Classification of microRNA Using Deep-learning Method

Project report of Computational Genomics Spring 2017

Han Den (Andrew Id: hdeng1)
Luyi Ma (Andrew Id: luyim)

Abstract

In recent years, deep learning methods, as a branch of machine learning, have become a powerful tool in many scientific fields. One of deep learning methods, convolutional neural network(CNN), which has an ability of extracting features of high-level abstraction from minimum preprocessing data, is the most widely used method in classification. In this project, we choose CNN to do microRNA classification and distinguish microRNA from non-microRNA. We use one-hot vector model to convert our input sequence into numeric matrices and use these matrices as input to CNN. At last we train as test our model using miRNA sequences from open-source database and the result shows that our model is efficient in doing miRNA classification.

Keywords: miRNA, Convolution Neural Network, Classification

1 Introduction

MicroRNA (abbreviated miRNA) is a small non-coding RNA molecule (containing about 22 nucleotides) found in plants, animals and some viruses, that functions in RNA silencing and post-transcriptional regulation of gene expression. While the majority of miRNAs are located within the cell, some miRNAs, commonly known as circulating miRNAs or extracellular miRNAs, have also been found in the extracellular environment, including various biological fluids and cell culture media. So it's of great value to distinguish miRNAs from the whole genome and make accurate classifications. [1]

In machine learning, a convolutional neural network (CNN, or ConvNet) is a type of feed-forward artificial neural network in which the connectivity pattern between its neurons is inspired by the organization of the animal visual cortex. Individual cortical neurons respond to stimuli in a restricted region of space known as the receptive field. The receptive fields of different neurons partially overlap such that they tile the visual field. The response of an individual neuron to stimuli within its receptive field can be approximated mathematically by a convolution operation. Convolutional networks were inspired by biological processes and are variations of multilayer perceptrons designed to use minimal amounts of preprocessing. They have wide applications in image and video recognition, recommender systems and natural language processing.[2][3]

The data we use to train and test our model is from miRBase. [4] MiRBase is the central online repository of miRNA nomenclature, sequence data, annotation and target pre- diction, which first appeared in Oct. 2002 Release 15 contains 14197 miRNA loci from 66 species. From version 5.0, miRBase began to classify miRNAs into different families. [5] We use miRNA sequences from this data base as our positive input, meaning that this sequence belongs to the miRNA family.

// TODO: fill this block

some possible topics

• miRNA structure

• miRNA database (introduction of this paper **miRFam: an effective automatic miRNA classification method based on n-grams and a multiclass SVM)**

• sequence recognizion and classification , some methods? CNN in sequence analysis

• Motivation of this project, next step is to classify miRNA sequence into subclasses...

2 Methods

Explain your computational approach. Describe your model and learning/inference methods in two different subsections. Define all variables and include self-sufficient equations.
// TODO: fill this block

- sequence vectorization
- CNN structure for this classification job.

3 Implementation details

Give enough detail so the results can be reproduced by someone familiar with the field. Include a description of data processing steps and how you selected constants and/or free parameters (if applicable). Include pseudocode if implementing a new algorithm.
// TODO: fill this block

3.1 Data Generation

3.2 Sequence Vectorization

3.3 Input Data Processing

3.4 Convolution Neural Networks

4 Results and Conclusions

Provide informative figures and legends, a summary of conclusions, limitations and future directions.
// TODO: fill this block

4.1 Training Record

4.2 Performance

4.3 Discussion and Conclusions

References

[1] hahahaha, haliluya