*Bioinformatics*, 2017, 1-2

doi: 10.1093/bioinformatics/boss

Advance Access Publication Date: 08 May 2017

Cool Things

|  |
| --- |
| Cool Things  HMMicro: predicting miRNA targets from epigenetic data  Wu, Felix\*,1 and Nick Giangreco\*,1  1Department of Systems Biology, Columbia University, New York, NY, USA  \*To whom correspondence should be addressed.  Associate Editor: Voldemort  Received on None; revised on of your; accepted on business  Abstract  **Motivation:** Multiple gene regulatory layers give rise to complex phenotypes, which include redundant patterns to maintain homeostasis and other functions. Identifying miRNA binding sites, another regulatory layer at the post-transcriptional level, is difficult and infeasible to obtain for any given condition and treatment.  **Results:** We developed a Hidden Markov Model (HMM) for predicting *de novo* miRNA binding sites in HEK293 cells using epigenetic information provided by ENCODE. Our model shows a low error rate and a high true positive rate, saying that leveraging known epigenetic information of a cell can be useful in predicting novel regulatory layers.  **Availability:** https://github.com/ngiangre/HMMicro  **Contact:** flw88@cumc.columbia.edu  **Supplementary information:** Supplementary data are available at *Bioinformatics* online. |

# Introduction

The central dogma of molecular biology tells us that DNA is transcribed to RNA, RNA is translated to protein, and information cannot not flow from protein back to the previous molecules. The population of proteins give rise to the complex and dynamic cellular phenotype that keeps homeostasis or gives rise to disease. The phenotype of a cell is a product of the reactions and relationships between many molecular layers, such as chromatin modifications, transcription factor binding, and chromatin confirmation. Each layer provides a regulatory logic, which necessarily robust for maintaining homeostasis when bombarded by its environment.

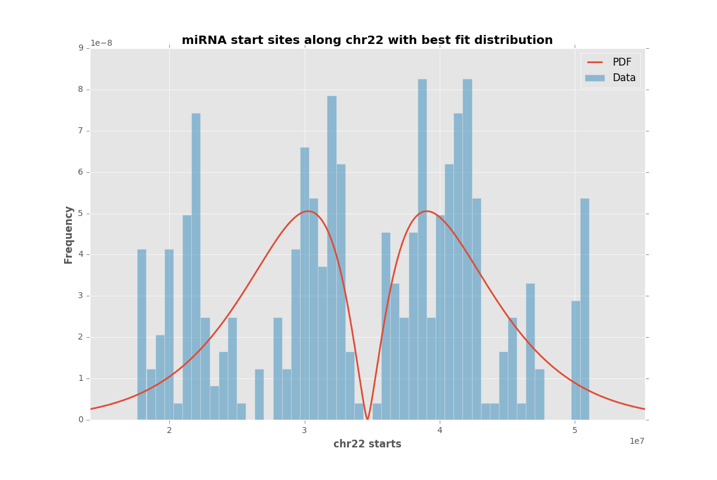
We hypothesize that there exists redundancy between the many layers that produce the population of proteins in a cell. This redundancy is found in molecular patterns that is present in all layers, allowing for robust phenotypes. Thus, integrating knowledge from all different molecular layers can give us a more succinct observation of the underlying phenotype of a cell.

With the advent of next generation sequencing and the popularity of high throughput experimentation, all molecular layers are not able to be assayed due to time and financial constraints. While understanding molecular phenotypes can help elucidate molecular mechanisms in health and disease, predicting patterns in other molecular layers can help in uncovering complex cellular phenotypes.

We present a HMM for *de novo* prediction of miRNA binding sites in HEK293 cell line. We train our model using various epigenetic experiments assayed through the ENCODE consortium, and test our model using experimentally validated miRNA binding sites for this model system. We obtain a low prediction error and high prediction of known miRNA binding sites. This method can be expanded upon for leveraging the vast amounts of existing information that give rise to a cellular phenotype.

