UNIT-2- Introduction to Bio-Database

2.1. Importance of Bio-database, Types of Biological data primary sequence databases, Composite sequence databases, Secondary databases

2.2. Nucleic acid sequence databases, Protein sequence data bases, structure databases

Metabolic pathway databases

2.3. Genome Browsers (Ensembl, NCBI map viewer, UCSC Genome Browse

2.4. Bioinformatics Database search engines, Bibliographic specialized genomic resources

data analysis packages

2.5. Taxonomic databases and biodiversity databases

2.1.Importance of Bio-database, Types of Biological data primary sequence databases, Composite sequence databases, Secondary databases

The explosion of biological database adoption among researchers, many in laboratories without dedicated informatics infrastructures, is driven in large part by need as the types and scope of data produced by modern technologies far outpaces our ability to properly collate the data. To illustrate this point, the Human Genome alone would occupy over 180,000 pages when printed out at a 4.5-point font, and finding meaningful information within it would require equally inefficient volumes of indexed data. Compounding the obviously unmanageable scale of data, there is the need to articulate an endless variety of data types, spanning character-based data, images and proprietary data types. The generic notion of a database is designed explicitly to mediate the centrality of these issues.

The drastic increase in database requirements coincided with the emergence of sophisticated open-source relational DBMS, such as MySQL and PostgreSQL. These systems brought free, robust, and flexible relational databases into the realm of the average biologist, effectively removing the need of costly unsupportable informatics overhead associated with proprietary systems such as Oracle or DB2. Biologists, in turn, began to effectively spread boutique bioinformatics databases with minimal entry requirements. The emergence of need and the ubiquitously standardized relational database has pushed researchers to adopt practices that only a decade ago seemed insurmountable. They have embraced a digitized life; gained an appreciation, albeit a subconscious one, of atomic data types; have rationalized the benefits of extensible data models; and have structured future experimentation planning around compatibility.

A database is an organized collection of data. For instance, a list with some of the movies that we like would be a movie database:

ID	**title**	**year**	**director**
movie1	The player	1992	Robert Altman
movie2	Cookie’s fortune	1999	Robert Altman
movie 3	The man who shot Liberty Valance	1962	John Ford

Vocabulary:

- *Entities*: The kind of things that we want to store in a database. E.g.: Genes, DNA sequences, bibliographical references.
- *Records*: The particular things stored in the database. E.g.: The gene BRCA1
- *Identifiers* or *key*: The unique name that identifies a record
- *Fields*: The properties that an entity has. E.g.: The name, sequence and mutations of the gene

In the previous movie examples the entities stored were movies, the records stored were: The player, Cookie’s fortune and The man who shot Liberty Valance. The unique idenfiers were: movie1, movie2 and movie3.

So if we think on the database as a table, the table would store information about one entity, the fields would be the column headers and the records would be the table rows.

It is quite common to store different entities in a database. For instance we could store movies, actors and directors or genes, sequences and mutations. In that case, the different entities could be stored in different tables and the records on those tables would be related by their unique identifiers. That structure would comprise a relational database.

The databases usually provide mechanisms to store, search, retrieve and modify the data.

2.1.1. Biological Databases

The data repositories more relevant to the biological sciences include:

- nucleotide and protein sequences
- protein structures
- genomes

- genetic expression
- bibliography

Main sequence databases:

- NCBI
- EMBL

Main protein databases:

- Uniprot
- PDB
- MMDB

Some genome databases:

- ENSEMBL (Human, mouse and others)
- SGD (Yeast)
- TAIR (Arabidopsis)

Bibliography:

- Pubmed
- Web of Science

Human diseases:

- OMIM

Metabolic pathways:

- KEGG

Biological databases emerged as a response to the huge data generated by low-cost DNA sequencing technologies. One of the first databases to emerge was GenBank, which is a collection of all available protein and DNA sequences. It is maintained by the National Institutes of Health (NIH) and the National Center for Biotechnology Information (NCBI). GenBank paved the way for the Human Genome Project (HGP). The HGP allowed complete sequencing and reading of the genetic blueprint. The data stored in biological databases is organized for optimal analysis and consists of two types: raw and curated (or annotated). Biological databases are complex, heterogeneous, dynamic, and yet inconsistent. The inconsistency is due to the lack of standards at the ontological level.

