

Mining Clusters for Knowledge:  
Finding Algorithm-Independent Groups in  
Microarray Data

by

Joshua E. Blumenstock  
Class of 2003

A thesis submitted to the  
faculty of Wesleyan University  
in partial fulfillment of the requirements for the  
Degree of Bachelor of Arts  
with Departmental Honors in Physics  
and the Computer Science Program

# Mining Clusters for Knowledge: Finding Algorithm-Independent Groups in Microarray Data

by

Joshua E. Blumenstock  
Class of 2003

## **ABSTRACT**

Using DNA microarray technology, it is now possible to measure the expression levels of tens of thousands of genes. Statistical analysis of these expression levels provides insight into the function of genes and their biological pathways, as well as information about the genomic underpinnings of many common diseases. Cluster analysis is a form of unsupervised learning commonly used to analyze microarray data, and there are several different types of cluster analysis to choose from. It is widely acknowledged that the different types of cluster analysis can produce vastly inconsistent results, yet there is no known way to deal with these inconsistencies. In this thesis, I present a novel approach to the cluster analysis of microarray data. The proposed methodology combines and distills the information generated by different types of cluster analysis, and produces a representative clustering structure. Several new statistics are developed to identify dominant clusters and assess consistency across clustering algorithms. Using real data from leukemia patients, the proposed methodology is shown to outperform the naïve choice of a single algorithm.

# TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	V
<b>1 INTRODUCTION.....</b>	<b>1</b>
1.1 Abstract.....	1
1.2 Organization of Thesis.....	2
<b>2 BACKGROUND ON MICROARRAY TECHNOLOGY .....</b>	<b>7</b>
2.1 Genetic Background .....	7
2.2 Microarrays and Expression Levels.....	8
2.3 Quantitative Aspects of Microarray Data.....	9
<b>3 CLUSTER ANALYSIS OF MICROARRAY EXPERIMENTS .....</b>	<b>12</b>
3.1 Introduction to Cluster Analysis.....	12
3.2 Clustering Algorithms .....	15
3.2.1 <i>Hierarchical Algorithms</i> – Agglomerative Hierarchical Clustering...	16
3.2.2 <i>Partitioning Algorithms</i> - <i>K</i> -Means .....	18
3.3 Major Sources of Variability in Clustering .....	20
3.3.1 Choice of Algorithm .....	20
3.3.2 Choice of Parameters .....	20
3.3.3 Choice of Input Data.....	21
3.4 Summary.....	22
<b>4 METHODS .....</b>	<b>24</b>
4.1 Motivation & Prior Work .....	24
4.2 Overview of Methods .....	28
4.3 Producing A Large Collection of Clusters: ‘Shotgun’ Clustering.....	31
4.4 Measuring Global Consistency of the Collection: $\lambda$ .....	33
4.5 Identifying Prevalent Clusters: $p_{\mathcal{D}}$ .....	34
4.6 Clustering the Clusters: ‘Condensation’ Clustering .....	36
4.6.1 Clustering Clusters.....	37
4.6.2 Merging Clusters to Form Multisets of Samples .....	38
4.6.3 Clustering Multisets .....	39
4.6.4 Distance Between Multisets.....	40
4.6.5 Condensation Algorithm.....	42
<b>5 RESULTS .....</b>	<b>44</b>
5.1 Data.....	44
5.2 Shotgun Clustering .....	46
5.3 Diagnostic Tests.....	47
5.3.1 Applications of $\lambda$ .....	47
5.3.2 Distribution of Prevalent Clusters.....	52
5.4 Biologically-Relevant Results .....	53
5.4.1 Single Clustering Configurations Miss the ‘Actual’ Partition.....	54

