An analysis of the relationship between AP-1 Transcription Factors and the Phenotype Proteins of a Melanoma Cell

Prediction of the phenotype indicator NGFR by the level of its most predictive AP-1 transcription factor

Group 4: Aaminah Ghumman, Michelle Wang, Charlene Kamba, Junhee Lee

December 8, 2022

Introduction

The dataset we are studying is from Dr. Heman Shaker's research into stopping the cancerous progression of cells. It contains measurements of the protein levels of phenotype proteins and AP-1 transcription factors (TFs) in individual cells across different experimental conditions (drugs and drug doses) and timepoints.

The cellular phenotype of melanoma cells can be classified into four different stages, undifferentiated, neural crest-like, transitory, and melanocytic, where undifferentiated is the only stable state. These phenotype stages can be determined by the protein level combination of the four phenotype proteins, MiTFg, Sox10, NGFR and AXL. These four proteins are seen as 'outcomes', while the TFs are thought of as their 'causes'. Therefore, if we are able to find a predictive relationship between the TFs and phenotype proteins, through future research into drugs influencing these factors, we can eventually affect the cellular phenotype and fight cancer.

Objectives

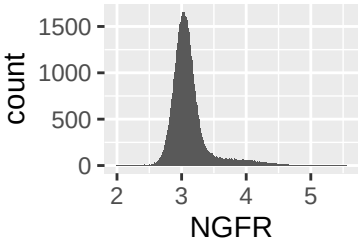
To address this big picture, we decided to study the interdependence between one phenotype protein, NGFR, and a predictive transcription factor.

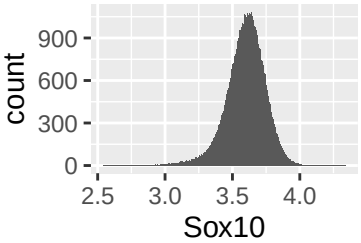
Research Questions:

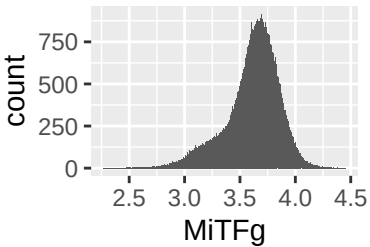
Under an experimental condition of 3.16 uM Vem:

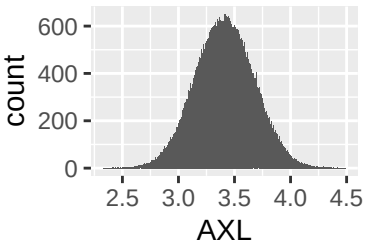
1. At 6 hours, which TF is most predictive of the protein levels of NGFR? - Classification
2. At 6 hours, what is the correlation between NGFR and its most predictive TF from question 1? Is this correlation consistent across timepoints? - Correlation
3. Through multivariate linear regression, can we predict the protein level and state of NGFR using this TF? - Multivariate Regression

1. **Determine the HIGH/LOW distinctions for phenotype proteins.**









- In analyzing the distributions of the four phenotype indicator proteins (MiTFg, AXL, Sox10 and NGFR), we observed the following:
  - The distributions are skewed.
  - The median would be more 'robust' and less influenced by outliers that skew a distribution.
- Therefore, we use the medians of the protein levels to create cutoffs for the HIGH and LOW distinctions for the phenotype proteins.
- We set HIGH as greater than or equal to the median, and LOW as below the median.

	NGFR	MiTFg	AXL	Sox10
mean	3.110165	3.617631	3.412165	3.597957
median	3.055596	3.650493	3.41105	3.607437

2. Classify the cells by phenotypes based on the HIGH/LOW balance of the four phenotype proteins.

- Creating a new column, 'cell_phenotype', that classifies the cells under the four cellular phenotypes based on the HIGH/LOW balance of the four phenotype indicators, as shown below.

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

- Omitting unclassified cells because we are focusing on predicting whether a cell is undifferentiated, so all observations need to have a distinct cellular phenotype.
- As a result of cleaning the data, a cell with low NGFR is immediately classified as undifferentiated, so we will only focus on NGFR rather than other phenotype proteins in our analysis.
- Therefore, if we can predict the protein level for NGFR, then we can also predict the healthiness of the cell.

3. Fix the experimental condition.

- Filter the observations with drug Vem and dosage 3.16 uM.

