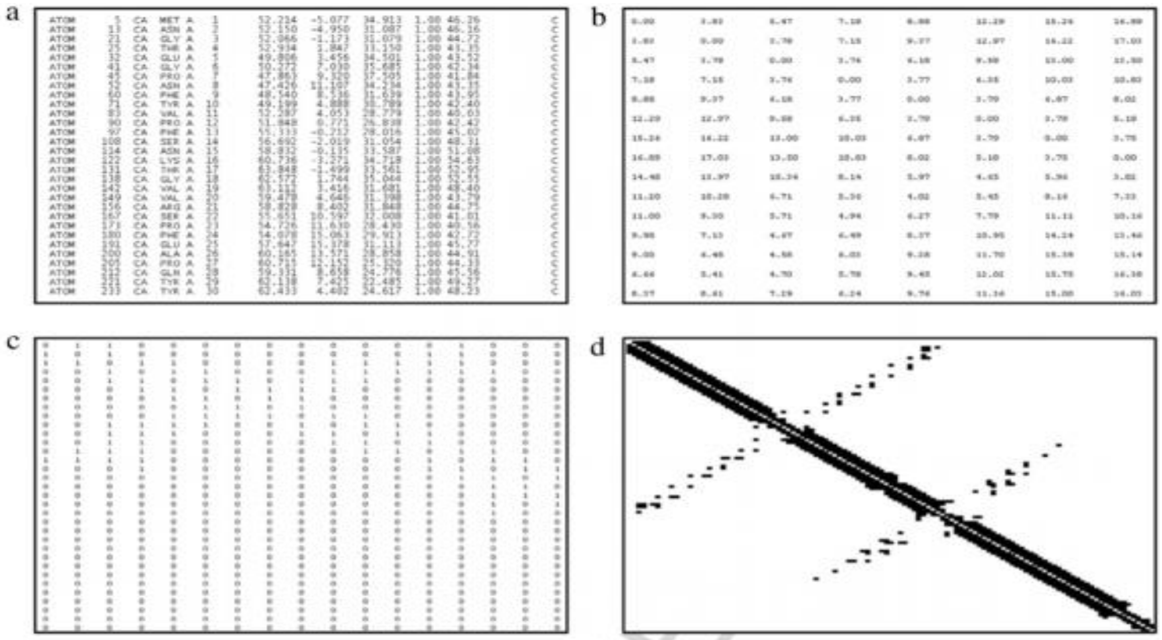
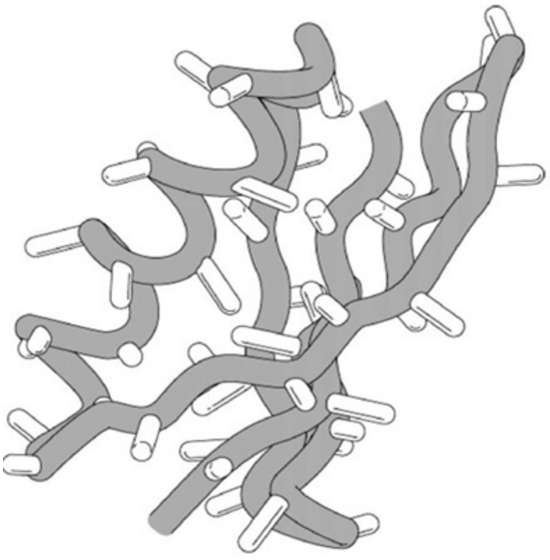
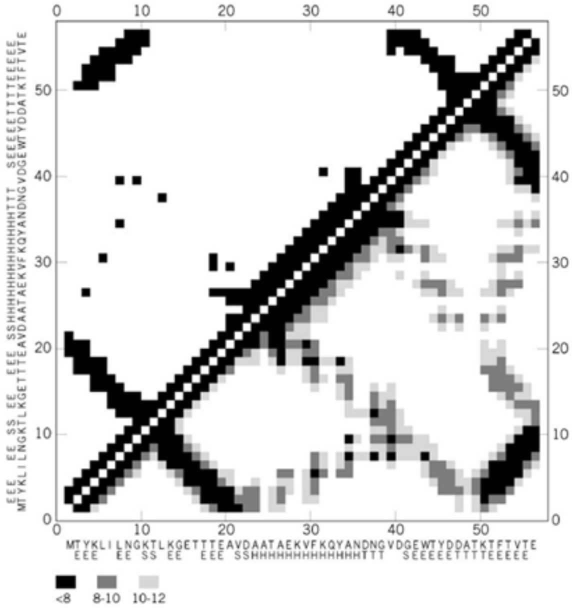


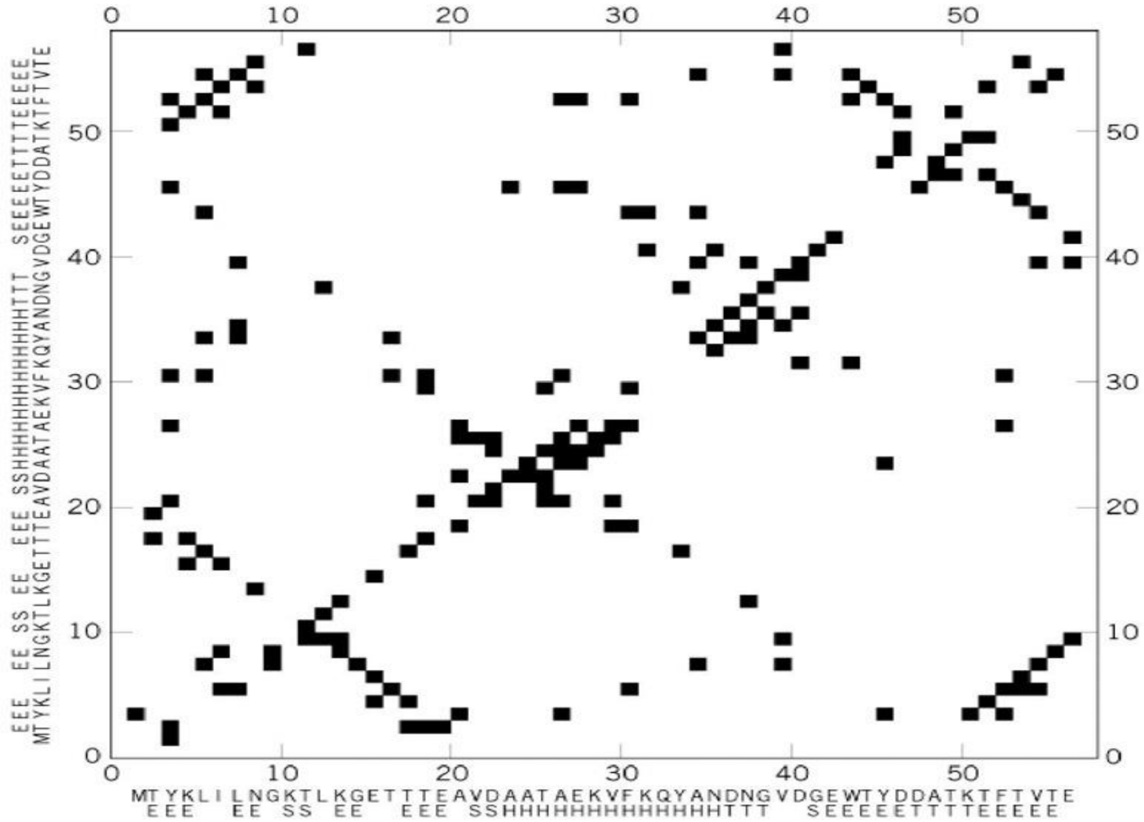
Fig. 1. Construction of protein contact map. (a) PDB atomic coordinates for the protein bovine rhodopsin (PDB_ID: 1U19a). (b) Distance matrix calculated using Euclidean distance. (c) Protein contact map obtained using a cut-off distance. (d) Visualization of protein contact map.



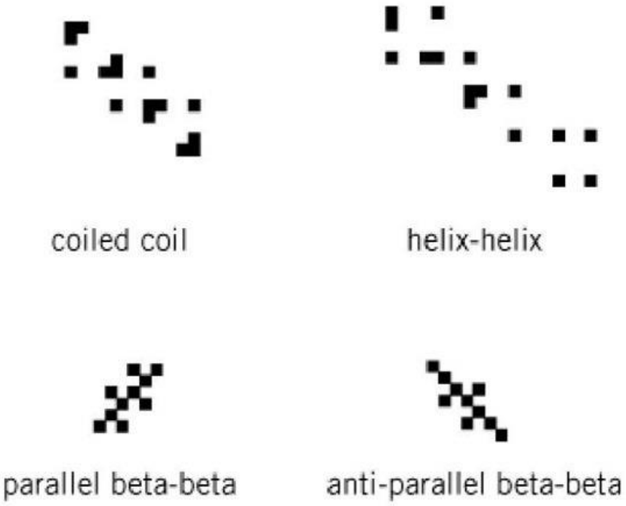
Three-dimensional structure of the B domain



Ca-based contact maps: above the diagonal, the black and white map with a cutoff distance of 8 A; below the diagonal is the gray-scale map where various shades of gray correspond to three values of the cutoff distance, 8 (darkest), 10, and 12 A.



