Client–Server environment for high-performance gene expression data analysis

Alexander Sturn, Bernhard Mlecnik, Roland Pieler, Johannes Rainer, Thomas Truskaller and Zlatko Trajanoski*

Institute of Biomedical Engineering, Graz University of Technology, Christian Doppler Laboratory for Genomics and Bioinformatics, Krenngasse 37, 8010 Graz, Austria

Received on September 2, 2002; revised on October 23, 2002; November 18, 2002; accepted on December 16, 2002

ABSTRACT

Summary: We have developed a platform independent, flexible and scalable Java environment for high-performance large-scale gene expression data analysis, which integrates various computational intensive hierarchical and non-hierarchical clustering algorithms. The environment includes a powerful client for data preparation and results visualization, an application server for computation and an additional administration tool. The package is available free of charge for academic and non-profit institutions.

Availability: http://genome.tugraz.at/Software

Contact: zlatko.trajanoski@tugraz.at

INTRODUCTION

High-throughput gene expression analysis using oligonucleotide or cDNA microarrays is becoming increasingly important in many areas of basic and applied biomedical research. The microarray technology itself is developing rapidly, leading to an increasing density of the elements spotted onto a single slide. However, these genome-wide microarrays pose significant challenges on the data analysis tools. Many gene expression data mining algorithms utilize a similarity matrix as a starting point, in which the distances between all genes are calculated on the basis of a similarity function (Eisen et al., 1998). The similarity matrix is a triangular matrix containing \((n^2 - n)/2\) elements, where \(n\) is the number of genes. Consequently, the similarity matrix of a genome-wide array with 30,000 genes requires almost 1.7 GB (230 b) of RAM, assuming that each cell is represented by a floating point value of 4 B. Moreover, this is just one of many matrices, lists, and lookup tables mandatory for the calculation of a gene expression clustering or classification. It is noteworthy that the Java Virtual Machine on 32-bit computer architectures like Personal and Apple Computers is limited to 2 GB of memory. Thus, more demanding jobs using some of the popular cluster analysis tools (Sturn et al., 2002) require costly 64-bit soft- and hardware architecture. Due to these constraints, data analysis of genomic scale microarrays becomes impractical or even impossible to perform on commonly used workstations. Computer architecture, CPU performance, amount of addressable and available memory, and costs are the limiting factors. Consequently, memory and calculation intensive tasks have to be outsourced to high-performance servers. We have therefore further developed our gene expression data analysis suite Genesis (Sturn et al., 2002) to be capable of using the advantages of outsourcing the calculations to in-house or remote application servers.

PROGRAM OVERVIEW

The client–server environment (Fig. 1) consists of a versatile, platform independent, and easy to use Java client for data preprocessing and results visualization (Genesis Client), an application server (Genesis Server) for computation of Hierarchical Clustering (HCL; Eisen et al., 1998), Self Organizing Maps (SOM; Tamayo et al., 1999), k-means Clustering (KMC; Tavazoie et al., 1999), and Support Vector Machines (SVM; Brown et al., 2000), as well as an additional administration tool for statistics, job handling, and user management (Genesis Server Client). Data analysis is prepared in Genesis Client and the jobs are distributed to an available Genesis Server, where calculation is started and results are stored until they are fetched by the client. At all times the client is informed about status and progress of the calculation task. Nevertheless, all server jobs are completely independent from the client, so that the client may be turned off during calculation and restarted again later to retrieve the computed results. The user management system of the server warrants that only enrolled users have the rights to submit jobs and get their progress information and results. Additionally, it provides the functionality to specify the

*To whom correspondence should be addressed.
number of calculation tasks each user is allowed to calculate simultaneously and in total. For controlling the server we have enclosed the standalone application Genesis Server Client, which enables system administrators to add or change user accounts in a straightforward manner, observe the server status, and abort specific calculation tasks if necessary. It also provides information on all calculated jobs by accessing the database incorporated into the Genesis Server. The latter is used to handle jobs, user accounts, and results in a reliable and secure environment. Our implementation uses the free available application server JBoss (http://www.jboss.org), is completely developed in Java, and available free of charge to academic and non-profit organizations. This renders it, to the best of our knowledge, the most cost effective solution for distributed high-performance gene expression data analysis. The Genesis Server environment is also scalable to high-performance multiprocessor servers. Up to date, the package has been tested on Windows 2000/XP, Linux (2 Intel PIII, 2 GB RAM), Solaris (Sun Fire V880, 4 UltraSPARC III, 8 GB RAM) and Tru64 Unix (AlphaServer ES45, 4 Alpha processors, 16 GB RAM) platforms.

FUTURE DEVELOPMENT

Present and future work will focus on porting the server to computer cluster environments to parallelize the huge computational tasks of gene expression clustering using bootstrapping and automatic parameter fitting. Additionally a job queuing system is in development to further improve performance and usability.

ACKNOWLEDGEMENTS

We thank our informatics staff and faculty for valuable comments and contributions. This work was supported by a grant F718 (SFB Biomembranes) from the Austrian Science Fund and by an Academic Equipment Grant from SUN Microsystems.

REFERENCES


