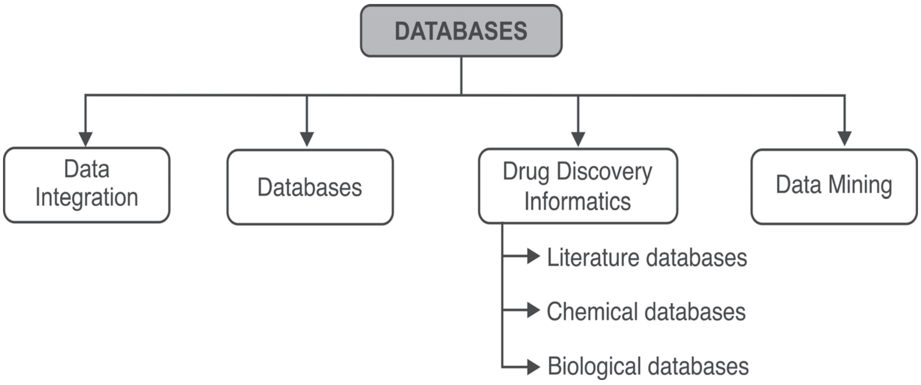
CHAPTER

1 Drug Discovery Informatics

Chapter Overview



DATA INTEGRATION

The information and knowledge pertaining to therapeutic candidate research accumulated over many years is basis for the current drug research process. The data includes chemical, generic, bio-chemical, pharmacological, physiological and related informations. Today the drug discovery process requires technology to access and manipulate large quantities of data. The field informatics (information technology) refers to the utility of computers in storage and retrieval of information. It organizes the data (information) and enables the analysis of information to produce knowledge.

- ***Information*:** The raw data, a list of fact with different meanings in different context. The term information is a component of knowledge.
- ***Knowledge*:** It includes the retrieval of information and analysis (processing). The facts, information and skills acquired through experience and or education (theoretical and practical). The knowledge can be derived from existing knowledge.

The increased recognition of informatics in drug discovery is witnessed by cheminformatics and bioinformatics.

Informatics, the utility of computer in storage and retrieval of information.

DATABASES

Databases are useful in
- Patient healthcare
- Patient surveillance
- Treatment advice
- Drug design
- Clinical procedures

In everyday life, we are in need of informations ranging from basic to scientific. Print and electronic indices (books and journals) were developed to retain the information. Databases are physical repositories for text, numerical, graphical and structural information. These indices facilitate the efficient retrieval of relevant information. The paper-based traditional database systems require large storage cabinets. Data duplication, loss or damage of data and time consuming search are the limitations of traditional databases. The computer based databases were developed to overcome these issues. These databases utilize algorithms for the extraction of data from the large pool. Computerized databases reduce the data redundancy (data inconsistencies). It offers data integrity, data sharing and data migration between systems, and security restriction.

These databases find wide range of applications for patient health care. It includes patient health care recording, surveillance of patient status and treatment advice. The role of these databases in drug design and related clinical procedures are significant. MEDLINE, DrugBank, Chemical Abstracts, Protein Data Bank (PDB) and BindingDatabase (BD) are the few note worthy databases.

Database Structure

1. **Database Schema:** The overall description of a database is called as database schema. Three types of database schemas are known.

   (a) ***Internal schema:*** It contains definitions of the stored records. It specifies about the data storage mechanism.

   (b) ***External schema***: It describes external views of the data. It restricts the user to access the un-authorized and excludes irrelevant data.

   (c) ***Conceptual schema***: It describes the types and relationships between the databases.

2. **Database Architecture:** The database architecture defines the nature and structure of the data. The database architecture specifies set of rules and processes for the data storage and the data access. It includes data types, relationships and naming conventions. It maintains the integrity, reliability, scalability and database performance. The key features of database architecture are listed below.

   (a) Numerous input forms of different citation formats (e.g., journals, books, theses).

   (b) Records download from external indexes and catalogs.

Database Management System (DBMS)

Database management involves generating, updating and deleting the data. The database system refers to the collection and management of relevant information.

The software tools that perform these functions are called as database management system (DBMS). The larger size (several giga bytes) of data poses greatest challenge for the management of these data (information). This demands the development of information management systems such as cheminformatics and bioinformatics. Physical collection of logically related records (data) and their retrieval process are facilitated through database management systems. It attempts to make the physical data non-redundant, maintains the data integrity and ensures the data independence.

- **Data redundancy:** The quality of the data is of the primary importance as it determines the prediction accuracy than the quantity. The quality control and data curation process will ensure the quality of data. Identical data stored in two or more files are known as data redundancy. The dependencies between attribute (column) causes data redundancy. It occurs to database systems which have a field that is repeated in two or more tables. It leads to data anomalies and corruption, hence should be avoided.

  Redundancy of sequence information in databases can be useful as quality control. Data redundancy is a major concern in streaming and optimizing the databases. Two sequences of the same gene with base differences (one / more) due to sequence errors cause data redundancy. Both cloning artifacts and sequencing errors creates unique sequence entries (has no biological relevance). Difference in technique (cloning, sequencing) creates redundant data.

- **Data normalization:** The non-redundancy property of databases ensures data integrity through database normalization. This permits data sharing and security restriction. Normalization is the process of simplifying the relationship between data elements in the records. Database normalization is a critical part of good database architecture, which ensures data integrity and avoids data redundancy.

Normalization, process which simplifies the data relationship and removes data redundancy.

Data redundancy, identical data in files.

DRUG DISCOVERY INFORMATICS

The informatics assists in the integration of literature, chemical, biological information. This technological advancement assists in the design of molecules for therapeutic intervention. The integrated technology (informatics) accelerates and strengthens the drug discovery and development processes.

Bioinformatics: The National Center for Biotechnology Information (NCBI) defines, bioinformatics as a field of science in which biology, computer science and information technology merge into a single discipline. It has profound application in the identification of new drug targets with the help of molecular biology and bio-physical techniques (X-ray crystallography and NMR spectroscopy). It encompasses the various methods and algorithms for analyzing and extracting biologically relevant information from the rapidly growing biological and essential sequence databases.

- ***Genome informatics*:** Genome informatics is helpful in unveiling the complexity of gene expression. It resolves the genetic variation at the genomic and cellular level.
- ***Proteome informatics*:** The principle application of proteome informatics include target discovery, target validation, natural protein therapeutics discovery, mode of action studies and toxicology.

Chem-informatics, scientific methods to store, retreive, and analyze molecular data.

Bioinformatics, molecular biology tool to identify, analyze and extract drug target data.

Cheminformatics: Cheminformatics deals with the scientific methods to store, retrieve and analyze the immense amount of molecular data. Cheminformatics has become an integral part of the drug discovery process (rational drug design).

- ***Chemical reaction informatics:*** Chemical reaction informatics enable chemist to explore synthetic pathways, design and record completely new experiments from scratch. The chemical reaction databases namely CASReact, ChemReact, CrossFire Plus, etc are very significant.
- ***Toxicoinformatics**:* It predicts the toxicity of chemical molecules in bio-phase. TOPKAT, MULTICASE and COMPACT are the most useful databases.

In future, these technological developments are expected to grow both in terms of their reliability and scope. Thus, the emerging informatics integrated with pharmaceutical sciences is becoming an essential component of drug discovery.

Components of Drug Discovery Informatics

I. ***Information Resources***

1. Literature databases
2. Chemical databases
3. Biological databases

II. ***Software Tools***

1. Predictive analytics
2. Structure (two-dimensional (2D) and three-dimensional, (3D)) search tools
3. Pharmacokinetic/Pharmacodynamic (PK/PD) data analysis
4. Assay data analytics
5. Pharmacophore modeling tool
6. Genomics data analytics
7. Proteomics data analytics

Table 1.1 Drug discovery databases, data classes and data types

Database	Data class	Data type
Literature	Articles (research and review)	Textual, graphical, structural, clinical and bibliographical details about the molecules.
	Patents	Property rights granted by a sovereign state to the inventor of novel, non-obvious and useful invention.
	Reports	The process, progress and results of scientific research.
Chemical	Physical	• Physicochemical (e.g., Log P) details. • Purification methods, experimental methods (X-ray, NMR) and related details.
	Structural	Structure, topographical and conformational analysis details.
Biological	Assay methods and biochemical process	Procedures and processes for the routine pharmacology and molecular biology investigations to assess the biological activity potential of phyto-chemicals and synthetic chemicals.
	Pharmacokinetics and pharmacodynamic	Log P, Log S, Caco-2 permeability, pKa, AUC, Cmax and related data.
	Toxicological	Carcinogenicity, teratogenicity, genotoxicity and mutagenicity data.
	Sequence	Molecular residue sequence details and chemical atom composition.
	Structural biology	• Genomics, proteomics and transcriptomics details about the target. • Sequence, gene location, gene structure and single nucleotide proteins (SNPs) and functional annotation details about the target. • Purification methods, experimental methods (X-ray, NMR) and related details.

LITERATURE DATABASES

Earlier print and electronic indices retain control over the phenomenal growth in the quantity of published information (books and journals). The electronic indices were developed for the storage, retrieval and dissemination of informations. The literature search and information retrieval previously dominated by librarians is now available through database search engines. The scientists from research laboratories and academic research institutes utilizes these engines for information retrieval. The internet has number of literature databases (database for information), which are useful for biological scientists including pharmaceutical scientists.

Pharmaceutical literature databases

MEDical Literature Analysis and Retrieval System (MEDLARS) is a bibliographic database. It is developed by National Library of Medicine (NLM), a unit of National

Centre for Biotechnology Information (NCBI), US. MEDLARS made revolution in the literature search process. It maintains an index of medical science articles. The most useful databases for the pharmacists and the pharmaceutical scientists are listed below.

