NIEHS
National Institute of Environmental Health Sciences

Bioinformatics E-Bulletin

a semi-annual communication of cutting-edge bioinformatics/computational biology support entities, scientific software, collaborative research and trends in the field

Jun 03, 2013 Volume 5, Issue 1

In This Issue

Useful hyperlinks

NIEHS Genome Browser

NIEHS Galaxy Server

Bioinformatics support experts

Send comments to:

Editor: Pierre R. Bushel

bushel@niehs.nih.gov

Leveraging the RGPU for BigData analysis

Contributed by Mr. Jianying Li

Towards the end of the last century, a new term GPU (graphics processing unit) became popular owing to Nvidia's development of "the world's first GPU" marketed as GeForce 256. It was initially designed to rapidly accelerate the creation of images via a specialized electronic circuit. In less than 15 years since the official debut of such a new design, the GPU has not only seen its dramatic growth in its use to boost 3D imaging, but it is also widely used in the data analysis community. With the fast growing "BigData" concept and increasing needs for computational power, the GPU is playing a pivotal role in providing unique benefits for running large analytical jobs with the least amount of code re-writing.

Using the K-means clustering algorithm as an example, it was reported that the GPU-accelerated version increased processing speed 200x-400x faster than a conventional desktop and 6x-12x faster than a high-end 8 core CPU workstation (www.azintablog.com). As for other popular data mining algorithms, which are computationally intensive processes (hidden Markov models, support vector machines, and Bayesian mixture models for example), all of these jobs can be sped up significantly if they are run on GPU cores. A clustering job using a GTX280 GPU with 240 cores that takes 26 minutes to run could take as long as 6 days on a single-core CPU. One can expect a faster speed with the latest Femi GPUs of 480 cores, which should be released in the near future.

In the emerging BigData era, a statistics programming language – R spearheaded among its counterparts and became one of the main trends for data science analytical processing in many fields, i.e. financial market, telecommunication, digital TV streaming and internet browsing. Owing to its unique advantage, R is poised to serve the scientific community in the decades to come. For example, in 2009, a group of pioneers helped to spinoff an entity out of Rosetta Inpharmatics and founded Sage Bionetwork. The goal was to transform biomedical research approaches into practical healthcare discoveries in order to speed up biomedical practices. The route to achieve the goal was to foster open community development for analytical protocols, outcome prediction models and deliverable components.
To leverage the ability that GPUs provide, an R/GPU project was developed at the Netherland Bioinformatics Centre (https://trac.nbic.nl/rgpu/) to produce a package to run R processing "transparent" to an NVIDA GPU via a CUDA core. Chi Yan (www.r-tutor.com) also published an R package (rpud), and version 2 was just released in February 2013. In the documentation, it details a benchmark comparison between the use of a CPU and a GPU on computing the Euclidean distance of a 120 x 4500 dimensional matrix. The CPU usage time was demonstrated on a desktop computer with an AMD Phenom II X4

CPU in the following block:

```
> test.data <- function(dim, num, seed=17) {
+     set.seed(seed)
+     matrix(rnorm(dim * num), nrow=num)
+ }
> m <- test.data(120, 4500)
> system.time(dist(m))
   user  system elapsed
 14.343   0.185  14.533
```

