INTRODUCTION TO DATA MINING AND GENOMICS

**Materials – readings and videos** (carefully curated by Dr. Welch to help students to learn the important concepts for week 1):

* Dr. Welch’s Overview of the Data Mining Course
  + <https://www.youtube.com/watch?v=DyYAgXgVaTc&feature=youtu.be>
* Data mining:
  + Chapter 1 - Introduction to data mining, *Discovering Knowledge in Data*. D.R. Larose and C.D. Larose. Wiley. 2014.

<https://alice.library.ohio.edu/record=b5187242?> (click on the link “Connect to resource OhioLink”)

* Genomics:
  + <https://www.genome.gov/About-Genomics/Introduction-to-Genomics>
  + <http://www.chromosomewalk.ch/en/list-of-chromosomes/>
  + <http://www.chromosomewalk.ch/en/what-are-we-made-out-of/>
* Genetic mutations:
  + <https://www.khanacademy.org/science/biology/x324d1dcc:metabolism/x324d1dcc:mutations/v/an-introduction-to-genetic-mutations>
* TCGA:
  + <https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas>

**Concepts to learn from the materials**:

1. The definition of the term *data mining*.
2. The *phases* of data mining.
3. *Fallacies* of data mining.
4. Basic *genomics concepts*, including the following:
5. Chromosome
6. DNA
7. Nucleotide
8. Gene
9. Genome
10. Genetic mutation
11. Protein
12. Amino acid
13. RNA
14. Ribosome
15. Genetic disease
16. The definition of the term *precision medicine*.
17. What is **TCGA**. What is the overarching goal of **TCGA**.

**Quiz**:

After learning the concepts listed above, complete the blackboard quiz no later than Wednesday August 25, 9:39 am. The quiz will cover your understanding of the *concepts to learn from the materials* (see above). The quiz may include multiple choice, true-false, fill-in-the-blank, and/or matching questions.

**Supplementary Materials – readings and videos** (carefully curated by Dr. Welch for students who would like to delve more deeply into the important concepts for week 1):

* Data mining:
  + <https://en.wikipedia.org/wiki/Data_mining>
* Genomics:
  + <https://www.youtube.com/watch?v=_xJXZBCOWMY>
* Genetic mutations:
  + <https://en.wikipedia.org/wiki/Genetics>
* TCGA:
  + <https://en.wikipedia.org/wiki/The_Cancer_Genome_Atlas>

Data Mining Activity: (*to be started after you complete the quiz*)

**Due date**: no later than Thursday August 26, 11:59 pm

(submit by email to welch@ohio.edu)

Throughout the course, we will analyze a data file that lists the genetic mutations of individuals who participated in a research study (the data file is provided on Blackboard in the week 1 folder). We will employ data mining techniques to discover possible genetic causes of cancer by analyzing the information in the data file. This week, you will get familiar with the contents of the file and will perform exploratory analysis of the data.

Some individuals who participated in the research study have cancer and some individuals do not have cancer. Each row of the file contains the binary *mutation vector* for one individual. The first element in each row is a unique identifier, which consists of

1. a class identifier:
   * ‘C’ if the individual has cancer,
   * ‘NC’ if the individual does not have cancer
2. a numeric patient identifier (patients in each class are assigned a unique number).

Each remaining element in a row contains a ‘1’ or a ‘0’, to indicate the presence or absence, respectively, of a specific mutation in a particular gene.

The first row of the matrix lists the names of the specific genetic mutations considered in the study.

A portion of the matrix (containing the first 3 individuals and the first 3 genetic mutations) is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **FARP1\_GRCh37\_13:99092237-99092237\_Frame-Shift-Del\_DEL\_G-G--** | **QKI\_GRCh37\_6:163987695-163987695\_Intron\_DEL\_T-T--** | **NHLRC2\_GRCh37\_10:115662308-115662308\_Frame-Shift-Del\_DEL\_A-A--** |
| **C1** | 0 | 0 | 0 |
| **C2** | 1 | 1 | 1 |
| **NC1** | 0 | 0 | 0 |

The first row indicates the genetic mutations for individual ‘C1’, which corresponds to person number 1 in the group that has cancer.

The third row lists the genetic mutations for individual ‘NC1’, which corresponds to person number 1 in the group that does not have cancer.

Each column in the matrix represents a specific mutation in a particular gene. Column one represents a mutation in the gene named ‘FARP1’, and columns two and three represent mutations in the ‘QKI’ and ‘NHLRC2’ genes, respectively.

Individual ‘C2’ has mutations in the ‘FARP1’, ‘QKI’ and ‘NHLRC2’ genes.

Individuals ‘C1’ and ‘NC1’ do not have mutations in the ‘FARP1’, ‘QKI’ and ‘NHLRC2’ genes.

Obtain the data file from blackboard and calculate the following (either use a spreadsheet program (e.g., Excel) or write a computer program):

* Number of unique mutations contained in the data file (each column represents a unique mutation)
* Number of individual samples contained in the data file (each row represents a patient sample)
* Number of mutations for individual C1
* Number of mutations for individual NC1
* Average number of mutations per individual
* The minimum and maximum number of mutations per individual
* Number of individuals who have a mutation in the ‘FARP1’ gene
* Number of individuals who have a mutation in the ‘QKI’ gene
* Average number of individuals per mutation (i.e., the number of individuals expected to have a randomly selected mutation). This can be computed as follows:
  + For each mutation *i*,compute *mi*, the number of individuals who have mutation *i*
  + Compute *m\_sum* = the sum of *mi* for all *i*
  + The average number of individuals per mutation = *m\_sum / number of mutations*
* The minimum and maximum number of individuals per mutation

Submit an email to [welch@ohio.edu](mailto:welch@ohio.edu) that contains a brief report of the computed values. Additionally, attach either (a) the .csv file (it should contain the original data, along with the calculated values), or (b) the computer program that you developed for this activity and the output of your program (either a screenshot or a file).

**NOTE**: I may respond to your email submissions with questions about your methods, results, and/or interpretation. Please respond promptly to my questions.