COMPUTATIONAL GENOMICS AND BIOENERGY: MODELING AND CLUSTERING OF RNA-SEQ DATA

by

JAMES WOODWARD WEIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF

BACHELOR OF SCIENCE IN COMPUTATIONAL BIOLOGY
WITH HONORS IN COMPUTER SCIENCE

at

BROWN UNIVERSITY

May 2012

Author .....
Department of Computer Science
May 3rd, 2012

Certified by .....
Sorin Istrail, Ph.D.
Professor of Computer Science
Thesis Supervisor

Certified by .....
Casey Dunn, Ph.D.
Assistant Professor of Biology
Thesis Supervisor

Accepted by .....
Tom Doeppner, Ph.D.
Director of Undergraduate Studies

Computational Genomics and Bioenergy: Modeling and Clustering of RNA-Seq Data

by

James Woodward Weis

Submitted to the Department of Computer Science
on May 3rd, 2012, in partial fulfillment of the
requirements for the degree of
Bachelor of Science in Computational Biology
with Honors in Computer Science

Abstract

In this thesis, I implement a program to simulate RNA-Seq datasets by sampling from a model based on the negative binomial distribution. I then describe and implement three popular clustering algorithms, along with some minor improvements, for use with this generated RNA-Seq data. Finally, I discuss three metrics of cluster goodness and apply these metrics to clusters produced by the discussed algorithms for a variety of different parameter settings.

Thesis Supervisor: Sorin Istrail, Ph.D.
Title: Professor of Computer Science

Thesis Supervisor: Casey Dunn, Ph.D.
Title: Assistant Professor of Biology

Acknowledgments

The development of this thesis would have been impossible without the support of several key individuals. First, I would like to thank Professor Sorin Istrail for his invaluable guidance and support over the past year.

I would like to thank Professor Casey Dunn for agreeing to be my second reader—his readership and review found important areas for improvement within my original drafts, and has had a strong impact on this final version.

I greatly appreciate the hospitality of Professor Jeremy Smith for inviting me to visit his Department of Molecular Biophysics at Oak Ridge National Laboratories, greatly expanding my understanding of the role computer science must play within the biofuel problem.

I would like to thank Professor Tom Doeppner of the Computer Science department for his time and review of my thesis materials. Finally, I extend gratitude to the whole of the Istrail Lab, my friends, and family for their review and help over the past year.

Thank you!

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Research Overview

RNA-Seq, or "Whole Transcriptome Shotgun Sequencing," is a relatively new tool for transcriptomics that involves the use of deep sequencing technology. Studies using RNA-Seq have already changed our understanding of the transcriptome [41]. Although RNA-Seq provides a method of *expression profiling* that is far more precise than other methods currently available [41], it also has a new suite of interesting bioinformatics problems that must be solved before it is able to entirely reach its promise. As an example, an RNA-Seq experiment necessitates the mapping of RNA reads back to a reference genome, or alternatively, their assembly into contigs, to reveal transcriptome structure. However, these reads are not uniformly sampled from the transcriptome, and often contain extensive alternative and *trans*-splicing, making the problem much more difficult.

1.1 Problems Investigated

This thesis focuses on one of the problems encountered during RNA-Seq experiments—namely, the development of algorithms to cluster genes that exhibit similar expression patterns. Because of the challenges and complexity of real-life RNA-Seq datasets, this is a very difficult problem and an active area of research [35].

1.1.1 The Simulation of RNA-Seq Experiments

In order to test and eventually develop transcriptome clustering algorithms, it is helpful to have a method of simulating RNA-Seq datasets. With such a tool, these clustering algorithms may be evaluated over a wide range of possible RNA-Seq experiments. Further, using such a tool makes it possible to know the "true" clusters from a simulated dataset, consequently making the analysis of these clustering algorithms much more straightforward and accurate.

The first work accomplished in this thesis is the development of a program that is capable of generating simulated RNA-Seq datasets. While based on a model that has been shown to fit RNA-Seq data well, my program does not intend to exactly simulate biological datasets—at this point the level of complexity within such datasets make them very difficult to simulate and well as analyze. Instead, I aim to produce simplified datasets for which the correct clustering may be feasibly obtained. This can be seen as a first step towards the analysis of real-life datasets.

1.1.2 Clustering Algorithms for RNA-Seq Data

The second work in this thesis is the implementation of algorithms to cluster RNA-Seq datasets. The algorithms implemented are a slightly altered versions of Lloyd's algorithm (k-means), the unweighted pair group method with arithmetic mean, and a self-organizing map.

Finally, the performance of these algorithms on the simulate dataset is analyzed with the use of three quantitative methods of cluster goodness: the adjusted Rand index, normalized mutual information, and the Silhouette validation metric. This analysis shows that the implemented and slightly altered version of k-means, along with the self-organizing map, perform best under almost all experiment parameterizations. Interestingly, the hierarchical clustering algorithm performs strictly worse than both other algorithms in the vast majority of simulations.

It must be noted that these results depend on the accuracy of the simulated data, which is a greatly simplified version of reality. However, this work comprises the

beginning of the development of a framework that may be used in the future for the realistic simulation of RNA-Seq data, as well as the validation, selection, and development of RNA-Seq clustering algorithms.

Chapter 2

Research Background

2.1 Relevant Molecular Biology

A review of some relevant biology is helpful to understand the models and algorithms contained within the remainder of this thesis. The following is a very brief introduction to the genome and transcriptome, as well as common methods of transcriptome quantification. This chapter is by no means a biological treatise; biology abounds with details and exceptions, most of which are abstracted away for the purpose of simplicity, feasibility, clarity, and relevance to the problem at hand.

2.1.1 The Genome and Transcriptome

The *genome* contains an organism's hereditary information in the form of DNA. The genome can be understood as a set of *chromosomes*, $G = \{G_1, G_2, ..., G_c\}$ where c is the number of chromosomes. For the human genome, $c = 23$. Each G_i is a sequence of *coding* and *non-coding regions* such that $G_i = (C^i_{0,a}, C^i_{a,b}, ..., C^i_{y,n})$. A coding or non-coding region C^i_{xy} is the substring of chromosome i from symbol x to symbol y

To become RNA, the region C^i_{xy} must first undergo transcription, $C^i_{xy} \rightarrow R^i_{xy}$, which produces the RNA complement of coding sequence C^i_{xy}. RNA is a string over alphabet $\Sigma_{RNA} = \{A, C, G, U\}$ and transcription is symbol-by-symbol bijective mapping from $\Sigma_{DNA} \rightarrow \Sigma_{RNA}$.

