Power Point 001 - Bioinformatics Topics

The purpose of this talk is to consider the major components of Bioinformatics.

The intention here is not to be comprehensive, but to suggest a frame work in which to fit the topics covered in your course.

<Click>

Here, I have divided the Bioinformatic topics to be considered into those that are primarily "informatics" issues

<Click>

and those that are directly concerned with the management of Biological data.

<Click>

The most fundamental Informatics issue is the Operating system.

One must be able to control the computer before one can expect it to do anything useful!

<Click>

Currently, and many would argue unfortunately, Windows is still the most common operating system, followed ever more closely by the Macintosh Operating system.

Both Windows and the Macintosh offer very sophisticated and intuitive Graphical User Interfaces (or GUIs) with which most people under 35 have been familiar since birth!

<Click>

The Linux operating system is becoming increasingly popular.

These days, Linux comes with very good graphical user Interfaces that, distressingly to some, ever more closely mimic Windows.

Linux, with a good Graphical User Interface should seem very familiar to all of you.

I suggest it is safe to assume familiarity with any of these systems with a good GUI.

<Click>

The real issue comes when it is necessary to use Linux from the command line.

That is to say ... no graphics, no mouse, just a keyboard!

This can be a bit of a culture shock for some, but it really is not as difficult as it at first seems.

Especially, after a gentle introduction and a bit of practice such as will be delivered as part of your current training programme.

<Click>

If you find command line Linux a struggle, it should be comforting to appreciate that it is conceptually identical to any GUI based Operating System.

Any conversation you might have with any operating system must exclusively concern the management of files within a hierarchy of directories (or folders) or which program you want to run next.

You can do this using a fancy GUI, or by typing instructions with a keyboard.

It is just the same.

<Click>

Command line Linux really is the only option for some tasks in Bioinformatics.

Specifically, it would be crazy to write programs to process high volume sequence datasets for a clumsy bloated OS such as Windows!

Compute and memory intensive software, required to run in a liveable time period, are Linux only.

Generally, it is not considered a worthwhile use of programming resources to provide friendly interfaces for such software.

Its command line or nothing I fear.

<Click>

A basic understanding of the principles of programming has become ever more important in recent years.

<Click>

It is generally not necessary for a Bioinformatics "user" to become a proficient programmer as the need for an individual researcher to write software from scratch is rare.

<Click>

However, it is of great benefit for many users to be able to construct simple scripts (or programs) to manage the large datasets now so common.

<Click>

Also, the capacity to write scripts to customise the order and fashion in which a series of programs might be invoked (that is, to construct simple pipelines) is commonly of value.

<Click>

Currently, the most popular programming language for Bioinformatics is Python.

The basics of Python can be mastered by most in just a few days of training and a little practice.

Learning Python is outside the scope of this course, but would be well worth the effort at some future date.

<Click>

Minimally, an understanding of programming that would enable a user to construct small programs from scratch, understand and adapt slightly larger programs and communicate meaningfully with a specialist programmer would be a reasonable objective.

<Click>

As more biological research becomes focused around the analysis of huge datasets, a good grasp of Statistics becomes ever more essential.

Like programming, this is beyond the remit of your current training program, but also might be worth some effort at a later date.

<Click>

Statistical evalution of an experiment is vital from the design stage.

It is too easy to invest resources in experiments that have no hope of generating results that are statistically meaningful.

It is always a very good plan to take careful note of the oft quoted advice of Sir Ronald Fisher, first offered way back in 1938!

<Click>

Having conducted a well designed experiment, generating vast quantities of raw data, Statistical analysis, implemented with quality Statistical software (that is a Statistical Package), is unavoidable.

The Statistical Package R is widely used in Bioinformatics. It is comprehensive, excellent quality, includes libraries specific to Bioinformatics ... and it is free!

<Click>

Effectively all evaluations generated by Bioinformatics software are statistical judgements.

That is, the selection of the most probable answer from a set of many possible answers to a given question.

This point will be reinforced as we consider various individual analytical possibilities.

<Click>

The obvious first "Biology" topics has to be Data Generation. Not much Bioinformatics can happen before there is something to make it happen upon!!

