ABDULLAH FAROUK

PREDICTING PROTEIN LEVELS USING MRNA LEVELS

April 16, 2018

Student ID: 71244370
University of British Columbia
Department of Statistics

SUMMARY

Measuring protein levels is challenging due to the complex biological processes that govern its formation in cells. This makes it difficult to gather data for proteomics studies. One potential solution to this problem, as suggested by Wilhelm et al., is to use gene expression data as a proxy for protein levels. In this report we asses the merits of this claim. First, we fit a linear mixed effects model to determine if there is a linear association between protein levels and mRNA levels within each gene across the 20 tissue/cell line samples in the datasets under study. We then predict protein levels for each gene using the model suggested by Wilhelm et al.. While we find a significant association between mRNA and protein levels, using mRNA levels alone results in poor prediction of protein levels within genes.

INTRODUCTION

A very widely debated issue in Molecular Biology is the use of mRNA levels of a particular gene to predict its corresponding protein levels. If this is possible, it would allow protein prediction studies to be carried out using genome wide transcriptomics data.

In this report we seek to test the claim made above. We do this in two parts;

- First we asses if mRNA levels can explain the variation observed in protein levels for a given gene-tissue combination.
- We then test the model proposed by Wilhelm et al. which predicts protein levels across gene-tissue combinations using modified mRNA levels.

In the upcoming sections we proceed as follows. First, we describe the datasets under study. We then detail the methods used to verify each of the points above. Finally we discuss our results and their interpretability in light of the limitations of the datasets provided.

PROPOSED STATISTICAL METHODS

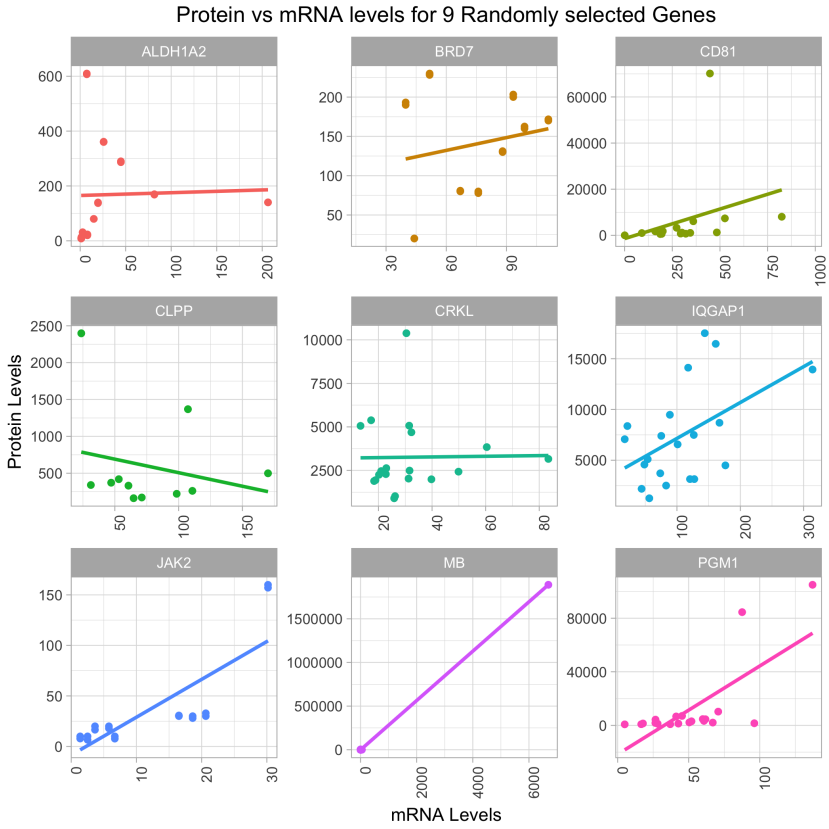
Data Description

There are three datasets under analysis. Each of them are tables with 55 rows and 21 columns. The rows correspond to genes and the columns correspond to tissue (11 samples) and cell line (9 samples) samples . The first table has data on TPM values (proxies for mRNA levels),

whilst the second table contains data on protein levels. The third table contains values of the gene specific conversion factor (RTP), for every gene-tissue combination possible. We also notice a large number of missing observations in the datasets of RTP and protein levels. Approximately 20% (224 out 1100) of gene-tissue combinations are missing values.

Linear Mixed Effects Models (LME)

In this section, we seek to identify if there is a linear association between protein and mRNA levels for each gene in our datasets. To check if this is the case, we plot the relationship between these variables for a few randomly selected genes.