The most useful literature databases are
1. MEDLINE
2. EMBASE
3. International Pharmaceutical Abstract (IPA)
4. Chemical Abstracts
5. TOXLINE
6. DART

1. **MEDLINE:** The online version of MEDLARS is named as MEDLINE (MEDLARS online). It is the prominent bio-medical database, which covers more than 25 million citations. This database gives access to medical information (clinical and therapeutic topics). It provides access to references from INDEX MEDICUS back to 1951 and earlier. PubMed is the search engine available for search against MEDLINE. The PreMEDLINE provides access to the articles which are not indexed in MEDLINE database. It maps user terms through NLMs, Unified Medical Language Systems (UMLS) metathesaurus.

   Literature Selection Technical Review Committee (LSRTC): The great majority of journals are included in the MEDLINE based on the recommendation of the Literature Selection Technical Review Committee (LSRTC) of NCBI. The subjects namely biomedicine, behavioural sciences, chemical sciences, bioengineering, clinical care, public health and health policy development are given preference in MEDLINE. It also covers biology, environmental science, marine biology, plant and animal science and biophysics.

2. **EMBASE:** Elsevier maintained European literature based database for drugs and human medicine related information. It provides access to journals dated back 1947. It covers more than 32 million records including MEDLINE titles. This database is useful in uncovering drug-disease relationship and drug-drug interactions. It is very useful in identifying the adverse-drug events.

3. **International Pharmaceutical Abstract (IPA):** The American Society of Health-System Pharmacists (ASHP) developed the database called as International Pharmaceutical Abstract (IPA). It provides access to comprehensive collection of information on drug usage. It is a primary source of drug-related health literature for pharmacists, drug and poison information centers, industry (pharmaceutical and cosmetic), health practitioners, pharmacologists, medical librarians, toxicologists and litigators. It covers pharmacy trade magazines, pharmacy journals and the abstracts of pharmacy related associations.

4. **Chemical Abstracts:** Chemical Abstract Service (CAS) is a division of the American Chemical Society (ACS). It covers scientific journals, patents, technical reports, books, conference proceedings and dissertations. This database is very useful in drug discovery and development process.

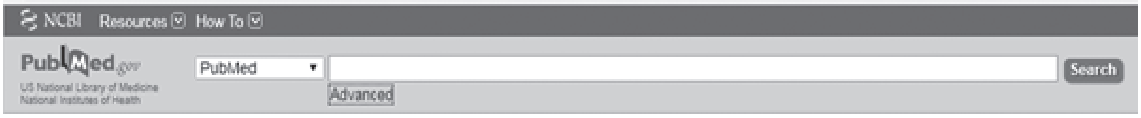
5. **TOXLINE:** The toxicology subset of MEDLINE/PubMed. It provides

bibliographic information covering the biochemical, pharmacological, physiological and toxicological effects of drugs and other chemicals.

6. **Development and Reproductive Toxicology Database (DART):** It contains references to reproductive and developmental toxicology literature.

PubMed

It is a primary electronic searching tool for the retrieval of bio-medical literature from PreMEDLINE and MEDLINE. PubMed provides a broad up-to-date and efficient search interface.


NCBI
Resources
How To
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Search
Advanced

Important PubMed Features

1. **My NCBI:** A personal workspace on NCBI databases including PubMed (registration is required, available free). The search query, search results and citations can be stored using save search link. The updates for the previously searched items will be droped to the registered e-mail. This feature helps in the storage of articles in the ready reference to later stage.

2. **Medical Subject Headings (MeSH):** MeSH is a controlled vocabulary thesaurus used to index PubMed database articles. The MeSH thesaurus is useful in indexing articles of the biomedical journals for the MEDLINE database. MeSH provides details about the information discussed in the article. The National Library of Medicine (NLM) cataloging database also utilizes the MeSH terms for the efficient search. The NLM experts assign appropriate terminologies to define these topics and articles. These terminologies can be used in the more accurate and comprehensive extraction of information. The MeSH thesaurus avoids the use of synonyms and acronyms.

Medical Subject Heading (MeSH), a controlled vocabulary thesaurus used to index PubMed database articles.

   (a) ***Subject heading (descriptor)*:** The MeSH terms are referred as subject headings or descriptors of the search.

   (b) ***Entry terms*:** These terms are synonyms of MeSH subject heading. It helps in finding the most appropriate MeSH term, e.g., Telehealth and eHealth.

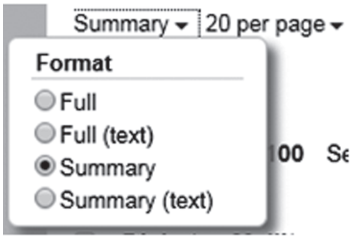
(c) ***Subheading (qualifier)*****:** This assists in describing the search results to particular aspect of subject heading. The search term diabetes is subject heading (MeSH term) and diagnosis is subheading.

(d) ***MeSH major topic*****:** Asterisks on MeSH heading and subheading designates that they are the major topics of the article.

The PubMed search for the key term Diabetes mellitus and subheading genetics include "Diabetes mellitus/genetics" (MeSH) for the information retrieval.

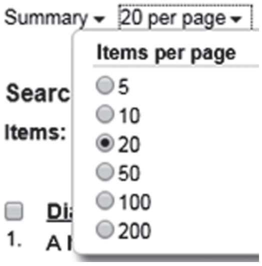
3. **Limits:** The search can be performed based on the MeSH terms, MeSH unique ID, record type, registry number, scope note, substance name and text word.

4. **Advanced Search:** It allows searching by field tags available in the builder section. The filed tags include affiliation, author name (first, full, last), book, date, journal, MeSH terms and many other related items. The Boolean operators namely AND, OR and NOT also can be used for the effective search. The more filters can be introduced based on the requirement.


Summary ▾ 20 per page ▾

Format

○ Full
○ Full (text)
● Summary
○ Summary (text)


Summary ▾ 20 per page ▾

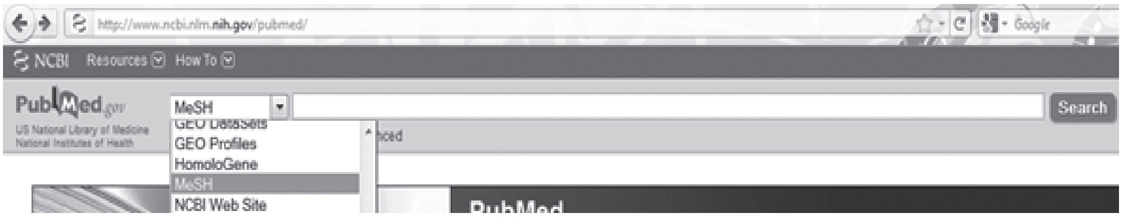
Items per page

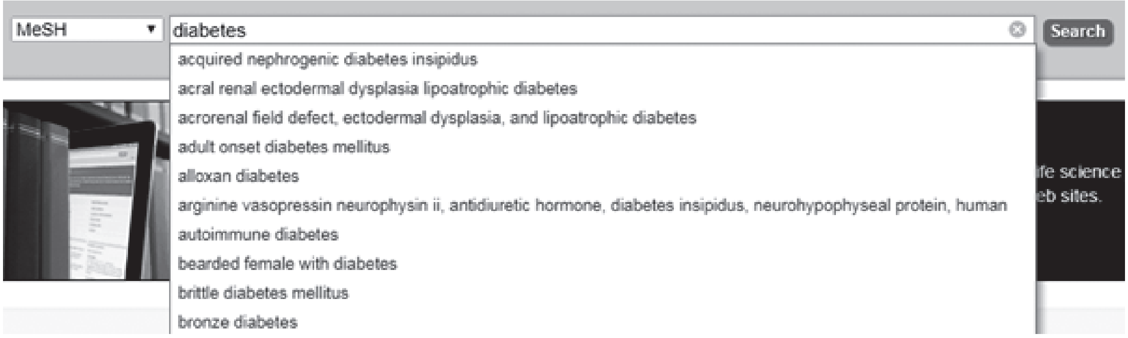
○ 5
○ 10
● 20
○ 50
○ 100
○ 200

5. **Automatic term mapping**: The PubMed categorizes the search key term into 'qualified' and 'unqualified'. The qualified terms restricts the PubMed search to specific fileds. The PubMed search for unqualified terms occurs through Automatic term mapping (ATM). This pre-processing step helps in identifying the correct MeSH term for the search key term. The ATM feature considers the search key term 'parkinson' as a disease (not as a person's name).

Information Search in MEDLINE through PubMed

1. **Information search for key words**: The key term (key words) based search can be performed in PubMed. The entry of key word for the search intended in the search box enables the search. It offers search options (simple and advanced) to enable the effective search retrieval process.
2. **Information search through MeSH terms:** Medical Subject Heading (MeSH) terms enhance the information retrieval efficiency. The MeSH browser verifies the correctness of the key term and suggests the closely associated key terms.

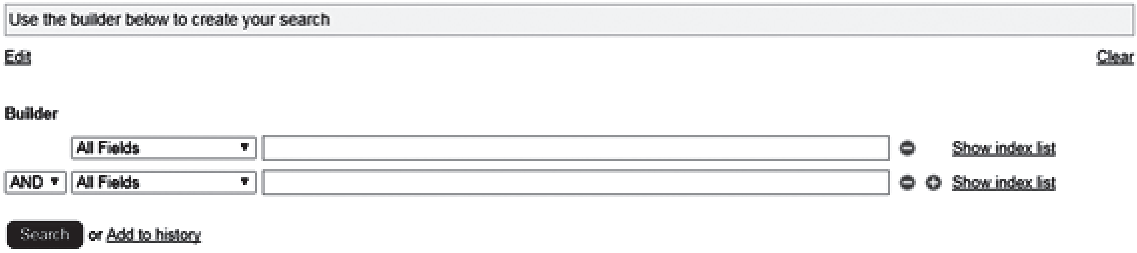



MeSH

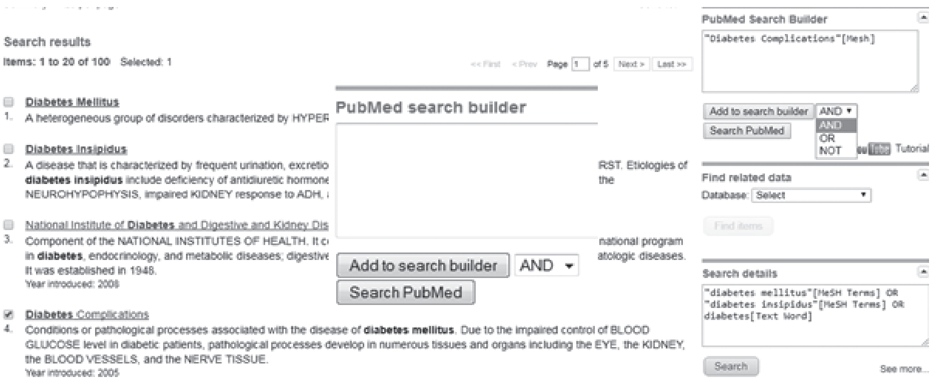
diabetes

Search
acquired nephrogenic diabetes insipidus
acral renal ectodermal dysplasia lipoatrophic diabetes
acrorenal field defect, ectodermal dysplasia, and lipoatrophic diabetes
adult onset diabetes mellitus
alloxan diabetes
arginine vasopressin neurophysin ii, antidiuretic hormone, diabetes insipidus, neurohypophyseal protein, human
autoimmune diabetes
bearded female with diabetes
brittle diabetes mellitus
bronze diabetes
ife science
eb sites.

