

Biological Databases and file formats



T: +27 51 401 9111 | E: info@ufs.ac.za | www.ufs.ac.za

 UFSUV |  UFSweb |  UFSweb |  ufsuv

*Inspiring excellence.
Transforming lives.*

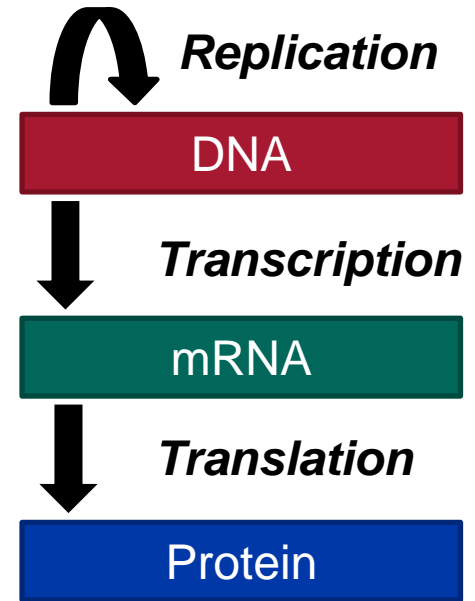
UNIVERSITY OF THE
FREE STATE
UNIVERSITEIT VAN DIE
VRYSTAAT
YUNIVESITHI YA
FREISTATA



UFS
INFORMATION AND
COMMUNICATION
TECHNOLOGY SERVICES
(ICT SERVICES)

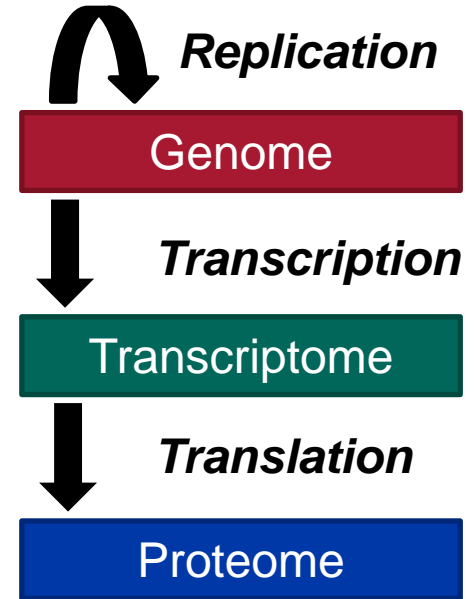
Central Dogma of Molecular Biology

- Technological limitations → single molecules
- Development of simple paradigm
- Last 20 years → Rapid technological development



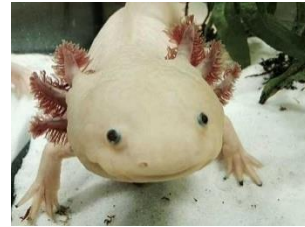
Central Dogma of Molecular Biology 2.0

- New Technologies
 - Next Generation Sequencing (NGS)
 - Advances in Mass Spectrometry
 - Advances in NMR technologies
- Single molecule → Entire complement of a cell's biomolecules
- -Omics → Birth of OMICS



Introduction – A problem of abundance

- Omics fields produce TONS of data
 - Mexican axolotl (*Ambystoma mexicanum*) (Salamander) genome
 - 32 giga-base-pairs (Gbp)
 - Human genome = 3272 Mbp
 - Genome + transcriptomics data = 8.5 GB (Gigabytes)
 - Largest genome sequenced thus far.





