EN 600.439/639: Computational Genomics

Homework 1

Fall 2014, Prof. Langmead

Please review the syllabus for guidelines regarding homework submission. I include a few reminders here, but they will not be included in future homeworks.

Work individually. Use Piazza to ask questions or report issues. When viewing a problem on `rosalind.info`, remember to click the grey link that says "click to expand" at the top. This reveals helpful discussion and links.

Develop and test your programs wherever you like. Your programs on the JHU CS undergraduate or graduate cluster. We recommend you test your solution on one of those clusters before submitting. *Important:* Email solutions to `cs439@cs.jhu.edu` and include the following in the subject line: "EN 600.439/639 Homework 1 submission". Also, say in your submission email which cluster, undergraduate or graduate, we should test on.

For rosalind.info problems, submit your program source, a compiled version (if you used a compiled language), and the command(s) you used to compile it. Whatever language you use, your program should accept input on the standard-in file handle, and write output on the standard-out file handle. In other words, when the grader tests your solution, they will do this:

```
your-executable < input.txt > output.txt
```

And you will be graded based on the content of `output.txt`. See your language's documentation to learn how to access the standard-in and standard-out file handles (it's usually very straightforward). Points will be deducted if your program does not work as above.

This homework is out of 35 points:

1. (2 pts) Do rosalind.info problem "Counting DNA Nucleotides":
   http://rosalind.info/problems/dna/
2. (2 pts) Do rosalind.info problem "Transcribing DNA into RNA":
   http://rosalind.info/problems/rna/
3. (2 pts) Do rosalind.info problem "Complementing a Strand of DNA":
   http://rosalind.info/problems/revc/
4. (2 pts) Do rosalind.info problem "Counting Point Mutations":
   http://rosalind.info/problems/hamm/
5. (2 pts) Do rosalind.info problem "Protein Translation"
   http://rosalind.info/problems/prot/
6. (2 pts) Do rosalind.info problem "Finding a Motif in DNA"
   http://rosalind.info/problems/subs/

7. (2 pts) Do rosalind.info problem "Computing GC Content"

   http://rosalind.info/problems/gc/

8. (3 pts) Say you take a genetic test, and one of the results is that you have "reduced odds" of developing lung cancer. It would be more medically useful, of course, if the test told you definitively whether or not you will get lung cancer. Give two reasons why the test cannot be so definitive.

9. The Central Dogma of molecular biology is discussed in your readings and in class.

   (a) (2 pt) What is transcription?
   (b) (2 pt) What is translation?
   (c) (2 pt) How does the Central Dogma explain the link between genotype and phenotype?

10. Download this file:

    http://www.cs.jhu.edu/~langmea/resources/hq1_reads.fastq

    This FASTQ file contains 1000 real-world DNA sequence reads from a human whole-genome resequencing project [1]. See the following iPython notebook for details of the FASTQ format:

    http://nbviewer.ipython.org/gist/BenLangmead/8376306

    Implement software to parse a FASTQ file (you can reuse Python code from the notebook above) and print the following summary:

    (a) Each row of the summary corresponds to a position in the read. That is, row 1 corresponds to the first nucleotide position, row 2 corresponds to second nucleotide position, etc. Your summary will have 100 rows.
    (b) Each row contains seven integers separated by spaces. The seven integers to print, in left-to-right order, are:
        i. The total number of A nucleotides observed across all reads at this position
        ii. Same for C nucleotides
        iii. Same for G nucleotides
        iv. Same for T nucleotides
        v. Same for any other characters observed besides A, C, G or T (yes, this happens)
        vi. The number of Phred-scaled quality values less than 20 across all reads at this position
        vii. Same for Phred-scaled quality values ≥ 20

    (7 pts) Submit the summary you calculated for the given data and the program you used to make the summary.

    (5 pts) Answer the following:

    (a) Do you see anything suspicious about any rows of the summary?
    (b) What trend do you observe in the number of quality values less than 20 over the length of the read? Explain briefly what could cause this.

References

[1] S.S. Ajay, S.C.J. Parker, H.O. Abaan, K.V.F. Fajardo, and E.H. Margulies. Accurate and comprehensive sequencing of personal genomes. *Genome research*, 21(9):1498–1505, 2011.