

# 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 this intriguing protein can be seen below. 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.

Notes on the lab:

1. Work in pairs (3 max)
2. **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.**
3. Copy answers into this document as you go along as this document is for marks.
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!
  - a. Where is this gene expressed in other plants?
  - b. How is plant development changed when you knock these genes out?
  - c. How is plant development changed when you over-express these genes?
  - d. What resources (mutants, over-expression lines, transcriptional reporters, etc.) are available for me to study the role of this gene in pest/disease resistance or abscission in Arabidopsis?
  - e. How is the expression of other genes affected by this gene?
  - f. What is known about potential protein-protein interactions?

Auxin **AXR1**

1 MQAVKRSRRH VEEPTMVEP KTKYDRQLRI WGEVGQAAL EASICLLNCG  
51 PTGSEALKNL VLGGVGSITV VDGSKVQFGD LGNNFMVDAK SVGQSKAKSV  
101 CAFLQELNDS VNAKFIEENP DTLITTNPFS FSQFTLVIAT QLVEDSMLKL  
151 DRICRDANVK LVLVRSYGLA GFVRISVKEH PIIDSKPDHF LDDLRLNNPW  
201 PELKSFVETI DLNVSEPAAL 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 AANYNKFPQG  
451 FDGGMEDIS RLKTTALSLL TDLGCNGSVL PDDLIHEMCR FGASEIHVVS  
501 AFVGGIASQE VIKLVTKQFV PMLGTYIFNG IDHKSQLLKL

Cytokinin **AXR3**

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

## NPA TIR1

1 MQKRIALSFP EEVLEHVFSF IQLDKDRNSV SLVCKSWYEI ERWCRRKVI  
51 GNCAVSPAT VIRRFKVRV VELKGKPHFA DFNLVPDGGW GYVYPWIEAM  
101 SSSYTWLEEI RLKRMVVTDD CLELIAKSF NFKVLVLSSC EGFSTDGLAA  
151 IAATCRNLKE LDLRESDVDD VSGHWLSHFP DTYTSLVSLN ISCLASEVSF  
201 SALERLVTRC PNLKSLKLN AVPLEKLATL LQRAQLEEL GTGGYTAQVR  
251 PDVYSGLSVA LSGCKELRCL SGFWDAVPAY LPAVYVCSR LTTNLNSYAT  
301 VQSYDLVKLL CQCPKLQRLW VLDYIEDAGL EVLASTCKDL RELRVFPSEP  
351 FVMEPNVALT EQGLVSVMG CPKLESVLYF CRQMTNAALI TIARNRPNMT  
401 RFRLCIIEPK APDYLTLEPL DIGFGAIVEH CKDLRRLSLS GLLTDKVFY  
451 IGTAKKMEM LSAFAGDSD LGMHHVLSGC DSLRKLEIRD CPFQDKALLA  
501 NASKLETMR 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 (axr1/3) or 2(tir3) "superfamily domains" in your protein – what are they? What does this tell you about your gene function? 2 marks**

Answer:

**AXR1 – E1\_enzyme\_family superfamily**

**AXR3 – AUX\_IAA superfamily**

**TIR1 – two AMN1 superfamily domains and an F-box superfamily domain**

Talk to student about protein domains. Suggested questions to ask:

What is meant by a protein domain? Are particular domains associated with particular functions? Do proteins which share the same domains perform the same function?

1 mark for correct answer, 1 mark for discussion about gene function

**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? 2 marks**

**Answer:**

**Tricky this, because for all genes the most likely candidate isn't the top result. I'm going to be a little lax (hopefully they focused on the proper genes from the past tutorial)**

