Assignment_04_2024

Your name:
Your email address:

Once you are done with the exercise, email your answers as an attachment to gogarten@uconn.edu **and** daniel.s.phillips@uconn.edu

Objectives:

- Know about bibliography software.
- Know about the advantage of the databanks accessible through NCBI's Entrez.
- Be able to perform literature databank searches at Google scholar, scopus and pubmed.
- Know how to retrieve full length manuscripts.
- Appreciate usefulness of # of publications, # of citations, and the H-index
- Know how to find manuscripts that are similar to one that you know is relevant to you.
- Appreciate that GenBank is highly redundant.
- Be able to perform blast searches via web interfaces and to evaluate the results

Today, we will focus on searches of databanks at NCBI's entrez, via Google Scholar, and Scopus. Related to the literature databank searches is bibliography software. If you write scientific or other academic articles, you frequently need to cite the literature. If you see writing scientific papers, thesis, or similar, get used to using a bibliography program ASAP. This makes it easier to incorporate citations into an article, to reformat the bibliography, and to download citations from the internet. Popular choices are Endnote, Refwork, Zotero and Mendeley.
Endnote is popular, but expensive, and old versions usually stop working when you update your operating system. Also, citations incorporated into a text document cannot be used by other citation programs. Refwork also is a commercial software, but UConn has a subscription.

Mendeley and Zotero are similar, but sadly no longer compatible with one another. - You still can export your personal library from one program and load it into the other, but the two softwares can no longer be used on the same document. Mendeley is from Elsevier, but free, and popular. Like Zotero, it also can be used to keep track of pdf versions of articles, and it is updated within reasonable time, when Microsoft office or the operating system becomes incompatible with an older version.
Your database of references is stored online, you can share folders with others, and you can use your references from different computers. Both, Mendeley or Zotero , come with three important features:

A) a bookmark for your browser that automatically downloads citation (and pdf, if available) from the cite you are visiting (e.g. pubmed, the article on a journal page, scopus, ... *),

B) a plug-in for microsoft word, the allows you to insert citations into the text you are writing,

and
C) a desktop application that allows access to your references, and different bibliography style sheets.

I strongly encourage you to install either Mendeley or Zotero on your personal computer.

*: This sometimes does not work as seamlessly as it should, because different sites use different Journal abbreviations, capitalization for the title and author initials. These inconsistencies disappear, when you always use the same source (pubmed) to fetch the references.

1. (less than 10minutes)
   Use Pubmed in NCBI's Entrez to find an article written by Carl R. Woese (famous scientist, co-discoverer of the Archaea), published in the journal Proceedings of the National Academy of Sciences of the United States of America with the words primary kingdoms in the title of the paper. Try to use Boolean operators (**AND**, **OR**, **NOT**) and field tags; if you cannot recall the tags, use the pull-down menus under "advanced" (link below the search text window).

   What query did find the 1977 article?
   Your answer --->

Aside: The above link https://pubmed.ncbi.nlm.nih.gov/?**otool=uconnlib** lets pubmed know that you are at UConn, and provides links on how to download articles that are behind a paywall.
If and article is behind a paywall that UConn does not have a contract for,

- you can ask the library to provide you with a pdf for the article (the library has a fast turnaround time),
- you can try to use GOOGLE to find the article, or
- go to SciHub. As the SciHub servers frequently change, the best way is to use the latest link on Wikipedia. Currently https://sci-hub.se works. The best is to use a document identification number (doi) as a search term.

2. (less than 5 minutes) If your search resulted in multiple matches, click on the link to the PNAS article. You should see a page with the title, the abstract and a listing of similar articles. How many similar articles are linked to this article? (click on the "Similar articles" link in the right-hand bar, scroll down and click on "See all similar articles" link)
   Your answer --->

   Sort the similar articles by publication date (on the top of the page). When was the most recent article published?
   Your answer --->

   How many articles cite the Woese 1977 PNAS paper, how many of these are publicly available? Click on "Cited by" in the right bar, then select "See all Cited by articles".
   Then in the bar on the left hand select "Free Full Text"

Your answers ---> # of citations of Woese's 1977 PNAS paper:
# of citing articles that are publicly available:

3. (15 minutes)
As a student at UConn you have access to different databases (e.g., Scopus). In practice pubmed, and google scholar, usually are all you need.

