Bioinformation Lab03

Summary

object : get location of gene region from sequence using HMM

author : Soojeong Shim (2016025532)

submit : 2018.11.18

environment : mac os & python3.6

Algorithm & Theory

Markov model

- stochastic model used to model randomly changing system

HMM

- statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved states

Viterbi Algorithm

- dynamic programming algorithm for finding the most likely sequence of hidden states that results in a sequence of observed events

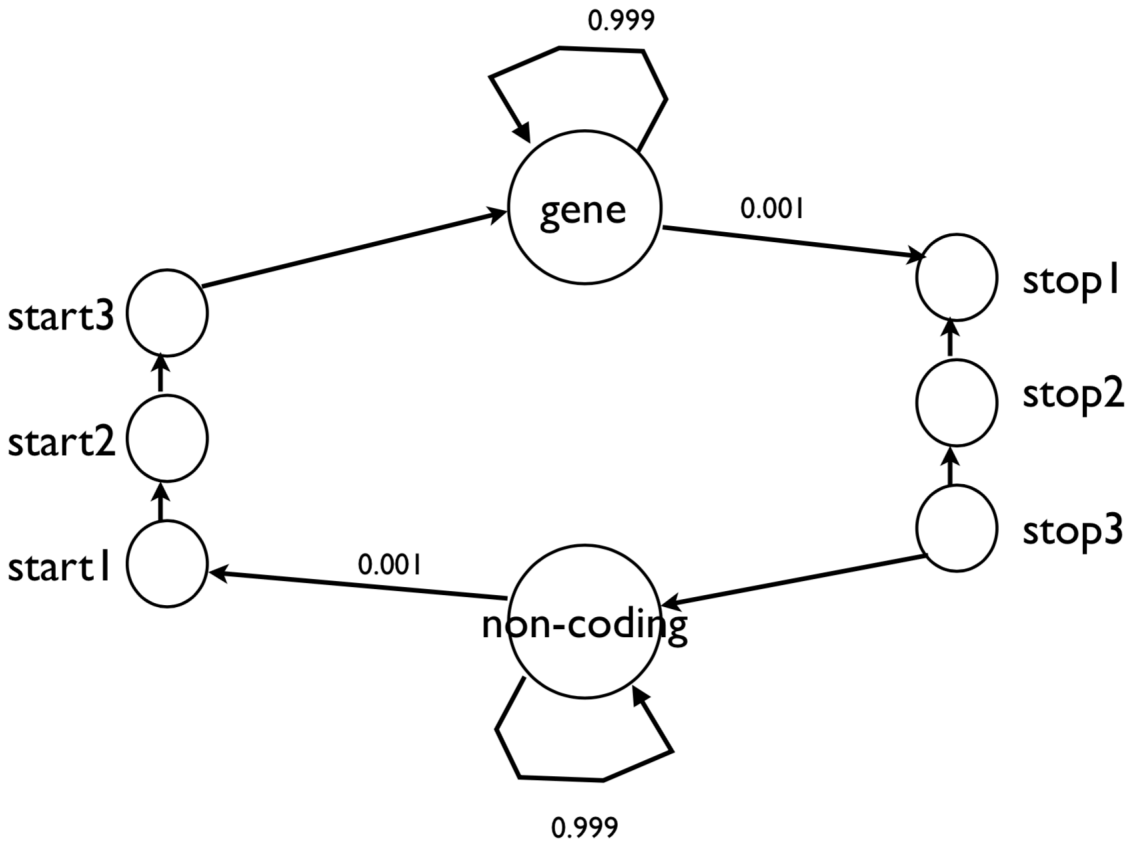
Confusion Matrix

- specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one
- some evaluation standards

- Accuracy : rate of accurate data of all data
- Recall : rate of truely expected positive data of al real positive data
- Precision : rate of truely expected positive data of all expected positive data

Model

transition probability and state



Result

Confusion Matrix

observed | predicted	Positive	Negative
Positive	959623	195339
Negative	143649	274741

- Accuracy = (TP + TN) / (P + N) = (959623 + 274741) / 1573352 = 0.7845440817
- Recall = TP / P = 959623 / (959623 + 195339) = 0.8308697602

- Precision = TP / (TP + FP) = 959623 / (959623 + 143649) = 0.8697972939

Analysis

- Training data is not given by training and only train emission probability, accuracy is not high.
- This program can find truely positive data well.