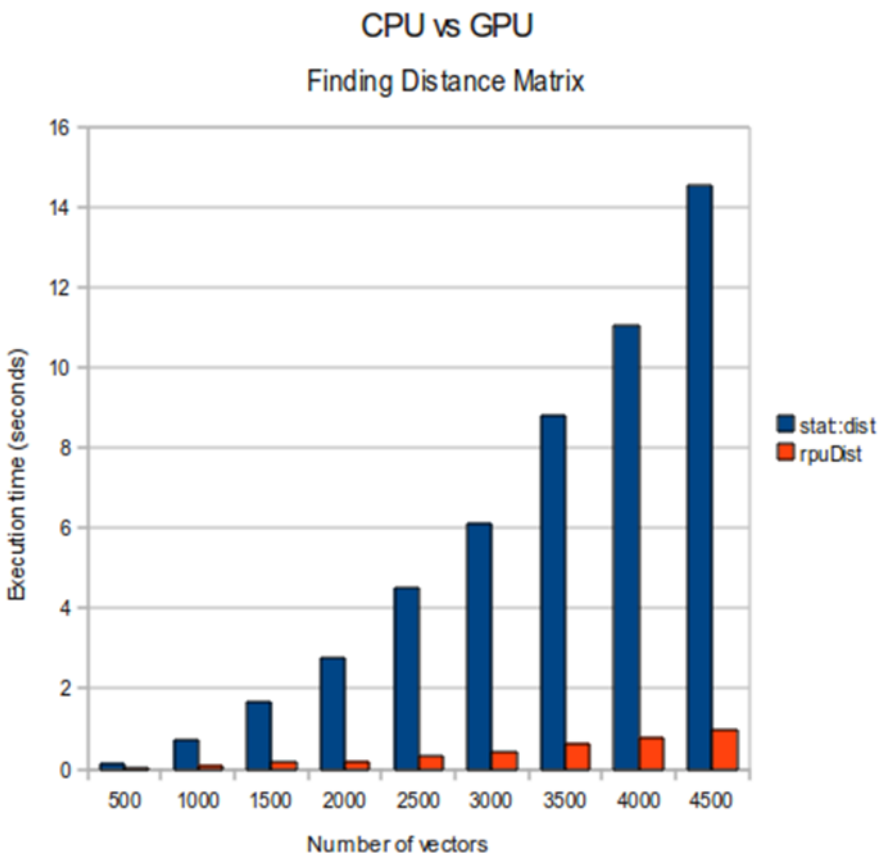
It could be longer (~37 seconds) even on a server with an AMD Opteron(tm) Processor 6168, 1.9 GHz

```
> system.time(dist(m))
   user  system elapsed
 37.140   0.223  37.370
```

With the rpud package and computing the same distance matrix using the rpuDist method (running on NVIDIA GTX 460 GPU), the execution time is about 1 second. Even better, with the rpudplus add-on, one can compute distance matrices for larger data sets that fit inside the system RAM available to R.

```
> library(rpud)                 # load rpud with rpudplus
> system.time(rpuDist(m))
   user  system elapsed
  0.674   0.305   0.980
```

At Yau's r-tutoring sites, he displayed a chart that compares the performance between the CPU and GPU on a series of increasing number of vectors



At NIEHS, the scientific computing team has equipped a few servers with GPU capabilities. They have been configured to support GPU for scientific computing jobs and will be configured to test the RGPU libraries. If all goes well, the RGPU framework will permit a flexible and efficient environment for intensive computational analysis of large data sets.

Bioinformatics By the Way

NanoMiner: Integrative Human Transcriptomics Data Resource for Nanoparticle

The NanoMiner webresource (http://nanominer.cs.tut.fi/) contains 404 human transcriptome samples exposed to various types of nanoparticles. All the samples in NanoMiner have been annotated, preprocessed and normalized using standard methods that ensure the quality of the data analyses and enable the users to utilize the database systematically across the different experimental setups and platforms. Contact Reija Autio at the Tampere University of Technology for access (reija.autio@tut.fi).

Center for Stem Cell Bioinformatics

The Center for Stem Cell Bioinformatics provides analytical support and a common environment for storage, sharing and analysis of stem cell research data for the Harvard Stem Cell Institute (HSCI). It promotes computational

integration and a community of sharing and discovery in stem cell biology. The Stem Cell Discovery Engine provides integrated access to tissue and cancer stem cell experimental information and molecular profiling analysis tools.

http://stemcellcommons.hsci.harvard.edu/CSCB/
http://discovery.hsci.harvard.edu/

Division of Bioinformatics and Biostatistics (DBB) at NCTR/FDA

The Division of Bioinformatics and Biostatistics (DBB) at the Food and Drug Administration (FDA) National Center for Toxicological Research (NCTR) was established in May 2012. It combines bioinformatics, biostatistics and scientific computing activities at NCTR to capitalize on the synergies of these diverse talents for enhanced research and support of the FDA missions. The Division develops integrated bioinformatics and biostatistics capabilities to address increasing demands in biomarker development, drug safety, drug repositioning, personalized medicine, and risk assessment. Its capability is directed towards integration with FDA business processes to ensure that NCTR linkages with FDA Product Centers are strengthened, and that NCTR informatics capabilities continue to become more diverse, robust, and capable of meeting future requirements of the FDA. The Division is directed by Dr. Weida Tong (weida.tong@fda.hhs.gov).

http://www.fda.gov/AboutFDA/CentersOffices/OC/OfficeofScientificandMedicalPrograms/NCTR/WhatWeDo/ResearchDivisions/ucm305786.htm

