Cancer! Predictable?

LEC0101, TUT0112 Group 2

Michelle Nguyen, Yushan Chen, Jinquan (Angi) Guo, Sami Ahmed

December 8, 2022

Introduction

The data contains 22 transcription factors and 4 phenotype proteins. These 22 transcription factors can be thought of as the causes which make us get the 4 “outcomes”. We aim to find a relationship between them and answer a few questions. The main dataset we will be working with looks like the following:

Phenotype Indicators	MiTFg	Sox10	NGFR	AXL	
AP-1 transcription factors	ATF2	ATF3	ATF4	ATF5	Phospho_ATF1
	ATF6	JunB	c_Jun	JunD	Phospho_S6
NF_kappaB	Fra1	Fra2	c_Fos	Ki_67	Phospho_Fra1

Objectives

Do experimental conditions change whether we can predict phenotypical outcome from transcription factors?

To answer the above question, we will first try to answer the following questions:

- Do protein levels in experimental condition x change over time t?
- At time t in experimental condition x , what is the relationship between different proteins?
- Can we predict cellular phenotypical outcomes (Y) values/states from transcription factors (TF)?

Statistical Methods

The following data analysis methods were used:

- Two Sample Hypothesis Testing
- Correlation Estimation
- Regression/Classification

Question 1

At 0.5 h time point in experimental condition (Vem drug at 0.1μM dose), what is the relationship between different proteins?

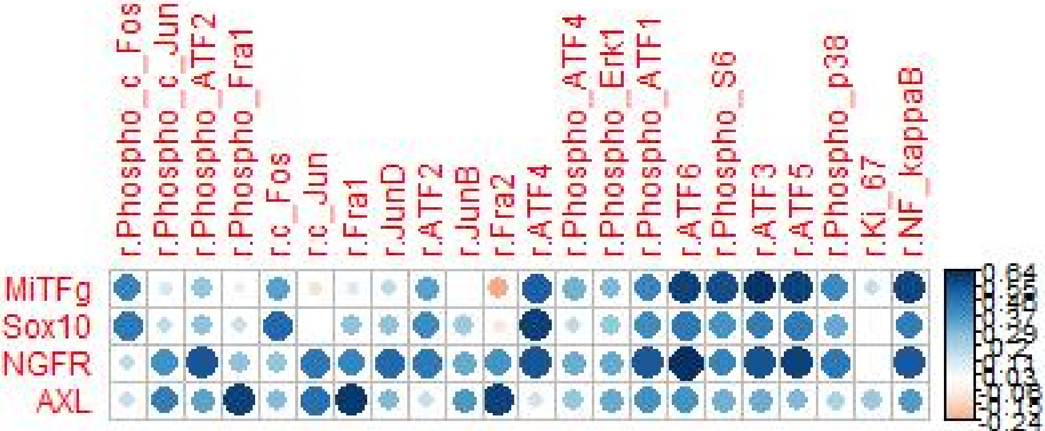
Statistical Method

- An correlation estimation was carried out to determine the relationship between invidiual transcription factors vs. individual outcome proteins
- A correlation matrix was plotted to demonstrate varying degrees of correlation

Data Wrangling

- Data fitting specific condition were extracted into a smaller subset using “subset” function
- Columns not needed for the correlation estimation were also removed

Question 1 - Data visualization



Question 1 - Conclusion

- Correlation coefficient r^2 values range from -0.24 to 0.64.
- Most transcription factors slightly correlate postively with the outcomes.

Transcription factor **ATF3**, *due to its positive correlation* **MiTFg** *in Vem drug at 0.1μM dose, has been chosen to carry out additional statistical test.*

Question 2

Do protein levels in experimental condition (Vem drug at 0.1μM dose) change over time t?

