Thesis

Sindhuja Sridharan

26^{th}December 2017

1 Introduction

Cancer is a disease of the genome: sequential accumulation of DNA mutations lead to both direct and indirect changes in the structure and abundance of the proteins that perform most of the functional activities within the cell. Thus the genetic changes that are the cause/origin of disease ultimately have their functional impact by changing the proteome which leads to tumour growth, spread, response to therapy and ultimately lethality.

While methodologies for the analysis of cancer genomes and transcriptomes have undergone rapid benchmarking and standardization, our understanding of how best to analyze the cancer proteome remains less-developed. In particular, there are key questions remaining in how to infer the abundances of peptides not detected in a subset of samples, in optimizing database searches to detect cancer-specific peptides caused by point-mutations, alternative transcript isoforms or fusion genes, in understanding the association between DNA, mRNA and protein data, and in performing robust absolute quantitation.

1.1 Predict Protein Abundances based on Genomics Information

Proteins are almost always the key functional biomolecule in the cell. Despite the central dogma of information transfer from RNA to protein, many groups have shown that RNA abundances are only weakly predictive of protein abundance (R2 = 0.1-0.4). Further, for cost and technical reasons, RNA and DNA data are both cheaper to generate and more widely available than protein data across the research community as a whole. As a result, functional inference would be greatly improved by robust models to predict protein abundance from RNA and DNA data. We will give participants an abundance matrix of K proteins along with a matrix of RNA expression and copy number alterations. The participants are asked to create models to predict the abundances of the K proteins from gene expression and copy number alterations.

2 Data Description

For this challenge, ovarian tumor samples quantitatively measured at four biological levels (proteomics, phosphoproteomics, transcriptomics (mRNA) and copy number alterations (CNA)) will be used as training data. Prospective samples of the cancer types with all four level measurement will be generated and used as testing data for performance evaluation. Sample size varies between different platforms due to the availability and quality of original tumor samples at the time of the study. The Mass-Spectrometry based proteomic and phosphoproteomic characterization of these tumor samples previously analyzed by TCGA yields more than hundred of thousand protein and phosphosite identifications combined, which will serve as the target to be predicted in the sub-challenges.

2.1 Overview of data set:

Ovarian cancer (subchallenge 2 and 3): Proteome from PNNL: 7061 proteins for 84 patients Proteome from JHU: 7061 proteins for 122 patients Phosphoproteome: 10057 phosphosites for 69 patients CNA: 11859 genes for 559 patients

mRNA(Array): 15121 genes for 569 patients mRNA(RNA-seq): 15121 genes for 294 patients

For ovarian proteome, there are 206 samples from 174 unique patients (84 from Pacific Northwest National Laboratory (PNNL), 122 from Johns Hopkins University (JHU) and 32 measured by both centers). We provide participants with both proteome collections for training to cover the maximum number of samples for subchallenge 2. However, for subchallenge 3, please note that ovarian phosphoproteome of 69 patients were measured exclusively by PNNL.

3 Methods: Regression Analysis

Regression analysis is a statistical analysis method to determine the quantitative relationship between two or more variables. On the basisof the number of independent variables, regression analysis can be divided intosimple regression analysis and multiple regression analysis. According to the type of relationship between the independent variables and the dependent variable, it can be divided into linear regression and nonlinear regression analysis. In regression analysis, only one independent variable and one dependent variable, and the relationship between the two can be represented bya linear approximation, which is called one-dimensional linear regression. If the regression analysis includes two or more than two independent variables, and the dependent variable and the independent variable is linear relationship, the linear regression analysis is called multiple linear regression.

3.1 Partial least squares regression

Partial least squares regression is an extension of the multiple linear regression model (see, e.g., Multiple Regression or General Stepwise Regression). In its simplest form, a linear model specifies the (linear) relationship between a dependent (response) variable Y, and a set of predictor variables, the X's, so that

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p$$

In this equation b0 is the regression coefficient for the intercept and the bi values are the regression coefficients (for variables 1 through p) computed from the data.

The multiple linear regression model has been extended in a number of ways to address more sophisticated data analysis problems. The multiple linear regression model serves as the basis for a number of multivariate methods such as discriminant analysis (i.e., the prediction of group membership from the levels of continuous predictor variables), principal components regression (i.e., the prediction of responses on the dependent variables from factors underlying the levels of the predictor variables), and canonical correlation (i.e., the prediction of factors underlying responses on the dependent variables from factors underlying the levels of the predictor variables). These multivariate methods all have two important properties in common. These methods impose restrictions such that (1) factors underlying the Y and X variables are extracted from the Y'Y and X'X matrices, respectively, and never from cross-product matrices involving both the Y and X variables, and (2) the number of prediction functions can never exceed the minimum of the number of Y variables and X variables.

