Practical Bioinformatics for Biologists (BIOS441/641) Course Project 1

Project title:
Sequence and expression analysis of glycosyltransferase family 2 (GT2) enzymes in Arabidopsis

Goals:
1) Practice databases and web-based tools that we have learned

2) Learn how to design a bioinformatics workflow to answer biology/evolution questions

3) Learn how to identify useful tools, datasets and existing knowledge in the research papers to help design the workflow

Background:
There are 9 CesA/Csl families (see Figure 1), all belonging to a same protein superfamily, meaning that they share a common ancestor but with low sequence similarity. They are important because they encode enzymes sitting on the plasma membrane or Golgi membrane for the synthesis of polysaccharides, glycoproteins, glycolipids and other glycol-molecules in all organisms. They are particularly important for plants because plants are carbohydrate-rich. Previous studies have shown that CesA is for cellulose synthesis, CslA for mannan synthesis, CslC for xyloglucan synthesis, CslF and CslH both for mixed linkage glucan synthesis. All these polysaccharides are the major components of plant cell walls, the dry materials of plants, the food for animals and the feedstock for the cellulosic biofuels.

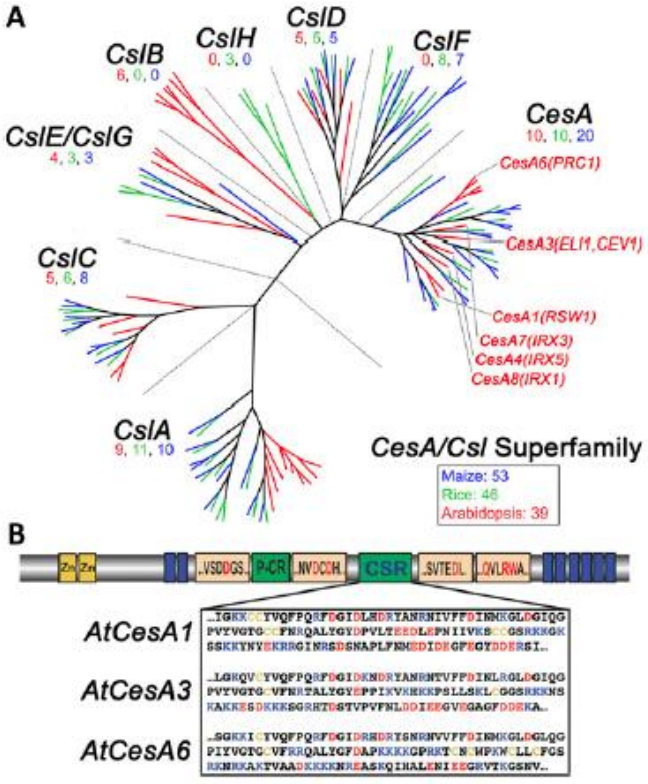

***Figure 1**: (A) nine CesA/Csl families in three plants; (B) domain diagram of CesA proteins, with known sequence motifs shown.*

Detailed Figure legend is available in the reference below

Plant Physiol. 2011 January; 155(1): 171–184
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3075763/pdf/171.pdf

This project is to study these Csl families in Arabidopsis in order to answer the following questions:

1. How are these families phylogenetically related (Figure 1 has already addressed this but we want to build our own phylogeny)? Put the family you are responsible in the context.
2. What conserved sequence motifs are shared between your family and a structurally characterized protein (visualize and examine the alignments, check literature)?
3. What is the expression profile of genes in your family (microarray data analysis and presentation)? For example, which genes are the most highly expressed in which tissue by looking at the expression profile graph?
4. Do green algae have CesA/Csl genes, how are they related to the Arabidopsis ones and do they have the conserved motifs too (BLAST and phylogeny clustering)?

Detailed steps:

1. Get a list of GT2 accession numbers from CAZy website (www.cazy.org)
   - Browse the CAZy website and check the NAR paper (http://www.ncbi.nlm.nih.gov/pubmed?cmd=search&term=18838391) to describe what CAZy is about
   - Find Arabidopsis thaliana genome page and copy the table to excel
   - Sort the excel based on family col.
   - Copy **only the GT2 rows** to a new excel sheet; this is **Table 1** for the report. The protein name column has the AGI (Arabidopsis gene identifier) ID. Create a new column named AGI in the table and put the AGI IDs in the col. Use http://www.online-utility.org/text/grep.jsp to extract the AGI if needed.
   - From **Table 1**, copy the AGI IDs and use UniProt website's Retrieve/ID Mapping tool to retrieve the gene names info.
   - Just keep the Reviewed entries and download the fasta sequences for these entries. Add the Gene name (the bold name) to **Table 1**. Sort the table based on this gene name col.

2. Edit the fasta sequence file to replace the description line with the gene name from **Table 1**. This is the **Data S1** for the report.

3. Take **Data S1** as input to build phylogeny at phylogeny.fr and locate your responsible family (see below table) for subsequent analyses. Save the radial view of phylogram as an image, which is **Figure 1** for the report.

