Programming for Bioinformatics

BIOL 8803B

November 23rd, 2015

Concepts:

threads

Exercises:

1.) Memory usage in the overlap script

There are some really, really big BED files these days. They're going to get bigger. This means that eventually we won't be able to just read them all into memory at once. We could just buy more RAM, but that's cheating. Modify your **overlap script from week 9**, or mine from T-Square, so that it can handle really big files without running out of memory. The best way to do this would be to only read in lines that you are going to work with, and to get rid of the ones you no longer need. `Tie::File` could also be of use in this exercise. What kind of performance change do you get?

2.) Study of threading the overlap script

Make some version of the **overlap script from week 11**, yours or mine, threaded. I would suggest doing this by giving each chromosome a thread. Make the number of threads that can be run at a given time a command line argument, i.e. if you give '1' for the number of threads to run, just run one chromosome at a time, '2' run two at a time, etc.. Experiment with different numbers of threads.

Deliverables:

- **Code:**
  - `overlapLargeBed.pl`
  - `overlapMT.pl`
- **A Readme file** which reports:

- For `overlapLargeBed.pl` report:
    a) The file size and number of lines of your input bed file
    b) Compare the running time of your original overlap script from week 9 and `overlapLargeBed.pl`
    c) What *other* differences would you expect by comparing the two scripts?
- For `overlapMT.pl` report:
    a) The file size and number of lines of your input bed files
    b) What were the running times when you used different number of threads?
    c) Go over your results from b) and try to answer the following questions: Is performance better than the single thread version of your script from week 11? Does increasing the number of threads improve performance? Does the running time increase at a certain point?