3. The correct key term from the MeSH list can be selected for the search. The Add to search builder option (located in the left side of the window) can be used to combine more than one MeSH terms. The Boolean operators AND, OR and NOT also can be used for the efficient search.
   - (a) ***AND***: It can be used to retrieve articles that are indexed under both topics.
   - (b) ***OR***: It can be used to retrieve articles that are indexed using either one or other topic.
   - (c) ***NOT***: It can be used to remove any articles that are indexed using that terminology.


Use the builder below to create your search

4. **Search:** The Search PubMed button can be clicked to execute search against MEDLINE.
5. **PubMed results**
   - Number of article in view can be changed by using Entrez date limit option).
   - Number of research and review articles can be viewed through filtered search.



Search results

Items: 1 to 20 of 100 Selected: 1

□ Diabetes Mellitus
1. A heterogeneous group of disorders characterized by HYPER

□ Diabetes Insipidus
2. A disease that is characterized by frequent urination, excretio ... RST. Etiologies of **diabetes insipidus** include deficiency of antidiuretic hormone ... the NEUROHYPOPHYSIS, impaired KIDNEY response to ADH, ...

□ National Institute of **Diabetes** and Digestive and Kidney Dis
3. Component of the NATIONAL INSTITUTES OF HEALTH. It c ... national program in **diabetes**, endocrinology, and metabolic diseases; digestive ... atologic diseases. It was established in 1948.
Year introduced: 2008

☑ Diabetes Complications
4. Conditions or pathological processes associated with the disease of **diabetes mellitus**. Due to the impaired control of BLOOD GLUCOSE level in diabetic patients, pathological processes develop in numerous tissues and organs including the EYE, the KIDNEY, the BLOOD VESSELS, and the NERVE TISSUE.
Year introduced: 2005

PubMed Search Builder

"Diabetes Complications"[Mesh]

Find related data

Database: Select

Search details

"diabetes mellitus"[MeSH Terms] OR "diabetes insipidus"[MeSH Terms] OR diabetes[Text Word]

PubMed mobile: The feature PubMed mobile provides better access for the smart phone users.

CHEMICAL DATABASES

Chemical databases contain encoded chemical structures along with their molecular and atomic data. Most chemical databases store two dimensional (2D) and three dimensional (3D) structural models in different domain. The chemical databases are of two types.

1. *Analytical (experimental) database*: It contains libraries of reference materials with defined techniques.
2. *Computed models*: It contains structures derived from the quantum mechanic and molecular mechanic methods.

Chem-informatics (*in silico*) techniques have wide range of applications in rational drug design process.

The vast number of molecules in the chemical databases and the associated data requires sophisticated information systems. Modern drug discovery procedures require systems that have the ability to access and manipulate large quantities of chemical data. The computer technologies increase the storage and retrieval process of chemical information. The use of computer-based chemical databases in drug design is known as cheminformatics. It encompasses the design, creation, organization, storage, management, retrieval, analysis, dissemination, visualization and use of chemical information.

Cheminformatics uses computer and information techniques to solve the problems in the field of chemistry. These *in silco* techniques have wide range of applications in rational drug design process.

In general, the data present in cheminformatics database fall into four categories:

1. ***Structural***: The molecular structure data are the most unique aspect of chemical databases. Molecular structure refers to the one-dimensional (1D), two-dimensional (2D) and three-dimensional (3D) representations of molecules. It differentiates cheminformatics from other database applications.
2. ***Numerical***: Numerical data includes biological activity (IC_{50}, LD_{50}), p*K*a, log P and analytical results (NMR and IR).
3. ***Annotation / text***: It includes informations such as experimental notes associated with the structure. The protocols are categorized by the type of task: for example, substructure / similarity searches, database access and property calculators.
4. ***Graphical***: It includes spectra or plots of any structure or data point (graphical representations).

Chemical Nomenclature System for Database Search

The chemical nomenclature system provides consistent, unambiguous and reproducible descriptions about the chemicals. It provides descriptions about the atom-to-atom connections, open chain, fuzed ring systems and branching through chemical notation.

Chemical notation: A specialized and abbreviated system of signs and symbols used in chemistry.

Signs: Alpha numericals.

Symbols: Δ, →.

The computerized nomenclature system includes:

(a) ***Chemical Abstracts Service***: It provides nomenclature to the structures of organic and biological chemistry. Every chemical structure is assigned with unique chemical abstract service (CAS) registry number.

(b) ***DARC***: This notation represents the molecular structures using fragments reduced to an envionment which is limited (FREL) concept. The valency of the atoms (hybridized state) are represented through open connections, residual valency index (RVI) and free site index (FSI).

(c) ***Dysons system:*** A molecular representation using lines and symbols. Different symbols are assigned to indicate the different chemicals. The methyl alcohol is indicated through C.Q and phenol as B6.Q.

(d) ***Wisweser Line Notation (WLN):*** This notation represents the compounds by structural formula (short combination of numeral and capital letters) and punctuation marks. The functional groups are indicated by capital letters (e.g., R indicates OH group). The numerals represent the number of carbon atoms (e.g., 2 indicates the ethyl group).

Chemical nomenclature system provides consistent, unambiguous and reproducible descriptions for chemicals.

(e) ***SYBYL line notation:*** It is an ASCII language used to represent organic chemical structures. It describes the chemical structures, specifies the elemental atoms, bonds, branches, ring closures, attributes and macro atoms.

(f) ***Simplified Molecular Input Line Entry System (SMILES)*****:** SMILES is a computer language for pharmaceutical and other chemical processes research. It is useful in

- Generation of unique notation
- Structural depiction
- Optimal data retrieval
- Substructure recognition

(g) ***Smiles Arbitary Target Specifications (SMART):*** This notation represents the molecular structures as atomic and bond symbols. It provides flexible and efficient structure and substructure search.

(h) ***International Chemical Identifier (InChI):*** The International Union of Pure and Applied Chemistry (IUPAC) has recommended this molecular representation. This notation provides connectivity, tautometic, isotopic, stereo chemical and electronic informations.

(i) ***Boolean:*** The structural key provides information about the frequencies of particular chemical feature. It describes the molecular chemical composition. It uses structural and molecular fingerprints for the database search.

Important

1. **Chemical Abstracts Service (CAS):** The American Chemical Society (ACS) has developed world's largest chemical database for chemical structures and reactions. It contains more than 132 million of compounds including unique organic and inorganic chemicals (alkyl minerals, polymers) along with their property data. Chemical abstracts (database) provide nomenclature to the structures of organic and biological chemistry.

   CAS registry number**:** Every chemical structure in the database is assigned to unique chemical abstract service (CAS) registry number called as Numeric identifier (CAS ID). The CAS ID for paracetamol is 103-90-2.

   SCIFINDER: The best database for chemistry-related topics. This research discovery application provides unlimited access to the world's most comprehensive and authoritative source of references and reactions in chemistry and related sciences. It provides access to CAS databases and MEDLINE.

2. **ChemBank:** It is a public, web-based informatics originally developed by Harvard Institute of Chemistry and Cell Biology. At present the Chemical Biology Program and the Chemical Biology Platform at Broad Institute of

MIT and Harvard maintains this database. This database utilizes Daylight Chemical Information Systems and is freely available. It contains molecular properties of small molecules and provides information about biochemical assays. This database permits text based and structure based search. The search can be restricted to selective subsets (e.g., natural products; FDA approved drugs).

3. **Chemical Entities of Biological Interest (ChEBI)**: ChEBI incorporates an ontological classification, and explains the relationships between molecular entities and their parent and or children. The molecular entities include atom, molecule, ion, ion pair, radical, radical ion, complex and conformer.

4. **Developmental Therapeutics Program (DTP)**: Developmental Therapeutics Program (DTP) of National Cancer Institute (NCI) provides services and resources to the academic and private-sector research communities worldwide to facilitate the discovery and development of new cancer therapeutic agents. The anti-cancer drugs namely, paclitaxel, romidepsin, eribulin, sipuleucel-T and dinutuximab were developed with the support from DTP.

5. **Comparative Toxicogenomics Database (CTD):** The National Institutes of Environmental Health Sciences (NIEHS) at North Carolina State University (NCSU) developed this database. It provides insight into complex chemical-gene and protein interaction networks. It describes the molecular mechanisms underlying variable susceptibility and environmentally influenced diseases.

6. **DrugBank:** It contains chemical, pharmaceutical, medical and molecular biological information about targets and drugs. It provides more than 200 data field for each drug and include chemical structures. It contains FDA approved small molecule drugs and peptide drugs along with nutraceuticals and experimental drugs. DrugBank can be searched through text based and structure based methods. DrugBank currently contains 11,924 drug entries including 2,538 approved small molecule drugs. DrugBank information are useful in target identification, along with biological activity screening (virtual, *in silico*) and drug metabolism predictions.