Side-chain-based contact map. Above the diagonal, the dark squares correspond to the pairs of side chains for which the distance between at least one pair of heavy atoms is less than 5 A. Below the diagonal, a 6.25-A cutoff criterion has been applied to the centers of mass of the side chains.

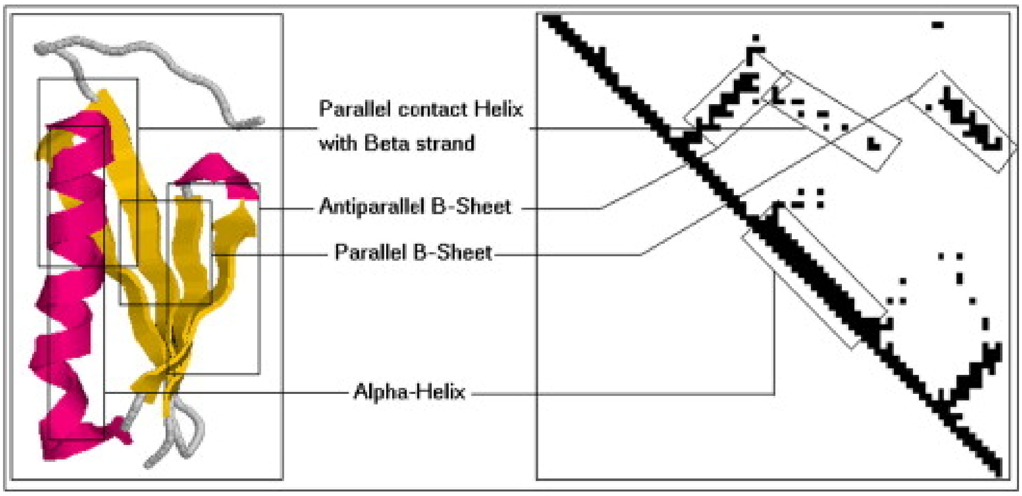


Representative patterns of side-chain contact maps describing interactions between a-helices (top) and between b-strands in parallel and antiparallel beta-sheets (bottom).

- Contact Maps as a Fingerprint of Protein Three-Dimensional Structure
- A contact map constitutes a structural "fingerprint" of a protein.
- Each protein can be identified based on its contact map.
- The secondary structure, fold topology, and side-chain packing patterns (for side-chain contact maps) can be visualized conveniently and read from the contact map.
- Furthermore, structural similarity between a pair of proteins is immediately apparent by a very pronounced similarity of their contact maps; in comparing two protein structures, there is no need to search all their possible relative orientations.
- The reconstruction of a protein structure from its contact map is more complex, although low-to-moderate resolution three-dimensional models can be easily built, even from a fragmentary contact map.
- The accuracy of the model depends on the type of contact map and the computational tools employed.
- A combination of protein nuclear magnetic resonance NMR spectra (see NOESY Spectrum; COSY Spectrum) constitutes a hybrid contact map of a protein, and model building from these data is an example of a map-to-structure modeling procedure

Ca-Based Contact Maps

- Ca-based **contact maps and distance matrices were perhaps the first commonly used maps for visualization of protein structures.**
- These contact maps reflect well the overall topology of the protein fold, but only rather coarse structural details can be read from them.
- This is due to the fact that the Ca-Ca distance distributions extracted from protein structures have several convoluted peaks. These peaks correspond to various distances between pairs of various secondary structure elements (a-helices, beta-strands, etc.).
- Thus, a single cutoff distance is always inadequate: Too small a value would miss some helix-to-helix contacts, while too large a value may create some problems with identification of the secondary structure patterns.
- Gray scale maps (several cutoff ranges) communicate much more detailed structural information.



Side-Chain Contact Maps

Side-chain contact maps contain much richer information, not only about the topology of a protein fold and its secondary structure, but also many fine details about the packing patterns of the protein side-chains.

Various conventions can be used to build side-chain contact maps.
In one case, two residues are assumed to be in contact when any two heavy atoms (ie, all except hydrogen) are a shorter distance from each other than some assumed cutoff.

Due to the comparable size of all the united atom types constituting the side chains (eg, CH2, CH, NH^, etc.), a good choice of cutoff distance is between 4.5 and 5.0 A .

In this range, the number of detected contacts is not sensitive to the particular choice of cutoff, and the packing pattern of the side chains is always described with high fidelity.

Characteristic patterns of contacts between elements of secondary structure are an important and useful feature of these contact maps.

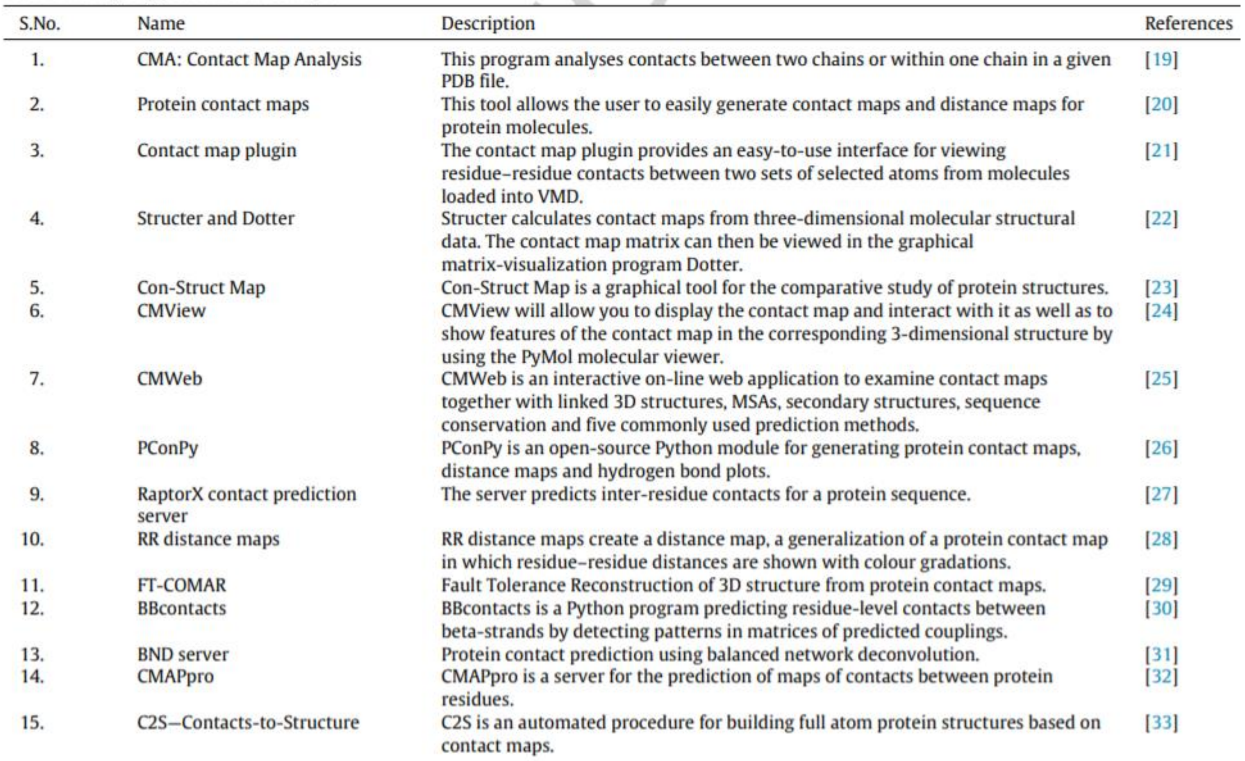
Alternatively, one may build a side-chain contact map using the side-chain centers of mass as a reference. In this case, a larger value of the cutoff distance needs to be used.

These cutoff values for the distance between side-chain centers of mass could be made specific for certain amino acid pairs, on the basis of the different sizes of the side chains, which produce different average contact distances for various pairs. the two approaches (atom based and center of mass based) lead to very similar protein representations. The patterns of the atom-based contact maps are slightly better defined.

