# BIO353 Lab 4: Online Research using (other peoples) BIG DATA

## Objectives

1. Learn how online genomic and data visualization resources can be used to answer biology research questions and make novel discoveries.
2. Experience how a research project can be broken down into smaller questions that can be answered experiments or by finding and analysing publicly available data.
3. Generating information for your final poster presentation.

## Long blurb that is important to read

Greetings developmental biologists, welcome to the lab. Today you are a virtual grad student again (yay). Your virtual selves performed proteomic work and were successfully able to isolate the sequence of your protein! The sequence of your protein can be seen below (next page, choose protein depending on your screen). The amino acids in the polypeptide chain are denoted by their letter codes.

Now that you have your protein there are so many molecular methods available to you and so many questions you can answer! The list is almost endless, and so we're going to direct you towards a few questions during these three hours

You are going to answer some of the above questions in the next 3 hours. Bonus: You will not need to know any code or learn a programming language (yay again). This lab will take you through the steps using the online resources at your disposal. These resources are tools commonly used by plant science (and some more generally in biology) researchers (academic and industrial) and you will often be expected to be familiar with them and their use.

1. Work in pairs (3 max)
2. **<u>At the end of each step (question section) talk to the TA/Prof to confirm that you have completed it successfully and discuss the implications of your discoveries before you move to the next step. Feel free to ask questions at any other point too.</u>**
3. Copy answers into this document as you go along as this document is marked and contributes to your grade.
4. There are important questions that we don't guide you through in this document, so be sure to ask how to use these tools to answer and other questions you want to address in your final presentation!
    - Where is this gene expressed in other plants?
    - How is plant development changed when you knock these genes out?
    - How is plant development changed when you over-express these genes?
    - What resources (mutants, over-expression lines, transcriptional reporters, etc.) are available for me to study the role of this gene?
    - How is the expression of other genes affected by this gene?
    - What is known about potential protein-protein interactions?

Auxin

1 MQAVKRSRRH VEEEPTMVEP KTKYDRQLRI WGEVGQAALE EASICLLNCG
51 PTGSEALKNL VLGGVGSITV VDGSKVQFGD LGNNFMVDAK SVGQSKAKSV
101 CAFLQELNDS VNAKFIEENP DTLITTNPSF FSQFTLVIAT QLVEDSMLKL
151 DRICRDANVK LVLVRSYGLA GFVRISVKEH PIIDSKPDHF LDDLRLNNPW
201 PELKSFVETI DLNVSEPAAA HKHIPYVVIL VKMAEEWAQS HSGNLPSTRE
251 EKKEFKDLVK SKMVSTDEDN YKEAIEAAFK VFAPRGISSE VQKLINDSCA
301 EVNSNSSAFW VMVAALKEFV LNEGGGEAPL EGSIPDMTSS TEHYINLQKI
351 YLAKAEADFL VIEERVKNIL KKIGRDPSSI PKPTIKSFCK NARKLKLCRY
401 RMVEDEFRNP SVTEIQKYLA DEDYSGAMGF YILLRAADRF AANYNKFPGQ
451 FDGGMDEDIS RLKTTALSLL TDLGCNGSVL PDDLIHEMCR FGASEIHVVS
501 AFVGGIASQE VIKLVTKQFV PMLGTYIFNG IDHKSQLLKL

Cytokinin

1 MMGSVELNLR ETELCLGLPG GDTVAPVTGN KRGFSETVDL KLNLNNEPAN
51 KEGSTTHDVV TFDSKEKSAC PKDPAKPPAK AQVVGWPPVR SYRKNVMVSC
101 QKSSGGPEAA AFVKVSMDGA PYLRKIDLRM YKSYDELSNA LSNMFSSFTM
151 GKHGGEEGMI DFMNERKLMD LVNSWDYVPS YEDKDGDWML VGDVPWPMFV
201 DTCKRLRLMK GSDAIGLAPR AMEKCKSRA

