GENERAL

Course web page: https://alpclk.github.io//mlabst/index
Discussion forum: https://www.synapse.org/#!Synapse:syn10153359/discussion/default

BIOINFO: DATA SOURCES FOR THE ASSIGNMENT

AMP-AD data
How to Access Synapse:
http://docs.synapse.org/articles/getting_started.html#installing-synapse-clients
AMP-AD description: https://www.synapse.org/#!Synapse:syn2580853/wiki/409846
RNA-seq data suggested (can choose also from the two other consortia):
https://www.synapse.org/#!Synapse:syn3163039

- Thanneer's lecture covered raw data acquisition and discussed normalization and covariates adjustment - if you combine results across three studies, consider these before deconvolution (molecular task) or choose appropriate differential expression result as basis for your analysis (imaging task) .

- **Scientific hypothesis:** Differential enrichment of different cell types are the results of neuro-degenerative diseases.
- **Question**: Which cell types are up regulated, down regulated and unchanged in AD
- **Data**: RNASeq from postmortem tissue samples collected across 7 different brain regions. Associated metadata (clinical and technical)

- Molecular data to integrate : single cell RNA-seq
- Imaging data to integrate: Allen Brain atlas in situ hybridization

---

MOLECULAR DATA

Cell type deconvolution task

The aim is to predict the proportion of different cell types from bulk RNASeq using clever approach utilizing data from single cell RNASeq

Background reading

Deconvolution problem

Biologically motivated review https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3874291/
http://web.cbio.uct.ac.za/~renaud/CRAN/web/CellMix/_PAGE-Algorithms.html

Technically motivated review https://arxiv.org/pdf/1510.04583.pdf

CIBERSORT method covered in Petri's lecture
http://www.nature.com/nmeth/journal/v12/n5/abs/nmeth.3337.html

This R package supporting RNA-seq data could be helpful
https://bioconductor.org/packages/devel/bioc/vignettes/DeconRNASeq/inst/doc/DeconRNASeq.pdf

Single cell RNA-seq

Data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67835
Code: https://github.com/seandavi/awesome-single-cell
Neural network application:
https://academic.oup.com/nar/article/doi/10.1093/nar/gkx681/4056711/Using-neural-networks-for-reducing-the-dimensions

Proposed solution

Option 1

You may use an existing implementation of cell type deconvolution and focus on innovating how to use scRNAseq in a clever way e.g. to refine the cell type identification (new marker genes). See above for suggested tools. One plausible approach could be to investigate unsupervised methods to get initial grouping of scRNAseq data into similar cells (possibly representing " cell types / states"). To quantify how well these groupings reflect the original sample grouping (in high dimensional space) consider the metrics presented by Juha.

Option 2

You may focus on proposing a new solution to the problem using neural networks.
One solution could be to start with an autoencoder model that learns cell type classification given scRNAseq data.

The simplest such model could be one that gets as input the expression values and as output attempts to reconstruct these values and associate the levels with predicted cell type. This thesis work addressed this problem and could give you insight into the problem: http://www.diva-portal.org/smash/get/diva2:942241/FULLTEXT01.pdf

You would have similar data from human brain as was used there from mouse

However, at this point your model does not yet solve the deconvolution problem. To pass the course, you would in minimum need to discuss how to proceed.

To get a good rank on the leaderboard, you should think a bit further. Given the single cell profiles, you could generate simulated data by pooling a known subset of cells together (e.g. reads from 50 neurons mixed with reads from 50 endothelial cells). This data and the known mixed proportions could serve as training data for training a deconvolution model. We will post more hints during the 2nd week how to proceed here!

Evaluation

- clever choice of methodology
- clever use of data
- demonstrating preliminary success, minimum requirement is that your approach is able to detect neuronal loss in AD

IMAGING DATA - Allen Brain Atlas

Background reading

http://www.nature.com/nbt/journal/v33/n5/abs/nbt.3209.html

Data

Allen Brain Atlas http://www.brain-map.org

Proposed solution

Option 1

You could study this example approach proposed in mouse image study with CNN as a starting point: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0553-9 and try to adapt this to human brain atlas, i.e. you would need to demonstrate that you can annotate gene expression patterns with human data.

To download images we have so far tested only one-by-one retrieval (you would need to find out if downloading all is even possible!), check instructions in:
http://help.brain-map.org/display/api/Downloading+an+Image
you need to find the desired imageIDs.

- Do any search in http://human.brain-map.org/ish/search
- Click a SpecimenID in the search results
- Click the experiment ID in the "Selection Information"-box
- There is a small button in the top right corner of the image. Howering the mouse over it displays a text "Launch a high resolution viewer in a new window". Click it.
- You can either download the images here using the download button or asinstructed in:
http://help.brain-map.org/display/api/Downloading+an+Image
and download the image using the imageId shown in the web address of this window.
- For every ISH-image there is an image showing gene expression areas and a Nissil-reference image of the shown area. Those images can also be viewed in this window.

Properties of search interface are explained here http://human.brain-map.org/ish/search

The next step (which you can complete even w/o image data) is to critically evaluate the example approach proposed in mouse image study with CNN:
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0553-9
To get it up and running for doing some tests, here some hints:

- VLFeat can be used via Matlab: http://www.vlfeat.org/install-matlab.html
- SIFT-algorithm to extract feature vectors. User guide:
http://www.vlfeat.org/overview/sift.html

- Bag-of-words model generated from the image using random sampling + obtained feature vectors -> K-means-algorithm -> visual codebook
- The OverFeat-neural network structure and source code is available in github (https://github.com/sermanet/OverFeat). It comes with python and torch APIs.
- The model comes with pre-trained weights.
- Example of the usage (https://github.com/sermanet/OverFeat/blob/master/API/python/sample.py)
- The OverFeat-model's input need to be resized or cropped to 231x231 and when using python it needs to be given to the model in a numpy array with dtype=numpy.float32 in the shape of 3xHxW (RGB image with size HxW). Images with the size HxWx3 can be transposed to desired shape.

Option 2

Implement your own network model, motivated by earlier literature and what you learned so far. Here you would also need to describe how to obtain training and test data (but if your focus is on method development, we will not require so much time spent getting the data, unless you aim for the leader board where we will give higher rank to full solutions)
- For gene localization check above instructions for Specimen section's expression images (ground truth)
- Implement network e.g. in python using keras with tensorflow. You can make something like the OverFeat model's architechture (Figure 2 in https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0553-9) or implement your own.
- Possible tricky parts: shape of tissue sections varies and also location where section was cut from. You might need to normalize your data somehow. You should propose some solutions here.
- Compared to Nissil-reference images some of the Allen Brain Atlas images are very dim -> you may need to smooth the color data. Again, propose some solutions here.

To get a good rank on the leaderboard, you should benchmark your solution with simulations and/or real data.

Evaluation

- clever choice of methodology
- clever use of data
- demonstrating preliminary success, minimum requirement is that your approach is able to detect neuronal loss in AD