PB607 Final Homework : Entrez & PSI-BLAST

Tutorial.

If you have not done the online Entrez and blast tutorials, now would be the moment to do them.

ENTREZ.

- The online Entrez tutorial for this class is available at http://germ.cit.cornell.edu/~cmlarota/PB607/PB607_introductory/Intro_contents.html#Entrez
- The online tutorial from NCBI is available at http://www.ncbi.nlm.nih.gov/Database/tut1.html

Exercise

Using the Web-Entrez interface at www.ncbi.nlm.nih.gov/entrez, answers to the following questions:

1. Use Entrez to find the number of nucleotide sequences in GenBank having the keyword "yeast". How many sequences are found by searching for "yeast" in "Organism" field, compared to searching this term in "All Fields"? Do you find the same number of sequences? If not, why not?

2. What is the taxonomic classification for the plant "cassava" (*Manihot esculenta*) as defined in the Entrez-taxonomy database? How many protein and nucleotide sequences are available in the databases for this species, respectively? Give a list of accession numbers for those proteins related to starch biosynthesis.

Find out whether a publicly available genetic map with molecular markers exists for this species (hint: search for published papers on the subject). Find a suitable species with cDNAs in the NCBI nucleotide database that would be likely to cross-hybridize with the cassava genome. Use criteria such as taxonomic relatedness and number of available sequences in the databases to justify the choice of this species as a source of candidate markers for constructing a new genetic map of cassava.

3. Using Entrez, download all the EST sequences (in FASTA format) from rice (Oryza sativa) that have been deposited in dbEST since September 1, 2000. Make sure you are dealing only with ESTs and not GSS or other sequence types. Report the settings and query string(s) that you used to conduct the search, the number of sequences that you downloaded, and the gi-number/accession from the **first** and the **last** sequences that you downloaded. (Don't print or include the whole downloaded list!!).

Repeat the same search, but this time request only the list of gi-numbers and save the downloaded list on your desktop. Open the list with a word processor or spreadsheet and examine the list to find whether all of the gi numbers are sequential (indicative of a single bulk submission to GenBank), or if there are several series of consecutivel gi-numbers (typically indicative of more than one independent submission). Partition the list so that you can identify the different sequential numbers into discrete groups. You can do this easily by simply adding a line (press 'return' key) or inserting a cell (in your spreadsheet program) between each of the groups. How many discrete groups do you identify?

Select two gi-numbers at random from each discrete group of sequences and make a separate "short" list with them. Download the GenBank formatted records for these gi-numbers (you can either download them or view them directly from Entrez). Look at the GenBank records to determine which laboratories made the submissions. Report the name of the submitters.

4. There are several records of cellulose synthase proteins in Genbank. A few kingdoms are represented, from higher plants to bacteria. Go to the Entrez protein database website and search for "cellulose synthase" in "All fields". How many records do you find? How many of these records are from fungi, bacteria, and higher plants, respectively? You will probably find it helpful to use the taxonomy database to guide the construction of your queries, including the correct spelling of various taxonomic divisions.

Now move down in the taxonomy tree to lower levels (such as family and genus) from within each of the categories you had identified at higher levels. Are there any species that are being over-represented relative to all others? If so, can you explain why? Does this means that these species have more genes coding for cellulose synthases than other species?

BLAST.

This link to an abstract of a paper will give you a very fast introduction to cellulose synthase biochemistry inside (and outside) the cell. (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=10874579&dopt=Abstract).

5. Use the pairwise BLAST server at http://www.ncbi.nlm.nih.gov/gorf/bl2.html to align record for gi numbers 7270145 and 9622890, the cellulose synthase catalytic subunit (RSW1) from *Arabidopsis thaliana* and the maize cellulose synthase-9, respectively. Leave all settings at their default values. These are two purportedly homologous sequences, and the pairwise alignment gives an idea of the level of similarity between these two closely related proteins. Notice numerous differences. Which portions of the proteins appear to be conserved (write down the amino acid coordinates of the regions) and write down the overall scores and E-value, %identity, %similarity and length (amino acids) of each HSP.

In the pairwise alignment, there are some regions of the query that were masked out with "X". These regions have "low complexity". The default value is to mask them. Try to realign these two sequences without masking the low complexity regions of the query sequence. To do this, uncheck the box next to "filter" and re-align. Do you see a difference? Explain. Under what circumstances would one want to unmask low complexity regions? You will probably need to read the tutorials for the answer.

6. Do an "advanced" BLASTP search of the *Arabidopsis* sequence from the previous question as the query against the non-redundant protein database. Modify parameters to view as many as 500 resulting alignments, leaving all other parameters at their default settings.

Follow the link to browse the taxonomy reports window (the link is right above the graph of the distribution of hits). The first part of the taxonomy report window shows the "lineage report" where results are sorted by taxonomic relationships rather that by the significance of the hits. There may be some biologically irrelevant hits. List the gi numbers of any that you find. Do some of these non-relevant hits have a higher similarity score (and E-value) that other hits annotated to be cellulose synthases? If so, explain how this paradoxical result could be obtained.

Now scroll down to the second part of the taxonomy report. Here the BLAST hits are listed by species, and sorted by the highest score reported for a hit from those species. Since you setup the program to report all hits, it will report non-significant hits as well. How would you set the cutoff parameters to separate the significant hits from non-significant hits?

