



# A Divide-and-conquer Strategy to Solve the Out-of-memory Problem of Processing Thousands of Affymetrix Microarrays

Cheryl Lee – Masters Student  
Computational Biology & Bioinformatics

Friday, November 30, 2007  
2:00 pm – 3:00 pm  
COOK 31 18 A&B

## ABSTRACT:

An extremely large amount of microarray data has been produced and accumulated. As of October 2007, 173,486 arrays have been deposited into the NCBI GEO database. It is anticipated that more information can be discovered from large-scale experiments or large compilations of experiments than from small experiments with just a handful arrays. However, processing such a huge volume of data requires tremendous computing resources, which is far beyond the capacity of most research labs. A frequently encountered issue is the out-of-memory problem when processing thousands of CEL files generated by the Affymetrix platform using Bioconductor. We propose a divide-and-conquer strategy to solve this problem. It works recursively by breaking down a problem into many sub-problems of the same type, until they become simple enough to be solved directly in the memory. The solutions to the sub-problems are then combined to give a solution to the original problem. We used the CAMDA 2007 META-analysis data set, which contains 5,896 microarrays, to test our approach. The results were validated against a gold standard data set obtained by using a main frame computer to run 5,896 arrays on a computer with 1TB of physical memory. In summary, this study is aimed at developing a general strategy to run any established Affymetrix pre-processing algorithms in the Bioconductor package on a commodity computer cluster (32-bit CPU and 1GB of memory for each CPU).

## Research Advisor:

Simon Lin, Feinberg School of Medicine