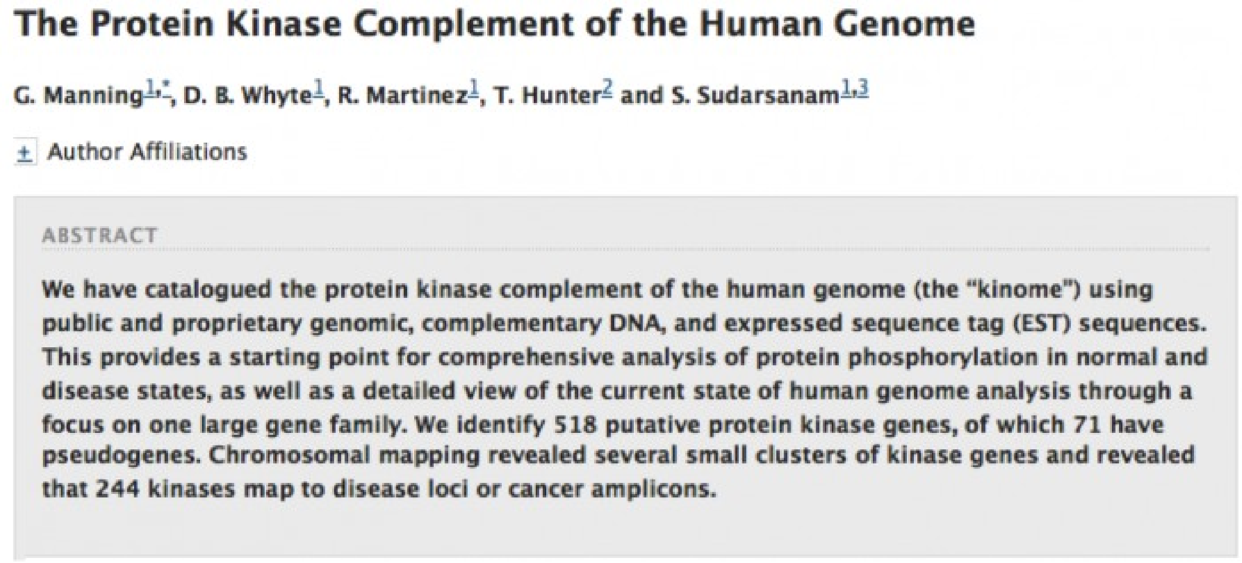
Practical Session 1: Introduction to databases and homology-based functional inference

A- UniProt

Exercise 1

You were flipping through a past issue of Science and came across the following article:



The Protein Kinase Complement of the Human Genome

G. Manning[1,*], D. B. Whyte[1], R. Martinez[1], T. Hunter[2] and S. Sudarsanam[1,3]

± Author Affiliations

ABSTRACT

We have catalogued the protein kinase complement of the human genome (the "kinome") using public and proprietary genomic, complementary DNA, and expressed sequence tag (EST) sequences. This provides a starting point for comprehensive analysis of protein phosphorylation in normal and disease states, as well as a detailed view of the current state of human genome analysis through a focus on one large gene family. We identify 518 putative protein kinase genes, of which 71 have pseudogenes. Chromosomal mapping revealed several small clusters of kinase genes and revealed that 244 kinases map to disease loci or cancer amplicons.

You want to find all human protein kinases in UniProt that have a 3D structure associated with them.

A- How would you do this?
B- How many proteins have you found?

Exercise 2

You are helping colleagues analyse the protein products of the gene tp53 in different organisms. They have sent you a list of gene accessions that they are interested in. They are interested in analysing the corresponding protein products for these genes.

Genes (Identifiers from GenBank)
X99952
AP002032
CP002688
AY056186
AY096713
AY087674
X02469
M13121
M13112

M13113
M13114
M13115
M13116
M13117
M13118
M13119
M13120
K03199

How would you download the corresponding UniProt protein sequences for these gene accessions?

Quiz 1

1- UniProtKB consists of two types of entries, Reviewed (Swiss-Prot) and ________.