Nature Journals Checklist

Nature journals are making efforts to improve the reporting and reproducibility of published results. As of May 2013, Nature journals require authors of life sciences research papers that are sent for external review to include in their manuscripts relevant details about several elements of experimental and analytical design. This initiative aims to improve the transparency of reporting and the reproducibility of published results. A section of the checklist pertains to statistics and general methods addressing aspects such as power of detection, sample size estimation, outlier detection, randomization and statistical tests employed.

http://www.nature.com/nbt/journal/v31/n5/full/nbt.2588.html

http://www.nature.com/authors/policies/checklist.pdf

Student interns in bioinformatics and biostatistics

Summers of Discovery students

Michael Falk – Undergraduate at Georgia Institute of Technology majoring in materials sciences and engineering. Co-mentors are Dr. Pierre Bushel and Dr. David Fargo

Jenny Sun – Undergraduate at the University of North Carolina, Chapel Hill majoring in biostatistics. Mentor is Dr. Clare Weinberg

NIEHS Scholars Connect Program

Mia Burks – Undergraduate at Saint Augustine's University majoring in public health and minoring in biology. Mentor is Dr. Pierre Bushel

Summer 2013 Introduction to Biostatistics and Bioinformatics Short Courses

Course Number	NIEHS Staff Registration Date	Non-NIEHS Staff Registration Date
1 and 2	May 23, 2013	May 30, 2013
3 and 4	May 30, 2013	June 6, 2013
5 and 6	June 6, 2013	June 13, 2013
6 and 7	June 26, 2013	July 3, 2013
8 and 9	July 3, 2013	July 10, 2013

NOTE:

Course #7 will be held in 101/A262 beginning at 1PM.
Course #9 will be held in 101/A262 and is an all-day course from 9am-5pm.
All other courses will be held in NIEHS Rall 101A beginning at 1PM.
Registration for NIEHS staff will open approximately 2 weeks before the course date.

If space remains available at 1 week before the course registration will open to non-NIEHS staff. Non-NIEHS staff are required to present a valid government issued ID when entering our campus. You will be contacted about the process as part of your registration confirmation.

Course handouts (PowerPoint slides, sample data, etc.) is typically available as part of each course description a few days before the presentation date.

Please contact Bill (qb) Quattlebaum (quattleb@niehs.nih.gov) if you have questions.

See the following web site for information about registering and for course materials:
http://junction.niehs.nih.gov/divisions/dir/resources/analysis/ib/courses/index.htm

1. **Wednesday, 5-June, 2013 – 1:00 PM – 2:30 PM AND Thursday, 6-June - 1:00 PM - 2:30 PM**

   Introduction to statistics and experimental design

   Presented by Grace Kissling

   NOTE: This is a 2-part course. You will be register for and plan to attend both parts of this course.

Abstract: This class provides an introduction to statistics and experimental design. Topics include experimental design, levels of measurement, numerical and graphical summarization of data, and sample size determination. The principles of statistical inference are also discussed, including confidence intervals and hypothesis testing. Examples are used throughout to illustrate the utility of many of the methods

2. **Monday, 10-June, 2013 - 1:00 PM – 2:30 PM**

Testing statistical hypotheses

Presented by Min Shi

Abstract: An approximate list of topics: Formulation of null and alternative hypotheses. Power and sample size calculations. Some common parametric and nonparametric tests. Resampling based methods: Permutation tests, bootstrap methodology. Basic principles of multiple hypothesis testing:

1. Two common notions of false positive rates - family wise error rate (FWER) and false discovery rate (FDR).
2. Common methods for controlling FWER and FDR.

3. **Friday, 14-June, 2013 - 1:00 PM – 3:30 PM**

Introduction to the R programming and graphing environment for advanced Excel users

Presented by Les Klimczak

Abstract: This course is targeted toward biologists who have achieved high proficiency in Excel and are facing its limitations with regard to automatic processing and analysis of large-scale datasets. The R programming language will be introduced as a more powerful alternative with emphasis on explaining the distinctions of operating in a command-driven environment. Basic data processing and analysis steps in R will be presented in comparison with their equivalents in Excel. Programmatic generation of graphs in R will be introduced as well.

