

HW 2: due Tuesday, October 19th

Readings:

- Green, E.D. (2001). Strategies for the systematic sequencing of complex genomes. *Nature Reviews Genetics*, 2:573-83.

Problems: Submit hard copy of answers to problems 1-4. For Problem 5, please use `provide` to submit your output file as well as any program files. This works from any EECS machine by typing:

```
conbrio% provide comp150b hw1 myfilename1.here myfilename2.here ..
```

1. Find a gene related to Alzheimer's Disease on Human Chromosome 21. List (by URL or by common name, i.e., "Golden Path Gene Viewer") which web sites you used to answer each question.
 - (a) What is its official gene symbol and name?
 - (b) What region of Chromosome 21 does it map to? Print out and submit a picture of this part of the Chromosome 21 map; circle your gene's name on the printout.
 - (c) Name another disease associated with this gene.
2. In one or two sentences, contrast the advantages and disadvantages of whole-genome shotgun sequencing versus a clone-by-clone approach. Which method do you think might be best for sequencing the genomes of each of the following organisms, and why?
 - (a) a previously unsequenced bacterium
 - (b) the laboratory rat (commonly used in medical research)
 - (c) a rare endangered bird

3. Draw the overlap graph for the following sequences. Find a maximum weight Hamiltonian path in this graph, and show the assembly this corresponds to.
 - (a) ACCA
 - (b) CAGGG
 - (c) CCAATA
 - (d) CGCC
4. Draw the suffix tree associated with the sequence ATCCATTATG.
5. (Inspired by S. Salzberg) In this programming assignment, you are to write a overlap detection program that could be used as part of a sequence assembler. The program should assemble the sequences in the text file “hw2.reads” available from the course web site. To make this easier, the sequences are all from the same strand (you do not need to reverse-complement any of them) and there are no errors.

Sequence A is considered to overlap sequence B if there is a suffix of A at least 40bp in length that exactly matches a prefix of B. Note that this relation is not necessarily symmetric; if A overlaps B, B may not overlap A!

You should submit both your source code and an output file showing all the overlaps detected in the sequence file. Sort the file by ID number of each of the sequences, and for each sequence the file should contain EXACTLY one line. That line should contain the ID of the sequence followed by the IDs of all sequences that it overlaps. The list of overlapping sequences should also be sorted in order by ID number. For example: R1 R24 R165 R175 R2 R33 R109 R138 ... etc. We will be comparing your files to the correct answer using 'diff' so the format should match exactly. Put exactly one space between successive ID numbers and no whitespace after the last ID number in each line. We will also check your program on another data set.