RNA that codes for a protein is called $mRNA$. However, RNA can have a variety of other biological functions. For example, $tRNA$ are molecules that are used during translation, $rRNA$ forms part of the ribosome, the protein synthesis enzyme, and $snRNA$ is involved in a variety of biological processes such as RNA splicing, transcriptional regulation, and telomere maintenance.

Transcripted coding regions are then translated to proteins, which are biochemical compounds consisting of folded *polypeptides* that have some biological function. Polypeptides are the strings of *amino acids* that are the result of translation, a mapping from $mRNA$ to protein.

The transcriptome at sample time i, T_i, is the set of all RNA molecules present in a set of cells, along with their count. If we define $N^i_{x,y}$ as the number of copies of $R^i_{x,y}$ present in the cell population, then the transcriptome is

$$T_i = \{(R^i_{x,y}, N^i_{x,y}) \mid C^i_{x,y} \in G_i\}.$$

Unlike the genome, which is mostly invariant over time, the transcriptome is a function of external conditions, and thus reflects the genes or coding regions that are actively expressed or transcribed.

2.1.2 Expression Profiling using RNA-Seq

Transcriptomics, also called *expression profiling*, is the attempt to discover information about how T changes over time and with new environments. Given k experiments, samples, or time points, expression profiling seeks to discover

$$T = \{T_i \mid 0 \leq i \leq k\}.$$

Expression profiling has been historically most often accomplished through the use of *microarray* technology. However, expression profiling using next-generation sequencing technology, or RNA sequencing, is known as RNA-Seq.

RNA-Seq is the next-generation of expression profiling tools. RNA-Seq provides

researchers with a way to experimentally quantify the transcriptome, and offers many advantages over DNA microarray technologies, as well as a new set of difficulties.

In a typical RNA-Seq experiment a population of RNA is converted into a library of cDNA fragments with adapters. Each cDNA molecule is then sequenced (occasionally after amplification) using next-generation sequencing technology to obtain sequences from the ends of the molecule inwards, typically 30 to 400 base pairs. After sequencing, these reads are maps to a reference genome or assembled *de novo* to produce a genetic map of the transcriptome, consisting both of the transcriptome structure, as well as some function of the level of expression for each gene.

	454 Sequencing	Illumina	SOLiD
Sequence chemistry	Pyrosequencing	Sequence-by-synthesis	Ligation-based
Amplification approach	Emulsion PCR	Bridge amplif.	Emulsion PCR
Paired end separation	3 kb	200bp	3 kb
Mb per run	100 Mb	300 Gb	3,000 Mb
Time per paired end run	7 hours	7-14 days	5 days
Read length	250 bp	100 bp	35 bp
Cost per run	$8,438 USD	$11,750 USD	$17,447 USD
Cost per Mb	$84.39 USD	$1.00 USD	$5.81 USD

Table 2.1: A comparison of the most popular next-generation DNA sequencers used for RNA-Seq experiments.

While RNA-Seq is still a developing technology, it presents some key advantages over other methods of transcriptome profiling. First among these is that RNA-Seq is capable of assembling transcriptomes for which there exists no existing genomic sequence or reference genome. Second, RNA-Seq has much lower background signal than DNA microarrays [41]. This is because each DNA sequence can be mapped directly to unique regions of the genome, and there is no upper limit for the number of reads that can be mapped in this way. Finally, because it does not any cloning steps, an RNA-Seq experiment requires less DNA than a corresponding DNA microarray experiment [41].

Due to its status as a new technology, RNA-Seq also has an interesting array of new challenges to confront. First, an RNA-Seq experiment requires the building

of a cDNA library, the necessary manipulations of which complicate the analysis of the experiment results. For example, amplification can introduce PCR artifacts that are difficult to distinguish from small RNA reads. Second, RNA-Seq experiments involve the same bioinformatic problems that are encountered in high-throughput sequencing experiments, including the necessity for efficient methods of manipulating large datasets, mapping the reads to a reference genome or assembling them into contigs. Further complicating the matter are problems such as alternative splicing and polymorphisms.

2.2 Motivations for Research

People have relied on biological sources of energy for the majority of human history, from wood combustion to muscle contractions. The switch to so-called non-renewable energy sources has occurred relatively recently. This can be seen in figure 2-1, produced using historical data from the United States Energy Information Association.

The increase in energy-hungry technological development, as well as a growing global population, has vastly increased the worlds energy requirements. As expected by economic theory, the rapid increase in use of non-renewable energy sources, such as coal, petroleum, natural gas, and nuclear materials, has resulted in a corresponding decrease in supply and increase in cost. This drop in supply is coupled with a decrease in energy security resulting from regional conflicts, as well as alarming environmental devastation and greenhouse gas emissions (GHG) [23]. The net effect is a rising cost of energy, despite arguably exponential increases in energy-enabling technology; in 1970, energy cost, on average, \$1.65[1]per million British thermal units (Btus). By 2008, that value had risen to \$12.93.

There are several possible outcomes from the currently unstable energy situation. One one extreme are Malthusian catastrophe-esque scenarios[2] On the other extreme

[1]This value includes taxes and is not adjusted for inflation; in 2008 dollars, this would be \$9.05. As mentioned above, this becomes especially significant when one considers the exponential pace of technological development, which would be expected to exponentially *decrease* the cost of energy.

