

Data-driven identification of transcription factor networks over time in indicating a shifted cellular homeostasis

STA130 - Final Project

Umi Yamaguchi, Ralisha Woodhouse, Yifei Hong

University of Toronto

December 8, 2022

Goals And Motivation

- Decrease or reduce the growth and expansion of malignant cells
- Controlling the transcription factor network using the data set given [1]
- Identify movement from deleterious to healthy phenotype overtime

Research Questions & Categories of Analysis

- 1] Can we predict cellular phenotype outcomes 'Y' values from transcription factors (TF)? → Linear Regression
- 2] At time 't' in experimental condition, what TF are most predictive of cellular values/states? → Classification Trees
- 3] Do protein levels in experimental condition 'X' change over time 't'? → Two Sample Hypothesis Testing

Data Set

- Overall 22 Levels of Transcription Factors (AP-1)
- 4 Phenotype Indicators (MiTFg, Sox10, NGFR, AXL)
- Others: Time point, Drugs type, Dose id, Dosage, Repetition

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

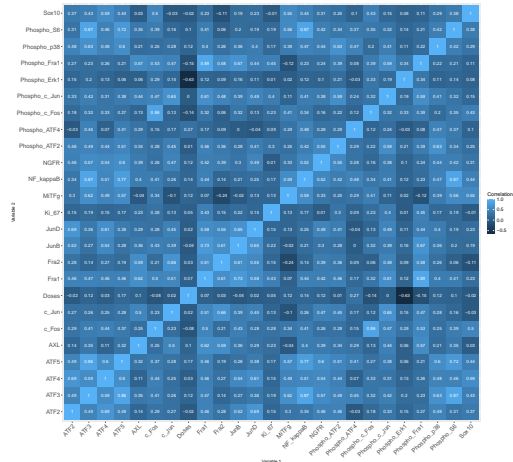
- Total Observations: 540792 (exclude NA's)
- Time point (exclude any characters)
- Doses [exclude any unit ('uM' and '+' sign)]
- Change the name of drugs to 0 and 1. (0 for 'Vem' and 1 for 'Vem + Tram')

Initial Data Analysis

- Empirically determining HIGH/LOW levels of the specific proteins that determine the phenotype of a melanoma cell:

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 (Drug 'Vem' at Timepoint 0.5 h)



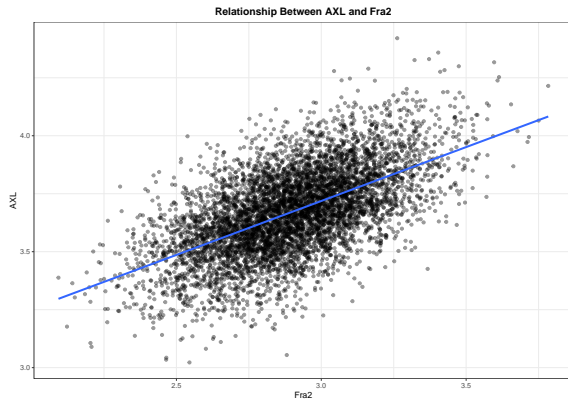
Research Question:

Can we predict cellular phenotype outcomes 'Y' values from transcription factors (TF)?

Methods:

- 1 Subset the data frame to include 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 the mostly correlated with each of the 4 specific proteins (using the correlation matrix)
- 4 Use 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 Train-test to assess accuracy of model

Identifying Predictors (Linear Regressions)

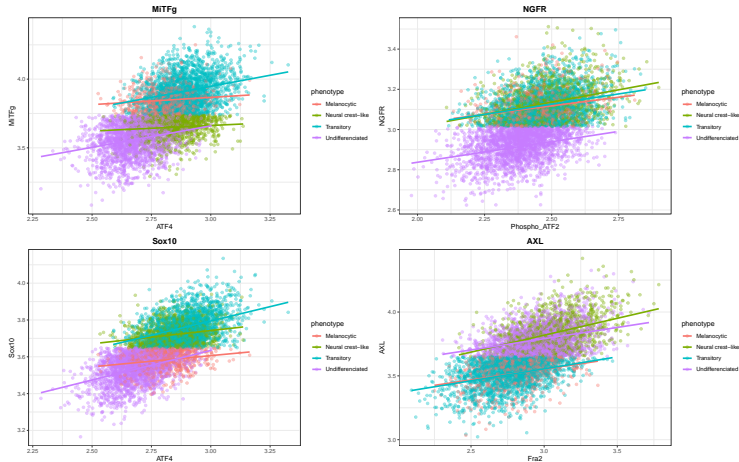


Equation:

- Initial Condition for each Drug

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

Results and Prediction (Linear Regressions)



What the results tell us:

- MiTFg:
- NGFR
- Sox10
- AXL

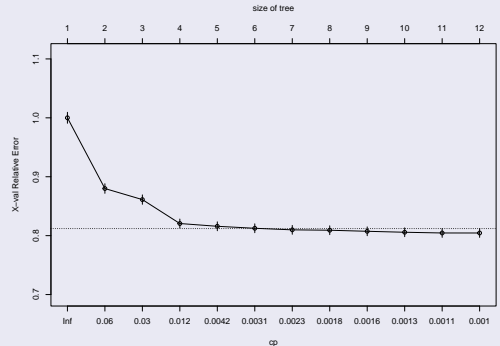
Classification Trees

Research Question:

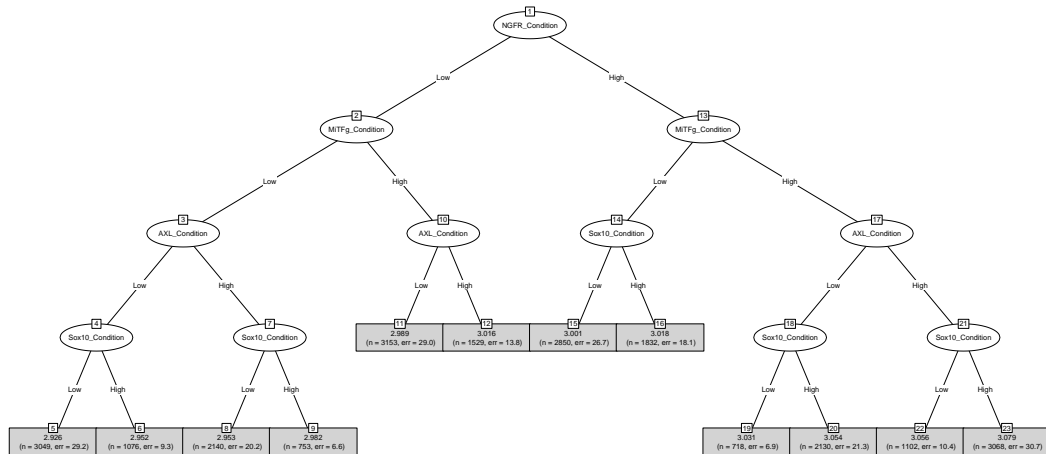
- At time 't' in experimental condition, what TF are most predictive of cellular values/states?

Methods:

- 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 example : Phospho_p38



Results (Classification Trees)

- 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

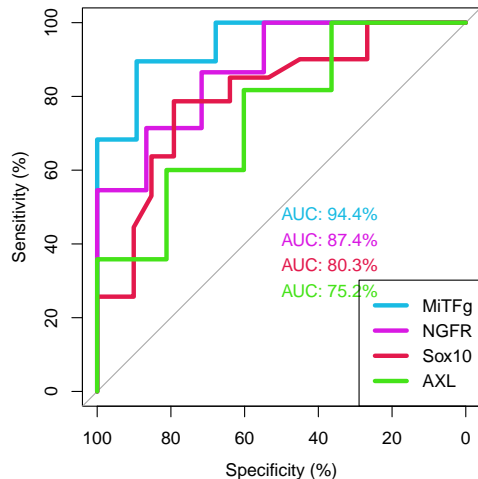
AP-1	Error	Condition	#Observations
Phospho_ATF2	17.2	Neural crest-like	1340
Phospho_p38	19.5	Transitory	2065
Phospho_Fra1	24.7	Melanocytic	793
NF_kappaB	7.4	Melanocytic	793

What the results tell us:

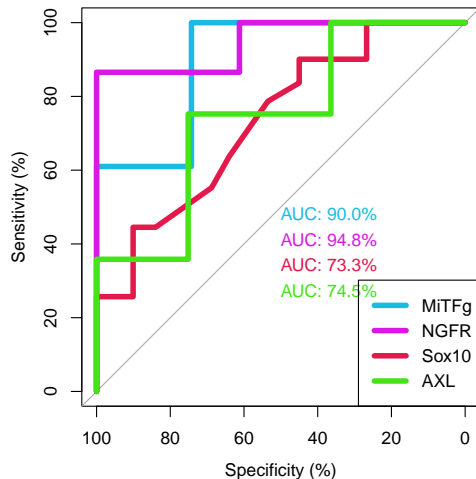
- Initial Condition for each Drug
- High accuracy becomes Melanocytic States
- Lot of # Transitory States found on Drug Vem + Tram

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

NF_kappaB's ROC curves



Phospho_p38's ROC curves



Two Hypothesis Teting

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_ATF2	Undifferentiated
Phospho_p38	Melanocytic
NF_kappaB	Melanocytic

- Time point = 0.5h Drug = 1

AP-1	Condition
Phospho_p38	Transitory
Phospho_Fra1	Melanocytic
NF_kappaB	Melanocytic

Process (Two Hypothesis Testing)

Methods:

- Categorize the 4 genes as high or low.
- Identify the Cellular Phenotype using given information of genes.
- Calculate the test static: the mean difference between two time periods.
- 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 and Prediction (Two Hypothesis Testing)

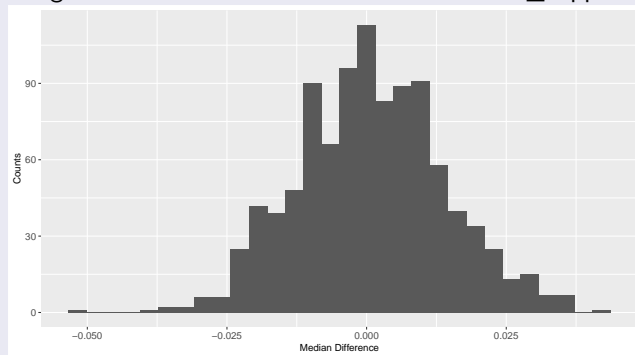
Results

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

Prediction:

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 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

Limitation

Example of limitation

Linear Regression

- Confounding variables - multicollinearity and bias

Classification Trees

- Bias and over fitting

Two hypothesis Testing

- Type I and type II error

Conclusion with Future Perspective

Conclusions

-
-
-

How this results will help?

- 1
- 2
- 3

[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