

# Spring 2008

## 10-810 / MSCBIO2070: Computational Genomics

### Problem Set 2: Genome Assembly and Sequence Evolution

**Due date: Feb. 19, 2008, before class**

The contact TA for this assignment is Jacob Joseph (jmjoseph@andrew.cmu.edu). Jacob's office hours are at Mellon Institute 646C, on Wednesdays, 4-5pm, or by appointment.

Where implementation is required, supply source code in Python, Java, MatLab, C, C++, or R. Code should be portable; it should execute on a UNIX/Linux platform. Please include any auxiliary files and short description of how to execute your code. Submission of this source code may be by email to the contact TA.

Collaboration is permitted, but solutions must be completed individually. Include a list of your collaborators. Refer to the course website for complete policies.

---

1. (a) A library of overlapping human genome fragments is to be assembled. Given a genome size  $G=3.2 \times 10^9$ , and lambda clone inserts with length 20kb, how many clones would be required to assure with probability 0.98 that a particular base be covered by at least one fragment?
- (b) A whole genome of 150Mb is to be sequenced. A small-plasmid library (average insert size 3kb) providing 5x insert coverage (=clone coverage) and a BAC library (average insert size 150kb) providing 15x insert coverage are prepared for sequencing. Sequencing reads of length 700nt are generated from both ends of all inserts in both libraries. Compute the average sequence read coverage for the genome.
- (c) Detail, in a brief paragraph, the general principles of top-down and bottom-up genome sequencing techniques. Describe the suitability of each to example sequencing problems, and mention the applicability of methods which integrate both techniques.
- (d) Read the Celera and ARACHNE papers and describe the similarities and differences between these approaches to genome sequencing. How does ARACHNE correct for sequencing errors? How do each handle repeats, and chimeric reads?
  - Batzaoglou et al. *ARACHNE: A Whole-Genome Shotgun Assembler*. Genome Research. Vol. 12, Issue 1, 177-189, January 2002.  
<http://www.genome.org/cgi/content/abstract/12/1/177>
  - Meyers et al. *A Whole-Genome Assembly of Drosophila* Vol. 287. no. 5461, pp. 2196 - 2204.  
<http://www.sciencemag.org/cgi/content/abstract/287/5461/2196?ck=nck>

2. (a) Implement an efficient program to construct contigs (i.e., segments comprised of two or more sequence fragments) from an input set of reads. For simplicity, assume the sequences are error-free, and consider only exact overlaps of 40bp or more. A set of 300 sequences in FASTA format is provided on the course website. Include output of the nature:

```
readID1  readID2  orientation  offset
007      025      F            -116
007      056      R            302
```

Where orientation is either F (forward) or R, indicating the direction of readID2, and offset is the signed number of bases that the left end of readID2 is from the left end of readID1. Please sort in ascending order of readID1, then readID2.

- (b) Extend your implementation to assemble the contigs. Assume pairwise sequence identity of  $\geq 90\%$  in overlapping regions.
3. (a) What properties should a similarity metric have to be suitable for use in Smith-Waterman alignment?
- (b) Consider the paper:  
 Mahony, Auron, and Benos. *DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies*. PLoS Computational Biology Vol. 3, No. 3, e61 doi:10.1371/journal.pcbi.003000  
<http://compbiol.plosjournals.org/perlserv/?request=cite-builder&doi=10.1371/journal.pcbi.0030061>  
 Figure 2 compares the resulting score distributions of the five main similarity metrics evaluated. Empirically compare the distributions for the ALLR and AKL metrics. Referring to the formulas in Table 1, suggest improvements to the AKL metric so it may be used with Smith-Waterman alignment.
- (c) A number of methods for multiple alignment have been discussed, including Smith-Waterman and progressive alignment (as in CLUSTALW). Hidden Markov Models may also be used for this purpose. Describe, in pseudo-code, how such an alignment may be constructed. Assume HMM operations such as Baum-Welch and Viterbi have already implemented.
4. (a) In maximum parsimony tree reconstruction, a minimal number of single, independent base mutations is used to construct a tree. From the multiple alignment of five taxa below, reconstruct the optimal tree. What is the parsimony score of this tree?

```
1: ATGC
2: ATTC
3: GTTC
4: GACT
5: GTCC
```

- (b) While maximum parsimony methods identify the minimal set of changes between present day sequences, a particular base may undergo several mutations, possibly to the original base (e.g.,  $A \rightarrow T$ ,  $T \rightarrow C$ ,  $C \rightarrow A$ ), rendering the change invisible. Differentiate and describe how the Kimura 2-parameter and Jukes Cantor models may be used to correct for such silent mutations.
- (c) Provide a brief comparison of how parsimony-based tree reconstruction methods differ from distance-based methods. When is each most applicable?