Regularities of the Contact Maps Reflect Regularities of Protein Structures

- Different types of contact maps reflect different aspects of the regularities seen in protein structures.
- The side-chain-based contact maps are a very good example.
- Near the diagonal of the map, the distinct features of the protein secondary structure can be easily read. Indeed, for extended fragments of the polypeptide chain, only residues i and i + 2 can be in contact. For a-helices, the i, i + 3 and i, i + 4 patterns of contacts are well pronounced.
- Furthermore, characteristic clusters of contacts further away from the diagonal reflect the packing between particular pairs of b-strands within the b-sheets.
- Parallel and antiparallel structures have very different features on the contact patterns.
- Very characteristic patterns could also be observed for other pairs of secondary-structure elements.
- It is even very easy to distinguish between the patterns for two helices in a helical or a/b protein and in the coiled-coil structural motifs.

Tools for analysing protein contact maps.



S.No.	Name	Description	References
1.	CMA: Contact Map Analysis	This program analyses contacts between two chains or within one chain in a given PDB file.	[19]
2.	Protein contact maps	This tool allows the user to easily generate contact maps and distance maps for protein molecules.	[20]
3.	Contact map plugin	The contact map plugin provides an easy-to-use interface for viewing residue–residue contacts between two sets of selected atoms from molecules loaded into VMD.	[21]
4.	Structer and Dotter	Structer calculates contact maps from three-dimensional molecular structural data. The contact map matrix can then be viewed in the graphical matrix-visualization program Dotter.	[22]
5.	Con-Struct Map	Con-Struct Map is a graphical tool for the comparative study of protein structures.	[23]
6.	CMView	CMView will allow you to display the contact map and interact with it as well as to show features of the contact map in the corresponding 3-dimensional structure by using the PyMol molecular viewer.	[24]
7.	CMWeb	CMWeb is an interactive on-line web application to examine contact maps together with linked 3D structures, MSAs, secondary structures, sequence conservation and five commonly used prediction methods.	[25]
8.	PConPy	PConPy is an open-source Python module for generating protein contact maps, distance maps and hydrogen bond plots.	[26]
9.	RaptorX contact prediction server	The server predicts inter-residue contacts for a protein sequence.	[27]
10.	RR distance maps	RR distance maps create a distance map, a generalization of a protein contact map in which residue–residue distances are shown with colour gradations.	[28]
11.	FT-COMAR	Fault Tolerance Reconstruction of 3D structure from protein contact maps.	[29]
12.	BBcontacts	BBcontacts is a Python program predicting residue-level contacts between beta-strands by detecting patterns in matrices of predicted couplings.	[30]
13.	BND server	Protein contact prediction using balanced network deconvolution.	[31]
14.	CMAPpro	CMAPpro is a server for the prediction of maps of contacts between protein residues.	[32]
15.	C2S—Contacts-to-Structure	C2S is an automated procedure for building full atom protein structures based on contact maps.	[33]

The Biomolecular Interaction Network Database (BIND)

The Biomolecular Interaction Network Database (BIND) (http://bind.ca) archives biomolecular interaction, reaction, complex and pathway information. Theri aim is to curate the details about molecular interactions that arise from published experimental research and to provide this information, as well as tools to enable data analysis, freely to researchers worldwide.

BINDdata are curated into acomprehensive machinereadable archive of computable information and providesuserswithmethods to discover interactionsand molecular mechanisms. BIND has worked to develop new methods for visualization that amplify the underlying annotation of genes and proteins to facilitate the study of molecular interaction networks. BIND has maintained an open database policy since its inception in 1999. Data growth has proceeded at a tremendous rate, approaching over 100 000 records. New services provided include a new BIND Query and Submission interface, a Standard Object Access Protocol service and the Small Molecule Interaction Database (http://smid.blueprint.org) that allows users to determine probable small molecule binding sites of new sequences and examine conserved binding residues.

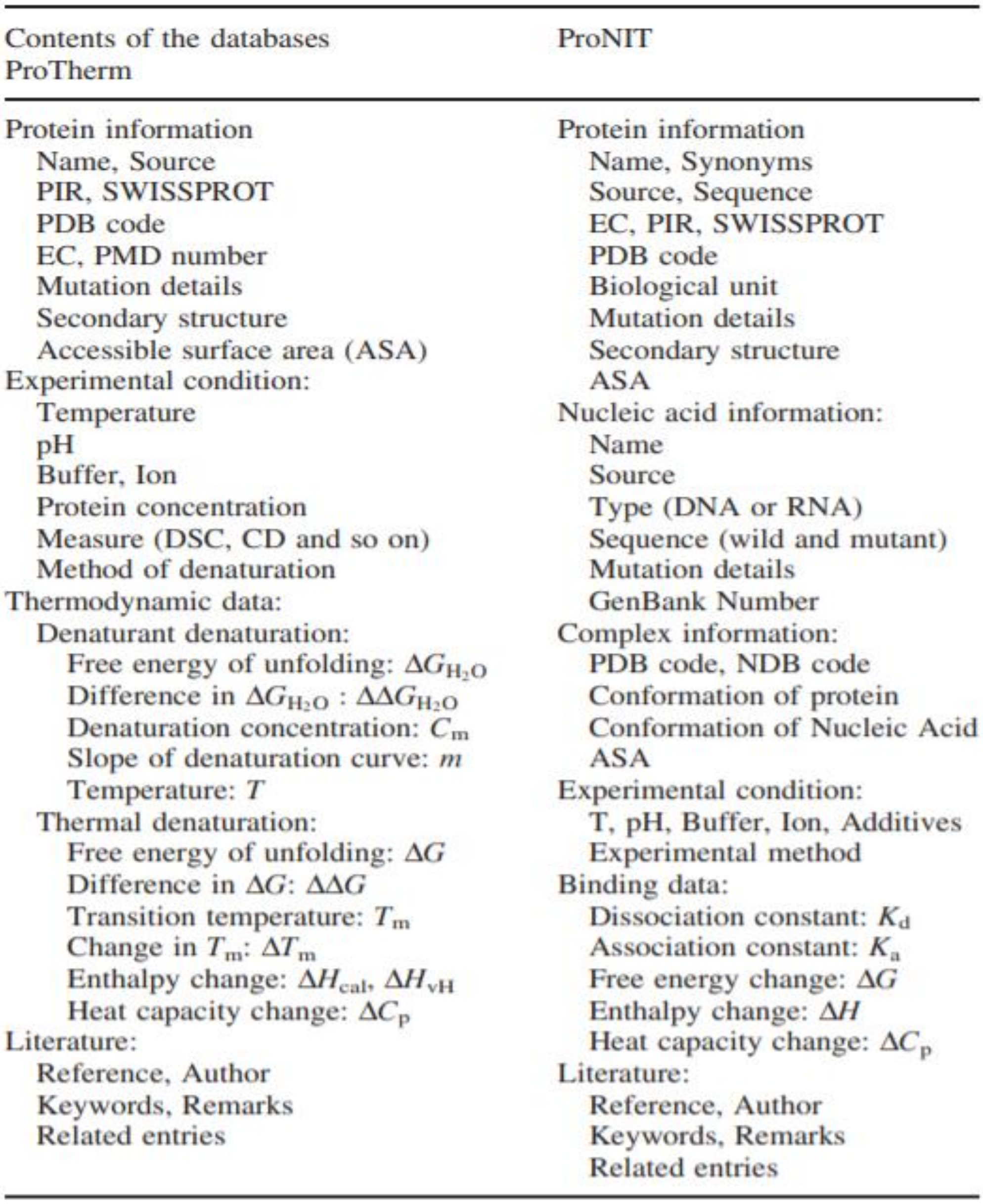
What you can do:

Find biomolecular interaction, complex and pathway information.

Highlights:

- BIND is a collection of records documenting molecular interactions, including high-throughput data submissions and hand-curated information gathered from the scientific literature.
- A BIND record represents an interaction between two or more objects that is believed to occur in a living organism. A biological object can be a protein, DNA, RNA, ligand, molecular complex, gene, photon or an unclassified biological entity.
- BIND records are created for interactions which have been shown experimentally and published in at least one peer-reviewed journal. A record also references any papers with experimental evidence that support or dispute the associated interaction.
- Data from the PDB and a number of large-scale interaction and complex mapping experiments using yeast two hybrid, mass spectrometry, genetic interactions and phage display are added.
- A new graphical analysis tool provides users with a view of the domain composition of proteins in interaction and complex records to help relate functional domains to protein interactions.

