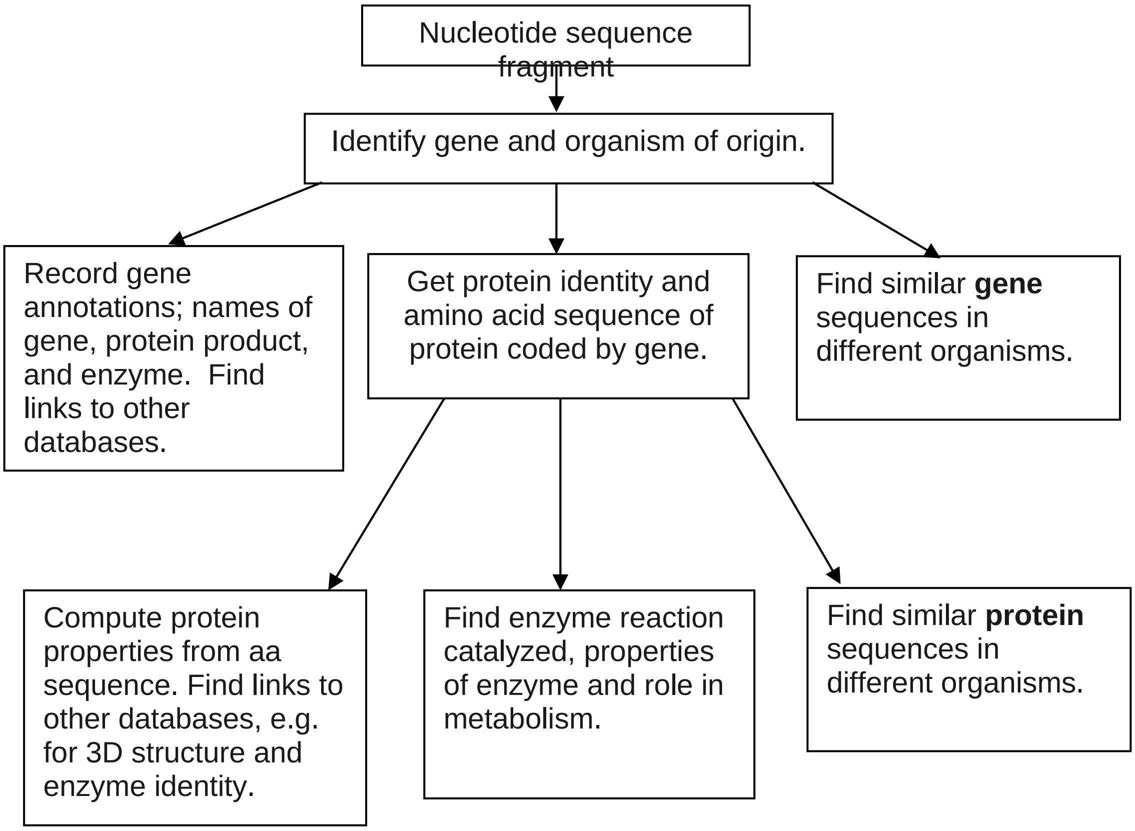
Bioinformatics Tutorial Session and Coursework

This is a small project involving the use of various internet-based databases & applications to identify and characterise a gene product starting with an identified DNA sequence. All the resources we will use are accessible using an internet browser such as Firefox, Netscape or Internet Explorer.

Aims

1. To gain an introduction to the contents and layout of the major web-accessible databases holding information on genes, proteins and enzymes.
2. To navigate between these databases to discover the source, identity and attributes of an enzyme given a fragment of its gene sequence, as follows:


Nucleotide sequence fragment
Identify gene and organism of origin.
Record gene annotations; names of gene, protein product, and enzyme. Find links to other databases.
Get protein identity and amino acid sequence of protein coded by gene.
Find similar **gene** sequences in different organisms.
Compute protein properties from aa sequence. Find links to other databases, e.g. for 3D structure and enzyme identity.
Find enzyme reaction catalyzed, properties of enzyme and role in metabolism.
Find similar **protein** sequences in different organisms.

About these instructions

Two warnings about these instructions:

- The web pages that access the databases change frequently in design and layout. Things might have changed by the time you look at them, in which case, search the pages for the most similar options to those specified!
- The way in which the data is recorded in the databases, and the amount of information available, differs for different genes and proteins, so again it is not possible to be absolutely definitive about where you should look for it, or whether you will find everything.

The instructions should be followed in conjunction with the page of web links provided for this exercise under BioinfLinks on Mudshark.

You might like to save pages as files for future reference, and also to have a Word processor page open so that you can paste in useful results from your searches that will be needed for your report, so that you don’t have to go back and repeat things.

Identifying the gene

Each of you will be allocated the name of a plain text file containing an un-identified fragment of a DNA sequence (seqn*n*.txt).