Partial least squares regression extends multiple linear regression without imposing the restrictions employed by discriminant analysis, principal components regression, and canonical correlation. In partial least squares regression, prediction functions are represented by factors extracted from the Y'XX'Y matrix. The number of such prediction functions that can be extracted typically will exceed the maximum of the number of Y and X variables.

In short, partial least squares regression is probably the least restrictive of the various multivariate extensions of the multiple linear regression model. This flexibility allows it to be used in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than predictor variables. Furthermore, partial least squares regression can be used as an exploratory analysis tool to select suitable predictor variables and to identify outliers before classical linear regression.

3.2 Principal Component Regression

Principal Components Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, principal components regression reduces the standard errors. It is hoped that the net effect will be to give more reliable estimates.

3.3 Procedure

From the ovarian cancer data,Only the RNA and Protein data were downloaded.

Mean imputation was done on the protein and RNA data.Mean imputation is a method in which the missing value on a certain variable is replaced by the mean of the available cases. This method maintains the sample size and is easy to use, but the variability in the data is reduced, so that the standard deviations and the variance estimates tend to be were underestimated.Then the data was normalized.

We tried to fit a PCR and PLS model for the data. Then the MSE results of PCR and PLS model were compared to the MSE obtained from PCR and PLS models after shuffling the RNA data. If we shuffle the RNA data randomly, then we will destroy any association between the data.

Further then the fold change was calculated.Fold change is calculated as the ratio of the difference between the final value and the initial value over the original value.

The y true and y predicted values were compared and found out that y predicted has a range of -1 to 1 but y true ranges from -2 to 2. So to get the prediction of the whole range and to find the solution for scaling issues we start with the raw anlysis again.

Since there were scaling issues with the rna data, we again started to analyze the raw data Histograms with the density lines were plotted column wise for the

rna and protein data.

From the histograms the protein data may follow the normal distribution.

The mean and variance of the RNA and protein were found out out and after that the mean data was sorted and plotted.The low variance filter was done by eliminating variance values less than .5

Then the PCA scree and loadings plots were plotted after the scaling issue was solved.

3.4 Results

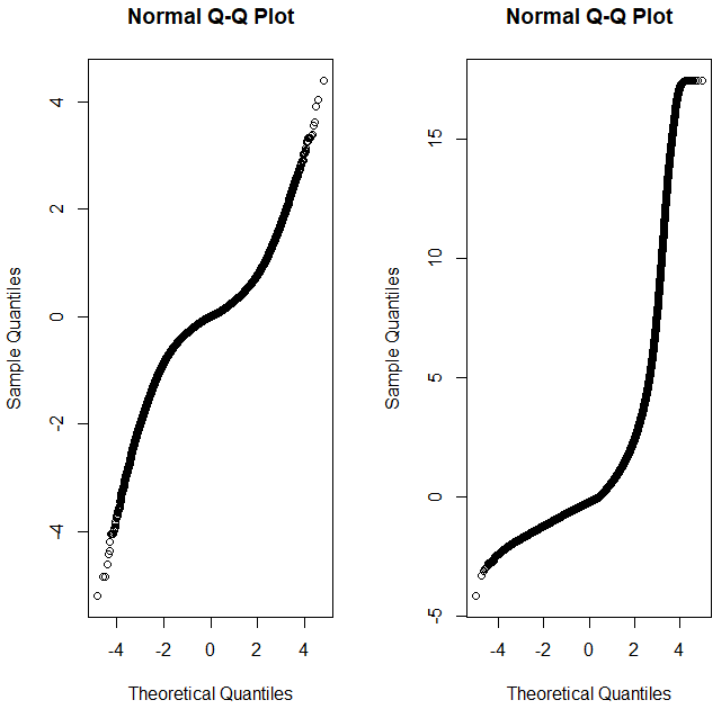
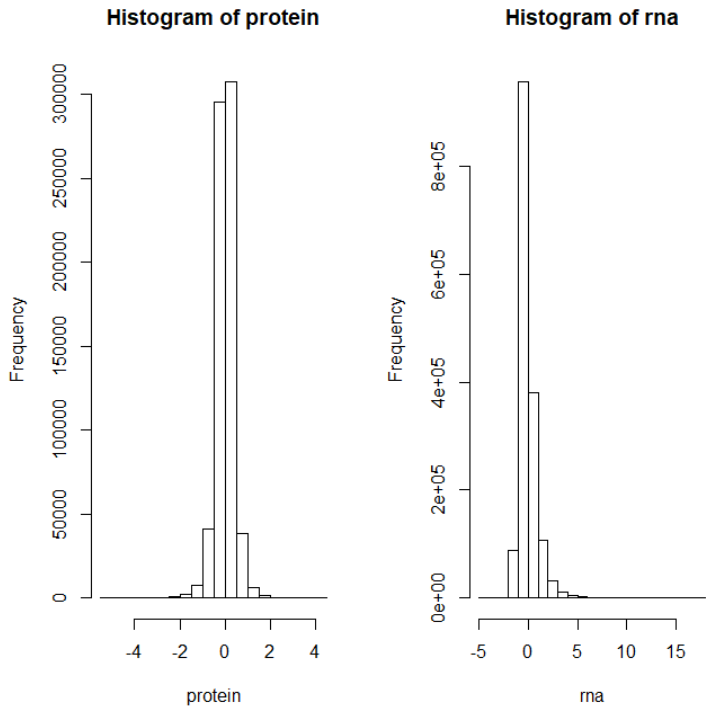


