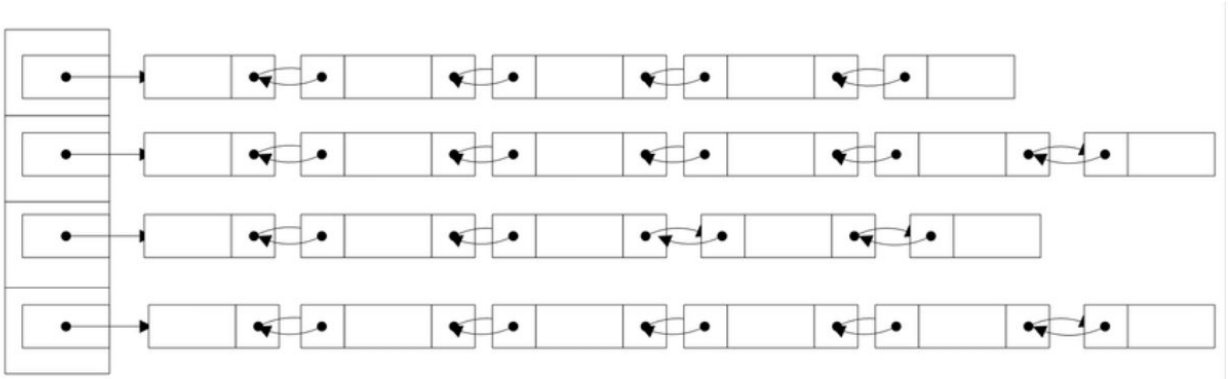
BioData Hackathon Project Plan

Overview: pipeline with nodal structure: Use of bucket data structure/linked list



A bucket data structure.

Operating Procedure for new pipeline:

1. Run CyTOF pipeline as is, generate key features from data (HAL).

2. Incorporate these features into a node list with keys that include tissue type, cellular phenotype, marker type, disease status, etc.

3. Map these nodes to “adjacent” nodes in the Tab. M. data; these will be datasets with minimum 1 degree of similarity (i.e (same tissue, same cell type), (different tissue, same cell type)).
    I. Heuristic to determine the best Tab. M. nodes: **expression pattern similarity**. This will tell which Tab. M. data best correlates to the cellular trends that had classification success within the corresponding CyTOF nodes. Using pattern similarity, we will classify the Tab. M node based on its likelihood of being related to the CyTOF node using a modified classifier that is customized to handle cellular data (if needed, general structure can come from sklearn source code, for convenience).

        a. CyTOF conclusion validations: We can compare diseased nodes to healthy nodes with no other conditions varied, analyzing and classifying the relationship between nodes. (for example, 85% prob related, 15% prob is not related). If the conclusion is deviant from the expected relationship based on the node keys, we can analyze further. This is basic (and hopefully fast) neighbour algorithm.

b. This will allow for whole body consideration: challenge phenotypes, establish connections between tissues

II. Statistical validation. We need to be careful about this, as our data dimensionality is going to increase significantly.
    a. Extensive cross validation
    b. Statistical multiple test correction measures
    c. High ‘significance’ value when algorithm is assessing if a conclusion is valid

In this manner, we may challenge conventional phenotypes and establish new physiological connections within a disease, as well as more carefully validate our CyTOF conclusions.

Specific Scripts Needed:

1. CyTOF pipeline
2. Cellular phenotyping algorithm (decision tree structure)
3. Script(s) to convert to nodes with specific keys that are compatible with Tab.M labels *
4. Script(s) to identify adjacent nodes in Tab. M data*
5. Script to handle nodal interactions*
6. Modified classifier to analyze the relationship between nodes
7. Script for statistical variation
8. Script to check marker/gene relationships
9. Final script to save/return all relevant data

* May be most efficient to implement in Java or C and incorporate with python. Quicker as well