Databases

- Computer Science → Databases
- Shared, integrated computer structures that stores a collection of:
 - End-user data (raw facts – e.g. Biological Sequences)
 - Metadata (data about data) → used to integrate and manage end-user data. E.g. Accession Number

Biological Databases

- Databases that store biological data
 - Publicly available
- Initially created to store sequence data (Sanger sequencing – era)
- Evolved → Integrated and interactive platforms for lab scientists
- Many biological databases → specialized
 - Organisms
 - Diseases

Biological Databases - Examples

Name	Website	Description
GenBank (NCBI)	www.ncbi.nlm.nih.gov/genbank	International nucleotide sequence databases and repositories
ENA (EMBL-EBI)	www.ebi.ac.uk/ena	-
DDBJ	www.ddbj.nig.ac.jp	-
UniProt	www.uniprot.org	Protein database, sequence and functional annotation
Ensembl	www.ensembl.org	Vertebrate and eukaryotic genomes
Ensembl genomes	www.ensemblgenomes.org	Genome-scale data for bacteria, protists, fungi, plants and invertebrate metazoa
InterPro	www.ebi.ac.uk/interpro	Functional analysis of protein sequences
Pfam	pfam.xfam.org	Manually curated collection of protein domain families
RCSB PDB	www.rscb.org	3D structure data for large biological molecules

Biological Databases – How do they grow?

- Incentives for data to be submitted/deposited:
 - Required for journal publication – Data → Publicly available
 - Required by funding agencies (especially public funders)
 - Data-exchange between databases

Types of Biological Databases – Primary DB

- Usually repositories of experimental data
- Good example → Sequence databases
 - Genbank, ENA, DDBJ → nucleic acid sequences
 - Uniprot → Protein sequences
- PDB → Structural data (X-ray Crystallography / NMR)

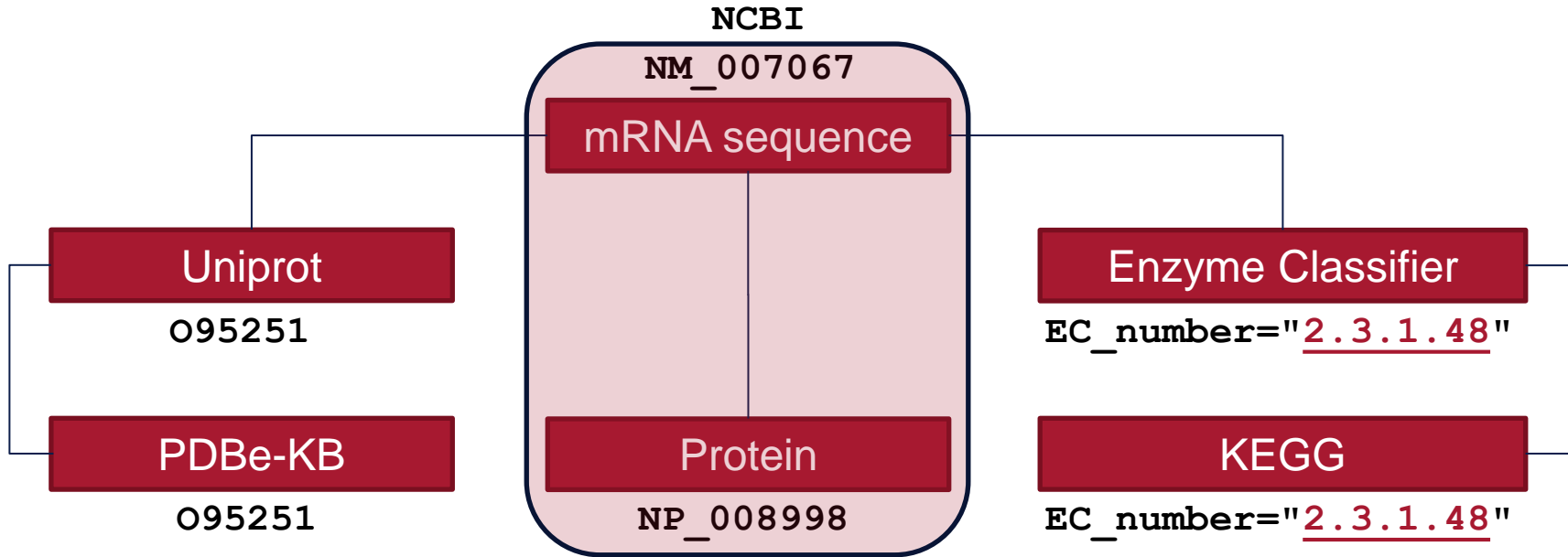
Types of Biological Databases – Secondary DB

- Layered on top of primary databases
- Provides additional information about records in the primary databases
- InterPro and Pfam → Classify protein structures into families based on structure/function

Types of Biological Databases – How are the connected

- Linked by accession numbers
 - Unique codes for records stored in a particular DB
- What would happen if the primary entry is corrupted / incorrect?

