Examining Shifted Cellular Homeostasis Through Data-driven Identification of Transcription Factor Networks Over Time

STA130 - Final Project

Umi Yamaguchi, Ralisha Woodhouse, Yifei Hong

University of Toronto

December 8, 2022

Introduction

Goals & Motivation

The goal of a cancer treatment is to decrease or reduce the growth and/or expansion of malignant cells. Overall, our goal is to:

- Control the transcription factor network using the data set given [1]
- Identify changes from deleterious to healthy phenotypes (cellular states) over time

Research Questions

1. Can we predict cellular phenotype outcomes (Y) values from transcription factors (TF)? $\rightarrow$ Linear Regression

Justification: This question is relevant as it will determine relationships between proteins and phenotype indicators which will allow us to predict the cellular phenotype. In terms of cancer treatment this will grant us information on what proteins to target to transform deleterious to good homeostasis.

2. At time t in experimental condition, what TF are most predictive of cellular values/states (Y)? $\rightarrow$ Classification Trees

Justification: This question is relevant to our goal because it aims to investigate which transcription factors are highly relevant in detecting cancerous cellular states.

3. Do protein levels in experimental condition X change over time t? $\rightarrow$ Two Sample Hypothesis Testing

Justification: This question is worthy of investigation because it will indicate whether the drugs in the data set are working or not.

Data Set & Data Wrangling

Data Set

- Overall 22 Levels of AP-1 Transcription Factors [TF]
- 4 Phenotype Indicators (MiTFg, Sox10, NGFR, AXL)
- Others: Timepoint, Drug type (Vem, Vem+Tram), Dose id, Dosage, Repetition

Data Wrangling

Table 1: Data Wrangling

AXL	Sox10	Timepoint	Drugs	dose_id	Doses	Rep
3.536432	3.686878	0.5	0	1	0	1
3.732794	3.668114	0.5	0	1	0	1
3.609001	3.781692	0.5	0	1	0	1
3.223876	3.700308	0.5	0	1	0	1
3.600571	3.755307	0.5	0	1	0	1

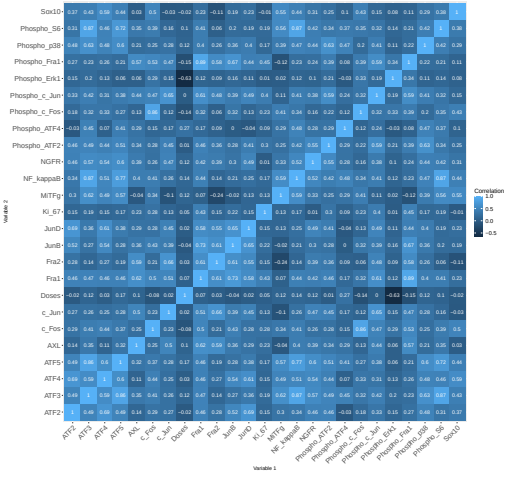
- Total Observations: 540792 (exclude NAs & exclude any TFs with incomplete data)
- Timepoint & Doses (exclude any characters or units)
- Change the name of Drugs to 0 and 1. (0 for 'Vem' and 1 for 'Vem + Tram')

Initial Data Analysis

- Empirically determining HIGH/LOW levels of the specific proteins that determine the phenotype of a melanoma cell:
  - HIGH: protein level is greater than or equal to the median of the subset of data
  - LOW: protein level is less than the median of the subset of data

Cellular Phenotype	MiTFg	NGFR	SOX10	AXL
Undifferentiated	LOW	LOW	LOW	HIGH
Neural crest-like	LOW	HIGH	HIGH	HIGH
Transitory	HIGH	HIGH	HIGH	LOW
Melanocytic	HIGH	LOW	LOW	LOW

Correlation Matrix for Drug 'Vem' at Timepoint 0.5 h



Linear Regression

Research Question 1:

Can we predict cellular phenotype outcomes (Y) values from transcription factors (TF) at experimental condition X?

