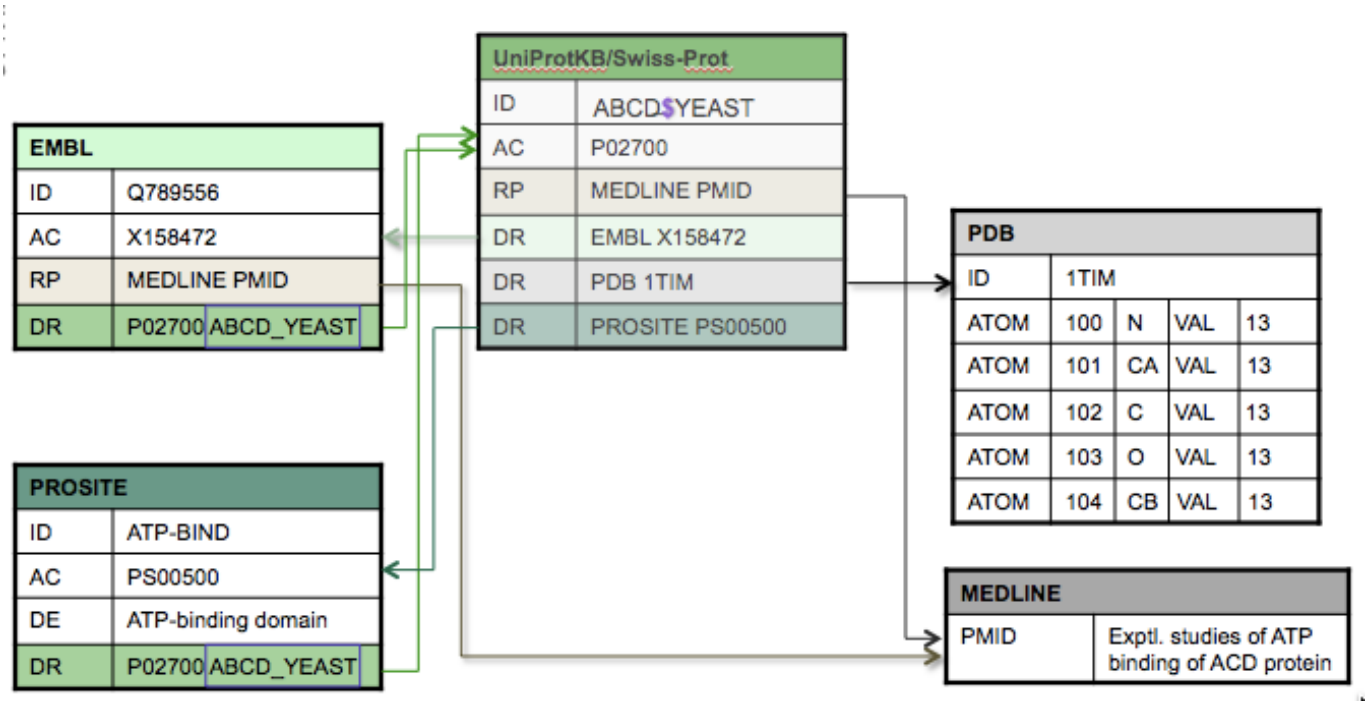
# BIOLOGICAL DATABASES


## Looking at flat-files



# Using & citing GOBLET materials


## Authors

Teresa K. Attwood

## Affiliation

The University of Manchester, the EMBnet Foundation, the GOBLET Foundation





## GOBLET Stichting

CMBI Radboud University
Nijmegen Medical Centre
Geert Grooteplein 26-28
6581 GB Nijmegen





## Funding

This material was produced with the support of the GOBLET and EMBnet Foundations

# BIOLOGICAL DATABASES


## Overview

This module introduces the need for biological databases, and examines features of a widely used **flat-file** format.

**Key terms**
**Flat file**: a plain-text file containing a number of data entries that lack structured inter-relationships
**Database record**: an entry in a database pertaining to a specific item of information

## Learning objectives

On completion of this module, given a UniProtKB/ Swiss-Prot flat-file, you will be able to:

- recognise key fields within the file
  - explain what they mean & why they're important
- retrieve specific information from different fields
  - deduce structural & functional properties of the sequence

## Introduction

Central to the discipline of bioinformatics is the need to store biological information systematically in structured databases. The first biological databases were simple formatted text or '**flat files**'. These were organised so that particular '**records**' within them could be easily identified and linked. However, as the complexity and types of available data have grown, database architectures and their user interfaces have become more sophisticated.

