Appendix 1  
Instructions of the Programs used in the Final Project

Tai-Hsien Ou Yang and Kaiyi Zhu

Department of Electrical Engineering

Columbia University

New York, NY, USA

e-mail: {to2232, kz2232}@columbia.edu

1. **The Concordance Index (CI) Toolkit**

The source code and the executable jar is located at ./CI/

* 1. Compute the CI of the genes in TCGA data set
     1. Download the pancan12 gene expression matrix from <http://128.59.65.250/html/pancan12.ge.rda>

and make sure Hadoop HDFS is properly running.

* + 1. Run ci\_preprocess.R in R to convert the features into CSV files.

source("ci\_preprocess.R")

* + 1. (Optional) Compile the java source code

ant compile jar

* + 1. Convert the CSV files into SeqenceFiles.

cd hashedfeature/

java -cp "../ci/lib/\*" corr.util.DummyDataToSeqFile .

* + 1. Upload the sequenceFiles to HDFS /hash/

mkdir hashedfeature.seq

mv ./hashedfeature/\*.seq ./hashedfeature.seq

{$HADOOP\_HOME}/bin/hdfs dfs -mkdir /hash

{$HADOOP\_HOME}/bin/hdfs dfs -put ./hashedfeature.seq/\* /hash

* + 1. Cmpute the CI of a single Gene (Algorithm B)

To compute CI using Algorithm B, run

{$HADOOP\_HOME}/bin/hadoop jar ./lib/ci.jar corr.job.CIJob /hash/SUSD3.seq

* + 1. Compute the CI of multiple selected genes in list.txt (Algorithm B)

for file in $(<list.txt)

do

{$HADOOP\_HOME}/bin/hadoop jar ./lib/ci.jar corr.job.CIJob "/hash/${file}.seq"

done

* + 1. Compute the CI of a single Gene (Algorithm A)
       1. Run the following in shell  
          {$HADOOP\_HOME}/bin/hadoop jar ./lib/ci.jar corr.job.CIAJob /hash/SUSD3.seq

hdfs dfs -get /user/taihsien/\*

* + - 1. Run ci\_a.R in R in the result directory downloaded from HDFS.
    1. Download the result from HDFS using

{$HADOOP\_HOME}/bin/hdfs dfs -get /user/USER\_NAME/hash/\* /YOUR\_DEST/

The complete results are in hash\_ci\_12121804.tar.gz.

1. **Identification of the outcome-associated genes and the Bayesian Network**

The scripts are in ./Bayes/

* 1. Install bnlearn package in R.

install.packages("bnlearn")

* 1. Go to the resulting CI directory in the previous step and run topci\_bayes.R in R, the resulting D3.js graph will be printed on the console.
  2. The output files are:

ciList.csv: The CIs of all genes.

ciListList.top30.csv: The genes with Top 30 CIs.

pancan12\_comp\_ci.rda: The covariate for the treatment recommendation engine.

1. **Performance evaluation for the treatment recommendation engine**

The script is in ./Recommendation/

* 1. Run evaluator.R in R. Since the default similarity metric is Kendall's tau, if you would like to evaluate the performance of the Pearson correlation version, remove ",method="kendall" in the file.

1. **HBase data storage**

The programs are available at ./Hbase/

* 1. Create the schema
     1. Make sure the Hadoop-Hbase environment is setup properly and download TCGA data from <http://128.59.65.250/html/TCGA.csv>
     2. Start HBase (run in pseudo-distributed mode) and initialize a table

cd $HBASE\_HOME

./bin/start-hbase.sh

./bin/hbase shell

>create ‘TCGA’,’mRNA’

>list

* 1. Import data from a CSV file to a HBase table
     1. Start Eclipse and create a Java Project for HTableExample.java.
     2. Add External JARs from to build path. The list of required JARs can be downloaded from <http://128.59.65.250/html/lib.zip>
     3. Select imported data file TCGA.csv in the source code.  
        csvArray = readCSV(TCGA.csv");
     4. Run “HTableExample.java”, and the updated HBase table “TCGA” can be found using HBase shell.
  2. Export data from a HBase Table to a CSV file
     1. Set $JAVA\_HOME:

export JAVA\_HOME=/usr/lib/jvm/java-7-openjdk-amd64/

* + 1. Run pig file export.pig:

cd $PIG\_HOME

./bin/pig -x local ../export.pig

* + 1. The CSV formatted file(s) “part-m-0000X” will be found in folder {$PIG\_HOME}/gene-data.csv/
    2. To convert the exported CSV file into the Hash-sorting format, run ci\_preprocess.R in R to convert the features into CSV files.

1. **Web service deployment**The entire website is located at ./Web/
   1. Make sure R and Apache-PHP (LAMP) are installed and configured. R can be installed by sudo apt-get install R-base. And LAMP environment can be configured following the instruction:  
      <https://www.digitalocean.com/community/tutorials/how-to-install-linux-apache-mysql-php-lamp-stack-on-ubuntu>
   2. Move the files in ./Web/site to the Apache web page directory /var/www/html
   3. Move pancan12\_comp\_ci.rda to /var/www/ and ciList.csv to /var/www/result
   4. Change the permission of /result and /uploads to 755

chmod 755 /var/www/result

chmod 755 /var/www/uploads

* 1. The web service can be accessed at http://127.0.0.1/