[2]First formulated by Thomas Malthus in 1798 in *An Essay on the Principle of Population*, a

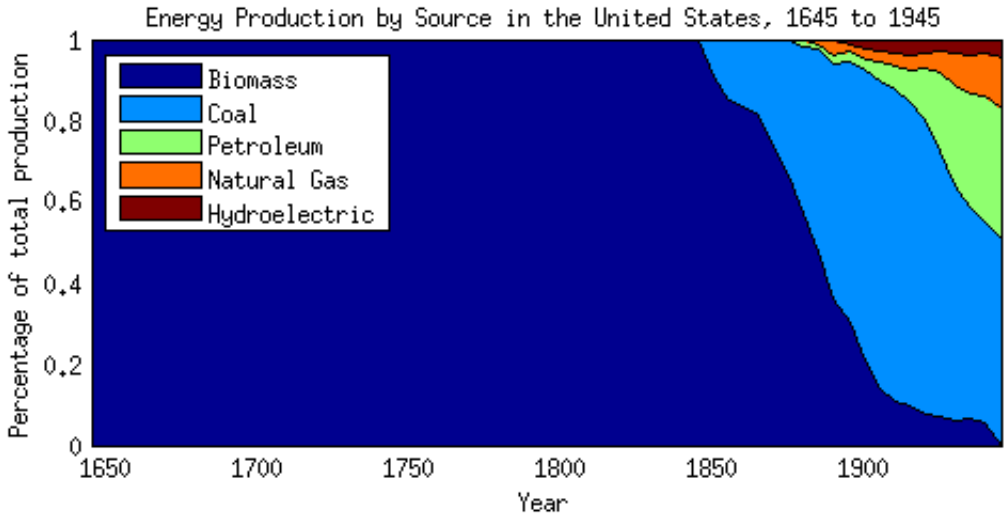

Figure 2-1: The relative contribution of energy sources to total energy production in the United States, from 1645 to 1945.

is the development of an energy source that approaches a perpetual motion machine[3]; a clean, renewable, inexpensive, and efficient energy source.

While renewable energy current makes up only 8% of total energy production, in the future that number is expected to rise considerable [39]. The United Nations Directive on Renewable Energy has mandated that by 2020, 20% of all energy produced by member states be from a renewable source. Even the United States Air Force and Department of Defense have began programs to sharply reduce their dependencies

Malthusian catastrophe is the result of a rate of population growth that outpaces agricultural development. This disparity would force a return to subsistence farming lifestyle, further aggravated by the effect of population growth on the environment. This theory is especially interesting and relevant to this thesis as it can be generalized to apply to energy, rather than agricultural, development. In this case, the rate of population growth supersedes the rate of advancements in energy production, resulting in a return to basic energy sources as the energy demands of the growing population far outweigh the energy generating potential of even rapidly developing energy generating technologies.

[3]A perpetual motion machine describes a hypothetical mechanism that generates more energy that it consumes, thus producing an infinite source of energy with no inputs or side effects. The development of such a machine would violate the first law of thermodynamics and is impossible, given current understandings of physics. An interesting and relevant thought experiment popularized by Richard Feynman is as follows: Consider a paddle connected to a ratchet, so that its rotation is limited to a single direction. Brownian motion—random particle movements—would cause the surrounding molecules to strike the paddle in a random fashion, but the ratchet would allow it to turn in only one direction [30]. The fallacy in this experiment quickly becomes obvious. For criticism of this theoretical machine, see Parrondo and Espanol (1996).

on foreign oil through research in oil alternatives [9, 15]. The development of new sources of biological energy holds enormous potential in helping meet these goals and providing a clean, carbon-neutral, environmentally-friendly, and inexpensive energy source.

2.2.1 Bioenergy

In this thesis, bioenergy is defined as energy produced from relatively recent, biologically-induced carbon fixation. Note that while fossil fuels are the consequence of biological carbon fixation, they are not considered biofuels because they are considered to have been "out" of the carbon cycle for a long time and also because their rate of accumulation is inconsequential when compared with their rate of usage. Biofuel, while most often used to refer to liquid bioenergy, can be considered a synonym of bioenergy, and is used as such in this thesis. By this definition, the set of energy sources that may be considered biofuel is very large, including, for example, solid biomass such as switchgrass and wood, liquid fuels such as bioethanol and biodiesel, and even biogasses[4].

First-generation biofuels are defined as those produced from a specific subset of bioenergy sources: sugar, starch, and vegetable oil [23]. These biofuels are in relatively common use worldwide, especially in Europe and Brazil. However, while first-generation biofuels offer certain advantages, their production and deployment has been hindered by a poor distribution infrastructure, low consumer demand, an unavailability of compatible vehicles, and a lack of a coordinated expansion plan [15]. While there are many issues and disadvantages with biofuels in general, first-generation biofuels often offer so many disadvantages so as to provide zero or even negative net utility. [15]. One such issue is the energy sources for first-generation

[4]Bioethanol refers to the alcohol produced by fermentation of carbohydrates produced in a sugar or starch crop. Biodisel refers to a biofuel made from animal fat and vegetable oils through transesterification, which is often used as a diesel additive, although it may also be utilized in a pure form. Biodiesel is the most common biofuel in Europe. Biogas refers to a gas formed during the anaerobid fermentation of organic matter. The most prevalent biogasses are methane, CH_4, carbon dioxide, CO_2, and hydrogen sulphide, H_2S.

biofuels, which are largely grain-based to allow for easy extraction and fermentation of simple sugars. However, such energy sources are themselves costly to produce, and often compete for resources with the production of foodstuffs such as corn [8]. Economically, the production of first-generation biofuels often relies largely on governmental subsidies, and in most cases is limited by a production threshold above which further biofuel production becomes threatening to food supplies, the environment, and biodiversity. Further, the amount of carbon emissions produced by the usage of these biofuels, including production and transportation, often approaches or even exceeds that of traditional fossil fuels [13].

Second-generation biofuels are the latest attempts to produce clean, efficient energy by harnessing the energy in waste and byproduct materials and non-food biomass. In this thesis, second-generation biofuels are defined as fuels produced from sustainable feed-stock, where sustainability is a function of factors including availability, GHG emissions, biodiversity impact, and land use. An especially large amount of research and funding has gone towards the use of *lignocellulosic biomass*, defined as biomass composed of cellulose, hemicellulose, and lignin [7]. While second-generation biofuels are often much more promising than their first-generation counterparts, they too suffer from a host of problems that must also be considered.

2.2.2 The Upsides of Biofuels

There are economic incentives driving the adaptation of second-generation bioenergy. On a macro scale, these include a $23.1 billion increase in United States gross domestic product (GDP), the creation of over 163,000 new jobs, a $6.7 billion increase in mean household income, and tax revenue increases of $2.7 billion and $2.2 billion for the federal and state governments, respectively [14].

This economic incentive has spurred interest and billion of dollars in funding in the public and private sectors. These investments include the formation of companies such as Amyris, Solazyme, Gevo, and KiOR[5], all of which have attracted international venture capital and media attention, although their success is, so far, very debatable.

