Introduction to bioinformatics (databases)

Course Code –BOTY 4204

Course Title- Techniques in plant sciences , biostatistics and bioinformatics

By – Dr. Alok Kumar Shrivastava

Department of Botany

Mahatma Gandhi Central University, Motihari

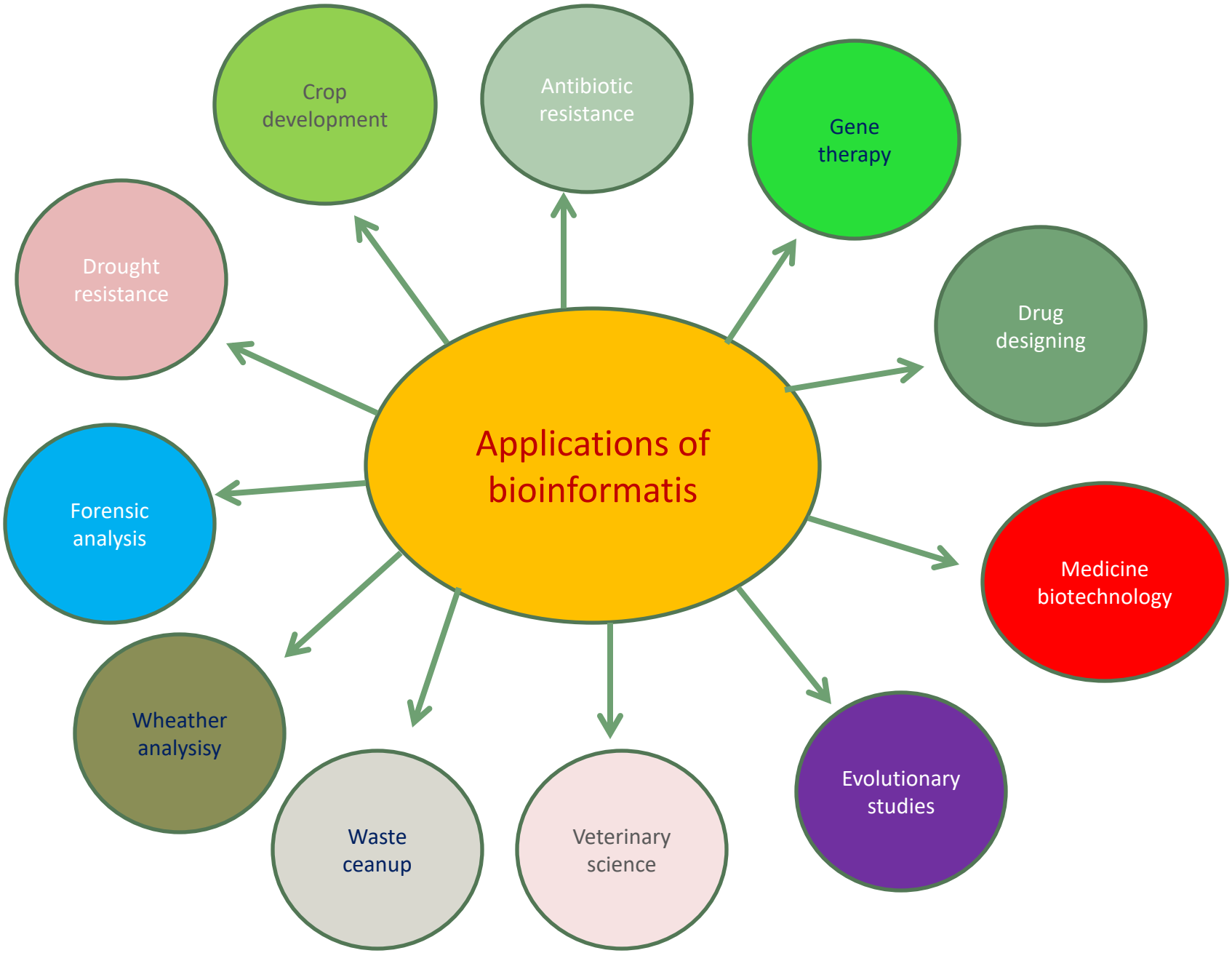
What is bioinformatics ?

- In biology, bioinformatics is defined as, “the use of computer to store, retrieve, analyse or predict the composition or structure of bio-molecules” . Bioinformatics is the application of computational techniques and information technology to the organisation and management of biological data. Classical bioinformatics deals primarily with sequence analysis.