<Click>

Types of data generated directly by Biological experiments include:

- Sequences (from small RNA sequences to entire Genomes)
... Commonly, this will involve Next-Generation
(or High Throughput) Sequencing (that is NGS) technologies.

Recent technological advances have radically changed the nature of DNA sequencing.

Older techniques, such as Sanger Sequencing were relatively slow, low volume and expensive.
Best suited to small scale, single gene projects.

NGS technologies have enabled high volume sequencing to be both affordable and widely available.

With greatly expanded sequence data volumes, the possible application of sequencing experiments have broadened enormously.

<Click>

- 3D Protein Structures ... from either X-ray crystallography
or Nuclear Magnetic Resonance (NMR) experiments.

3D protein Structures will be considered in one of the modules of this training course.

<Click>

- Gene Expression Data ... from Micro-Array experiments.

Arguably, the heyday of micro-arrays is behind us.
However, they do still have a significant role to play in Bioinformatics and there are extensive databases of very useful micro-array data available to users.

Recently, it has become quite common to use NGS experiments to investigate Gene Expression, rather than micro-arrays.

<Click>

In the context of this talk, Data Analysis is the process by which meaning and interpretation are added to raw experimental data.

There are a very wide variety of analyses that can be applied to experimental data.
Many involve the use of computer programs and thus can be considered to be Bioinformatics.

To try and make a comprehensive list and describe all possible analyses, would not really be too useful.
I suggest a concentration on those analyses that will be covered in your course.

<Click>

Starting with the comparison of pairs of DNA or Protein sequences representing "Homologous" entities(that is entities that are evolved from a common ancestor).

The software will align ANY two sequences, however, the alignment algorithms do assume Homology!

As the purpose is, generally, to investigate the effects of evolution by examining sequences, there can rarely, if ever, be purpose in trying to align unrelated sequences.

<Click>

Pairwise Sequence analysis is the operation underlying the vast majority of sequence analysis by computer programs.

<Click>

Pairwise Sequence analysis starts then with 2 homologous sequences

<Click>

and a putative acenstor sequence from which those sequences may have evolved

<Click>

The software computes the most probable alignment, matching residues of the homologous sequences with those of the putative ancestor sequence.

Where required, padding characters, here I use minus signs, are introduced to achieve full correspondence between sequences.

Biologically, the padding characters represent insertions or deletions between the sequences being aligned and their putative ancestor.

<Click>

The proposed alignment represents the most probable combination of Substitutions,

<Click>

Deletions (relative to the Ancestor Sequence),

<Click>

Insertions (relative to the Ancestor Sequence),

<Click>

and fully Conserved regions,

assuming some given set of statistical assumptions particular to the sequences being aligned.

<Click>

Multiple Sequence Alignment (or MSA) is the logical expansion Pairwise Alignment to enable the

alignment of families of homologous sequences of (theoretically at least) unlimited size.

The best algorithm for MSA would be a straight forward extension of that used for Pairwise Alignment.

Sadly this is not practical, and so MSAs are usually constructed in a series of pairwise alignment steps.

I leave further algorithmic detail to the MSA module of your course.

As will be touched on later, MSAs, represented mathematically rather than in the intuitive fashion to be offered here, are fundamental to many other forms of Bioinformatics analyses.

<Click>

The first step for the computation of a Multiple Sequence Alignment is, clearly, to find a set of suitable sequences.

<Click>

Then the software must align corresponding sequence residues by introducing gap characters (here I use minus signs) in the fashion that most closely fits the selected assumptions.

This is an entirely analogous process to that employed by Pairwise Alignment software.

The gap characters, again, represent either Insertions of Deletions, relative to the putative Ancestor sequence of the family.

It is impossible to distinguish between Insertions and Deletions unless the Ancestor sequence is know, which, in general, it is not.

Hence the term "InDel" which means "a gap representing either an Insertion or a Deletion".

<Click>

Note the alignment columns where the conservation is less than perfect, that is, columns that include at least one InDel or one Substitution.

<Click>

Then identify the fully conserved regions of the MSA.