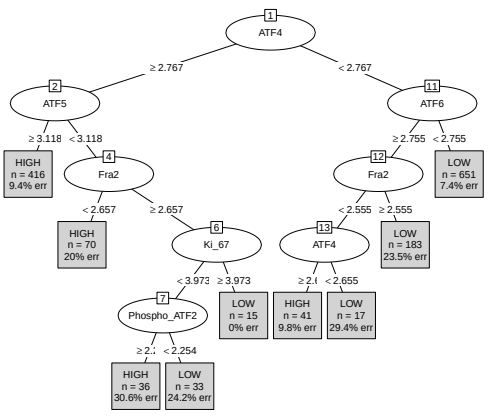
Research Question 1 - Statistical Method

Under an experimental condition of 3.16 uM Vem, at 6 hours, which TF is most predictive of the protein levels of NGFR?

Method: Classification Decision Tree

- Fixing the time point to 6 hours, we created a random 80-20 Train-Test Split for the data. 80% of the data will be used to fit the model and the remaining 20% will be used to 'score' the model.
- Using the 22 AP-1 Transcription Factors as our covariates, we created a binary decision tree to classify values as either HIGH or LOW, as categorized by the NGFR protein levels.
- Placing the decision tree through the variable importance function to find the TFs that are most important in predicting the HIGH/LOW protein levels of NGFR.
- A confusion matrix was then scored on the Test sample to test the performance of the model.

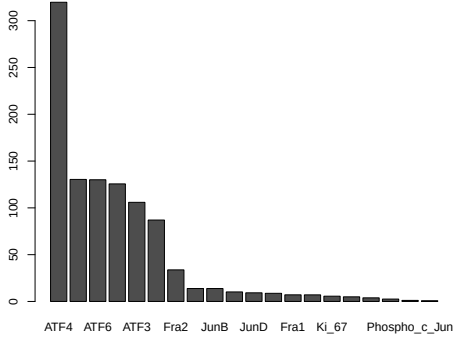
Decision Tree



- We created a binary decision tree that classified the observations into HIGH or LOW NGFR levels by using the transcription factors. Its root nodes include the AP-1 TFs that form the best split, ie. the TFs that produce the lowest gini/impurity nodes are selected. There are 5 levels in the decision tree with 8 root nodes.

Research Question 1 - Results

Variable Importance



- We found that the most important AP-1 TF is ATF4, with an importance of over 300 as shown in the figure below.
- However, a limitation of this model, as we are only using 80% of the data in the training dataset to model the decision tree, is that changing the seed will change the outcome of the decision tree, along with the TF with the highest importance.
- To overcome this problem of randomness, we did a confusion matrix to score our model on the remaining 20% of data, which is our test data.

Research Question 1 - Results

Confusion Matrix

Measures	Results
accuracy	0.819672131147541
precision	0.81896551724137
sensitivity	0.88785046728972
specificity	0.723684210526316

- Specificity is most important in evaluating our model as a false negative can result in potentially cancerous cells passing undetected.
- The model scores well for all the measures on the test data, hence it is representative of the whole data and is very likely to produce good classification and prediction.
- Therefore, our decision to further analyse the relationship between ATF4 and NGFR has a good basis since the conclusion that ATF4 is NGFR's most predictive TF holds credibility.

Research Question 2 - Statistical Method

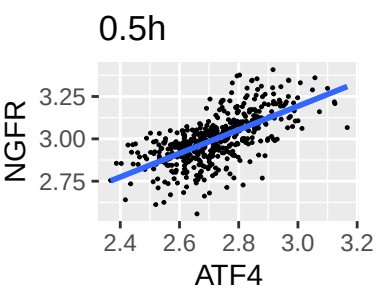
Under the experimental condition of 3.16 uM Vem, at 6 hours, what is the correlation between NGFR and its most predictive TF (ATF4) from question 1? Is this correlation consistent across all time points?

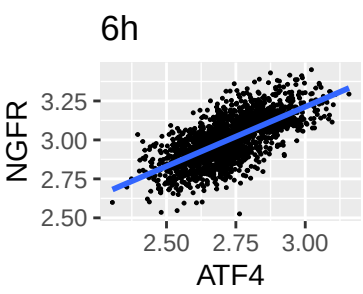
Method: Correlation Estimation

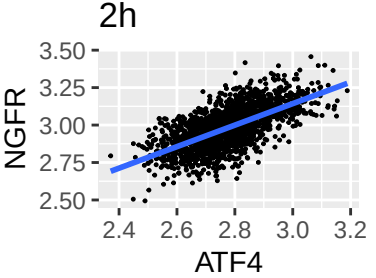
- As shown in the decision tree analysis, ATF4 is the most predictive AP-1 TF for NGFR at our fixed experimental condition.
- To further analyze the relationship between ATF4 and NGFR under this fixed condition, we did correlation estimation at the time point of 6 hours.
- To see whether this relationship is consistent over time, we repeated this correlation estimation for all time points.

