Power Point 01 - Bioinformatics Topics - Part 02 - Databases

And so to databases! Their creation and use.

<Click>

First Raw Data is generated.

Then what can be discovered is revealed by various forms of analysis.

Next comes the time to put Data together with its Interpretation. The process of "Annotation".

Data, properly associated with its Annotation forms "Information". Truly a case where, as that fine fellow Aristotle might have put it, possibly as he ran a bath for his dear friend Archimedes no doubt:

> "The Whole is truly Greater than the Sum of its Parts".

<Click>

Here it is the process of assembling Annotated Data, that is Information, into freely accessible Databases that will be considered.

<Click>

Commencing with the earliest viable DNA Sequence Databases of the early 1980s.

At this time, Sequencing was relatively low volume. All sequences were deemed precious.

Annotation was left to the submitter, and in consequence, inconsistent and often of low quality.

It seems almost anything that could be published was accepted.

These databases were not curated and some very strange entries remained in place for much longer than was perhaps ideal.

<Click>

The three earliest DNA Sequence Databases were:

The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database.

The GenBank Sequence Database – from America.

The DNA Data Bank of Japan

<Click>

Later in the 1980s, these three databases merged to the extent that data submitted to any of the three would be passed on to the other two.

<Click>

This act of co-operation was celebrated with the impressive title “International Nucleotide Sequence Database Collaboration” (or INSDC to its close friends).

<Click>

The equivalent story for Primary Protein Sequence Databases begins with the Atlas of Protein Sequence and Structure.

This was edited from 1965 to 1978 by Margaret Dayhoff, who contributed so much to this field.

<Click>

The Protein Information Resource (PIR) grew out of Dayhoff's work in 1982.

<Click>

Swiss-Prot was created in 1986 by Amos Bairoch.

The aim was to produce a comprehensive collection of carefully annotated and reliable protein sequences.

<Click>

The least achievable of Swiss-Prot's objectives was “comprehensive”.

It became ever more easy to produce more protein sequences than Amos and all his friends could possibly annotated to an acceptable standard.

And so … TrEMBL (and, a very similar American Database called GenPept).

Never mind the quality, just pick out anything in any of the Primary DNA Sequence Databases that claims it codes for protein. Translate it and annotate it as well as it is possible from what is known of the coding DNA sequence.

There are, as you might imagine, some pretty questionable entries in TrEMBL!

<Click>

BUT … with TrEMBL ones has close to a complete collection of all available protein sequences.

In 2002, PIR, Swiss-Prot and TrEMBL were merged together to form UniProtKB.

The KB stands for Knowledge Base (as opposed to Database). This to emphasise the importance of the extensive annotation that enriches the raw protein sequences.

The quality of annotation varies very significantly between entries originating in PIR or Swiss-Prot compared to those from TrEMBL.

Accordingly, UniProtKB is divided into two sections:

UniProtKB/Swiss-Prot - High quality, reviewed annotation
UniProtKB/TrEMBL – unreviewed computer generated annotation

UniProtKB/TrEMBL are constantly being reviewed and either reject or promoted to UniProtKB/Swiss-Prot. However, even more unreviewed sequences are ever being generated.

UniProtKB/TrEMBL is by far the larger section, and despite the worthy efforts of very many annotators, is likely to remain so for the foreseeable future.

<Click>

By the first years of this century, sequencing had become much cheaper and quicker.

The volumes of sequence now available, including many entire genomes, was such that more organised Databases, maximising quality of data and annotation and minimising duplication, were in demand.

These newer databases were largely derived from existing data in Primary Databases.

<Click>

The RefSeq Database, first released in 2003, is a good illustrative example of such a database.

RefSeq is derived from the older INSDC DNA sequence databases, primarily GenBank (as RefSeq is made by the same Institution).

RefSeq aspires to be non-redundant and well annotated. Not something that can be universally claimed for the INSDC databases.

RefSeq includes DNA, RNA and Protein sequence entries from many a wide range of organisms.

<Click>

It has been nicely put that the relationship between GenBank and RefSeq in analogous to that between individual Research Papers and retrospective Reviews of a topic.

The Research Literature will include everything published on a topic, independent on quality or veracity.