4. **Tuesday, 18-June, 2013 - 1:00 PM – 2:30 PM**

DNA Microarray Analysis

Presented by Keith Shockley

Abstract: This introduction to DNA microarray data analysis will begin with an overview of microarray technology and the quality control and normalization of microarray data. Data preprocessing, experimental design and gene expression clustering will be discussed. Analysis of variance (ANOVA) will be introduced to compare two or more treatment groups. The class will incorporate numerous examples and include references to widely available commercial and open-source software.

5. **Friday, 21-June, 2013 - 1:00 PM – 2:30 PM**

Introduction to NGS tech and ChIP-seq

Presented by James Ward

Abstract: We will start by looking at the evolution of sequencing technology and briefly go over the chemistry and principles behind first, second, and third generation sequencers. We will then delve into the methods and applications for ChIP-seq analysis. ChIP-Seq stands for Chromatin Immunoprecipitation followed by Sequencing, and as the name suggests, uses immunoprecipitation to assess the genomic loci of DNA binding proteins such as transcription factors, histone modifications, CTCF etc. This section will include an overview of the ChIP-seq protocol, experimental requirements,

statistical analyses, data formats and visualization. We will end with a demonstration of ChIP-seq data analysis, starting with raw data from the sequencer to the visualization of “Chipped” peaks in the UCSC genome browser. Please note that due to time limitations, we will not be able to perform a real-time demonstration and will instead rely on a recently analyzed example. Bringing a laptop is not necessary.

6. **Wednesday, 10-July, 2013 - 1:00 PM – 3:00 PM**

NGS QC and RNA-seq

Presented by Sara Grimm

Abstract: This module will cover QC of NGS data and analysis of RNAseq data. Some basic aspects of NGS quality control of raw sequencing data will be discussed, focusing on common problems to look for and how to implement work-arounds when possible. A general background on the uses and analysis strategies for RNAseq will be discussed, and the most popularly used tools for mapping, quantification, and visualization of RNAseq data will be described. A simple experimental example will be used to demonstrate these tools & their output.

7. **Friday, 12-July, 2013- 1:00 PM – 2:30 PM**

Pathways analysis

Presented by Jianying Li

NOTE: in room 101/A262 AND limited to 14 registrants. Laptops will be provided.

Abstract: This short course will provide an overview of a few of the current knowledge-based pathway analysis approaches widely used in genomic research. It covers basic information on frequently used databases (i.e. gene ontology categories), as well as curated knowledge bases from scientific literature and the public domain. It also focuses on two commonly used statistical approaches (Hypergeometric/Fisher exact test and Kolmogorov Smirnov) in pathway analysis with a hint of theoretical illustration. In this course, we will introduce a few (publicly available/free access and license-based) applications which implement either of the two statistical approaches. In the end, we will touch on some questions about the pros and cons in the pathway analysis package(s) and explore ways to deal with such concerns. The goal is to provide the scientists at the institute with a solid understanding of the fundamental concepts behind commonly used analytical tools for pathway analysis in order to empower them to maximize the analysis of their data and enhance interpretation of their results.

8. **Wednesday, 17-July, 1:00pm – 3:00 PM**

Introduction to UNIX and Perl for Biologists

Presented by Adam Burkholder

Abstract: This course is intended as an introduction to the UNIX computing platform for biologists with no prior experience. It will include an overview of the scientific computing resources available at NIEHS, and cover the most commonly used and most useful built-in UNIX commands. Additionally, the Perl programming language will be introduced in the form of easy-to-use "one-liners". This course will consist primarily of live demonstrations using Mac OS X, so attendees with access to a Macintosh laptop are encouraged to bring one.

9. **Thursday, 18-July, 2013- 9:00 AM - 5:00 PM**

Galaxy

Presented by Dave Clements (Galaxy Project for Emory University)

http://wiki.galaxyproject.org/DaveClements

Hosted by Thomas Randall

NOTE: in room 101/A262 AND limited to 14 registrants. Laptops will be provided.

Abstract: Galaxy is a web based tool that allows users to input a variety of datasets locally available or from public sources, including genome wide tracks from the UCSC Browser and BioMart, and manipulate and analyze these datasets in a free and publicly available point and click environment. A variety of types of sequence/text manipulation tools are available. These include a variety of common Linux/Unix type functions and all of the EMBOSS tools. Its developing role is in enabling a variety of genomics level manipulations on custom large datasets, including tools for Next Generation Sequencing data (RNAseq, ChIPseq). There are also some statistical and phylogenetic tools available. Galaxy has a workflow generation capability so that users can set up custom pipelines with user-set parameters based on tools within Galaxy, save, share, and re-use them.