2.1.2. Why are these Important?

Earlier, databases and databanks were considered quite different. However, over the time, database became a preferable term. Data is submitted directly to biological databases for indexing, organization, and data optimization. They help researchers find relevant biological data by making it available in a format that is readable on a computer. All biological information is readily accessible through data mining tools that save time and resources. Biological databases can be broadly classified as sequence and structure databases. Structure databases are for protein structures, while sequence databases are for nucleic acid and protein sequences.

Kinds of Biological Databases

2.1.3. Biological databases can be further classified as primary, secondary, and composite databases.

Primary databases contain information for sequence or structure only. Examples of primary biological databases include:

- Swiss-Prot and PIR for protein sequences
- GenBank and DDBJ for genome sequences
- Protein Databank for protein structures

Secondary databases contain information derived from primary databases. Secondary databases store information such as conserved sequences, active site residues, and signature sequences. Protein Databank data is stored in secondary databases. Examples include:

- SCOP at Cambridge University
- CATH at the University College of London
- PROSITE of the Swiss Institute of Bioinformatics
- eMOTIF at Stanford

Composite databases contain a variety of primary databases, which eliminates the need to search each one separately. Each composite database has different search algorithms and data structures. The NCBI hosts these databases, where links to the Online Mendelian Inheritance in Man (OMIM) is found.

2.1.4.The Future

Because of high-performance computational platforms, these databases have become important in providing the infrastructure needed for biological research, from data preparation to data extraction. The simulation of biological systems also requires computational platforms, which further underscores the need for biological databases. The future of biological databases looks bright, in part due to the digital world.

In terms of research, bioinformatics tools should be streamlined for analyzing the growing amount of data generated from genomics, metabolomics, proteomics, and metagenomics. Another future trend will be the annotation of existing data and better integration of databases.

With a large number of biological databases available, the need for integration, advancements, and improvements in bioinformatics is paramount. Bioinformatics will steadily advance when problems about nomenclature and standardization are addressed. The growth of biological databases will pave the way for further studies on proteins and nucleic acids, impacting therapeutics, biomedical, and related fields. If you use biological databases and would like to share any insights, comment in the section below

2.2. Nucleic acid sequence databases, Protein sequence data bases, structure databases

Metabolic pathway databases

The Nucleic Acid Database (NDB) was established in 1991 as a resource for specialists in the field of nucleic acid structure. Its purpose was to gather all the structural information about oligonucleotides that had been obtained from x-ray crystallographic experiments and to organize them in such a way that it would be easy to retrieve the coordinates, the information about the experimental conditions used to derive these coordinates, and the structural information that could be derived from these coordinates. It was clear from the beginning that many of the users of these data would not themselves be crystallographers, and that the information provided by the database had to be presented in such a way as to maximize its utility for various types of modeling and structure prediction.

As the project progressed, many new technologies developed that presented challenges and opportunities. These include the development of the standard interchange format for handling crystallographic data, called the Macromolecular Crystallographic Information File (mmCIF), and the explosive use of the World Wide Web (WWW).

2.3. Database Contents

Structures available in the NDB include RNA and DNA oligonucleotides with two or more bases. These oligonucleotides may be complexed with drugs and ions. Structures of larger nucleic acid containing crystals, including protein-DNA and protein-RNA structures, are also curated and included in the archive. Table 1 shows the current holdings of the NDB.

he Nucleic Acid Database (NDB) was founded in 1991 to assemble and distribute structural information about nucleic acids (1). In addition to the primary structural data that are contained in the archival Protein Data Bank (PDB) (2), the NDB contains annotations specific to nucleic acid structure and function, as well as tools that enable users to search, download, analyze and learn more about nucleic acids. NDB is thus a value-added database providing services specifically for the nucleic acid community.

When the NDB was first established, the focus was on DNA structural biology. As more RNA structures have been determined (Figure 1), tools and annotations were developed to address the features of these molecules.