This module looks at the evolution of biological databases, exploring key features of a commonly used flat-file format, and showing how these facilitate linkage between disparate resources.

# BIOLOGICAL DATABASES


Complete **genome sequencing** was a major achievement. However, just amassing more data doesn't instantly make us more knowledgeable or provide miraculous understanding of the information we're collecting. Gaining biological and biomedical insights from raw genomic data is a complex process: *e.g.*, it requires:

- genes to be located and their structures to be properly assembled
- **coding regions** to be translated
- functions to be assigned to genes and their products
- disease associations to be identified, *etc*.

Online access to databases gave scientists the ability to use public data in their own research. Databases thus became invaluable as stores of biological information. But they are also important because they allow logical connections to be made to information in related resources via their **annotations**.

Annotations are the 'intelligence' or clues we attach to raw data to make them meaningful to, and reusable by, other researchers: linear strings of nucleotide bases or amino acid residues are virtually useless on their own, but allied with information about their evolutionary relationships, biological functions, roles, interactions, disease associations, *etc*., they become building-blocks of knowledge.

**Key terms**
**Genome sequencing**: the lab process of determining the complete DNA sequence of an organism's genome
**Coding region**: the region of an mRNA sequence that is translatable into a polypeptide
**Annotation**: notes added to database entries to make them useful and re-usable

# BIOLOGICAL DATABASES


## Flat-file databases

Database annotations add value to **raw data**, in principle allowing them to be reused quickly and conveniently. The more annotations provided, the richer the database content. The problem is, the more information added, the greater the need for disciplined approaches to data archiving – if computers are to be able to access particular annotations reliably, they must be stored in a structured way. This begs two questions:

- what kinds of annotation are vital?
- how should they be organised?

Clearly, for sequence data, adding notes about biological relationships, functions, *etc*., is useful. Database-specific details, like when the sequence was submitted and when the database entry was last updated, add further value; links to the literature (*e.g.*, to articles describing the function of a protein) and cross-references to information in related databases are also helpful.

Structuring such information systematically to facilitate computer access is challenging. Plain text is meaningful to humans, but not to computers; yet the earliest data-bases were created as flat-files. This required particular bits of information to be pin-pointed with specific **tags** to allow the data being stored there to be identified.

**Key terms**
**Raw data**: experimental data of any description that bears no annotation
**Database tag**: a code identifying the type of data that appears in a given database field (*e.g.*, AC identifies an entry's accession number)

# BIOLOGICAL DATABASES

Inevitably, several different flat-file formats evolved to store different types of biological data. Of these, one particular format became popular (because its structure was relatively simple) and was adapted for a variety of different data-types by a number of databases that are still in use today (*e.g.*, **UniProtKB/Swiss-Prot**, **UniProtKB/TrEMBL**, **PROSITE**). This simple flat-file format was the one originally devised to store nucleotide sequences in the **EMBL** data library.

**Key terms**
**UniProtKB/Swiss-Prot**: a manually curated protein sequence database
**UniProtKB/TrEMBL**: a computer-curated protein sequence database
**PROSITE**: a protein family database
**EMBL:** the core European nucleotide sequence database