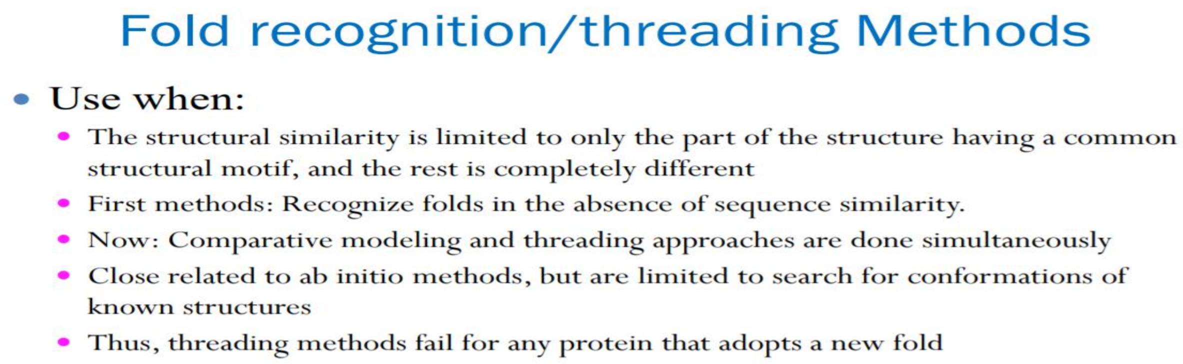
- In light of the vast scientific resources made available through genomics, the science of deciphering molecular mechanisms is expanding rapidly. Scientists who once hunted for disease genes or sought to distinguish key concepts in evolution are now turning their attention to the details of molecular assembly and mechanism to further understand medicine and the key concepts underlying biology.
- The Biomolecular Interaction Network Database (BIND) was designed to store complete information about molecular assembly through a database structure in order to archive interactions and reactions arising from biopolymers (protein, RNA and DNA), as well as small molecules, lipids and carbohydrates.
- **ProNIT**
- **ProTherm and ProNIT** are two thermodynamic databases that contain experimentally determined thermodynamic parameters of protein stability and protein–nucleic acid interactions, respectively.
- Thermodynamic database for proteins and mutants (ProTherm) and thermodynamic database for protein–nucleic acid interactions (ProNIT) are two comprehensive, integrated databases that document experimentally determined thermodynamic parameters published in the literature
- Both ProTherm and ProNIT include several thermodynamic parameters along with sequence and structural information, experimental methods and conditions, and literature information.



Contents of the databases ProTherm	ProNIT
Protein information	Protein information
Name, Source	Name, Synonyms
PIR, SWISSPROT	Source, Sequence
PDB code	EC, PIR, SWISSPROT
EC, PMD number	PDB code
Mutation details	Biological unit
Secondary structure	Mutation details
Accessible surface area (ASA)	Secondary structure
Experimental condition:	ASA
Temperature	Nucleic acid information:
pH	Name
Buffer, Ion	Source
Protein concentration	Type (DNA or RNA)
Measure (DSC, CD and so on)	Sequence (wild and mutant)
Method of denaturation	Mutation details
Thermodynamic data:	GenBank Number
Denaturant denaturation:	Complex information:
Free energy of unfolding: ΔG_{H_2O}	PDB code, NDB code
Difference in ΔG_{H_2O} : $\Delta\Delta G_{H_2O}$	Conformation of protein
Denaturation concentration: C_m	Conformation of Nucleic Acid
Slope of denaturation curve: m	ASA
Temperature: T	Experimental condition:
Thermal denaturation:	T, pH, Buffer, Ion, Additives
Free energy of unfolding: ΔG	Experimental method
Difference in ΔG: $\Delta\Delta G$	Binding data:
Transition temperature: T_m	Dissociation constant: K_d
Change in T_m: ΔT_m	Association constant: K_a
Enthalpy change: ΔH_{cal}, ΔH_{vH}	Free energy change: ΔG
Heat capacity change: ΔC_p	Enthalpy change: ΔH
Literature:	Heat capacity change: ΔC_p
Reference, Author	Literature:
Keywords, Remarks	Reference, Author
Related entries	Keywords, Remarks
	Related entries

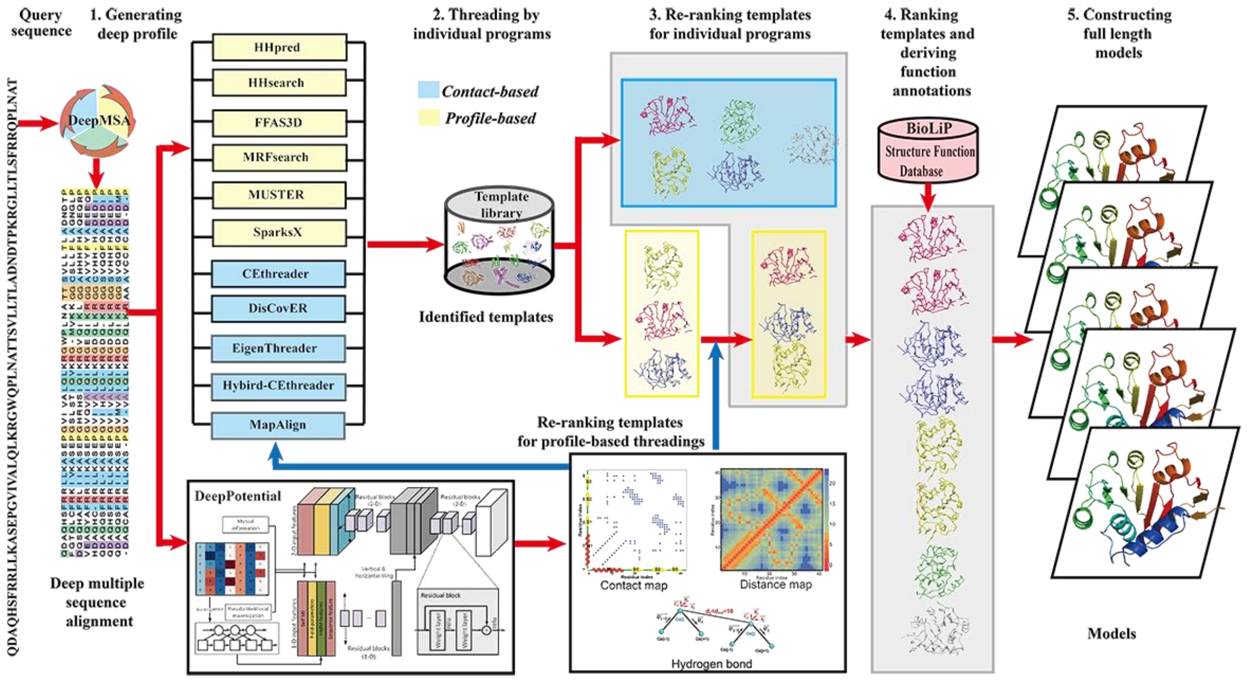
Protein threading

Protein threading (fold recognition) is protein modeling method done for those proteins whose folds are same as some known proteins , but they don't have homologous protein with known structure in protein data bank.



Fold recognition/threading Methods

- Use when:
  - The structural similarity is limited to only the part of the structure having a common structural motif, and the rest is completely different
  - First methods: Recognize folds in the absence of sequence similarity.
  - Now: Comparative modeling and threading approaches are done simultaneously
  - Close related to ab initio methods, but are limited to search for conformations of known structures
  - Thus, threading methods fail for any protein that adopts a new fold


Query sequence
1. Generating deep profile
2. Threading by individual programs
3. Re-ranking templates for individual programs
4. Ranking templates and deriving function annotations
5. Constructing full length models
QDAQHSFRRLLKASEPGVIVALQLKRGWQPLNATTSVLLTLADNDTPKRGLLTSFRRQPLNAT
DeepMSA
Deep multiple sequence alignment
HHpred
HHsearch
FFAS3D
MRFsearch
MUSTER
SparksX
CEthreader
DisCovER
EigenThreader
Hybird-CEthreader
MapAlign
Contact-based
Profile-based
Template library
Identified templates
BioLiP
Structure Function Database
Re-ranking templates for profile-based threadings
DeepPotential
Contact map
Distance map
Hydrogen bond
Models

- LOMETS (LOcal MEta-Threading-Server, version 3) is a meta-server method for protein structure prediction and function annotation.
- It generates protein structure predictions by ranking and selecting models from multiple state-of-the-art threading programs.
- Starting from a query sequence, deep multiple sequence alignments (MSAs) are generated by iterative sequence homology searches through multiple sequence databases.
- These MSAs are used as inputs into 11 threading programs, which are all locally installed on our cluster, to identify structural templates from the PDB library.
- The MSAs are also used to predict residue-residue contacts, distances, and hydrogen bond geometries, that are used in the 5 contact-based threading programs.

