Modeling gene expression using five histone modifications

Fifth Annual Primes MIT Conference

Lalita Devadas

Mentor: Angela Yen

May 17, 2015

Outline

1. Biological Background
2. Method
3. Results
4. Moving Forward

Outline

1 Biological Background

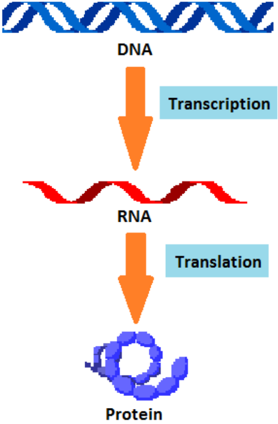
2 Method

3 Results

4 Moving Forward

Gene Expression

Central dogma of molecular biology



Gene Expression

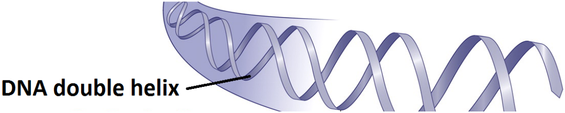
Relevance

- Important to understanding biological activity
- Crucial to advances in medicine
- Detection, prevention, and treatment of disease

Gene Expression

Regulation

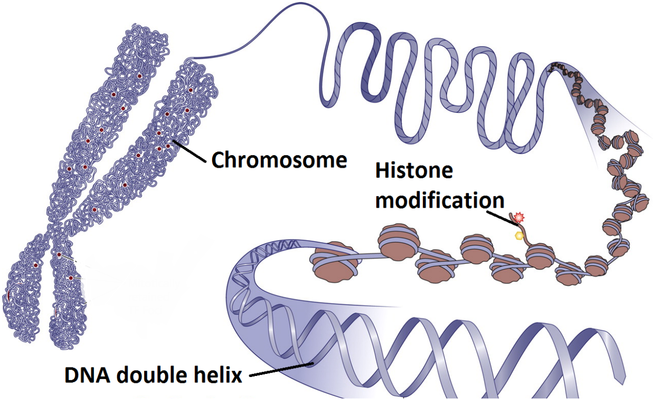
- Genetic
  - Sequences of nucleotides (ACTG)



Gene Expression

Regulation

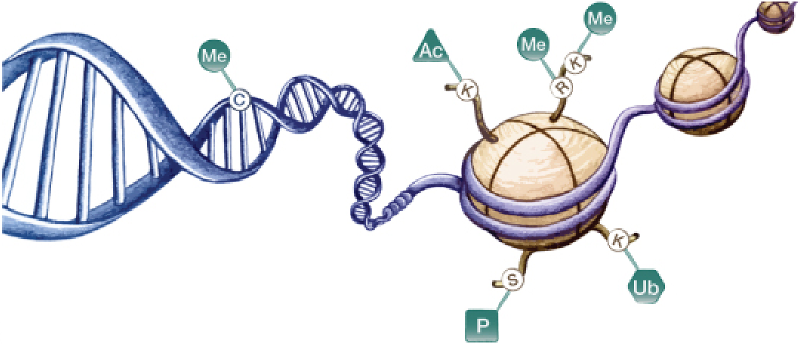
- Epigenetic
  - Changes to environment surrounding DNA



