The Fault Tolerant Epigenome
&
its Data Profile

Somali Chaterji, PhD

https://www.cs.purdue.edu/homes/schaterj/

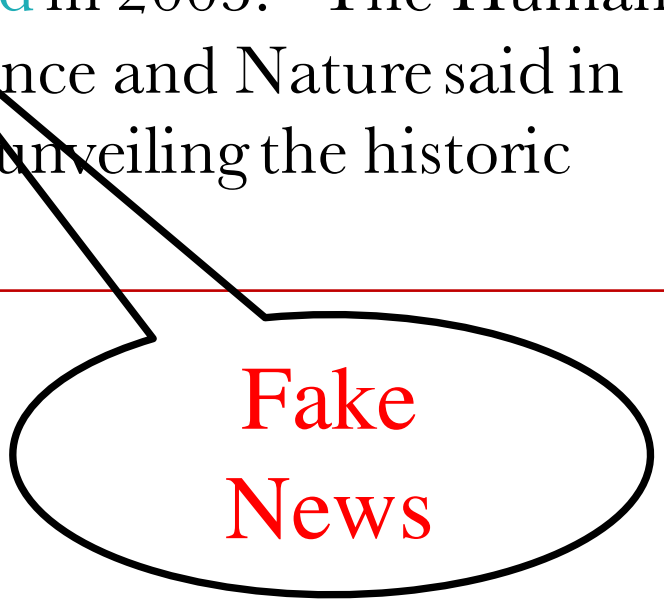
https://bitbucket.org/cellsandmachines/

Department of Computer Science, Purdue University

PURDUE UNIVERSITY

2003

The feat made headlines around the world: "Scientists Say Human Genome is Complete," the New York Times announced in 2003. "The Human Genome," the journals Science and Nature said in identical ta-dah cover lines unveiling the historic achievement.



There was one little problem.

"It's very fair to say the human genome was never fully sequenced," **Craig Venter**

Human variation

It turns out, in the grand scheme, we're all very, very similar, genetically: **99.9 percent of people's genes are identical**. It's in that last one-tenth of 1 percent where we find all of human variation.

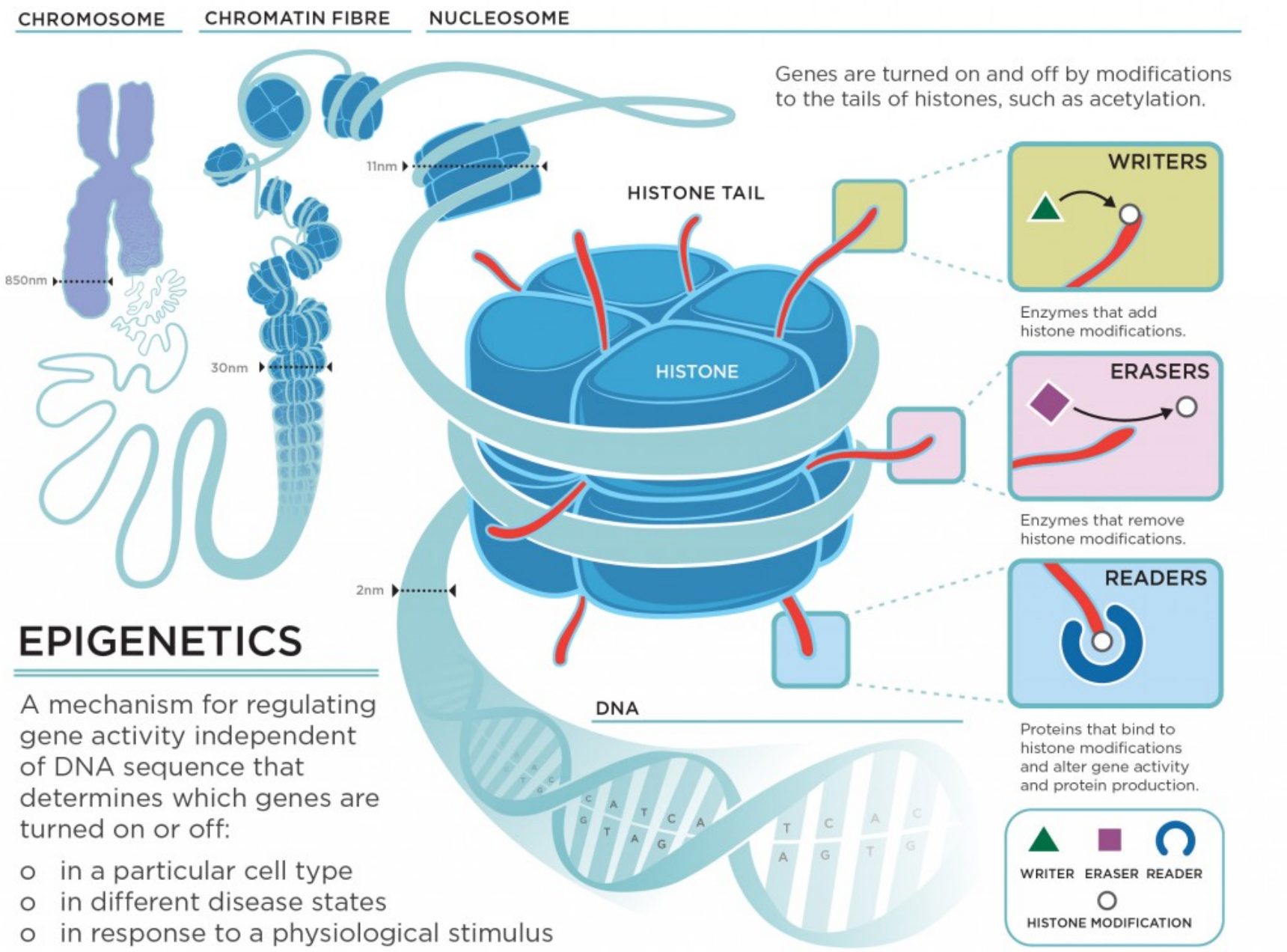
More than 98% of the human genome does not encode protein sequences, including most sequences within introns and most intergenic DNA.

