Graphics processing unit implementations of relative expression analysis algorithms enable dramatic computational speedup

Andrew T. Magis1, John C. Earls2, Youn-Hee Ko2, James A. Eddy3

1Center for Biophysics and Computational Biology, 2Department of Computer Science, 3Department of Bioengineering and 4Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, IL 61801, USA

ABSTRACT

Summary: The top-scoring pair (TSP) and top-scoring triplet (TST) algorithms are powerful methods for classification from expression data, but analysis of all combinations across thousands of human transcriptome samples is computationally intensive, and has not yet been achieved for TST. Implementation of these algorithms for the graphics processing unit results in dramatic speedup of two orders of magnitude, greatly increasing the searchable combinations and accelerating the pace of discovery.

Availability: http://www.igb.illinois.edu/labs/price/downloads/

Contact: ndprice@illinois.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on October 26, 2010; revised on December 19, 2010; accepted on January 15, 2011

1 INTRODUCTION

Rapidly improving technologies have made large amounts of gene expression data available for analysis and classification. The NCBI Gene Expression Omnibus (GEO) database contains hundreds of thousands of samples representing a wide range of diseases and healthy tissue for which gene expression has been measured. As next-generation RNA sequencing technology (Cloonan et al., 2008) becomes ubiquitous, GEO and other databases will further increase in size and accuracy of information. Researchers have sought to use this expression data to identify distinct gene relationships that classify disease states, allowing for accurate diagnosis of diseases given the expression patterns of a few genes. Such methods include support vector machines (Brown et al., 2000), decision trees (Zhang et al., 2003) and neural networks (Khan et al., 2001). The top-scoring pair (TSP) algorithm and its variants have demonstrated similar accuracies to these methods while remaining relatively simple, resistant to overfitting and consistent across data normalization methods (Geman et al., 2004; Lin et al., 2009; Price et al., 2007; Tan et al., 2005). Classifiers identified using these algorithms have been used to predict cancer outcomes and model disease progression (Eddy et al., 2010). Despite these advantages, the TSP and especially the top-scoring triplet (TST) algorithm are computationally intensive and therefore slow. Because increasing the accuracy of predictions may require analysis of thousands of samples across tens of thousands of transcripts, it is important to improve the speed of these algorithms. Faster algorithms also enable more comparisons to be made, including automated data mining across many sample sets.

While primarily known for gaming applications, the graphics processing unit (GPU) is increasingly applied to computationally challenging scientific problems including molecular dynamics simulations (Stone et al., 2010), quantum chemistry (Ufimtsev and Martinez, 2008), and medical imaging (Stone et al., 2008). The GPU is designed for massive parallelism involving thousands of simultaneous executing threads, but requires different coding than that which runs on CPUs. Algorithms well suited for such parallelism can run tens to hundreds of times faster on GPUs than a corresponding CPU implementation. GPUs are also now widely available to researchers via National Center for Supercomputing Applications (NCSA) clusters and businesses such as Amazon Web Services EC2 Cloud Computing. Here, we present implementations of the TSP algorithm and the TST algorithm on the GPU. As the TST algorithm is particularly computationally demanding, this GPU implementation enables the first comprehensive search of all possible TSTs for classification.

2 ALGORITHMS

Given two classes of samples \( C = \{C_1, C_2\} \) with expression values for \( N \) genes \( \{x_1, \ldots, x_N\} \), the TSP algorithm identifies the marker gene pair \( (x_i, x_j) \) in which the TSP score

\[
\Delta_{ij} = |P(x_i < x_j | C = C_1) - P(x_j < x_i | C = C_2)|, \quad i \neq j
\]

is maximized. The algorithm performs all pairwise comparisons between the genes of the dataset and calculates the TSP score for each, then selects the maximum or set of maximum scores. Multiple pairs may then be combined to improve classification (Tan et al., 2005). The TSP algorithm exhibits \( O(N^2) \) time complexity.

The TST algorithm extends the TSP algorithm to triplets of genes (Lin et al., 2009). The TST algorithm is calculated similarly to TSP: all possible triplets of gene ranks are compared, and the probabilities of each permutation within each triplet are calculated and scored. The algorithm finds the maximum TST score from all triplets in \( O(N^3) \) time complexity. Extended details of both algorithms can be found in the Supplementary Material.

3 METHODS

Modern (compute 1.2 and higher) NVIDIA CUDA GPU architectures contain multiple (2–30) streaming multiprocessors (SM), each consisting...
The algorithms described above have been implemented in C with RXA enables the discovery of accurate marker gene pairs and triplets that algorithm on this dataset was completed in <9 h. Using the GPU take over 6.5 months, while the GPU implementation of the TST GTX480 executes the algorithms 228X to 455X faster. Processing 255X faster than the corresponding CPU implementations, and the computation times. The Tesla T10 executes the algorithms 77X to 83X (Tesla T10) and 301X (GTX480) (right). The dataset used had 350 samples in Class 1 and 306 samples in Class 2. All data points are the mean of three independent runs of the software. Dotted lines indicate extrapolation from previous data points using a quadratic (TSP) or cubic (TST) polynomial fit.

are infeasible with the GPU implementations, while also allowing more stringent error estimation methods than are currently possible due to previous computational time constraints.

ACKNOWLEDGEMENTS

The authors acknowledge the National Center for Supercomputing Applications (NCSA) AC cluster for CPU and GPU speed tests.

Funding: National Institutes of Health Howard Temin Pathway to Independence Award in Cancer Research; the Grand Duchy of Luxembourg; the Roy J. Carver Charitable Trust.

Conflict of Interest: none declared.

REFERENCES


Lin,X. et al. (2009) The ordering of expression among a few genes can provide simple cancer biomarkers and signal BRCA1 mutations. BMC Bioinformatics, 10, 256.


