CARNEGIE MELLON UNIVERSITY - COMPUTATIONAL BIOLOGY DEPARTMENT

April 23, 2017

Predicting Pseudouridylation of Individual RNA Bases Using Machine Learning

Christine Baek
christib@andrew.cmu.edu

Kevin Chon
khchon@andrew.cmu.edu

Deepank Korandla
dkorandl@andrew.cmu.edu

Tianqi Tang
tianqit1@andrew.cmu.edu

TODO: CHRISTINE abstract here. CB will complete this section after all other sections are written

1 INTRODUCTION

TODO: DEEPANK. 3 sections minimum covering 1) RNA modification - what is it, biological significance. 2) PseudoUracil - what is it, maybe include figure of pseudouracil 3) What has been done with PseudoUracil research. DO NOT FORGET TO CITE.

example for Deepank :

This is how you cite stuff [2].

2 METHODS

TODO: CHRISTINE - brief overview Main challenge of this project is that we have to build a per-base/position predictor of whether it is likely to be a post-transcription modification, while having limited labeled data, and diverse set of inputs. Instead of trying to build a model that attempts to cover all species, we will build models for specific species, and scale up from there. Overall approach is, build a multiple sequence alignment of multiple RNA reads from

closely related organisms. At this point, each base (independently, or in conjunction with its neighbors) can be used as a feature

2.1 Relevant Domain Knowledge

TODO: CHRISTINE - overall view of why we picked certain stuff, and logic for our design

2.2 Data

2.2.1 Data Collection

3 types of data
TODO: CHRISTINE

1. human chromosome rRNA (insert citation)
2. human rRNA (insert citation)
3. non-human rRNA (insert citation)

2.2.2 Data Processing

TODO: CHRISTINE : obtaining the actual data, modifying it

2.3 Feature Selection & Engineering

TODO: CHRISTINE : GC content, sliding window
TODO: TIANQI : discuss stuff like vienna/RNA fold

2.4 Machine Learning Algorithm

TODO: CHRISTINE : Brief intro of using two learning strategies

2.4.1 SVM - TIANQI feel free to modify the section title

TODO: TIANQI - talk about why you picked SVM, what SVM is, and some settings you may be using and why.

2.4.2 Random Forest - KEVIN feel free to modify the section title

TODO: KEVIN - talk about why you picked SVM, what SVM is, and some settings you may be using and why.

3 Results

3.1 SVM Results

TODO: TIANQI - discuss the results, and provide the table of performances where columns are input data type, and rows are various parameters used. Please use some form of heat map for the resulting table. I recommend matlab. Kevin & Tianqi - talk to each other so that the two tables look simliar (colorscheme, format, etc.) for uniformity

3.2 Random Forest Results

TODO: KEVIN - discuss the results, and provide the table of performances where columns are input data type, and rows are various parameters used. Please use some form of heat map for the resulting table. I recommend matlab. Kevin & Tianqi - talk to each other so that the two tables look simliar (colorscheme, format, etc.) for uniformity

4 Conclusion

TODO: CHRISTINE - done after all other sections are filled out

References

[1] AUTHOR1, AUTHOR2, et al. *TITLE* http://LINK, DATE OF PUBLICATION

[2] Schraga Schwartz and Douglas A. Bernstein and Maxwell R. Mumbach and Marko Jovanovic and Rebecca H. Herbst and Brian X. Leon-Ricardo and Jesse M. Engreitz and Mitchell Guttman and Rahul Satija and Eric S. Lander and Gerald Fink and Aviv Regev *Transcriptome-wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of ncRNA and mRNA* http://doi.org/10.1016/j.cell.2014.08.028, 2014