**Summary of Research**

Jong Youl Choi

School of Informatics and Computing

Indiana University – Bloomington, IN

jychoi@cs.indiana.edu

I have accomplished solid and original research work in pursuit of my degree. With large-scale and data-intensive computing in mind, I have been working on my dissertation research primarily in two directions: i) developing novel parallel data mining and machine learning algorithms for large and high-dimensional data analysis, and ii) implementing distributed and parallel algorithms for processing large-scale data with efficient use of distributed computing resources, including high-performance clusters, grids, and cloud systems.

The first part of my work was about developing novel dimension reduction algorithms for visual data analysis. Data point visualization is one of the most important and basic methods in scientific data analysis. Thus, developing efficient visualization algorithms based on dimension reduction (also known as low-dimensional embedding) for large-scale and high-dimensional data is a crucial part for the modern data-intensive scientific discovery.

Among many dimension reduction algorithms, including traditional Principle Component Analysis (PCA) and Self Organizing Map (SOM), I have focused on Generative Topographic Mapping (GTM). Due to its solid theoretical foundation, GTM is superior to other methods in visualization. One drawback, however, is its use of Expectation-Maximization (EM), an optimization method, which can easily be trapped in local optimum, not finding global solutions. I have improved and developed a new variant of GTM, called DA-GTM by applying a novel optimization method called Deterministic Annealing (DA) [[1](#_ENREF_1)]. The new algorithm is more robust against the local optima problem. To the best of my knowledge, this work is the first attempt ever made to apply the Deterministic Annealing (DA) method to GTM.

To take further steps to cope with large-scale data problem, I have also improved the GTM algorithm with parallelization [[2](#_ENREF_2)] and interpolation [[3](#_ENREF_3)], respectively. Both of them are designed to maximally utilize a large number of computing nodes concurrently in a distributed cluster system [[4](#_ENREF_4)].

All of my research work were successfully applied as real life-science applications in bioinformatics and cheminformatics and the results were presented and published in many high-quality conferences and journals (eg. HPDC, CCGrid, ICCS).

In a separate but related project, I also worked on developing the cloud systems for parallel bioinformatics applications, named Virtual Collaborative Lab for Protein Sequence Analysis (V-Lab-Protein) and Microarray Data Analysis (V-Lab-Microarray). These were designed to provide virtual and flexible cloud computing resources by using Amazon EC2 and S3 services. The V-lab-Protein system is now recognized as one of the first science applications using cloud computing infrastructures and it was introduced in the October 2007 issue of Nature’s News [[5](#_ENREF_5)].

[1] J. Y. Choi, J. Qiu, M. Pierce, and G. Fox, "Generative Topographic Mapping by Deterministic Annealing," presented at the ICCS 2010, Amsterdam, The Netherlands, 2010.

[2] J. Y. Choi, S.-H. Bae, X. Qiu, and G. Fox, "High Performance Dimension Reduction and Visualization for Large High-dimensional Data Analysis," presented at the The 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid2010), Melbourne, Australia, 2010.

[3] S.-H. Bae, J. Y. Choi, J. Qiu, and G. C. Fox, "Dimension Reduction and Visualization of Large High-dimensional Data via Interpolation," presented at the HPDC'10, Chicago, Illinois USA, 2010.

[4] J. Y. Choi, S.-H. Bae, J. Qiu, B. Chen, and D. Wild, "Browsing large-scale cheminformatics data with dimension reduction," *Concurrency and Computation: Practice and Experience,* 2011.

[5] E. Hand, "Head in the clouds," *Nature,* vol. 449, pp. 963-963, Oct 25 2007.