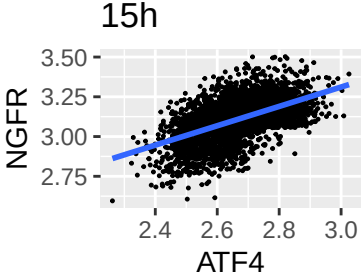
Research Question 2 - Results

Timepoint (hours)	n	Correlation
0.5	466	0.6471733
2	1748	0.6515669
6	1828	0.6919023
15	4277	0.5303693
24	2044	0.2670015
72	114	-0.2569150
120	123	-0.2295300

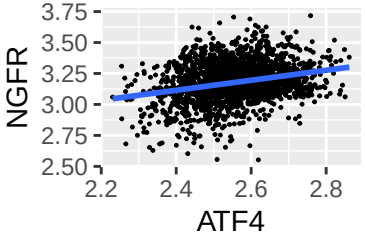

0.5h
NGFR
3.25
3.00
2.75
2.4
2.6
2.8
3.0
3.2
ATF4


6h
NGFR
3.25
3.00
2.75
2.50
2.50
2.75
3.00
ATF4

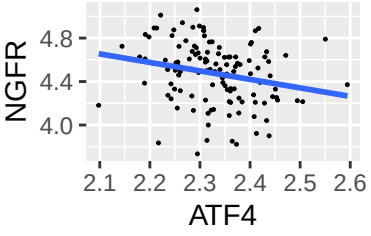

2h
NGFR
3.50
3.25
3.00
2.75
2.50
2.4
2.6
2.8
3.0
3.2
ATF4


15h
NGFR
3.50
3.25
3.00
2.75
2.4
2.6
2.8
3.0
ATF4

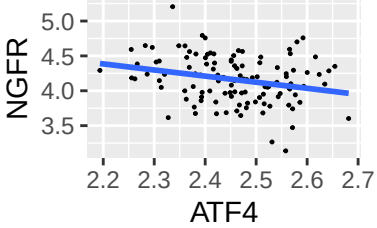
24h


NGFR
3.75
3.50
3.25
3.00
2.75
2.50
2.2
2.4
2.6
2.8
ATF4

120h


NGFR
4.8
4.4
4.0
2.1
2.2
2.3
2.4
2.5
2.6
ATF4

72h


NGFR
5.0
4.5
4.0
3.5
2.2
2.3
2.4
2.5
2.6
2.7
ATF4

Research Question 2 - Results

- At 6 hours, a correlation coefficient of 0.69 was calculated. This moderately strong positive correlation between the protein levels of ATF4 and NGFR is corroborated by the decision tree.
- A similarly strong correlation can be found at three other timepoints, 0.5 hours, 2 hours and 15 hours.
  - Thus, it is highly likely that ATF4 is one of the most predictive transcription factors for these time points as well, and can be used to predict the level of NGFR for these timepoints.
- In contrast to this, however, the correlation coefficients of timepoints beyond 15 hours (24 hours, 72 hours and 120 hours) demonstrated weak positive relationships between the two proteins, which gradually changed to a weak negative relationship.
  - Therefore, ATF4 is not as predictive towards the level of NGFR in these timepoints, and it may be more accurate to look at other TFs here.

Research Question 3 - Statistical Method

Through multivariate linear regression, can we predict the protein level and state of NGFR using this TF (ATF4) under the experimental condition of 3.16 uM Vem?

Multivariate regression

```
proj_data_03 <- proj_data_01 %>% select(NGFR, ATF4, Timepoint)
least_squares_fit <- lm(NGFR ~ ATF4*Timepoint, data=proj_data_03)
```

Through multivariate linear regression, we get the equation of the model, which allows us to predict the protein levels of NGFR using ATF4 levels.

