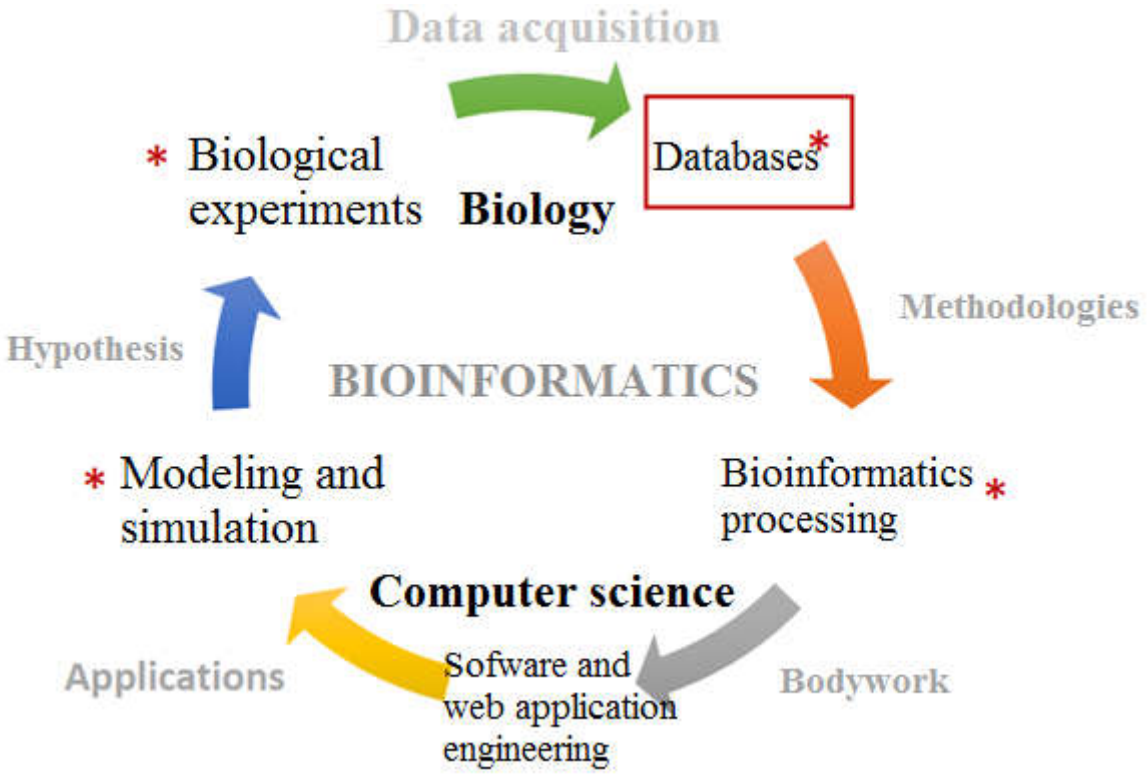
CHAPTER II.

BIOLOGICAL DATA BANKS

Introduction

Databases containing biological information and data are widely disseminated via the internet. They are generally interconnected through links. There are a large number of databases of biological interest. Their main mission is to make publicly available the sequences that have been determined; thus, one of the primary interests of these banks is the vast amount of sequences they contain. Among other things, they are responsible for archiving, storing, disseminating, and exploiting biological data.



II. Databases

A database is a structured and organized set that allows for the storage of large amounts of information, making it easier to use (adding, updating, searching, and possibly analyzing).

We can distinguish between two types of databases: those that aim for the most exhaustive data collection possible, resulting in a rather heterogeneous set of information (**general databases**), and those that focus on more homogeneous data built around a specific theme (**specialized databases**), providing added value through a particular technique or interest driven by a group of scientists.

Table 1. Some general nucleotide sequence databases

→ General nucleotide sequence databases			
Name	**Link**	**Creation Date**	**Description**
EMBL	http://www.ebi.ac.uk/embl/	1980	European database (European Molecular Biology Laboratory) distributed by EBI (European Bioinformatics Institute, Cambridge)
GenBank	http://www.ncbi.nlm.nih.gov/	1982	American database distributed by NCBI (National Center for Biotechnology Information, Los Alamos)
DDBJ	http://www.ddbj.nig.ac.jp/	1986	DNA Data Bank of Japan distributed by NIG (National Institute of Genetics)
→ General protein sequence databases			
Name	**Link**	**Creation Date**	**Description**
UniProt	https://www.uniprot.org/	1986	Annotated sequences & coding sequences translated from EMBL

Table 2. Some specialized databases

→ Specialized databases		
Name	**Link**	**Description**
Ensembl	https://www.ensembl.org/index.html	Integrative genomic database
Prosite	http://prosite.expasy.org/	Lists protein motifs with biological significance
Reactome	https://reactome.org/PathwayBrowser/	Integrative metabolic database
Kegg Pathway	http://www.genome.jp/kegg/pathway.html	Molecular interactions and reactions
PFAM	http://xfam.org/	Protein domains
Interpro	http://www.ebi.ac.uk/interpro/	Combines several existing databases
PDB	http://www.rcsb.org/pdb/home/home.do	3D structures of proteins, amino acids, and biological molecules
PubMed	https://www.ncbi.nlm.nih.gov/pubmed	Citations, abstracts, and articles (bibliographic research)

II.1. Nucleotide databases

The data stored in these databases comes from DNA and RNA sequencing. Three well-known nucleotide databases share information and therefore contain almost identical sets of sequences. These three databases have been systematically exchanging their content since 1987 and have adopted a common convention system: "DDBJ/EMBL/GenBank":

- **EMBL database**: Created in 1980 and funded by the EMBO (European Molecular Biology Organization), developed within the European Molecular Biology Laboratory located in Heidelberg, Germany, and now distributed by EBI: http://www.ebi.ac.uk/embl/. As of February 24, 2014, the database contains 369.5 million sequences.
- **GenBank (Genetic Sequence Databank)**: Created in 1982 by IntelliGenetics and now distributed by NCBI (National Center for Biotechnology Information): http://www.ncbi.nlm.nih.gov/. In February 2014, the database contained 171,123,749 sequences. GenBank includes a protein sub-database, a translation of nucleotide sequences, called GenPept.
- **DDBJ (DNA Databank of Japan)**: Created in 1986 and distributed by NIG (National Institute of Genetics, Japan), it recorded a total of 81,994,905 DNA sequences as of December 2019 (DDBJ 2019).

II.2. Protein databases

The data stored in these databases are derived from the translation of DNA sequences or through protein sequencing (which is rare because it is time-consuming and expensive):

- **SwissProt database**: A protein database created in 1986 at the University of Geneva and maintained since 1987 as part of collaboration between this university (via ExPASy, Expert Protein Analysis System) and EBI. It also includes annotated sequences from the PIR-NBRF database as well as coding sequences translated from EMBL. As of February 2014, the database contains 542,503 sequences comprising 192,888,369 amino acids.

II.3. Structural databases

These are specialized databases for 2D and 3D protein structures. Several well-known databases exist in this context; here we cite the **PDB database** as an example:

- **PDB (Protein Data Bank)**: Created in 1971, it is the reference database for protein structures obtained experimentally by X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (the most recently used technique). The coordinates of the atoms forming the structure of a protein, the sequence details, and crystallization conditions are the main information available for each structure in the database.