7. **Carcinogenic Potency Database (CPDB):** This database was developed at the University of California and Lawrence Berkeley Laboratory. It provides results of chronic and long-term animal cancer tests published in literature.

8. **Non-Redundant Database of Small Molecules (NRDBSM):** It is useful for virtual high-throughput screening of small molecules. It has special consideration to physicochemical properties and Lipinski's rule of five (RO5).

9. **Cambridge Structural Database (CSD):** A highly curated and comprehensive chemical database. It contains experimentally determined molecular structures of small organic molecules and coordination compounds from published literature. It also contains data published directly through the CSD (not available anywhere).

10. **CrossFire Beilstein database:** The Beilstein Institute developed this oldest database based on the Beilstein hand book of organic chemistry. This database covers compounds earlier to 1960. It contains data back to 1771 and provides structures, Beilstein and CAS registry numbers, names, formula, natural product isolations, chemical derivatives and citation data (author, journal and patent). It includes Gmelin database and Elsevier's patent chemistry database. It also provides spectral, thermodynamic, biological and toxic property data along with their uses.

11. **DRUGDEX System:** The DRUGDEX system provides independently reviewed data (from major drug centers and pharmacology services) and also include informations gathered by MICROMEDEX editorial board. It can be retrieved by generic or brand name and indications. It contains details about drug dosage, pharmacokinetics, interactions, comparative drug efficacy, clinical applications and adverse effects. It delivers unbiased drug information to physicians, pharmacists, and other health care professionals.

12. **Therapeutic Target Database (TTD):** It includes the information on the primary targets and mode of action of the drugs. Signal transduction, perturbation in metabolic reactions and metabolic pathway links are provided in this database. The affinity of the drug molecule to the macromolecule can be studied from this database.

13. **Kyoto Encyclopedia of Genes and Genomes (KEGG):** It contains genomic, chemical and network / pathway information for most organisms and chemical compounds. This provides organism specific metabolic pathways, gene signalling, protein interactions and also the structures of metabolites.

14. **MDL-Drug Data Report (MDDR)**: It is developed by Accelrys, Inc., and includes the compound details from the patent literature, journals, meetings and conferences. It contains biologically relevant compounds with demonstrated biological activity. It permits the structure search and also provides other fields.

15. **MDDR**: MDDR is developed jointly by BIOVIA and Clarivate Analytics. It covers the patent literature, journals, meetings and congresses. It provides chemical structure, calculated properties, literature references, patent informations and 3D models. It also predicts the biological activity based on the Porous classification system.

Table 1.2 Major chemical databases and their web addresses

Database	Web address
ChemBank	http://chembank.broad.harvard.edu/
Chemical Entites of Biological Interest (ChEBI)	https://www.ebi.ac.uk./chebi
Comparative Toxicogenomics Database (CTD)	htttp://ctdbase.org/
Cambridge Structural Database (CSD)	https;//www.ccdc.cam.ac.uk
Developmental Therapeutic Program (DTP)	http://dtp.cancer.gov/databases_tools/data_search.htm
Directory of useful Decoys (DUD-E)	http://dude/docking.org/
DrugBank	http://drugbank.ca/
E molecules	http://www.emolecules.com/
KEGG Drug database	http://www.genome.jp/kegg/drug/
Ligand Expo	https://ligand-expo.rscb.org/
Non-Redundant Database of Small Molecules (NRDBSM)	http://www.scfbio-iitd.res.in/ software/nrdbsm/ index.jsp
Protein Data Bank (PDB)	https://www.rscb.org
PubChem	http://pubchem.ncbi.nih.gov
Therapeutic Target Database (TTD)	https://db.idrblab.org/ttd/
Zinc	zinc.docking.org

16. **Protein Data Bank (PDB)**: Protein Data Bank (PDB) is developed by Brookhaven National Laboratories and is managed by the Research Collaboratory for Structural Bioinformatics (RCSB). It maintains the repositories of experimentally determined 3D structures of macro molecules and ligands.

17. **PubChem**: It is an open chemistry database developed and maintained by National Institutes of Health (NIH). It provides structures, identifiers, properties (chemical and physical), biological activities, patents, health, safety, toxicity data for the small molecules. It includes nucleotides (including siRNAs and miRNAs), carbohydrates, lipids, peptides and chemically modified macromolecules. It uses a 881 bits long finger point to rank database molecules against a query compound.

18. **TOXICOLOGY NETWORK (TOXNET):** The Toxicology and Environmental Health Information Program (TEHIP), a division of the National library of Medicine (NLM) developed this database. It refers to a group of chemical databases. It covers environmental health, poisoning, risk assessment and regulations and toxicology.

19. **ZINC:** It stands for *Zinc is not Commercial (ZINC)*. It is a database of small molecules for docking. The database molecule complies with Lipinski rule of five (RO5). ZINC has docking, substructure searching and compounds purchasing features. The ZINC molecular structures are annotated for the drug-like properties. The molecules can be searched based

on their molecular weight, log P, number of rotatable bonds, hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), chiral centres, desolvation energy (polar and apolar), net charge, rigid fragments and molecular function data.

Table 1.3 Chemical Database Search Engines

Search engine	Web address
PubChem	https://pubchem.ncbi.nlm.nih.gov/
Chemspider	http://www.chemspider.com
ChemID plus	https://chem.nlm.nih.gov/chemidplus/chemidlite.jsp
ChEMBL	https://www.ebi.ac.uk/chembl

Applications

Cheminformatics has become an integral part of the drug-discovery process, from lead identification. Cheminformatics approaches include database creation, physicochemical property calculations and quantitative structure activity relationships (QSAR) analysis.

- ***Library design*:** Design of small molecules with expected therapeutic potential.
- ***Similarity searching and clustering*:** It is useful in finding the closely related molecules. The substructure search identifies chemical structures with equal and larger in size (sub similarity and super similarity). This process is also known as “leapfrogging” or “scaffold hopping”. Superposition of the molecules helps in identifying the chemical similarity among the molecules.
- ***Structure searching*:** 2D structure searching and 3D structure searching (rigid and flexible 3D searching) are useful in exploring molecules of interest from databases. The substructure filters are useful in eliminating selective functional group expected to produce toxic functions.
- ***Pharmacophore modeling*:** The conserved structural unit of the molecule are known as pharmacophore. Pharmacophore modeling has been used in library design. A pharmacophore model is useful in identifying ligands with the desired biological effect.

BIOLOGICAL DATABASES

Biological databases contain well organized and persistent biological data. It includes the sequence (nucleotide and amino acid), macromolecular structure and biochemical characteristics of organisms. The increasing pool of biological data demanded the development of biological databases. GenBank and Protein Databank

(PDB) are important biological databases. The most important biological databases are conveniently grouped into:

1. Sequence databases
2. Structure databases

SEQUENCE DATABASES

The sequence of nucleotides and amino acids represents a particular gene or protein. The advancements in the molecular biology and related fields have facilitated the sequencing of genes and proteins. The most significant databases are listed in the tables given below.

Biological databases are broadly grouped into sequence and structure databases.

Table 1.4 Nucleotide sequence databases

Database	Web address
GenBank	http://www.ncbi.nlm.nih.gov/Genbank/
EMBL	http://www.ebi.ac.uk/embl
DDBJ	http://www.ddbj.nig.ac.jp
Ensembl	http://www.ensembl.org
UniGene	http://www.ncbi.nlm.nih.gov/UniGene/
Human genome	http://www.ncbi.nlm.nih.gov/genome/guide/
Nucleic acid search	http://www.oup.co.uk/har/database/c/
Mol Biol Net	http://www.molbiol.net/

Table 1.5 Protein sequence databases

Database	Web address
SwissProt	http://www.expasy.ch/sprot/
PIR	http://pir.georgetown.edu
PFAM	http://www.sanger.ac.uk/Software/Pfam http://www.cgr.ki.se/Pfam
InterPro	http://www.ebi.ac.uk/interpro/
Prosite	http://expasy.ch/prosite/

Sequence Letter Codes

The sequence analysis programs require sequence files in a particular format (e.g., FASTA format). These formats uses the standard single letter codes for describing the sequences.

- The nomenclature committee of the International Union of Biochemistry (IUB) has established a standard single-letter code to represent the nucleic acid bases.

Table 1.6 Nucleotide single letter code

Nucleotides	Single letter code
Adenine	A
Guanine	G
Cytosine	C
Thymine	T
Uracil	U

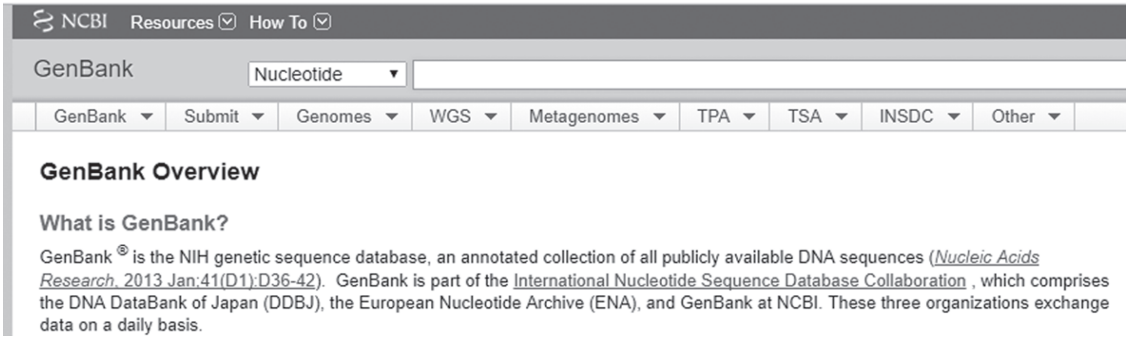
- The Joint International Committee has established single-letter amino acid codes.

