CARNEGIE MELLON UNIVERSITY - COMPUTATIONAL BIOLOGY DEPARTMENT

April 5, 2017

Team 1 - Design Document

Christine Baek
christib@andrew.cmu.edu

Kevin Chon
khchon@andrew.cmu.edu

Deepank Korandla
dkorandl@andrew.cmu.edu

Tianqi Tang
tianqit1@andrew.cmu.edu

1 INTRODUCTION

In this project, we seek to build a model by applying machine learning algorithms, which takes in an RNA sequences and outputs the probability of each base or position as a site of either m1a or pseudouracil modification.

2 METHODS

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.

2.1 DOMAIN KNOWLEDGE

2.2 DATA

We plan to utilize known positive examples of m1a and pseudouracil for testing and building model. In addition to this, additional relevant informations ("metadata") for such sequences will be used as additional features in the learning process.

2.2.1 Data Definition

- Sequence : referes to the sequencing results/reads
- Metadata : Including, but not limited to :
  - GC content (of the host organism)
  - GC content (of the read)
  - RT-stop frequency from RNA-seq
  - ,

2.2.2 Data Collection

2.2.3 Data Processing

2.3 Learning

2.4 Feature Selection

2.5 Machine Learning Algorithm

3 Milestones

3.1

References