- These predicted terms along with the profile score from original profile-based threading are used to re-rank the templates detected by the individual threading programs.
- The top templates are ranked and selected by a score that combines the alignment Z-score, program-specific confidence scores and the sequence identity to the query.
- The functional annotations (including gene ontology terms, enzyme commission number, and ligand binding pockets) are generated by searching the template structures through the BioLiP function library.
- Then, the 5 full-length models are constructed by MODELLER from top 5 templates for homologous targets, or by L-BFGS for non-homologous targets using the distance restraints predicted by DeepPotential and calculated from top templates.
- FG-MD and FASPR will be used to refine the global topology and re-pack the side-chain conformation of the final models.
- Finally, structural analogs in PDB are detected by TM-align by matching the first LOMETS3 model to all structures in the PDB library.

LOMETS3 reports the top 10 proteins from the PDB that have the closest structural similarity, i.e., the highest TM-score, to the predicted model, associated with the functional annotation
TM-score is a metric for assessing the topological similarity of protein structures.
It is designed to solve two major problems in traditional metrics such as root-mean-square deviation (RMSD):

(1) TM-score weights smaller distance errors stronger than larger distance errors and makes the score value more sensitive to the global fold similarity than to the local structural variations;
(2) TM-score introduces a length-dependent scale to normalize the distance errors and makes the magnitude of TM-score length-independent for random structure pairs.

TM-score has the value in (0,1], where 1 indicates a perfect match between two structures. Following strict statistics of structures in the PDB, scores below 0.17 correspond to randomly chosen unrelated proteins whereas structures with a score higher than 0.5 assume generally the same fold in SCOP/CATH.

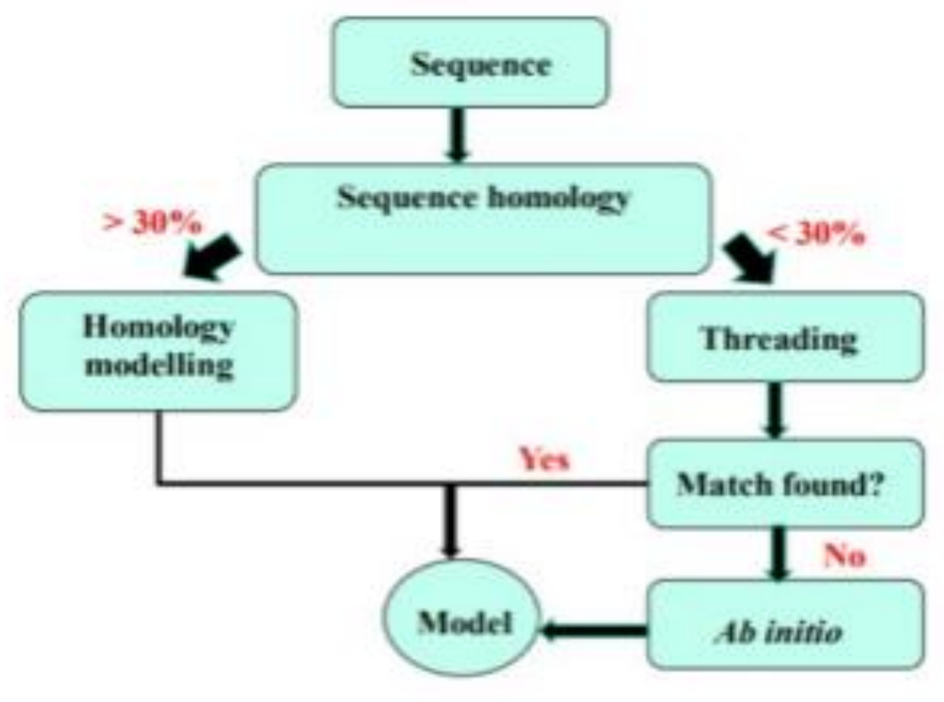
- I-TASSER (Iterative Threading ASSEmbly Refinement) is a hierarchical approach to protein structure prediction and structure-based function annotation.
- It first identifies structural templates from the PDB by multiple threading approach LOMETS, with full-length atomic models constructed by iterative template-based fragment assembly simulations.
- Function insights of the target are then derived by re-threading the 3D models through protein function database BioLiP.
- I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent community-wide CASP7, CASP8, CASP9, CASP10, CASP11, CASP12, CASP13, and CASP14 experiments.
- It was also ranked the best for function prediction in CASP9.
- The server is in active development with the goal to provide the most accurate protein structure and function predictions using state-of-the-art algorithms.

What is difference and relationship between C-score and TM-score?

- TM-score (or RMSD) is a known standard for measuring structural similarity between two structures which are usually used to measure the accuracy of structure modeling when the native structure is known, while C-score is a metric that I-TASSER developed to estimate the confidence of the modeling.
- In case where the native structure is not known, it becomes necessary to predict the quality of the modeling prediction, i.e. what is the distance between the predicted model and the native structures?

- To answer this question, predicted the TM-score and RMSD of the predicted models relative the native structures based on the C-score.
- In a benchmark test set of 500 non-homologous proteins, they found that C-score is highly correlated with TM-score and RMSD.
- Correlation coefficient of C-score of the first model with TM-score to the native structure is 0.91, while the coefficient of C-score with RMSD to the native structure is 0.75.
- These data lay the base for the reliable prediction of the TM-score and RMSD using C-score.
- In the output section, I-TASSER only reports the quality prediction (TM-score and RMSD) for the first model, because it was found that the correlation between C-score and TM-score is weak for lower rank models.
- However, the C-score is listed for all models just for a reference.

Ab Initio Protein Structure Prediction



If protein templates are not available, we have to build the 3D models from scratch.

This procedure has been given different names,

ab initio modeling ;
de novo modeling ;
physics-based modeling
free modelling.
Typically, ab initio modeling conducts a conformational search under the guidance of a designed energy function.
This procedure usually generates a number of possible conformations **(also called structure decoys),** and final models are selected from them.

Therefore, a successful ab initio modeling depends on three factors:

(1) an accurate energy function with which the native structure of a protein corresponds to the most thermodynamically stable state, compared to **all possible decoy structures**

(2) an efficient search method which can quickly identify the low-energy states through conformational search

(3) a strategy that can select near-native models from a pool of decoy structures

Energy functions used for ab initio modeling

We classify the energy functions into two groups:

(a) Physics-based energy functions

(b) Knowledge-based energy functions, depending on whether they make use of statistics from the existing protein 3D structures in the PDB

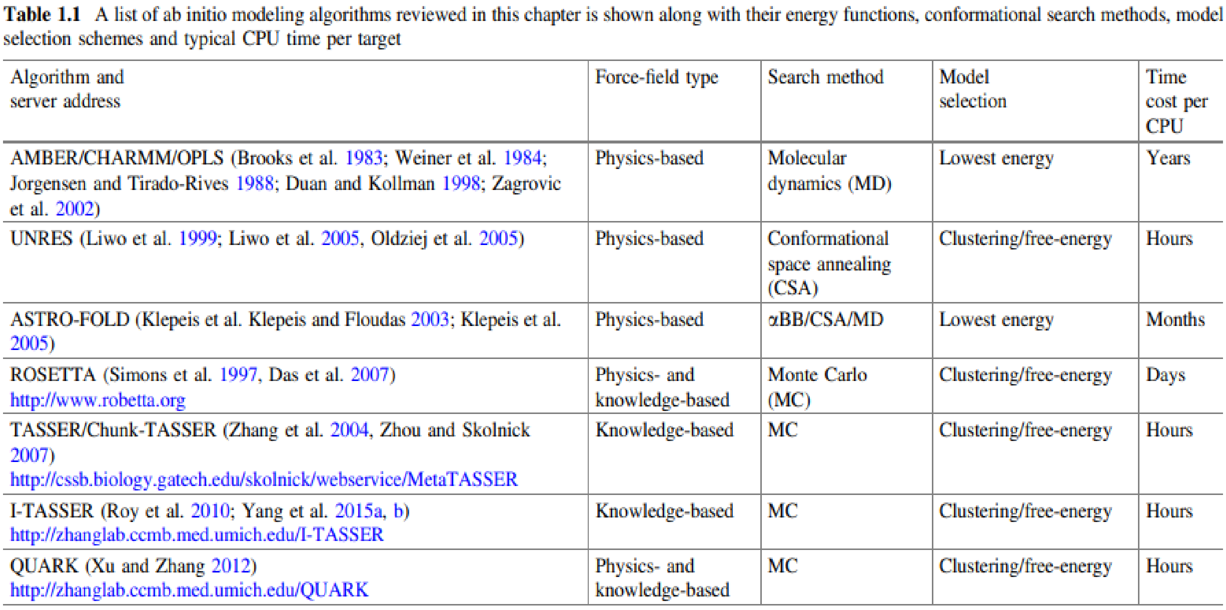
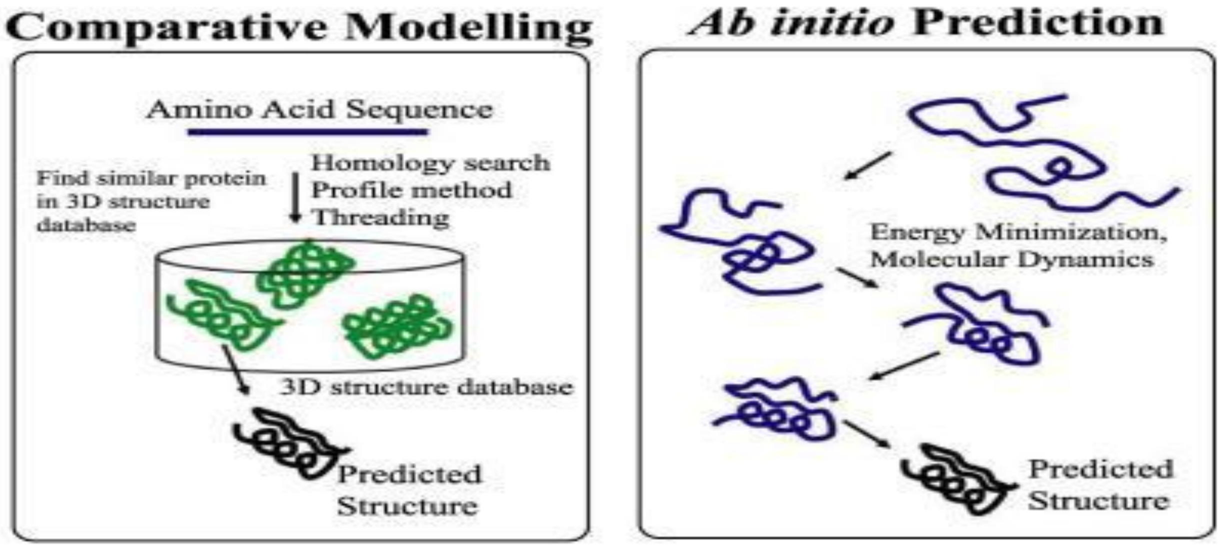