Table 1.7 Amino acid single letter codes

Amino acid	Single letter code		Amino acid	Single letter code
Glycine	G		Tryptophan	W
Alanine	A		Cysteine	C
Valine	V		Lysine	K
Leucine	L		Arginine	R
Isoleucine	I		Histidine	H
Methionine	M		Aspartic acid	D
Serine	S		Glutamic acid	E
Threonine	T		Asparagine	N
Phenyl alanine	F		Glutamine	Q
Tyrosine	Y		Proline	P

GenBank

GenBank is a National Institutes of Health (NIH) genetic sequence database. It is a comprehensive public database of nucleotide and protein sequences. GenBank is the fastest growing repositories of known genetic sequences. GenBank files provides accession numbers, gene names, phylogenetic classification and references. GenBank contains more than 27 billion nucleotide bases from more than 20 million different sequences.



GenBank Overview

What is GenBank?

GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research. 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration , which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

FASTA sequence format: The FASTA sequence format begins with > (greater than) symbol followed by single line description about the organism. This text-based format represents nucleotide and protein sequences using standard single-letter codes. This format accepts lower-case single-letter codes and maps them into upper-case. A single hyphen (-) in the sequence format represents a gap in the sequence.

GenBank sequence format: NCBI developed this standardised informational format for sequences. This human readable format contains both bibliographic and feature informations about a sequence. GenBank records are delimited by a pair of forward slashes // on a single line. It starts with the word LOCUS, which follows annotation lines. The second line defines the organism type and it follows the ACCESSION NUMBER. The word ORIGIN denotes the start of sequence. The sequence ends with the double forward slashes (//).

Abstract Syntax Notation one (ASN.1) sequence format: It is an International Standards Organization (ISO) data representation format. ASN.1 is a formal computer data description language, and the NCBI uses this code for the storage and retrieval of data. The single letter code sequence with the character * (asterisk) indicates the end of the sequence, has been adopted to encode sequence data. It is useful in retrieving the taxonomic informations, molecular structures and bibliographic informations. It also include literature references, sequence functions, mRNA location, coding region and mutations.

Genetics Computer Group (GCG) sequence format: It begins with annotation lines. Each line of the sequence format starts with two-letter code (similar to EMBL and SwissProt format). This format include informations about the sequence in the GenBank entry. The information lines are terminated by two period, which mark the end of information and start of the sequence on the next line.

European Molecular Biology Laboratory (EMBL)

The European Molecular Biology Laboratory (EMBL) is a comprehensive database for DNA and RNA sequences. It is collected from the scientific literature and patent applications. The data collaboration with GenBank and DNA Database of Japan (DDBJ) is the important feature.

EMBL Sequence format: The European Bioinformatics Institute (EBI) has developed this sequence format. This line based sequence format consists of a two character code (identifier) line followed by three spaces (blank lines). It follows the annotation lines. The sequence header contains the two letter code SQ (sequence header) and it follows a set of sequence lines. The sequence record ends with two forward slashes (//).

DNA Data Bank of Japan (DDBJ)

National Institute of Genetics, Japan, established DNA Data Bank of Japan (DDBJ). It exchanges the molecular data with EMBL of EBI and GenBank at NCBI. DDBJ accepts data primarily from Japanese researchers (also from other countries). It assigns accession number for each entry.

International Nucleotide Sequence Database Collaboration (INSDC)

The collaboration between the DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory (EMBL) and GenBank at NCBI has developed International Nucleotide Sequence Database Collaboration (INSDC).

SwissProt

The department of Medical Biochemistry, a division of the University of Geneva (Swiss Institute of Bioinformatics (SIB)) and the European Molecular Biology Laboratory (EMBL) together established this protein sequence database. SwissProt is the curated protein sequence database. It provides a high level of annotations such as functions, domain structure and post-transitional modifications (PTMs). The distinct features of SwissProt are listed below.

1. ***Core data***: It provides the sequence data, citation information and taxonomic data.
2. ***Annotation***: It includes protein functions, domain, post-transitional modifications (PTMs), structure (secondary and quarternary), sequence similarities and diseases.
3. ***Minimum redundancy***: The SwissProt database merges protein sequence data from different literature reports to provide non-redundant database.
4. ***Integretity***: It offers high levels of integration with other databases.

Protein ID: The protein ID code format in SwissProt is X_Y (mnemonic).

(a) ***X*:** It indicates the protein name with maximum four characters (X). The code PGH2 stands for prostaglandin G/H synthetase 2.

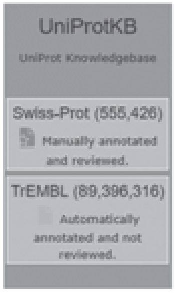
(b) ***Y*:** The genus and species are indicated by maximum five character code (Y). The first three letters of the genus and first two letters of the species. The code CAVPO stands for ***Ca**via **po**rcellus* (guinea pig). Alternatively the common name is assigned for the Y code. The name HUMAN is used in case of *Homo sapiens* and BOVINE for *Bos taurus*.

SwisProt Sequence format: It provides high level of annotation, including the physical and biochemical properties of the protein. Each line of the sequence format begins with a two character line code. The two character code describes the type of data. The code ID refers to identification, SQ for sequence header, OS for organism species and AC for accession number. It also provides post-translational modification (PTM) and domain structure.

UniProt knowledgebase (UnitProtKB) is the collection of functional information on proteins. It includes amino acid sequence, protein description, taxonomic data, citation and other experimental and computational data. It consist of two sections:

(a) **UniProtKB/SwissProt:** A high quality, manually annotated and non-redundant protein sequence database.

(b) **UniProt/TrEMBL:** TrEMBL contains computationally generated annotation, which are not integrated in SwissProt. It contains all the translations of EMBL nucleotide sequence (excluding coding sequences) in SwissProt format. The TrEMBL entries are progressively merged with SwissProt entries.

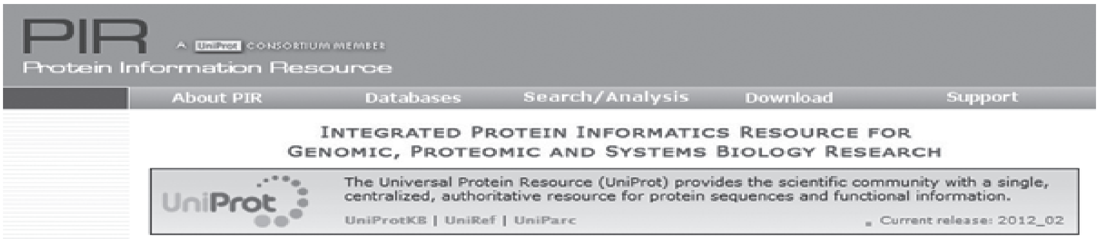


Swiss-Prot (555,426)
Manually annotated and reviewed.

TrEMBL (89,396,316)
Automatically annotated and not reviewed.

Protein Information Resource (PIR)

The National Biochemical Research Foundation (NBRF) established the Protein Information Resource (PIR). It assists in the molecular evolution, functional genomics and computational biology analysis. PIR in collaboration with Munich Information Center for Protein Sequences (MIPS) and Japan Information Database developed PIR International Protein Sequence Database (PIR-PSD). The primary data sources for PIR-PSD are GeneBank of NCBI, EMBL of EBI, DDBJ translations, published reports and direct submissions to PIR.


PIR
A UniProt CONSORTIUM MEMBER
Protein Information Resource
About PIR
Databases
Search/Analysis
Download
Support
UniProt
UniProtKB | UniRef | UniParc

INTEGRATED PROTEIN INFORMATICS RESOURCE FOR GENOMIC, PROTEOMIC AND SYSTEMS BIOLOGY RESEARCH

The Universal Protein Resource (UniProt) provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information.

Current release: 2012_02

PIR has four distinct sections.

PIR-1	It contains fully classified and annotated entries.
PIR-2	It includes preliminary entries.
PIR-3	It includes unverified entries which have not been reviewed.
PIR-4	• Conceptual translations of artifactual sequences. • Conceptual translations of sequences that are not transcribed or translated. • Protein sequences or conceptual translations that are extensively and genetically engineered. • Sequences that are not genetically encoded and not produced on ribosome.

Protein Information Resource (PIR) / CODATA sequence format: The National Biomedical Research Foundation (NBRF) and Protein information Resource (PIR) programs jointly developed this sequence format. It resembles FASTA format by the symbol > (greater than) at the beginning. It follows a two-letter code to describe the sequence type (e.g., P1; F1). The presence of character P refers the complete sequence and the character and F indicates the fragments. The numeral one (1) indicates linear sequence and two (2) refers to circular sequence. A semicolon and four to six character unique name for the entry is included after the alphanumerical characters. The second line contains the sequence name and species of origin (connected by hyphen).

PFAM

Pfam is EMBL based database of protein families that includes their annotations and multiple sequence alignments generated using Hidden Markov Model (HMM). It provides a complete and accurate classification of protein families and domains with wide coverage of proteins. This database provides sequence alignment for protein domains and conserved protein region for better homology detection. It also generates profile-HMM. It has two divisions, namely

(a) **PFAM-A**: It contains curated families with associated profile HMMs.

(b) **PFAM-B**: It includes the cluster of sequence segments, which are not included in PFAM-A.

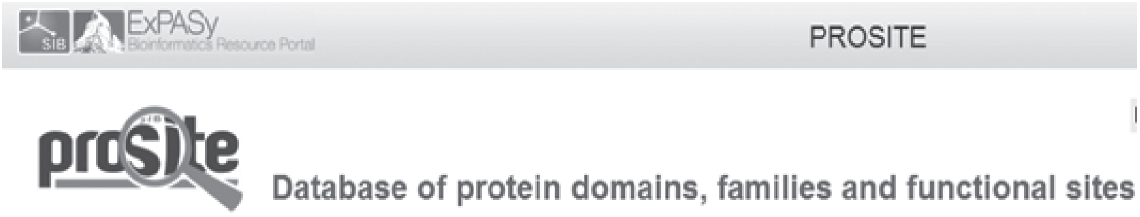
INTERPRO

InterPro is a EMBL based database for functional analysis of protein sequences. It classifies protein into superfamily, family and subfamily levels. It offers prediction analysis for the presence of functional domain and repeats.