Research Question 3 - Statistical Method

Regression model

- y_i: NGFR level of a cell for i^{th} experiment
- x_{1i}: ATF4 level of a cell for i^{th} experiment
- x_{2i}: time point of i^{th} experiment
- β_0: intercept parameter
- β_i: slope parameter

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} \mathrm{I}(x_{2i} = 2\mathrm{h}) + \beta_3 x_{1i} \mathrm{I}(x_{2i} = 6\mathrm{h}) + \beta_3 x_{1i} \mathrm{I}(x_{2i} = 15\mathrm{h}) + \beta_4 x_{1i} \mathrm{I}(x_{2i} = 24\mathrm{h}) + \beta_5 x_{1i} \mathrm{I}(x_{2i} = 72\mathrm{h}) + \beta_6 x_{1i} \mathrm{I}(x_{2i} = 150\mathrm{h})$$

Research Question 3 - Results

```
##          (Intercept)                 ATF4      Timepoint120 h       Timepoint15 h
##           1.10295332           0.69662251           5.19175817           0.38179654
##         Timepoint2 h       Timepoint24 h        Timepoint6 h       Timepoint72 h
##          -0.11363482           1.04264643          -0.18176496           5.20853912
## ATF4:Timepoint120 h ATF4:Timepoint15 h  ATF4:Timepoint2 h ATF4:Timepoint24 h
##          -1.47787669          -0.08776639           0.02128113          -0.29286894
##   ATF4:Timepoint6 h ATF4:Timepoint72 h
##           0.06711219          -1.57212144
```

As observed in the results of the previous research question, the correlation between ATF4 and NGFR is not consistent over time, more importantly, it begins to weaken past the time point 24 hours. By looking at the beta values, this regression analysis also confirms the interaction between ATF4 and time points against NGFR.

Prediction model

- $\hat{y}_i$: predicted NGFR level of a cell for i^{th} experiment
- x_{1i}: ATF4 level of a cell for i^{th} experiment
- x_{2i}: time point of i^{th} experiment

$$\hat{y}_i \approx 1.103 + 0.697x_{1i} + 0.021x_{1i}\mathrm{I}(x_{2i} = 2\mathrm{h}) + 0.067x_{1i}\mathrm{I}(x_{2i} = 6\mathrm{h}) - 0.088x_{1i}\mathrm{I}(x_{2i} = 15\mathrm{h}) - 0.293x_{1i}\mathrm{I}(x_{2i} = 24\mathrm{h}) - 1.572x_{1i}\mathrm{I}(x_{2i} = 72\mathrm{h}) - 1.478x_{1i}\mathrm{I}(x_{2i} = 120\mathrm{h})$$

The results of the multivariate regression substantiate our findings from the correlation estimations as it fits all of the slopes seen in the plots. The indicator variable of the regression is the time point, and the model uses the level of ATF4 as a predictor, the time point as an indicator, and the outcome is the level of NGFR.

To answer our 3 research questions:

Under the experimental condition of 3.16 uM drug Vem:

- At 6 hours, ATF4 is most predictive of the protein levels of the phenotype indicator NGFR.
- Between 0.5h to 15h, there exists a fairly strong positive correlation between ATF4 and NGFR. However, starting at 24h, this changes to an inconclusive weak correlation.
- Using the equations from multivariate regression, we can use ATF4 to predict the level of NGFR of a cell at different timepoints.

Conclusion

Overall Conclusion:

- Through our conclusions made, we have shown the predictive relationship between ATF4 and NGFR under the condition of 3.6uM Vem. Since the state of NGFR directly influences the phenotype of the cell, ATF4 may act as an indicator when estimating the cellular phenotype for a potentially cancerous cell.
- Further research into drugs and dosages that could control the protein level of ATF4 would bring us one step closer to finding a way to lower NGFR levels and therefore, change a bad cellular homeostasis to a good cellular homeostasis. By doing this, we are approaching our goal of fighting cancer.

Limitations:

- There are possible confounding variables, such as other transcription factors.
- Correlation is not strong for timepoints from 24 hours onwards.
- Our analysis is fixed to the experimental condition of 3.16 uM Vem.

References and Acknowledgements

Comandante-Lou, N., Baumann D. G., Fallahi-Sichani M. (2021, December 7). "AP-1 transcription factor network explains diverse patterns of cellular plasticity in melanoma". bioRxiv. https://www.biorxiv.org/content/10.1101/2021.12.06.471514v1.full

Research dataset provided by Professor Scott and Dr. Heman Shaker from UVA.

Code for defining 'parameter=argument' option from Professor Scott's slides.

Special thank you to Professor Scott for his help and support throughout the process of this presentation and thank you to TA Amin for his help with the code for categorizing phenotypic outcomes and data cleaning.

Thank You for Listening!