- The first step is to open this file. Go to Practical 1/Sequences on Mudshark. Open your assigned sequence file (it is convenient to do this in a separate tab; e.g. in Firefox, right-click on the filename). The text should be selected (select all) and copied to the clipboard. If you choose to download the file for future reference, you can open it subsequently with a text viewer such as Notepad (**not a Word processor such as Word**).
- We now need to access an application which will search the nucleotide databases (EMBL & GenBank) for closely related sequences. We will use nucleotide BLAST (Blastn) from NCBI (because you are starting from a nucleotide sequence). Use the Bioinformatics links on Mudshark to go to NCBI Blast and under ‘Basic Blast, choose nucleotide blast’.
    - Place cursor over the ‘Query sequence – FASTA sequence’ box and paste in your sequence from the clipboard (CONTROL V).
    - Under the ‘Database options’ choose ‘Others’ and select the ‘nucleotide collection nr/nt’ which is a combined database of all ‘non-redundant’ sequences.
    - Start with the option for searching for highly similar sequences and press ‘BLAST’ to start the search. You might find it convenient to use the option for returning the results in a separate window.
- When the process is complete we should have a page containing the BLAST output. It is worth saving this to disc for future reference. From this information you should be able to find the identity of your gene. In the table of text results, reference to a gene or a CDS (codinq sequence) should give a clear indication about the protein that gave the match. Note that:
    - The set of top hits may, include the organism’s complete genome sequence, in which case you have to jump to the right part, otherwise you can find every gene in the organism! The simplest way to find the matching part is to move over the line diagram of the colour-coded matches and click on the appropriate one.
    - Record these initial results in the table in the results proforma. Some of the results are clearly duplicates or only slight variants (e.g. different strains of the same organism) so choose a selection of slightly different matches, i.e. try to avoid listing the exact same sequence over and over just because different research groups have uploaded the sequence as part of different experiments.
- Search one of your top matches for the link to the gene, and follow the link to a **protein**. Note the protein accession number (Genbank or, even better,

SwissProt/UniProt if available), and also copy to the clipboard the translated protein sequence (which you may choose to save in a textfile using Notepad or similar. Your protein is an enzyme, so also note the EC number and the enzyme name for future reference.

Protein structure

We are now in a position to investigate the structure and properties of the protein coded by this gene and determine more about its structure and function. The ExPASy site in Switzerland is a very useful site, which gives us access to the UniProtKB (formerly SwissProt) database and many analytical programmes.

If we have the UniProt ID or Accession Number (AC) (or TrEMBL ID – where TrEMBL is an extension with computer-generated identities, of lower reliability) of our protein we can start to investigate using this resource.

- Go to the UniProt top-page and take the accession number (AC) route.
    - Enter the accession number and search. Save this page to file for future reference. If you have not yet got a UniProt number, use the ‘ID mapping’ function to search for the UniProt correspondence to your GenBank/Entrez reference numbers.
    - If it cannot find your accession number, or you have not got one, go further down the page and paste in your translated amino acid sequence from the GenBank search and run blastp (for a protein sequence search). You will get a set of matching proteins; at the moment you are interested in the best match, but the other matches are needed as well (see the section on the protein similarity search below).
- Either way, you should now have a page describing your protein with links to other databases and you can start to investigate the structural information there. There may be a features list – this indicates those areas of the sequence which fit known ‘profiles’ for domains which have identified functions. These can include: transmembrane regions, glycosylation sites, ATP binding sites etc. This will give you clues to the possible function of your protein. The FT viewer gives a graphical representation of these domains.
    - If the features are not shown here, you can use a number of applications to identify possible ‘profiles’. We can use PROSITE. This can be found on the Proteomics tools page on ExPASy.
    - Using the UniProt AC, start the scan for possible sites. Again, save the output. And follow the links to find out more about the various domains, including their consensus sequences.

Protein properties

There are also on ExPASy a number of programmes to measure properties such as hydrophobicity.We will use ProtParam and ProtScale, which are accessible from the Tools and Software packages section.

- From ProtParam obtain the pI, the overall charge and the *Grand Average of Hydropathicity* (GRAVY). Note these for your protein on the report sheet.
- We can then use ProtScale to look at the whole sequence, to determine the distribution of a wide variety of properties. One of the most commonly used is the Kyte and Doolittle measure of hydophobicity. This can be particularly useful for indicating sequences that may have transmembrane targeting. Using your protein's AC obtain a KD plot and save this.

Secondary Structure Prediction

We can obtain some useful information about the likely conformation of the protein if we know the extent of secondary structure such as α-helix and β-sheet in the sequence. There are a number of applications, some of them listed on the Proteomics page, which attempt to predict the location of secondary sequence. It is best to try a variety and look for consensus, since they use different methods and can give quite different results. SOPMA gives a consensus view.