In the conserved regions, the most interesting meaning(s) of the MSA are most likely to be found.

In reality, one would not restrict one's attention just to fully conserved regions, as here.

Substitutions between similar amino acids, plus a modest number of InDels may not necessarily eliminate the possibility of shared properties for any section of an MSA.

<Click>

Finally, seek any message the MSA might have to offer!

For this stage, Bioinformatics must take a back seat.

Used sensibly, MSA software, generally, does a reasonable job.

Not so good, however, that it should be accepted without question! Particularly where a user has a good understanding of the data, it is very sensible to make good use of software tools that allow MSAs to be viewed intuitively and EDITED!

Interpretation and fine tuning is for the informed Biologist!

With certain notable exceptions, it is ... and I sincerely hope, will always be ... true that "People are better that Computers"!

Bioinformatics can process the data and generate draft information.

It remains the role of humans to interpret and adjust ... to do the real Biology, that is.

Once a "Message" is identified, it night be a good strategy to link ever onwards in the hope of confirmation and enhancement!

<Click>

Searching a database for sequences that might be homologous to a query sequence (or MSA if suitably formatted) is the most common form of Bioinformatics analysis.

What could be more obvious than to ask of any sequence data not yet fully understood: "I wonder if there are more fully annotated homologous data available?"

To answer this question, the primary strategy has to be to compare the "query sequence or MSA" with every sequence in appropriate database(s).

<Click>

Clearly this involves **Pairwise Sequence Alignment** repeated, potentially, millions of times!

<Click>

The slow and careful optimal methods used for single pairs of homologous sequences are generally not appropriate, they are far too slow.

For database searching, it is necessary to use very crude (but quick and effective) strategies in order to get through the ever expanding databases before the next Ice Age is upon us.

Details, I will leave to the appropriate module of your course.

<Click>

The majority of alignments between a Query Sequence and the entries of a database will be between totally unrelated sequences, and so meaningless.

The software will generate a list of the more promising alignments it computes, ranked according to

statistics predicting how likely each may have biological justification.

That is, the alignment least likely (according to the software) to be between two unrelated sequences, will be at the top of the list.

<Click>

It is vital to understand that the computer will seek "similar" matches, whereas the user typically seeks "homologous" matches.

Not at all the same thing! 2 sequences can appear very similar but represent entities that are very different.

<Click>

Consider a protein example.

<Click>

The query is very "similar" to the matched Database entry.
Matches being indicated by the matched amino acid code, frequently "accepted substitutions" being represented by Plus signs, less commonly "accepted substitutions" being linked by Spaces.

<Click>

But ... is this similarity detected by the software to be interpreted as the discovery of two Homologous proteins ... or two proteins that both just happen to include an unusual number of Prolines?

<Click>

Consider also a nucleotide sequence example.

This time a vertical bar indicates a match and mismatches are represented by Spaces.

<Click>

The sequences are certainly similar beyond what would be expected by chance, but ... does this indicate interesting homology, or just any two mRNA sequences both complete with poly-A tail?

To be honest, the better database searching software tools have clever mechanisms to avoid such obviously misleading matches as I have illustrated here, which is not to say that more subtly examples cannot occur.

It is the case that the software detects "Similarity" (the evidence) it is up to the user to decide whether that the matches detected are meaningful or not (the interpretation).

Almost always, real biology must follow Bioinformatics analysis.

<Click>

For a final example, consider this pair of nucleotide sequences.

<Click>

Particularly considering there are only four possible letters! So a random alignment might be expected to be around 25% identical.

This is very far from convincing. The possibility of Homology would seem remote on this evidence.

<Click>

However, if the nucleotide sequences were coding for protein, they could be viably translated into amino acid sequences before alignment.

In this case, a terribly unconvincing nucleotide alignment is transformed into a perfect amino acid alignment!

Due to the redundancy in the Genetic Code, dissimilar DNA sequences can often translate into highly conserved Protein sequences.

The better database searching software will translate DNA Query sequences before comparing with a protein sequence database.

<Click>

When dealing with coding DNA, it always makes sense to translate and align at the amino acid level.

Protein sequences represent so much more information than do DNA sequences.