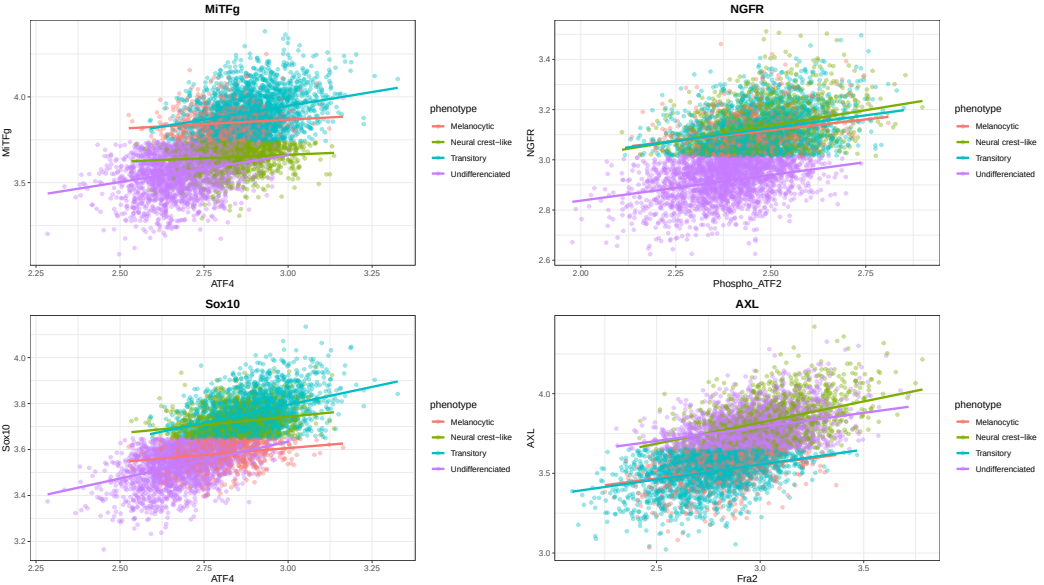
Method:

1. Subset the data frame to the experimental condition X which is indicated by time point 0.5 hours and Drug Vem
2. Categorize the phenotypic outcomes using empirically determined HIGH/LOW distinctions of the 4 specific proteins
3. Extract the transcription factor(s) that are highly correlated with the 4 specific proteins (using the correlation matrix)
4. Perform the linear regression method to identify the most significant transcription factor
5. Plot the transcription factor against the protein, observing the effects on phenotypic outcome
6. Conduct linear regression again to construct equations for the fitted lines of the phenotype indicators

Determining best predictor ex. Sox10:

Term	Estimate	p-value	Correlation
(Intercept)	0.66037444	2.657616e-104	1
Fra1	0.54898424	2.247034e-269	0.62
Fra2	0.27987459	0.000000e+00	0.49
Phospho_Fra1	0.02822788	4.107424e-03	0.57

Linear Regression: Data Visualisation



Linear Regression: Results & Interpretation

Theoretical equation:

$$Y_i = \beta_0 + \beta_1 x_{TF} + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \epsilon_i$$

Hypothesis test for multivariate linear regression:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_1 : \text{At least one of} \beta_1, \beta_2, \beta_3, \beta_4 \neq 0$$

Fitted Regression Lines:

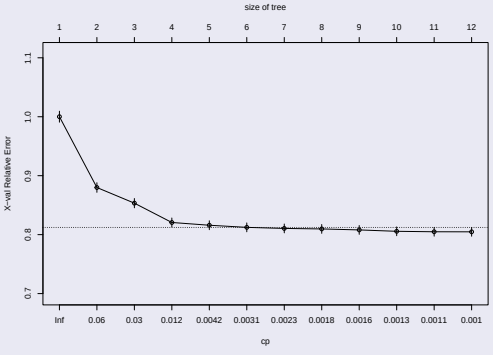
1. $$\widehat{MiTFg} = 3.14 + 0.25x_{ATF4} - 0.21x_N + 0.04x_T - 0.25x_U$$
2. $$\widehat{NGFR} = 2.6 + 0.21x_{Phospho_ATF2} + 0.016x_N - 0.18x_U$$
3. $$\widehat{Sox10} = 2.84 + 0.26x_{ATF4} + 0.13x_N + 0.16x_T - 0.017x_U$$
4. $$\widehat{AXL} = 2.96 + 0.2x_{Fra2} + 0.26x_N + 0.23x_U$$

Research Question 2:

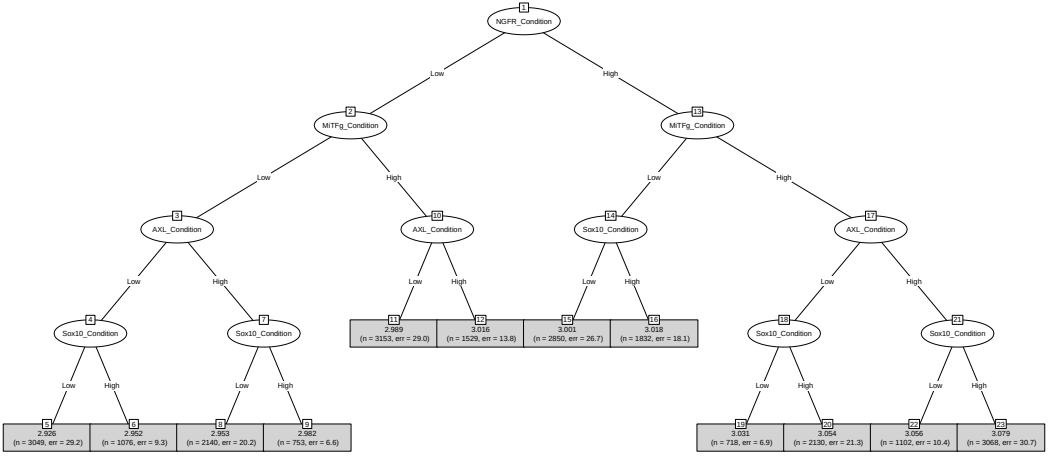
At time t in experimental condition, what TF are most predictive of cellular values/states (Y)?

Method:

1. Separate the data frame for time point 0.5 and 15 hours
2. Separate the time point into drugs 0 and 1
3. Each drug 0 and 1 creates the new columns for 4 phenotype indicators conditions (High/Low)
4. Check all phenotypes of 0.5h drugs 0 and 1 to identify any condition matches



Classification Trees: Visualisation ex. Phospho_p38



- Time point = 0.5h Drug = 0

AP-1	Error	Condition	#Observations
Phospho_ATF2	26.3	Undifferentiated	2140
Phospho_p38	6.9	Melanocytic	718
NF_kappaB	5.7	Melanocytic	718

- Time point = 0.5h Drug = 1

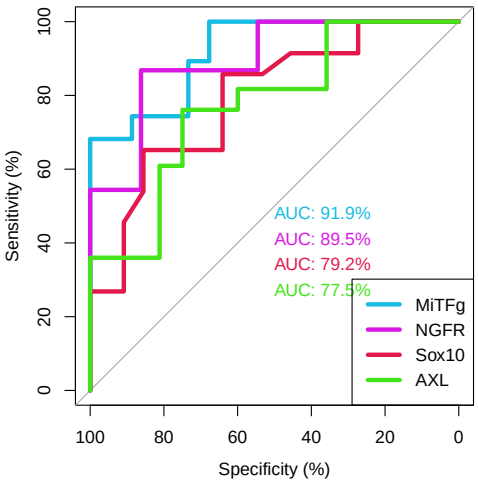
AP-1	Error	Condition	#Observations
Phospho_ATF2	17.2	Neural crest-like	1340
Phospho_p38	19.5	Transitory	2065
Phospho_Fra1	24.7	Melanocytic	793
NF_kappaB	7.4	Melanocytic	793

What the results tell us:

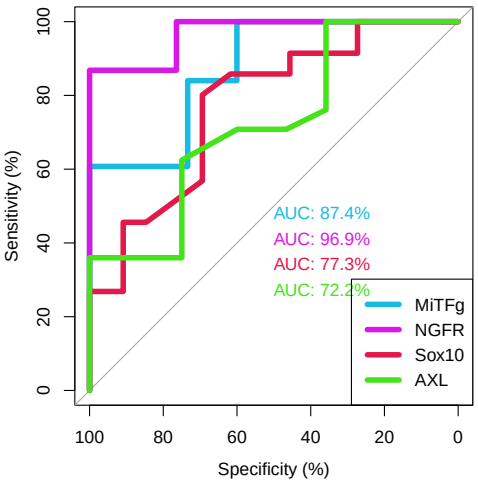
- Initial Condition for each Drug
- High accuracy becomes Melanocytic States
- Lot of # Tranistory States found on Drug Vem + Tram

Prediction (Classification Trees) : ROC curves (Sensitivity vs. Specificity)

NF_kappaB



Phospho_p38



Research Question:

Do protein levels in experimental condition X change over time t?