For a scientist of your choice (e.g., your advisor, or someone who publishes in your field of interest), go to Google scholar to search for his profile – if your choice of scientist doesn't have a profile, use a different scientist. In a google scholar search profiles are marked by an inkpot and a feather.
Which scientist did you choose? How many publications, and how many citations are listed on the profile, and which H-factor are reported. What does the H-factor signify? (if you hover the pointer over the word H-factor in the little table on the right, an definition pops-up)

Your answers ---> name of scientist:
number of publications:
total number of citations:
H-factor (total):
What does the H-factor signify:

Go to Scopus, do an author search (on top of the page) for the same scientist, click on the name of the scientist, How many articles, total citations, citing documents, and which H-factor does Scopus report?
Your answers ---> number of publications:
total number of citations:
number of citing documents:
H-factor (total):

4. (15 minutes)
Using Entrez, search Protein (use drop-down box to select the Protein database) for WP_010886039.1 (this is an accession number, which have replaced the gi numbers previously in use, see historical note)
Run a BLAST search (select "run blast" in the right hand column)
In the form that pops up,
under database select *refseq select proteins* "
under organism enter the taxon *Euryarchaeota (taxid:28890)* (you need to

click on the offered choices after you start typing)
Click on Algorithm Parameters and select 250 Maximum Target sequences
Click on **BLAST**

Once the search is done (the results are displayed piece by piece, you need to wait until the graphics tab shows up),
Select the graphic tab.

Do you notice anything interesting about the alignments? (think intein)

Your answers --->

If you hover the pointer over the red lines in the graphic representation, the name of the organism pops up.

Do all the sequences with the insertion belong to the same genus as the query? If not, which genera possess the insertion?

Your answers --->

Are there any sequences from the same genus as the query among the sequences that do not have the insertion? If yes, give an example.

Your answers --->

Under the "*alignments*" tab, scroll through the first few alignments. Which part seems to be least conserved?

Your answers --->

5. **INTEINS IN PHAGE GENES** (to be completed next week).
For the next exercise we will use the phagesDB database. This database at present contains 5033 genomes from phages (=viruses that infect actinobacteria), and 31225 families of phage proteins.
A long-term goal of the Gogarten Lab is to study the distribution and evolution of inteins in these phages.

A list of intein sequences is in this shared document:
https://docs.google.com/document/d/1Qrcve31lVn2ZHDEPP0X1E624tNh3snDDL5qreSv1vFQ/edit?usp=sharing

Select one of the intein sequences and copy it into your notebook. If you are doing the CURE project, choose the intein assigned to you (the last seven in the listing).

Do a BLASTP search on PhagesDB with "your" intein sequence. Copy the intein sequence (including the annotation line) onto the clipboard of your computer. Go to

phagesDB, click on BLAST in the header, select **blastp**, then copy and paste the intein sequence into the sequence window (the sequence needs to have an annotation line). Select 500 description and alignments. Click on "Blast'
The "top" match should be the sequence (extein + intein) that contains the particular intein. If many identical intein sequences your sequence may not the top one, but is should have the same alignment score.
Write down (or copy) the name of the "top" hit.

Your answers --->
which intein did you choose:

give the annotation of the hit to the phage of interest (include the gene number):

Go to this spreadsheet
https://docs.google.com/spreadsheets/d/1YxJXTiexyXG3hlKdpf2na2TOyKuOHv8jNajrIpA7nsw/edit?usp=sharing
and enter your findings.

A. **The Genome of the phage that contains the gene with the putative intein.** The genes are labeled by the name of the phage (names were picked by the student who isolated the phage) and the number of the gene in the genome. E.g., Alice_147 refers to the 147th Open Reading Frame in the genome of phage Alice.
We want to get the complete protein sequence encoded by this gene. On the home page of phagesDB in the upper left-hand corner, enter the name of the phage and hit return. This brings you to the page for the phage.
Scroll down the page.