Aims of bioinformatics

- ✓ Development of database containing all biological information.
- ✓ Development of better tools for data designing, annotation and mining.
- ✓ Design and development of drugs by using simulation software.
- ✓ Design and development of software tools for protein structure prediction function, annotation and docking analysis.
- ✓ Creation and development of software to improve tools for analysing sequences for their function and similarity with other sequences

Applications of bioinformatics

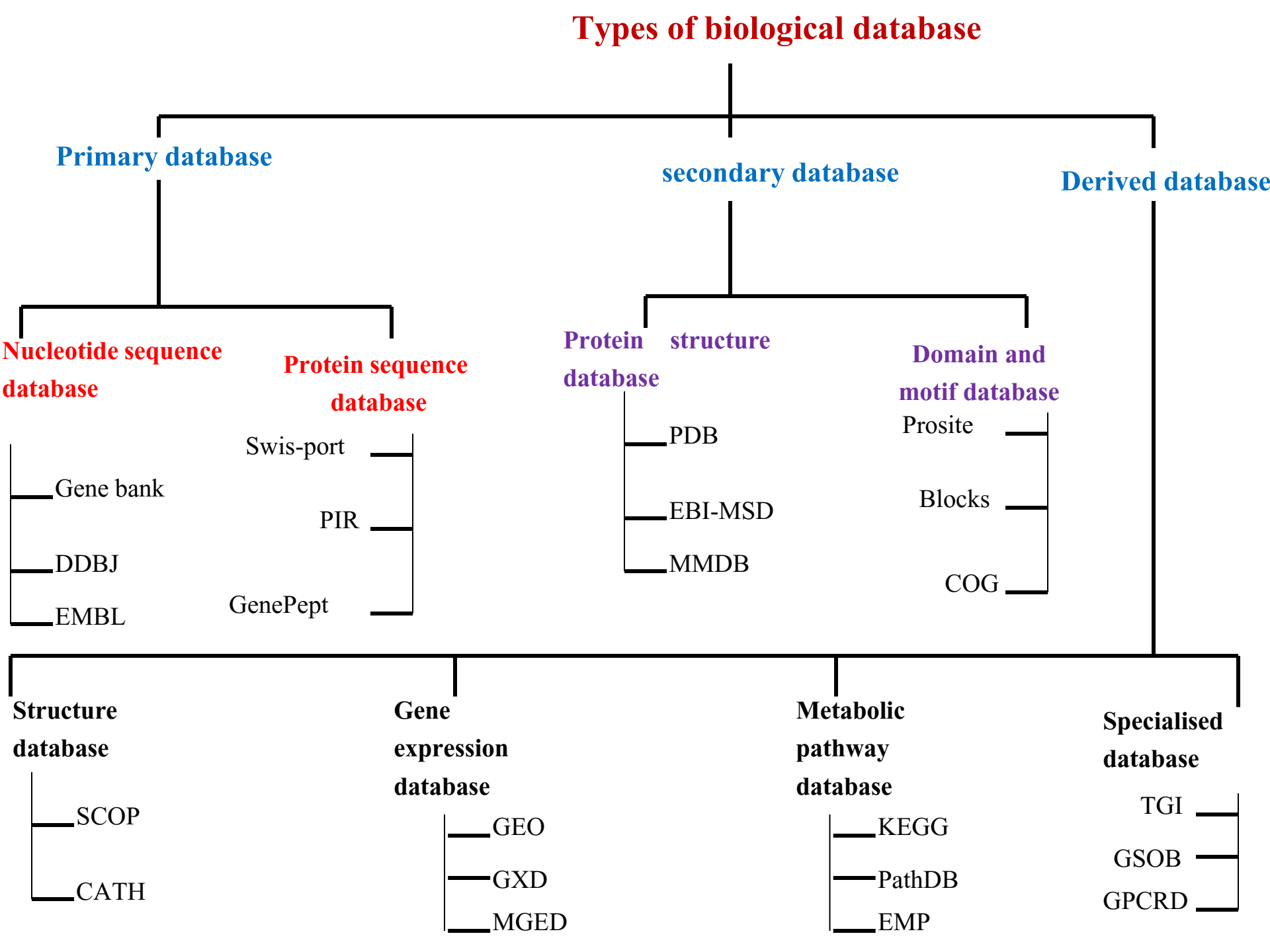


Biological databases

- Biological data are complex, exception-ridden, vast and incomplete. Therefore several databases has been created and interpreted to ensure unambiguous results. A collection of biological data arranged in computer readable form that enhances the speed of search and retrieval and convenient to use is called biological database. A good database must have updated information.

