MRS-EMBOSS Sequence mining tutorial-DAY 3:

These exercises are designed to enforce the knowledge presented on the Sequence Retrieval Systems lecture . At this point, you should have prepared your online UNIX (workstation) environment during the tutorial session of Day 1. For some of the questions there is not a single right answer. If you do not understand the point of the question, please ask your instructor.

1)Visit the URL/Web page: http://cnkeeper.uio.no:8080/mrs-web/
This exercise aims to familiarize you with aspects of the MRS (version 4) visual web interface:

> BEFORE YOU BEGIN THIS EXERCISE, MAKE SURE THAT THE MRS-4 INTERFACE IS SET IN 'RED-BARN STYLE' MODE. THIS IS THE DEFAULT MODE IN MRS-4. IF YOU DO NOT SEE A GREEN TOP MENU BAR, FOLLOWED BY A RED SEARCH BAR IMMEDIATELY BELOW, ASK YOUR INSTRUCTOR TO RESET YOUR INTERFACE.

a)From the top bar menu (green toolbar), choose the 'Databanks' entry and then the 'Indexed' option (**Databanks->Indexed**). This should give you a list of all available databases on the MRS server.

b)For each of the Indexed databases, click on the **'ID'** or **'Name'** column. What you get are two sections:

- A section of statistics for the database, mentioning useful metadata information such as number of entries, file names and references ('**Additional Information**' row).
- A section where the index names are mentioned (Recall from the lecture notes that some index names are common across different databases, but you should be familiar with the index naming conventions of the databases you are interested, especially when you construct complex queries.)

c)Notice that when you select a particular database, you have on the green top bar an additional '**Databank:database_name'** indicator appearing. This is a useful indicator to watch for, in order to keep reference of which database you are working with.

d)Go back to the home page of MRS by selecting Home. In the main page, follow the 'Tip' text to make sure you enter MRS on cnkeeper.uio.no as part of your search engines in your browser (preferrably you should be using Firefox).

2)It is now time to start issuing some queries and use MRS. Go under the green top bar and you will find the search bar in red color. Select the **EMBL** database from the **'Search'** drop down menu on the left and see if:
a)You can locate **human** related H1N1 sequences by typing a suitable MRS query.

b)After entering a valid query in step i), MRS will return a list of sequence results sorted by ID and relevance. Between the '**Nr**' and '**Relevance**' result columns, you will find a column with tick boxes. Make sure you select some of the results sequences of interest (left click) and then click on the ▼ symbol at the top of the column. You will then see various options popping up ('Select All', 'Deselect

All', 'Export', 'Export Fasta'). Try to play with these options and see if you can export the result sequences to your favorite text editor.

c)Go back to the result view of your query (your browser's back button should work). Make sure you spend some time to click on each result entry and use the View button to familiarize yourself with the different formats you can view the results.

d)See if you can locate all H1N1 sequences published by 'Subbarao' and are complete CDS of a 'nucleoprotein gene'.

e)Try to repeat the queries by entering search terms without specifying indexes. Does it work? When you combine these sort of search terms in your search, which Boolean operator is implied to exist between the search terms?

3)This exercise will teach you how to use the MRS add-ons BLAST and CLUSTALW:
In MRS, it is possible to blast protein sequences against protein databases. It is also possible to perform sequence alignments using clustalw. These utilities are integrated to MRS.

a)Try to use MRS to locate protein sequences for Human Immunodeficiency Virus (HIV) Gag polyprotein. Click on one of the result entries and then click on the 'Blast' button (at the top of the result). You will get into a screen of familiar blast parameter details. The selected sequence should already be in the query field. Make sure that you blast the sequence against the 'Uniprot KB' database. Your query will be in a queue for execution. In the meantime, select other sequences to blast against the same database (**notice that your results and queries persist during your browser session. This means that you can continue doing other things in MRS and return back to the 'Blast Results' field in the top green bar**). Make sure you spend some time looking at the way the results are presented.

b)Now, we are going to see an MRS feature called '**dynamic database filtering**' query. This is useful to reduce the execution time and the search space of BLAST queries, making your searches more targeted and efficient. Go back to the query screen (Home). Select the same query as in a). Select one of the resulting protein sequences to blast it against the UniprotKB. However, this time, we are going to reduce the search space dynamically, before pressing the 'Run Blast' button. Type in the field under to the Blast Database selection ('Optionally enter an MRS query to limit the search space') the following: os:human . What do you achieve with that secondary search query?

c)Select some of the results of the blast query (say the first six entries). Left click on the ▼ result column symbol, and from the options select Clustalw. You will immediately get to a clustalw query screen. The selected protein sequences should already be in the input field. Then just press, 'Run Clustalw' (top right).

d)Use the command-line 'mrs-query' application (and some PERL/Unix scripting) to create a file containing HIV Gag polyprotein sequences and another file containing non HIV related polyprotein sequences. Create two separate blast databases from the two.

4)OMIM is a database that contains a collection of disease related genes. The famous physician Dr House comes towards your desk and says to you: “I have a patient with an atypical case of hepatitis. I can detect no viral infection, he is no drinker, not receiving any medication that could affect his liver... I am running out of ideas. I also noticed hand tremor and his family reported occasional changes in his mood”.
a)Can you use the OMIM's MRS indices to give the physician a hint about what might be wrong with the patient?

b)Hopefully you found the right disease in step a). Which polypeptide encoding gene **allele** is often mutated and causes the disease in question?

c)Use the EMBL (or the Genbank) database to download the gene allele. Then use EMBOSS to extract from the Feature Table the CDS.

d)Use MRS to blast the obtained CDS protein against UniProt's 'E.Coli' sequences (BLAST query filter reduction, as described in exercise 3). From the best hits you get, can you deduce any conclusions (supported by literature)?

George Magklaras

-If you are interested in becoming EMBnet Norway users, please visit the following web page:

http://www.no.embnet.org/register.html

Our services include an entire array of computational life science tools provisioned in powerful centralized computer servers and are described here:

http://www.no.embnet.org/services.html