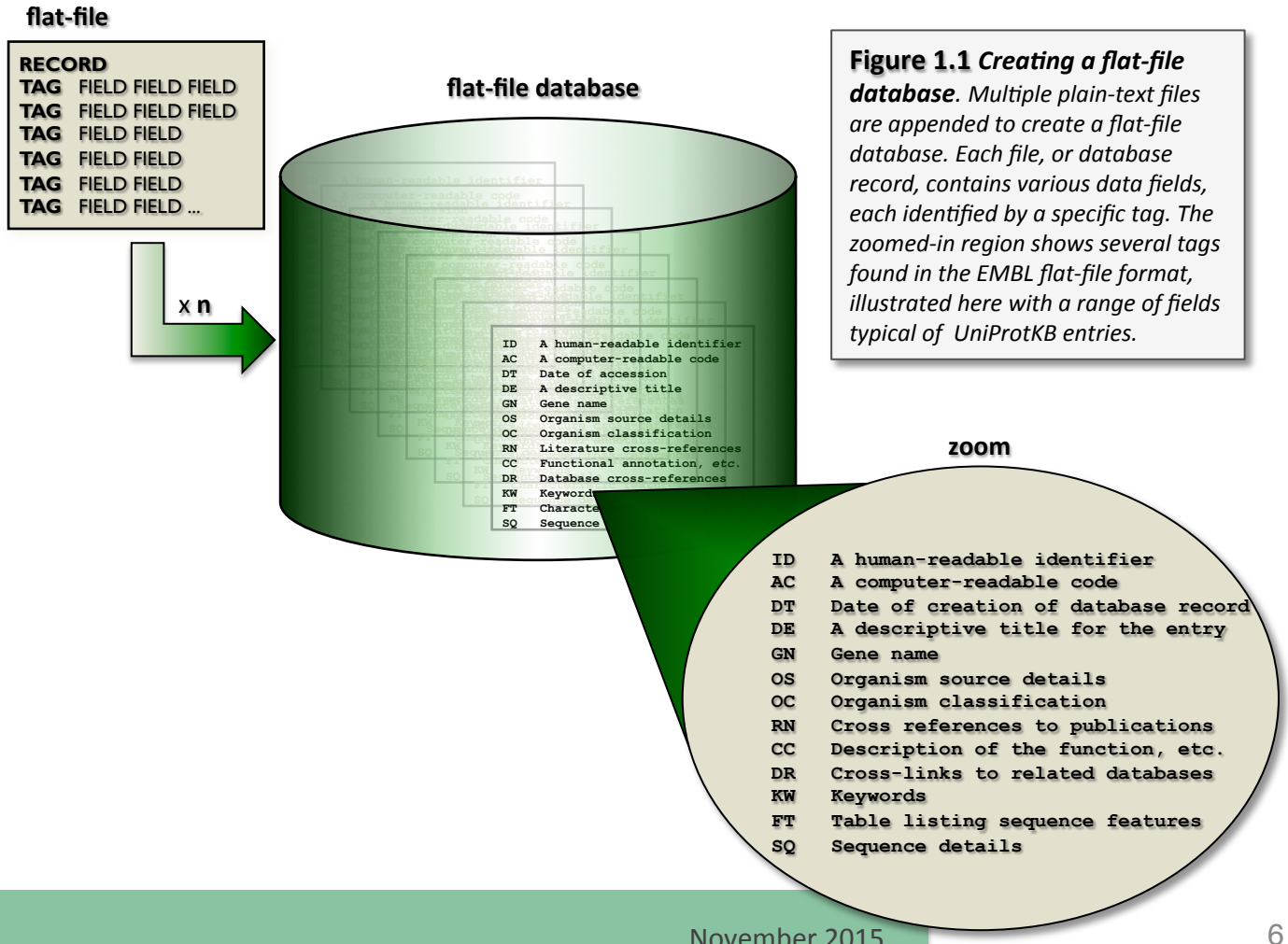



**Figure 1.1** ***Creating a flat-file database****. Multiple plain-text files are appended to create a flat-file database. Each file, or database record, contains various data fields, each identified by a specific tag. The zoomed-in region shows several tags found in the EMBL flat-file format, illustrated here with a range of fields typical of UniProtKB entries.*


## Flat-file fields & tags

The EMBL flat-file format uses a series of two-letter tags to describe the data stored on each line of the file, as shown in **Figure 1.2**.