Variables

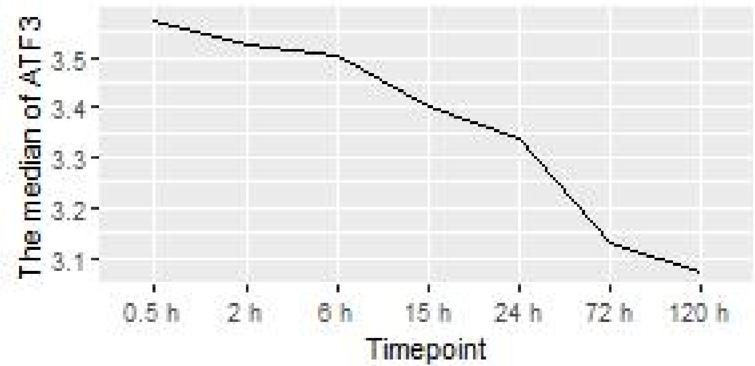
- Protein levels (Ordinal quantitative variable)
- Time (Ordinal categorical variable)

Data Wrangling

- Only columns needed for the analysis was selected using the "select" function.
- Data fitting the experimental condition were extracted using "filter" function.
- "factor" function was nested inside "mutate" function to rearrange timepoint in the right order.

Question 2 - Data visualization

The median of ATF3's protein level at different time point with Vem drug at 0.1μM dose



The median of ATF3 under the condition of Vem drug at 0.1μM dose decreases from around 3.57 to approximately 3.06 at 0.5h and 120h timepoint respectively.

Question 2 - Statistical Method

AP-1 transcription factor: ATF3

Variables:

- Experimental condition x: Vem drug at 0.1 μM doses
- Time t: 0.5h timepoint and 120h timepoint

Hypothesis testing

- NULL Hypothesis: The median of ATF3 under the condition of Vem drug at 0.1 μM dose in 0.5h is equal to the median of ATF3 under the condition of Vem drug at 0.1 μM dose.

$$H_0\colon\ \mu_{0.5h} = \mu_{120h}$$

- Alternative Hypothesis: The median of ATF3 under the condition of Vem drug at 0.1 μM dose 0.5h is not equal to the median of ATF3 under the condition of Vem drug at 0.1 μM dose in 120h.

$$H_1\colon\ \mu_{0.5h} \neq \mu_{120h}$$

- Set α-significance level at 0.05.

Question 2: Result

test statistic = -0.2169056
p-value = 0

p-value: The probability of observing a test statistic that is equal to or larger than the one we got when the null hypothesis is True.

- p-value is 0 which is smaller than the α-significance level. There is strong evidence to reject the Null hypothesis.
- We are, however, at risk of type I “false positive” error where we wrongly reject the Null hypothesis, but it actually is true.

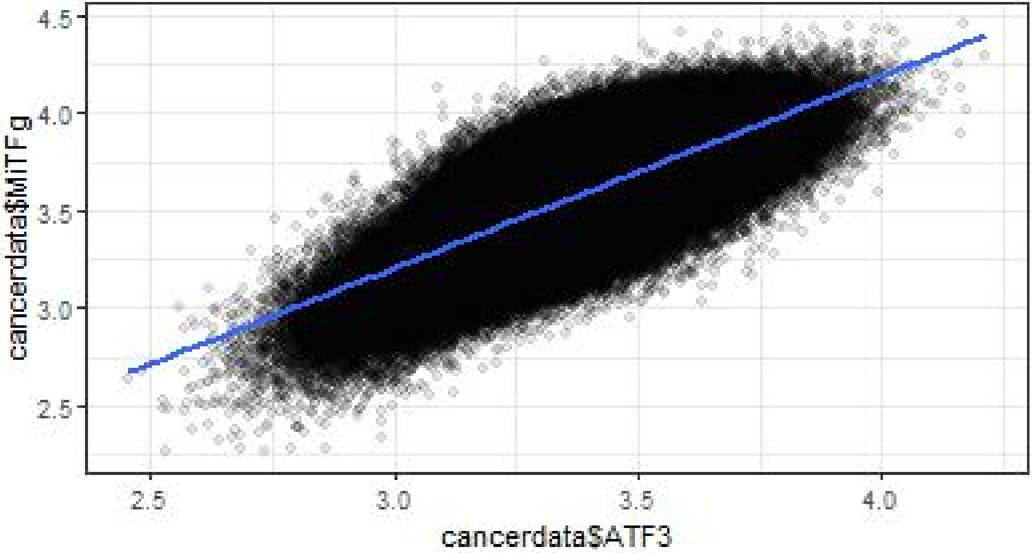
Question 3

Can we predict cellular phenotypical outcomes (Y) values/states from transcription factors (TF)?