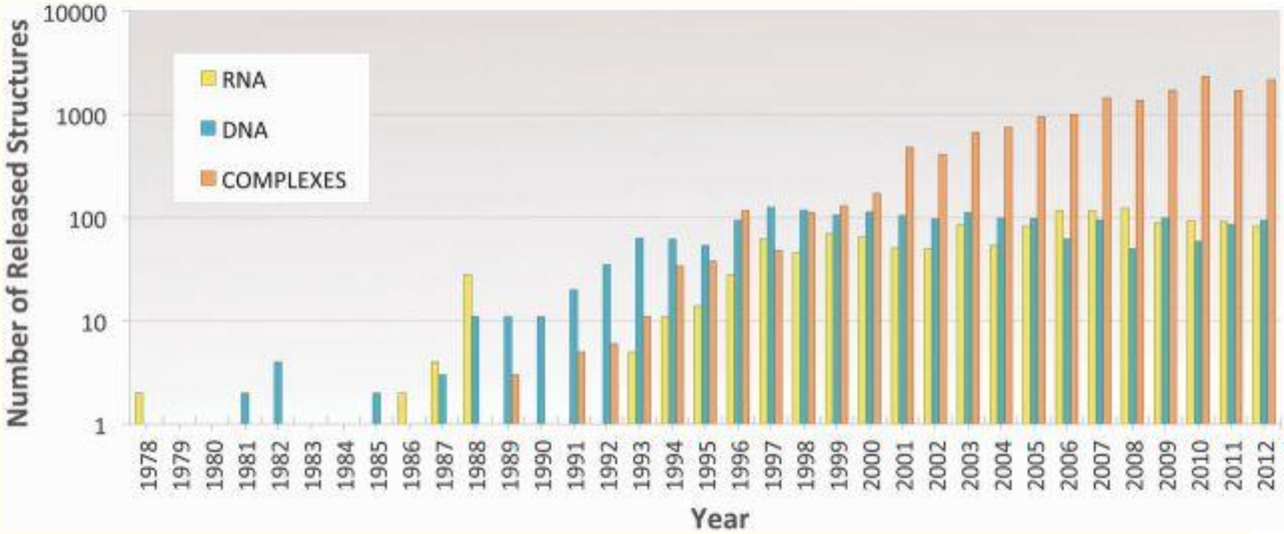


Figure 2.1. Sequence data

Growth of the number of nucleic acid structures in NDB. The total number of structures released in log scale per year for RNA (yellow), DNA (blue) and protein-nucleic acid complexes (orange) is shown.

The NDB seeks to be a central source for nucleic acid structural information and annotations that evolves with the science. In this article we describe the recent redesign of the NDB Web site with special emphasis on new RNA-derived data and annotations and their implementation and integration into the search capabilities.

2.4. Metabolic pathway

- **BRENDA**, the enzyme database, has comprehensive information on enzymes and enzymatic reactions. It is one of several databases nested within the metabolic pathway database set of the **SRS5 sequence retreival system** at EBI.

- **KEGG Metabolic Pathways** include graphical pathway maps for all known metabolic pathways from various organisms. Ortholog group tables, containing conserved, functional units in a molecular pathway or assembly as well comparative lists of genes for a given functional unit in different organisms, are also available.

- **The WIT Metabolic Reconstruction project** produces metabolic reconstructions for sequenced, or partially sequenced, genomes. It currently provides a set of over 25 such reconstructions in varying states of completion. Over 2900 pathway diagrams are available, associated with functional roles and linked to ORFs.

- **EcoCyc** describes the genome and the biochemical machinery of E. coli. It provides a molecular and functional catalog of the *E. coli* cell to facilitates system-level understanding. Its Pathway/Genome Navigator user interface visualizes the layout of genes, of individual biochemical reactions, or of complete pathways. It also supports computational studies of the metabolism, such as pathway design, evolutionary studies, and simulations. A related metabolic database is **Metalgen**.

- **Boehringer Mannheim - Biochemical Pathways** is a searchable database of metabolic pathways, enzymes, substrates and products. Based on a given search, it produces a graphic representation of the relevant pathway(s) within the context of an enormous metabolic map. Neighboring metabolic reactions can then be viewed through links to adjacent maps.

