





Microarrays data analysis of cancer disease on a cluster of computers

Thesis submitted to the Department of Scientific Computing
Faculty of Computer and Information Sciences
Ain Shams University

In partial fulfillment of the requirements for the degree
of Master in Computer and Information Sciences

By

Dina Mohamed Elsayad

B.Sc. in Computer and Information Sciences (2007)
Ain Shams University – Cairo

Under the supervision of

Prof. Dr. El-Sayed M. El-Horbaty

Professor and Head of Computer Science Department
Faculty of Computer and Information Sciences
Ain Shams University

Prof. Dr. Mohammed Essam Khalifa

Professor of Basic Sciences
Faculty of Computer and Information Sciences
Ain Shams University

Dr. Amal Said Khalifa

Lecturer of Scientific Computing
Faculty of Computer and Information Sciences
Ain Shams University

Cairo-2013



Scientific Computing Department
Faculty of Computer & Information Sciences
Ain Shams University

Microarrays data analysis of cancer disease on a cluster of computers

This thesis submitted to the Department of Scientific Computing, Faculty of Computer and Information Sciences, Ain Shams University, in partial fulfillment of the requirements for the degree of Master of Computer and Information Sciences.

BY

Dina Mohamed Elsayad Anwar

Approved by the discussion committee:

Prof. Dr. El-Sayed M.El-Horbaty

Member & Supervisor

Professor & Head of Computer Sciences Department.

Faculty of Computer and Information Sciences, Ain Shams University.

Prof. Dr. Mohammed Essam Khalifa

Member & Supervisor

Professor of Basic Sciences

Faculty of Computer and Information Sciences, Ain Shams University.

Prof. Dr. Aboul Ella Hassanien

Member

Professor of Information Technology

Faculty of Computer and Information Sciences, Cairo University

Prof. Dr. Mohamed El-Sayed Waheed

Member

Dean of the Faculty of Computers and Informatics - Suez Canal University



Scientific Computing Department - Faculty of Computer and Information Sciences
Ain Shams University
Cairo – 2013

Acknowledgement

I acknowledge my deep gratitude to ALLAH for all gifts in my life and for helping me to complete this work on a level that I hope will please the reader.

Special thanks are due to my supervisors; Prof. Dr. El-Sayed M. El-Horbaty, Prof. Dr. Mohammed Essam Khalifa and Dr. Amal Said Khalifa for their brilliant guidance, encouragement, constructive feedback, resourceful ideas and suggestions that helped in enhancing my work.

Finally; With all my appreciation and love that no words can express, I dedicate this Master to my beloved family members for their love and support, my dear parents , my brother and my sisters who always provided me with love, prayers, blessings, and care. They are behind any success in my life.

Abstract

Microarray is one of the hottest fields related to bioinformatics. The microarray is an exceptionally powerful genomic tool, capable of examining thousands of genes simultaneously. By examining microarray data, researchers can infer the relationships between genes and thus gain a better understanding of how genes interact within an organism.

One type of the microarrays is DNA microarrays where the biological material is DNA fragments. Extracting biologically relevant information from DNA microarrays is a very important task for drug development and test, function annotation, and cancer diagnosis. Various clustering methods have been proposed for the analysis of gene expression data. Because of the huge size of the microarrays data the clustering process is time consuming. So, Olman et al. presented a parallel MST-based clustering technique called CLUMP (**Clustering through MST in Parallel**). Although, this algorithm has enhanced the speed of MST construction using a parallel algorithm, the speed of the algorithm can be more enhanced especially the MST construction phase which is still the most time consuming phase.

In this work three enhanced versions of the CLUMP algorithm were proposed iCLUMP, iCLUMP-2 and hiCLUMP. iCLUMP (**Improved CLUMP**) enhanced the speedup of MST

construction phase using cover tree data structure. While, iCLUMP-2 (**iCLUMP** 2nd version) enhanced the edge inconsistency measure to detect more types of clusters and hence improve the cluster quality. Finally, hiCLUMP (**H**ybrid **I**mplementation of **CLUMP**) enhanced the MST construction phase through increasing the computing power. It used the hybrid architecture to reduce the MST construction time.

The performance of the implemented algorithms was tested on a 45 processing nodes cluster using various cancer microarrays datasets. The results showed that the iCLUMP-2 algorithm outperformed all of CLUMP, iCLUMP and hiCLUMP providing better speedup and efficiency. While, iCLUMP outperformed both of CLUMP and hiCLUMP in terms of speedup and efficiency. On the other hand, hiCLUMP is outperformed CLUMP providing better speedup and efficiency. Furthermore the quality of cluster produced by the iCLUMP-2 algorithm is much better than those produced by all of CLUMP, iCLUMP and hiCLUMP.

