THE 'TTIME' PACKAGE: PERFORMANCE EVALUATION IN A CLUSTER COMPUTING ENVIRONMENT

***Marico Howe*[1]**, Daniel Berleant and Albert Everett

(University of Arkansas at Little Rock, Little Rock, AR, USA)

ABSTRACT: The objective of translating developmental event time across mammalian species is to gain an understanding of the timing of human developmental events based on known time of those events in animals. The potential benefits include improvements to diagnostic and intervention capabilities. The CRAN 'ttime' package provides the functionality to infer unknown event timings and investigate phylogenetic proximity utilizing hierarchical clustering of both known and predicted event timings. The original generic mammalian model included nine eutherian mammals: *Felis domestica* (cat), *Mustela putorius furo* (ferret), *Mesocricetus auratus* (hamster), *Macaca mulatta* (monkey), *Homo sapiens* (humans), *Mus musculus* (mouse), *Oryctolagus cuniculus* (rabbit), *Rattus norvegicus* (rat), and *Acomys cahirinus* (spiny mouse). However, the data for this model is expected to grow as more data about developmental events is identified and incorporated into the analysis. Performance evaluation of the 'ttime' package across a cluster computing environment versus a comparative analysis in a serial computing environment provides an important computational performance assessment. A theoretical analysis is the first stage of a process in which the second stage, if justified by the theoretical analysis, is to investigate an actual implementation of the 'ttime' package in a cluster computing environment and to understand the parallelization process that underlies implementation.

INTRODUCTION

The original intent of the statistical model was to create, on one hand, a scale of developmental events such that early events score low and later events score high, and on the other hand, an additional scale of species where fast-developing species score low and slow-developing species score high; therefore the two numbers, event and species scores, could be combined to infer the potentially unknown time of an event in a specified species.

Neural events associated with the existing data set are comprised of onsets, peaks, and offsets of neurogenesis related to structures, which include but are not limited to peaks of neuronal death and components of process maturation. A constant *k* found in the statistical model accounts for organizational events such as implantation, blastulation, and differentiation of the primitive germ layers that have been found to be consistent across investigated mammals. A value of 5.37 was found for constant *k,* as shown in Equation 1:

$$Y = \ln (day - 5.37) \quad (1)$$

where *Y* is comprised of the species score, the event score as well as the primate interaction, which equates to a log transform of postconceptional days; *day* is thus the postconceptional (PC) day. This multiple regression model fits a function of time such that the estimation of the day of any developmental event in any of the nine species can be algebraically determined. The model predicts species-event dates when empirical data may be missing. The statistical model also accounts for primate limbic events, which occur earlier by adjusting *k* for such events by

[1] Please send comments to the authors at mchowe@ualr.edu.

adding 0.248683 for primate cortical events and subtracting 0.079280 for primate limbic events (Clancy et al. 2007a, 2007b).

This research has evolved into a web-based tool with a user-friendly front-end interface for researchers and clinicians to utilize and submit new data points. The back end is a MySQL database with 102X10 potential data points. Here 10 represents the number of species while 102 represents the number of events. The user has the option to predict neurodevelopmental events or translate neurodevelopmental events across mammalian species. The translation is in postconception (PC) and postnatal (PN) time, where the first 24-hour period after conception is often denoted as PC1, whereas PN0 denotes the first 24-hour period after birth.

The predictions fall within the 90% confidence limits of the predictions. Error in the mathematical model could stem from interaction terms other than the primate cortical and limbic terms. Additionally, there is potential variability among different individuals of the same species in timing of events. Nonetheless, the model can be enhanced through continued data collection by interested parties in the scientific and industrial communities (Clancy et al. 2000, 2001).

MATERIALS AND METHODS

We conducted a baseline performance evaluation of the '*ttime'* package in two environments, one Windows-based and one Linux-based. This open source package uses R version 2.12.0 and the '*pvclust'* package (Nagarajan 2010). The Windows machine operates Windows Vista 32-Bit Edition with Service Pack 2. The system is a Compaq Presario CQ60 Notebook PC (Personal Computer) manufactured by Hewlett-Packard. Its processor is an Intel Pentium Dual CPU (Central Processing Unit) T3400 at 2.16 and 2.17 GHz (Gigahertz) with 2 GB (Gigabytes) of RAM (Random Access Memory).

Ares, a workstation, was also utilized for this baseline. It is a Dell PowerEdge 6950, running a 4 Dual-Core AMD Opteron Processor 8220 with 32 GB of RAM. It runs Cluster Rocks release 4.3 that includes CentOS version 5.5, 64-Bit Edition. Cluster Rocks is a freeware version of Linux.

