Homework 5 – Class Project Proposal, Conceptual Design Phase

Homework 5 is a required proposal of your project. You will submit the topic you have selected along with the following detailed information.

In the conceptual phase of a database project you consider the scientific researcher's *use case* – what does s/he want to accomplish. In this assignment you will have to **formulate 5 questions** of scientific interest, then produce **a work flow diagram** that includes data acquisition and processing steps (For example, download file ABC from Chloroplast DB, unzip, verify unique file names, separate sequence and quality score values- these are ready to upload to the Sequence entity as attributes FASTA and PHRED , and B) **an ER diagram** that includes the attributes that belong to each entity for your class project, and your first pass at establishing the relationships across the entities.

In addition you will include **a brief write up** emphasizing your reasons selecting the topic, why it is of scientific interest, the significance of the questions you propose. How will creating a database work to help solve the scientific problem you present. Include information about what data will be needed and how it ties into the scientific ontology you select.

You must turn in the following:

1. What are the 5 (minimally) scientific questions your topic, for which sequence data is available, that your database will allow you to answer?
2. Provide a high-level workflow diagram showing any transformations you will have to perform to create attributes you need from the files available (see examples below).
3. Create an ER diagram that includes
    a. The main entities and the attributes required by your questions and thus present in some form your workflow diagram
    b. The relationships connecting them that will be needed to hold and query the information required to answer your questions.
4. A write up discussing your topic, its scientific relevance, the ontology that relates to your project, the data you have collected and how you will use this data to answer the questions you propose.

Notes:

For step 1 turn in the queries as a scientist would ask them, not as SQL, and explain what data will be needed to answer them. There can be more than 5 questions initially, until you figure out which ones are going to be productive. Then make sure you can find the needed data sets to answer at least some of the questions.

For step 2 provide a workflow diagram similar to those shown below for the acquisition and manipulation of the data files that allows you to use them

For step 3 turn in a graphical representation of the model using IE notation (not hand-drawn, use some CAD tool).

For step 4 find an ontology that seem to map to your own entities, and see whether the relationships match those you defined.

For step 5 explain the source database and file names for the data you will use to populate your database, and verify that you are able to download the file, open it, and on inspection can identify the elements you need.

Note: there will undoubtedly be many changes between this design step and your final product, but having a game plan and knowing where work will be required is a first step towards success. I am asking you to do this in part because most of you will have far too ambitious a plan and I want to be sure it is sized appropriately.

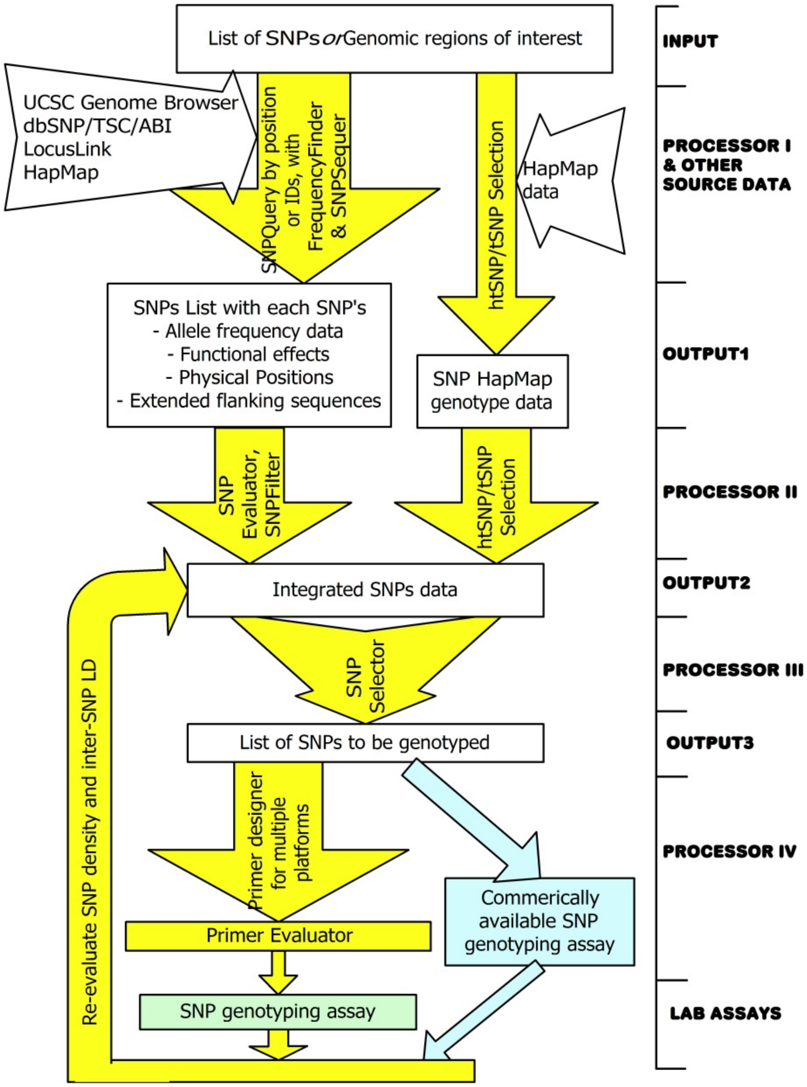
Here is an example of a workflow for SNP data mining from http://www.bioapp.psych.uic.edu

