Practical Bioinformatics for Biologists (BIOS493) Course Project 2

Project title:
Data mining green algae SRA reads for Cellulose synthase-like (Csl) genes

Two groups:
Group 1 (*Chara vulgaris*): **Bill, Brenda, Steve, Tom**
Group 2 (*Spirogyra pratensis*): **Jenny, Matt, Shannon, William**

Task assignment: Within each group, you may discuss who will be responsible for which steps of analysis (see detailed method steps below) based on your skills or interests. For example, each group has at least one person able to program in perl or python (yes you may use python if you want to). These people may want to take care of writing the needed scripts.

Project presentation: There will also be a presentation time in the class on **May 02** for each group to present the project. Each group may elect one person to do the presentation in 20-30 minutes. In the presentation, you will tell me who contributed which analyses of the project.

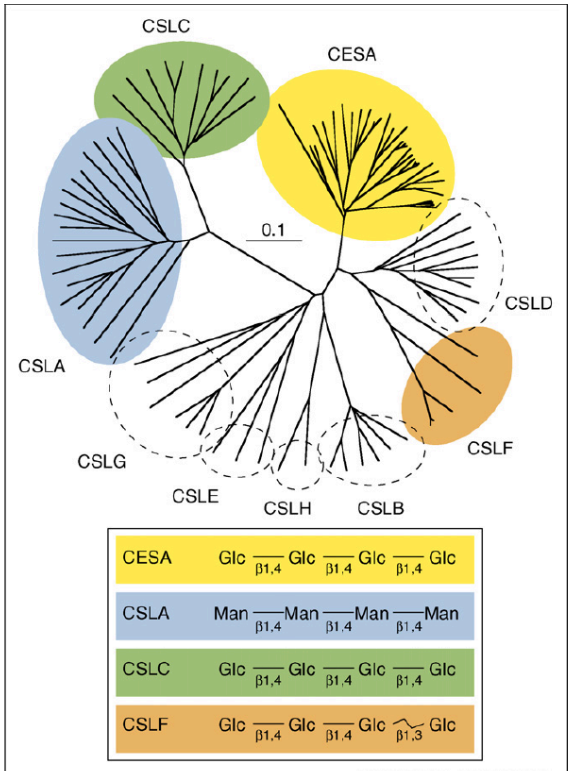
Project report: At the end of this class (**May 7th**), each student will submit a project report, summarizing your contributions to the project as well as the overall design, the research background (literature), the detailed methods, the major findings (with figures and tables) and the discussion of results. Each student should create a folder under your home@glu named project2 and put your files (data, tools, scripts, intermediate files and history files there). You may choose to do all analyses on your own Ubuntu computer, but make sure to copy things to project2 folder@glu after May 7th. I will check your folder when I grade your reports.

Goals:
1) Practice the bioinformatics knowledge and computing skills by applying them to a research project, which might lead to novel research findings

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

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

Introduction:
What: Cellulose synthase like (Csl) genes encode enzymes involved in the synthesis of celluloses and hemicelluloses, two most important classes of polysaccharides for


CSLC
CESA
0.1
CSLD
CSLA
CSLG
CSLE
CSLH
CSLB
CSLF
CESA Glc β1,4 Glc β1,4 Glc β1,4 Glc
CSLA Man β1,4 Man β1,4 Man β1,4 Man
CSLC Glc β1,4 Glc β1,4 Glc β1,4 Glc
CSLF Glc β1,4 Glc β1,4 Glc β1,3 Glc

Current Opinion in Plant Biology

biofuel production that comprise of 50-70% of the dry weight of plant biomass. According to published research (Current Opinion in Plant Biology 2006, 9:621–630), the Csl genes form nine different sequence clusters/subfamilies based on phylogeny analysis; one of the subfamilies contains proteins for cellulose synthesis (CesA) while the rest are thought to be hemicellulose synthases (CslA, CslC, CslF, CslH).

Why: We identified Csl homologs in fully sequenced plants and lower green algae (unicellular *chlorophyta*) in 2009. Our goals were to answer the questions: how wide are different Csl families distributed taxonomically and how have they evolved? With NGS transcriptome data available for a few higher green algae (multicellular *charophyta*), we want to expand our previous homology search to these new data.

Where: Two *charophyta* (*Chara vulgaris* and *Spirogyra pratensis*) transcriptome data sets sequenced by 454

How: Use TBLASTN/TFASTY to query published Csl proteins against the above 454 EST data and then assemble homologous reads into longer contigs

Detailed methods and steps:

1. MUST READ references:
   http://www.biomedcentral.com/qc/1471-2229/9/99/
   http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/21501468/

   Optional references:
   http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3390603/
   http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/22253761/

2. Download datasets:
   2.1 Databases: Search NCBI SRA web database using the species name; each species has multiple data sets (locate the links and then wget)
   2.2 Query sequences: Csl protein sequences in BMC Plant Biology 2009 (http://www.biomedcentral.com/qc/1471-2229/9/99/additional, additional file 3, wget)

3. SRA read format convert for each SRA dataset
    - 3.1 Use the fastq-dump command in the SRA tool Box
        - 3.1.1 If use glu, add /home/mrupani/sratoolkit.2.1.16-centos_linux64/bin/ to your PATH environment variable (edit .bashrc file)
        - 3.1.2 If use your own laptop, go to http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software to download the tool and install it
        - 3.1.3 Run the fastq-dump command to convert SRA format files to FASTQ format files
    - 3.2 Use the FastqTo454.pl script in the NGSQC toolkit
        - 3.2.1 If use glu, add /home/mrupani/ngs-qc/NGSQCToolkit_v2.3/Format-converter/ to your PATH environment variable
        - 3.2.2 If use your own laptop, go to http://59.163.192.90:8080/ngsqctoolkit/ to download the tool and unpack it
        - 3.2.3 Run the FastqTo454.pl script to convert FASTQ format file to fasta + quality files

These jobs may take a couple of hours; try to use nohup and & on command line

4. Homology search:
    - 4.1 TBLASTN search using 2.2 protein fasta sequences as query and 3.2.3 nucleotide fasta sequences as database, output as tabular format without alignment
    - 4.2 Also try TFASTY search and compare which one gives more hits and which one runs faster (add time before command line), output as tabular format without alignment

I suggest using the output from 4.2 for downstream analysis

5. Parse the output
    - 5.1 Use Shell commands and pipe to construct command line to extract matched hit IDs
    - 5.2 Write a perl script (using bioperl modules) to take the hit ids (from 5.1) and the original databases (from 3.2.3) as the input files to extract the fasta sequences

6. Combine multiple fasta sequence files in 5.2 into one using cat, because each alga has multiple datasets

7. Csl homologous reads assembly
    - 7.1 Go to http://seq.cs.iastate.edu/cap3.html and download & unpack cap3 tool
    - 7.2 Run cap3 to assemble the resulting file from step 6 (use -o 60 -p 97 options)

8. TFASTY to search 2.2 file against the contig file resulted from 7.2