5.4.2	Mining Clusters for Information With Prevalence $p_D$ .....	56
5.4.3	Mining for Patterns Using Condensation Clustering .....	59
5.5	Summary of Results.....	63
<b>6</b>	<b>CONCLUSIONS .....</b>	<b>67</b>
6.1	Conclusions.....	67
6.2	Directions for Future Research.....	68
	<b>REFERENCES.....</b>	<b>70</b>
	<b>APPENDIX A: SOURCES OF VARIABILITY AND INDETERMINACY IN THE CLUSTERING PROCESS .....</b>	<b>76</b>
A.1	Choice of Algorithm .....	77
A.2	Choice of Parameters .....	81
A.2.1	Distance Metric .....	81
A.2.2	Linkage Function (hierarchical).....	85
A.2.3	Choice of k-value, and the placement of the seeds (for k-means) .....	89
A.3	Choice of Input Data.....	90
	<b>APPENDIX B: IMPLEMENTATION.....</b>	<b>94</b>
B.1	Overview .....	94
B.2	Database Schema and Code .....	96
B.3	Shotgun Stage .....	99
B.4	Consistency and Prevalence .....	101
B.5	Condensation Stage.....	102
B.6	Results & Visualization .....	104

# ACKNOWLEDGEMENTS

If anyone is to be acknowledged first, it must be my parents. I thank them for giving me love and getting me to a place from which I can publicly acknowledge them.

With respect to this thesis, I owe an enormous debt to my advisors Michael Rice and Rick Jensen. Dr. Rice provided tremendous feedback and direction, especially in the last few weeks as the thesis went from a mess of code to a document with paragraphs and page numbers. Dr. Jensen has been an inspiration for many years now, and was always ready to help with an inexhaustible supply of ideas and insight.

I also thank Dr. Adam Fieldsteel for helping me wade through the depths of probability and the murky waters of statistics, Dr. Michael Keane for much-needed discussion and criticism, and Dr. Danny Krizanc for letting me scribble all over his chalkboard.

Lastly, thanks to Dan, Alex, Shana and the other cohorts in PacLab, mainly just for being idiots with me at 7am. To Loveland and everyone else who asked me to explain this thesis, thank you, without that continual grounding these pages would surely not make sense.

# 1 INTRODUCTION

## 1.1 Abstract

Using DNA microarray technology, it is now possible to measure the expression levels of tens of thousands of genes. Statistical analysis of these expression levels provides insight into the function of genes and their biological pathways, as well as information about the genomic underpinnings of many common diseases. Cluster analysis is a form of unsupervised learning commonly used to analyze microarray data, and there are several different types of cluster analysis to choose from. It is widely acknowledged that the different types of cluster analysis can produce vastly inconsistent results, yet there is no known way to deal with these inconsistencies. In this thesis, I present a novel approach to the cluster analysis of microarray data. The proposed methodology combines and distills the information generated by different types of cluster analysis, and produces a representative clustering structure. Several new statistics are developed to identify dominant clusters and assess consistency across clustering algorithms. Using real data from leukemia patients, the proposed methodology is shown to outperform the naïve choice of a single algorithm.

## **1.2 Organization of Thesis**

### ***Section 2: Background on Microarray Technology***

With DNA microarray technology we can now simultaneously measure the expression levels of tens of thousands of genes. Each microarray corresponds to a specified set of experimental conditions (e.g. different patients or different cell lines), and the thousands of measurements on a microarray give an expression profile for the conditions. By comparing expression profiles across experiments, it is often possible to understand the relationship between gene expression and external factors. Hundreds of such experiments have uncovered biologically-relevant patterns in gene expression. Gene expression information has been used, for example, to classify tissue types [8, 49] or differentiate between different forms of cancer [5, 26, 42, 80, 93, 101]. Additionally, such expression information often reveals information about the biological underpinnings of different conditions [4, 8, 51, 64, 76]. Section 2 presents this technology and some of the more relevant quantitative aspects of the data.