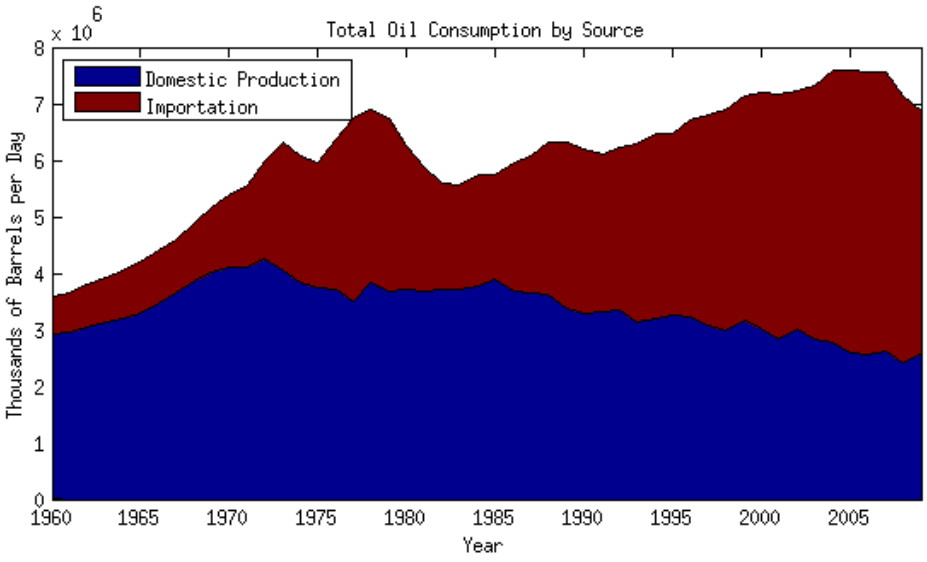

Figure 2-2: Total oil consumption in the United States by original source. Data from U.S. Energy Information Association (2011).

There also exist political advantages in the adoption of biofuels, primarily due to a potentially decreased reliance on imported oil. Many nations are highly dependent on imported energy, and the United States is the top importer of petroleum in the world. While foreign oil dependence peaked in 2005 and has since decreased, see figure 2-2, the United States remains highly dependent on foreign oil, consuming almost 7 billion barrels a day, of which 49% is imported [39].

Further, as is shown in figure 2-3, much of this oil is imported from nations that are either unstable or are located in troublesome regions of the world. As a result, securing production requires substantial U.S. investment of both economic and political capital. Further, this reliance on importation results in price volatility, decreasing net U.S. efficiency and GDP. Biofuels, as well as any other type of domestic energy production, are a potential route to ameliorating this dependence.

[5] Amyris and Solazyme were both founded in 2003 and have market capitalization of $170 million and $687 million, respectively. Gevo was founded in 2005 and has a current market capitalization of $243 million, and KiOR was founded in 2007 and has market capitalization of $1.11 billion [29].

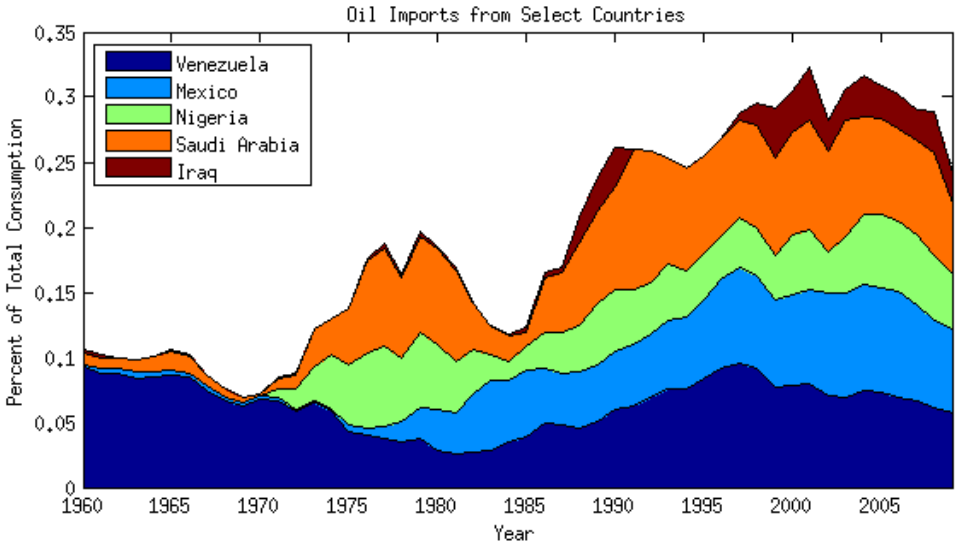

Figure 2-3: Oil imports from select countries as a percentage of total U.S. consumption. Data from the U.S. Energy Information Association (2011).

2.2.3 The Downsides of Biofuels

One must also, however, explore the potential downsides of bioenergy, of which there are many. When considering the billions of dollars in funding that is going into bioenergy research in both the public and private sectors, these various disadvantages become very serious. Addressing these problems is equally as important a research topic as enabling the technology in the first place. Some of the key issues with the production and use of biofuels include competition with food production, the often high carbon emissions associated with the production and transportation of biofuels, deforestation, and pollution [21].

A serious problem with the development of biofuels that rely on agriculture for their feedstock production is the risk of diverting farmland from food production. This introduces competition for resources, driving up the prices of the food stables that are typically produced on such land [3]. In fact, it has been shown that the expansion of the use of ethanol as a fuel in the United States increased maize prices by 21% in 2009, in comparison with what they would have been had ethanol stayed at 2004 levels of production [3]. Another, even more recent study has shown the production of biofuels and their subsidies to be the leading causes of shocks in the

agricultural market. Perhaps the most negative outcome from this conflict between food and fuel is the decrease in the amount of food that can be purchased by the most impoverished and at-risk sector of society. The counter-argument is that new, second-generation biofuels use parts of corn or other plant matters that are not suitable for human consumption, and that this competition for land use only exists in relation to first-generation biofuels.

