Practical Bioinformatics for Biologists (BIOS493) Course Project 1

Project title:
In silico identification of GH family 9 and 10 homologs in metagenomes

Two groups:
Group 1 (GH9): **Bill, Brenda, Steve, Tom**
Group 2 (GH10): **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 **Apr 30** 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 project1 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 project1 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: Glycoside hydrolase (GH) families 9 and 10 (http://www.cazy.org/GH9.html and http://www.cazy.org/GH10.html) are protein families containing enzymes with

a variety of substrates. These enzymes include **cellulases** - enzymes degrading celluloses, and **xylanases** – enzymes degrading xylans, as well as enzymes breaking down other polysaccharides.

Why: Celluloses and xylans are the two most abundant biopolymers found in nature. They represent the most promising resources for producing biofuels, as degradation of celluloses and xylans will release glucloses and xyloses, which could be further fermented to yield ethanol and other biofuels. Therefore studying cellulases and xylanases of GH families is of great significance to the bioenergy research.

Where: In biomass-rich environments such as animal guts and biomass composts, bacterial and fungal communities produce numerous enzymes to degrade plant biomass, the foods of microbes to acquire carbon and energy. Recent metagenome sequencing projects have deposited millions of proteins into the public databases. **These metagenomes very likely contain many novel GH9/10 homologs and subfamilies.**

How: Here we want to identify GH9/10 homologs in metagenomes and further look for novel subfamilies that are not represented in completed microbial genomes. By doing homology search, we can **identify homologs of known GH9/10 proteins** in the public metagenome databases such as JGI metagenomes and published animal gut metagenomes. We will further combine metagenome homologs and known/annotated GH9/10 proteins (from CAZy database, see below) and perform **phylogeny analysis**. By showing the two datasets in different colors in the phylogram, we can identify clades containing only or primarily metagenome homologs; these clades will represent novel GH9/10 subfamilies. By locating biochemically characterized NCBI-nr proteins (from CAZy database) on the tree, we can predict functions for neighboring subfamilies. Member proteins of the novel subfamilies will be further analyzed by checking what environments they are from, what other functional domain they may have, what are the characteristic and distinct motifs, or predicted protein 3D structure for selected proteins.

Detailed methods and steps:

1. MUST READ references:
   GH5: http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/22992189/
   GH48: http://cys.bios.niu.edu/yyin/teach/PBB/GH48.pdf

   Optional references:
   CAZy database: http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/18838391/
   dbCAN: http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/22645317/

2. Download datasets:
   2.1 Databases:
      2.1.1 JGI metagenomes (http://genome.jgi.doe.gov/)
      Use wget to download *.gene.faa.gz files
      Might need to write a shell script to automate

2.1.2 Uncompress (gzip, unzip, tar) the downloaded files

These might be not necessary, as I have downloaded them and the file size is huge (46GB); here is the path: /home/yyin/db/jgi-v3.5.pr.fa

2.2 Query sequences and models:

2.2.1 Known GH9/10 proteins:
http://www.cazy.org/GH10.html
http://www.cazy.org/GH9.html
Download the GenBank IDs first and then batch entrez to get sequences or use bioperl module: Bio::DB::GenBank

2.2.2 GH9/10 hidden markov models (HMMs):
http://cys.bios.niu.edu/dbCAN/family.php?ID=GH9
http://cys.bios.niu.edu/dbCAN/family.php?ID=GH10

3. Homology search:
   3.1 *hmmsearch* using GH9/10 HMMs as the query against the JGI metagenome protein datasets (2.1) and the NCBI-nr proteins (2.2.1)

4. Parse the *hmmsearch* output
   4.1 Write a perl script to extract the hit id, e-value, bit score, query start, query end, hit start, hit end as a tabular file (the output of the script); the input is the *hmmsearch* raw output file
      4.1.1 Try to use e-value and coverage [calculate as (query end – query start)/query length; note here the query is the HMM; the model length could be extracted from files downloaded in 2.2.2] to filter the hit list, e.g. only keep hits with e-value<1e-5 and coverage >0.8
   4.2 From the tabular file, use Shell commands with pipe to extract all the hit ids
   4.3 Write a perl script (using bioperl modules) to take the hit ids (from 4.2) and the original databases (from 2.1 and 2.2.1) as the input files to extract the full length protein fasta sequences

5. Write a perl script to retrieve the domain regions
   5.1 The input files are the tabular file from 4.1 and the full-length protein fasta file from 4.3. The output will be the matched domain sequences.

6. Combine the homologous domain sequences from 5.1 from NCBI-nr and from JGI metagenomes using *cat*

7. If dataset size is too large, try USEARCH to remove very similar sequence to reduce the data size using sequence identity threshold (may not necessary after you applied the e-value and coverage cutoff in 4.1.1)

8. Build multiple sequence alignment (mafft) and phylogeny (fasttree) and prepare color definition file (Shell commands and perl) to show NCBI-nr and JGI sequences differently

9. Use iTOL to visualize the tree

10. Inspect the tree topology to identify metagenome-specific subfamilies