InterPro protein sequence analysis & classification

PROSITE

A database of biologically significant sites, patterns and profiles. It determines the function of proteins transcribed from genomic or cDNA sequence. It includes concise description of the protein family / domain along with summary of the development of pattern or profile.



Database of protein domains, families and functional sites

iProClass

It provides value-added information reports for UniProtKB and unique NCBI Entrez protein sequences in UniParc. It provides links to more than 175 biological databases. It include databases for protein families, functions and pathways, interactions, structures and structural classifications, genes and genomes, ontologies, literature and taxonomy. It supports in protein sequence annotation, genomic and proteomic research.

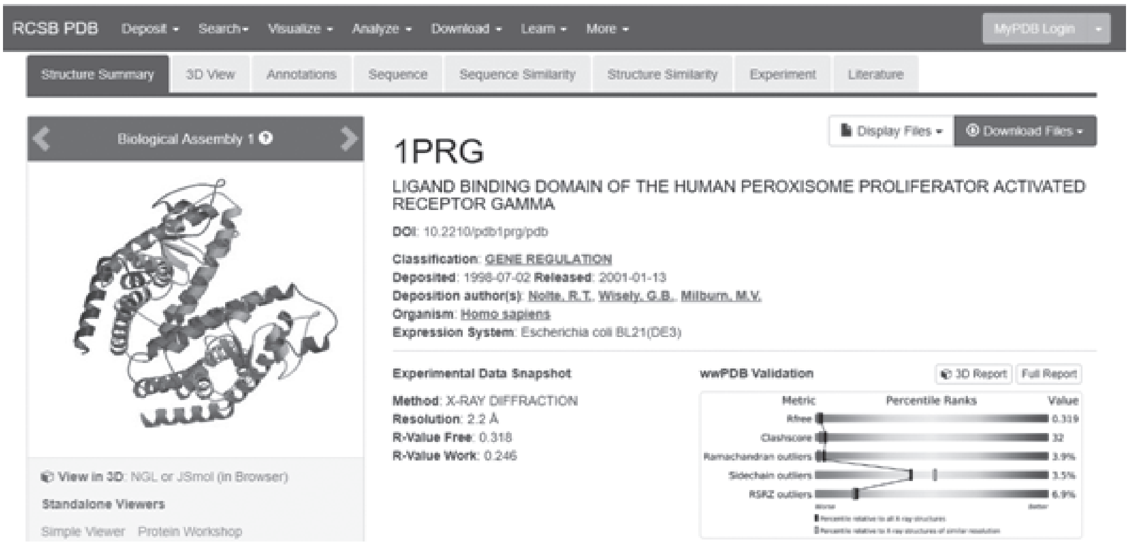
STRUCTURE DATABASES

The two-dimensional (2D) structures are not sufficient to explain the macromolecular properties. The three-dimensional (3D) molecular framework are essential for explaining their functions (in books and articles also). The deeper insight to the structure can be obtained from the three-dimensional (3D) structures. The molecular biology and biophysical techniques (X-ray crystallography and NMR spectroscopy) has generated three-dimensional (3D) structure of macromolecules. These 3D structures are deposited in the structural repositories called as structure databases. The most important structure databases are listed below and described in the following section.

1. PDB	http://www.rcsb.org
2. CATH	http://www.biochem.ucl.ac.uk/bsm/cath
3. SCOP	http://scop.mrc.imb.cam.ac.uk/scop
4. MMDB	http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml
5. DNA databank	http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml

PROTEIN DATA BANK (PDB)

Protein Data Bank (PDB) is developed by Brookhaven National Laboratories and is managed by the Research Collaboratory for Structural Bioinformatics (RCSB). It maintains the repositories of experimentally determined 3D structures of macro molecules. It provides access to the sequence details, atomic coordinates, crystallization conditions, 3D structure neighbors, compute methods, geometric data, structural factor, 3D images, etc. Each protein is assigned with a unique four (4) character alphanumeric identifier code (PDB-ID / PDB code). The alphanumeric code for the human peroxisome-proliferator activated receptor gamma is 1PRG (2PRG also available). The protein 3D structure of interest can be searched against PDB by providing key words (text based), PDB-ID, author name and by querying for the deposition data. Advanced search options provide more efficient and selective search.



1PRG

LIGAND BINDING DOMAIN OF THE HUMAN PEROXISOME PROLIFERATOR ACTIVATED RECEPTOR GAMMA

DOI: 10.2210/pdb1prg/pdb

Classification: GENE REGULATION
Deposited: 1998-07-02 **Released**: 2001-01-13
Deposition author(s): Nolte, R.T., Wisely, G.B., Milburn, M.V.
Organism: Homo sapiens
Expression System: Escherichia coli BL21(DE3)

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.2 Å
R-Value Free: 0.318
R-Value Work: 0.246

wwPDB Validation

View in 3D: NGL or JSmol (in Browser)

Standalone Viewers

Simple Viewer Protein Workshop

PDBbind: The PDBbind database provides experimental binding affinity data and computational molecular recognition details for the biomolecular complexes.

Search Methods

1. ***Search through PDB-ID***: The unique four character PDB-ID of the molecule can be used as key word for search.
2. ***Search through molecular name***: The molecular name can be used in place of PDB-ID, when there is no details available.
3. ***Search through author***: The author name can be used when there is no details available.
4. ***Search through deposition data***: The year of deposition and technique used in the molecular structure elucidation can be used in search.

Search options in PDB: The search against the PDB can be performed using the advanced search, Drilldown search, unreleased and new entries, sequences, ligands, drugs and drug targets and browse by annotation options.

PDB Informations

1. **Structure summary:** It contains a description about the research group (authors), publication details (journal), article abstract and PubMed-ID. It also provides taxonomical classification, deposition details, molecular weight, chain details (number and type), fragments, organism type and UniProtKB details.
2. **3D view:** The JS mol, jmol and NGL (webGL) tools are available for the molecular visualization. It provides the following molecular view options.
   (a) ***Style***: Cartoon, back-bone, space fill, ball and stick, trace, ribbon, ligands, ligands and pockets.

(b) ***Colour***: Secondary structure, rainbow sequences, subunits, symmetry and amino acids.

(c) ***Surface***: Solvent accessible and solvent excluded cavities.

3. **Sequence:** It provides sequence display with number of amino acids, chains, domain assignment and secondary structure features. The files download is available for further analysis.
4. **Annotations:** It includes the domain info, class (α- and β-proteins, peptides), fold, super family, family, domain and species. It provides access to domain annotations (SCOP and CATH), protein family annotation (Pfam) and gene product annotation (gene ontology consortium).
5. **Sequence similarity:** Tools namely SS cluster and BLAST for alignment of protein sequences are available.
6. **Structure similarity:** The homology analysis identifies the similar 3D structures for the query (entry).
7. **Experiment:** X-ray crystallography, NMR spectroscopy and other techniques such as electron microscopy details are available. It includes crystal data, instrument parameter and resolution.
8. **Literature:** The information regarding the research team, journal, MeSH details and related articles can be accessed from this section.
9. **Biology and chemical report:** The biology and chemical nature of the molecules can be retrieved from this link.
10. **Methods:** The methods applied in the structure elucidation can be accessed.
11. **Geometry:** The bond length, bond angle and dihedral angle along with the graphical representations of proteins are loaded in this section (Ramachandran plot).

Data deposit in PDB: It offers prepare data, validate data and deposit data features for the research scientists.

Visualization: The tools available for the molecular visualizations are NGL, Jmol, RSCB viewer, Ligand explorer, pose view, protein feature view, human gene view and pathway view.

Analysis: PDB offers the following molecular sequence analysis tools

(a) Sequence and structure alignments

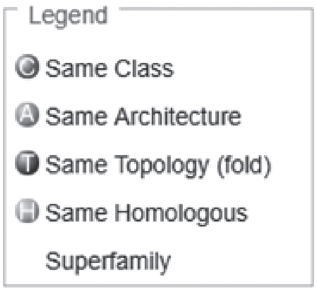
(b) Protein symmetry

(c) Structural quality

(d) Map genomic position to protein

CATH

CATH is a four level hierarchical classification of protein domain structures. The database CATH can be expanded as Class, Architecture, Topology and Homologous super family.



Legend

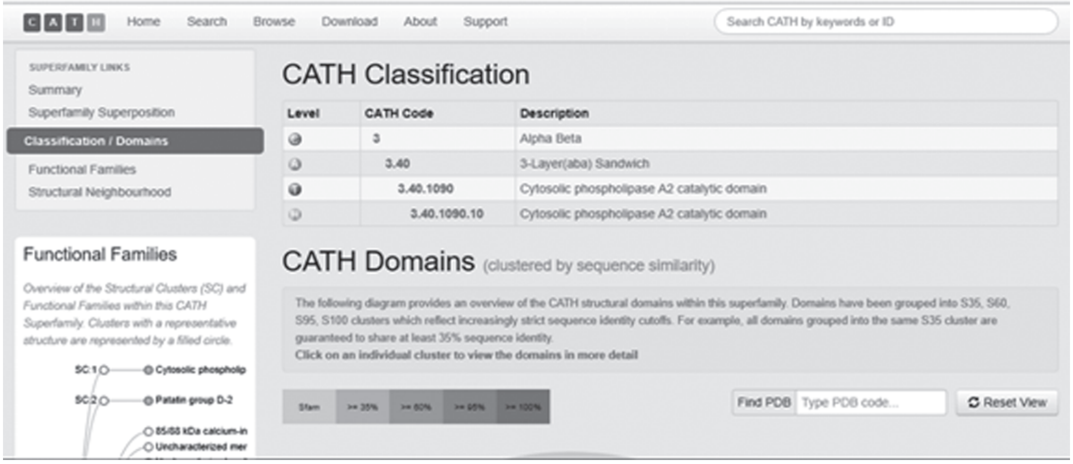
C Same Class

A Same Architecture

T Same Topology (fold)

H Same Homologous Superfamily

1. ***Class***: Proteins can be clustered based on their similar secondary structure content (all α all β, α / β etc). It refers to secondary structure content.
2. ***Architecture***: It refers to the general arrangement of the secondary structural elements, irrespective of the topological connection type.
3. ***Topology:*** It explains about the shape and topological connectivity of structural elements.
4. ***Homologous super family***: It demonstrates the evolutionary relationship among the different proteins.