While the goal of biofuel research is the creation of an energy source that is carbon negative, or at least carbon neutral, it has been observed that some biofuels fall far short from this goal, sometimes producing as much or even more carbon emissions than their corresponding fossil fuels. When carbon emissions are calculated using a *life cycle analysis* (LCA), all carbon emitted during the production of biofuels are summed. This includes everything from tractor emissions during planting to transportation costs. Many such studies have been done for many different types of biofuels, with varying results (see figure 2-4) [40, 14, 37]. One interesting, hidden aspect of carbon emissions discussed by a study from Princeton University is the fixed carbon released during the clearing of new land for the production of biofuels [37]. While these issues are well characterized for first-generation biofuels, accepted LCAs for second-generation biofuels are still to be produced.

Another, related criticism of biofuels is that they result in large-scale deforestation of old, mature trees. These trees are then usually replaced with sugar cane or some other feedstock crop. However, these crops have a much lower capacity for CO_2 removal through photosynthesis than the large trees they are replacing. This also contributes to high atmospheric levels of greenhouse gasses and, consequently, to global warming and a reduction in biodiversity both on land and sea [31]. Further, the removal of cellulosic biomass for biofuel production, which is called for in second-generation biofuels, is claimed to deplete soils of their necessary nutrients.

The effect of biofuels on chemical pollution has also been studied, to alarming results. The chemicals that have received the most attention in this area are the *aldehydes*, especially formaldehyde and acetaldehyde. Most aldehydes are toxic to living cells—formaldehyde cross-links amino acids and has been banned by the Eu-

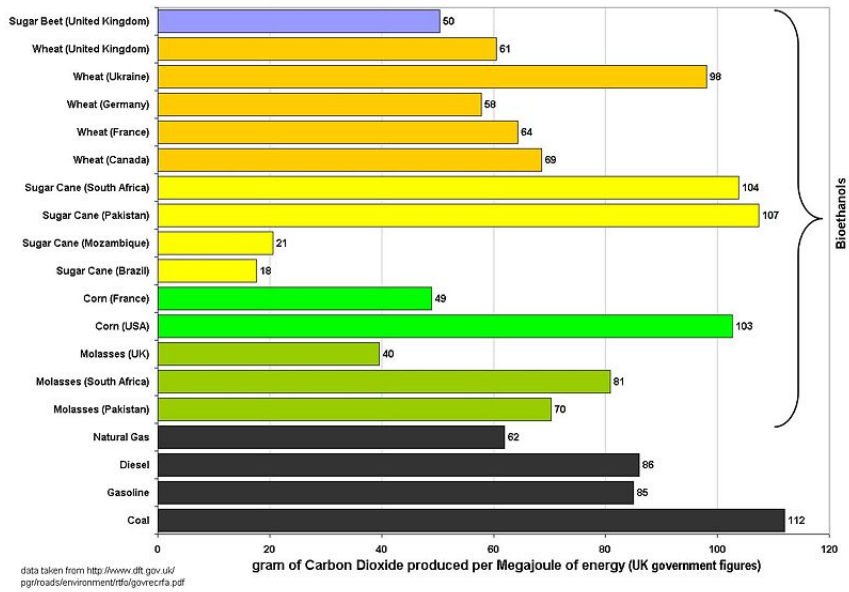

Figure 2-4: Carbon emissions per Megajoule of energy produced for various biofuel sources. From *Carbon and Sustainability Reporting Within the Renewable Transport Fuel Obligation*.

ropean Union, while acetaldehyde is carcinogenic and mutagenic. Alarmingly, only a 10% mixture of ethanol in gasoline (as in E10 gasohol) increases aldehyde emissions by 40% [17]. Studies using gas chromatography in São Paulo, Brazil, which uses ethanol fuel, and Osaka Japan, which does not use ethanol fuel, were performed to measure atmospheric levels of aldehydes [6]. It was found that atmospheric formaldehyde was 160% higher in Brazil, while acetaldehyde was 260% higher. However, these findings are controversial, and some claim that the lowered sulfer content of biofuels actually lowers net acetaldehyde levels [17].

Clearly, while biofuels have great promise, they also hold great controversy. It is likely that the portrayal of first or even second-generation biofuels as the "holy grail" of renewable energy are far overblown. It is also likely that the more alarming of the anti-bioenergy reports are also exaggerated. In any case, there is certainly a need for further research and study to reduce the downsides of biofuels, improving their efficiency and making them a more viable alternative energy source.

2.2.4 The Biofuel Problem

On the most high level, biofuel researchers seek an algorithm to solve the *biofuel problem*, outlined in algorithm 2.1. That is, on a very high level, an algorithm that takes as input a sustainable energy source derived from relatively recent carbon fixation and returns as output a biofuel with the largest amount of energy possible and the minimum number and magnitude of undesirable effects, where undesirable effects include issues such as deforestation, impact on food prices, carbon emissions, decrease in biodiversity, and impact on water resources.

Algorithm 2.1 The biofuel problem

Require: An energy source derived from recent carbon fixation
Ensure: A biofuel with maximum energy and minimum negative side effects

In the popular and specific case of biofuels derived from lignocellulosic biomass, the biofuel problem can be divided into three sub-problems. The first such sub-problem is the design of feedstocks that are optimized for use as a biofuel energy source, where feedstocks are defined as the biomass crop and are generally warm season perennial grasses, fast growing woody grops, and common garden or agricultural waste [18]. A solution to the feedstock design problem is an algorithm that outputs a biofuel energy source, such as sugar cane, with fermentable sugars that are more easily deconstructed. Research in this area seeks to produce plants with optimized production of easily fermentable sugars through genetic engineering and manipulation of the genes and enzymes involved in the creation of lignocellulose. These engineering efforts may also be aimed at improving the plant itself; for example, making it more disease resistant or tolerant of low-nutrient conditions.