98% of human genome is "**Junk DNA**"

Wait there is more: Epigenome!

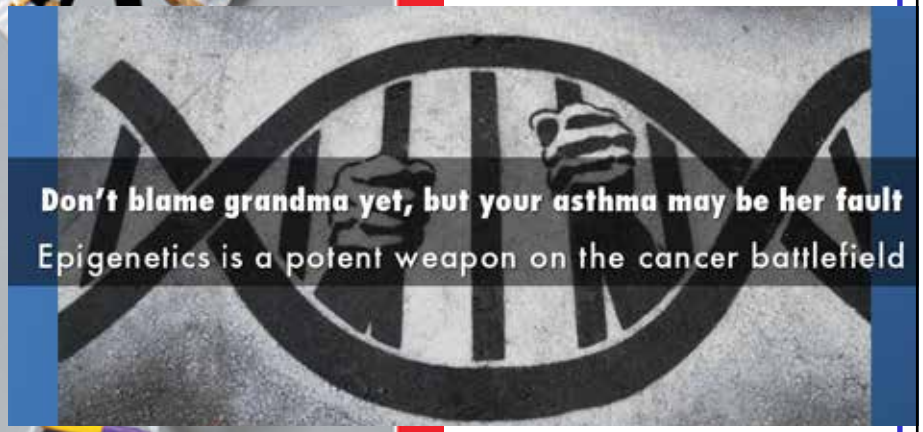
Fake News

Epigenetics: What does the epigenome look like?



PURDUE UNIVERSITY

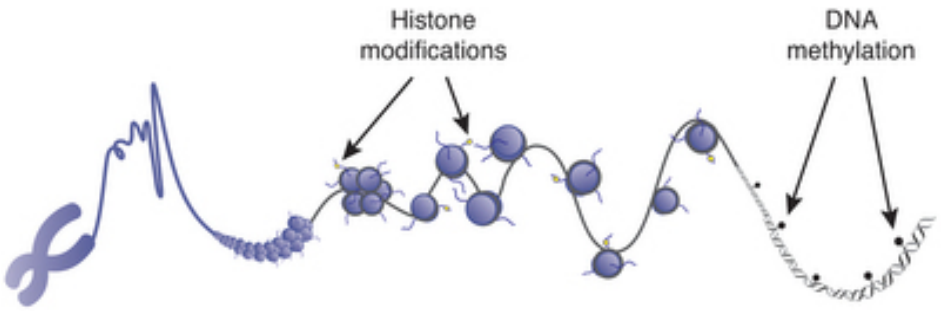

JANUARY 18, 2010
Joe Klein: The CIA's Afghan Disaster
Yemen: The New Center Of Terror
Why the Recession Hasn't Been Cool To Teens
TIME
WHY YOUR DNA ISN'T YOUR DESTINY
The new science of epigenetics reveals how the choices you make can change your genes —and those of your kids
BY JOHN CLOUD


Don't blame grandma yet, but your asthma may be her fault
Epigenetics is a potent weapon on the cancer battlefield


nature video
Epigenome: The symphony in your cells
0:06 / 5:05
CC

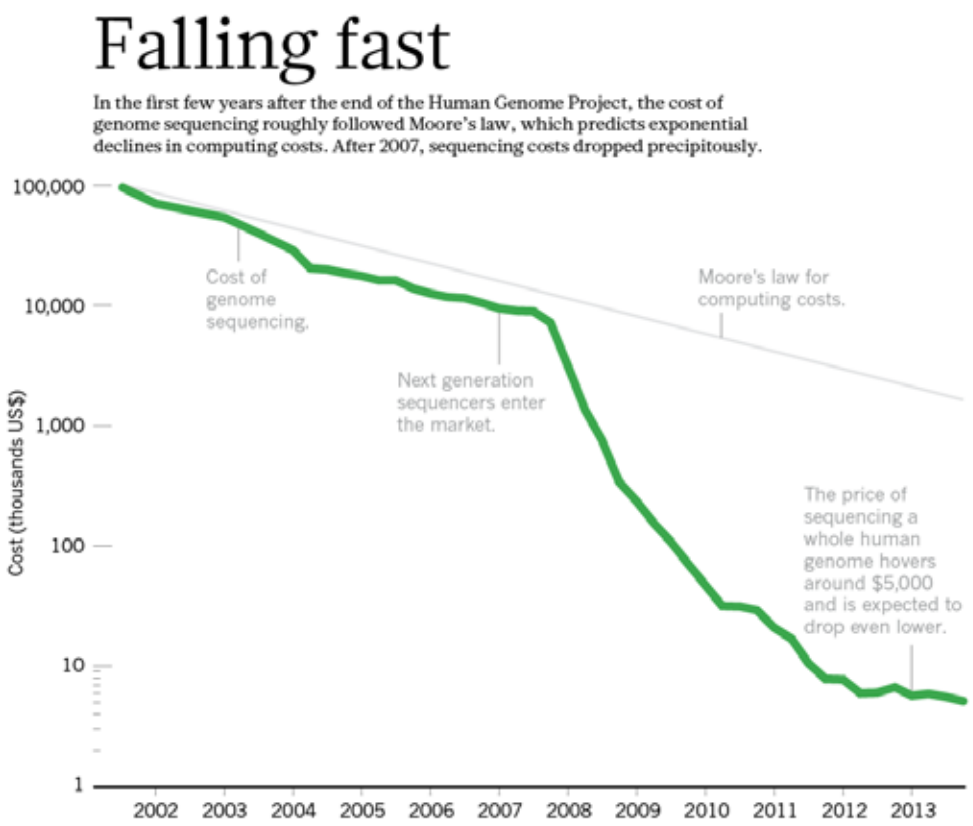
Why do we need to analyze the epigenome now?