<Click>

Searching for simple patterns in either DNA or protein sequences is another common use of Bioinformatics.

<Click>

For DNA sequences the implementation of such searches is computationally trivial.

It is simply a matter of finding matches to a short string of characters within a larger string.

<Click>

Pattern searching is the strategy employed to locate the Recognition and Cut sites of Restriction enzymes to create Restriction Maps.

<Click>

The only real complication is that so few Restriction Enzyme Recognition Sites have no ambiguous positions!

<Click>

What is needed is an alphabet for DNA that includes codes for every possible ambiguity.

The International Union of Pure and Applied Chemistry (IUPAC) provides such an alphabet.

<Click>

Unambiguous Restriction Enzyme Recognition Sites can be represented using just the charaters A, C, C, G and T.

The forward slash indicates the Cut site. So, in the case of EcoRI, the Enzyme will cut the DNA between the G and the first A of the Recognition Site.

<Click>

Using the IUPAC DNA Alphabet, any site can be represented.

<Click>

The Restriction Mapping software operates by sliding the Recognition Site of each Enzyme to be mapped along the DNA under investigation.

Where the Recognition site matches the DNA sequence, given any allowed interpretation of all ambiguity codes, a Cut position is recorded.

<Click>

Simple text patterns are also used to identify features in proteins.

<Click>

Typically, patterns are manual determined to represent interestingly conserved regions of Multiple Sequence Alignments.

To varying degrees, patterns derived in this fashion could be said to represent the protein property, or feature, that provided the evolutionary pressure that resulted in the conserves region.

<Click>

Pattern design is trivial when conservation is perfect.

<Click>

But far less so when there are substitutions and InDels to consider.

<Click>

It would not be practical to construct an amino acid alphabet including all ambiguities.

This worked for the simple 4 letter alphabet for DNA, but would not be far from practical for the 20 letter amino acid alphabet.

<Click>

The solution is to use a simple pattern syntax to express variations where necessary.

The detail is not difficult, but let us not tackle it here.

I have embedded a link to a definition of the most commonly used syntax, for those who enjoy such things.

Hopefully the example conveys the gist?

<Click>

As previously, the pattern searching software will slide the pattern along the sequence under investigation, looking for matches.

Matches might indicate the an instance of the feature that gave rise to the conserved region of the MSA from which the pattern was derived.

Or they might be false positives. Chance matches. As always, it is up to the user to decide which.

Protein patterns are of very variable stringency. False positives are not uncommon.

<Click>

In truth, protein patterns are very limited in their ability to accurately represent conserved alignment.

<Click>

A meaningful protein pattern can only be designed for very highly conserved regions

<Click>

Patterns cannot weight possibilities. That is, it is possible to state that in a particular position an F or a Y are both acceptable (using square brackets, [FY]), but it not possible to record that one amino acid is more likely than the other given the evidence of the MSA.

<Click>

Pattern are exclusively based on just alphabet. They cannot reflect amino acid properties. For example, a fully conserved F in an MSA would be normally be represented by an F in a pattern. That is, any match must have an F in that position. However, Fs and Ys are known to frequently substitute for each other successfully. Should not some allowance be made for a potential match having a Y where the F was consistent in the MSA?

<Click>

Happily, there are more sophisticated solutions.

<Click>

Still using an MSA representing the feature to be modelled as a starting point, create a more

complete Model of the conserved feature.

Slide the Model across all proteins under investigation and come to a decision as to how meaningful that match might be.

<Click>

Various simple Models (or Profiles) were implemented in earlier software, including Position Weight Matrices .

These Profiles are used to detect various properties of both DNA and Proteins.

<Click>

Currently, most Profiles are Hidden Markov Models (HMMs).

In very broad principle, these are not so difficult to understand, but beyond the scope of this talk.

Suffice it to say that they include probabilities for all 20 amino acids in all positions of the MSA segment represented.

Also they include probabilities estimations for a Deletion and for an Insertion at each position!

To the casual observer, an HMM is a mass of numbers. Far from the intuitive patterns discussed previously, but many times more effective at detecting features.