Example: *Homo sapiens* lysine acetyltransferase 7



Types of Biological Databases – Secondary DB

- Layered on top of primary databases
- Provides additional information about records in the primary databases
- InterPro and Pfam → Classify protein structures into families based on structure/function

Database Curation

- Accuracy of information in biological databases critical
 - Many downstream secondary databases rely on data to generate information used by lab scientists
- Curated in two ways:
 - Manually by humans → Slow but high quality
 - Automatically → Quick but less reliable
- Always double-check data entries manually in cases of unusual results

Biological Data Formats

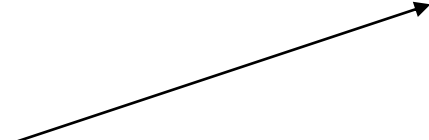
- Downloading data from biological databases
 - Specific file formats. Ex. Sequences in FASTA format
 - Designed to be read by computer software (i.e. predictable)
 - Usually text files → can be viewed in text editors
- Some important formats:
 - FASTA (sequence storage format)
 - GenBank (sequence and annotation format)
 - CLUSTAL (sequence alignment format)

FASTA File Format

- Common extensions → **.fasta** , **.fa** , **.faa**
- Single or Multiple sequences / records
- Records start with header:
 - “>” followed by a text description → can contain accession number
- Next line(s) contains sequence
 - 70 characters per line until the end of sequence
- For multiple sequence files:
 - Start next sequence with a new header line

FASTA File Format Example

- This is only one line! (Look at the line-numbers)
- Good editors wrap the lines for you for easier reading
- Thus it is a good idea to enable line numbers in your app of choice



```
1 >NC_045512.2:21563-25384 S [organism=Severe acute respiratory syndrome coronavirus 2] [GeneID=43740568]
   [chromosome=]
2 ATGTTTGTTCCTTGTTCCTTATGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACAACCAGAACTCAAT
3 TACCCCCTGCATACACTAATTCTTTTCACACGTGGTGTTCCTTACCCCTGACAAAGTTTTTCAGATCCCTCAGT
4 TTTACATTCAACTCAGGACTTGTTCCTTACCTTTCTTTTCCAATGTTACTTGGTTCATGCTATAACATGTC
5 TCTGGGACCAATGGTACTAAGAGGTTTGATAACCCTGTCCTACCATTTAATGATGGTGTTCCTTATTTGCTT
6 CCACTGAGAAAGTCTAACATAATAAGAGGCTGGATTTTGGTACTACTTTAGATTTCGAAGACCCAGTCCCT
7 ACTTATTGTTAATAACGCTACTAATGTTGTTATTTAAAGTCTGTGAATTTCAATTTTGTAATGATCCATTT
8 TTGGGTGTTTATTACCACAAAAACAACAAAAGTTGGATGGAAAGTGAGTTCAGAGTTTATTCTAGTGCGA
9 ATAATTGCACTTTTGAATATGTCTCTCAGCCTTTCTTATGGACCTTGAAGGAAAACAGGGTAATTTCAA
10 AAATCTTAGGGAATTTGTGTTTAAAGAATATTGATGGTATTTTAAAATATATTCTAAGCACACGCCTATT
11 AATTTAGTGCGTGATCTCCCTCAGGGTTTTTCGGCTTTAGAACCATTTGGTAGATTTGCCAATAGGTATTA
12 ACATCAGTACCTTTCAAACTTTACTTTCCTTACATACAACTTATTTTCACTCCCTCCCTCATTCTTCTTCAAC
```