- To use many of these you will need to paste your protein sequence from the clipboard (CONTROL V).
- Use one or more of these programmes to obtain secondary structure prediction. Save the output for your report.

Protein similarity search

Now that the identity of the gene product has been established, it is possible to find out what similar proteins have been identified, and whether this protein is a member of a known family, if you haven't already run this search.

- You can use the BLAST programme again for this, but we will need to use 'blastp' this time as we are querying the protein databases. You should know how to do this by now. The output will be long and complicated and will probably list some sequences which are nearly/identical to each other. Limit the output to 25 or 30 matches.
    - The databases inevitably contain duplicated material, so try to identify a list of different sequences (e.g. from different species) to fill in the table on the report form.
    - The degree of identity and the number of positive matches are at the bottom of the output in the same order as the summary table.

3D structure

- If you are lucky, someone has already solved the 3D structure of your protein, or a closely-related one. If this is the case, it will be stored on a database, probably the Brookhaven Protein Data Bank or one of its associated sites. These files are .pdb files and contain the x, y & z coordinates of every atom. There are likely to have been links to the relevant files on your earlier results pages. If there is no structure for the protein from the species containing your gene, widen your search to the homologous protein from the most closely-related species you can find.

- If you can identify such a file put this information on the form. It is possible to view these files with a number of excellent programmes such as Rasmol or Swiss-PDBviewer. However, some of these require installation on your computer, so the option most likely to work is to use the RCSB site and the Jmol viewer. This allows you to choose view angles and display options (Select ‘all’ first).
    - The RCSB pages have a ‘sequence details’ tab that allows you to see the secondary structure elements. How do these experimentally-derived results compare with the sequence-based prediction you obtained above? Save a picture of your protein for your report.
- The protein structure pages are also a starting point for finding out about identified domains in the structure.

Enzyme function

You should by now have come across a reference to your enzyme’s EC number.

- Follow the link to the Enzyme catalogue on ExPaSy and note the information about the reaction catalysed etc on the report form.
- If you follow a link to the BRENDA database, similar information will be available, but also information on the kinetic properties, with references to the literature describing the experiments that led to this. At the top of the Brenda results page, choose the option to limit the results to your organism (or the most similar species for which information is available).
    - o Find out if there are Km values for the substrates, and information on effectors or inhibitors.
- Use the EC number of your enzyme and the link to the BioCyc database to identify the pathways the enzyme participates in.

Originally devised by Alan Betteridge, 1999, and modified by David Fell, 2005-2010.

Extensively revised, 2011.

School of Life Sciences

Coursework Submission Sheet

To be completed by the student	
Name:	Student number:
Academic advisor:	Module number and title: **U14525 Mammalian Biochemistry**
Module leader: **Prof David Fell**	Title of work: **Bioinformatics report**
Work set by: **Prof David Fell**	Practical set: (circle) 1 2 3 4
I have read and understood the University's regulations on academic good practice, as set out in the 'Cheating' statement, and confirm that this piece of work is, and all future pieces of work will be, my own original work. Signed .... Date	

Summary feedback on key issues

Assessor:
Mark:
You have identified your enzyme and its properties correctly / partly correctly / incorrectly.
Your investigations and searches are complete / mostly complete / incomplete.
You have provided extensive / adequate / inadequate supporting information.
Your report demonstrates good / satisfactory / partial / inadequate understanding of the web site outputs.

Bioinformatics Report

Sequence file number:	.txt

1. **Results from BLASTn search** (giving only **one** example of multiple matches to the same sequence from the same species/strain under different accessions, i,e whole genome, isolated gene, cDNA):

GenBank ID	**Description**	**Score**	**Frequency (E) value**

2. **Identity of gene & protein product,** taken from best match above.

Protein name, species	**EMBL/Genbank code**	**UniProt/PIR code**

3. **Physical properties:**

pI:	**Overall charge (i.e. +/- balance):**	**GRAVY value:**

Attach a print-out of a Kyte and Doolittle plot showing the hydrophobicity of your protein.

4. **Protein similarity (blastp) search results** (**selecting only one match per species**):

UniProt Code	Protein name	Species	Score	Frequency (E value)	Identities	positives

5. **Protein structure**

   a. **Protein domains, motifs etc.** Enter brief details of any information found:

   b. **Secondary structure prediction**
   What are the **predicted** percentages of:

Helix ………………………………….

β Sheet ……………………………

β turn ………………………………….

Random coil …………………………..?

6. 3D structure

Relevant PDB filename:	.pdb

a. Attach an image of the 3D structure (using a related organism if you cannot find a structure for the protein corresponding to your gene).

b. Is the predicted secondary structure content similar to that seen in the observed structure?

7. Enzyme identity

a. **EC number** ………………………………

b. **Name** ………………………………………………………………….

c. **Reaction:** …………

d. **Metabolic role** ……………………………………………………………………

…………………………………………………………………………………….

……………………………………………………………………………………..