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. 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

Most important factor here is what is known about *how* m1a and pseudouracil arises (ex: certain combination of sequences results in specific editing), and what impact they have on the reads. More information on how these post-transcription modifications came to be is greatly helpful in processing our data, and feature engineering. Each base is not independent of one another - adjacent bases, as well as far away (3D structure) may have impact on the function and potential editing of another base. It is known that at least for RNA editing (during-transcription RNA modification), there is bias for certain types of substitution. For simplicity, we will initially build our model by considering only direct neighbor bases for each position, but potentially expanding and taking further neighbors into consideration.

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 : refers to the sequencing results/reads
- *metadata* : Non-sequence data, including, but not limited to :
  - GC content (of the host organism)
  - GC content (of the read)
  - RT-stop frequency from RNA-seq
  - sequence motif

2.2.2 Data Collection

1. choose few species that are well-represented in terms of labeled m1a/pseudouracil RNA sequencing reads based on tRNAdb and MODOMICS databases
2. download population SNP data for chosen species

2.2.3 Data Processing

Perform multiple sequence alignment on the collected RNA sequence reads with various SNPs

2.3 Learning

2.4 Feature Selection & Engineering

Given the multiple sequence alignment of RNA reads of closely related species, features can be built by outputting various subsequence of each base and its neighboring base as it is not yet known how much and which neighboring bases have an impact on post-transcription RNA modification, if any. This will be used in addition to the various *metadata*, as defined earlier.

2.5 Machine Learning Algorithm

Given that most of our features are discrete rather than continuous, decisions tree/random forest would be a fast and potentially robust algorithm to apply to this problem. Another potential algorithm to explore is logistic regression. Use cross-validation for measuring performance of algorithm. Utilize python/numpy/scikit-learn for learning once appropriate algorithm is decided on.

3 Milestones

3.1 Week 1 - April 5th, 2017

- Design a overview of the project
- Define how and which data to collect
- Decide on learning algorithms

3.2 Week 2 - April 12th, 2017

- Collect data, process, feature engineer and build model with *one* chosen species, with few specific genes
- Continue collecting data for other species as well
- Experiment with various approaches to feature selection/engineering and determine which ones are most useful for learning

3.3 Week 3 - April 19th, 2017

Week of mini demo

- Demo using one or two species
- Finish processing/engineering and model building with the remainder of species

3.4 Week 4 - April 24th, 2017

Deliver solution and manuals

- Attempt to combine the independently built models, for a more generalizable predictor of RNA modification
- Clean up code and documentation

References