Dr. Weller's Example Submission:

Scientific questions could include (this is not for a prokaryote that causes a disease so it is not directly analogous to your project requirements).

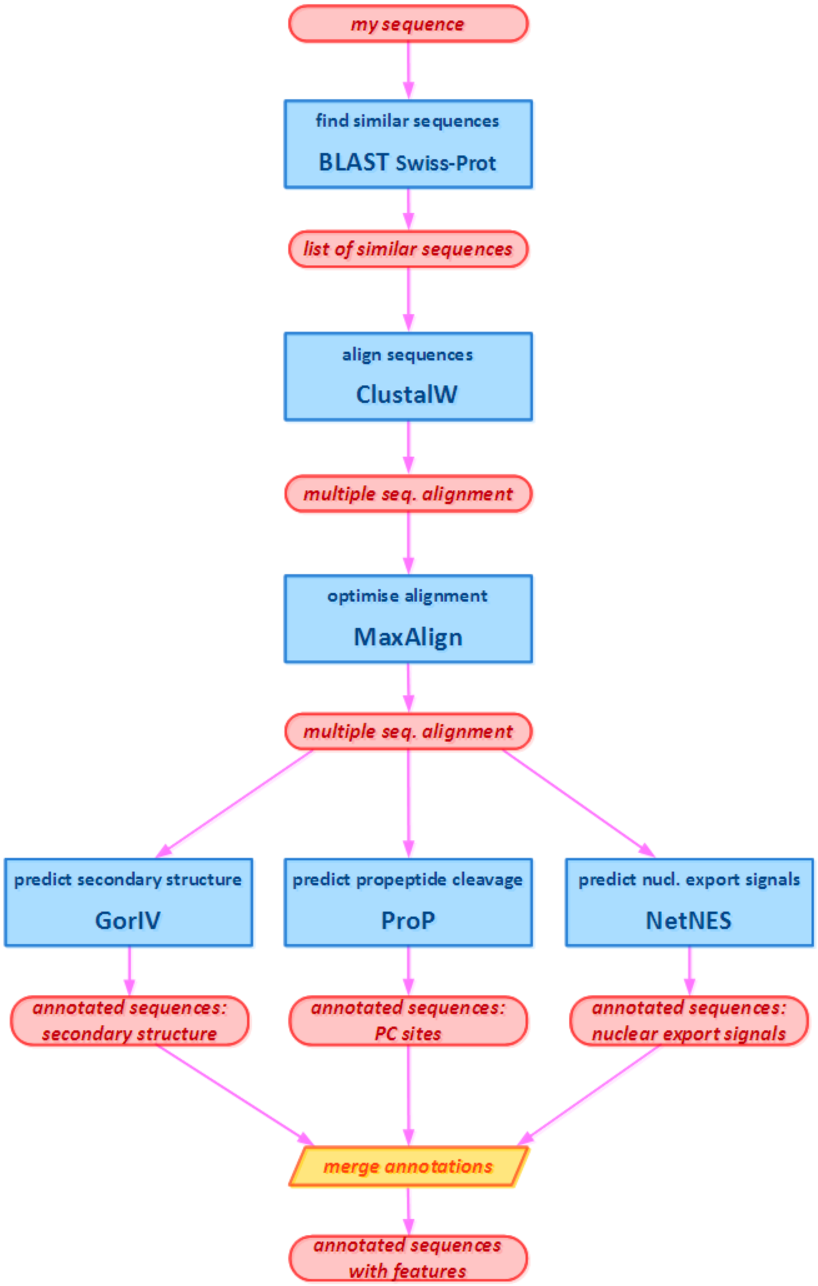
For a specific psychiatric disorder, what genes seem to be associated with the disorder when there are particular coding sequence variants present (variants could include SNPs or CNVs)? Do any of these lead to changes in the protein primary structure, or to protein processing differences? What pathways do those genes belong to? Are the compensating variants, so that a change in one gene removes the effect of a bad variant in another gene? Are these variants present in any mouse models with similar disorders? How conserved are these genes evolutionarily?