Figure 1: QQplots

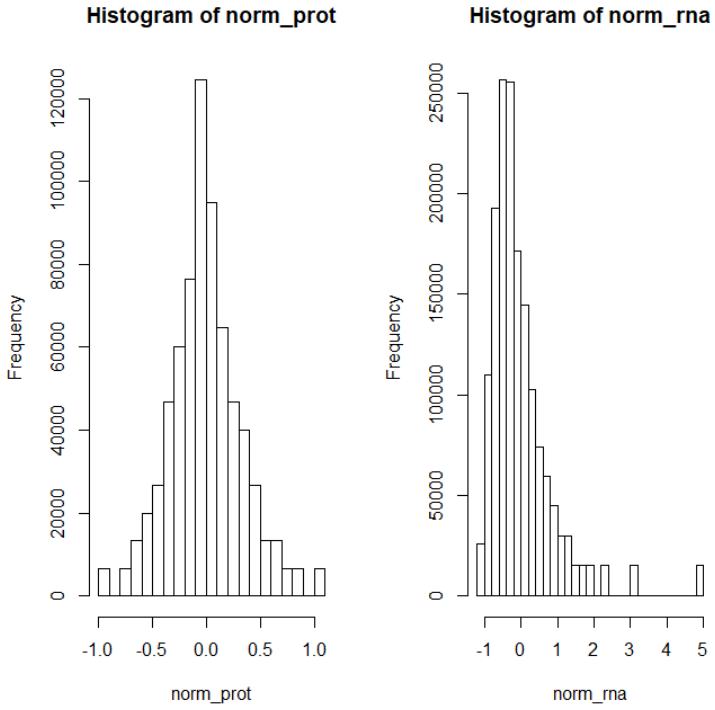
Despite the central dogma of information transfer from RNA to protein, it



data.png

Figure 2: Imputed data

shows that RNA abundances are weakly predictive of protein abundance.