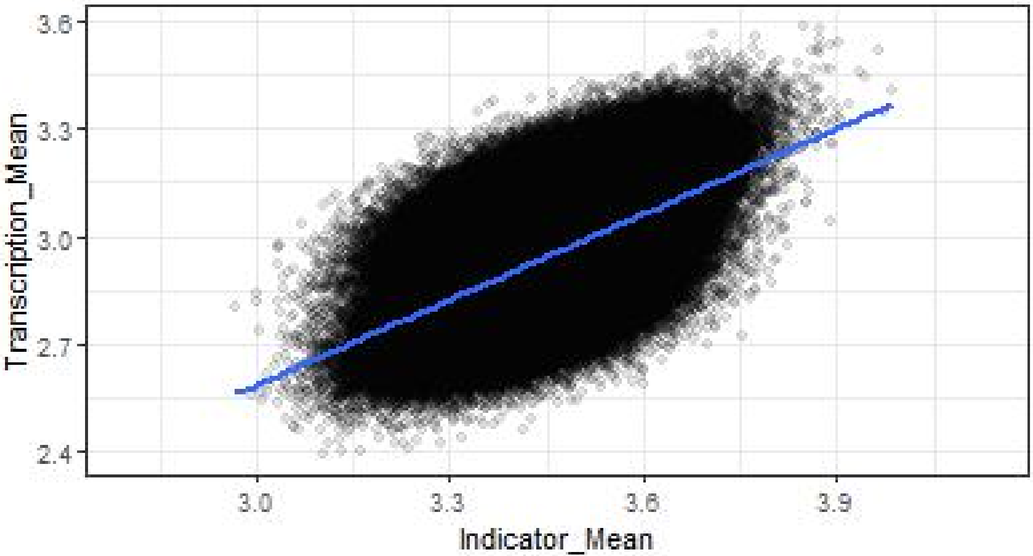
Two graphs were plotted in an attempt to answer this question:

- The first: A linear regression between phenotype indicator: MiTFg and transcription factor: ATF3
- The second: A linear regression between mean phenotype indicators and mean transcription factors
- The latter was done in order to come to a generalised conclusion.

Question 3 - Data visualization 1



Question 3 - Data visulization 2



Question 3 - Regression Equations

$$\texttt{MiTFg}_i = \beta_0 + \beta_1 \texttt{ATF3}_i + \epsilon_i$$

$$\texttt{MiTFg}_i = 0.261425 + \beta_1 0.982256_i$$

$$\texttt{Transcription}_i = \beta_0 + \beta_1 \texttt{Indicator}_i + \epsilon_i$$

$$\texttt{Transcription}_i = 0.195810 + \beta_1 0.796468_i$$

Results

From our analysis we can conclude the following:

- Certain transcription factors have higher correlations with certain phenotype indicators than others. (Question 1)
- As a result, ATF3 and MiTFg were chosen for further analysis due to their high correlation (Question 1)
- The protein levels for transcription factor ATF3 decreases from 0.5h to 120h (Question 2)
- There appears to be a strong to moderate positive correlation between the transcription factor ATF3 and phenotype indicator MiTFg (Question 3)
- There appears to be a moderate positve correlation between mean transcription factors and mean phenotype indicators (Question 3)

Conclusion

The above analyses has led us to conclude the following

- Since there is a correlation between transcription factors and phenotype indicators, phenotypical outcomes can be predicted in certain cases.
- Since protein levels appear to decrease, the base scenario with one drug and one dose was chosen.
- Since ATF3 and MiTFg are strongly related, this makes us conclude that other transcription factors and the remaining three phenotypical indicators will have similar relationships.
- Using the above, the levels of phenotype indicators as they correspond with transcription factors can be used to find the phenotypical outcome, *Undifferentiated, Neural crest-like, Transitory* or *Melanocytic* for instance.

Limitations

- The Cellular phenotypes as they correspond to the Indicator level weren't analyzed.
- The transcription factors with lower correlation weren't used in this analysis.
- All 22 transcription factors weren't used to come to a conclusion.
- (Changes in) relationship/correlation between proteins in different experimental conditions have not been analyzed.

Acknowledgements

Our group would like to thank Professor Scott Schwartz and Dr Heman Shakeri for providing us with the data set. We would also like to thank our TA Anita Pilehrood for her help, support and suggestions.