2.3.Genome Browsers (Ensembl, NCBI map viewer, UCSC Genome Brows

2.3.1. Genome browser

In bioinformatics, a genome browser is a graphical interface for display of information from a biological database for genomic data. Genome browsers enable researchers to visualize and browse entire genomes with annotated data including gene prediction and structure, proteins, expression, regulation, variation, comparative analysis, etc. Annotated data is usually from multiple diverse sources. They differ from ordinary biological databases in that they display data in a graphical format, with genome coordinates on one axis with annotations or space-filling graphics to show analyses of the genes, such as the frequency of the genes, their expression profiles, etc.

large number of genome browsers are available, many of them free and with database accessible online. Among the best known are the UCSC Genome Browser, Ensembl Genome Browser and NCBI's Genome Data Viewer. These genome browsers may support multiple genomes, however, other genome browsers may be specific for particular species. These browsers may provide summary of data from genomic databases and comparative assessment of different genetic sequences across multiple species, and allow the data to be visualised in various ways to facilitate assessment and interpretation of these complex data

2.3.2.Ensembl is a genome browser

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 100 (April 2020)

Update to GENCODE 34 (human) and GENCODE M25 (mouse)

Update of gnomAD genomic allele frequencies to version 3

New genomes: 3 mammals, 7 fish, 6 birds, 4 reptiles

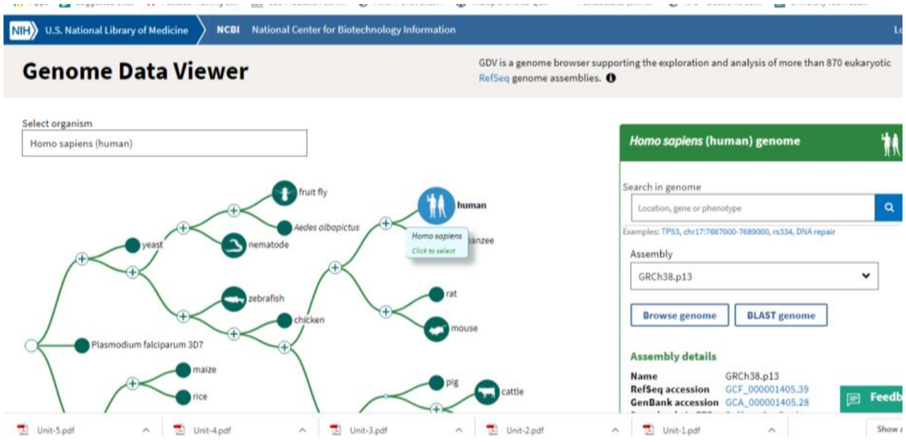
Updated genomes: Platypus and Northern Pike

New interface for configuration of multidimensional track hubs

The advent of the human genome project and subsequent projects to sequence genomes of other species and multiple individuals has driven the need for tools that can visualize vast amounts of genomics data. Software for genome browsing has had a vast impact in the arenas of human medical and genetics research, enabling researchers to process and integrate different data types from a large variety of sources. Three major genome browsers are freely accessible online — the University of California, Santa Cruz (UCSC) Genome Browser, the Wellcome Trust Sanger Institute (WTSI)/European Bioinformatics Institute (EBI) Ensembl browser and the National Center for Biotechnology Information (NCBI) MapViewer. The UCSC Genome Browser is a key part of the UCSC Genome Bioinformatics suite of integrated tools that facilitate data mining, together with allowing users to visualize and query their own data in the context of the existing Genome Browser annotations. The chapter provides an overview of the types of annotation data displayed by the Genome Browser, as well as step-by-step examples illustrating how to create custom tracks and query both the Genome Browser and Table Browser. The Genome Browser offers links to several programs: BLAT for performing fast sequence alignment to genomes; the In Silico PCR tool for aligning primers to the genome, and liftOver for converting genomic coordinates from one assembly to another. Other tools in the suite include the Gene Sorter for sorting genes based on their relationships such as expression profiles and genomic proximity; the Proteome Browser, which shows protein-related information in graphical form and links out to external protein-related sites; and Genome Graphs, which allows the user to display genome-wide datasets such as those from SNP association studies, linkage studies and homozygosity mapping. The suite of UCSC Genome Bioinformatics tools, data downloads, extensive documentation and links to further training materials can be found at http://genome.ucsc.edu.

2.3.3. NCBI MAP VIEWER

Introduction to the Map Viewer, describing how to perform a simple text-based search of genome annotations to view the genomic context of a gene, navigate along a chromosome, zoom in and out, and change the displayed maps to hide and show information. It also describes some of NCBI's sequence-analysis tools, which are provided as links from the Map Viewer. The Alternate Protocols describe different ways to query the genome sequence, and also illustrate additional features of the Map Viewer. Alternate Protocol 1 shows how to perform and interpret the results of a BLAST search against the human genome. Alternate Protocol 2 demonstrates how to retrieve a list of all genes between two STS markers. Finally, Alternate Protocol 3 shows how to find all annotated members of a gene family.