The high performance computing (HPC) cluster environment at the University of Arkansas at Little Rock Computer Science Department was also employed. It consists of three types of nodes: a front end node, a login node, and compute nodes. The front end node, hpc1-cpsc.host.ualr.edu, contains the cluster setup and node definitions. It runs a torque resource manager and moab scheduler. It shares /home via NFS (Network File Server), and has approximately 3.3 terabytes (TB) available. The login node, hpc1-cpsc-login.host.ualr.edu, is where users login to compile programs and submit computing jobs. The /home is mounted from the front end node via NFS. The nodes compute 1-1 through compute 2-32 (2 racks, 32 nodes per rack). Each node has 2 CPU sockets, each with a quad core Intel Xeon 2.66 GHz CPU. Each node has 16 GB of RAM.

The HPC environment also utilizes three types of networks: a public network, a private 1000 Mb network, and a private Infiniband network. The public network is for front end and login node usage only. It has an eth1 interface, with the following Internet Protocol (IP) address: 144.167.99.0/24. The private network has a NFS and Message Passing Interface (MPI) ring assembly, accessed at IP address 192.168.1.0/24. The private Infiniband network is for MPI message passing. Its use is for higher speed and lower latency than a 1000 MB Ethernet. The nodes in the cluster communicate using the standard network protocol TCP/IP (Transmission Control Protocol and Internet Protocol) over high-speed 10 networks. The TCP/IP is accomplished via Internet Protocol Over Infiniband (IPoIB), which uses an ib0 interface with IP address 192.168.2.0/24, required by some applications. Native IB is faster than IPoIB, hence, is preferred for MPI applications. The compute nodes currently run on the Linux Operating System, which includes CentOS version 4.5 (Everett 2010).

R is an open-source integrated software suite, which was designed for data manipulation, computations, and graphical display associated with statistical computing. It runs on the Unix, Windows, and Mac families of operating systems, some of which are free of charge (such as Linux (Venables and Smith, 2010). The analyses were developed using both Windows and Linux versions of R version 2.12.0.

By default, R utilizes a single processor. Thus, consumption of time with and without dual processors was measured to verify performance. The performance of the 'ttime' package was measured using the system clock by executing the following code from the console:

```
library(ttime);
data(event_data);
npsp <- 1;
system.time(pred_vals <- translate(event_data, npsp), gcFirst = TRUE)
system.time(phylo(pred_vals), gcFirst = TRUE)
```

where line 1 loads the 'ttime' library into the current workspace. In line 2, the event data for known and unknown neurodevelopmental events across species is passed as a parameter to the data function, which will later be returned by the translate function. There are 10 species, 8 of which are non-primates; therefore, in line 3 shown above, only 1 species is evaluated. However, for the compute times shown in Tables 1 and 2 the number of non-primate species (npsp) is incremented by 1 to 8 to accommodate, accordingly.

The system time to predict events is shown in line 4, which generates a scatter plot of the translated neurodevelopmental event timings across species. The second parameter ensures that garbage collection is performed prior to the evaluation of the expression in the first parameter. Line 5 generates a phylogenetic tree via hierarchical clustering. It also returns the system time to generate a dendrogram and performs garbage collection prior to the process.

RESULTS

On a Windows Operating System (OS), the performance of a single processor when disabling the dual core is approximately half the compute time of an enabled dual core processor as shown in Tables 1 and 2. The time varies by approximately 20 seconds more for the system time utilized to generate the scatter plot as compared to the dendrogram whether it is in a single processor or dual processor mode on a Windows OS. This also holds true for the compute times on the Ares workstation as shown in Table 3. The metrics of a multi-core processor are actually negatively skewed because of the communication overhead and redundant computation between the two cores. Additionally, using the garbage collection option tends to add overhead because garbage collection takes precedence to allow for better performance of the function call.

In the HPC cluster environment compute times shown in Table 4, the time to generate a scatter plot shown in Figure 1 is almost 1/3 higher than the time to generate a dendrogram shown in Figure 2. It is important to note that the tasks in the HPC environment are executed on a single compute node, compute1-1. Although the workstation and the compute nodes both run on Linux, the compute nodes run version 4.5 of CentOS. (However, the HPC environment is scheduled for an upgrade to version 5.5.)

DISCUSSION

Our testing has shown that a single processor in the current environment is effective in generating both scatter plots and phylogenetic trees in a Windows environment. In a performance test on Windows and Linux operating systems, the compute node in the HPC

environment is slower; however, each compute node is operating at 2.66 GHz. The core speeds between Ares and compute nodes are similiar. CPUs are not getting faster but hardware manufacturers are adding better hardware subsystems such as faster buses and more.