Importance of biological database

- A range of information like biological sequences, structures, binding sites, metabolic interactions, molecular action, functional relationships, protein families, motifs and homologous can be retrieved by using biological databases. The main purpose of a biological database is to store and manage biological data and information in computer readable forms.


Types of biological database
Primary database
secondary database
Derived database
Nucleotide sequence database
Protein sequence database
Protein structure database
Domain and motif database
Gene bank
DDBJ
EMBL
Swis-port
PIR
GenePept
PDB
EBI-MSD
MMDB
Prosite
Blocks
COG
Structure database
Gene expression database
Metabolic pathway database
Specialised database
SCOP
CATH
GEO
GXD
MGED
KEGG
PathDB
EMP
TGI
GSOB
GPCRD

Primary database vs. secondary database

- A primary database contains only sequence or structural information.
- The database derived from the analysis or treatment of primary data are secondary database. It is very important for interfering protein function.

Examples of some primary biological database

GeneBank

- ✓ One of the fastest growing repositories of known nucleotide sequences, GeneBank (Genetic Sequence Databank), has a flat file structure. It is an ASCII text file, readable by both humans and computers. Besides sequence data, GeneBank files contain information such as accession numbers and gene names, phylogenetic classification and references to published literature.

- ✓ This database has been developed and maintained at the NCBI, Bethesda, MD, USA, as a part of International Sequence Database Collaboration (INSDC).

- ✓ It is an open access sequence database.

- ✓ It coordinates with individual laboratories and other sequence databases like EMBL and DDBJ.

Continue……..

✓It is an annotated collection of all nucleotide sequences that are available to the public.

✓The nucleotide database was divided into three databases at NCBI: CoreNucleotide database, Expressed Sequence Tag (EST) and Genome Survey Sequence (GSS).

✓CoreNucleotide database has most of the nucleotide sequences used. It also encloses all nucleotide records that are not in the EST and GSS databases.

✓Submission of sequences to GeneBank can be done using BankIt, Sequin and tbl2asn tools.

EMBL(European Molecular Biology Laboratory)

- A comprehensive database of DNA and RNA sequences, EMBL nucleotide sequence database is collected from scientific literature, patient offices and is directly submitted by researchers. EMBL has been prepared in collaboration with GeneBank (USA) and the DNA Database of Japan (DDBJ).
- It is established in 1980.
- It is maintained by EBI (European Bioinformatics Institute)

Swiss-Port

- ✓ This is a curated protein sequence database that offers a high level of integration with other databases and also has a very low level of redundancy. Swiss-Port strives to provide protein sequences with a high level of annotation (for instance, the description of protein function, domain structure and post translational modifications, etc.).
- ✓ It is established in 1986 and maintained collaboratively , since 1987, by the department of Medical Biochemistry of the University of Geneva and the EMBL data Library.
- ✓ TrEMBL is a computer–annotated supplement of Swiss-Port that contains all translations of EMBL nucleotide sequence entries, which is not yet integrated in Swiss-Port.
- ✓ Currently Swiss-Port have 0.5 and TrEMBL have 7.6 milliom sequences.

Protein Information Resource(PIR)

- PIR is an integrated public bioinformatics resource to support genomic and proteomic research and scientific studies. Nowadays, PIR offers a wide variety of resources mainly oriented to assisting the propagation and consistency of protein annotations like PIRSF, ProClass and ProLINK.

Examples of Some Secondary Biological Database

Motif Databases

- Protein sequence motif is a set of conserved amino acid residues that are important for protein function and are located within a certain distance from one another. These motifs usually provide clues to the functions of otherwise uncharacterised proteins.
- The PROSITE database consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them.
- PRINT is a database for protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family.

Domain Database

- A protein domain is an independently folded, structurally compact unit that forms a steady three- dimensional structure and shows a certain level of evolutionary conservation. Typically , a conserved domain contains one or more motifs.
- ProDom is a protein domain database automatically generated from the Swiss-Port and TrEMBL sequence database.
- SMART is a highly reliable and sensitive tool for domain identification.
- COG is a database and a convenient tool for motif and domain identification.