The file begins with an identifying (ID) code and an accession (AC) number: the AC number (*e.g.*, P04156) is designed for computers to read; the ID code is more intelligible to humans – *e.g.*, here, PRIO_ HUMAN denotes the human **prion protein**. The AC and ID codes specify a given database entry. In principle, the AC number is invariant so that this sequence can always be tracked in any version of the database. Other important pieces of information include:

- the date a sequence entered the database, and when changes were last made to its entry – DT
- the description or title of the stored entity (here, the major protein precursor) – DE
- the source gene name (here, *prnp*) – GN
- a more precise specification of the organism species (here, *Homo sapiens*) – OS
- a more precise specification of the organism classification (Eukaryota, Metazoa, Chordata, *etc.*) – OC

In addition, the file includes bibliographic citations:

- RN is the reference number
- RP gives the subject
- RM the literature database (**PubMed**) cross-reference
- RA the authors
- RL the place of publication

**Key terms**

**Prion protein:** a membrane protein encoded in the mammalian genome; in its misfolded state, it aggregates in the brain and neural tissues, leading to neurodegenerative diseases

**PubMed:** the online interface to the MEDLINE biomedical literature database, which includes citations from life science journals, books, *etc.*

**Further reading**

More information about flat-file database formats can be found in Chapter 3 of *Introduction to Bioinformatics* (Attwood and Parry-Smith, 1999), Prentice Hall.

# BIOLOGICAL DATABASES


```
ID   PRIO_HUMAN     STANDARD;      PRT;   253 AA.
AC   P04156;
DT   01-NOV-1986 (REL. 03, CREATED)
DT   01-NOV-1986 (REL. 03, LAST SEQUENCE UPDATE)
DT   01-NOV-1991 (REL. 20, LAST ANNOTATION UPDATE)
DE   MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33-35C)
GN   PRNP.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE FROM N.A.
RM   86300093
RA   KRETZSCHMAR H.A., STOWRING L.E., WESTAWAY D., STUBBLEBINE W.H.,
RA   PRUSINER S.B., DEARMOND S.J.;
RL   DNA 5:315-324(1986).
…
CC   -!- FUNCTION: THE FUNCTION OF PRP IS NOT KNOWN. PRP IS ENCODED IN THE
CC       HOST GENOME AND IS EXPRESSED BOTH IN NORMAL AND INFECTED CELLS.
CC   -!- DISEASE: PRP IS FOUND IN HIGH QUANTITY IN THE BRAIN OF HUMANS AND
CC       ANIMALS INFECTED WITH THE DEGENERATIVE NEUROLOGICAL DISEASES KURU,
CC       CREUTZFELDT-JACOB DISEASE (CJD), GERSTMANN-STRAUSSLER SYNDROME
CC       (GSS), SCRAPIE, BOVINE SPONGIFORM ENCEPHALOPATHY (BSE), ETC.
CC   -!- SUBUNIT: PRP HAS A TENDENCY TO AGGREGATE YIELDING POLYMERS CALLED
CC       "RODS".
CC   -!- PRP CONTAINS 5 TANDEM REPEATS OF AN OCTAPEPTIDE P-H-G-G-G-W-G-Q.
CC   -!- SUBCELLULAR LOCATION: PRP IS ATTACHED TO THE EXTRACELLULAR SIDE OF
CC       THE CELL MEMBRANE BY A GPI-ANCHOR.
DR   EMBL; M13667; HSPRP0A.
DR   EMBL; M13899; HSPRP.
DR   PIR; A05017; A05017.
DR   PIR; A24173; A24173.
DR   MIM; 176640; NINTH EDITION.
…
DR   PROSITE; PS00291; PRION.
KW   PRION; BRAIN; GLYCOPROTEIN; GPI-ANCHOR; TANDEM REPEAT; SIGNAL.
FT   SIGNAL        1     22
FT   CHAIN        23    253       PRION PROTEIN.
FT   DOMAIN       90    234       PRP27-30 (PROTEASE RESISTANT CORE).
…
SQ   SEQUENCE   253 AA;  27661 MW;  354945 CN;
         MANLGCWMLV LFVATWSDLG LCKKRPKPGG WNTGGSRYPG QGSPGGNRYP PQGGGGWGQP
     HGGGWGQPHG GGWGQPHGGG WGQPHGGGWG QGGGTHSQWN KPSKPKTNMK HMAGAAAAGA
     VVGGLGGYML GSAMSRPIIH FGSDYEDRYY RENMHRYPNQ VYYRPMDEYS NQNNFVHDCV
     NITIKQHTVT TTTKGENFTE TDVKMMERVV EQMCITQYER ESQAYYQRGS SMVLFSSPPV
     ILLISFLIFL IVG
//
```

