PATH 828 – Bioinformatics for Cancer Research

The course is designed to train biomedical scientists in proper data analysis techniques and statistical approaches to enable students to perform their own data analysis, regardless of data type, using statistics, data mining and machine learning methods. It consists of four assignments, each addressing a specific aspect of data analysis such as data preprocessing, sequence alignment, statistics and feature selection. The course project requires students to apply all the skills learned through the assignments and lectures and perform data analysis from scratch on data of their choice. Students may use data from their own research or a publicly available data set that is relevant to their thesis.

The Fall 2020 course will be delivered fully online.

PATH 828 course syllabus

1. **Introduction**
   General introduction, course outline, course schedule, grading, required tools, resources

2. **Data and data processing**
   a) Brief description of different types of data: microarrays, RNA-seq, miRNA-seq, qRT-PCR, NanoString, WGS, 16S rRNA, proteomic, flow cytometry, clinical/pathological, etc
   b) Public data repositories
      a. TCGA, ICGC, GEO, proteomics, etc
   c) Basic programming skills – in MATLAB
      a. Data variables: numeric (integer, double, boolean), string, character, arrays, matrices, cell arrays, struct
      b. Operators: mathematical, logical, relational
      c. If/else statements, loops, switch statement
      d. Functions/Scripts: how to create functions, how to pass variables
      e. Variable scope
      f. Software development cycle

g. Basic program validation and error/exception handling
h. How to use a debugger
i. Comments

d) Analysis tools
   a. How to choose appropriate tools, learn their main advantages and disadvantages
   b. Be aware of free programs/packages
   c. Resources for some of the available tools for data analysis

e) Data processing
   a. Pipelines for data processing of: microarrays, RNA-seq, miRNA-seq, NanoSting, 16S rRNA, etc
   b. Generic data pre-processing pipeline
      i. Data visualization
         1. perform and interpret basic visualization techniques such as boxplots, scatter plots, clustering
         2. How to check for outliers and batch effects
         3. How to deal with outliers and batch effects
      ii. Normalization – for RNA-seq, miRNA-seq, NanoString, qRT-PCR
      iii. Standardization
      iv. Filtering of features with low expressed and/or low variability

f) Study design
   a. minimizing technical variability
   b. following simple sample randomization practices
   c. replicates in data collection

3. Center for Advanced Computing (CAC) resources
   a. Advantages of having and using a CAC account
   b. How to create an account
   c. How to login
   d. Execute basic Unix commands for working with directory and files
   e. How to upload or download files to/from the CAC account

4. Machine Learning
   a. Motivation
   b. Visualization
      1) Basic, one/ two dimensional
      2) High dimensional

1. Parallel Coordinates plots
2. Andrews plots
3. Glyphplot
4. Heatmaps/hierarchical clustering
5. Self-Organizing Maps
6. Principle Component Analysis (PCA)
7. Multidimensional Scaling (MDS)
8. T-SNE

c. Dimensionality reduction
d. Unsupervised learning / clustering analysis
    1) Hierarchical clustering
    2) Self Organizing Map (SOM)
    3) K-mean clustering
e. Supervised learning / classification
    1) Terminology
    2) Performance evaluation
    3) Dimensionality/overfitting/Cross-validation
    4) Feature selection
    5) Feature reduction
    6) Different families of classifier: Tree-base, SVM, k-means, etc
    7) No free-lunch theorem
f. Deep Learning
g. Application in MATLAB

5. Introductory statistics with application in SPSS

a. Random sample
b. Variable types: categorical, nominal, ordinal and continuous variables.
c. Summarizing data: visualization, mean, median, quantiles, min, max, mode, range, standard deviation, interquartile range,
d. Distributions: normal, skewed, Chi-squared,
e. Basic terms: type1 error (alpha), type 2 error (betta), p-value, Ho (null hypothesis), Ha (alternative hypothesis), false discovery rate (FDR), parametric vs non-parametric tests
f. Test assumptions
g. Statistical tests
    1) Descriptive statistics

2) T-Test / Mann-Whitney U test / Wilcoxon ranked sum test
3) ANOVA / Kruskal-Wallis test
4) Chi2 test
5) Regression (linear, logistic, multiple)
6) Correlation (Pearson, Spearman)
7) Survival analysis and Cox (hazard ratio)
8) Multivariable analysis
9) Power analysis

h. Practical in SPSS, how to perform, interpret and properly report the results of various statistical tests
i. Differential expression
j. Fold change

6. Paper critiques review

Course Evaluation

Assignments (4) – 44%
Paper critique (1) – 6%
RAT and class participation/engagement – 5%
Course project – 45%
    Proposal – 5%
    Presentation – 10%
    Final report – 30%

Assignment1 – RNAseq data alignment
Assignment2 – Data pre-processing
Assignment3 – Feature selection
Assignment4 – Statistical analysis

Course project requirements
Using your own data set or any other publicly available data perform full data analysis that consists of:

Data visualization
Data preparation
    normalization (if necessary), checking/correcting for outliers/batch effects, filtering
Analysis using at least 2 statistical tests from the tests that we learned (e.g. Survival analysis and hazard ratio analysis, Kruskal Wallis test, Correlation, etc)
Analysis using at least one Unsupervised learning approach and at least one Supervised learning approach

Learning outcomes
By the end of this course, students will be able to:

- describe and follow a general data analysis pipeline
- perform sequence alignment for RNAseq, miRNAseq
- perform, interpret and properly report results of most commonly used statistical tests
- perform basic programming in MATLAB
- perform unsupervised and supervised analyses
- critic computational methods in the literature