GDV is a genome browser supporting the exploration and analysis of more than 870 eukaryotic RefSeq genome assemblies.

Select organism: Homo sapiens (human)

Search in genome: Location, gene or phenotype

Examples: TP53, chr17:7667000-7689000, rs334, DNA repair

Assembly: GRCh38.p13

Assembly details

Name: GRCh38.p13

RefSeq accession: GCF_000001405.39

GenBank accession: GCA_000001405.28

Fig.2.2 Genome data viewer

We will use as an example the glial cell derived neurotrophic factor GDNF. GDNF may contribute to Hirschsprung disease when mutated. Search for "GDNF" in the Map Viewer. Access the Human Map Viewer page by clicking on Map Viewer. Further, access the current human genome assembly by selecting the "Homo sapiens (human) Build 36" link. Enter GDNF as a search term and click on the "Find" button. Filter your search result for gene maps by checking the "Gene" box in the "Quick Filter" menu.

Name the chromosome on which this gene is located. Click on the Genes_seq map link of the map element GDNF in the reference assembly.What is the map that is displayed? Turn on the ruler line for the map through the Maps & Options link. What are the nucleotide locations for the gene on the chromosome? What is the orientation of the gene on the chromosome? Download the gene sequence by using the "dl" link. (Change the strand, if necessary). In order to search for promoter elements, you may also download 5000 nucleotides of upstream sequence by adjusting the locations for the upstream 5000 nucleotidesAdd the Clone, Component and Contig maps for this region. Name the contig and GenBank accession numbers for the sequence covering this region. Are the sequences in the finished form? Is there a clone mapped to this region? If so, how can you order it?Remove all the maps except Genes_seq and add the Ab initio (model) and Transcript (RNA) maps. Does the gene prediction match the current gene annotation? How many alternatively spliced transcripts have been annotated for the gene? Display the current data as

"Data As Table View".Using the Model Maker (mm), obtain a possible alternatively spliced product and its translated amino acid sequence. Search for similar proteins by using BLAST.

Go back the Map Viewer report. Remove all the maps except Genes_seq and add the Gene maps for mouse, chimp and rat. Are the gene structures in the three organisms similar?Remove all the maps except the human Gene map, and add the phenotype map. Name the disease with which the GDNF gene is associated. Obtain more information about the disease by linking to the corresponding OMIM record.

2.3.5. Repeat the above procedure for the human PRNP gene.

Search for "PRNP" in the Map Viewer. Access the Human Map Viewer page by clicking on Map Viewer. Further, access the current human genome assembly by selecting the "Homo sapiens (human) Build 36" link. Enter PRNP as a search term and click on the "Find" button. Filter your search result for gene maps by checking the "Gene" box in the "Quick Filter" menu.

Name the chromosome on which this gene is located. Click on the Genes_seq map link of the map element PRNP in the reference assembly.

What is the map that is displayed? Turn on the ruler line for the map through the Maps & Options link. What are the nucleotide locations for the gene on the chromosome? What is the orientation of the gene on the chromosome?

Download the gene sequence by using the "dl" link. (Change the strand, if necessary). In order to search for promoter elements, you may also download 5000 nucleotides of upstream sequence by adjusting the locations for the upstream 5000 nucleotides.

Add the Clone, Component and Contig maps for this region. Name the contig and GenBank accession numbers for the sequence covering this region. Are the sequences in the finished form? Is there a clone mapped to this region? If so, how can you order it?

Remove all the maps except Genes_seq and add the Ab initio (model) and Transcript (RNA) maps. Does the gene prediction match the current gene annotation? How many alternatively spliced transcripts have been annotated for the gene? Display the current data as "Data As Table View".

Using the Model Maker (mm), obtain a possible alternatively spliced product and its translated amino acid sequence. Search for similar proteins by using BLAST.

Go back the Map Viewer report. Remove all the maps except Genes_seq and add the Gene maps for mouse, chimp and rat. Are the gene structures in the three organisms similar?

Remove all the maps except the human Gene map, and add the phenotype map. Name the disease with which the PRNP gene is associated. Obtain more information about the disease by linking to the corresponding OMIM record.

