Gene expression

The gputools package enables GPU computing in R

Joshua Buckner1,*, Justin Wilson1, Mark Seligman3, Brian Athey2, Stanley Watson1 and Fan Meng1,2

1Department of Psychiatry and Molecular and Behavioral Neuroscience Institute, 2National Center for Integrative Biomedical Informatics, University of Michigan, Ann Arbor, MI 48109 and 3Rapid Biologics, P.O. Box 31989, Seattle, WA 98103, USA

Received on July 8, 2009; revised on October 7, 2009; accepted on October 17, 2009

Advance Access publication October 22, 2009

Associate Editor: John Quackenbush

1 INTRODUCTION

Microarray gene expression analysis often involves CPU intensive computation requiring days or even months on mainstream desktop computers. Examples are the Granger test for detecting causal relationships in gene expression time course data (Kaminski et al., 2001), pairwise similarity calculations involving permutations, and machine learning-based approaches for microarray sample classification.

R is the most popular statistical environment for microarray data analysis. By default, R does not perform calculations in parallel. Modern graphics processing units (GPUs) have made parallel processing on a single desktop computer practical. Using a GPU, parallel processing may be applied to data from high-throughput biological experiments to increase analysis efficiency (Mateos-Duran et al., 2009). To facilitate the application of parallel processing using a GPU to such data, we developed the gputools package from Nvidia, the CUDA toolkit from Nvidia, we have implemented several functions commonly used in microarray gene expression analysis for GPU-equipped computers.

2 DESCRIPTION OF THE PACKAGE

Table 1 contains a summary of the primary and secondary functions of the package. The primary functions include the functions gpUCor, gpuGranger, gpuHclust, gpuSvmTrain and gpuMi. For example, gpUCor calculates Pearson and Kendall correlation coefficients and may be used as a substitute for R’s cor function for large datasets. The function gpuGranger applies the Granger causality test (Kaminski et al., 2001) to time sequences of variables and is comparable with the granger.test function from the MSBVAR R package. Refer to Table 1 for descriptions of the rest of the primary functions.

The secondary functions include functions that extend or assist primary functions: gpuDistClus, gpuSvmPredict, getAucEstimate (which is not GPU enabled), gpUTest, chooseGpu and getGpuId.

The secondary functions also include functions for GPU assisted matrix algebra: gpuMatMult, gpUCor and gpUSolve. The gpUCor and gpUSolve functions are relatively inefficient. The two functions have been retained in the package in the hope that others may find the source code instructive.

3 RESULTS

We tested GPU enabled functions against non-GPU enabled versions with biomedical data on a desktop computer. The desktop computer has an Intel Core i7 920 processor and an Nvidia GeForce GTX 295 GPU card. The desktop computer’s operating system is the CentOS 5.3 Linux distribution.

The primary functions of the gputools package were tested against corresponding non-GPU enabled R solutions (Table 2). We chose to use a single thread of the R environment in our test, since this is the way that most users interact with R. The function gpuMi was tested against the same algorithm coded in the C programming language with an R interface and without any parallelism.

The gputools package was tested using a dataset from a P19 cell line differentiation gene expression profiling study from Dr Michael Uhler’s laboratory of the University of Michigan (unpublished data) with 12 time points using the Illumina MouseRef-8 BeadChip. Each trial of the gputools function used a subset of the 28 843 probes in the chip. The other functions were tested with subsets of GSE6306 (Li et al., 2006).
Table 1. Functions of the gputools R package

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>gpuGranger</td>
<td>Granger-causality (Kaminski et al., 2001)</td>
</tr>
<tr>
<td>gpuHclust</td>
<td>Hierarchical clustering</td>
</tr>
<tr>
<td>gpuCor</td>
<td>Pearson and Kendall correlation coefficients</td>
</tr>
<tr>
<td>gpuSvmTrain</td>
<td>Support vector machine training (Carpenter, 2009)</td>
</tr>
<tr>
<td>gpuDist</td>
<td>Distance between vectors</td>
</tr>
<tr>
<td>gpuDistClust</td>
<td>Correlation coefficients from vector data</td>
</tr>
<tr>
<td>getAncEstimate</td>
<td>Estimate the quality of an svm (Hand and Till, 2001)</td>
</tr>
<tr>
<td>gpuTest</td>
<td>Perform Student’s t-test</td>
</tr>
<tr>
<td>gpuMatMult</td>
<td>Matrix multiplication</td>
</tr>
<tr>
<td>gpuQr</td>
<td>QR decomposition (Bjorck, 1996)</td>
</tr>
<tr>
<td>gpuSolve</td>
<td>Solve linear systems and find matrix inverses</td>
</tr>
<tr>
<td>chooseGpu</td>
<td>Select a GPU in case the desktop has more than one</td>
</tr>
<tr>
<td>getCpu</td>
<td>Get the number of the currently selected GPU</td>
</tr>
</tbody>
</table>

Table 2. Primary functions and their pre-existing R solutions

<table>
<thead>
<tr>
<th>Function</th>
<th>Pre-existing R counterpart</th>
</tr>
</thead>
<tbody>
<tr>
<td>gpuGranger</td>
<td>granger.test of package MSBVAR</td>
</tr>
<tr>
<td>gpuHclust</td>
<td>hclust</td>
</tr>
<tr>
<td>gpuCor</td>
<td>cor</td>
</tr>
<tr>
<td>gpuSvmTrain</td>
<td>svm of package e1071</td>
</tr>
<tr>
<td>gpuMi</td>
<td>none known</td>
</tr>
</tbody>
</table>

Table 3. Test data sizes

<table>
<thead>
<tr>
<th>Test Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The performance test results are shown in Figure 1. For gpuCor, the test number in Table 3 and Figure 1 indicates the number of components of each of 10,000 vectors. For the rest of the primary functions, the test number indicates the number of variables. In the gpuGranger test, lag was set to 2 and each variable had 12 samples (Kaminski et al., 2001). In the gpuCor test, 10,000 vectors were used for Kendall correlation for each trial. In the gpuMi test, each variables had 100 samples, there were 10 bins of equal width, and spline order was set to 3 (Daub et al., 2004). In the gpuSvmTrain test, each variable had 500 features (Carpenter, 2009).

4 CONCLUSION

Our approach significantly speeds up several algorithms frequently used in gene expression data analysis. Since the code for the gputools package is open source and distributed with the package, researchers may customize it to suit individual needs. We welcome collaboration with the research community and contributions to the package from other researchers and developers.

Funding: J. Buckner, J. Wilson, and F. Meng are members of the Pritzker Neuropsychiatric Disorders Research Consortium, which is supported by the Pritzker Neuropsychiatric Disorders Research Fund L.L.C. This work is also partly supported by the National Center for Integrated Biomedical Informatics through National Institutes of Health grant 1U54DA021519-01A1 to the University of Michigan.

Conflict of Interest: none declared.

REFERENCES