<Click>

Phylogeny is the estimation of evolution from evidence from a variety of sources including palaeontology and comparative anatomy.

This topic will be covered fully in one of the modules of your course.

<Click>

In the context of Bioinformatics, the evidence upon which estimations are based is invariably carefully computed MSAs.

High quality MSAs are essential as they will be assumed perfect by the phylogenetic software.

<Click>

Estimated phylogenies are commonly represented as “Evolutionary Trees”.

<Click>

All very convincing, apart from the consistent insistence of placing HUMAN so dubiously close to MOUSE!

<Click>

Phylogeny is yet another example of a use of MSAs.

<Click>

A very important phylogenetic strategy is to compute the most likely tree given the MSA evidence.

<Click>

Thus emphasising the central role of Statistics in Bioinformatics.

<Click>

Protein Structure Prediction will be covered in a dedicated module of your course.

Here I aim simply to fit this into the pattern of the other Bioinformatics Topics covered in this talk.

<Click>

First, consider the prediction of Secondary Structure from Primary Structure (that is, the Protein Sequence).

In essence, the software is trying to locate the Alpha Helices and Beta Sheets of a protein.

Also, typically with rather less success, the Turns.

<Click>

The best modern methods use Machine Learning to generate Artificial Neural Networks.

Very simplistically, Machine Learning involves the evolution of a model of “something” simply by observation of many examples of that “something” (Training Sets). Analogously, so it is claimed, to the way that humans “learn”. Trial and error?

Models generated by this sort of process are called Artificial Neural Networks. Analogous, so it is claimed, to real Neural Networks.

Used appropriately, this approach has proved effective beyond more conventional, “rule-based” strategies.

So better models? But once those models are determined, their implementation is not dissimilar to to those described above for patterns, HMMs and similar.

<Click>

Secondary Structure can be predicted for an individual protein sequence, or for an MSA of homologous protein sequences.

The latter is far preferable as an MSA represents a far richer source of information than an individual protein sequence.

To crudely justify this claim, in this context, it would be reasonable to expect structural regions (Alpha Helices or Beta Sheets) to be highly conserved between a set of homologous proteins.

Frequent InDels and Substitutions would not generally be tolerated in a structural region.

The software only has these extra clues if it has an MSA to analyse rather than a single protein sequence.

<Click>

The best systems, when offered a single protein sequence to analyse, will automatically generate an MSA by searching and aligning similar sequences detected in appropriate sequence databases.

Thus the system insists on only analysing MSAs, in preference to single protein sequences.

<Click>

Predicting Tertiary Protein Structure is rather more difficult to achieve.

<Click>

De Novo Protein Structure Prediction is the prediction of Tertiary Structure directly from Primary Structure (that is the Protein Sequence).

This is the only option where no Homologous protein(s) of known structure are available for use as a template.

Simplistically, all possible folds would be evaluated. The conformation with the lowest overall energy being “The Winner”.

Due to the vast number of possible folds to consider, this would not generally be a practical approach.

However, the potential benefits of De Novo Protein Prediction are such that it is a very active field of research. Maybe one day, with better algorithms and increased computing capacity, it will become common practice.

<Click>

Homology Modelling (also referred to as Template-Based or Comparative Modelling) is possible only when reliable structure(s) exist for proteins homologous to those under investigation.

Very broadly, where a suitable template exists, the tertiary structure of a protein can be predicted by comparing it with that template structure using the corresponding template protein sequence as a guide.

<Click>

The number of available, experimentally determined, protein structures grows at a much slower rate than (for example) the volume of available sequence data.

Also, some proteins are less amenable to experimental structure determination than others, so the variety of available structures is not uniform.

But, never fear! Things can only get better!

<Click>

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.

<Click>

And now, once more, it is your turn! To think though a few simple issues.

There is no intention here to be tricky. In fact, I have embedded some links to what I hope will be “helpful hints” for you.

The intention is to ensure that all the simple issues of these presentations are understood and to invite you all to be more than passive receptors of the wise thoughts paraded before thee!

<Click>

Yes indeed ... There is more!

<Click>

and more …

<Click>

and yet more!! … keep on clicking, it does end eventually!