The second sub-problem within the biofuel problem is the development and improvements of methods to convert lignocellulosic biomass into fermentable sugars. There are three enzymes that are known to be required for the conversion of cellulose to glucose: endoglucanases, exoglucanases, and cellobiases. However, lignocellulosic biomass is recalcitrant to hydrolysis by these enzymes, partially due to its complicated cell wall matrix shown in figure 2-5, and the direct conversion of this biomass

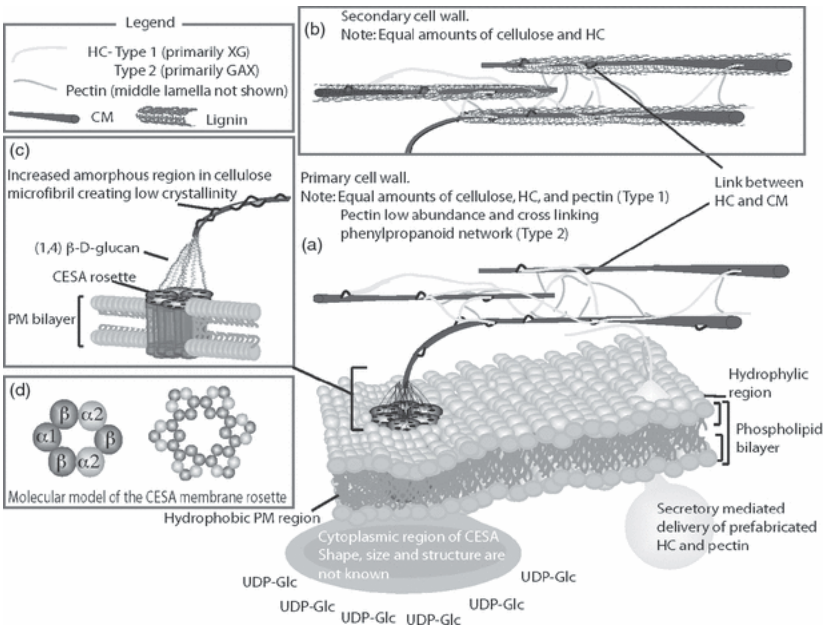

Figure 2-5: A simplified diagram of the plant cell wall matrix structure with an emphasis on cellulose. From Harris (2010).

to fermentable sugars requires a large imput of energy. Thus, feedstocks are often pre-treated with a chemical cocktail to lessen the bonds between the lignin, cellulose, and other polysaccharides, as well as to increase the surface area available for enzyme function [18].

Research in this area is aimed at developing an understanding of the chemical and biological environments that occur during the feedstock deconstruction. This includes the development of optimized enzyme cocktails or novel pretreatment methods [26]. However, perphaps the most active reaserch in this area is devoted to the discover, design, and optimization of deconstruction enzymes. This involves both searching within novel environments, such as tropical rain forests and compost piles, for existing enzymes that excel at the deconstruction of plant biomass[6], as well as optimizing the structure and catalytic efficiency of these enzymes so as to improve efficiency [1, 4, 26].

[6]The search of useful organic compounds, biological or genetic, in this case novel enzymes, is called *bioprospecting*. Often, bioprospecting involves the testing and sequencing organic compounds in microrganisms, plants, and fungi that grow in extreme environments such as hot springs, deserts, rainforests, or even landfills.

As mentioned before, first-generation biofuels are based mostly on the use of starch-based biomasses because the sugars within such biomasses are easily fermentable by microbes such as yeast. On the other hand, the sugars resulting from the deconstruction of lignocellulose are complex and often contain chemicals that prevent their fermentation by yeast. Research in this area involves tools such as *synthetic biology* and mathematical models of cell regulation and metabolism in an attempt to create new microbes that are optimized for the fermentation of these more complex sugars [22].

Work in this area requires a very strong understanding not only of genetic structure of various microbes, but also the ability to manipulate microbe metabolism to increase overall biofuel yield, as well as dealing with the side effects of such alterations. For example, the creation of microbes to convert biomass to ethanol requires microbes with unnaturally high levels of ethanol tolerance. This can be resolved through knowledge and of a microbe's natural pathways, which may lead to genetic engineering to artificially regulate networks related to ethanol tolerance in a new way, or proteomic work to increase or decrease the efficiency of enzymes related in ethanol tolerance, such as dehydrogenases [4].

2.2.5 RNA-Seq Clustering

As mentioned earlier, many governments including the United States have funded research in the development and implementation of biofuels. The U.S. Department of Energy has established three Bioenergy Research Centers (BRCs) in 2007. The goal of these centers is to investigate high-risk, high-return biological solutions for the biofuel problem, with research categorized under the three sub-problems outlined above: the development of next-generation bioenergy crops, the development of novel enzymes and microbes with exceptional biomass-degrading abilities, and the creation of microbe-based biofuel production strategies. These BRCs are the Joint BioEnergy Institute in Emeryville, California, the Great Lakes Bioenergy Research Center in Madison, Wisconsin, and the BioEnergy Science Center at Oak Ridge National Laboratories in Oak Ridge, Tennessee.

I was given the opportunity to visit one of these BRCs, the BioEnergy Science Center at Oak Ridge National Laboratories in Tennessee. While there, I visited Jeremy Smith, Ph.D., the Director of the Department of Molecular Biophysics, to better understand the role that computational genomics must play in the solving of the biofuel problem.

Much of the research in the Department of Molecular Biophysics is in structural simulations of protein interactions using physical models and the massive computing power available at Oak Ridge National Laboratories. These simulations are often focused on modeling enzymes involved in the biofuel problem, particularly those involved in biomass construction. However, it is sometimes helpful to understand the role of these enzymes in terms of the organisms transcriptional structure—what type of expression pattern it contains, and what other expressed regions contain a similar expression pattern [36].

To this end, there is a need for computational algorithms that find regions with similar *changes in expression* over different time points. Such algorithms would take as input a set of genes and regulatory information about each gene, and output the genes, partitioned into groups such that each group shares regulatory behavior. These groups then imply correlation of regulation. For example, if genes known to be involved in the cellulose degradation pathway of some microbe are outputted within a partition with other genes, then those genes would merit further attention and inquiry, such as biophysical modeling or biological experimentation to determine their function. While such clustering does not imply causation, with sufficient data points and a sufficient number of treatments, the associations will likely prove interesting, and often do reflect true biological relationships [36, 11, 43, 42].

Of special interest is the clustering of variability in gene expression among biological samples from a single group, rather than simple mean expression levels [43]. Some notable investigations that utilized clustering analysis on transcriptome data are outlined below.