Experimental condition:

- Time point = 0.5h Drug = 0

AP-1	Condition
Phospho_p38	Melanocytic
NF_kappaB	Melanocytic

- Time point = 0.5h Drug = 1

AP-1	Condition
Phospho_p38	Transitory
Phospho_Fra1	Melanocytic
NF_kappaB	Melanocytic

Two Sample Hypothesis Testing: Method

Method:

1. Categorize the 4 genes as high or low.
2. Identify the Cellular Phenotype using given information of genes.
3. Calculate the test static: the mean difference between two time periods.
4. Perform the two sample hypothesis test.

Hypothesis:

$$H_0 : M_{0.5} = M_{15}$$

$$H_A : M_{0.5} \neq M_{15}$$

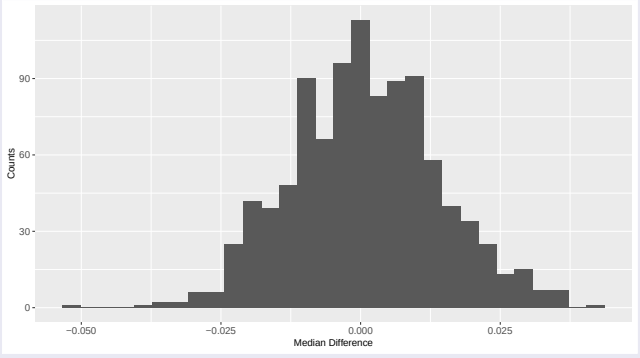
`Significance value alpha=0.01`

Results

p-value	Reject/Not
$P_1 = 0$	Reject
$P_2 = 0$	Reject
$P_3 = 0$	Reject
$P_4 = 0$	Reject
$P_5 = 0$	Reject

- Assume Null Hypothesis is true
- No single value that is as or more extreme than the test statistic
- Reject all the null hypothesis

Drug0: Difference between 0.5 and 15h of NF_kappaB



Overall Results

Linear Regression

- Detected the Good predictor for each Phenotype Indicator

AP-1	Good Predictor for
ATF4	MiTFg,Sox10
Phospho_ATF2	NGFR
Fra2	AXL

Decision Trees

- Identified some phenotype at 0.5 hour time point
- Initial Melanocytic condition found in 2 main TF
- Detected as high sensitivity and specificity

Two Sample Hypothesis Test

- Detected that there's difference between 0.5 and 15 hours
- Rejected the Null Hypothesis testing: $\alpha = 0.01$
- It may have some effect from drugs

Limitations

Linear Regression

- Confounding variables $\rightarrow$ multicollinearity of transcription factors, some extent of bias when choosing the predictor.

Classification Trees

- Overfitting that may lead to do wrong prediction, bias from a subjective observer.

Two hypothesis Testing

- Type I and Type II error(unlikely), not providing enough information for causal relationship.

Conclusions & "The Bigger Picture"

Recall...

Our goal was to control the transcription factor network to identify changes in homeostasis over time.

What did we achieve?

Overall, our exploration of such a dynamical system allowed us to detect deviations away from healthy cellular function; through linear regression, classification trees and two sample hypothesis testing, we were able to predict phenotypic outcome based on certain transcription factors and experimental conditions.

What does this mean?

From identifying phenotypic outcome and their predictors, we know when and how to intervene the progression of cancer before it manages to establish a deleterious cellular homeostasis. Thus, our results provide inside cancer treatments

Meta: What is "good" cellular homeostasis, and how can "bad" cellular homeostasis be changed to be "good"?

In terms of reversing deleterious homeostasis, we understand the phenotypic outcomes result from a combination of transcription factors. Therefore, in the future, an investigation into this interdependence could provide insight into possibly changing to a'good' homeostasis condition.

Reference

[1] AP-1 transcription factor network explains diverse patterns of cellular plasticity in melanoma Natacha Comandante-Lou, Douglas G. Baumann, Mohammad Fallahi-Sichani bioRxiv 2021.12.06.471514; doi: https://doi.org/10.1101/2021.12.06.471514

[2] Tables for the Classification Trees: https://docs.google.com/spreadsheets/d/13_1a_-0V5JHSz73iriPjz50Bo9Y6YCpHou2AdM0XbMA/edit#gid=0