1 Brief Bio
My name is Xinlei Pan. I am a first year graduate student in bioengineering department. I’m interested in machine learning and its application in biological data analysis. Typically examples include gene expression data analysis to construct gene regulatory network, EEG data analysis to study brain function, etc. For the machine learning part, I have a particular interest in probabilistic graphical model and its application in computational genomics.

My general motivation for taking this course is to help with my current research in computational genomics. In recent years, Bayesian networks has been applied in systems biology to uncover genetic regulatory relationships. In this network, the nodes are modeled as genes and the associated variable is gene expression level measured under different experimental conditions. The question is learning the most probable graphical representation of the dependencies among those genes. This can be summarized as a structure learning problem. Exact structure learning of bayesian networks was shown to be NP-complete [1]. I’m interested in seeing how parallel computing can help with this problem.

2 Machine Learning and Parallel Computing
Bayesian network is a probabilistic graphical model which describes the causal relationship via directed acyclic graphs (DAG). The learning of Bayesian network structure from experimental data has been proven to be NP-complete [1]. Exact searching of optimal graph structure that best matches the input data entails searching on a structure space of all possible DAGs over a given set of variables, which not only requires long run-time but also large memory. Thus, heuristic approaches are being developed. Most heuristic approaches for Bayesian network structure learning can be characterized as either constraint-based or optimal score based, or a hybrid of the two. Constraint-based methods typically search on a limited space of graphs such that all constraints between variables are satisfied. Optimal score based method evaluates how well a proposed graph predicts dependencies in the input data set; and at each step the addition, deletion or reversal of edges are under the constraint of acyclicity as well as required to improve the graph score. Due to the large scale intrinsics of gene regulatory network, searching for the optimal structure is time consum-
ing. Therefore, parallel computing methods for structure learning are being developed these days.

One of this kind of research has been implemented using C++ and MPI[3], and to assess the performance and applicability of the algorithms, experiments were performed on a Cray system with two 12-core AMD 'ManyCores' 2.1-GHz processors and 32 GB DDR3 1333-MHz memory per node [3].

Another paper [2] implements their bayesian structure learning algorithm (score based algorithm) on a CPU-GPU combined platform, where CPU serves as the host and assigns tasks to and collect results from the GPU. The figure below shows the architecture of the GPU being used. This GPU contains several blocks connected in the form of a grid. Each block usually includes 256 threads. Each thread has a number of registers and a local memory. All threads have access to shared memory of that block as well as the global memory of the GPU. Experiments were run on GPP and GPU respectively and the GPU implementation achieves a significant speedup over the GPP implementation.

Figure 1: The Architecture of GPU [2].

In addition, MapReduce, a functional programming model suited for parallel computation also demonstrated its advantage over traditional non-parallel computing algorithms [4]. MapReduce has been proved to have better performance over large dataset’s Bayesian network structure learning problem. However, when the dataset is small, traditional centralized algorithm needs less running time than distributed algorithm with MapReduce since the start up and tear down costs are primary performance bottlenecks of MapReduce.
References