The plots indicate that the relationship between the 2 variables are approximately linear across the genes plotted. They also reveal some interesting insights about our data. First, we note that some plots have fewer points due to missing protein values in specific gene-tissue combinations for that particular gene. Second, we notice that the fitted lines in the diagrams above ***do not*** go through most points in the plot but are pulled instead towards points at the

extreme top or bottom right corners of the plots. This suggests that these datasets contain influential outlying observations that can pull the line of best fit towards it. Now that we have visually established that the relationship between these two variables is linear, we move on to the model fitting process.

LMEs are extensions of simple linear models that allow us to use both fixed and random effects. A fixed effect is a variable that affects our response variable in a systematic and predictable manner. Random effects on the other hand are effects that we expect to have a non-systematic or random influence on our response variable. We feel LMEs are an appropriate choice because

- We have 20 observations of protein and mRNA levels for each gene.
- The 20 observations for each gene are not independent. This is because some tissue and cell lines are related to one and another (e.g. HEK and Kidney).
- From the plot above it seems like each gene requires a different intercept and slope.
- Some genes are missing protein values in specific tissue/cell samples. LMEs let us use these genes in our model by dropping only those observations missing for that gene.

In our LME model we treat genes as random effects and mRNA levels for each gene as fixed effects. Furthermore, our model is a random slope random intercept model. That is, each gene is fitted with a different slope and intercept under our model.

We find mRNA levels to be statistically significant in explaining the variability observed in protein levels within genes. To test the statistical significance of mRNA levels we used a likelihood ratio test. This test is based on the concept of likelihood (i.e. how likely are we to observe the data under study if our model actually generated the data at hand). It works as follows. First, it computes the likelihood of observing the data under the model without mRNA levels, then the model with it. It then checks if the difference between the likelihood of these two models is significant. If it is then it finds the variable of interest significant.

We feel skeptical about the findings of our model. This is because our initial data exploration showed potential influential observations in our dataset along with a large number of missing observations for some genes. We were not able to deal with these issues as they lie outside the scope of this report.

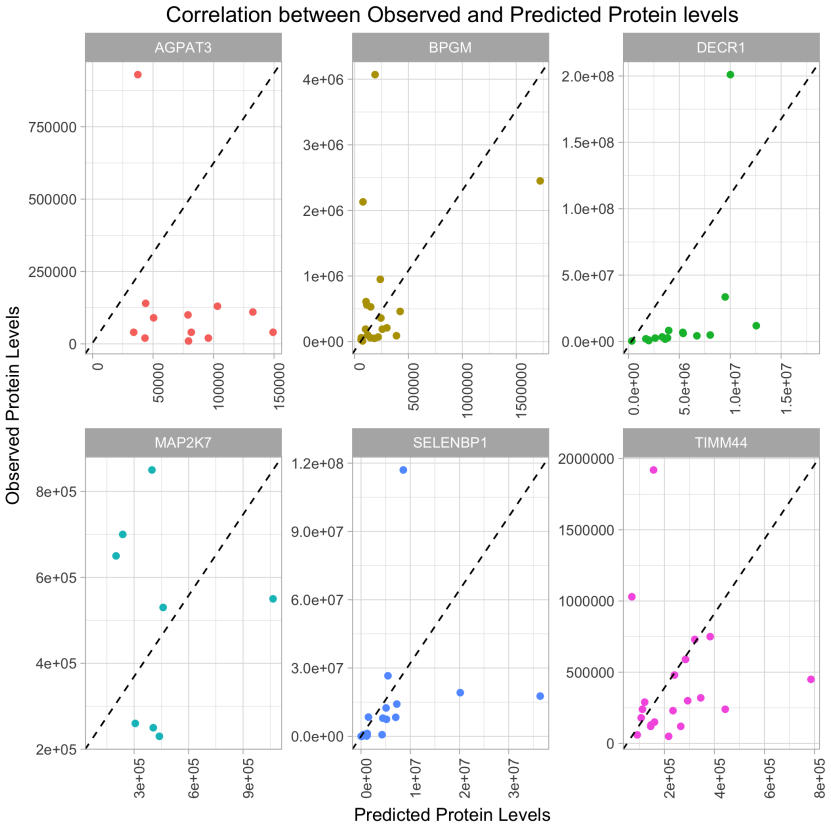
Predicting Protein levels using mRNA levels

In the previous section we found evidence of a linear relationship between mRNA and protein levels. The next step in our analysis is determining if mRNA levels can be solely used to predict protein levels for a given gene. To do this, we compute predictions of protein levels, for every gene-tissue combination in our dataset, using the relation suggested by Wilhelm et al.