- Lots and lots of epigenetic data



- Promise that this data can be mined to create personalized models of health and disease
- Ability to create very flexible models
  - Nonparametric Bayesian Models
  - Neural Networks
- Enough computing power
- Powerful priors that can defeat the curse of dimensionality

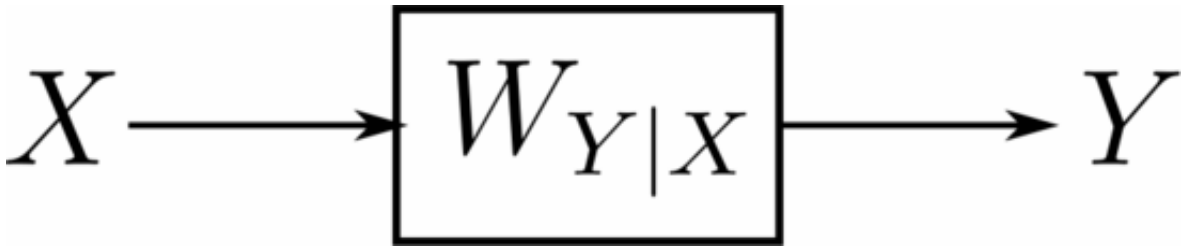
Driver: Big Data Explosion in Genomics



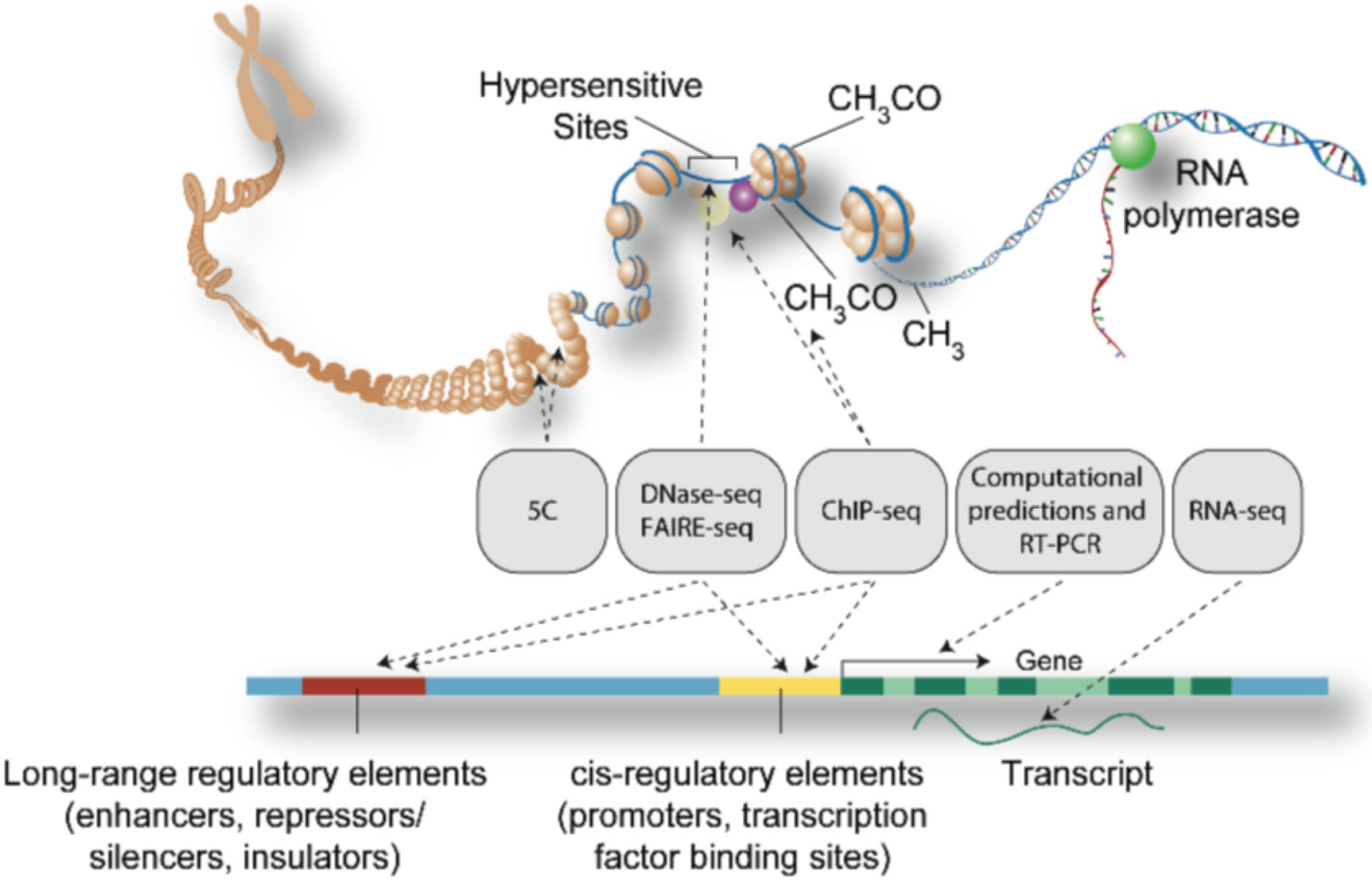
- Amount of genomics data is increasing rapidly
- Can we use this data to make personalized medicine approaches more disciplined?
- What are the ML classifiers best suited to this problem area
- What does the input space look like?
- Can we speed up the slow training process?
  - New datasets are being generated through genomics experiments at a fast rate
  - Diverse datasets need separate models to be trained
- Can we make use of large distributed clusters to speed up training?
- Can we make the overall development of computational genomics algorithms speedier and more efficient

Epigenomic data for personalized medicine

- Create different models for classes of individuals or cell types based on their features (demographics, -omics data, …) (*Training phase*)
- Use model on hitherto unseen individual or cell type – then predict the individual's predilection for disease, etc. (*Prediction phase*)



Epigenomic big data



Precision epigenome-based therapeutics

- Which are the hotspots of genetic variants?
  - Genomic enhancers (Non-coding regulatory DNA)
- Which are the mobile, endogenous fine-tuners of gene expression?
  - MicroRNA (Non-coding regulatory RNA)

