Problem-Based Learning in Bioinformatics (2020)
Data Standards and Data Integration

Facilitator: Ryan Brinkman, rbrinkman@bccrc.ca**; 604-675-8132, BCCRC room 12.113**

Bad news! Its grant renewal time in your position as head of bioinformatics at the leading provincial institute doing cancer research as part of a large consortium of biologists and clinicians that encompasses 20 research institute across the country. Your team is in charge of all bioinformatics analyses of the data generated.

Good news! A collaborator in Nunavut has invented a new methodology for analyzing the protein content of individual cells in a high throughput manner. The methodology analyzes single cells for the relative abundance of thousands of proteins at a time using Raman probes to tag individual proteins and protein complexes. Researchers there have miniaturized and automated sample-handling techniques so that they can now generate 1000 times more data in a week than they previously could. The data files also tend to be quite large in terms of the number of entities (rows of cells * columns of proteins). They have provided this technology to a few other laboratories that are now collaborating with your Centre in investigating the protein response in different diseases.

Bad news! One of your team comes into your office and complains that your group is frustrated with the data files provided for analyses: it's a pain to write parsers for all the different kinds of information users are putting in them. They are also having trouble matching up data output by the machine with information about the samples and how the data was generated (*e.g.,* is the data from a normal or disease person and other demographic information about the individual, that type of machine and instrument setup). Information about samples is also differentially encoded in terms of how things are recorded in free text fields, making matching updates in automated analysis really frustrating. As a result, your group is spending a great deal of time cleaning up the data before analysis and writing scripts to match up data, and feel they don't have enough time for analyses itself. Additionally, the commercial software tools being used for statistical analysis are not very flexible, and the data analysis pipeline is fragmented and not biologist-friendly which requires a lot of hands-on work from the bioinformatics group. It would be great to hand this off to the biologists as they tend to do canned analysis pipelines for the most part. Part of the analysis involves looking up information about each of the protein complexes and their known interactions, and there are different opinions about how to best access publicly available information. They ask to you to bring order to their chaos and develop a plan to help them more easily compare data between centers. Other data centers in other part of Canada would of course benefit if any solution you developed could be implemented or integrated within individual groups, which requires appropriate methods to share data generated across the different centers. Finally, the first manuscript is almost ready for publication, and there is disagreement about what and how much supplementary information to include.

You need to come up with a plan of how to solve the data/information management problems starting from when the data comes off your collaborators experimental machines to when visualizations are produced by the analysis software, which may involve integration of several components, software technologies, and data standards that you could use, extend and/or develop *de novo* to help your group cope.

Leaning Outcomes

- Understand importance of data standards & data integration
- Familiarity with terminology in data standards, integration & sharing
- Awareness of challenges in data integration, data standards and data sharing
- Knowledge of mechanisms to address these challenges

Questions to consider

(1) Terminology familiarity - definitions, examples, utility

Data standard
Databases (relational and else)
Data integration
Distributed data
Client
Server
Federated data
Registry
Ontology
Web service
Data
Meta data
Application Program Interface
XML
Schema

Upon completing this case, students should be familiar with data standards and integration efforts within the bioinformatics community, using answers to the discussion questions below as a guide:

1. What examples of bioinformatics data standards can you think of?

2. What kind of file formats are useful for implementing data standards and why? Which would be useful for different components in this pipeline?

3. How do you ensure the information on instrument settings and setup and sample information should be made available to someone analyzing a data file? What are some alternatives for storing these different information types for this problem and their relative advantages and disadvantages? Which would you chose for this problem and why?

4. Should raw data go into a database or file system? What about metadata? Why? How can you connect these

5. Would ontologies be useful? If so, how you get an ontology to work with your data (i.e., how can people use one when sitting in front of a computer to work with their data)? Where does it plug in and how (i.e., through what software hooks or components). In the case of alternative solution for managing data annotations, consider the mechanics of how this could be implemented. Much of this should have already been covered in previous class, but apply it here.

6. What is a Web service and how does it work?

7. Are there existing bioinformatics/genomics/high throughput data analysis pipeline tools you could adapt to this problem?

8. What are some alternatives programs for high throughput data analysis and what are the advantages/disadvantages of using these?

9. What are the advantages/disadvantages of federated versus centralized databases and their requirements?