$$Protein_{g,t} = Ratio_g \times mRNA_{g,t}$$

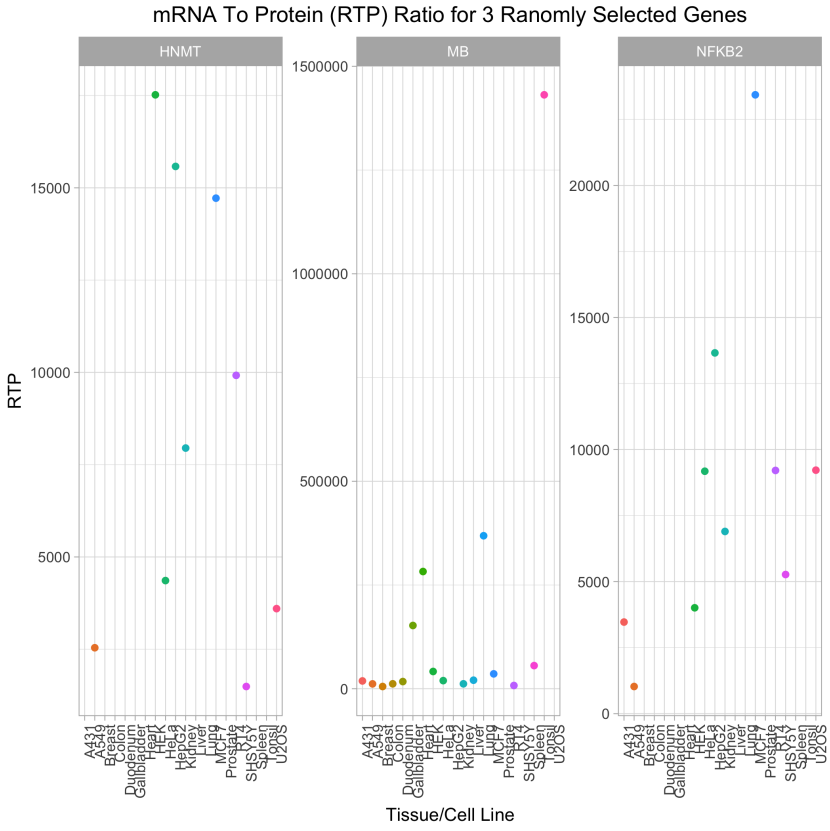
Wilhelm et al. claims this model works well as the ratio of protein to mRNA levels (RTP) is roughly constant across tissues for any given gene (a claim we examine later). Hence to use it to predict protein levels we first calculate the gene-specific conversion factor (Ratio). We compute Ratio as the median over 19 of the 20 tissue samples associated with each gene, excluding the tissue we are predicting for (as done by Wilhelm et al.).

Since this model uses a gene-specific conversion factor, we examine the accuracy of its predictions by plotting predicted vs observed protein levels for 6 randomly selected genes from our dataset.



If the model's predictions are accurate, we would expect the points in these plots to line up along the diagonal (dashed line). This seems to be the case for a few genes. For the rest, the model's predictions differ greatly from observed protein levels. Trying to verify this quantitatively is difficult due to the large number of missing observations in the data provided.

To understand why modified mRNA levels predict protein levels poorly we examine the central claim behind Wilhelm et al.'s model; the ratio of protein to mRNA levels (RTP) is roughly constant across tissues for any given gene. We test this claim by randomly picking 3 genes and plotting their RTP ratio across the different Tissue/Cell lines in our dataset.



We see a wide variation in RTP ratios for each gene, across the 20 tissue/cell line samples associated with it. This seems to suggest that their claim does not hold true and is perhaps the reason behind their model's poor predictive performance.

CONCLUSION

In this report we examine whether mRNA levels can be used to predict protein levels accurately. Our linear mixed effects model finds a significant association between mRNA levels and observed protein levels for the genes under study. We remain skeptical of this result due to the impact of influential and missing observations in our datasets. Next we predict protein levels, for each gene-tissue combination, using Wilhelm et al.'s model. We observe that this model predicts protein levels poorly. While we find the existence of a significant association between mRNA and protein levels plausible, we observe that mRNA levels ***alone*** cannot be used to accurately predict protein levels.

REFERENCES

Gene-specific correlation of RNA and protein levels in human cells and tissues [online] Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5081484/>

A Practical Guide to Mixed Models in R [online] Available at:<https://ase.tufts.edu/gsc/gradresources/guidetomixedmodelsinr/mixed%20model%20guide.html>

Power Analysis for Generalized Mixed Models in R [online] Available at: <http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12504/full>

A very basic tutorial for performing linear mixed effects analyses [online] Available at: <http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf>

Can we predict protein from mRNA levels? [online] Available at: <https://www.nature.com/articles/nature23293>