GenBank File Format

- Common extensions → **.gb** , **.genbank**
- Contains annotation section and the sequence
 - Larger Files
 - More information than sequence
- Starts with the LOCUS keyword and record ends with “//” after sequence
- Goto <https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html> for an interactive exploration of a genbank record

GenBank File Format Example

```
1 LOCUS SCU49845 5028 bp DNA linear PLN 29-OCT-2018
2 DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p
3 (AXL2) and Rev7p (REV7) genes, complete cds.
4 ACCESSION U49845
5 VERSION U49845.1 GI:1293613
6 KEYWORDS .
7 SOURCE Saccharomyces cerevisiae (baker's yeast)
8 ORGANISM Saccharomyces cerevisiae
9 Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
10 Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
11 Saccharomyces.
12 REFERENCE 1 (bases 1 to 5028)
13 AUTHORS Roemer,T., Madden,K., Chang,J. and Snyder,M.
14 TITLE Selection of axial growth sites in yeast requires Axl2p, a novel
15 plasma membrane glycoprotein
16 JOURNAL Genes Dev. 10 (7), 777-793 (1996)
17 PUBMED 8846915
18 REFERENCE 2 (bases 1 to 5028)
19 AUTHORS Roemer,T.
20 TITLE Direct Submission
21 JOURNAL Submitted (22-FEB-1996) Biology, Yale University, New Haven, CT
22 06520, USA
23 FEATURES Location/Qualifiers
24 source 1..5028
25 /organism="Saccharomvces cerevisiae"
```


ClustalW File Format

- Common extensions → **.aln**
- For sequence alignments
- Format:
 - Start with “CLUSTALW” or “CLUSTAL W” (all other info is ignored)
 - One or more new lines (empty lines)
 - One or more blocks of sequence

ClustalW File Format – Sequence Block

- One line representing each sequence in alignment
- Formatted as follows:
 - Sequence Name
 - White space (usually a TAB)
 - Sequence symbols (max 60 per line). Gaps represented by “-”
 - Optional: White space followed by position of last sequence unit
- One line representing degree of conservation for columns in block
- One or more new lines

ClustalW File Format – Conservation Symbols

- * → All residues are identical
- : → Conserved substitutions observed
- . → Semi-conserved substitutions observed
- “ “ → No match

CLUSTALW File Format Example

```
1 CLUSTAL W (1.82) multiple sequence alignment
2
3
4 FOSB_MOUSE      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
5 FOSB_HUMAN      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
6      |           |
7      |           |
8 FOSB_MOUSE      ITTSQDLQWLVPPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS 120
9 FOSB_HUMAN      ITTSQDLQWLVPPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS 120
10     |           |
11     |           |
12 FOSB_MOUSE      GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEEKRRVRRERNKLAALKCRNRREL 180
13 FOSB_HUMAN      GGPSTSGTTS GPGPARPARARPRRPREETLTPEEEEEKRRVRRERNKLAALKCRNRREL 180
14     |           |
15     |           |
16 FOSB_MOUSE      DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVR 240
17 FOSB_HUMAN      DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVR 240
18     |           |
19     |           |
20 FOSB_MOUSE      LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY 300
21 FOSB_HUMAN      LPGSAPAKEDGFSWLLPPPPPPPLPFQTSQDAPPNLTASLFTHSEVQVLGDPFPVVNPSY 300
22     |           |
23     |           |
24 FOSB_MOUSE      TSSFVLTCPVSAFAGAQR TSGSEQPSDPLNSP L L L L 338
25 FOSB_HUMAN      TSSFVLTCPVSAFAGAQR TSGSDQPSDPLNSP L L L L 338
26     |           |
27     |           |
```

Take Home Message

- Get a good text editor
- Know your format of choice → will make understanding what software can do easier
- Using bioinformatics software → Read input section
 - Parsing / input errors common
 - Some software applications may be strict or permissive with regards to format adherence
- Format converters
 - Understand what information may be discarded
 - Example → Coverting GenBank to FASTA