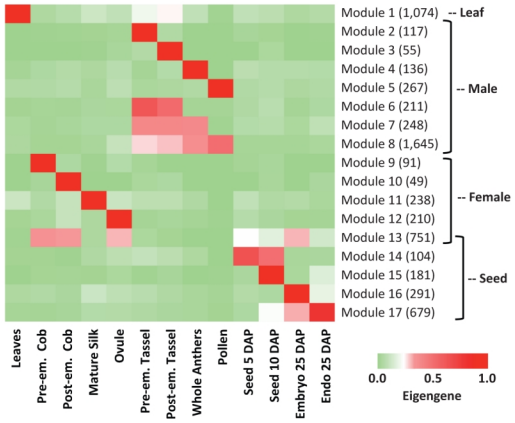

Figure 2-6: Heat map of the eigengenes representing each gene module. The columns represent tissue samples, and the rows represent eigengenes.

The maize reproductive transcriptome

Maize (*Zea mays L.*) is a vital worldwide crop, with over 36 billion hectares grown in the United States alone in 2010 [11]. Maize makes an interesting species to study as it is used for food, animal feed, and also biofuel production. As demand for all three of these sources increase, genetic engineering will be necessary to improve the currently available species. This, in turn, requires the use of computational algorithms to understand the maize genes and phenotypes.

Davidson et al. (2011) describes attempts to elucidate the functioning of some maize genes involved in reproduction through the use of clustering algorithms. As a result, see figure 2-6, modules of genes were found that were expressed in specific stages of maize development. Also discovered were sets of genes whose expression varied across male, female, or seed maize tissue. These analyses, combined with others within the same paper, offer a set of genes as candidates for further research into maize productive development and grain yield potential.

Functional transcriptome organization in the human brain

Microarray and, more recently, RNA sequencing technologies have been used to discover transcriptome organization across tissues, and also across species. Organization of the transcriptome typically involves creating several groups of genes, or gene

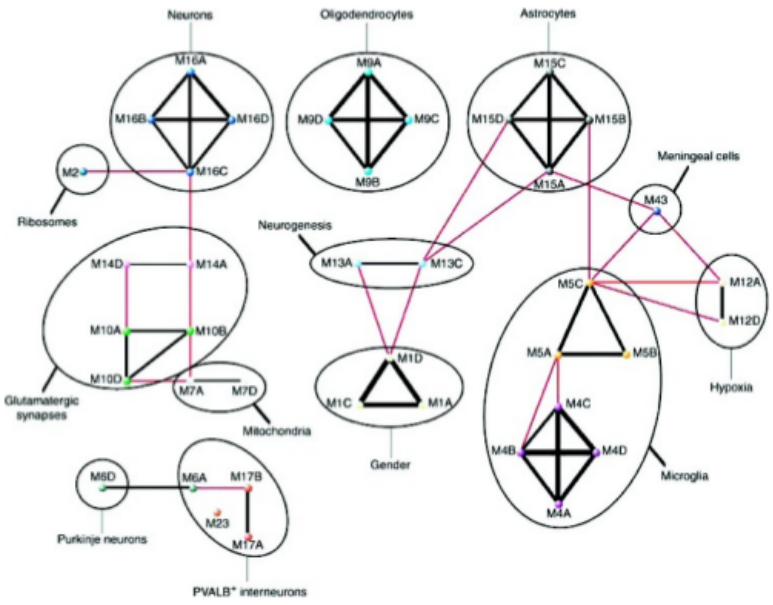

Figure 2-7: Meta-network of co-expressed gene modules in the human brain.

modules, whose expression levels are correlated across samples. Finding these relationships offers insight into the transcriptional regulation represented in the genome. Further, if these gene modules are capable of being reproduced across samples and each module is able to be organized into a clear functional category, then it is possible to use this clustering to elucidate the underlying organizational structure of the tissue being samples [42].

Winden et al. applied a type of cluster analysis to transcriptome data sets from human brain samples. Modules of co-expressed genes were identified via hierarchical clustering, and gene modules representing different brain tissues and cell types were identified. Using this information (see figure 2-7). Winden et al. were able to annotate novel genes based on their correlation with eigengenes from specific modules, and some genes were identified that had no previous report of neuronal function.

Drug targets in oncogenic signaling networks

One of the most lethal of all cancers, and also the most common primary malignant adult brain tumor, is glioblastoma [19]. New methodologies of treatment are drastically needed, and it has been shown that inhibition of the epidermal growth factor

receptor, which is commonly over-expressed in glioblastoma, promoted significant response in some glioblastoma patients. The next step in development a treatment was identifying possible drug targets.

Horvath et al. (2006) applied cluster analysis to identify a drug target. For this study, two datasets were analyzed using cluster analysis. The results showed highly conserved gene modules. Another dataset of breast cancer samples was introduced and analyzed, and all three samples showed a high degree of gene module preservation. The module representing genes involved in cell cycle and mitosis was selected, and the most highly connected genes in that module were investigated. The ASPM gene stood out because it has previously been reported as highly expressed in other types of cancers and because it is typically expressed at very low levels in most cells under the majority of conditions. Subsequent experiments showed ASPM-knockout cells had a strong, specific inhibition of proliferation.

Comparison of human and chimpanzee brains

Humans and chimpanzees are known to share a remarkable number of genes related to brain function. Using eighteen human and eighteen chimpanzee samples, Oldham et al. (2006) identified modules of similarly expressed genes in both samples. Upon analysis, it was found that several of these modules were highly conserved, and several were not. Further, it was found that the percentage of human-only connections in each module relate known evolutionary hierarchies. Many of these connections converged to the same hub genes. Gene expression patterns were found to have particularly strong differences in the cerebral cortex, which is consistent with rapid cerebral cortex expansion after human evolutionary divergence from chimpanzees.

This study was later expanded on, and the modules from an eigengene network analysis were found to have good agreement with human-only modules, which is to be expected as most human modules are preserved in chimpanzees. Overall, these studies resulted in novel insights into the differences between human and chimpanzee brain transcriptome arrangement.

Chapter 3

Modeling and Simulating RNA-Seq Data

Model selection is one of the most basic methods of scientific inquiry, and selection of a good model—where good is defined as a qualitative balance between simplicity and goodness of fit—is of paramount importance. In past research, RNA-Seq data has been modeled using the Poisson [5, 28] or negative binomial [33] distributions. While neither model is complicated, the goodness of fit of each model varies according to the dataset. The Poisson distribution is appropriate when technical replicates are used [28, 5], while the negative binomial is a better fit for datasets containing biological replicates [2]. This can be understood intuitively; biological replicates will likely introduce greater variability than is allowed for by the Poisson distribution (this has been called the "over-dispersion phenomenon" [2]). A negative binomial model that allows for over-dispersion has been several times applied to RNA-Seq data analysis [33, 2, 35]. Herein, I will describe a model for RNA-Seq datasets with biological replicates that is based on the negative binomial distribution and follows the work of Si and Liu (2011). I then implement and use this model to generate synthetic datasets which will be used to analyze the performance of clustering algorithms of use in the biofuel problem.