# Methods

We obtained HEK293 epigenetic signal at nucleotide resolution on chromosome 22 from ENCODE [ref]. We obfuscated retrieving and processing the raw data by using the R package *DeepBlueR* [ref]. This allowed us to download and process the data on a remote server. We obtained normalized read signal (wiggle file signal) for 8 experiments (7 ChIP-Seq against KAP1, POL2RA, TCFL2, ZNF263, CTFC, ELK4, and H3K4me3; 1 DNase experiment) at nucleotide resolution across chromosome 22. We also developed in-house R scripts for querying and performing additional processing.

The hidden markov model was developed using the *Pomegranate* python package [ref]. We developed the HMM in three steps: pre-processing, training and testing. In the processing step, we computed the principal components of the epigenetic matrix. Also, using experimentally validated miRNA binding sites [ref], we modeled the distribution of start sites across chromosome 22 to identify an appropriate prior distribution of the binding state. In the training step, we used the Baum-Welch algorithm to learn emission (epigenetic characteristics) and refine the prior transition (binding and non-binding miRNA states) probabilities from the training dimension-reduced epigenetic matrix. In the testing step, we applied a 5-fold cross validation to test the HMM. The code for these analyses are provided in multiple python scripts and jupyter notebooks.

# Results

**Fig. 1. Principal components analysis shows epigenetic variation captured by few di-mensions..** Reducing dimensions of the HEK293 epigenetic features showed majority of variance captured in the first few components.

## Multiple epigenetic layers contain redundant information

We investigated the amount of variation present in normalized epigenetic signal from 8 different ChIP-Seq and DNase across chromosome 22 in HEK293 cells. This will give an indication of the amount of redundancy across the epigenetic layers. In agreement with our hypothesis of redundant patterns, we observed the first principal component captured about 85% of the variation across the epigenetic layers (Fig. 1).

### 3.2 *De novo* miRNA binding site discovery using a Hidden Markov Model

### To model miRNA binding sites, we first identified an approximate prior distribution model by sampling all common distributions (Fig. 2). This best-fit model informed approximation of the binding state distribution in the HMM.

### We developed a HMM that would take into account multiple epigenetic features of our HEK293 cell model system to guide predicting of miRNA binding patterns.

**Fig. 2. Histogram and best-fit distribution for experimentally validated miRNA binding sites.** The distribution along chromosome 22 of the experimentally-validated miRNA start positions.

### To test our HMM, we conducted 5-fold cross-validation where each cross used 4-parts of the experimentally validated miRNA binding sites as training and 1-part as testing data. We found a low median error in our validation strategy and a high true positive rate.

### We identified several limitations and future work from our modeling strategy….

**Table 1.**Benchmark results of the cascade oscillators model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |S| | Predicted cost | Timing | Predicted speed | Speed |
| 1 | S219.20(100%) | 68m43s | 1.00 | 1.00 |
| 2 | 29.10+219.10(~50%) | 35m13s | 2.00 | 1.95 |
| 4 | 219.20(100%) | 68m43s | 1.00 | 1.00 |
| 10 | 29.10+219.10(~50%) | 35m13s | 2.00 | 1.95 |
| 20 | 219.20(100%) | 68m43s | 1.00 | 9.5 |

This is table foot note sample text This is table foot note sample text This is table foot note sample text

Acknowledgements

We’d like to thank Itsik Pe’er and the Spring 2017 Computational Genomics class at Columbia University for feedback and support.

Funding

This work has been supported by the training grant from the Integrated Program in Biomedical, Cellular, and Molecular Sciences at Columbia University Medical Center.

*Conflict of Interest:* none declared.

References

Hafner, M. (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1): 129-41.

ENCODE consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 489(7414): 0028-08.

Albrecht,F., List,M., Bock,C. and Lengauer,T. (2016) DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. *Nucleic Acids Research*, [doi:10.1093/nar/gkw211](http://dx.doi.org/10.1093/nar/gkw211)