data.png

Figure 3: Normalized data

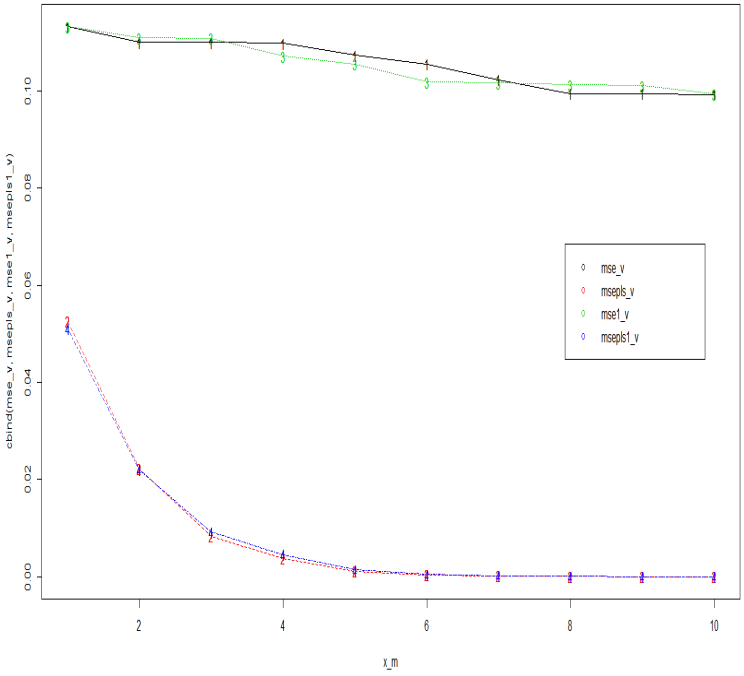

cbind(mse_v, msepls_v, mse1_v, msepls1_v)
0.00
0.02
0.04
0.06
0.08
0.10
2
4
6
8
10
x_m
mse_v
mseplS_v
mse1_v
mseplS1_v