<input checked="" type="checkbox"/>	<a href="#">AUX/IAA transcriptional regulator family protein [Arabidopsis thaliana]</a>	474	474	100%	6e-169	100.00%	<a href="#">NP_171921.1</a>
<input checked="" type="checkbox"/>	<a href="#">putative auxin-induced protein IAA17/AXR3-1 [Arabidopsis thaliana]</a>	471	471	99%	4e-168	100.00%	<a href="#">AAM64837.1</a>
<input checked="" type="checkbox"/>	<a href="#">IAA17/AXR3-1 protein [Arabidopsis thaliana]</a>	470	470	100%	1e-167	99.56%	<a href="#">AAC39440.1</a>
<input checked="" type="checkbox"/>	<a href="#">hypothetical protein ARALYDRAFT_470401 [Arabidopsis lyrata subsp. lyrata]</a>	466	466	100%	8e-166	97.38%	<a href="#">EFH65758.1</a>
<input checked="" type="checkbox"/>	<a href="#">auxin-responsive protein IAA17 [Arabidopsis lyrata subsp. lyrata]</a>	467	467	100%	2e-165	97.38%	<a href="#">XP_002889499.2</a>

<input checked="" type="checkbox"/>	<a href="#">F-box/RN1-like superfamily protein [Arabidopsis thaliana]</a>	1221	1221	100%	0.0	100.00%	<a href="#">NP_567135.1</a>
<input checked="" type="checkbox"/>	<a href="#">LOW QUALITY PROTEIN: protein TRANSPORT INHIBITOR RESPONSE 1 [Arabidopsis lyrata subsp. lyrata]</a>	1202	1202	100%	0.0	98.15%	<a href="#">XP_002878489.2</a>
<input checked="" type="checkbox"/>	<a href="#">protein TRANSPORT INHIBITOR RESPONSE 1 [Capsella rubella]</a>	1197	1197	100%	0.0	97.47%	<a href="#">XP_006290792.1</a>
<input checked="" type="checkbox"/>	<a href="#">PREDICTED: protein TRANSPORT INHIBITOR RESPONSE 1 isoform X1 [Camelina sativa]</a>	1194	1194	100%	0.0	97.31%	<a href="#">XP_010413212.1</a>
<input checked="" type="checkbox"/>	<a href="#">PREDICTED: protein TRANSPORT INHIBITOR RESPONSE 1 [Camelina sativa]</a>	1192	1192	100%	0.0	96.80%	<a href="#">XP_010512696.1</a>

<input checked="" type="checkbox"/>	<a href="#">unnamed protein product [Arabidopsis thaliana]</a>	1116	1116	100%	0.0	99.81%	<a href="#">VYS45060.1</a>
<input checked="" type="checkbox"/>	<a href="#">NAD(P)-binding Rossmann-fold superfamily protein [Arabidopsis thaliana]</a>	1116	1116	100%	0.0	100.00%	<a href="#">NP_172010.1</a>
<input checked="" type="checkbox"/>	<a href="#">NEDD8-activating enzyme E1 regulatory subunit AXR1 isoform X1 [Arabidopsis lyrata subsp. lyrata]</a>	1072	1072	100%	0.0	95.56%	<a href="#">XP_020867624.1</a>
<input checked="" type="checkbox"/>	<a href="#">PREDICTED: NEDD8-activating enzyme E1 regulatory subunit AXR1 [Camelina sativa]</a>	997	997	99%	0.0	89.44%	<a href="#">XP_010475191.1</a>
<input checked="" type="checkbox"/>	<a href="#">PREDICTED: NEDD8-activating enzyme E1 regulatory subunit AXR1 [Camelina sativa]</a>	996	996	99%	0.0	88.89%	<a href="#">XP_019090757.1</a>

**1 mark for "yes this is as expected" , 1 mark for a reasonable hypothesis for why not the first entry**

**What other species are in the top 5 match results from your protein blast? 1 mark**

**Answer:**

**AXR1 – Arabidopsis thaliana, Arabidopsis lyrata, Camelina sativa**

**AXR3 – Arabidopsis thaliana, Arabidopsis lyrata,**

**TIR1 – Arabidopsis thaliana, Arabidopsis lyrata, Capsella rubella, Camelina sativa**

**Select the top 5 results and then click on the taxonomy tab. How are these species related? (use specific taxonomic words) 2 marks**