AVISHKAR

MicroRNA-based targeting

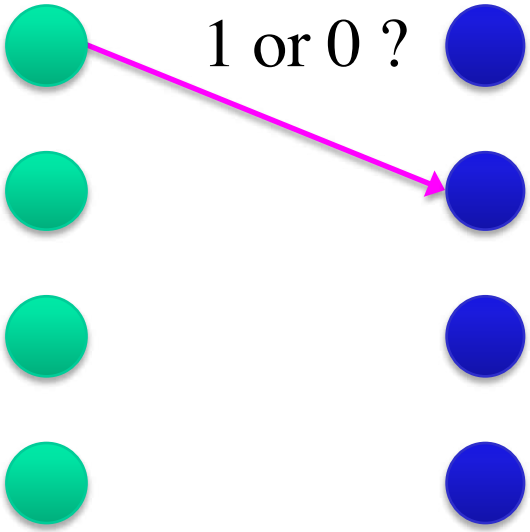
- miRNA are 22 nucleotide (nt) strings of RNA, base-pairing with messenger RNA (mRNA) to cause mRNA degradation or translational repression
- Can be thought of as biology's dark matter: small regulatory RNA that are abundant and encoded in the genome
- Dysregulation of miRNA may contribute to diverse diseases
- Canonical (i.e., exact) matches involve the **miRNA's seed region (nt 2-7)** and the 3' untranslated region (UTR) of mRNA and were thought of as the only form of interaction
- Recent high-throughput experimental studies have indicated the high-preponderance of "non-canonical" miRNA targets

Our Contributions in microRNA Target Prediction

- Many tools are available for miRNA target prediction
  - However, they require complicated configurations and are computationally expensive
- Our contributions:
  1. Most general microRNA targeting algorithms
  2. Distributed pattern mining algorithm
  3. Visualizing the predicted miRNA-mRNA mappings

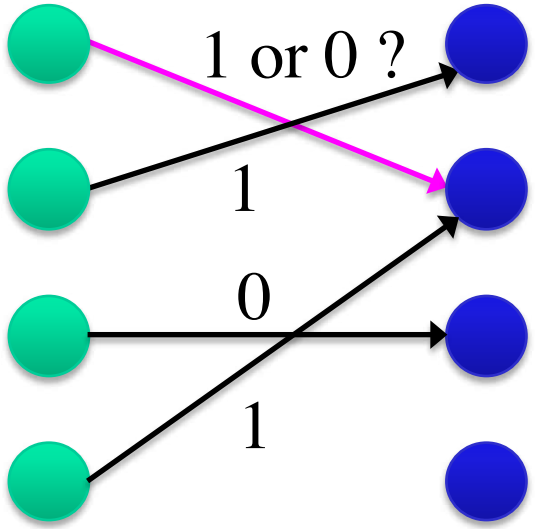
Problem Statement

- Predict if a miRNA targets an mRNA (segment)



miRNAs mRNA segments

- Ideally we would want some experimentally verified edge labels to train on

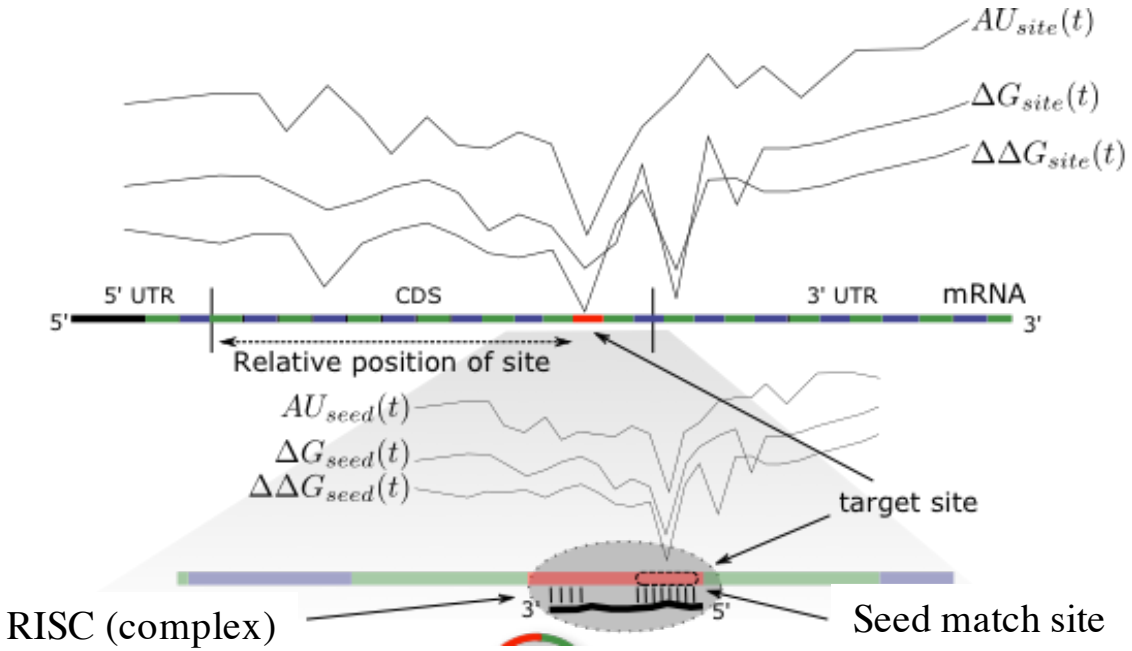


miRNAs mRNA segments

Edges with ground truth labels

Edges whose labels have to be predicted

Methods: Feature Construction



- The alternating blue and green regions denote the 13 consecutive windows around the miRNA target site (red). These are the windows where the average thermodynamic and sequence features are computed.
- Compute interaction profiles at two different resolutions
  - Window size of 46 and using the entire miRNA: **"site" curves**
  - Window size of 9 and only using the seed region of the miRNA: **"seed" curves**
- Use coefficients of B-spline basis functions as features for classifier
- We hypothesize that the curves are different for the positive and negative samples.

