

Speeding Up Clustering of Gene Expression Data

Supervisor: P. Fazendeiro, Co-Supervisor: P. Prata

1 Abstract

The main purpose of this project is to address the feasibility of a high performance system that performs gene expression data analysis at a relatively low cost. The gene expression data analysis task brings many challenges due essentially to the high-throughput and high-dimensional data. The proposed approach is based on addressing the problem from two different fronts in an integrated manner. From the hardware side, we advocate the exploration of affordable multi-core parallel architectures enriched with multiple graphical processing units (GPUs) dedicated for scientific computation. From the software side, we intend to develop a parallel version of a current fuzzy clustering algorithm. The potential applications to the type of system proposed for this project are manifold ranging from the identification of genetic diseases to the individual adaptation of the medical treatment.

2 Objectives and Tasks

T1 Study the problem

T2 Select the training and test data

T3 Implement parallel GPU versions of a clustering algorithm

T4 Compare the performance of the sequential and parallel versions

T5 Write up a final report

3 Timetable

T1 3 weeks

T2 2 weeks

T3 5 weeks

T4 3 weeks

T5 2 weeks

4 Expected Results

- 1 Software Prototype
- 1 Final Report

5 References

- Fuzzy Clustering Based Parallel Cultural Algorithm, Jihane Alami , L. Benameur and Imrani International Journal of Soft Computing Year: 2007, Volume: 2, Issue: 4, Page No.: 562-571

- Parallel Fuzzy c- Means Clustering for Large Data Sets Terence Kwok, Kate Smith, Sebastian Lozano and David Taniar COMPUTER SCIENCE EURO-PAR 2002 PARALLEL PROCESSING Lecture Notes in Computer Science, 2002, Volume 2400/2002, 27-58