Table of Contents

Acknowledgement	i
Abstract	iii
Table of Contents	v
List of Figures	viii
List of Tables	xi
List of Publications	xii
1. Introduction	3
1.1. Motivation	3
1.2. Problem Definition	3
1.3. Research Objectives	4
1.4. Thesis Organization	5
2. Background Overview	9
2.1. Biological preliminaries	9
2.1.1. DNA	9
2.1.2. RNA	11
2.1.3. Genes	12
2.1.4. Proteins	14
2.2. Central Dogma	18
2.3. Bioinformatics	22
2.4. Major research areas of Bioinformatics	24
2.4.1. Genome annotation	25
2.4.2. Sequence analysis	26
2.4.3. Computational evolutionary biology	28
2.4.4. Measuring biodiversity	30
2.4.5. Analysis of gene expression	31
2.4.6. Analysis of regulation	35
2.4.7. Analysis of mutations in cancer	36
2.4.8. Prediction of protein structure	37

2.4.9.	Modeling biological systems	38
2.4.10.	High-throughput image analysis	39
2.4.11.	Protein-protein docking.....	40
2.5.	Parallel Computing Paradigms.....	41
3.	Microarrays data analysis	47
3.1.	Microarrays Technology	47
3.1.1.	Microarray Experiment Workflow	52
3.1.2.	Microarray image color interpretation	55
3.2.	Microarrays Manufactures	56
3.3.	Microarrays data repository	58
3.4.	Microarrays Data analysis.....	62
3.5.	Graph representation of Microarrays data.....	64
3.6.	Clustering Techniques.....	66
3.6.1.	Fuzzy clustering.....	67
3.6.2.	Hard clustering.....	68
3.7.	Minimum spanning tree based clustering.....	75
3.8.	Cluster quality assessment techniques	78
4.	Clustering Methods	83
4.1.	Introduction	83
4.2.	CLUMP algorithm.....	84
4.2.1.	MST construction.....	84
4.2.2.	Inconsistent edges identification.....	93
4.2.3.	Clusters identification	94
4.3.	iCLUMP Algorithm	98
4.4.	iCLUMP-2 Algorithm.....	105
4.5.	hiCLUMP Algorithm	108
5.	Implementation and Results.....	115
5.1.	System architecture	115
5.2.	Datasets	115
5.3.	Performance Measures	117
5.4.	Experimental Results.....	118

5.4.1.	iCLUMP.....	119
5.4.2.	iCLUMP-2	121
5.4.3.	hiCLUMP.....	122
5.4.4.	Comparative Study.....	124
6.	Conclusions and Future Work	137
6.1.	Conclusions	137
6.2.	Future Work	138
	Bibliography	139

List of Figures

Figure 2.1 : The structure of DNA.....	11
Figure 2.2 : Gene in relation to the DNA and to a chromosome	14
Figure 2.3 : Amino Acids Types.....	15
Figure 2.4 : Protein Structure	16
Figure 2.5 : The Central Dogma of molecular biology	21
Figure 2.6 : Relation between Bioinformatics and Computational Biology.....	24
Figure 2.7 : Graphical representation for a WMM model	26
Figure 2.8 : Superposition of five proteins and a corresponding sequence alignment.....	28
Figure 2.9 : Phylogenetic tree of Life	30
Figure 2.10 : Graphical representation for biodiversity measure against the time	31
Figure 2.11 : Steps of SAGE.....	34
Figure 2.12 : parallel computing architecture	43
Figure 3.1 : Lab-on-a chip made of glass.....	49
Figure 3.2 : Microarray Scanner	49
Figure 3.3 : DNA Microarrays.....	50
Figure 3.4 : Two-Channel microarray	51
Figure 3.5 : Workflow of a DNA microarray experiment	53
Figure 3.6 : Microarray image color interpretation.....	56
Figure 3.7 : Geo Data structure	61
Figure 3.8 : Microarrays Data analysis tasks	62
Figure 3.9 : Graph representation of microarrays data	66
Figure 3.10 : Types of clustering algorithms	67

Figure 3.11 : Bi-clusters patterns	73
Figure 3.12 : MST-based clustering algorithm steps	76
Figure 4.1 : MST construction phase	92
Figure 4.2 : Bipartite graph	93
Figure 4.3 : Clusters identification phase	96
Figure 4.4 : Nesting invariant	100
Figure 4.5 : Covering invariant	101
Figure 4.6 : Separation invariant	101
Figure 4.7 : Cluster composed of sparse points	107
Figure 5.1 : The runtime of the CLUMP, and iCLUMP applied on dataset with accession number GSE6532	120
Figure 5.2 : The runtime of the CLUMP, and iCLUMP-2 applied on dataset with accession number GSE9195	122
Figure 5.3: The runtime of the CLUMP, and hiCLUMP applied on dataset with accession number GSE9195	124
Figure 5.4: The run of the CLUMP,hiCLUMP, iCLUMP and iCLUMP-2 applied on the dataset GSE6008	127
Figure 5.5: The run of the CLUMP,hiCLUMP, iCLUMP and iCLUMP-2 applied on the dataset GSE7390	127
Figure 5.6: The run of the CLUMP,hiCLUMP, iCLUMP and iCLUMP-2 applied on the dataset GSE2034	128
Figure 5.7: The run of the CLUMP,hiCLUMP, iCLUMP and iCLUMP-2 applied on the dataset GSE3494	128
Figure 5.8: The run of the CLUMP,hiCLUMP, iCLUMP and iCLUMP-2 applied on the dataset GSE9195	129
Figure 5.9: The run of the CLUMP,hiCLUMP, iCLUMP and iCLUMP-2 applied on the dataset GSE6532	129