**Figure 1.2** ***Key features of the EMBL flat-file format.*** *This excerpt from the UniProtKB/Swiss-Prot entry for the human prion protein shows a range of tags, such as those of the identifier (ID), accession number (AC), protein description (DE), etc.*

**Further reading**

Additional information on the structure of UniProtKB entries can be found in the manual: www.uniprot.org/help/?query=*&fil=section:manual

# BIOLOGICAL DATABASES


A rich set of annotations can be found in the comment (CC) field. In this example, we learn about the protein's structure and function, tissue-specificity, disease associations, and so on. To facilitate swift computational processing of the file, many of the terms used here are also included as keywords (KW). Links to related information in other databases are made in the DR lines (including, for example, EMBL, **MIM** and PROSITE).

Further enriching the entry, various characteristics of the sequence itself are documented in the Feature Table (FT): in this example, we discover the locations of several octapeptide repeats**,** of potential carbohy-drate attachment sites, and of sequence variations.

Finally, the **single-letter code** is used to store the amino acid sequence itself in the SQ field, together with attributes such as its length and molecular weight. The entry terminates with the // symbol.

Databases aren't useful unless computers can access their data and help humans to interrogate and analyse them. Doing this requires adherence to standard data formats.

Adherence to a common format, regulating the content, and vocabulary and syntax of documented features, improves data consistency and reliability, and facilitates computer access and database interoperation.

The principal means by which computers access database entries is via their AC numbers and ID codes. This allows data from very different resources to be connected (whether they contain protein sequences, families, structures, literature, *etc*.). The more internal cross-references a database stores, the greater the web of connectivity that's possible from it.

**Key terms**

**MIM or OMIM (Online Mendelian Inheritance in Man)**: a comprehensive database of human genes and genetic disorders.

**Single-letter code**: individual letters used to denote the amino acids – Q for glutamine, W for tryptophan, *etc*.

**Further reading**

Higgs, P. & Attwood, T. (2005) Bioinformatics and Molecular Evolution, Wiley-Blackwell. See Chapter 5 for more details of biological databases.

# BIOLOGICAL DATABASES

**Activity**

The history of the amino acid sequence of the human prion protein since its first appearance in Swiss-Prot in 1988 can be found here:
www.uniprot.org/uniprot/P04156?version=*

Scroll down and click on the first version (1.txt).

1. How many octapeptide repeats are there?
2. What are their locations?
3. With what diseases is the protein associated?
4. How many amino acid residues are there in the protease-resistant core?
5. What is the role of the octapeptide repeats (hint: you may need to explore more recent versions to answer this)?

**Take homes**

In this module, we saw that:

- Sequence information was originally stored in 'flat-files' (essentially plain-text files)
- The EMBL flat-file format was adopted by different databases because its structure made it easy to use and to adapt
- The EMBL format uses a series of two-letter tags to denote different database fields (ID and AC tags for the entry identifier and accession number, DE for the descriptive title, CC for comments, FT for the Feature Table, *etc*.)
- Tags allow the data in different parts of flat-file databases to be cross-linked to information in related databases
- The origins and evolution of all sequences in UniProtKB can be accessed via the history tab

# GOBLET

## Global Organisation for Bioinformatics Learning, Education & Training

### About GOBLET

GOBLET is a not-for-profit Foundation, legally registered in the Netherlands. It's mission is to

- *provide a global, sustainable support and networking structure for bioinformatics trainers and trainees (including a training portal for sharing materials, guidelines and best practice documents, train-the-trainers facilities, etc.)*
- *facilitate capacity development in bioinformatics in all countries*
- *develop standards and guidelines for bioinformatics education and training*
- *act as a hub for fund gathering*
- *reach out to, amongst others, high-school teachers, to bridge the gap to the next generation of bioinformaticians, and*
- *foster the international community of bioinformatics, biocuration, biocomputing and computational biology trainers*

Its ethos embraces:

- *Inclusivity*
- *sharing*
- *openness*
- *innovation and*
- *tolerance*

For further information about GOBLET's work, visit www.mygoblet.org

For general enquiries about GOBLET, contact us at info@mygoblet.org