Overview

The goals of the class project are to gain experience working on a (small) bioinformatics research problem using real biological data. In order to encourage creativity and ambitious projects, you will not be graded based on whether or not your idea works. Grades will be based on whether your project was thoughtful, executed carefully, and built upon sound computational and biological principles. Contributions to the bioinformatics software community (e.g. releasing your new method as open source software or contributing patches, new features, or documentation to existing software) will be viewed favorably.

In contrast to the homework, you are encouraged to use existing software and packages for the project. You are required to reference all data sources and software in your project reports. Each student will implement their own project individually. However, multiple students may elect to work on the same project as long as they work independently. You do not need to run your code on the biostat server, but your project must be reproducible. This means that you must provide the source code and sufficient documentation for running your code.

Types of projects include:

- Applied projects where one or more computational methods are run on a single dataset of interest and the evaluation focuses on the biological interpretation of the results.
- Comparative projects where multiple computational approaches are run on multiple datasets to assess their relative strengths.
- A computational project where a new algorithm is implemented or an existing algorithm is extended and compared with the original.
- An extension of your current research, in which you use a dataset that you are already studying but analyze it with new types of computational methods. *The project should not be something that you were already planning to do as part of your research.*

Specific ideas

- BMI/CS 776 Spring 2015 project ideas https://www.biostat.wisc.edu/bmi776/spring-15/project.html
- Implement a solution to a closed DREAM challenge http://dreamchallenges.org/project-list/closed/ Be sure to only use the training data when building your model, and choose a challenge for which the test set is already available so that you can evaluate your performance. Avoid reimplementing methods that have already been described in the challenge summary papers.
- PrecisionFDA Consistency Challenge https://precision.fda.gov/challenges/consistency
- Use Basset https://github.com/davek44/Basset to explore deep learning in genomics. The project could vary the network structure and hyperparameters to assess performance. The full multi-task model cannot be trained on a CPU so the project would require either access to a powerful GPU or training on a single ChIP-seq dataset. It may be possible to assess single-task versus multi-task performance for a small number of classes.

- Run multiple network algorithms to predict pathways from genome-wide datasets and compare the results. Potential network algorithms include:
  - ResponseNet, which we will learn in class
  - Forest from Omics Integrator https://github.com/sgosline/OmicsIntegrator
  - PathLinker https://github.com/Murali-group/PathLinker
  - TieDIE https://github.com/epaull/TieDIE
- Create a GenomeSpace recipe and associated tool to teach users how to conduct a particular bioinformatics analysis. Due to the atypical nature of this project, anyone interested in this type of project must seek prior approval from Professor Gitter one week before the project proposal deadline http://www.genomespace.org/blog/2016/02/01/announcing-the-new-genomespace-recipe-resource
- Compare RSEM with RNA-seq quantification algorithms that bypass exact alignment to the genome. See http://robpatro.com/blog/?p=248 for an introduction and specific methods.
- Develop a tool for visualizing genomic or other large-scale biological data. The visualization approach should not be a trivial modification of existing plotting libraries (e.g. a heatmap or Circos plot), and the results should center on how the visualization strategy facilities better biological interpretation of the data.
- Analyze genomic data using reference genome graphs. See http://sjcockell.me/2016/02/22/reference-genome-graphs/ for an introduction and specific methods.