How many Genes does the phage genome contain?
Is the phage lytic or lysogenic (it can form a prophage)?
Your answers --->

(Enter this also into the spreadsheet)

B. **Getting the sequence of the intein containing gene and its PHAM.** Scroll down the page and under **gene list** (in the left column) click on *Click to View*. This opens a pull-down table. Scroll down the table to the gene identified in the blastp search with the intein as query. Clicking on the opens the page for this gene. This page contains a list of all genes that were placed in the PHAMilie. **Download the amino-acid sequence encoded by the gene.** Click on *Click to View*, copy the complete aa sequence, and paste it into your notebook. In your notebook add an annotation line in the line above the sequence.

C. We also want to save the sequences from all the Phamily members. As with the gene numbers, PHAMs changes over time, including both the number of the PHAM and the members of the PHAM. It is best to identify things by their sequence. Click on the PHAM number (in a colored highlight), on the PHAM page that opens, click on

Download all sequences. Save the file in your notebook.

What is the PHAM number? How many members were in the PHAM? Which annotations (if any) did these have?

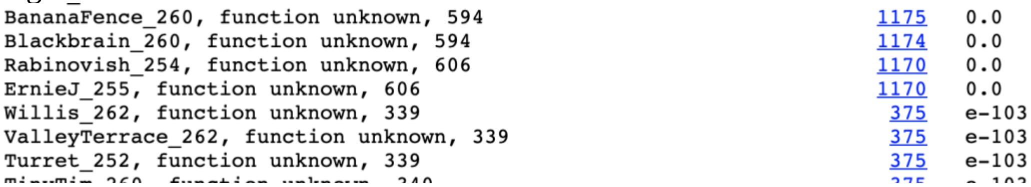
Your answers --->

(Enter this also into the spreadsheet)

D. Go back to the "Gene Details" page of the ORF identified by the blast search with the putative intein. (This is the page that had a Link to the sequence and to the PHAM report (in colored highlight)) On top of this page is a button labelled "*BLAST the gene product on PhagesDB*" click on it
At the bottom of the blast page increase the number of descriptions and alignments to 500.
Click on BLAST.
Do you see the typical pattern with
N-Extein (conserved in all matches) -- Intein (conserved in top matches --C- Extein (conserved in all matches)
in the graphical overview?

Your answers --->

Take a picture (command ctrl shift and the number 4 pressed simultaneously allows you to select part of the screen, which is copied onto the clipboard, paste it into your notebook). If not, a possible reason is that there are too many sequences with intein in the databank, and that the sequences without intein are at the bottom of the list not included in the graphical overview. If this is the case, scroll down the list of matches until you come to a dramatic drop in alignment score (or increase of the e-value). E.g.: _



```
BananaFence_260, function unknown, 594                  1175   0.0
Blackbrain_260, function unknown, 594                   1174   0.0
Rabinovish_254, function unknown, 606                   1170   0.0
ErnieJ_255, function unknown, 606                       1170   0.0
Willis_262, function unknown, 339                        375   e-103
ValleyTerrace_262, function unknown, 339                 375   e-103
Turret_252, function unknown, 339                        375   e-103
```