A Review of the Literature can ignore papers that have been proven to be inaccurate and can merge overlapping messages where appropriate. Not to mention the opportunities to apply evaluations only possible with the advantages of hindsight!

<Click>

A number of databases have been derived from Protein Sequence databases, whose entries represent Protein Domains and Motifs.

Generally, sets of sequences representing homologous proteins, sharing at least one domain or motif, are compiled using database searching.

The relevant regions of each sequence set are carefully aligned and a suitable model, almost invariably an HMM these days, is computed to represent each shared feature.

Each collection of domain or motif models can then be used to investigate other protein sequences. Each database model being compared along each query sequence in turn.
Where there is a significant match, there exists a potential domain of motif.

<Click>

Available, quality Domain or Motif databases currently available include those illustrated.

Each Database has its own Website and its own tailored search software.

None is difficult to use, but each must be search individually, which can be tedious when more than one judgement on a proteins properties is deemed necessary.

Follow the embedded links for more information on each.

<Click>

The various Protein Feature Databases are of very similar purpose.

In particular, the way they represent Features has converged over the years.

Nevertheless, they still vary significantly in detail of purpose, the type of feature for which they are optimised and in other ways.

In consequence, they will often not produce exactly the same answers to a given query.

Occasionally, one service will entirely miss a feature that other services detect! None of these databases are perfect.

For the best results, it is generally wise to use ALL appropriate Feature searching services for every query.

<Click>

Knowing which databases to choose for each circumstance and visiting each Website one at a time is not really practical.

<Click>

InterPro is a resource that allows easy access to the correct combination of Feature Searching services for any occasion.

InterPro does not have its own Domain or Motif Models.

Instead, for each potential Feature, it selects appropriate searches from a wide range of options (including all those mentioned in the previous slide), and runs them.

If sufficiently encouraging results are obtained, it reports a potential Feature site.

<Click>

As the sequences of more and more entire genomes became available, databases storing one or

more entire genome were an obvious “next step”.

A considerable number of these databases now exist, many specific to a particular organism, or related collection of organisms.

The near completion of the Human Genome, around the turn of the century, led to a considerable acceleration of interest in this sort of database.

<Click>

Each newly sequence genome is analysed and annotated from scratch.

Significant re-assemblies of existing genomes are also re-analysed.

<Click>

The vast majority of analyses used to interpret whole genomes are just those that would be used to analyse an individual gene.

Many have been touched upon in the course of this simple talk.

For example, locating the genes is central to understanding a genome.
What could be more effective for this purpose that searching the genome sequnce for near perfect matches with every available mRNA sequence fro the same (or a very similar) organism?

Pairwise Sequence Comparison as part of a Sequence Database search?

<Click>

Analysing a small region of DNA, a single gene say, could be undertaken by an individual researcher and conducted manually.

To apply the same analysis to all but the very smallest genomes, requires carefully constructed software pipelines to allow an enormous volume of computation to take place with minimal human intervention.

<Click>

The three Genome databases in most common general use are:

Ensembl, maintained in Europe
Map Viewer, maintained in America
The University of California, Santa Cruz (UCSC) Genome Browser

All three are not specific to any particular organism, or family of organisms.

All offer a consistent User Interface to a very large number of genomes.

For the Human Genome, all use the same assemblies of the genome sequence (raw data, that is).
Ensembl and Map Viewer analyse the assemblies independently.
The UCC offers an alternative (and very popular) interface to the Map Viewer Annotation.

Encouragingly, the differences, in most important respects, between the Ensembl and MapViewer interpretations of the Human Genome grow fewer with each reassembly.

<Click>

The software, for both Ensembl and the UCSC Genome Browser, can be downloaded and used to provided a front end to personal datasets.

<Click>

The primary Database for Protein Structures is The Protein Data Bank (PDB)

<Click>

Access to the same data collection, with different emphases and tools, is offered by other sites across the World, especially from:

The Research Collaboration for Structural Bioinformatics Protein Data Bank (RCSB PDB)

The Protein Data Bank Japan (PDBj)

The Protein Data Bank Europe (PDBe)

<Click>

Two databases attempting to classify the structure of the PDB database were established in the middle 1990s. They are:

The Structural Classification of Proteins (SCOP)
and
The CATH Protein Structure Classification.