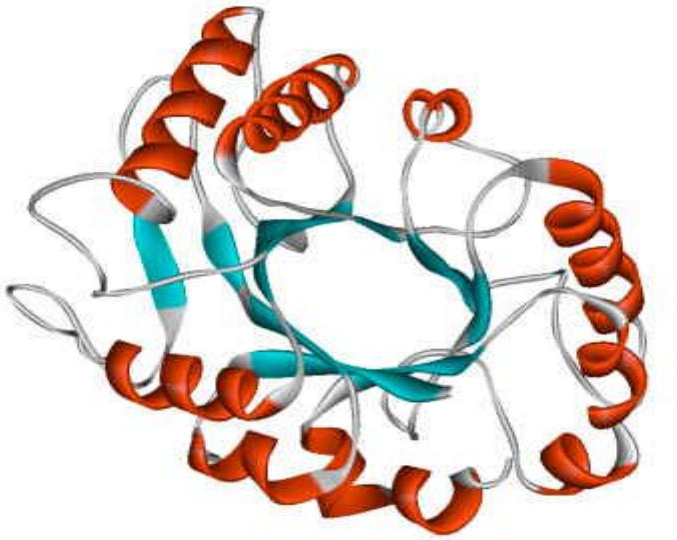
Table 1.1 A list of ab initio modeling algorithms reviewed in this chapter is shown along with their energy functions, conformational search methods, model selection schemes and typical CPU time per target

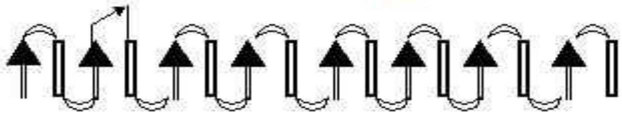
Algorithm and server address	Force-field type	Search method	Model selection	Time cost per CPU
AMBER/CHARMM/OPLS (Brooks et al. 1983; Weiner et al. 1984; Jorgensen and Tirado-Rives 1988; Duan and Kollman 1998; Zagrovic et al. 2002)	Physics-based	Molecular dynamics (MD)	Lowest energy	Years
UNRES (Liwo et al. 1999; Liwo et al. 2005, Oldziej et al. 2005)	Physics-based	Conformational space annealing (CSA)	Clustering/free-energy	Hours
ASTRO-FOLD (Klepeis et al. Klepeis and Floudas 2003; Klepeis et al. 2005)	Physics-based	αBB/CSA/MD	Lowest energy	Months
ROSETTA (Simons et al. 1997, Das et al. 2007) http://www.robetta.org	Physics- and knowledge-based	Monte Carlo (MC)	Clustering/free-energy	Days
TASSER/Chunk-TASSER (Zhang et al. 2004, Zhou and Skolnick 2007) http://cssb.biology.gatech.edu/skolnick/webservice/MetaTASSER	Knowledge-based	MC	Clustering/free-energy	Hours
I-TASSER (Roy et al. 2010; Yang et al. 2015a, b) http://zhanglab.ccmb.med.umich.edu/I-TASSER	Knowledge-based	MC	Clustering/free-energy	Hours
QUARK (Xu and Zhang 2012) http://zhanglab.ccmb.med.umich.edu/QUARK	Physics- and knowledge-based	MC	Clustering/free-energy	Hours



UNIT –V- Structural Bioinformatics – SBI1403

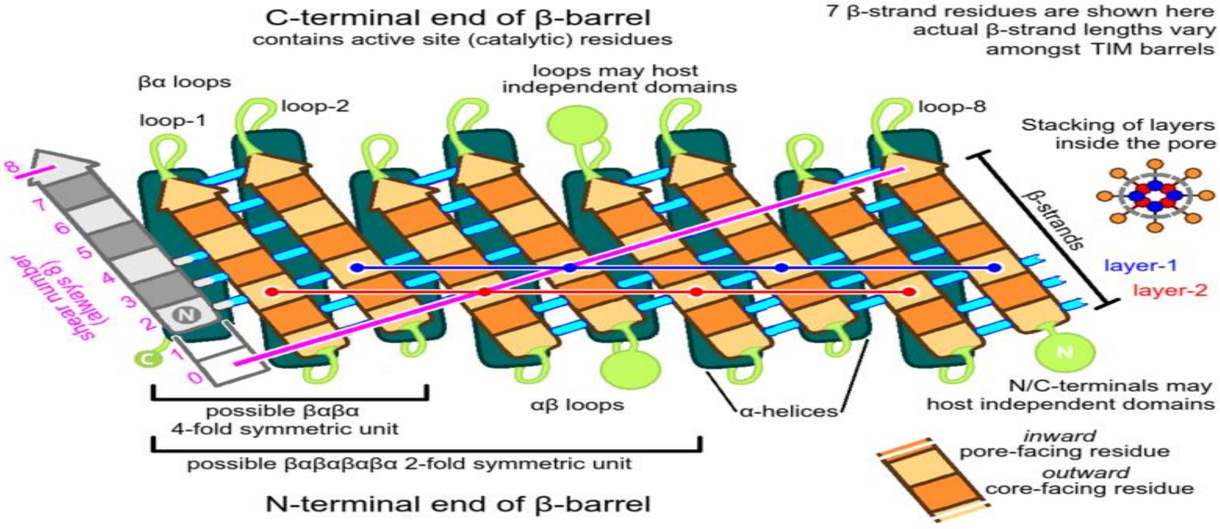
The a /b Barrel Domain



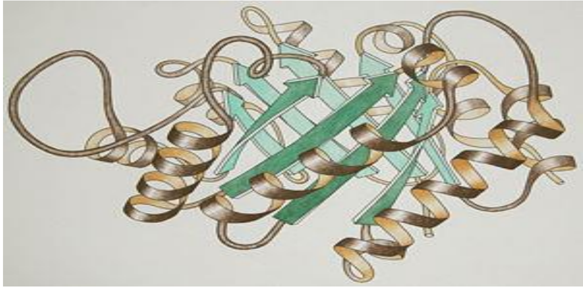


Topology diagram of Hevamine - one of the TIM barrel structures

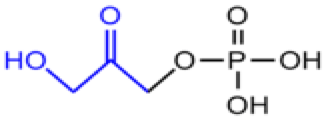
- The first protein that was discovered to have an eight-stranded a /b domain was Triose phosphate isomerase.
- This fold is characterized by a central barrel formed by parallel b -strands surrounded by seven or eight a helices which shield the barrel from solvent.
- The b -strands of the barrel form an intrinsic network of hydrogen bonds with the neighbouring strands and are oriented in the same direction.
- The overall twist associated with all the strands cause the first and the eighth strand to register in parallel held in place by hydrogen bonds causing the closure of the barrel.
- The overall sequence topology can therefore be either (b /a)$_{8.}$ where the protein begins with a strand or (a/b)$_{8.}$ where helix is the first secondary structure. In addition, many (a/b)$_8$ enzymes have additional domains that are not a part of this fold



