

Contents

LIST OF FIGURES	vii
LIST OF TABLES	ix
GLOSSARY	xi
ACKNOWLEDGMENT	xiii
ABSTRACT	xiv
1 Introduction	1
1.1 Thesis Overview	1
1.2 Thesis Objective	5
1.3 Thesis Organization	5
2 Biological Background	6
2.1 Introduction to Bioinformatics	6
2.1.1 Historical Development	7
2.1.2 The Need for Bioinformatics	7
2.1.3 Bioinformatics Impact on Health Life	8
2.1.4 Bioinformatics Research Area	8
2.2 Basic Biology	10
2.2.1 Prokaryotic and Eukaryotic Cell Types	10
2.2.2 Molecules of Life	11
2.2.2.1 Small Molecules	12
2.2.2.2 Proteins	13
2.2.2.3 DNA	14
2.2.2.4 RNA	17
2.2.2.5 Chromosomes and Genomes	18
2.2.3 Genes and Protein Synthesis	19
2.2.4 Gene Function	22
2.3 Data Acquisition Methodologies	23
2.3.1 DNA Sequencing Methodology	23
2.3.2 Microarray Technology	24
2.3.2.1 Microarray Concepts	25
2.3.2.2 Gene Expression Data Analysis	26
2.3.2.3 Microarray Data Limitations	27

2.4	Biological Database	28
2.5	Saccharomyces Cerevisiae	30
2.5.1	Why Yeast	31
2.5.2	Yeast Genes Features	32
2.5.3	Yeast Gene Naming	32
2.5.3.1	Standard Name	32
2.5.3.2	Systematic Names	34
2.5.4	Yeast Databases	34
2.5.4.1	Genome Database	34
2.5.4.2	Microarray Database	35
2.5.5	Spellman Cell Cycle Experiment	35
2.5.6	Gasch Environmental Changes Experiment	37
3	Gene Expression Data Analysis	38
3.1	Data Acquisition	38
3.2	Preprocessing	38
3.2.1	Update Genes List	39
3.2.1.1	Important <i>S. Cerevisiae</i> Genes Which Are Not Included In Gasch Dataset	41
3.2.2	Filtration	42
3.2.2.1	Conditions Filtration	43
3.2.2.2	Genes Filtration	44
3.2.2.3	Remove Genes with Large Missed Values	44
3.2.2.4	Remove Genes with Small Profile Variance	44
3.2.2.5	Remove Genes with Low Absolute Values	45
3.2.2.6	Remove Genes with Low Entropy	48
3.2.3	Imputation Missing Values	49
3.2.4	Normalization	51
3.2.5	Discertization	52
3.2.6	Data Denoising	52
3.3	Data Partitioning: Clustering Algorithms	54
3.3.1	What is Clustering?	55
3.3.2	K-means	56
3.3.3	Hierarchical clustering (HCL)	57
3.4	Biclustering Algorithms	57
3.4.1	Cheng and Church (CC)	58
3.4.2	Iterative Signature Algorithm (ISA)	59
3.4.3	Biclusters Inclusion Maximal (Bimax)	60
3.4.4	Order Preserving Submatrix(OPSM)	61
3.4.5	Maximum Similarity Bicluster(MSBE)	61
3.5	AGO:Automatic Gene Ontology Software Tool for Bicluster and Cluster Comparisons	61
3.5.1	Comparison Methodology	62
3.5.2	Gene Ontology	63
3.5.3	Hypergeometric Test	65
3.5.4	GO Enrichment Programs	66
3.5.5	AGO Implementation	66

3.5.6	AGO Testing: Case Study	68
3.6	BicAT-plus: An Automatic Comparative Java Tool For Bi/Clustering Algorithms Used In Analysis And Visualization of Gene Expression Data Obtained Using Microarrays	72
3.6.1	BicAT-Plus Development and Architecture	74
3.6.2	BicAT-Plus Comparison Process Steps	77
4	Bayesian Network	80
4.1	Reverse Engineering Approach	80
4.1.1	Boolean Network	82
4.1.2	Non Linear Definitional Equations	83
4.1.3	Stochastic Differential Equation	84
4.1.4	Association Network	85
4.2	Which Model Should I Select?	86
4.3	Data Sources and Requirements	87
4.4	Bayesian Network	88
4.4.1	Bayesian Networks Representation	88
4.4.2	Bayesian Networks Structure Learning	89
4.4.2.1	Scoring Function	90
4.4.2.2	Heuristic Search	91
4.4.2.3	Model Averaging	91
4.5	Performance Comparison of the Structure Learning Bayesian Network Algorithms Using Gene Expression Data	93
4.5.1	K2 algorithm	93
4.5.2	MCMC	93
4.5.3	BNPC	94
4.5.4	GSMES	94
4.5.5	The dataset	94
4.5.6	Comparison Methodology	95
4.6	Dream Project	98
5	Results	100
5.1	Comparison to Previous Algorithms	100
5.2	Comparison to Literature	102
5.3	Network Generation	104
5.3.1	Biclustering Phase	105
5.3.2	Learning Phase	105
5.4	Evaluation Methodology	107
5.5	Network Analysis and Validation	113
5.5.1	Network Topology	114
5.5.2	Finding Network Module	116
6	Conclusion and Future Work	119
	Bibliography	122