Biological Data Analysis (CSE 182): Final Project

Genome sequencing and annotation

You have leading the Bioinformatics part of a project on the whole genome sequencing of the eukaryotic organism *H. mysterium*. The genome of this organism is known to have only one chromosome. It also contains many repeats from the MLU family.

You will present your findings as a 20min.presentation in class June 1 or June 3. You may not consult with any one other than your designated project partner, TAs, or the instructor. Your presentation (Powerpoint or other suitable format) must be mailed to Gary Tedeschi and Qian Peng by midnight May 31, and cannot be changed subsequently. No late submission will be accepted.

Input: You will have the following data:

1. Fragments from a whole genome shotgun assembly.
2. A Physical map of *H. mysterium*. The physical map data file will be a collection of short unique sequences in the order in which they appear in the *H. mysterium* genome.

The data will be made available by May 14. You are to spend that time testing available Bioinformatics tools, and getting your scripts ready.

Output: Your presentation must contain the following information:

1. Genome (5pts):
   (a) Your estimate of the coverage. The coverage is the sum of the lengths of fragments in your assembly, divided by the size of the genome.

2. Assembly (20pts).
   (a) Number of contigs in your assembly, and the distribution of sizes.
   (b) Compute an ordering of the contigs based upon information in the physical map, and present evidence that your assembly is correct. How many contigs do not have any order associated with them.

3. Repeats (15pts):
   (a) Compute percentage of the genome that contains Repeats. Create an assembly with the Repeated sequences masked out (use Blastn to find the repeats), and a perl script to mask out repeats.

4. Gene Finding (20pts):
   (a) Compute number of protein coding genes. (Note: some genes may be partial because of gaps in the assembly).
   (b) Report the location of each of the genes.
   (c) Compute the protein sequence for each gene.

5. Protein sequence analysis (40pts): The questions below are only a suggestion. While you are only required to address these, additional analysis will earn extra credit. This could include validating your results using multiple tools and, and/or using our own tools, as well as other information that might be available on the web regarding these genes. It might also help to check if your results are confounded because of assembly errors, and poor gene prediction.
   (a) Identify domains and their locations on each of the protein sequences.
   (b) How many genes are membrane bound? For each of the membrane bound proteins, what is the number and location of transmembrane segments?
   (c) How many genes encode Transcription factors? What are the coordinates of the DNA binding domain?
   (d) Which proteins are secreted? Identify the signal sequence, and the cleavage site for each protein.
Tools: Links can be found on the web

1. Blast
2. Genscan (http://genes.mit.edu/GENSCAN.html)
3. Phrap (Assembly tool). Will be provided at the class web-site.