Bioinformatics I
Biological Databases

Dr Manaf A Guma
University Of Anbar- college of Applied sciences-Hit
Department of applied chemistry

1

Biological Knowledge is Stored in Global Databases

- The most important basis for applied bioinformatics is the collection of sequence data and its associated to biological information.
- For example, with genome sequencing projects such data are generated daily in very large quantities worldwide.
- Furthermore, for a number of databases, original articles describe their functions.
- This database issue, which is freely accessible also on the Web, is a good starting point for working with biological data- bases.

2

What are the data bases that are used for bioinformatics ?

1. Primary databases contain primary sequence information (nucleotide or protein) and accompanying annotation information regarding function, bibliographies, cross references to other databases, and so forth.
2. Secondary biological databases, however, summarize the results from analyses of primary protein sequence databases.

- The aim of these analyses is to derive common features for sequence classes, which in turn can be used for the classification of unknown sequences (annotation).

3

What are the Primary Databases?

- **First: Nucleotide Sequence Databases.**
- **1- GenBank** https://www.ncbi.nlm.nih.gov/genbank/
- The GenBank database [genbank] is perhaps the best-known nucleotide sequence database available at the U.S. National Center for Biotechnology Information (NCBI) [ncbi].
- GenBank is a public sequence database, which in its present version contains roughly 199 million sequence entries (up to 2016). Sequences can be entered into GenBank by anyone via a Web page [bankit] or by e-mail [sequin] when working with larger sequence sets.
- It is associated with other databases, for example the European Nucleotide Archive (ENA) or the DNA Database of Japan (DDBJ).

4

What is the accession number (AN)? Key word for a gene

- Each single database entry is provided with a unique identification tag, the accession number (AN).
- The AN is a permanent record that remains unchanged even if changes are subsequently made to the database record.
- In some cases, a new AN can be assigned to an existing number if, for example, an author adds a new database record that combines existing sequences.

5

How to find the primary sequence of a gene using ncbi?

- Open the link: https://www.ncbi.nlm.nih.gov/genbank/
- Search for a specific gene Troponin C, for example.
- What do you find?
- describe the results..?

6

The fig shows a GenBank entry

GenBank: AJ419175.1

FASTA Graphics

Go to:

```
LOCUS       AJ419175                470 bp    mRNA    linear   INV 26-JUL-2016
DEFINITION  Ostertagia ostertagi partial mRNA for troponin (trp gene).
ACCESSION   AJ419175
VERSION     AJ419175.1
KEYWORDS    troponin; trp gene.
SOURCE      Ostertagia ostertagi
  ORGANISM  Ostertagia ostertagi
            Eukaryota; Metazoa; Ecdysozoa; Nematoda; Chromadorea; Strongylida;
            Trichostrongyloidea; Haemonchidae; Ostertagia.
REFERENCE   1
  AUTHORS   Geldhof,P., Vercauteren,I., Knox,D., Demaere,V., Van Zeveren,A.,
            Berx,G., Vercruysse,J. and Claerebout,E.
  TITLE     Protein disulphide isomerase of Ostertagia ostertagi: an
            excretory-secretory product of L4 and adult worms?
  JOURNAL   Int. J. Parasitol. 33 (2), 129-136 (2003)
   PUBMED   12633650
REFERENCE   2  (bases 1 to 470)
  AUTHORS   Geldhof,P.B.
  TITLE     Direct Submission
  JOURNAL   Submitted (05-NOV-2001) Geldhof P.B., Parasitology, Ghent
            University, Salisburylaan 133, Merelbeke, BELGIUM
FEATURES             Location/Qualifiers
     source          1..470
                     /organism="Ostertagia ostertagi"
                     /mol_type="mRNA"
                     /db_xref="taxon:6317"
                     /dev_stage="adult"
     gene            <1..>470
                     /gene="trp"
     CDS             <1..>470
                     /gene="trp"
                     /codon_start=2
                     /product="troponin"
                     /protein_id="CAD11862.1"
                     /db_xref="InterPro:IPR001978"
                     /db_xref="UniProtKB/TrEMBL:Q95PN9"
                     /translation="NFKINSKGEQAAQFGNLAQGVKQDGQTKEQQEEAKAAFLAAVCR
                     SVDISSLLPNDLKERIKTLHNRICKLEADKYDLEKRHERQEYDMKELHERQRQVARNK
                     ALKKGLDPEEAASSQHPPKITTASKFDRQIDRRSYGDRRELFEHPVIKKPPTIA"
ORIGIN
//
```

```
ORIGIN
        1 caatttcaag atcaattcca aaggcgagca ggcggcgcag ttcggcaatc tggcacaagg
       61 agtaaaacaa gatggacaaa cgaaagaaca gcaagaagaa gccaaggcag cgttcttggc
      121 agccgtttgc cgttcagtgg atatctcgtc gctgcttccg aacgatctga aggagcgaat
      181 caaaacgttg cataaccgaa tctgtaaatt ggaggccgat aagtatgatc tggagaagcg
      241 ccatgagcgt caggaatatg acatgaaaga gctgcacgaa cgtcaacgcc aagttgccag
      301 gaacaaggcg ctcaaaaagg gactcgatcc tgaggaagcc gcttcatctc aacatcctcc
      361 aaaaatcact accgcttcca agtttgatcg tcagattgac agaaggtctt atggagatcg
      421 acgagagctg tttgagcatc cagtcatcaa gaagccaccc accattgccc
//
```


Describe the results..?

