CALIFORNIA STATE SCIENCE FAIR
2013 PROJECT SUMMARY

Name(s)

Manogna Vemulapati

Project Number

S1427

Project Title

Identification of CpG Islands in a DNA Sequence Using a Hidden Markov Model Trained in MapReduce

Abstract

Objectives/Goals

The purpose of my project was to design and develop a software program to identify the locations of CpG islands in a DNA sequence. Hidden Markov Model (HMM) is the most popular statistical model used to locate CpG islands in DNA sequences. The accuracy of this model is limited by the size of the training data. The engineering goal of my project was to parallelize the training of the HMM so that (a) the training could be performed on larger data sets in reasonable time and (b) the resulting trained HMM could be used to more accurately identify the CpG islands.

Methods/Materials

I modeled the CpG island detection problem as a HMM by identifying the hidden states and emission symbols. There were 8 hidden states: A+,C+,G+,T+ corresponding to the bases within an island and A-,C-,G-,T- corresponding to bases outside an island. The observable emission symbols were A,C,G,T. A state such as C+ or C- emits the symbol C with a probability of 1 and similarly for other states. I used the MapReduce enabled version of Baum Welch algorithm from Apache Mahout project to perform unsupervised training on a sample contig from a human chromosome. I used the Amazon Elastic MapReduce platform to run the training. From the resulting trained HMM, I obtained the 8x8 state transition probability matrix. I used the Viterbi algorithm from HMM package in R to decode the hidden state sequence for given test contig from another human chromosome sequence. From the decoded hidden state sequence, I obtained the start and end positions of each CpG island in the test sequence.

Results

Unsupervised training on contig NT_028395 from human chromosome 22 (hg19) completed in about 2 hours on the MapReduce platform. From the trained HMM, I was able to identify about 80 percent of CpG islands for some test sequences from human chromosome 8. The same training using a non MapReduce version of Baum Welch algorithm did not complete even after few hours of running.

Conclusions/Discussion

I have observed that training the HMM on MapReduce cluster facilitates training on larger data sets in shorter time and also results in a more accurate identification of CpG islands in DNA sequences.

Summary Statement

My project is about identifying CpG islands in a DNA sequence more accurately by training the HMM on large data sets.

Help Received

Father helped me with the setup and running jobs on Amazon MapReduce platform.