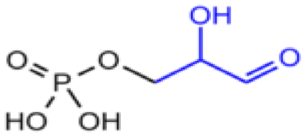
- **TIM barrel topology.**
- α-helices are colored teal, loops are colored green, and β-strands are colored in two shades of orange.
- Lighter shades indicate residues pointing inward, towards the barrel pore.
- Darker shades indicate residues pointing outward, towards the barrel core.
- Cyan lines depict an example backbone β-barrel hydrogen bonding network.
- Interior β-barrel residues (pore residues) display a 4-fold geometric symmetry, despite emerging from an 8-strand β-barrel.
- Each layer contains 4 residues that point towards the pore, and lie on the same plane perpendicular to the barrel axis. The shear number for TIM barrels is always 8, and is illustrated in magenta. Some TIM barrels naturally adopt, or are designed to adopt, two or four-fold symmetry.
- Triose phosphate isomerase (TIM)PDB 1wyi and 1hti) is a crucial enzyme in the glycolytic pathway.
- TIM reversibly converts the aldose Glyceraldehyde-3-phosphate (GAP) to the ketose Dihydroxyacetone phosphate (DHAP).
- The interconversion proceeds by an enediol intermediate.
- Triose phosphate isomerase is not directly regulated, but the enzyme two steps before it in the glycolytic pathway, phosphofructokinase, is a heavily regulated, irreversible enzyme.



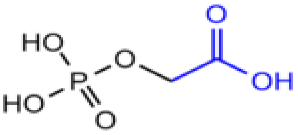
DHAP
Glycolytic substrate



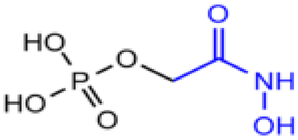
GAP
Glycolytic substrate



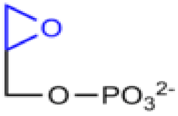
2-Phosphoglycolate
Transition state analog



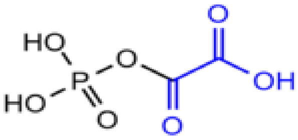
Phosphoglycolohydroxamate
Transition state analog



Glycidol phosphate
Irreversible inhibitor



Phosphoenol pyruvate
Weak reversible inhibitor

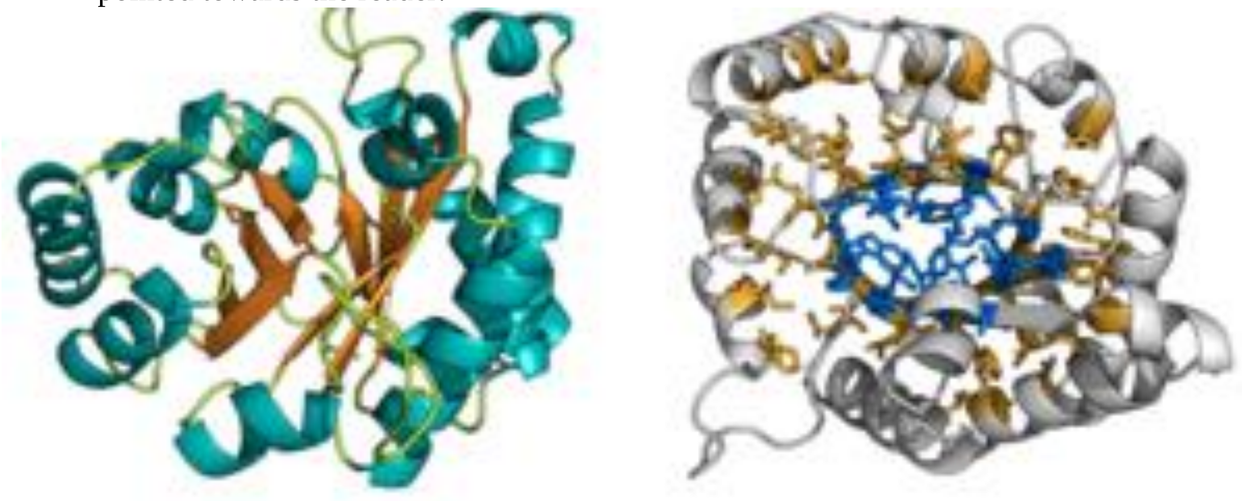


Structural Characteristics
The secondary structure consists of 14 alpha helices and 8 beta sheets per monomer, making it fall in the SCOP category of alpha and beta proteins.
The tertiary structure is a alpha-beta barrel, and it is the prototypical example of the "TIM barrel" fold.
The quaternary structure is a homodimer. The molecular weight of the enzyme is estimated at 57,400 Da.

Triose phosphate isomerase (TIM) isolated from chicken muscles (PDB: 1TIM), the archetypal TIM barrel enzyme.

(A) **Cartoon representation of the TIM barrel structure.** α-helices are colored teal, β-strands are colored orange, and loops are colored green.

(B) Core and pore regions are highlighted. Amino acid residues belonging to the pore are colored blue. Amino acid residues belonging to the core are colored orange. Note that the TIM barrel is depicted in a top-down view, where the C-terminal ends of the β-barrel are pointed towards the reader.

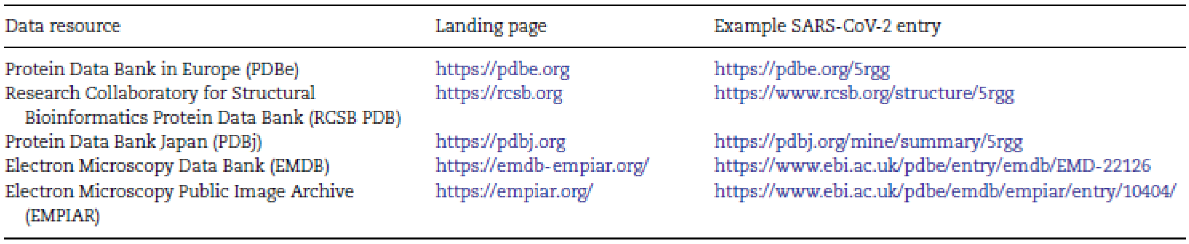


This fold has intrigued many researchers over the past years for two reasons :

- Because a /b barrel proteins catalyze a wide range of reactions, they are primary targets for Protein engineering and drug design.
- To unravel its evolutionary history. The lack of substantial sequence homology between members of this family makes it a challenging target for evolutionary analysis and to trace the ancestry of each member of this class. This fact, combined with geometric arguments concerning the barrel structure, has lead to suggestion that these proteins are related by convergent evolution to a stable fold. Although many studies have been undertaken in this direction, concrete evidence has been lacking and controversies remain. Answers to this questions can have allied benefits for a protein designer in understanding the reasons for such different sequences acquiring the same fold and thus designing a denovo sequence that can fold into an a /b barrel.