- Each entry starts with the keyword LOCUS followed by a locus name.
- Like the AN, the locus name is also unique. Unlike the AN, it may change after revisions of the database.
- The locus name consists of eight characters, including the first letter of the genus and species names, in addition to a six-digit AN.
- A sequence must have at least 50 base pairs to be entered into GenBank.
- Every GenBank entry must contain coherent sequence information of a single molecule type, that is, an entry cannot contain sequence information of both genomic DNA and RNA.
- The last column in the LOCUS line gives the date of the last entry modification. The end of the database record starts with the keyword ORIGIN.

What are the Primary Databases?

- **2- Entrez:** http://www.ncbi.nlm.nih.gov/Entrez/.
- Query of the GenBank database is carried out via the NCBI Entrez system [entrez],
- [entrez] is used to query all NCBI-associated databases.
- Entrez is an important and effective tool for the execution of both simple and complicated searches for genes.
- To use this search, follow the link beneath the Entrez search field.

9

Field IDs to restrict research terms to certain database fields in the Entrez system ?

Field ID	Database field
ACC	Accession number
AU	Author name
DP	Publication date
GENES	Gene name
ORGN	Scientific and common name of the organism
PT	Publication type, e.g., review, letter, technical publication
TA	Journal name, official abbreviation, or ISSN number

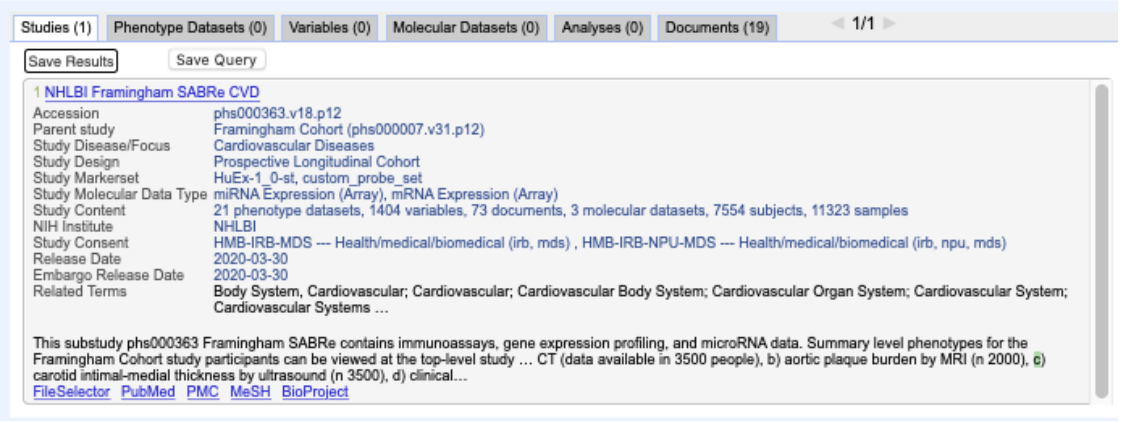
10

What are the Primary Databases?

- **3- EMBL and DDBJ** dbgap. http://www.ncbi.nlm.nih.gov/gap and ddbj. http://www.ddbj.nig.ac.jp/
- The European counterpart to GenBank is the ENA [ena], located at the European Bioinformatics Institute (EBI) [ebi].
- Another primary nucleotide sequence database, the DDBJ [ddbj], is operated by the National Institute of Genetics (NIG) [nig] in Japan and is the primary nucleotide sequence database for Asia.
- The three database opera- tors, NCBI, EBI, and NIG, compose the International Nucleotide Sequence Database Collaboration and synchronize their databases every 24 h.

11

The results of the DDBJ



While the database format of the DDBJ is identical to that of the NCBI, that of the ENA differs somewhat

12

What are the Primary Databases?

- **4- EMBL database:**
- https://www.ebi.ac.uk
- The most obvious difference is the use of two-letter codes instead of full keywords.
- Furthermore, there are small changes in the organization of the individual data fields.
- For example, the date of the last modification is not listed in the field ID (corresponding to the LOCUS field in GenBank) but appears in the field DT (database field).

13

What is the ENA web for?

- **5- ENA Online Retrieval https://www.ebi.ac.uk**
  The ENA offers several search forms.
- First is a simple search, which allows for text searches as well as for sequence retrieval .
- For text search, it is possible to search for accession numbers and for simple free text.
- ENA also allows for sequence searches using sequence comparisons.
- Basically, this is a BLAST search, which can either be carried out using standard BLAST parameters or which makes it possible to tweak BLAST parameters on the advanced search page.