Epigenetics

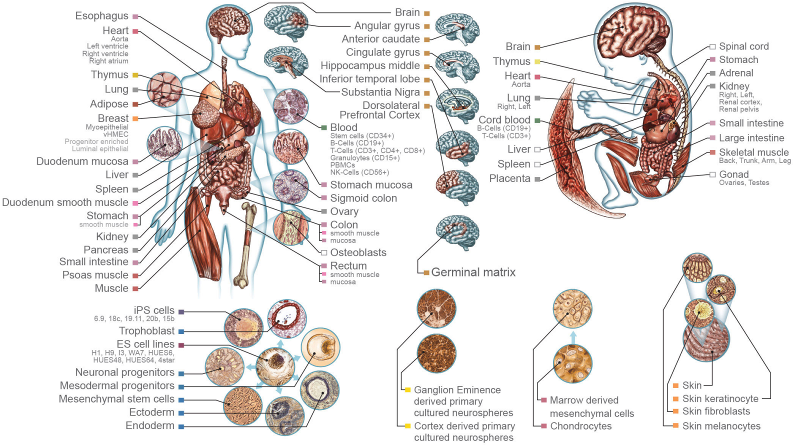
Histone modifications

- Chemical changes to histone protein core or protruding tail



Epigenomes

Roadmap Project



Outline

1 Biological Background

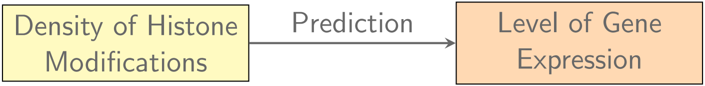
2 Method

3 Results

4 Moving Forward

Data Pipeline

Objective

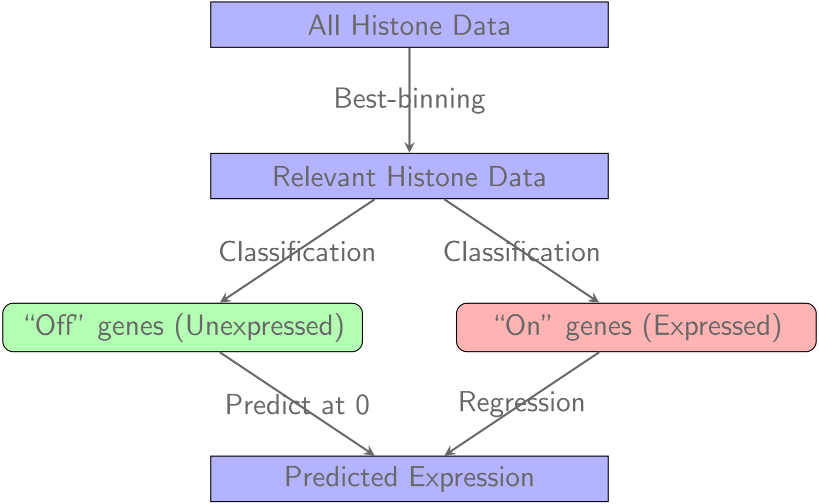

Density of Histone Modifications

Prediction

Level of Gene Expression

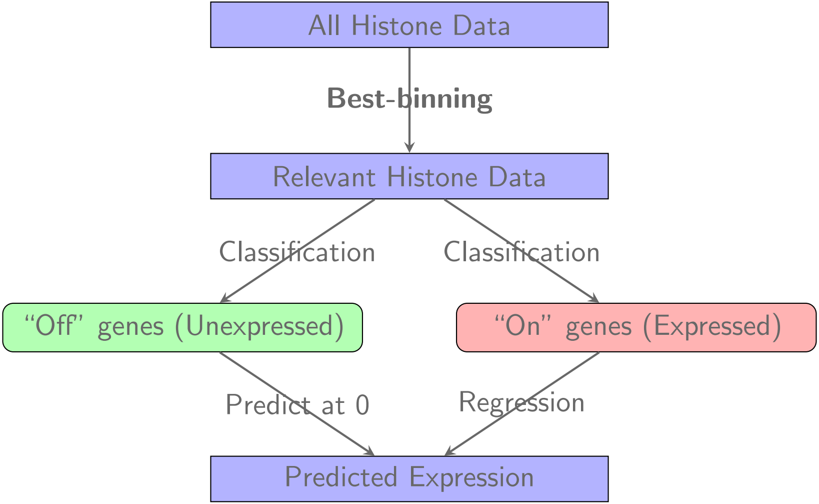