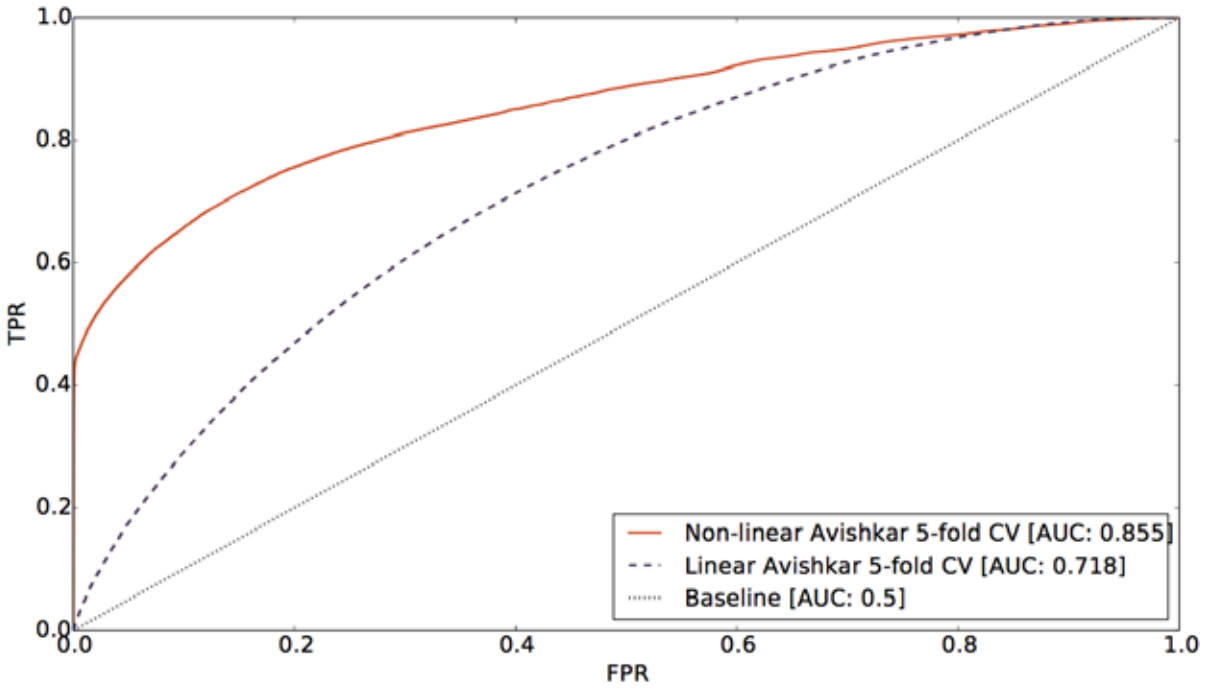
Improving Classification Performance with Kernel SVM

- Linear classifier suffers from high bias (large error even on training set)
- Solution: Use more complex learning model
  - Non-linear or Kernel SVM
- SVMs suffer from a widely recognized scalability problem in both memory use and compute time.
- Kernel SVM computational cost: $O(n^3)$
- Does not scale beyond a few thousand examples for feature vector of dimension ~ 150
- Running serial version on entire dataset (300 GB) will take 45.4×10^3 years!

Making Kernel SVM Scale Up

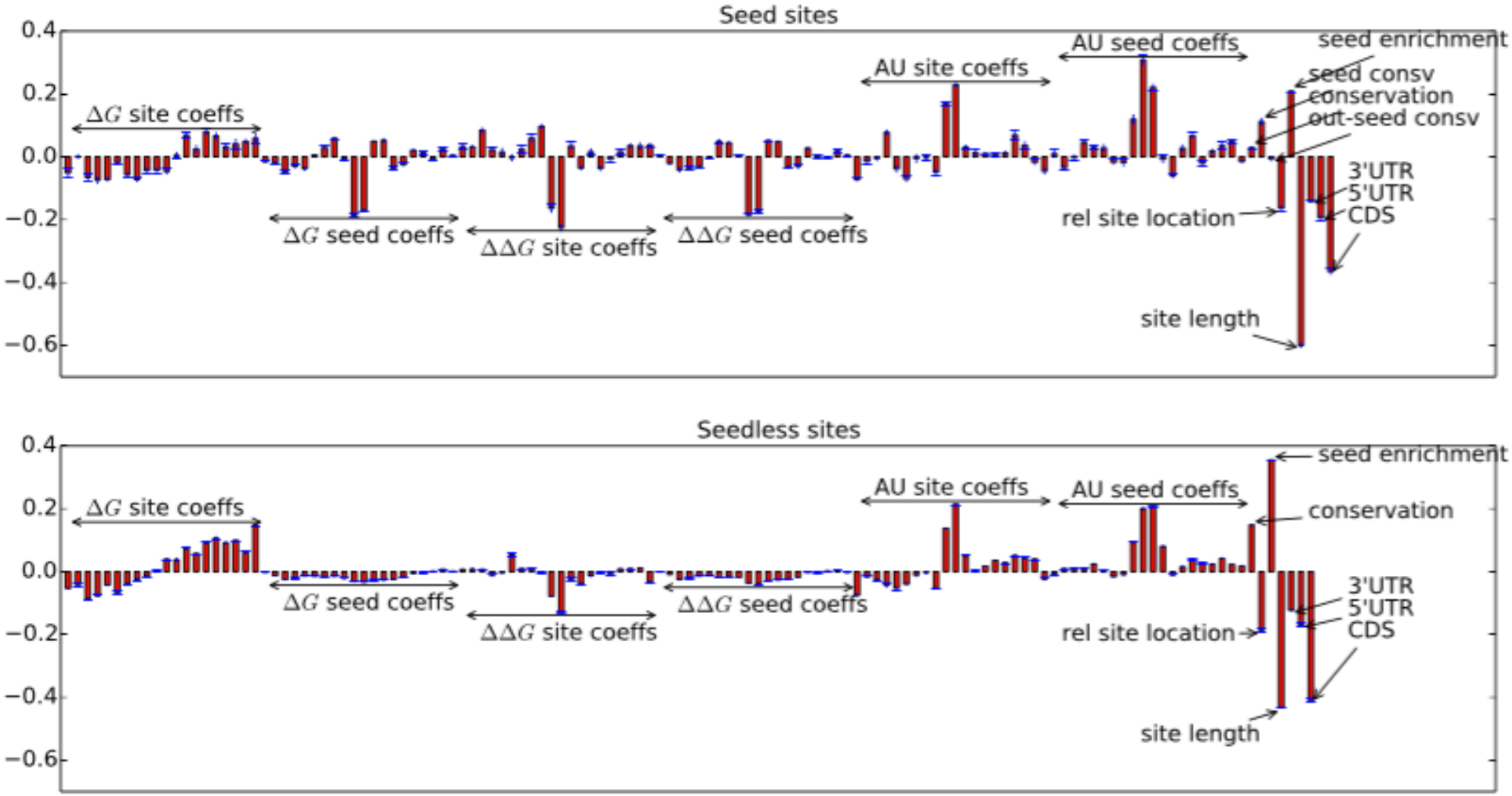
- Biological insight: miRNAs within an miRNA family share structural similarities
- Therefore, we create a separate non-linear classifier for each miRNA family
- Within each family, we train in parallel using Cascade SVM approach