3D Structure databases

- PDB (Protein Data bank) is the main primary database for 3D structures of biological macromolecules determined by X-ray, crystallography and NMR. It also accepts experimental data used to determine the structures and homology models.
- SCOP (Structural Classification of Protein database) classifies protein 3D structures in a hierarchical scheme of structure classes. All the protein structures in PDB are classified her, and the updated new structures are deposited in PDB.
- The CATH database (Class, Architecture, Topology, Homologous) contains a hierarchical classification of protein domain structure.

Protein data bank

- PDB (Protein data bank) is a repository for 3D structural data obtained by x-ray crystallography or NMR spectroscopy of proteins and nucleic acids.
- Research Collaboratory for Structural Bioinformatics (RCSB) PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationship with other sequences, its function and diseases caused if any .

Gene expression databases

- GEO or Gene Expression Omnibus is a curated online resource and a gene expression molecular abundance repository for gene expression data browsing, query and retrieval.
- GXD or Gene Expression Database is a community resource for gene expression information.
- MGED or Microarray Gene Expression data contains microarray data, generated by functional genomics and proteomics experiments.
- ArrayExpress from European Bioinformatics Institute is a repository for transcriptomics data.

Metabolic pathway databases

- KEGG PATHWAY Database contains graphical pathway maps for all known metabolic pathways from various organisms.
- EcoCyc is an *E. coli* database , stores information regarding the genome and biochemical machinery of *E. coli*.
- LIGAND is a chemical database for enzyme reactions at the Institute for Chemical Research, Kyoto. It is composite database currently consisting of the COMPOUND, DRUG, GLYCAN, REACTION, RPAIR and ENZYME databases.
- MetaCyc is a non-redundant, experimentally elucidated metabolic pathway database.
- BRENDA is an enzyme database tat contains information on all aspects of enzymes and enzymatic reactions.

Genome databases

- Genome databases give absolute information on the heritable properties of an organism. These databases help to identify genes and predict their functions. A few genome databases have links with specific organism databases.
- GOLD (Genomes Online Database at the University of Illinois, USA) contains a list of all the complete and ongoing genome projects worldwide.
- Genomes at NCBI (National Centre for Biotechnology Information, USA).
- TIGR database (TDB), at the institute for Genomic Research at Rockeville MD, USA.

Virological databases

- A virological database contains all the sequences and related information of viruses of animals, plants, bacteria, fungi and archea; for example, the HIV protease database. A committee called The International Committee on Taxonomy of Viruses(ICTV) authorises and organises the taxonomic classification of viruses. ICTVdB contains taxonomic information for over thousands of virus species.

World biodiversity databases

- Taxonomic databases are built to document all known species and make them available and accessible worldwide. These databases contain taxonomic hierarchies, species names, synonyms, descriptions, illustrations and references. For example: CCINFO, STRAIN and ALGAE.

Database for various model organisms

- *Escherichia coli-* *E. coli* Genome Centre(Wisconsin university, USA), The *E.coli* index (University of Birmingham, UK)
- *Arabidopsis thaliana*- TAIR (The *Arabidopsis* Information Resource)
- *Homo sapiens-* Human Genome Resources at NCBI, USA
- *Oryza sativa* (rice) -*RGP* (Rice Genome Research Programme, Japan)
- *Drosaphila* melanogaster -FlyBase (*Drosophila* Genome Database)
- *Mus musculus* (mouce)- Mouce Genome Informatics
- *Danio rerio*(zebrafish)- ZFIN (Zebrafish Information Network at the University of Oregon, USA)
- *S. cerevisiae* (Bakers yeast)- SGD ()Yeast Genome Database at Stanford, USA

Annotation of Gene ?????

In molecular biology, genomes make the basic genetic material and typically consist of DNA. Whereby, genome include the genes (coding) and non coding regions, of interest to us, are the coding regions as they actively influence basic life processes. The genes contain useful biological information that is required in building up and maintaining an organism. Gene annotation can be defined merely as the process of making nucleotide sequence meaningful.

Gene annotation involves the process of taking the raw DNA sequence produced by genome sequencing projects and adding layers of analysis and interpretation necessary to extracting biologically significant information and placing such derived details into context. Annotation is the process by which pertinent information about these raw DNA sequences is added to the databases.

Accession number

Accession numbers are unique identifiers which permanently identify sequences in the database. Accession numbers are assigned and communicated to authors within two working days of the receipt of submission.

Thank you