Data Pipeline

Overview



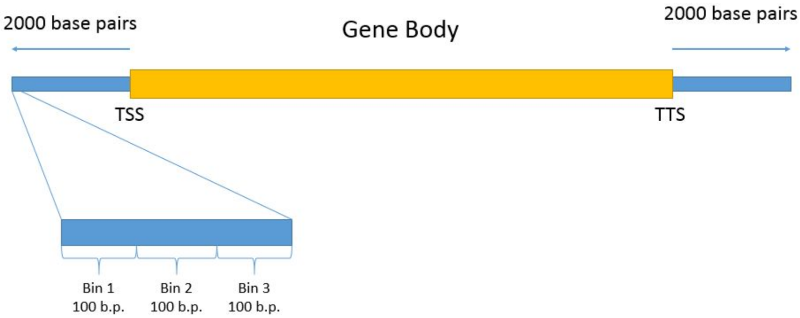
Data Pipeline

Best-bin approach



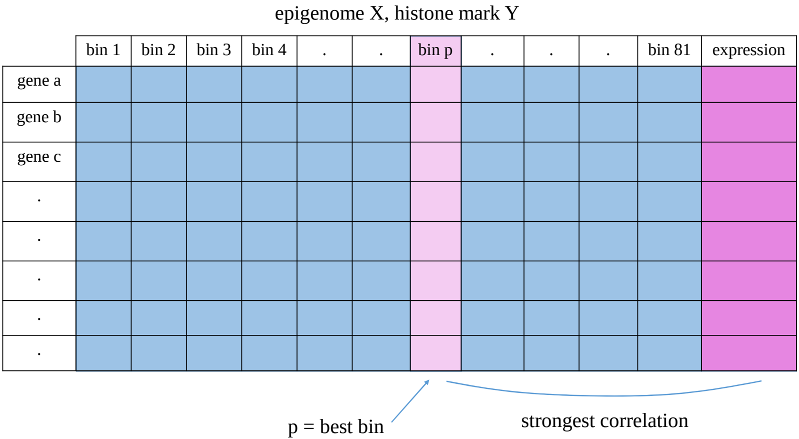
Best-bin approach

Dividing genes



Best-bin approach

Choosing best bin



epigenome X, histone mark Y

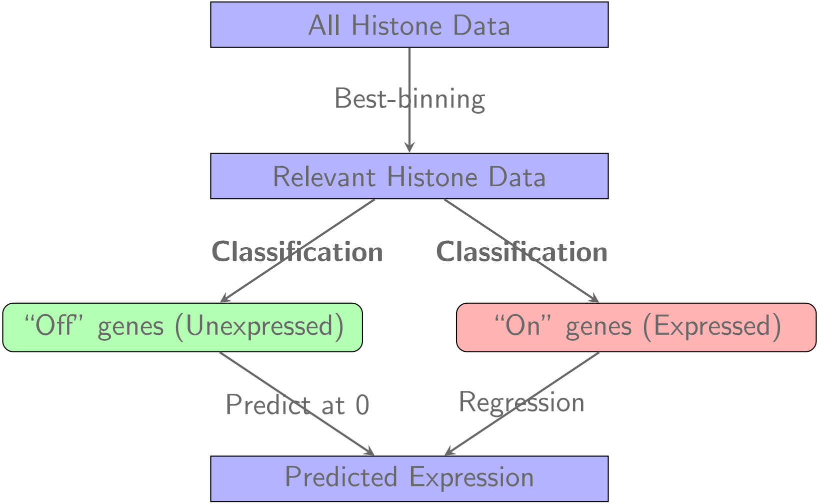
	bin 1	bin 2	bin 3	bin 4	.	.	bin p	.	.	.	bin 81	expression
gene a												
gene b												
gene c												
.												
.												
.												
.												
.												