### ***Section 3: Cluster Analysis of Microarray Experiments***

Finding biologically-relevant patterns in the enormous quantity of data, which represents millions of measurements in large experiments, requires large-scale data mining. This thesis is concerned with cluster analysis, a statistical tool widely used in fields ranging from economics and marketing [13, 45] to archaeology and signal processing[32]. Cluster analysis is commonly used in bioinformatics to discover

groups of similar genes, tissues and patients. However, many different algorithms for cluster analysis exist, and a standard technique has yet to emerge. In section 3, I present the algorithms most commonly used in microarray experiments, and describe major differences between algorithms. I also mention factors that lead to indeterminacy in the clustering process. An extensive discussion of these factors can be found in Appendix A.

#### ***Section 4: Methods***

The methodology developed in section 4 is designed to overcome the indeterminacy in the clustering process. It is assumed that the reader is familiar with the extent and sources of such variability. If this is not the case the author recommends reading Appendix A prior to reading section 4.

The approach taken here is to use the clusters generated by *many* algorithms to infer biological information. The central question addressed by this thesis can thus be stated as follows: *How can the results of different forms of cluster analysis be synthesized to produce biologically-relevant information?*

The first step is to use multiple clustering configurations to produce a large database of clusters. This collection of clusters is a robust source of information; the clusters reflect natural groupings in the data. However, a methodology does not currently exist to interpret this collection. In section 4, three mathematical techniques are developed to conduct exploratory data analysis on the collection of clusters. Just as existing statistical methods are used to mine ‘normal’ expression data for patterns, these tools can be used to mine the clusters for patterns. By combining the results of

many forms of cluster analysis, then identifying consistent and dominant patterns, I hope to extract meaningful information from the underlying biological data.

### ***Section 5: Results***

The methodology developed in section 4 represents a different approach to cluster analysis than is commonly used; for this reason many of the results described in section 5 are diagnostic in nature. Using random and real data, it is demonstrated that each tool discovers information consistent with known properties of both datasets. Being thus validated, the tools are used to infer new information about the general consistency and prevalence of patterns across different clustering algorithms.

In particular, one of the tools (the *prevalence* statistic) is able to identify patterns in leukemia microarray data that correspond to clinical differences between patients. Using this unsupervised technique, a single microarray is identified as being vastly different from the other microarrays in the experiment, and four groups of patients are discovered. The singled-out microarray, using widely-accepted standards, is found to have been the only defective microarray in the experiment. Checking the clinical information on the groups of patients reveals that they almost exactly match known subtypes of acute leukemia.

### ***Section 6: Conclusions***

In section 6, the ideas of the thesis are presented within the context of the field of bioinformatics. I conclude that much insight can be gained by mining clusters for

knowledge, and give specific suggestion for how to further extend the methods of this thesis.

### ***Appendix A: Sources of Variability and Indeterminacy in the Clustering Process***

Referring to the current state of microarray analysis, D’Haeseleer recently observed, “the field is in dire need of a comparison study of the main combinations for some of the standard applications.”[24] The methodology developed in this thesis is designed to overcome the inherent uncertainty involved in clustering. It assumes such variability exists, and is predicated on the assumption that there is no single best technique, that different techniques are informative in different ways. In Appendix A I substantiate both of these assumptions by demonstrating that the clusters produced by any algorithm are extremely sensitive to:

- A.1 The choice of algorithm
- A.2 The parameters passed to the algorithm
- A.3 The input data used as a basis for clustering.

### ***Appendix B: Implementation***

Computational concerns can impede effective cluster analysis of microarray data[10, 57]. The methodology presented in this thesis presents a significant computational challenge. Appendix B gives an outline of the author’s implementation of all statistical techniques and data storage schema.

To allow for the use of the “brute force” iteration over multiple configurations of multiple clustering algorithms, data at every stage of clustering is retained in a database. This frees memory for fast paging and optimizing algorithm performance, while allowing for hundreds of concurrent experiments. Statistical programming is done in Matlab, a mathematical programming language optimized for fast manipulation of arrays and matrices. As the complete project consists of over 3000 lines of code, only crucial sections are included in Appendix B. If code for a particular algorithm or statistic is included, it will be marked with a <sup>\*code</sup> in the text.