


# Sequence databases



M.Sc 2<sup>ND</sup> SEMESTER  
PAPER 204: BIostatistic, COMPUTER  
APPLICATION AND BIOINFORMATICS  
UNIT 4: SEQUENCE DATABASES

# INTRODUCTION

- A sequence database is a collection of DNA or protein sequences with some extra relevant information. The main sequence databases are [Genbank](#) and [EMBL](#).
- Originally they were just sequence collections, but they have grown to store different biological databases heavily interconnected and they provide powerful interfaces to search and browse the stored information.
- The sequences submitted to any of those databases are shared between them, so any sequence could be retrieved in the european or the american database. But they differ in the tools to search and browse the data and in some databases that provide extra information to the raw sequences like: mutations, coded proteins, bibliographical references, etc.



# PIR

- PIR - Protein Information Resource
- 1984 - National Biomedical Research Foundation (NBRF) - US
- PIR-PSD (PIR- Protein Sequence Database)



# EMBL/GENBANK/DDJB

- These 3 db contain mainly the same information (few differences in the format and syntax)
- Serve as **archives** containing all sequences (single genes, ESTs, complete genomes, etc.) derived from:
  - Genome projects and sequencing centers
  - Individual scientists
  - Patent offices (i.e. USPTO, EPO)
- Non-confidential data are exchanged daily
- Currently:  $2.5 \times 10^7$  sequences, over  $3.2 \times 10^{10}$  bp;
- Sequences from  $> 50,000$  different species;



# GENBANK

- GenBank is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation, built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM).
- NCBI makes the GenBank data available at no cost over the Internet and via a wide range of web-based retrieval and analysis services which operate on the GenBank data.
- NCBI builds GenBank primarily from the submission of sequence data from authors and from the bulk submission of expressed sequence tag (EST), genome survey sequence (GSS), and other high-throughput data from sequencing centers.
- The US Office of Patents and Trademarks also contributes sequences from issued patents.



# SWISS-PROT

- Annotated protein sequence database established in 1986 and maintained collaboratively since 1987, by the Department of Medical Biochemistry of the University of Geneva and EBI
- Complete, Curated, Non-redundant and cross-referenced with 34 other databases
- Highly cross-referenced
- Available from a variety of servers and through sequence analysis software tools
- More than 8,000 different species
- First 20 species represent about 42% of all sequences in the database
- More than 1,29,000 entries with  $4.7 \times 10^{10}$  amino acids
- More than 6,22,000 entries in TrEMBL



# TREMBL (TRANSLATION OF EMBL)

- Computer-annotated supplement to SWISS-PROT, as it is impossible to cope with the flow of data...
- Well-structure SWISS-PROT-like resource
- Derived from automated EMBL CDS translation maintained at the EBI, UK.
- TrEMBL is automatically generated and annotated using software tools (incompatible with the SWISS-PROT in terms of quality)
- TrEMBL contains all what is **not yet** in SWISS-PROT