NPA

1 MQKRIALSFP EEVLEHVFSF IQLDKDRNSV SLVCKSWYEI ERWCRRKVFI
51 GNCYAVSPAT VIRRFPKVRS VELKGKPHFA DFNLVPDGWG GYVYPWIEAM
101 SSSYTWLEEI RLKRMVVTDD CLELIAKSFK NFKVLVLSSC EGFSTDGLAA
151 IAATCRNLKE LDLRESDVDD VSGHWLSHFP DTYTSLVSLN ISCLASEVSF
201 SALERLVTRC PNLKSLKLNR AVPLEKLATL LQRAPQLEEL GTGGYTAEVR
251 PDVYSGLSVA LSGCKELRCL SGFWDAVPAY LPAVYSVCSR LTTLNLSYAT
301 VQSYDLVKLL CQCPKLQRLW VLDYIEDAGL EVLASTCKDL RELRVFPSEP
351 FVMEPNVALT EQGLVSVSMG CPKLESVLYF CRQMTNAALI TIARNRPNMT
401 RFRLCIIEPK APDYLTLEPL DIGFGAIVEH CKDLRRLSLS GLLTDKVFEY
451 IGTYAKKMEM LSVAFAGDSD LGMHHVLSGC DSLRKLEIRD CPFGDKALLA
501 NASKLETMRS LWMSSCSVSF GACKLLGQKM PKLNVEVIDE RGAPDSRPES
551 CPVERVFIYR TVAGPRFDMP GFVWNMDQDS TMRFSRQIIT TNGL

# Question 1: What is known about the domains of this protein?

Open up the internet browser of your choice, you are going to BLAST! Go to the NCBI BLAST page: http://blast.ncbi.nlm.nih.gov/Blast.cgi

According to the Biology Curriculum Map, you've used NCBI BLAST other courses. If not, do not fret. You are given some choices based on what you want to BLAST, within each you will find choices of different algorithms to use depending on your query and goal. We have a complete protein sequence to query with and we're going to use "protein blast". Select this from within the five "basic BLAST" options:

| | |
|---|---|
| nucleotide blast = | Search a nucleotide database using a nucleotide query |
| protein blast = | Search protein database using a protein query |
| blastx = | Search protein database using a translated nucleotide query |
| tblastn = | Search translated nucleotide database using a protein query |
| tblastx = | Search translated nucleotide database with translated nucleotide query |

In the web page that appears copy and paste your protein sequence (above) into the box entitled "Enter accession number(s), gi(s), or FASTA sequence(s)". This time around we don't need to choose any further options or different algorithms, just click on the blue "BLAST" button! **In the "graphic Summary" tab there are 1 or 2 "superfamily domains" in your protein – what are they? What does this tell you about your gene function?**

**Answer:**

**The description tab lists sequences with significant alignments. Is the gene you found in last week's tutorial the top entry? Top 5? Are the results from the BLAST as you expected?**

**Answer:**

**What other species are in the top 5 match results from your protein blast?**

**Answer:**

**Select the top 5 results and then click on the taxonomy tab. How are these species related? (use specific taxonomic words)**

**Answer:**

## Question 2: OK, so what do we know about these genes from existing research?

Remember TAIR? Let's return there: www.arabidopsis.org (or via the library). Before you proceed **ask your TA/Prof to confirm the gene** you are searching with from the end of the last section. In the search box at the top, type in the 4-character gene identifier you found in the last step (leave the search setting on "Gene") and click "search". **What's the Locus number (AGI) for this gene? What do these numbers tell you about the locus?**

**Answer:**

Follow the link to the locus page. We can learn a great deal from this page. As you know we can use it to identify and purchase mutant seeds, find publications, genomic and coding sequences, and gene ontologies (GO) for the gene in question. GO terms can give you lots of info about the processes a gene is likely involved in (including development), the subcellular location and the molecular function of the protein. Publications are going to answer a lot of the questions we posed in the introduction, e.g. "What is the loss/gain-of-function phenotype?", and so on. **What is the description for this gene?**

**Answer:**

# Question 3: Where is my gene expressed and where is the protein localized?

You will likely get very reliable information answering this question from the publications listed on the locus page. This could include *in-situ* hybridization studies (telling you exactly where the mRNA is on fixed and prepared samples), as well as visible reporters of transcription (e.g. promoter::GFP fusions) and the protein products (promoter::coding sequence-GFP), which may tell you what your gene/gene product are up to in real-time. However, because the published approaches done by other groups might not be completely exhaustive (e.g. testing expression under lots of different circumstances/treatments), and because we want to show you some other cool stuff, we are getting you to try out an eFP browser. An eFP browser gives you a visual output summarizing multiple global expression analysis experiments (hundreds of microarray or RNA-seq experiments) that is easy to interpret quickly without requiring expert knowledge. We'll look at a couple of other tools here too.