A- UniRef
B- Reference
C- Unreviewed (TrEMBL)
D- Proteomes

2- UniProt allows you to convert other database identifiers to UniProt identifiers but not vice versa.
A- True
B- False

3- Which of the following is not a UniProt dataset?
A- Taxonomy
B- Diseases
C- UniParc
D- Pathogens

4- To find functional information about a protein, which UniProt section should you consult?
A- UniProtKB
B- Uniref
C- UniParc
D- Taxonomy

5- UniProt provides complete proteome sets for organisms whose genomes have been completely sequenced.
A- True
B- False

6- Which of the following are tools provided on the UniProt website for protein sequence analysis? *Choose all that apply*
A- BLAST
B- FASTA
C- Clustal Omega alignment tool
D- Identifier mapping

B- KEGG

Exercise 3
Glycolysis is the process of converting glucose into pyruvate and generating small amounts of ATP (energy) and NADH (reducing power). It is a central pathway that produces important precursor metabolites.
Using KEGG you have to identify differences in the Glycolysis pathway between this fungi species: Penicillium rubens and Tremella mesenterica.

Exercise 4
Oxidative phosphorylation is the process in which ATP is formed as a result of the transfer of electrons from NADH or FADH 2to O 2 by a series of electron carriers. This process, which takes place in mitochondria, is the major source of ATP in aerobic organisms.
In eukaryotes, these redox reactions are carried out by a series of protein complexes within the inner membrane of the cell's mitochondria, whereas, in prokaryotes, these proteins are located in the cells' intermembrane space.

A- Which is the main difference in the Oxidative phosphorylation pathway between Human and the yeast *Saccharomyces cerevisiae*?
B- And between Human and the bacteria *Escherichia coli*?

C- InterProScan

Exercise 5
Find information about the protein in **my_protein.txt** using InterPro.
What family does this protein belong to?
What domains does it have?
What processes is it involved in?

Quiz 2
1- How can you know the type (family, domain, repeat, site) of an InterPro entry?
A- The entry type is indicated by a specific icon before the name and identifier on every IntePro entry page
B-By looking at the GO terms for the entry

C- InterPro entries do not have a type

2- When an IntePro entry consists of several signatures, this means:
A- Those signatures match exactly the same set of proteins
B- Those signatures were made by the same member database
C- Those signatures are predicting the same biological entity: a protein family, domain, repeat or site

3- If you have a novel uncharacterised protein sequence you can use InterPro:
A- To submit your sequence to a protein database
B- To predict the function of the protein and the presence of important domains or sites
C- To perform a structural alignment with others sequences of interest

D- GO enrichment

Exercise 6

Search for GO enrichment terms in Biological Process using the list of genes from Arabidopsis thaliana (Athaliana_identifiers.txt). To do this, you must use the Gene Ontology Consortium website http://geneontology.org/. Download the results.

REViGO (http://revigo.irb.hr) can take long lists of Gene Ontology terms and summarize them by removing redundant GO terms. The remaining terms can be visualized in semantic similarity-based scatterplots, interactive graphs, or tag clouds.

Using the GO terms and p-values obtained before you must obtain a Treemap from REViGO. To perform the search the GO IDs may be followed by p-values (or another quantity which describes the GO term in a way meaningful to you).

Exercise 7

A- Using the same gene list (Athaliana_identifiers.txt) you must search for GO enrichment terms in Molecular function.
B- Using Quick GO (https://www.ebi.ac.uk/QuickGO/), lists all terms that are direct descendants of the GO terms obtained from A

D- Ensembl

Exercise 8

C-C chemokine receptor type 5, also known as **CCR5** or CD195, is a protein on the surface of white blood cells that is involved in the immune system as it acts as a receptor for chemokines. This is the process by which T cells are attracted to specific tissue and organ targets.

You must search for CCR5 Human gene in ensembl and obtain the follow information:

Which is the ensembl gene identifier?
How many transcripts has this gene?
How many homologs are in chimpanzees?
Download the Genomic sequences.
Make a multiple sequence alignment between Human and Bonobo, Chimpanzee, Crab-eating macaque, Gibbon and Gorilla.

Exercise 9

Choose a protein from the fasta file ASPCL.fasta and use the web resources we have seen, in order to annotate that protein (you should get all the information you can).