“A study on simulating realistic gene expression microarray data”

PI: Guan-Hua Huang (National Chiao Tung University)
Role: Principal investigator

National Science Council of Taiwan: 08/01/2012-07/31/2014

Microarray gene expression analysis has become one of the most widely used functional genomics tools. Since that, many analytical methods have been proposed. It is desirable to develop realistic models that can be applied in simulating expression values of each gene, and can then be used to assess the analysis methods and testing approaches. In this project, we plan to download publicly available raw data of the Affymetrix HG-U133A platform for various tissues from two public repositories: Gene Expression Omnibus and ArrayExpress. Then, an empirical approach is developed to determine the distribution of expression intensity for each gene, which can be used to simulate realistic gene expression data. The proposed method has several unique features that resolve the shortage of previous research. To evaluate the proposed simulating approach, we will examine the distributions of housekeeping genes, compare the simulated and real gene expression data, and simulate gene expression intensities, which mimic the expression patterns shown in the HG-U133A tag spike-in dataset, to determine the sensitivity and specificity of various differential expression detecting methods. This project also attempts to use OpenMP and MPI parallel computing to reduce computing time when reprocessing the large amount of downloaded microarray raw data. We will compare the parallel efficiency of OpenMP and MPI in the high efficient personal workstation, the National Center for High-performance Computing and the Amazon EC2 cloud computing environment. The results and experiences gained from this experiment can be applied to future high-dimensional genomic data computation.

This study was proposed for a three-year project on year 2011, but we only obtained funding for one year. Our team has finished implementing high efficient parallel computing for reprocessing the large amount of downloaded microarray raw data. We are now planning to continue the un-done part of the study in the coming two years.

Keywords：Affymetrix GeneChip; cloud computing; gene expression microarray; microarray data archive; parallel computing; simulation.