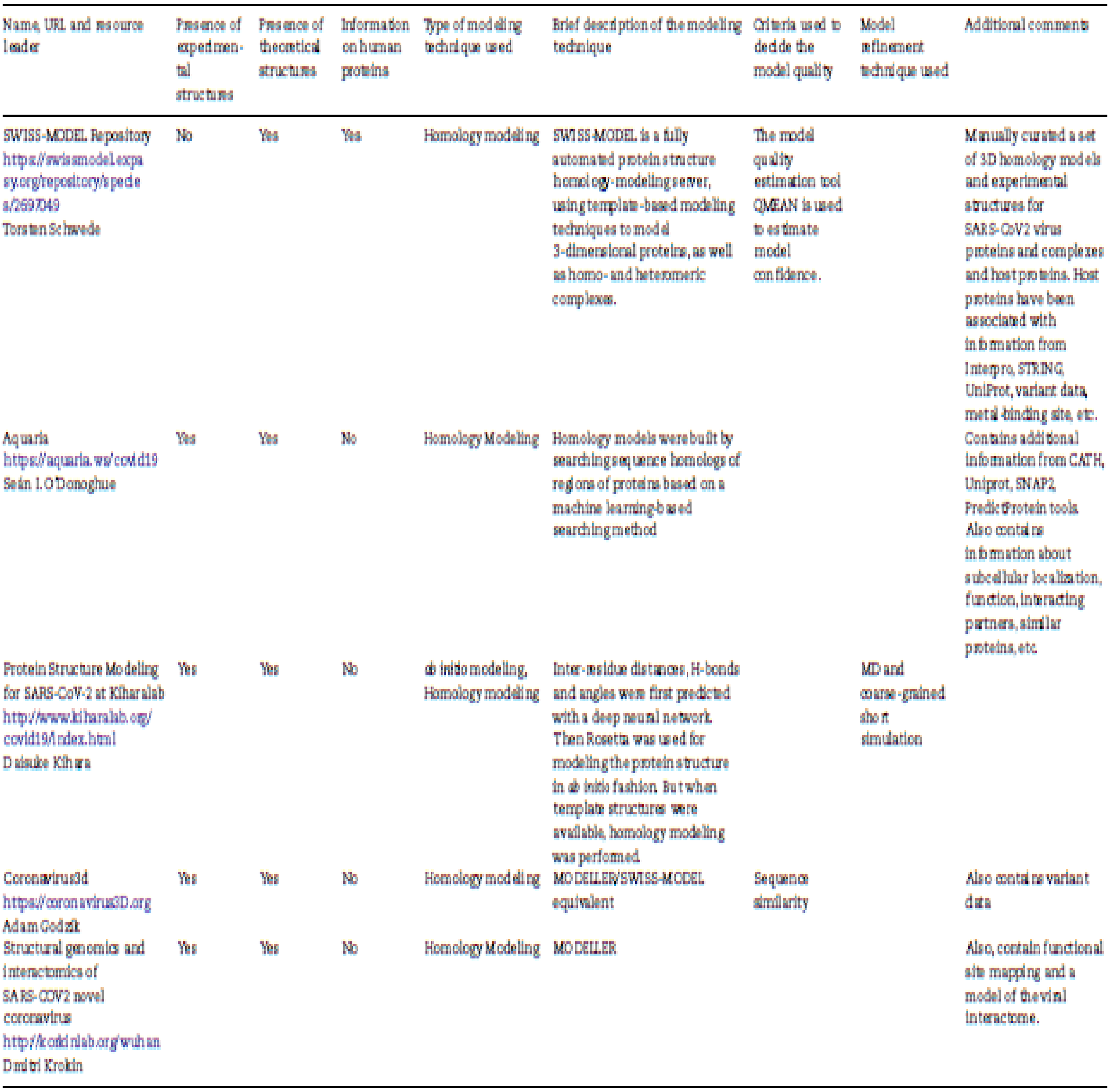
The impact of structural bioinformatics tools and resources on SARS-CoV-2 research and therapeutic strategies

- Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of coronavirus disease 19 (COVID-19), which is an ongoing pandemic, causing severe health and socioeconomic burden worldwide. As of 7 September, globally 26 763 217 COVID-19 infection cases, and 876 616 deaths are reported by the WHO

- SARS-CoV-2 (previously known as 2019-nCoV), is a single stranded positive-sense RNA virus, belonging to the genus Beta coronavirus and the family Coronaviridae
- Since the first sequenced genome of SARS-CoV-2 isolated in Wuhan City, China, over 90 000 genome sequences have been deposited in Global Initiative on Sharing All Influenza Data (GISAID, https://www.gisaid.org/). Three-dimensional (3D)structures were rapidly solved for the key target proteins in SARS-CoV-2 and host proteins, namely the spike protein,RNA-dependent RNA polymerase (RdRp), main protease (Mpro or 3CLpro), Papain-like protease (NSP3 or PLpro) and human angiotensin-converting enzyme 2 (hACE2)
- **Experimentally determined 3D structures of SARS-CoV-2 proteins**
- Experimentally determined structures of macromolecules play an essential role in the effort to discover and develop effective drug molecules against target viral organisms.
- The worldwide PDB (wwPDB) manages the global archive of macromolecular structures, the PDB , with over 165 000 protein and nucleic acid structures and over 30 000 interacting ligand molecules. As of October 2020, 21 of the 29 viral proteins of SARS-CoV-2 have over 300 experimentally determine structures.
- The overwhelming majority of these structures have canonical amino acid sequences, but a few structures have modified residues, such as PDB 7JR4 (N-Methyl Lysine) or PDB 6XB0
- (S-Hydroxycysteine), while a few structures have engineered mutations, such as PDB 6WRH. The majority of these structures focus on the Replicase polyprotein 1ab (over 250 entries covering 10 processed mature proteins to date: Host translation inhibitor, Papain-like proteinase, 3C-like proteinase, RNA-directed RNA polymerase, Helicase, Uridylate-specific endoribonuclease, 2 -Oribo semethyl transferase and non-structural proteins 7, 8 and 9), while most of the non-structural proteins lack experimentally determined structures. As of October 2020, 49% of the sequence of Replicase polyprotein 1ab is covered by experimentally determined structures, 76% of the Spike glycoprotein, 56% of the nucleoproteins, while the ORF proteins have coverages ranging between 39 and 86%. The archived structures are available to the public through theweb services of the wwPDB consortium members, na ely PDBe , RCSB PDB and PDBj . The electrostatic potential maps determined using electron microscopy are archived in the Electron Microscopy Data Bank (EMDB), with the raw EM data available from EMPIAR (**Table**), covering molecular structures from single proteins to organelles and cells
- 

Data resource	Landing page	Example SARS-CoV-2 entry
Protein Data Bank in Europe (PDBe)	https://pdbe.org	https://pdbe.org/5rgg
Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB)	https://rcsb.org	https://www.rcsb.org/structure/5rgg
Protein Data Bank Japan (PDBj)	https://pdbj.org	https://pdbj.org/mine/summary/5rgg
Electron Microscopy Data Bank (EMDB)	https://emdb-empiar.org/	https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-22126
Electron Microscopy Public Image Archive (EMPIAR)	https://empiar.org/	https://www.ebi.ac.uk/pdbe/emdb/empiar/entry/10404/

Service name	Access URL
PDBe COVID-19 Portal	https://www.ebi.ac.uk/pdbe/covid-19
RCSB PDB COVID-19 Page	https://rcsb.org/covid19
PDBj COVID-19 Page	https://pdbj.org/featured/covid-19?tab=all
BMRB COVID-19 Page	http://www.bmrb.wisc.edu/coronavirus.shtm
EMBL-EBI COVID-19 Data Portal	https://www.covid19dataportal.org
Swiss-Model COVID-19 Page	https://swissmodel.expasy.org/repository/species/2697049
Coronavirus3D	https://coronavirus3d.org
Complex Portal COVID-19 Page	https://www.ebi.ac.uk/complexportal
InterPro COVID-19 Page	https://www.ebi.ac.uk/interpro/proteome/uniprot/UP000464024
UniProt COVID-19 Entry Pages	https://covid-19.uniprot.org
3DBioNotes-WS COVID-19 Page	https://3dbionotes.cnb.csic.es/ws/covid19
PDBSum COVID-19 Page	https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/covid-19.html
Protepedia COVID-19 Page	http://proteopedia.org/wiki/index.php/Coronavirus_Disease_2019_&#x0025;28COVID-19&#x0025;29



Name, URL and resource leader	Presence of experimental structures	Presence of theoretical structures	Information on human proteins	Type of modeling technique used	Brief description of the modeling technique	Criteria used to decide the model quality	Model refinement technique used	Additional comments
SWISS-MODEL Repository https://swissmodel.expasy.org/repository/species/2697049 Torsten Schwede	No	Yes	Yes	Homology modeling	SWISS-MODEL is a fully automated protein structure homology-modeling server, using template-based modeling techniques to model 3-dimensional proteins, as well as homo- and heteromeric complexes.	The model quality estimation tool QMEAN is used to estimate model confidence.		Manually curated a set of 3D homology models and experimental structures for SARS-CoV2 virus proteins and complexes and host proteins. Host proteins have been associated with information from Interpro, STRING, UniProt, variant data, metal-binding site, etc.
Aquaria https://aquaria.ws/covid19 Seán I. O'Donoghue	Yes	Yes	No	Homology Modeling	Homology models were built by searching sequence homologs of regions of proteins based on a machine learning-based searching method			Contains additional information from CATH, Uniprot, SNAP2, PredictProtein tools. Also contains information about subcellular localization, function, interacting partners, similar proteins, etc.
Protein Structure Modeling for SARS-CoV-2 at Kiharalab http://www.kiharalab.org/covid19/index.html Daisuke Kihara	Yes	Yes	No	ab initio modeling, Homology modeling	Inter-residue distances, H-bonds and angles were first predicted with a deep neural network. Then Rosetta was used for modeling the protein structure in ab initio fashion. But when template structures were available, homology modeling was performed.		MD and coarse-grained short simulation	
Coronavirus3d https://coronavirus3D.org Adam Godzik	Yes	Yes	No	Homology modeling	MODELLER/SWISS-MODEL equivalent	Sequence similarity		Also contains variant data
Structural genomics and interactomics of SARS-COV2 novel coronavirus http://korkinlab.org/wuhan Dmitri Korkin	Yes	Yes	No	Homology Modeling	MODELLER			Also, contain functional site mapping and a model of the viral interactome.