Functional Families

Overview of the Structural Clusters (SC) and Functional Families within this CATH Superfamily. Clusters with a representative structure are represented by a filled circle.

CATH Classification

Level	CATH Code	Description
●	3	Alpha Beta
●	3.40	3-Layer(aba) Sandwich
●	3.40.1090	Cytosolic phospholipase A2 catalytic domain
●	3.40.1090.10	Cytosolic phospholipase A2 catalytic domain

CATH Domains (clustered by sequence similarity)

The following diagram provides an overview of the CATH structural domains within this superfamily. Domains have been grouped into S35, S60, S95, S100 clusters which reflect increasingly strict sequence identity cutoffs. For example, all domains grouped into the same S35 cluster are guaranteed to share at least 35% sequence identity.

Click on an individual cluster to view the domains in more detail

STRUCTURAL CLASSIFICATION OF PROTEINS (SCOP)

The knowledge of structural similarity relationship is the key factor in determining evolutionary pathway of sequences. The structural classification of proteins (SCOP) database is developed to describe the structural and evolutionary relationship between the protein, whose 3D structures are already characterized. The database is constructed using several automatic software tools meant for visualization. It contains experimentally resolved biomolecular structures of proteins, ribonucleic acid (RNA) and deoxyribonucleic acid (DNA) available in PDB. It provides explicits, chemical graphs, 3D domains, similar sequences, literature, bonds and ligands.

It utilizes four level hierarchies to explain the functions and evolutionary relationship among the proteins.

1. **Class:** A hierarchic structural classification of protein structural domains is based on their amino acid sequence similarities and three-dimensional structures.
   (a) *Class* α: It comprises a bundle of α-helices connected by loops on the surface of the protein.
   (b) *Class* β: It comprises an anti-parrallel β-sheets.
   (c) *Class* α / β: It comprises mainly parallel β-sheets with intervening α-helices.
   (d) *Class* α + β: It comprises mainly segregated α-helices and anti-parallel β-sheet.
   (e) *Multi domain* (α + β): It compares domains of more than one of the four classes (α, β α / β and α + β) mentioned earlier.

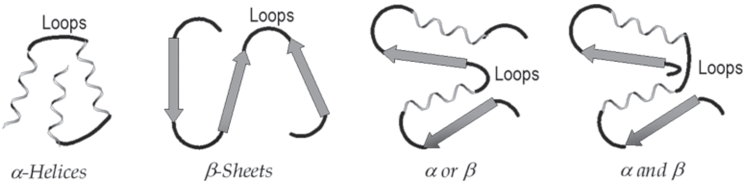


Fig. 1.1 Different classes of proteins in SCOP

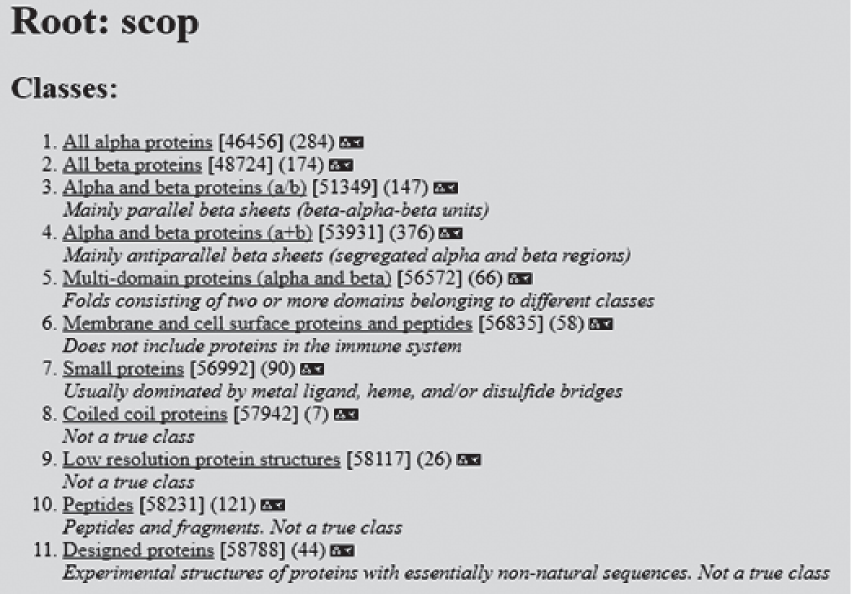
2. **Fold:** The protein with similar secondary structure features are designated as fold. These proteins share common topological connections. It offers no evidence for evolutionary relatedness.

3. **Superfamily:** The proteins with sufficient evidence, that they have diverged from common ancestor are grouped into superfamily. The proteins are clustered into superfamily based on their common functional features. The common functional features infer their structure and evolutionary relationship.

4. **Family:** The proteins with more than 30% residue identities (sequence similarity) and proteins with similar structures as well as function are clustered into single family. The evolutionary relatedness of the proteins can be detected through sequence similarity.

MOLECULAR MODELING DATABASE (MMDB)

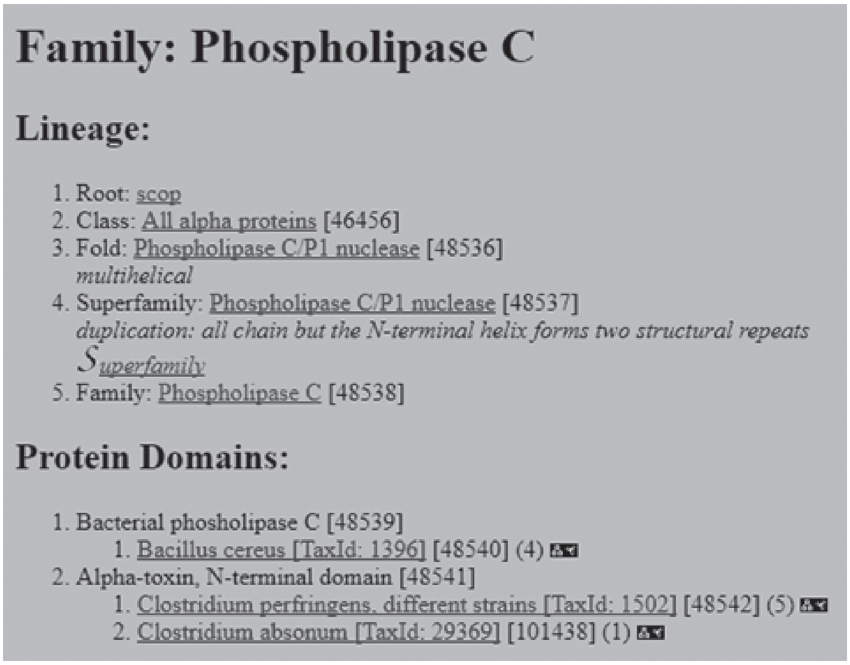
Molecular modeling database (MMDB) is an integral part of NCBI Entrez information retrieval system. The MMDB can be searched by providing key word, accession number, author name and journal name.



Root: scop

Classes:

1. All alpha proteins [46456] (284)
2. All beta proteins [48724] (174)
3. Alpha and beta proteins (a/b) [51349] (147)
   Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) [53931] (376)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) [56572] (66)
   Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides [56835] (58)
   Does not include proteins in the immune system
7. Small proteins [56992] (90)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins [57942] (7)
   Not a true class
9. Low resolution protein structures [58117] (26)
   Not a true class
10. Peptides [58231] (121)
    Peptides and fragments. Not a true class
11. Designed proteins [58788] (44)
    Experimental structures of proteins with essentially non-natural sequences. Not a true class



Family: Phospholipase C

Lineage:

1. Root: scop
2. Class: All alpha proteins [46456]
3. Fold: Phospholipase C/P1 nuclease [48536]
   multihelical
4. Superfamily: Phospholipase C/P1 nuclease [48537]
   duplication: all chain but the N-terminal helix forms two structural repeats
   Superfamily
5. Family: Phospholipase C [48538]

Protein Domains:

1. Bacterial phosholipase C [48539]
   1. Bacillus cereus [TaxId: 1396] [48540] (4)
2. Alpha-toxin, N-terminal domain [48541]
   1. Clostridium perfringens, different strains [TaxId: 1502] [48542] (5)
   2. Clostridium absonum [TaxId: 29369] [101438] (1)

DNA DATABANK

The National Institutes of Health (NIH) developed this database. DNA Data Bank maintains DNA clones of blood, saliva, hair, skin, muscle, liver and other tissues. DNA databank informations are useful in screening genetic diseases, paternity testing, criminal investigations (genetic fingerprinting) and genetic geneology.

COMBINED DNA INDEX SYSTEM (CODIS)

The United States Federal Bureau of Investigation (US-FBI) has developed this database. The Combined DNA Index System (CODIS) maintains a repository of reference DNA sample. It assists in law enforcement agencies in recovering the biological evidence.

DATA MINING

Most information regarding biology and medicine is hidden in text documents (articles). Database mining (literature mining) process helps identifying the scientific concepts buried in the articles. Database mining process discovers valid and potential patterns of information from database. It involves the non-trivial extraction of implicit data (previously unknown and potentially useful). The large amount of bio-molecular data generated from the advancement of molecular biology search increased the number and size of the databases. This growth exceeds human capacity to analyze the databases. The automated search for the non-trivial extraction of implicit, unknown and potentially useful information from these databases is very important.

Database mining, a process to discover valid and potential patterns of information from database. It is related to artificial intelligence (AI).

