CMPT 441/711, Fall 2017, List of Projects

Tentative timeline:

- Read through and think about the top 3 projects you would like to work on by Friday, November 3rd at 5PM.
- Teams of 3-4 people each will be assigned by the morning of November 6th.
- With your assigned group, prepare a plan of attack due Friday, November 10th at 5 PM.
- Feedback will be provided by the morning of November 13th.
- Preliminary meetings with the instructor will be held for each group the week of November 27th.
- Group presentations will be held on December 4th during class time (depending on the number of groups we may need to schedule an additional session as well).
- You will then have until December 15th at 5 PM to submit your final project report.

A. **Biomarker discovery from cancer omics data**

Easy goal: Read through and understand the contents of modules 1 and 2 of this workshop: *https://bioinformatics.ca/workshops/2009/clinical-biomarkers.* Then go through the lab (module 4) and perform the suggested analysis on a relevant paper of your choice.

Hard goal: Identify a paper that studies the same type of cancer under the same conditions, but using a different method, and apply the method of the first paper to the data of the second paper, and vice versa. Discuss any differences that you observe.

Stretch goal: Perform a comprehensive literature review and apply each method used to the data in each of the papers to see how reproducible the results are; report on the observed discrepancies and any biomarkers that persist under variation of methodology.

B. **Applying the Four Russians speed-up to alignment (thanks: Emre/Baraa)**

Easy goal: Read through and understand the technique known as the Four Russians, *https://en.wikipedia.org/wiki/Method_of_Four_Russians*, and how it can be applied to speed up global alignment.

Hard goal: Implement your own version of the global alignment algorithm using the Four Russians technique, with an adaptive choice of block size, in a language of your choice, and demonstrate a speed improvement relative to the basic version in the same language.

Stretch goal: Extend the approach to the Nussinov algorithm for RNA folding.

C. **Understanding regulatory networks and motifs**

Easy goal: Read through and understand the contents of modules 1 and 17 of this workshop: *https://bioinformatics.ca/workshops/2015/high-throughput-biology-sequence-networks-2015*. Then go through the materials and understand the principles on which different software packages work.

Hard goal: Go through a collection of gene sequences classified into two groups (for instance, circadian genes vs. non-circadian genes) and identify all motifs that appears to be over-represented in one group relative to the other, using a method of your choice; discuss your results.

Stretch goal: Incorporate the Gibbs sampling approach into your methodology and repeat the process until convergence. Discuss what you learned and the challenges you run into.

D. **Understanding metabolic network models**

Easy goal: Read through and understand the contents of modules 1 and 4 of this workshop: *https://bioinformatics.ca/workshops/2008/systems-network-biology#outline*. Then go through the lab for module 4 and perform the analysis of the metabolic network.

Hard goal: Also install the competing MONGOOSE software (I will provide you with the details) and compare the results of your analysis with the above; discuss why the differences arise.

Stretch goal: Repeat the process for all the models available from the In Silico Organisms repository, *http://sbrg.ucsd.edu/InSilicoOrganisms/OtherOrganisms*, and report on the results.

E. **Microbiome/metagenomics composition analysis**

Easy goal: Read through and understand the contents of modules 1 and 2 of this workshop: *https://bioinformatics.ca/workshops/2015/analysis-metagenomic-data-2015*, and work through the methodology used for taxonomic composition determination.

Hard goal: Apply the methodology to a public dataset, such as the one available at the GOLD (Genomes OnLine Database), and discuss your results in detail, including the main source of variability.

Stretch goal: Design an approach for discovering any genes other than 16S rRNA that could be suitable for taxonomic composition determination.

F. **Using the inside-outside algorithm to train an RNA folding model**

Easy goal: Read through and understand the contents of the papers “RNA sequence analysis using covariance models” and “Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction”.

Hard goal: Apply the methodology to a public dataset, such as the one available at the tmRDB, and discuss your results in detail, including the main source of variability in the parameters estimated.

Stretch goal: Extend your approach to either a) automatically extract the model structure (not just the model parameters) from data or b) ensure the sparsity of the parameter vector (so that as many coefficients as possible are 0).

G. **Cloud computing in computational biology (clarified)**

Easy goal: Read through and understand the contents of modules 1, 3 and 4 of this workshop: *https://bioinformatics.ca/workshops/bioinformatics-big-data-computing-human-genome-2016*.

Hard goal: Develop a Docker container that will allow a multiple sequence alignment of at least 20 sequences (e.g. the mammalian genome dataset) to be performed in a reproducible way, and compare the performance to what you would get on a local machine running the multiple alignment software directly.

Stretch goal: Develop a program that will semi-automatically create Docker containers for other computational biology tools, including the loading of appropriate dependencies and the setting of parameters.

H. **Estimating phylogenies with a CTMC taking indels into account**

Easy goal: Read through and understand the contents of the papers "Efficient Inference in Phylogenetic InDel Trees" and "A Poissonian Model of Indel Rate Variation for Phylogenetic Tree Inference".

Hard goal: Implement and test the method to estimate the posterior probability of different trees for a well-studied dataset (e.g. the mammalian genome dataset), and discuss the limitations and challenges.

Stretch goal: Either suggest a more efficient way to estimate the model parameters or compare its performance to the TKF91 model.

I. **Pick your own poison**

Design your own project, based on one of the workshops at *bioinformatics.ca* or another source, and structure it into an easy, hard and stretch goal. Send it to me for approval, along with a proposed team composition, by Friday November 3rd at 12 PM. If approved, this will be your project and your team; if not, you will still be responsible for selecting your top three projects out of the list proposed here.