p = best bin

strongest correlation

Data Pipeline

Classification

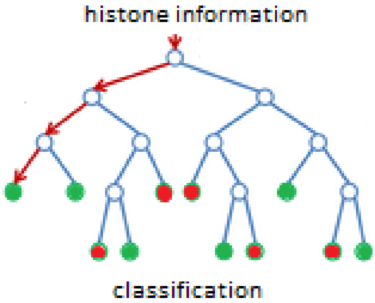


Types of Models

Random Forest

Random Forest model

Returns majority vote of classification determined by a group of decision trees

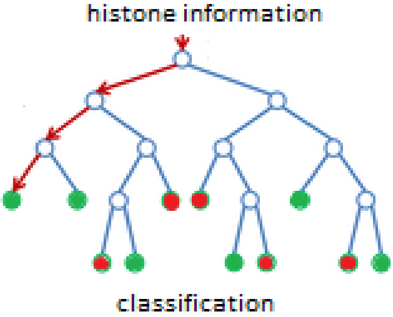


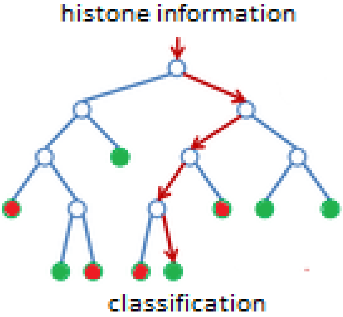
Types of Models

Random Forest

Random Forest model

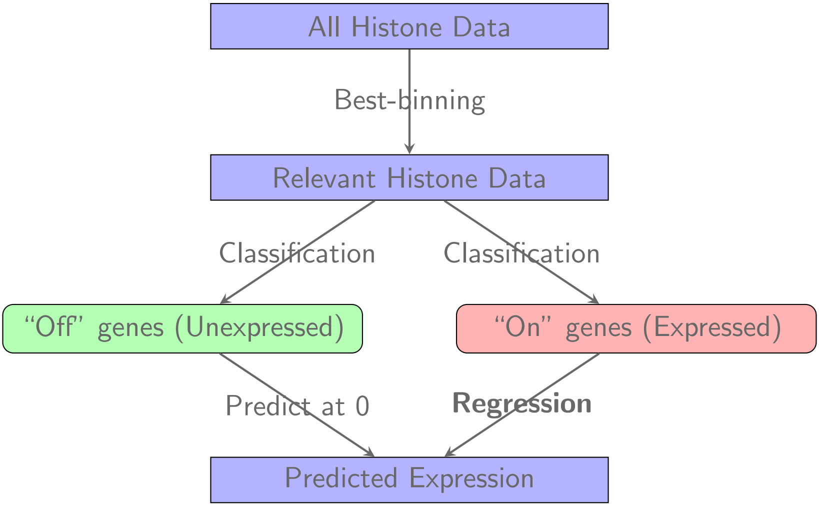
Returns majority vote of classification determined by a group of decision trees





