As a Biomedical Data Scientist working with researchers and students in the Department of Pathology and Molecular Medicine, I became aware of the types of genetic data analysis required. I also became aware of the gap in student's educational knowledge when it comes to genetic data processing and analysis. Therefore, I proposed to introduce a graduate course that would teach and train the new generation of biomedical scientists proper data analysis techniques and better statistical knowledge and practice. In this course, I want students to develop habits of mind that data analysis should be part of the experiment design, not, as it often is, an after-thought.

The course is designed to enable students to perform their own data analysis, regardless of genetic 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 the students to apply all the skills learned through the assignments and lectures and perform data analysis from scratch on data of their choice. Students are encouraged to use data from their own graduate work or a publicly available cancer data set that is relevant to their thesis.

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 genetic data: microarrays, RNA-seq, miRNA-seq, qRT-PCR, NanoString, WGS, 16S rRNA, clinical/Pathological
   b) Public data repositories
      a. TCGA, ICGC, GEO, proteomics,
   c) 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
   d) Data processing
      a. Pipelines for data processing of: microarrays, RNA-seq, miRNA-seq, NanoSting, 16S rRNA
      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 expression and/or low variability
   e) Study design
      a. minimize technical variability
      b. following simple sample randomization practices
      c. replicates in data collection
   f) 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

3. **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) Multivariate 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

4. Center for Advanced Computing (CAC) resources – guest lecture by Jeff Stafford

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

5. Data Mining

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. Application in MATLAB

6. **Next generation sequencing and mutation analysis**
   Variant Calling, quest lecture by Elina K. Cook

Course Evaluation

Assignments (4) – 40%
Paper critique (2) – 10%
Class participation/engagement – 5%
Course project – 45%
Proposal – 5%
Presentation – 10%
Final report – 30%

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

Course project requirements

Using your own data set or any other publically available data perform full data analysis that consists of:

- Data visualization
- Preprocessing (normalization (if necessary), filtering)
- Checking/correcting for outliers/batch effects
- 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 data mining approach and at least one Supervised learning data mining 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

Comments on course evaluation form for PATH 828

“I enjoyed the breadth and material covered and the assignments that had us apply what we have learned. Great course overall!”

“Amazing course, best I have ever taken. Difficult and challenged me but to my benefit. Would recommend to any graduate student exploring bioinformatics”

“I like the amount of resources provided. Even if we didn’t have time to cover things very in-depth in class, many resources were provided so that we could learn more about topics on our own. The assignments were very well-planned to prepare us for conducting a full analysis. This course should be mandatory for all students! Instructor and TA were extremely knowledgeable and always available to assist”

“The course was a great introduction to bioinformatics. Both the professor and the TA presented the material clearly and simply for someone with little to no prior background in bioinformatics. I especially liked the in-class tutorials because we got to work on the material together as a group”

“I really enjoyed this course. I liked that everything we learned was relevant and useful. I also really liked that we were able to incorporate our own data into the final project for the course. I also enjoyed the content and hands on learning involved with this course; it was quite different than any other course I’ve taken previously”

“I really enjoyed the depth of material covered and how the data analyses were put into context. Starting with no prior knowledge in this field I learned an enormous amount in 12 weeks thanks to great teaching and assignments that created the best learning opportunity”

“I learned a lot. This was one of the most useful courses that I’ve ever taken. Some of the results from my final project will go into my thesis”