Power Point 01 - Bioinformatics Topics – Part 01 - Analyses

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 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!