2.4. Bioinformatics Database search engines Bibliographic specialized genomic resources data analysis packages

2.4.1. Bibliographic specialized genomic resources

a. ArrayExpress

The ArrayExpress Archive is a database of functional genomics experiments including gene expression where you can query and download data collected to MIAME and MINSEQE standards. Gene Expression Atlas contains a subset of curated and re-annotated Archive data which can be queried for individual gene expression under different biological conditions across experiments.

b. Genome size database

Animal genome size data. Haploid DNA contents (C-values, in picograms) are currently available for 4972 species (3231 vertebrates and 1741 non-vertebrates) based on 6518 records from 669 published sources.

c. MorphoBank

MorphoBank displays dynamic phylogenetic matrices of morphological characters with labeled images demonstrating homology statements, and implements the data editing functions of widely used desktop programs (e.g., Mesquite, Nexus Data Editor) in a password protected environment.

d. NERC Environmental Bioinformatics Centre

The NEBC was established in 2002 to provide bioinformatics, data management and computing supporting to the NERC research community using 'omics technologies. NEBC now collaborates with and supports environmental scientists generating and using a range of molecular data types in their environmental research. Our collaborators are based both in the UK and internationally.

2.4.2. Data analysis packages

Bioinformatics tools can be used to obtain sequences of genes or proteins of interest, either from material obtained, labeled, prepared and examined in electric fields by individual researchers/groups or from repositories of sequences from previously investigated material.

Both types of sequence can then be analyzed in many ways with bioinformatics tools.

They can be assembled. Note that this is one of the occasions when the meaning of a biological term differs markedly from a computational one (see the amusing confusion over the issue at Web-based geek forum Slashdot). Computer scientists, banish from your mind any thought of assembly language. Sequencing can only be performed for relatively short stretches of a biomolecule and finished sequences are therefore prepared by arranging overlapping "reads" of monomers (single beads on a molecular chain) into a single continuous passage of "code". This is the bioinformatic sense of assembly.

They can be mapped***---that is, their sequences can be parsed to find sites where so-called "restriction enzymes" will cut them.

They can be compared, usually by aligning corresponding segments and looking for matching and mismatching letters in their sequences. Genes or proteins that are sufficiently similar are likely to be related and are therefore said to be "homologous" to each other---the whole truth is rather more complicated than this. Such cousins are called "homologs".If a homolog (a related molecule) exists, then a newly discovered protein may be modeled---that is the three dimensional structure of the gene product can be predicted without doing laboratory experiments.Bioinformatics is used in primer design. Primers are short sequences needed to make many copies of (amplify) a piece of DNA as used in PCR (the Polymerase Chain Reaction).Bioinformatics is used to attempt to predict the function of actual gene products.

Information about the similarity, and, by implication, the relatedness of proteins is used to trace the "family trees" of different molecules through evolutionary time.

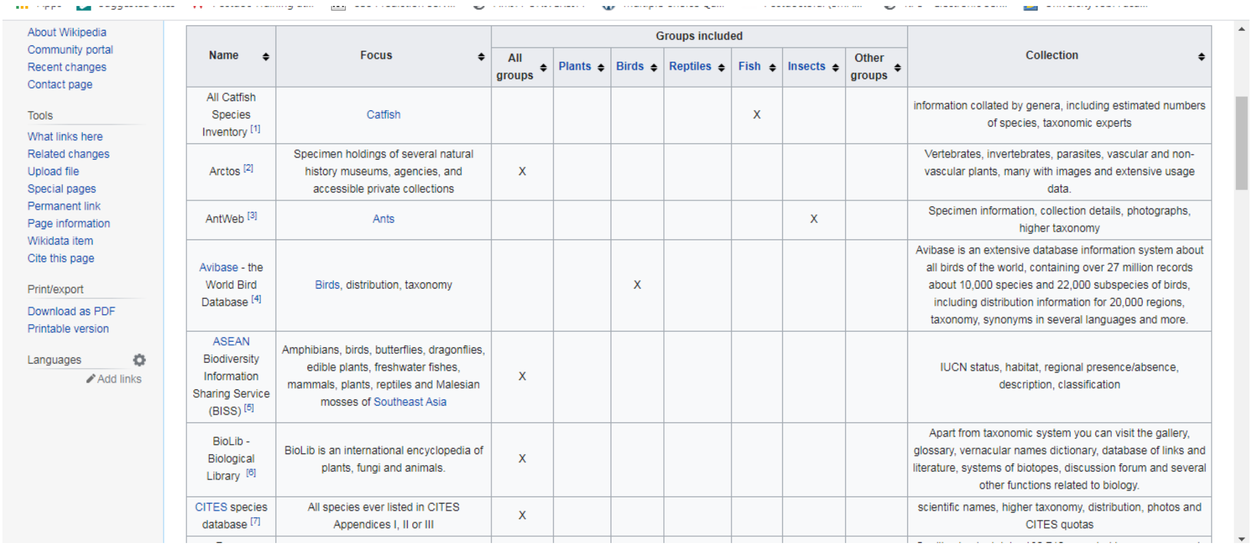
There are various other applications of computer analysis to sequence data, but, with so much raw data being generated by the Human Genome Project and other initiatives in biology, computers are presently essential for many biologists just to manage their day-to-day resultsMolecular modelling / structural biology is a growing field which can be considered part of bioinformatics. There are, for example, tools which allow you (often via the Net) to make pretty good predictions of the secondary structure of proteins arising from a given amino acid sequence, often based on known "solved" structures and other sequenced molecules acquired by structural biologists.