Data Pipeline

Regression

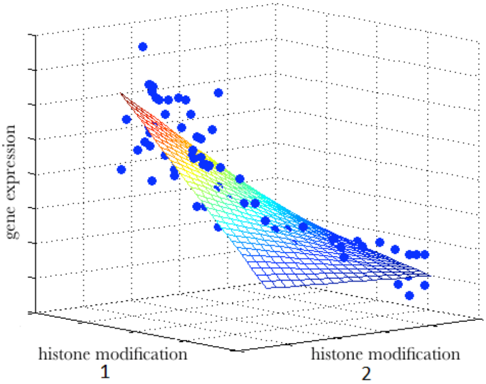


Types of Models

Linear Model

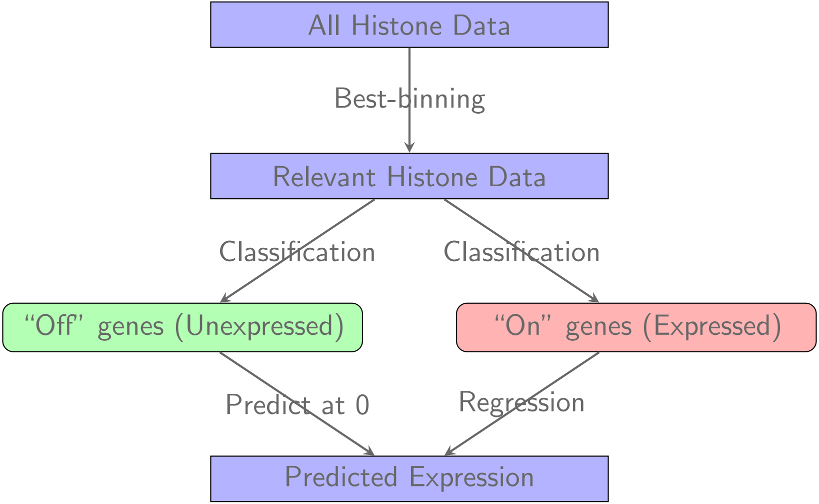
Linear model

Finds a linear correlation between predictors and response



Data Pipeline

Overview



Outline

1 Biological Background

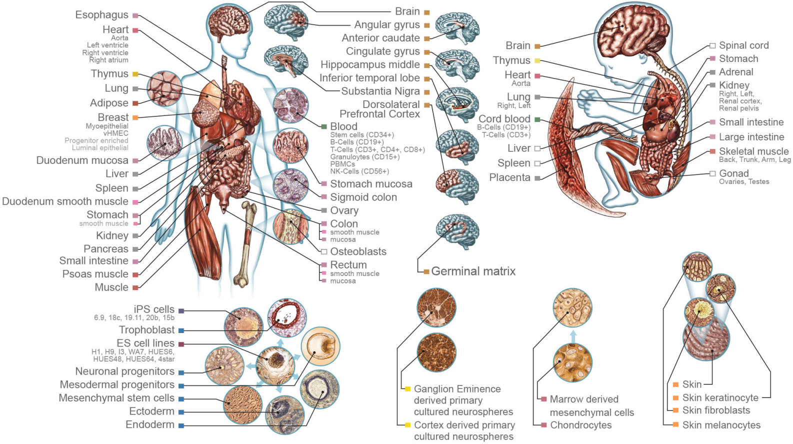
2 Method

3 Results

4 Moving Forward

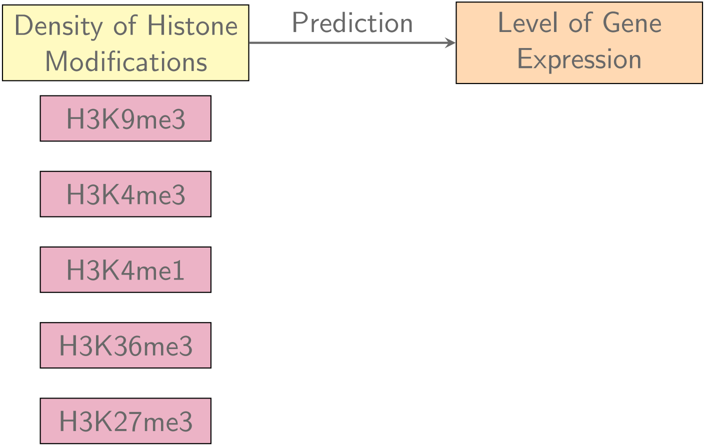
Epigenomes

Roadmap Project



Data Pipeline

Objective



Density of Histone Modifications

Prediction

Level of Gene Expression

H3K9me3

H3K4me3

H3K4me1

H3K36me3

H3K27me3

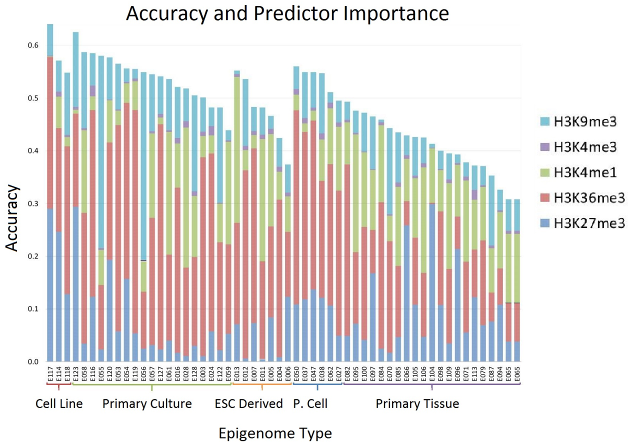
Results of Pipeline

Conclusions

- Models created for cultured epigenomes have a much higher predictive power than those created for tissue samples
- H3K36me3 is the most important histone mark used for prediction

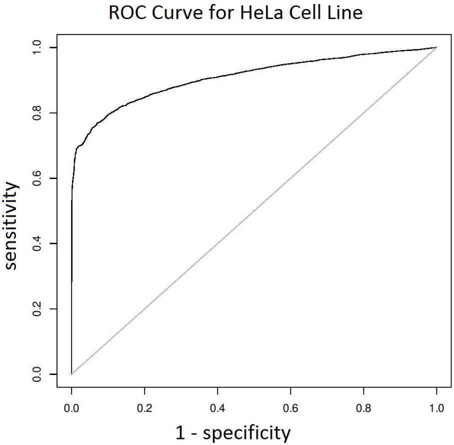
Results of Pipeline

Graph



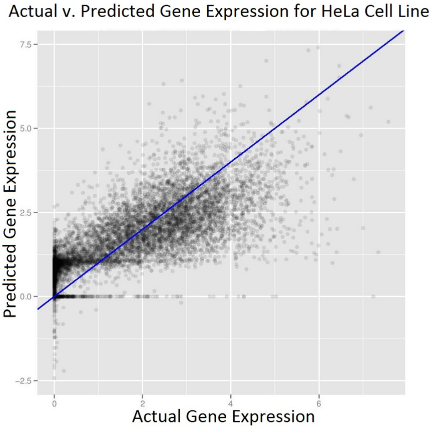
Specifics of Best Model

Classification Accuracy



Specifics of Best Model

Regression Accuracy



- every data point represents one gene
- r^2 value: 0.640

Outline

1 Biological Background

2 Method

3 Results

4 Moving Forward

Next Steps

- Improve predictive power
- Broaden scope of predictors and response
- Further analysis of current results
- Apply procedure to different data
- Release code as a tool for other researchers

Acknowledgements

I would like to thank:

- My mentor, Angela Yen
- Prof. Manolis Kellis
- Roadmap Project
- PRIMES program
- My family