Results: ROC Curve for Linear and Non-linear SVM



- ROC curves for the ensemble linear model and ensemble non-linear model, obtained by varying the probability threshold for the output of the SVM.
- One possible operating region is FPR = 0.2: TPR for the linear model is 0.469, while the TPR for the non-linear model is 0.756.
- TPR is 150% better than competition.

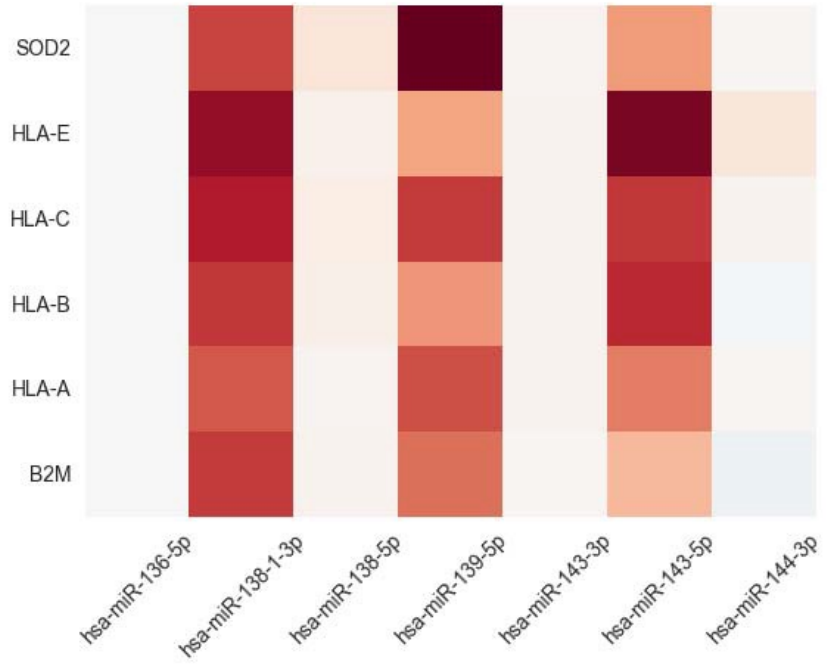
Results: Feature Importance



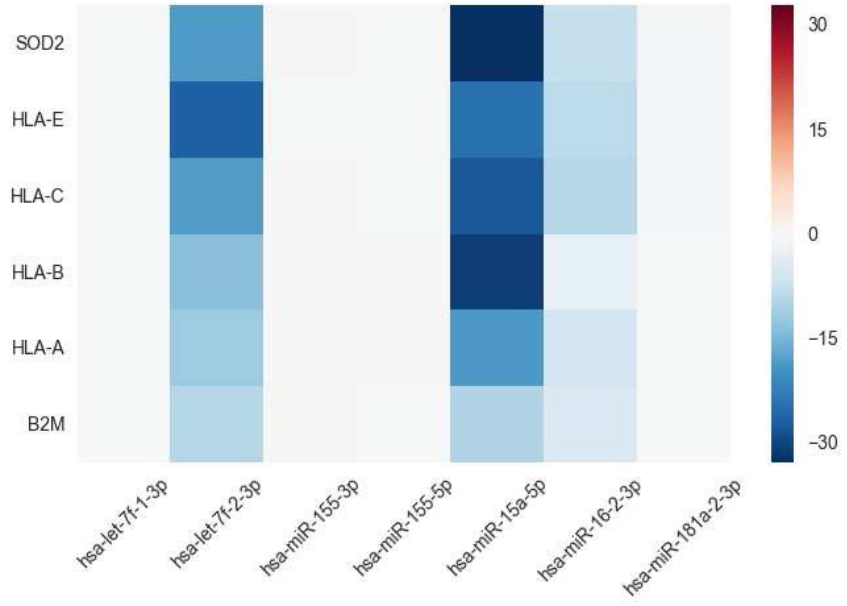
PURDUE UNIVERSITY

Visualization Snapshots

- Provided through a web server: http://cygnus.ecn.purdue.edu/



Upregulation



Downregulation

PURDUE UNIVERSITY

TIRESIAS

Context-Specific Prediction