**Answer:**

**AXR1 – They are all part of the Camelineae subpopulation under the Brassicaceae family**

**AXR3 – They are all part of the Arabidopsis genus (questions about cloning vectors)**

**TIR1** – They are all part of the Camelinae subpopulation under the Brassicaceae family

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

Remember TAIR? Let's return there: [www.arabidopsis.org](http://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?** **3 marks**

**Answer:**

**Species, chrom #, gene, something to do with descending order of loci on the chromosome.**

**AXR1 – AT1G05180**

**AXR3 – AT1G04250**

**TIR1 -- AT3G62980**

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?** **1 mark**

**Answer:**

**AXR1 –**

**AXR3 –**

**TIR1 --**

### 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? (1 mark)**

Answer:

**AXR1 – 24h Imbibed seed 325.04 SD = 0.78**

**AXR3 – Hypocotyl 782.3 SD = 34.95**

**TIR1 – VEGETATIVE ROSETTE 259.75 SD 7.67**

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? (3 marks)**

Answer:

**AXR1 – basal section of embryo**

**AXR3 – stigma and ovaries**

**TIR1 – rib meristem of the SAM**

## 2 marks for understanding why might we get higher number in smaller tissue

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? 1 mark**

**Answer:**


**AXR1 – nucleus**

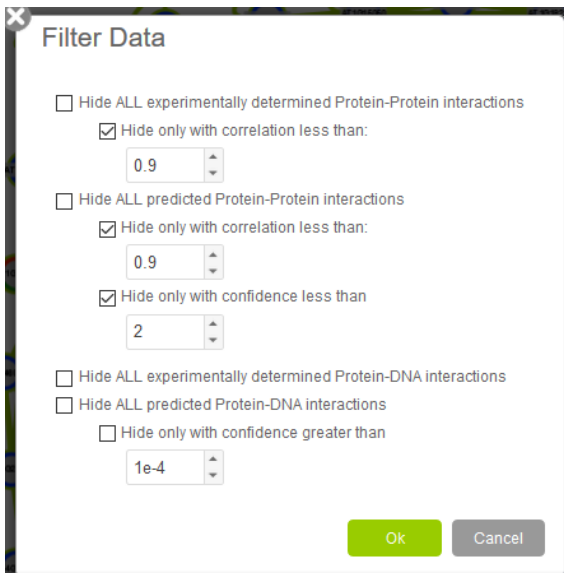
**AXR3 – nucleus**

**TIR1 – nucleus**

## 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 give you a brief summary (locus number, aliases, annotation) . Look through the results for those that seem directly related to your gene.



**Filter Data**

☐ Hide ALL experimentally determined Protein-Protein interactions  
☒ Hide only with correlation less than:

☐ Hide ALL predicted Protein-Protein interactions  
☒ Hide only with correlation less than:  
  
☒ Hide only with confidence less than

☐ Hide ALL experimentally determined Protein-DNA interactions  
☐ Hide ALL predicted Protein-DNA interactions  
☐ Hide only with confidence greater than

**Ok** **Cancel**

**How many results are related to auxin signaling? Does this make sense for the screen you performed? What might explain this? 5 marks**

Answer:

**AXR1 – 0**

- Surprising
- Why? Might be that info isn't great

**AXR3 – ~20**

**TIR3 – ~7**

- Not surprising
- Even though screen was not auxin related, auxin affect lot so things (including our phenotype of interest)

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? 3 marks**

Answer:

- **AXR1 –**
  - Honestly I'm just interested in what they respond with
- **AXR3 –**
  - AT1G04550 – aux/iaa transcriptional regulator family protein
  - AT1G04240 – AUX/IAA transcriptional regulator family protein
  - AT1G19850 – transcriptional factor B3 family protein / auxin-responsive factor AUX/IAA-related
  - AT2G33310 - auxin-induced protein 13
  - AT3G61830 - auxin response factor 18
  - AT4G23980 – AUXIN RESPONSE FACTOR 9
  - AT5G20730 - Transcriptional factor B3 family protein / auxin-responsive factor AUX/IAA-related
  - AT5G43700 - AUX/IAA transcriptional regulator family protein
  - AT5G25890 - indole-3-acetic acid inducible 28
  - AT4G14560 - indole-3-acetic acid inducible
  - AT3G04730 - indole-3-acetic acid inducible 16
  - AT3G15540 indole-3-acetic acid inducible 19
  - AT3G17600 - indole-3-acetic acid inducible 31
  - AT3G23030 indole-3-acetic acid inducible 2
  - AT2G22670 - indole-3-acetic acid inducible 8
  - AT1G52830 - indole-3-acetic acid 6



- AT1G15580 - indole-3-acetic acid inducible 5
- AT1G15050 - indole-3-acetic acid inducible 24
- AT1G15050 - indole-3-acetic acid inducible 34
- AT1G04100 - indoleacetic acid-induced protein 10
- 
- **TIR1 –**
  - AT1G04240 - AUX/IAA transcriptional regulator family protein
  - AT1G04250 - AUX/IAA transcriptional regulator family protein
  - AT1G04550 - AUX/IAA transcriptional regulator family protein
  - AT1G15580 - indole-3-acetic acid inducible 5
  - AT2G22670 - indole-3-acetic acid inducible 8
  - AT4G14560 - indole-3-acetic acid inducible
  - AT5G25890 - indole-3-acetic acid inducible 28

OK, so there are 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) 5 mark**

**Answer:**

**In general, I want them to understand that even if the function is irrelevant, its not necessarily an accident. Fully marks for some kind of guess as to how their gene COULD affect these genes.**

**2 marks for good annotations, 3 marks for a comment on why similar expression levels may/may not be due to chance**

**AXR1 – (all like 0.9 and up)**

- At1g03140 - PRP18a is one of two paralogs (the other being PRP18b) which are highly similar to the step II splicing factors in yeast. Loss of function mutations show defects in alternative splicing, mostly intron retention event
- At5g10710 - centromere protein

- At5g02050 - Mitochondrial glycoprotein family protein;
- At3g09650 - RNA binding protein involved in the processing of chloroplast psbB-psbT-psbH-petB-petD transcriptional UNIT
- At5g36950 - Encodes a putative DegP protease

**AXR3 – (note that none are above 0.8, all around 0.6)**

- At2g37170 - a member of the plasma membrane intrinsic protein subfamily PIP2. localizes to the plasma membrane and exhibits water transport activity in *Xenopus* oocyte. expressed specifically in the vascular bundles and protein level increases slightly during leaf development.
- At5g17330 - Encodes one of two isoforms of glutamate decarboxylase. The mRNA is cell-to-cell mobile
- At4g23690 - Encodes a homodimeric all-beta dirigent protein in the superfamily of calycins. Dirigent proteins impart stereoselectivity on the phenoxy radical coupling reaction yielding optically active lignans from two molecules of coniferyl alcohol
- At2g25810 - tonoplast intrinsic protein
- At1g09560 - Encodes a plasodesmata-located protein involved in regulating primary root growth by controlling phloem-mediated allocation of resources between the primary and lateral root meristems. The mRNA is cell-to-cell mobile

**TIR1 – .86-.90**

- At3g11830 - TCP-1/cpn60 chaperonin family protein
- At5g01020 - Protein kinase superfamily protein
- At2g41620 - Nucleoporin interacting component
- At5g16505 - Encodes a member of a domesticated transposable element gene family MUSTANG. Members of this family are derived from transposable elements genes but gained function in plant fitness and flower development
- At3g55460 - encodes an SC35-like splicing factor that is localized to nuclear specks.

**Which of the genes on the list were already identified in the interaction viewer?**

**Does this surprise you? 1 mark**

**Answer:**

**AXR1 – none**

**AXR3 – none**

**TIR3 – none**