14

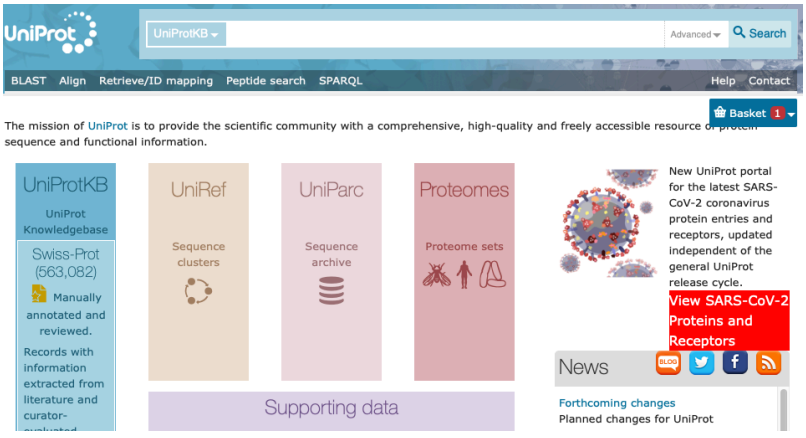
What are the Primary Databases?

- **6- UniProt** Protein Sequence Databases **https://www.uniprot.org** **(very important)**
- The information available for proteins continues to grow rapidly.
- UniProt consists of three parts, the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters Database (UniRef), and the UniProt Archive (UniPArc), a collection of protein sequences and their history.
- Besides sequence information, expression profiles can be examined, secondary structures predicted, and biological/biochemical function(s) analysed.
- The result is the Universal Protein Resource (UniProt) [uniprot], which unites the information in the three protein databases Swissprot, TrEMBL, and Protein Information Resource (PIR).
- Protein sequences and their annotations are stored in the UniProt Knowledgebase (UniProtKB), which is divided into two realms.

15

Theentry in the UniProtKB/SwissProt database website.

- At first glance the entry is similar to an ENA entry. Indeed, the two database formats are related.



16

What are the Primary Databases?

- **7- NCBI Protein Database https://www.ncbi.nlm.nih.gov (very important)**
- Another well-known protein sequence database is maintained at the NCBI. This data- base, however, is not a single database but a compilation of entries found in other protein sequence databases.
- For example, the NCBI database contains entries from Swissprot, the PIR database [pir], the Protein Data Bank (PDB) database [pdb], protein translations of the GenBank database, and several other sequence databases.
- Its format corresponds to that of GenBank, and queries are carried out analogously to those in GenBank via the Entrez system of NCBI.

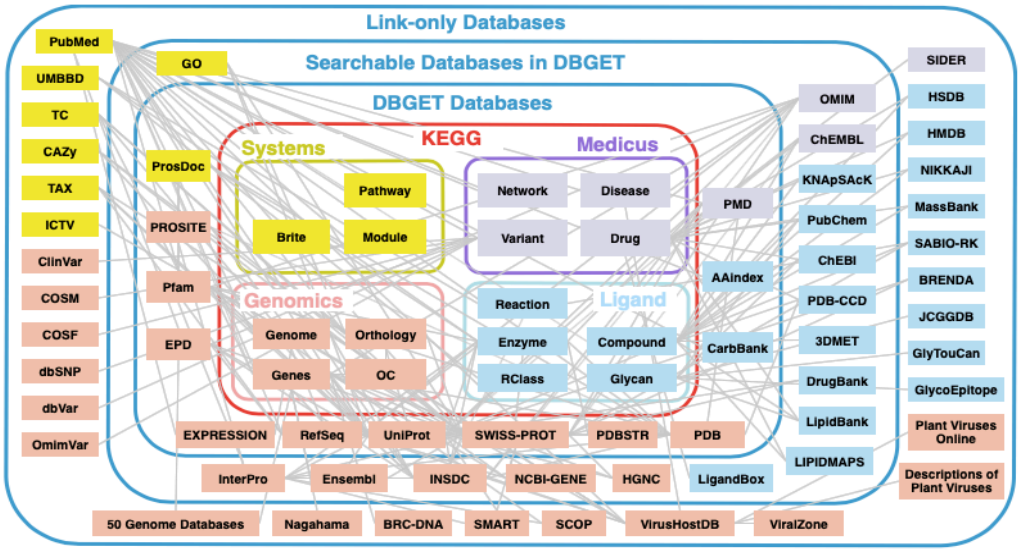
17

KEGG

- https://www.genome.jp/kegg/
- KEGG is a database resource for understanding high-level functions and utilities of the biological system.
- The biological system examples: the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.

18

Correlation Map



19

Exercise?

- Use the website **https://www.uniprot.org** to find the sequence of Tropomyosin TPM1. Download it a Fasta format.
- Use the website https://www.ncbi.nlm.nih.gov/protein/ to find the information about Tropomyosin TPM1. Download as much as you can and then discuses it.
- Use the website **https://www.ncbi.nlm.nih.gov/genbank/** to find information about Tropomyosin TPM1. Discuss it as you studied!

20