click on the score of the first sequence that has a lower score (in the example the first 375). This gets to the actual alignment of the query to the target sequence. In case the query contains an intein and target does not, this alignment should have two components (the N and the C-Extein), e.g.:

```
>Willis_262, function unknown, 339
          Length = 339

 Score =  375 bits (963), Expect = e-103
 Identities = 186/187 (99%), Positives = 186/187 (99%)

Query: 1   MGSGAWDSYTYTSHLAAKAAAGKSTFDYTDQIRSGQTSAKANILLDPKVKAGDGSSFAGK 60
           MGSGAWDSYTYTSHLAAKAAAGKSTFDYTDQIRSGQTSAKAN LLDPKVKAGDGSSFAGK
Sbjct: 1   MGSGAWDSYTYTSHLAAKAAAGKSTFDYTDQIRSGQTSAKANSLLDPKVKAGDGSSFAGK 60

Query: 61  VMREVVISDEHPNPTPIAIVLDVTGSNYTAAVAVHAKLPQLFGLLQRKGLIEDPQILIAA 120
           VMREVVISDEHPNPTPIAIVLDVTGSNYTAAVAVHAKLPQLFGLLQRKGLIEDPQILIAA
Sbjct: 61  VMREVVISDEHPNPTPIAIVLDVTGSNYTAAVAVHAKLPQLFGLLQRKGLIEDPQILIAA 120

Query: 121 TGDANSDRVPLQVGQFESDNRIDAMIEAMYLEGFGGGQAHETYELAAYFLARHTYLEPWH 180
           TGDANSDRVPLQVGQFESDNRIDAMIEAMYLEGFGGGQAHETYELAAYFLARHTYLEPWH
Sbjct: 121 TGDANSDRVPLQVGQFESDNRIDAMIEAMYLEGFGGGQAHETYELAAYFLARHTYLEPWH 180

Query: 181 KQGRKGY 187
           KQGRKGY
Sbjct: 181 KQGRKGY 187



 Score =  301 bits (770), Expect = 4e-81
 Identities = 150/151 (99%), Positives = 150/151 (99%)
```

```
Query: 444 IFIGDEKPYDRVKASQVRTHIGVDIEADVPTTQVFEELKEQYEPFFLFQKQGSYSEGQVL 503
           IFIGDEKPYDRVKASQVR HIGVDIEADVPTTQVFEELKEQYEPFFLFQKQGSYSEGQVL
Sbjct: 189 IFIGDEKPYDRVKASQVRAHIGVDIEADVPTTQVFEELKEQYEPFFLFQKQGSYSEGQVL 248

Query: 504 PSWRKLLNEQAVTLEDPNNVCEFIAGLLLLREGGLDLDEVEDELHDAGFDTTAIRSASKT 563
           PSWRKLLNEQAVTLEDPNNVCEFIAGLLLLREGGLDLDEVEDELHDAGFDTTAIRSASKT
Sbjct: 249 PSWRKLLNEQAVTLEDPNNVCEFIAGLLLLREGGLDLDEVEDELHDAGFDTTAIRSASKT 308

Query: 564 LALVGPGGSGGAVAKTDGSLGLDDSTGTDRL 594
           LALVGPGGSGGAVAKTDGSLGLDDSTGTDRL
Sbjct: 309 LALVGPGGSGGAVAKTDGSLGLDDSTGTDRL 339
```

Does the gene likely contain and intein? (*I.e.* Is the ORF recognizably separated into N-extein – intein –C-extein)? (Note: some genes are difficult, because they have been invated by multiple inteins)

Your answer --->

E. Very often genes invaded by an intein and those that were not invaded are placed in different PHAMS even though the extein sequences are identical. We also want to get the sequences from the PHAM without intein. In the result page from the blast search, identify the first match to a putatively intein free target. Note the phage name and gene number. In the above example “Willis” and gene# 262.

Your answers --->

What is the name and gene number of the intein free homolog?

Does it have an annotation, if yes, what?

On the home page of phagesDB, enter the name of this phage on the upper left. On the page of the phage, scroll down to “*Gene List*”. “*Click to view”* and then select the gene number. On the page of the gene the opens, click on the PHAM number (in a colored highlight), on the PHAM page that opens, click on Download all sequences. Save the file in your notebook. AND add a description of what these sequence are (intein free homologs to the exteins invaded by your intein.

Is this PHAM different from the PHAM you downloaded in part C (the PHAM containing the intein harboring genes)?

Do you think the putative intein you started out with is actually an intein?

Your answers --->

Copy your answers into the spreadsheet.

At the end of exercise #5 you should have

A. The putative intein sequence (or part of the intein sequence), *i.e.*, the sequence your started out with; label the file as phage_name_gene#_intein

B. The sequence of the protein that contains this intein (as a single sequence fasta file), label the file as phage_name_gene#

C. A homolog to this sequence that does not harbor the intein (as a single sequence fasta file), label the file as phage_name_gene#

D. The phamily to which the intein containing sequence belongs (as a multiple fasta file)

E. The phamily to which the intein free homolog belongs (as a multiple fasta file)