Data mining is an ideal approach developed for the automated extraction of patterns (hidden predictive information) from large databases. The data mining process is related to artificial intelligence (AI) called knowledge discovery and machine learning. They are concerned with the process of sorting and searching the data, which provides an ideal framework for the application to information retrieval. An attempt to classify molecules based on the pharmacophoric feature is typical example in the data mining process.

Fig. 1.2 Process of data mining.

Data Mining Algorithms

The algorithms precisely define the sequence of operations. Many computer programs for the database search contain algorithms that specify instructions for calculation, data processing and automated reasoning.

(a) ***Artificial neural networks* (*ANN*):** Non-linear predictive models that learn through and resemble biological neural networks in structure.

(b) ***Genetic algorithms (GA*: *evolutionary method)*:** This technique uses the processes such as genetic combination, mutation and natural selection in a design based on the concepts of evolution.

(c) ***Decision trees* (DT):** Tree-shaped structures that represent set of decisions to generate rules for the classification of a dataset. The most versatile tool employed in the extraction of patterns (mining). The algorithms useful are C4.5, C5.0, Classification and Regression Trees (CART) and Chi Automatic Interaction Detection (CAID).

(d) ***K-Nearest neighbor method* (*KNN*):** This technique classifies each record in a dataset based on a combination of the classes. It identifies natural groups of similar objects (clusters) in a data set of interest.

Data Mining Strategies

Search machines retrieves the information from the databases, indexer collects key words and stores them in index. The search machine matches the key word of the query to the index key words and returns the results based on the matches.

1. ***Keyword (phrase) search*:** It removes the stopwords, stems the remaining words and searches for the words in all searchable fields. The punctuation marks and stopwords are replaced with the Boolean AND, but the words in between are retained as phrases.

2. ***Subject headings search*:** It involves the identification of the best matching headings and further identifies the exact matches.

3. ***Concept based searching*:** This search strategy determines the correct query mean term. The query 'heart' apart from coronary, stroke, cholesterol and arteriosclerosis results also returns hits on love and passion.

Data Mining Applications

Data mining focus on the search and establishes the relationship between data types. It is widely used in the fields of bioinformatics, cheminformatics and related fields.

Data mining process describes the patterns of genes and proteins responsible for the particular disease and the therapeutic targets.

1. ***Data mining in Bioinformatics:*** Biological data are very amenable to the concepts of pattern discovery and recognition. Data mining process describes the patterns or collections of descriptors such as amino acids, in protein sequences. A genes and proteins responsible for the particular disease and the therapeutic targets that modulating the diseases can be retrieved. The binding site for the ligands also can be identified. Datamining is useful in mapping the relationship between the inter-individual variation in human DNA sequences and variability in disease susceptibility. It finds the individual DNA sequence changes that are related to the development of diseases such as cancer. It is also useful in the diagnosis, prevention and treatment of the diseases (multifactor dimensionality reduction). The algorithms useful in the prediction of binding pockets are listed below.

(a) ProFunc: It predicts protein function from 3D structure.

(b) Pocket-Finder: A binding pocket detection algorithm.

(c) Q-SiteFinder: A new method of ligand binding site prediction.

(d) P-CATS (Prediction of CATalytic residues in proteins): It predicts the catalytic residues in proteins from the atomic coordinates of ligands.

***Data mining in genetics*:**

(a) The inter-individual variation in human DNA sequences and variability in disease susceptibility can be mapped.

(b) The DNA sequence related to the development of common diseases can be identified.

2. ***Data mining in cheminformatics*:** The large chemical libraries (thousands of chemical compounds) can be utilized in the process of finding efficient molecules (with higher affinity) for the target of interest. The chemical database can be constructed by storing chemical data in searchable format. Pattern recognition supports in the identification of pharmacophore (chemical finger print), toxicophore, bio-isostere and molecular finger print.

   Data mining in adverse drug reaction (ADR) surveillance: It is used to screen drug safety issues in the World Health Organization (WHO) global database and produces a report of suspected adverse drug reactions (ADRs).

Pattern recognition supports in the identification of pharmacophore, toxicophore, bio-isostere and molecular finger print.

3. ***Data mining in literature search:*** It uses an initiating datum to determine how other information's are related to the initiating datum. PubMed contains millions of abstracts. The process of inferring implicit knowledge from biomedical concepts is known as literature based discovery (LBD).

- ***ABC model in LBD*:** A scientific discovery of substance A affects the functions of B (characteristics of disease C). The link between A and C can be revealed through establishing connection with B.
- ***Arrowsmith*:** An automated ABC model.
- ***BITOLA*:** It computes association rules between concepts extracted from MEDLINE.

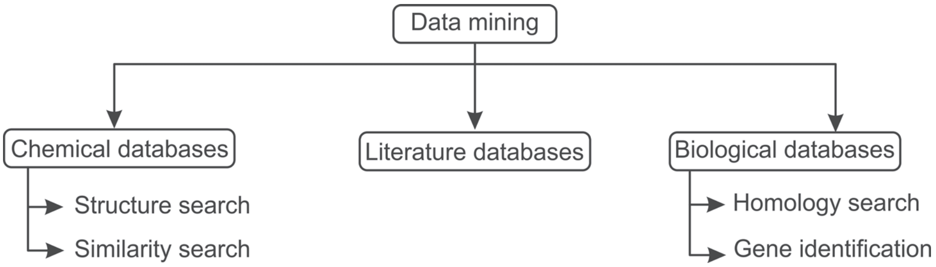


Fig. 1.3 Data mining: Applications of data mining process.

SUMMARY / KEY POINTS

(1) Information and knowledge pertaining to therapeutic candidate research accumulated over many years is basis for the current drug research process.

(2) The increasing pool of chemical and biological data demanded for the development of databases.

(3) Modern drug discovery procedure requires database system (informatics) to access and manipulate larger data.

(4) Informatics integrated with pharmaceutical science research is known as drug informatics (cheminformatics and bioinformatics) and became an essential component of drug discovery.

(5) It encompasses the various methods and algorithms for analyzing and extracting hidden information's from databases.

(6) Literature databases namely MEDLINE, EMBASE, CAS are useful for biological scientists including pharmaceutical scientists.

(7) Cheminformatics includes database creation, physicochemical property calculations and QSAR analysis.

(8) Integration of biology, computer science and information technology developed bioinformatics discipline. It has profound application in the identification of new drug targets with the help of molecular biology and bio-physical techniques (X-ray crystallography and NMR spectroscopy).

(9) Data mining, an artificial intelligence (AI) approach useful in the target identification, detection of disease susceptibility and pharmacophore identification.

REFERENCES

1. Andreeva A. Lessons from making the structural classification of proteins (SCOP) and their implications for protein modelling. Biochem Soc Trans. 2016; 44(3): 937–43.
2. Bajorath J. Improving data mining strategies for drug design. Future Med Chem. 2014; 6(3): 255–7.
3. Bellis LJ, Akhtar R, Al-Lazikani B, Atkinson F, Bento AP, Chambers J, et al. Collation and data-mining of literature bioactivity data for drug discovery. Biochem Soc Trans. 2011; 39(5): 1365-70.
4. Chen B, Wild DJ. PubChem BioAssays as a data source for predictive models. J Mol Graph Model. 2010; 28(5): 420–6.
5. Chen X, Liu M, Gilson MK. BindingDB: a web-accessible molecular recognition database. Comb Chem High-Throughput Screen. 2001; 4(8): 719–25.
6. Cheng A, Diller DJ, Dixon SL, Egan WJ, Lauri G, Merz KM Jr. Computation of the physiochemical properties and data mining of large molecular collections. J Comput Chem. 2002; 23(1): 172–83.
7. Donald AB, Lindberg M. Internet access to the national library of medicine. Eff Clin Pr. 2000; 4: 256–60.
8. Engels MF, Reijmers TH. Data mining applications in drug discovery. In: Bultinick P, de Winter H, Langenaeker W, Tollenaere JP, editors. In: Computational medicinal chemistry for drug discovery. New York: Marcel Dekker, Inc; 2004. p. 669–98.
9. Fatehi F, Gray LC, Wootton R. How to improve your PubMed/MEDLINE searches: 3. advanced searching, MeSH and My NCBI. J Telemed Telecare. 2014; 20(2): 102–12.
10. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong H. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 2016; 44(D1): D1045-1053.
11. Golovin A, Henrick K. Chemical substructure search in SQL. J Chem Inf Model. 2009; 49(1): 22–7.
12. Harper G, Pickett S. Methods for mining HTS data. Drug Discov Today. 2006; 11(15–16): 694–6.
13. Langer T. Bryant SD. Chapter 10 - In silico screening: Hit finding from database mining. In: Wermuth CG, editor. The practice of medicinal chemistry. San Diego: Academic Press; 2003. p. 210–27.
14. Lo Conte L, Ailey B, Hubbard TJ., Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res. 2000; 28(1): 257–9.

15. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database. 2011: Article ID baq03.
16. Mattavelli RGB. Dong JC. Lasseur S. Kopp S. Chapter 43 - Drug nomenclature. In: Wermuth CG, editor. The practice of medicinal chemistry. San Diego: Academic Press; 2003. p. 867–75.
17. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 2015; 43: D213–21.
18. Mitchell JB. Informatics, machine learning and computational medicinal chemistry. Future Med Chem. 2011; 3(4): 451–67.
19. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995; 247(4): 536–40.
20. Opera TI, Gottfries J, Sherbukhin V, Svensson P, Kuhler TC. Chemical information management in drug discovery: optimizing the computational and combinatorial chemistry interfaces. J Mol Graph Model. 2000; 18(4-5): 512–24.
21. Searls DB, Data integration: challenges for drug discovery. 2005; 4(1): 45-58.
22. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S. ChemBank: a small-molecule screening and cheminformatics resource database. Nucleic Acids Res. 2008; 36: D351–9.
23. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, et al. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 2002; 1611–8.
24. Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem. 2004; 47(12): 2977–80.
25. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006; 34: D668-672.
26. Xie X-Q, Chen J-Z. Data mining a small molecule drug screening representative subset from NIH PubChem. J Chem Inf Model. 2008; 48(3): 465–75.