EE 372: Data Science for High-Throughput Sequencing

Project Guidelines and Ideas Spring 2016

Projects for the course should be broadly related to the overarching theme of the class: sequencing. Projects can have a creative (or innovative component) or be expository. We expect expository projects also to have some component involving real data. Up to two students can collaborate on a project. Some datasets may be larger and therefore take longer to process, so start early! Please keep the following dates in mind: Students may also find GEO useful in obtaining datasets.

1. Discuss project with TAs (Wednesday, 27 April 2016)
2. Proposal deadline (Monday, 2 May 2016)
3. Milestone deadline (Friday, 20 May 2016)
4. Poster presentations (Monday, 6 June 2016)
5. Project write-up deadline (Wednesday, 8 June 2016)

1 Deliverables

1.1 TA meeting

Students are required to meet with the TAs during the Wednesday, 8 June 2016 class to make sure their project is appropriate for the course.

1.2 Proposal

The project proposal should be about a page in length (not including figures and references) and briefly explain the following:

1. Objective: Why is your problem interesting? What do you hope to accomplish?
2. Data: What is the dataset you will be working with?
3. Methodology: What is your approach?
4. Potential problems: What major obstacles do you foresee? What will you need to overcome these obstacles?

1.3 Milestone

The purpose of the milestone is to ensure that you have made at least some progress on your project well before the deadline. The milestone report should be about 2 pages in length (not including figures and references) and hit the following points:

1. Introduction: Brief summary of your proposal. What has changed?
2. Initial results: What have you tried? What has not worked? What relevant figures or equations have you produced? What do they mean?
3. Next steps: How will you overcome initial obstacles? What else do you want to accomplish for this project?

1.4 Poster Presentation

The poster presentation will take place during the finals block allocated by the university to this class (6 June 2016 3:30pm-6:30pm). The exact location is to be determined and will be announced later. The poster should summarize your findings for the project and be 3' by 2' in size.

1.5 Writeup

The write-up should be about 4 pages in length (not including figures and references) and include the following sections:

1. Introduction: Background information about your project. What was the objective of this project? Why is your problem interesting?
2. Data: What datasets did you look at for this project?
3. Methodology: How did you go about accomplishing your objective?
4. Results: What relevant figures or equations have you produced? What do they mean? What went wrong?
5. Future Work: How do you build upon the work you have done?

2 Project ideas

A project broadly related to any of the following topics will be appropriate for the course. Students can pick projects outside the ones listed below but should check suitability with the course staff.

1. **Peak calling for epigenetics**: While cells have roughly identical genomes, they express different parts of the genome due to epigenetic factors. Assays such as ChIP-seq and ATAC-seq help scientists study which parts of the genome are accessible for transcription. This problem involves determining which parts of the genome are accessible based on obtained reads. Two popular peak calling tools are MACS and ZINBA.
2. **Short Tandem Repeat (STR) calling**: STRs are short strings of DNA which appear one behind the other. These are known to be important for biological function, but calling them is challenging. A tool for performing STR calling is LoBSTR.
3. **Alignment of 2 random strings**: What is the edit distance between two random strings of A,G,C,T? Recent work shows via simulation that this is 0.518n. How does this change with the change in weights given for inserts and deletions?
4. **Short read error correction**: Short reads have about 1% error rates. How do the types of errors tie in with the base calling used? How does one model and correct these errors? The state of the art software used for this is QUAKE.
5. **Dealing with batch effects**: While scientists are careful when executing experimental protocols, there will always be differences between repeated experiments. Batch effects are observed when the experiment ID (e.g. day the experiment was done) overcomes signal from the biological variable of interest. Tools like ComBat have been designed to mitigate batch effects.
6. **Meta-genomics**: When one has the genomes of multiple organisms (or many genomes from the same type of organism), a natural question to ask is: how do these organisms relate to each other on a phylogenetic level? One can build a tree illustrating which organisms are more similar. This can provide insight on which genes are important for distinguishing two different organisms. Some interesting work done here include MetaGene and MetaBat.

7. **Multi-omics**: With RNA-Seq, ChIP-Seq, Hi-C-Seq, etc., we are often inundated with many different categories of data for the same type of biological sample. Could we possibly combine the information from these different types of data to discover something more?

8. **Hi-C-Seq**: Hi-C-Seq is an assay used to study the 2D and 3D structure of the chromatin, giving insights on how the genome folds and how distant regions of the genome can interact with each other.

9. **Long range information**: In class we have covered short reads and long reads. There are other read types including read clouds (10x Genomics) and reads with long range information (BioNano Genomics). Genes are regulated by both local elements and distant elements. From a theoretical perspective, what is possible and what is not with such reads?

10. **Single cell assays**: Recent technologies have allowed researchers to extract cell signatures at the resolution of individual cells (e.g. single-cell RNA-seq). What can we learn about cell populations from these datasets?

11. **Single-cell cancer genomics**: One advantage of the resolution granted by single-cell assays is that we can observe how a cellular population differentiates. Looking at single-cell datasets for cancer cells may provide insight on how cancer evolves and which genes are important for this evolution.

12. **Base-calling for fourth-generation sequencing**: This is an active area of research with interesting recent papers involving methods such as deep learning.

13. **Reanalysis of published data**: Several datasets related to the above topics already exist. What can you discover by reanalyzing one of these datasets using a novel method?