Figure 2. Workflow of SIMP

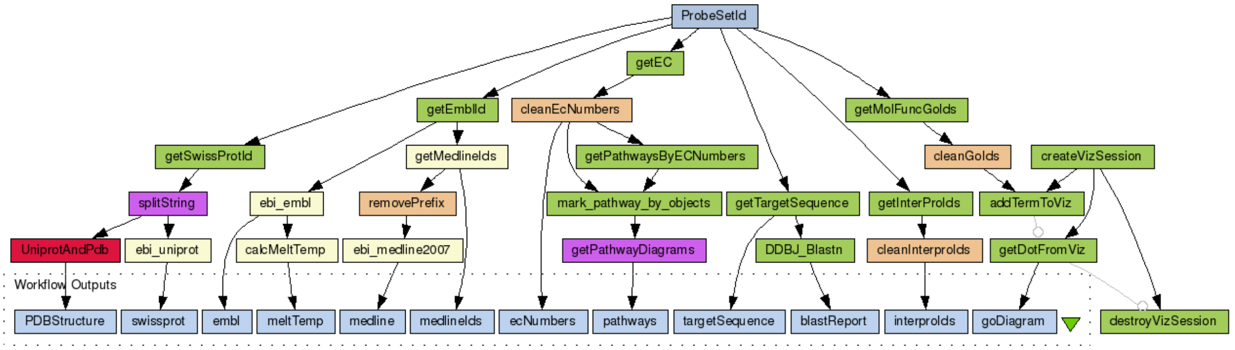


Note: SNPs discovered by local re-sequencing project can be mapped by DNannotator, evaluated for putative functional effects, combined with data from public sources, and enter into SIMP data flow

Note that by having the SNP workflow, I realize that I have to include some information about the difference assays used to measure sequence variation, because reports for 'genes' aren't always measuring the same variants. Some data has to be processed by an application to produce the next set of data.


my sequence
find similar sequences
BLAST Swiss-Prot
list of similar sequences
align sequences
ClustalW
multiple seq. alignment
optimise alignment
MaxAlign
multiple seq. alignment
predict secondary structure
GorIV
predict propeptide cleavage
ProP
predict nucl. export signals
NetNES
annotated sequences: secondary structure
annotated sequences: PC sites
annotated sequences: nuclear export signals
merge annotations
annotated sequences with features

This part of the workflow handles the problem of looking for multiple sequence alignments and figuring out if the changes in the sequences lead to property changes in the proteins of specific types. Boxes in blue are software applications that transform the input data.



(from http://www.myexperiment.org/workflows/28/versions/2/previews/full).

This part of the workflow shows how to find pathways and pathway diagrams starting with microarray data (ProbeSet ID is an Affymetrix concept), along with some other protein data that was not really part of my set of questions.

I have used other people's workflows to get an idea about what they have included that I might have neglected – this is fine so long as you give them credit, and so long as you integrate them specifically to suit your project.

Note that my first questions have much too wide a scope – I will end up with a database that is too complicated for a class project. By drawing out the workflow I can figure this out pretty fast, and then I can focus prune the queries to focus on one area that is appropriate for the scope of a project. If I don't work to a detailed level I won't realize this until I am under water.