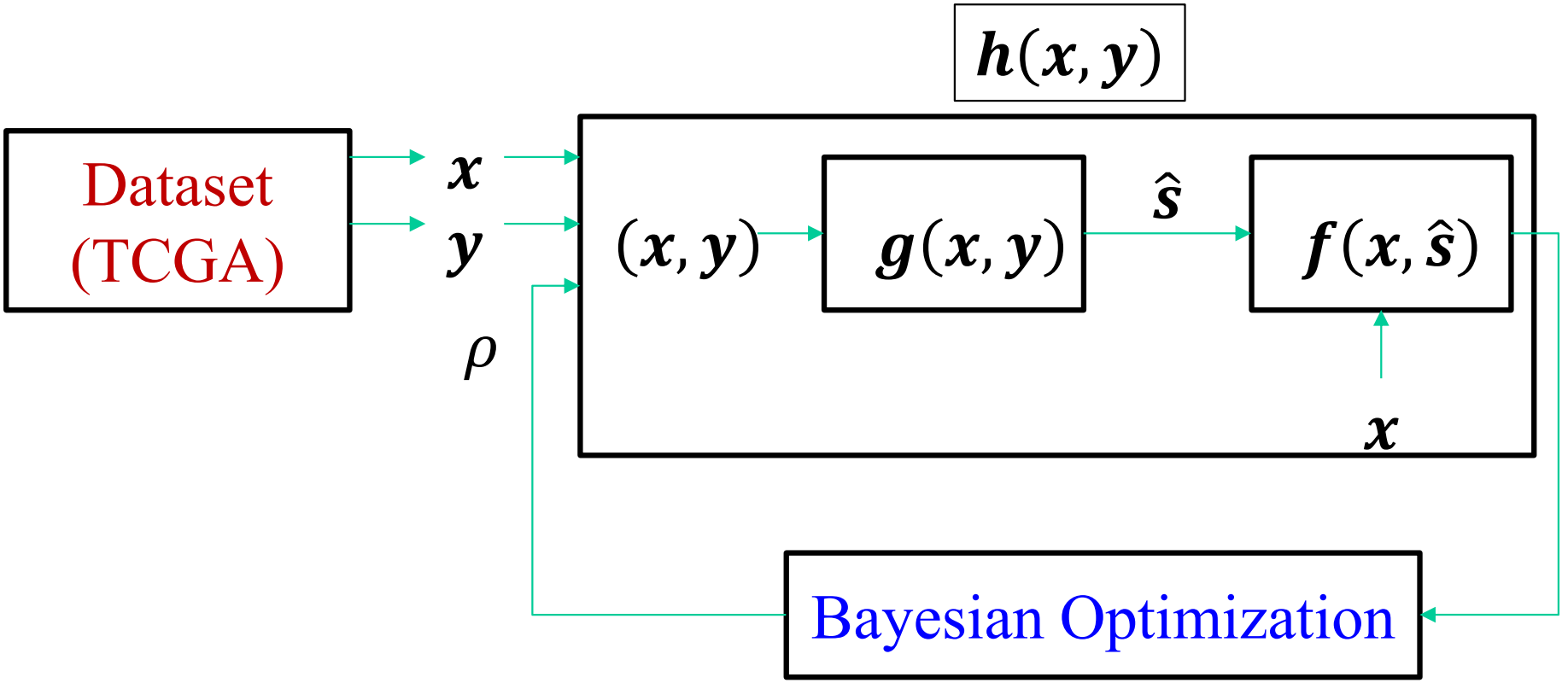
- Goal
  - Integrate expression data about miRNA-mRNA, with prior sequence data, for predictions
  - Perform prediction under dynamic conditions of interest, such as in disease conditions and in specific tissues
- Prior work
  - Little prior work can achieve the two goals above
  - [Bioinformatics 2013][§] can only handle linear effects and downregulation
  - Recent results show that the effects from multiple interacting miRNAs is non-linear
  - Recent results show that miRNA regulation can be upregulation

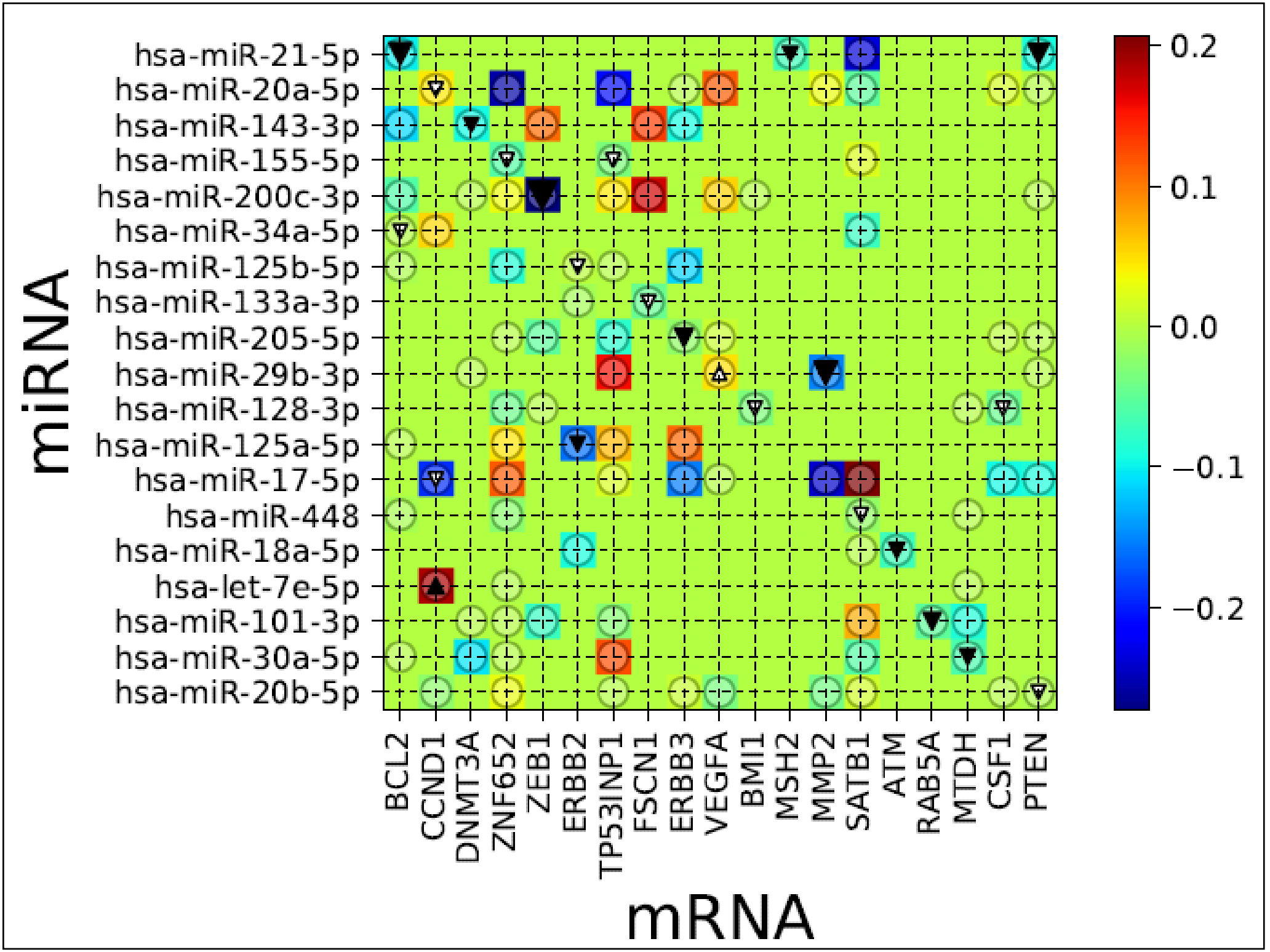
[§] Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation: Hai-Son Le, Ziv Bar-Joseph