Copy the locus number for the gene from question 3. Go to the BAR… not that bar, this one: http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi Paste the locus number into the 'Primary Gene ID' search box, or type in the 3 letter + 1 number gene name, with "Data Source = Development Map" and "Mode = Absolute" and click search. **Where and at what level was the expression of your gene highest, and what was the SD?**

**Answer:**

This is not the maximum level of expression seen for your gene though. Check out the 'tissue specific' map where data from laser micro-dissection or fluorescence-activated cell sorting experiments are shown. **Where can the highest levels of expression be found on this map? Why might we get different numbers in smaller tissue samples?**

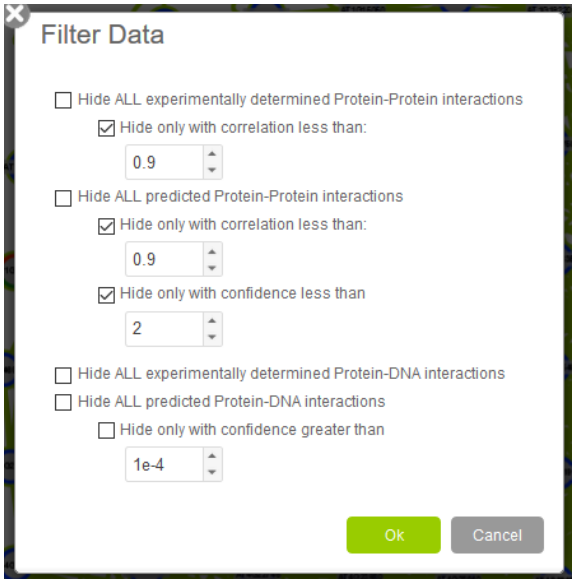
**Answer:**

Coolaboola, now click on the Bar homepage icon (top left hand corner), scroll down to "Gene Expression and Protein tools" and click on ePlant (new version). Input your gene name abbreviation '_ _ _ _'. When that is loaded click on the 'Cell eFP'. **Where is your protein localized in the cell?**

**Answer:**

## Question 4: Finding new targets: Are other genes expressed in a similar pattern to my gene and are other proteins known to interact with mine?

Now, go to the "Interaction viewer" and select that. The display will show predicted and experimentally supported interactions between your protein and other proteins or DNA. As it turns out, the gene you've selected have been researched extensively and so there are lots of known interactions. To narrow down the search to the strongest interactions click on the "filter" icon  at the top of the page and set the box parameters as seen below. Press Ok and you should have ~20 results left. Hovering over each result will gives you a brief summary (locus number, aliases, annotation) . Look through the results for those that seem directly related to your gene.



**How many results are related to auxin signaling? Does this make sense for the screen you performed? What might explain this?**

**Answer:**

While all of these interactions (auxin-related or not) can lead to future cool research, we only have a limited amount of time and resources. Which genes are the most relevant / interesting for you to explore? **Write down the locus number and annotations for 5 genes that you think have the most important /relevant interaction with your gene. Why did you select these?**

**Answer:**

OK, so there maybe some candidates for interaction with our protein that researchers already know about, let's try to find some new ones! Similar expression patterns of genes might reflect similar regulation of those genes and involvement in the same developmental pathways. Click on the top left-hand corner of your gene in the left hand panel and select "top 5 responses" and hit 'search'. These genes were identified based on the similarity of their expression patterns with your gene of interest the tissue-specific experiments (literally hundreds of samples characterised by microarrays or RNAseq, by a bunch of different research groups). **Record the locus number of the 5 genes below, as well as a BRIEF annotation/description provided by TAIR (1 sentence). Comment on these results: do you think the similar expression pattern is a product of chance? (1-2 sentences)**

**Answer:**

**Which of the genes on the list were already identified in the interaction viewer? Does this surprise you?**

**Answer:**