Since data for the translating time across mammalian species model is anticipated to expand as more data about developmental events becomes available and incorporated into the analysis, it is important to explore the options related to parallelization of the 'ttime' package. An important gauge for users is the performance of a web-based application, which is measured by the execution time. The ultimate goal is to ensure timely delivery of the requested information as the dataset grows.

npsp	Scatter Plot	Dendrogram
1	52.40	52.51
2	52.54	49.89
3	52.73	55.46
4	52.56	50.03
5	53.06	50.59
6	52.52	50.51
7	52.57	49.89
8	52.34	49.60
Total	**420.72**	**408.48**

TABLE 1. 'ttime' Package Performance on Single Processor (Windows).

npsp	Scatter Plot	Dendrogram
1	104.20	100.00
2	102.50	99.59
3	101.66	100.15
4	102.58	99.81
5	101.51	100.26
6	102.08	100.51
7	106.33	100.6
8	102.43	101.93
Total	**823.29**	**802.85**

TABLE 2. 'ttime' Package Performance on Dual Processor (Windows).

npsp	Scatter Plot	Dendrogram
1	46.006	35.689
2	46.568	35.314
3	45.937	35.317
4	47.373	35.222
5	47.345	35.234
6	46.823	35.355
7	46.954	35.356
8	46.702	35.351
Total	**373.708**	**282.838**

TABLE 3. 'ttime' Package Performance on Dual Processor (Linux on Ares Workstation).

npsp	Scatter Plot	Dendrogram
1	83.951	57.852
2	82.904	57.373

3	82.675	55.653
4	85.914	56.047
5	79.39	55.932
6	85.197	57.372
7	84.962	56.621
8	85.838	56.809
Total	**670.831**	**453.659**

TABLE 4. 'ttime' Package Performance on HPC (Linux).

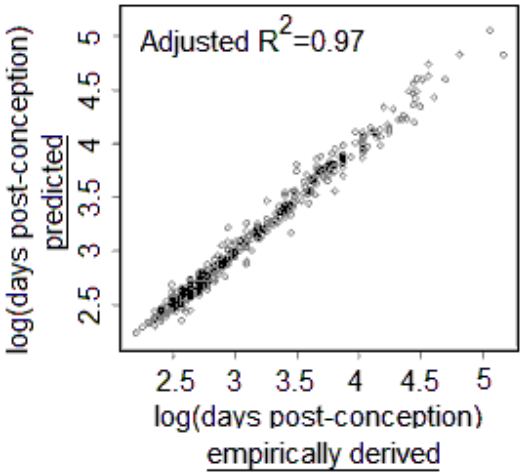

FIGURE 1. Scatter Plot.

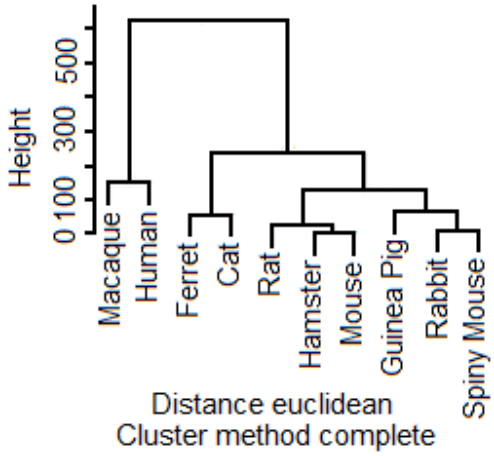

FIGURE 2. Cluster Dendrogram.

CONCLUSION

At this point, parallelization is an efficient means to overcome the speed bottleneck of a single processor, hence our investigation of the implementation and use of parallel computers. Compute nodes can run on a single machine with a single or several processors, or on multiple machines connected through a communications network. Future plans include the implementation of a parallelized version of the '*ttime'* package. The R parallelization will be achieved using the high performance computing (HPC) cluster environment at the University of

Arkansas at Little Rock Computer Science Department. Evaluation of the parallelization as a function of the number of clusters will be investigated.

The aforementioned is important because the translating developmental event time across mammalian species research has evolved into a web-based tool for researchers and clinicians. We request bona fida access to this research to benefit from the best performance available. This type of analysis is critical to the success of the project as more data about developmental events is identified and incorporated into the analysis.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0849626.

REFERENCES

1. (a) Clancy, B., B. Kersh, J. Darlington, K. Anand, and B. Finlay. 2007. “Web-based method for translating neurodevelopment from laboratory species to humans.” *Neuroinformatics* 5(1):79-94.
2. (b) Clancy, B., B. Finlay, R. Darlington, and K. Anand. 2007. “Extrapolating brain development from experimental species to humans.” *NeuroToxicology* 28(5):931-937.
3. Clancy, B., R. Darlington, and B. Finlay. 2000. “The course of human events: predicting the timing of primate neural development.” *Developmental Science* 3:57-66.
4. Clancy, B., R. Darlington, and B. Finlay. 2001. “Translating developmental time across mammalian species.” *Neuroscience* 105(1):7-17.
5. Everett, A. 2010. *General information about the hpc1-cpsc cluster*. Retrieved August 24 from https://plone2.git.ualr.edu/hpc1-cpsc-users/cluster-docs/general.
6. Nagarajan, R. 2010. *Package 'ttime'*. Retrieved August 23 from http://cran.r-project.org/web/packages/ttime/ttime.pdf.
7. Venables, W.N. and D.M. Smith. 2010. *An Introduction to R*. The R Core Development Team, retrieved August 23 from http://www.r-project.org/.