Figure 4: MSE

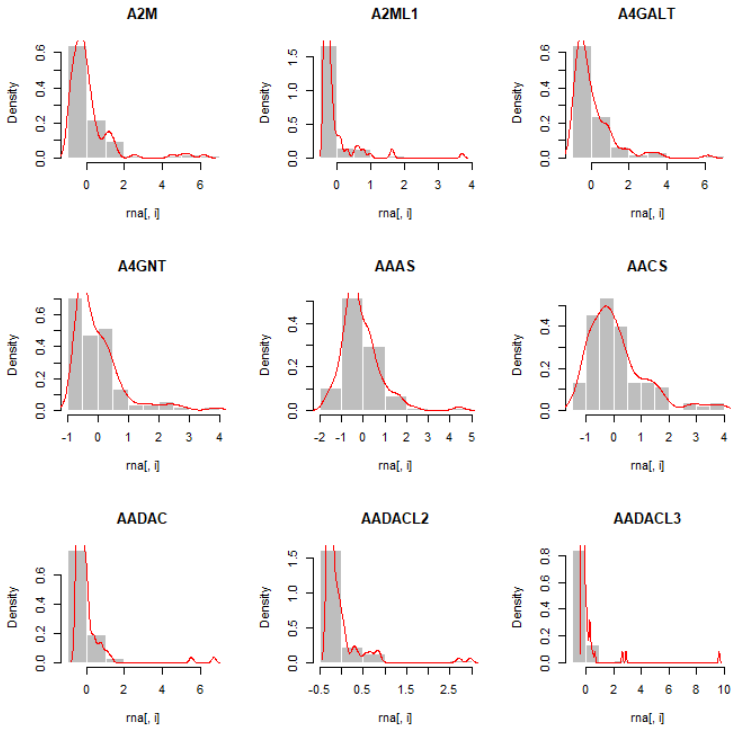


Figure 5: Rawrna

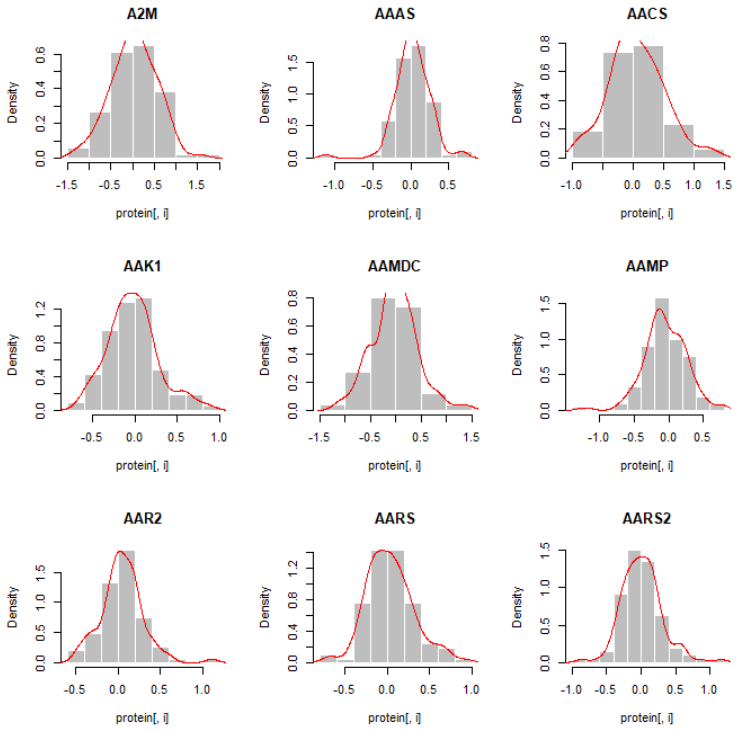


Figure 6: rawprotein

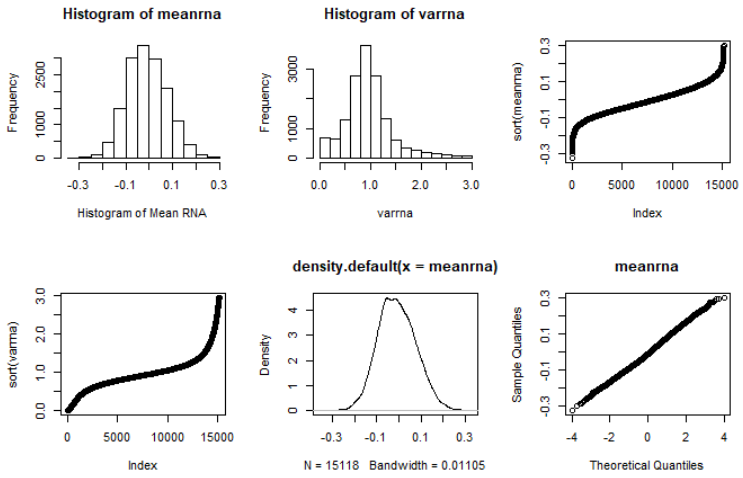


Figure 7: MeanRna

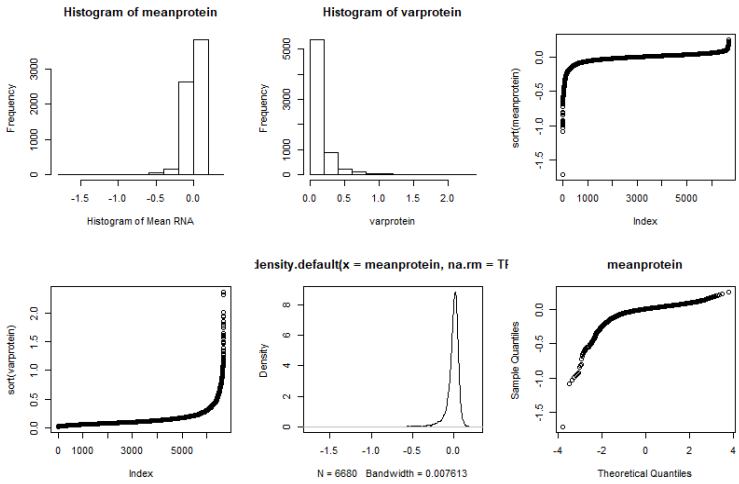


Figure 8: proteinmean

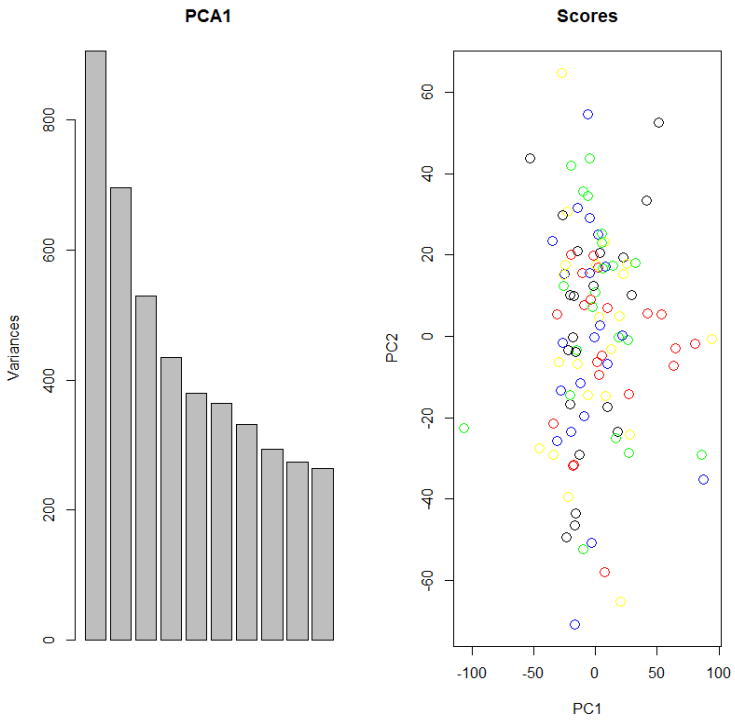

PCA1
Variances
0
200
400
600
800
Scores
PC2
-60
-40
-20
0
20
40
60
-100
-50
0
50
100
PC1

Figure 9: PCA