7. Go back to the original report window. Explore the graphic representation of the distribution of hits by moving the mouse through the window to see the position of HSPs. Note, the names are displayed in top box, red lines denote hits with very high scores, then pink, green, blue, etc..

The report reveals hits of Arabidopsis cellulose synthase RSW1 to transcription factors in soybean and fava beans, e.g.

```
gi|7488867|pir||T12093  TGACG-motif binding protein - fava b...   191  4e-47
gi|7488717|pir||T08591  TGACG-motif binding protein STF1 - s...   188  2e-46
gi|7488718|pir||T08592  TGACG-motif-binding protein STF2 - s...   175  3e-42
```

These have very high scores and low E-values. Try to locate these hits in the graphic of the distribution of hits (this gives you an overall spatial reference to compare to the other hits). Do the alignments span the entire length of the shorter of the query and subject sequence? Do these results suggest that these sequences are homologous to the Arabidopsis query protein, or that they share one or more domains? Could we suggest that the Arabidopsis cellulose synthase is able to bind DNA? (hint: check the suspicious soybean and favabean protein records to find where the DNA binding domain is located in those proteins)

8. The BLAST report also reveals hits to bacterial cellulose synthases, although these are hits have much lower score and E-value. A list of species and their hits is shown next:

```
 Escherichia coli [enterobacteria] taxid 562
gi|2851646|sp|P37653|YHJO_ECOLI HYPOTHETICAL 78.6 KD PROTE...    65  4e-09
gi|1073471|pir||S47754 hypothetical protein f692 - Escheri...    65  4e-09

 Acetobacter xylinus [a-proteobacteria] taxid 28448
gi|4827167|dbj|BAA77593.1| (AB015803) bcsABII-A [Acetobact...    61  5e-08
gi|4827175|dbj|BAA77600.1| (AB015804) bcsABII-B [Acetobact...    61  5e-08
gi|7521917|pir||T31338 cellulose synthase (UDP-forming) (E...    51  4e-05
gi|3298349|dbj|BAA31463.1| (AB010645) cellulose synthase s...    51  4e-05
gi|114889|sp|P19449|ACSA_ACEXY CELLULOSE SYNTHASE CATALYTI...    51  5e-05


 Agrobacterium tumefaciens [a-proteobacteria] taxid 362
gi|2120777|pir||I39714 cellulose synthase - Agrobacterium ...    55  3e-06
gi|710492|gb|AAC41435.1| (L38609) cellulose synthase [Agro...    55  3e-06
```

What would you say about the similarity of these bacterial sequences relative to the Arabidopsis query protein? What information regarding function could you extract from these alignments? Do the HSPs for these bacterial proteins occur in the conserved regions identified in question 5? What is the nature of the functional motif(s) shared by plant and bacterial CelluloseSynth proteins?

PSI-BLAST

We could see whether using a profiling method (such as PSI-BLAST) will find better statistical significance to non plant cellulose synthases and perhaps to be able to detect other biologically relevant sequences not reported here. There is an interactive tutorial for PSI-BLAST on the NCBI WebPages that follows the exercise of trying to identify the function of a rare unknown bacterial protein that gets few informational hits from normal BLAST runs. You should take a look at it to make sure you understand PSI-BLAST's web interface and what the program can do for you. To go there, follow this link: (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/psi1.html). Notice how they make use of any biological information available to create hypotheses and in some cases, test them by separate (experimentally).

9. Using the same query sequence, gi 7270145 from *Arabidopsis*, start a PSI-BLAST search at http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi. Select the "nr" database, change the cutoff E-value to 0.01 from the default of 10, filter low complexity sequence, request the graphical overview. What you get as your first report is a run of a normal BLASTP against the database.

What substitution matrices should be used in this step?

This first set of alignments is used to create the first profile matrix to be used as query in the first iteration. What you include in the first iterations has a determinant effect in the final result of PSI-BLAST. If you allowed the wrong sequences to be included at the beginning, after several iterations your results may be misleading. This is where the biological knowledge comes into play, when deciding which is you threshold and deciding which sequences might be false positives (remember all this is just statistics) and which ones might be worth leaving in. Take a look at the results; note the sequences, scores and E-values. For this exercise let the program choose the sequences for you (a check mark to the left indicates that it will form part of the set of sequences used to calculate the profile matrix). Run the first iteration.

Take a look at the results of your first PSI-BLAST iteration. Look at the scores and E-values of matches. Notice any changes? How were the scores and E-values of previously seen hits affected this time (compare to hits from question 8)? is the order of sequences the same? Remember that hits are sorted by E-value. Anything new has been found? Describe them.

Repeat the experiment (do another iteration) and try to answer the same questions given this new data (if new at all).

Is there a point, after several iterations where nothing new comes out in the report? How many iterations did you need to reach this "convergence" point? What could affect the number of iterations needed to reach a convergence point?

After you have answered the previous questions, check these abstracts from pubMed and tell us what bells it rings on you (you don't need to modify your answers to the homework if this paper says something you missed):

From 1996:
(http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=8901635&dopt=Abstract)
From 1994:
(http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=7815941&dopt=Abstract)
And from 1992:
(http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=1472720&dopt=Abstract)

Bibliography

Altschul, S. F., T. L. Madden, et al. (1997). “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.” Nucleic Acids Res **25**(17): 3389-402.