Parallelized preprocessing algorithms for high-density oligonucleotide array data

M. Schmidberger and U. Mansmann

Chair of Biometrics and Bioinformatics, IBE, University of Munich, Germany
schmidb@ibe.med.uni-muenchen.de

Abstract. Studies of gene expression using high-density oligonucleotide microarrays have become standard in a variety of biological contexts. The data recorded using the microarray technique are characterized by high levels of noise and bias. These failures have to be removed, therefore preprocessing of raw-data has been a research topic of high priority over the past few years.

Actual research and computations are limited by the available computer hardware. For many researchers the available main memory limits the number of arrays that may be processed. Furthermore most of the existing preprocessing methods are very time consuming and therefore not useful for first and fast checks in laboratories. To solve these problems, the potential of parallel computing should be used. In microarray technologies and statistical computing parallel computing does not appear to have been used extensively. For parallelization on multicomputers, message passing (MPI) methods and the R language will be used.

Ideas for parallelization of VSN and FARMS as well as a large project in applied bioinformatics (> 5000 microarrays) will be discussed. Furthermore this presentation proposes the new BioConductor package `affyPara` for parallelized preprocessing of high-density oligonucleotide microarray data. Partition of data could be done on arrays and therefore parallelization of algorithms gets intuitive possible. In view of machine accuracy, the same results as serialized methods will be achieved. The partition of data and distribution to several nodes solves the main memory problems and accelerates the methods by up to the factor ten.

References

R. Gentleman, et all. (2005): Bioinformatics and Computational Biology Solutions Using R and Bioconductor. *Springer, Statistics for Biology and Health*

A. Rossini (2003): Simple Parallel Statistical Computing in R. *UW Biostatistics Working Paper Series, 193*

H. Sevcikova (2003): Statistical Simulations on Parallel Computers. *Journal of Computational and Graphical Statistics, 13, pp. 886-906*

R. A. Irizarry, et all. (2004): Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4, Apr, Nr. 2, 249264*

M. Schmidberger, U. Mansmann (2008): Parallelized preprocessing algorithms for high-density oligonucleotide array data. *22th International Parallel and Distributed Processing Symposium (IPDPS 2008), Proceedings, 14-18 April 2008, Miami, Florida, USA. IEEE 2008 (in press)*