Family	# of proteins	Students
CesA	10	Catherine, Sean, Preethi
CslA	9	Xueqiong, Jessica, Stephan
CslB	6	Lauren, Jay, Sandesh, Tim
CslC	5	Nipin, Hattie, Sarah, Beckymarie
CslD	5	Aleisha, Jason, Daniela
CslE/G	4	Hannah, Amol, Tegan, Kyrsten

4. Search and download GSE607 expression data matrix at GEO database
   - Find the GPL file associated with this GSE data series, which contains the AGI ID to probe ID mapping info and other annotation info. Change the file extension to xls so that excel can open it. This is **Data S2** for the report. From **Data S2** you will be able to match AGI ID to probe ID. Add this probe ID info as a new col to **Table 1 (just for your family)**
   - Download GSE607 series matrix file, which contains the expression values. The file is a zipped file with .gz extension. You need install 7-Zip (http://www.7-zip.org/) to uncompress it. Or you can

get the already unzipped file at http://cys.bios.niu.edu/yyin/teach/PBB/GSE607_series_matrix.txt. Change the file extension to xls so that excel can open it. This is **Data S3** for the report. From this file you can get the expression values for each gene

- **Just for your family:** Use **Table 1** to find the probe IDs and extract expression values from **Data S3** and put them in a new excel sheet as **Table 2** for the report. Also include the gene name and AGI ID as two new cols to this **Table 2**. Header line includes the GSM IDs and indicates the tissue info, e.g. leaf (GSM9223). Discard genes that do not have expression values in Data S3.
- **Just for your family:** Plot expression profiles for genes in your family. This is **Figure 2** for the report (see example at http://cys.bios.niu.edu/yyin/teach/PBB/example-exp.xlsx)

5. Build alignment

Just for proteins of your responsible family: copy the sequences from **Data S1** to a separate plain text file; also include a bacterial protein as the first sequence in the file: search 4HG6_A using NCBI entrez and select protein database. This is **Data S4** for the report. Build the alignment at EBI MAFFT server and save it in plain text file, which is **Data S5** for the report. Use ESpript server to visualize **Data S5**, save it as a PDF, which is **Figure 3** for the report. Also save the sequence identity matrix at EBI showing show the pairwise sequence identity. This is **Table 3** for the report.

6. Look for conserved motifs: check Figure 2a of the Nature paper (http://cys.bios.niu.edu/yyin/teach/nature11744.pdf, which reported PDB structure of 4HG6_A). Locate the seven motifs in **Figure 3** and then in **Data S5**, copy paste these regions to WebLogo server and make sequence logo for each of these motifs. Create **Table 4** with three cols: motif position, motif seq in 4HG6_A (bacteria CesA) and motif sequence logo in your plant Csl family.

7. Take the **first plant sequence** in **Data S4** and TBLASTN search against TSA database, select charophytes (taxid: 3146) as organism for homologs at NCBI Blast server. Take the best hit mRNA. Extract its fasta sequence (nucleotide) and use its species name as the ID for this seq. Put the fasta sequence in a plain text file as **Data S6.**

8. Translate nucleotide sequence using transeq of EMBOSS into amino acid sequence and determine the correct open reading frame (hit: has fewest *). Also verify the correct from by checking the alignment of the query protein and the hit mRNA in step 7. From the determined frame in transeq output, copy the longest string of letters between two *s and put them in a plain text file as **Data S7**; also need a description line for creating the fasta format: use the determined frame's description line.

9. Concatenate **Data S7** and **Data S1** as **Data S8.** Use it as input to build phylogeny using MEGA. Determine what family does this charophyte algal seq belong to or is closest to? This is **Figure 4**.

10. Use Needle to align this charophyte algal seq and 4HG6_A to determine if the motifs in step 6 exist in the algal seq? Save the needle result as **Data S9.**

Check list for your report:

Figure 1: Phylogeny of all Arabidopsis Csl/CesA proteins
Figure 2: Expression profiles of genes of your responsible Csl family
Figure 3: Sequence alignment of Csl proteins (your family) plus 4HG6_A
Figure 4: Phylogeny of all Arabidopsis Csl/CesA proteins plus one charophyte algal protein
Table 1: Arabidopsis GT2 GenBank accessions, AGIs, gene names, probe IDs, etc.
Table 2: Microarray expression values of Csl proteins (your family)
Table 3: Sequence identity matrix for Csl proteins (your family) plus 4HG6_A
Table 4: Sequence logo of seven motifs in Csl proteins (your family)

Data S1: Fasta sequences of all Arabidopsis CesA/Csl proteins
Data S2: GPL xls file from GEO
Data S3: GSE data matrix xls file from GEO
Data S4: Fasta sequences of proteins of your responsible family plus 4HG6_A
Data S5: MSA of S4 from MAFFT
Data S6: Fasta sequence of charophyte mRNA from TBLAST search
Data S7: Fasta sequence of peptide translated from S6
Data S8: S1+S7
Data S9: Needle output of aligning 4HG6_A vs S7

Report format:
Introduction: Use one sentence to explain CAZy, GT2, CesA, Csl, respectively. See references above. Explain what you want to study in this project: see questions to be answered above.

Methods: Feel free to use the above detailed steps with your own modification/addition

Results: Explain in details each Figure, Table and Data that you have made. What is the data in there and what the data tells you?

Conclusion: Use one sentence to answer each of the questions above.

Include figures and tables in the report text. Send Data files together with the report. Screen shots are encouraged but not required.

Report should be in word doc file.

Check out one example from previous student Bill Wysocki:
http://cys.bios.niu.edu/yyin/teach/PBB/WysockiProject1Writeup.pdf

Report is due through email by the midnight of 11/05. Total 15 points.

Use office hours or email me for questions.

DO NOT COPY OTHERS WORK!!!