Structural biologists use "bioinformatics" to handle the vast and complex data from X-ray crystallography, nuclear magnetic resonance (NMR) and electron microscopy investigations and create the 3-D models of molecules that seem to be everywhere in the media

2.5.Taxonomic databases and biodiversity databases

2.5.1. Taxonomic databases & Biodiversity databases

This is a list of biodiversity databases. Biodiversity databases store taxonomic information alone or more commonly also other information like distribution (spatial) data and ecological data, which provide information on the biodiversity of a particular area or group of living organisms. They may store specimen-level information, species-level information, information on nomenclature, or any combination of the above. Most are available online.Specimen-focused databases contain data about individual specimens, as represented by vouchered museum specimens, collections of specimen photographs, data on field-based specimen observations and morphological or genetic data. Species-focused databases contain information summarised at the species-level. Some species-focused databases attempt to compile comprehensive data about particular species (FishBase), while others focus on particular species attributes, such as checklists of species in a given area (FEOW) or the conservation status of species (CITES or IUCN Red List). Nomenclators act as summaries of taxonomic revisions and set a key between specimen-focused and species-focused databases. They do this because taxonomic revisions use specimen data to determine species limits.



Name	Focus	Groups included							Collection
		All groups	Plants	Birds	Reptiles	Fish	Insects	Other groups	
All Catfish Species Inventory [1]	Catfish					X			information collated by genera, including estimated numbers of species, taxonomic experts
Arctos [2]	Specimen holdings of several natural history museums, agencies, and accessible private collections	X							Vertebrates, invertebrates, parasites, vascular and non-vascular plants, many with images and extensive usage data.
AntWeb [3]	Ants						X		Specimen information, collection details, photographs, higher taxonomy
Avibase - the World Bird Database [4]	Birds, distribution, taxonomy			X					Avibase is an extensive database information system about all birds of the world, containing over 27 million records about 10,000 species and 22,000 subspecies of birds, including distribution information for 20,000 regions, taxonomy, synonyms in several languages and more.
ASEAN Biodiversity Information Sharing Service (BISS) [5]	Amphibians, birds, butterflies, dragonflies, edible plants, freshwater fishes, mammals, plants, reptiles and Malesian mosses of Southeast Asia	X							IUCN status, habitat, regional presence/absence, description, classification
BioLib - Biological Library [6]	BioLib is an international encyclopedia of plants, fungi and animals.	X							Apart from taxonomic system you can visit the gallery, glossary, vernacular names dictionary, database of links and literature, systems of biotopes, discussion forum and several other functions related to biology.
CITES species database [7]	All species ever listed in CITES Appendices I, II or III	X							scientific names, higher taxonomy, distribution, photos and CITES quotas

References:

1.https://bioinf.comav.upv.es/courses/biotech3/theory/databases.html

2.B Thiagarajan, Pa Rajalakshmi,Computational Biology,MJP Publisher, 12-Jun-2019 – Science

3.https://en.wikipedia.org/wiki/List_of_biodiversity_databases

4.Gordon B. Curry, Chris J. Humphries,Biodiversity Databases: Techniques, Politics, and Applications,CRC Press, 19-Apr-2016 - Computers