Our Solution Approach: *Tiresias*

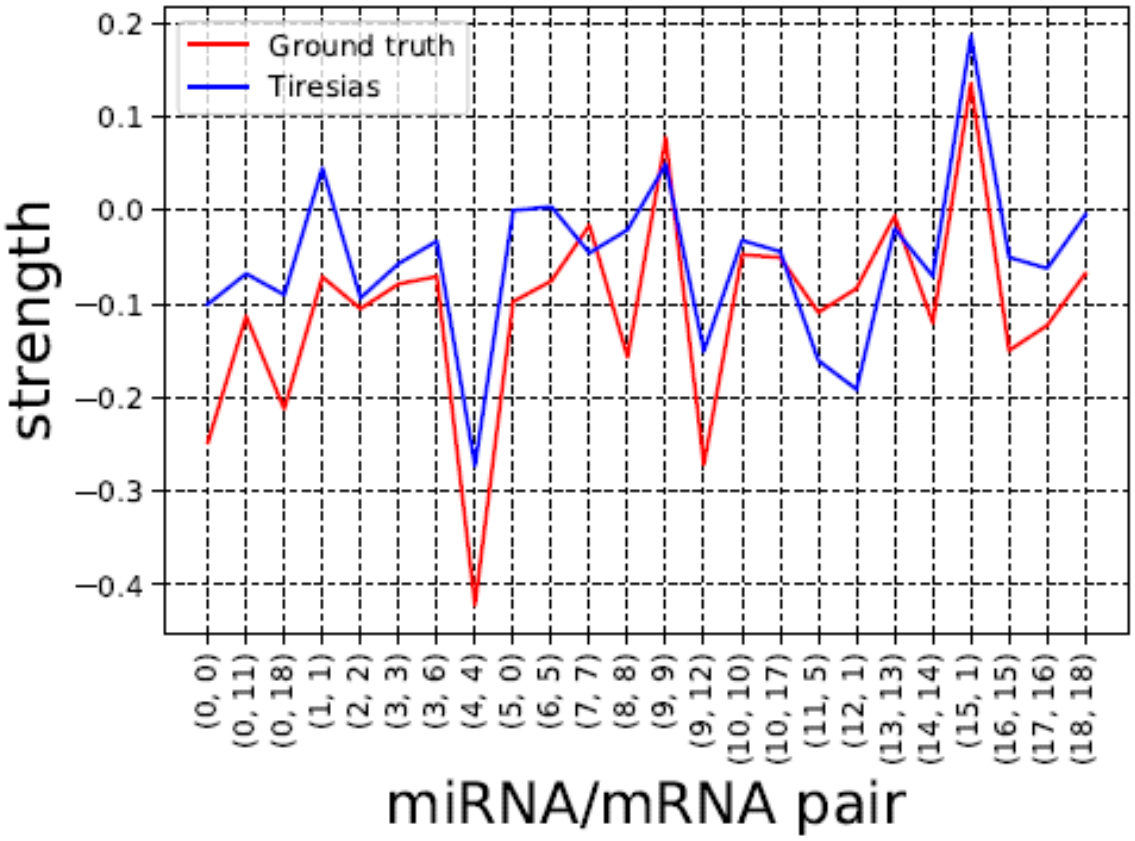
- Tiresias computationally predicts miRNA targets under context-specific conditions by incorporating expression-level data into sequence-based prediction results
- Tiresias decouples the problem making the learning easier
  - The first stage estimates miRNA targets (as in *Avishkar*)
  - The second stage estimates regulation weights based on the previous stage's outputs
- Tiresias considers up-regulation and down-regulation simultaneously
- Tiresias extends prediction to a complex non-linear regulation model using two Artificial Neural Networks (ANNs)
- Tiresias characterizes the density of miRNA-mRNA interactions by one single hyper-parameter (ρ) unlike prior work

Our Solution Approach: *Tiresias*




miRNA
hsa-miR-21-5p
hsa-miR-20a-5p
hsa-miR-143-3p
hsa-miR-155-5p
hsa-miR-200c-3p
hsa-miR-34a-5p
hsa-miR-125b-5p
hsa-miR-133a-3p
hsa-miR-205-5p
hsa-miR-29b-3p
hsa-miR-128-3p
hsa-miR-125a-5p
hsa-miR-17-5p
hsa-miR-448
hsa-miR-18a-5p
hsa-let-7e-5p
hsa-miR-101-3p
hsa-miR-30a-5p
hsa-miR-20b-5p
BCL2
CCND1
DNMT3A
ZNF652
ZEB1
ERBB2
TP53INP1
FSCN1
ERBB3
VEGFA
BMI1
MSH2
MMP2
SATB1
ATM
RAB5A
MTDH
CSF1
PTEN
mRNA
0.2
0.1
0.0
−0.1
−0.2

Regulation strength and direction



Experimental results showed that Tiresias performs better than existing computational methods such as GenMiR++, Elastic Net, and PIMiM.

For the TCGA breast cancer dataset, Tiresias showed a true positive rate of about 88% in recovering the ground truth regulatory interactions between miRNAs and mRNAs.

Expertise

- Computational biology
  - Mining epigenomics and metagenomics data
  - Data structures for efficient computing in genomics
  - Federation of infrastructures for genomics
  - Faster and more efficient evolution of new algorithms
  - Predictive analytics for genome editing and therapeutic targeting
- Data-driven cell engineering
  - Efficient and precise genome editing
  - Bridging the gap between systems biology and synthetic biology
- Contact: Somali Chaterji; schaterj@purdue.edu