Both have, broadly, the objective of identify sets of proteins that are assumed distantly homologous on the evidence of shared structure, rather than shared function or easily detectable sequence conservation.

How this objective is achieved is where these two databases differ.

In passing, and because I hate acronyms that have no explanation, CATH stands for Class, Architecture, Topology, and Homology. These are the 4 levels (in order) used to classify proteins by the CATH system.

<Click>

Both SCOP and CATH offer Domain databases based on HMMs.

SCOP classifications are represented in a database called Superfamily.

CATH classifications are represented in a database called Gene3D.

Both of these databases define very general domain relationships.

For example, one Superfamily classification is likely to encompass several domain families of Pfam, or any other of the databases derived from sequence conservation (that is MSAs).

Both Superfamily and Gene3d are included in the Interpro Consortium.

For more detail, wait with baited breath for the module of your course dedicated to this topic.

<Click>

There are a number of databases available that represent the way individuals and species vary from each other.

<Click>

Many are incorporated in Genome databases, making it possible for a researcher to easily discover what common variations have been discovered around any given region.

<Click>

Since is has been possible to sequence in such high volumes and reasonable costs, the number of individual genomes available has escalated enormously.

Clearly, it becomes more possible to discover variations as the number of fully sequenced genomes rises.

The size, scope and importance of Genetic Variation Databases has grown accordingly.

<Click>

The most famous and all encompassing Variation Database has to be “The Single Nucleotide Polymorphism Database (dbSNP)”, founded in 1998 in America.

Originally, dbSNP was essentially a database of Human Single Nucleotide Polymorphisms. Hence the name.

Now dbSNP includes other type of short genetic variations such as InDels and microsatellite repeats. To reflect this broadening of content, the name of the database was changed to “The database of Short Genetic Variation”. Happily, the acronym dbSNP was retained to avoid simple folk getting confused.

<Click>

DbSNP is also no longer so focussed on Human data. It now includes information for a wide range of organisms. Variations between species as well as between individuals are covered.

Originally, dbSNP was intended as a tool primarily for research into population genetics. “Interesting” variations were those that were sustained in a population and reasonably common. Increasingly now, recording the relationships between variation and phenotype has become an important role for dbSNP.

<Click>

Other relevant databases include those storing Data generated from Microarray experiments.

<Click>

A number of these exist, many of which are commercial.

<Click>
The two most used Public Domain Microarray Database are The Gene Expression Omnibus (GEO) based in America, and the European Database ArrayExpress.

<Click>

Both GEO and ArrayExpress initially stored only Microarray data.

Relatively recently, High Throughput Sequencing (HTS) has begun to take over from the use of Microarrays.

Accordingly both databases now also manage HTS data sets.

<Click>

The two databases work co-operatively. Specifically, ArrayExpress regularly imports data from GEO.

<Click>

There are a number of Science orientated Literature resources available via the INTERNET.

They must have only a mention here.

They will will be covered properly in another part of your course.

<Click>

The Gene Ontology (GO) Project set out to provide a means to unambiguously describe genes and there products and so enable effective Searching of Databases by Keyword search.

<Click>

It has already been noted above that sequence annotation, specially in the older Databases, is disorganised and inconsistent.

<Click>

Annotation was often entirely the responsibility of the submitter.

Once accepted into a database, sequence annotation was not curated, so poor annotation remained poor annotation.

<Click>

Database Searching by comparing Keywords to Annotation was, in consequence, far from reliable.

<Click>

The Gene Ontology (GO) Project provides a hierarchy of universally accepted terms to describe gene products accurately and unambiguously.

<Click>

Searching with GO terms is by far the most effective way to search Sequence Databases accurately.

A more complete look at how the Gene Ontology works, is well worth the effort, but beyond this simple talk or you current training programme.

<Click>

Finally, just consider the "Biology" topics for a moment.

<Click>

Spread them out a trifle.

<Click>

Add an extra topic "Data Annotation" and split Data & Information Storage/ & Access into two separate processes.

By "Data Annotation", I mean combining the information obtained by analysis with the appropriate experimental data ready for submission to a suitable Database.

<Click>

Now add an ordering and one has a rather simplistic, but hopefully useful, representation of how all the process discussed link to together.

Particularly how the end of the chain (that is information from Databases) can be fed back into the start of the system to form a loop.