3.1 Modeling RNA-Seq Data

RNA-Seq data can be extremely high dimensional and complex. In the model that follows, I opt for simplicity and clarity over biological accuracy. As this simulation and subsequent analysis is the first step towards more complex models and algorithms for RNA-Seq data, the successful implementation and testing of this model is a landmark on the way to biological accuracy.

3.1.1 Negative binomial distribution

Consider a sequence of Bernoulli trials, parameterized by the probability of success p. The distribution of the number of successes before r failures follows a negative binomial distribution. This number of successes, Y, defines a negative binomial random variable, which is notated $Y \sim \mathrm{NB}(r, p)$. For $k \in \mathbb{N}_1$, the negative binomial distribution has the probability mass function

$$f(y \mid r, p) = \binom{y + r - 1}{y} (1 - p)^r \, p^y.$$

If instead of r and p the known parameters are the mean, λ, and dispersion, ϕ, an alternative expression is necessary. Because this is the case in the following model, I follow instead the alternative parametrization of Robinson and Smyth (2008). Let Y be a negative binomial random variable with mean λ and dispersion ϕ. The probability mass function becomes

$$f(y \mid \lambda, \phi) = \frac{\Gamma\left(y + \phi^{-1}\right)}{\Gamma\left(\phi^{-1}\right) \Gamma\left(y + 1\right)} \left(\frac{1}{1 + \lambda\phi}\right)^{\phi^{-1}} \left(\frac{\lambda}{\phi^{-1} + \lambda}\right)^{y},$$

giving $\mathrm{E}\left[Y\right] = \lambda$ and $\mathrm{Var}\left[Y\right] = \lambda + \phi\lambda^2$. Given the standard parameters p and r, it can be derived that

$$p = \frac{\phi}{\phi + \lambda}, \quad r = \phi,$$

which is used below. Note that as ϕ approaches 0, this model approaches a Poisson model.

3.1.2 Model

Define the random variable N, where N_{gij} as the number of reads mapped to gene $g = 1, ..., G$ during treatment $i = 1, ..., I$ and replicate $j = 1, ..., J$. Thus, G is the total number of genes, I is the number of different treatment groups, and J is the number replicates per treatment group. Define α_g as the mean expression level for gene g, and let β_{gi} be the expression level of gene g in treatment group i relative to α_g. Let s_{gij} be the normalization value, which can be a function of many factors such as gene length and the total number of mapped reads in a library [35]. Assuming that N_{gij} follows a negative binomial distribution as defined above, our model is represented by the probability mass function

$$f(N_{gij} \mid \lambda_{gij}, \phi_g) = \frac{\Gamma\left(N_{gij} + \phi_g^{-1}\right)}{\Gamma\left(\phi_g^{-1}\right)\Gamma\left(N_{gij}+1\right)} \left(\frac{1}{1+\lambda_{gij}\phi_g}\right)^{\phi_g^{-1}} \left(\frac{\lambda_{gij}}{\phi_g^{-1}+\lambda_{gij}}\right)^{N_{gij}}, \quad (3.1)$$

where $\lambda_{gij} = \mathrm{E}\left[N_{gij}\right] = \exp(\alpha_g + \beta_{gi} + s_{gij})$, and also $\mathrm{Var}\left[N_{gij}\right] = \mathrm{E}\left[N_{gij}\right] + \phi_g \mathrm{E}\left[N_{gij}\right]^2$.

Likelihood and independence

To gain a better understanding of our model, it is helpful to analyze its likelihood. Assume that we have K distinct clusters. Define the center of each cluster $k \in K$ as $\mu_k = (\mu_{k1}, ..., \mu_{kI})$, which is a vector of the cluster centers for all I treatments of a specific gene, where for $k \in \{1, ..., K\}$ and $\sum_{i=1}^{I} \mu_{ji} = 0$. The likelihood that gene g belongs in the kth cluster is then $f(N_g \mid \alpha_g, \beta_g = \mu_g)$. Adding mixing probabilities for each cluster, $p_k \geq 0$ and $\sum_{k=1}^{K} = 1$, the entire likelihood function is

$$\prod_g \sum_k p_k f(N_g \mid \alpha_g, \beta_g = \mu_g). \quad (3.2)$$

The likelihood function makes explicit the assumption of independence among genes. This is clearly an inaccurate assumption, and one of the drawbacks of this model. However, the relationships between many thousands of genes is extremely difficult to capture. This problem is discussed later on.

3.2 Simulating RNA-Seq Data

This model can be used to generate simulated data from theoretical RNA-Seq experiments, which can then be used to analyze the performance of various RNA-Seq-focused algorithms. In this section, I first describe a program developed to sample RNA-Seq data from the model described above. Several parameters are introduced to control various aspects of the simulated dataset, from the magnitude of expression changes across treatments to the level of fluctuation around gene clusters. I then discuss the implementation of this algorithm and present some sample results and visualizations.

3.2.1 Algorithm

To sample from the model discussed above and produce mapped read counts in a similar way to real RNA-Seq data, it is necessary to introduce several new parameters and concepts. These parameters, as well as all variables used in the algorithm, are described in table 3.1.

The most significant of these new parameters are $K, \{\delta_{ki}\}, \tau_\mu, \tau_\epsilon, \tau_\alpha$, and τ_ϕ, as well as the parameters defining the normal and gamma distributions that are sampled from, $\epsilon_\mu, \epsilon_{\sigma^2}, \alpha_\mu, \alpha_{\sigma^2}$, and ϕ_α and ϕ_β. The parameter K controls the number of distinct expression patters in the simulated data, and δ_{ki} for $k = 1, ..., K$ and $i = 1, ..., I$ determines the type of expression change for expression pattern k during treatment i. The τ parameters control characteristics of the data being analyzed by RNA-Seq and in many ways determine the difficulty of analyzing the data; τ_μ controls the level of change in gene expression across treatments, τ_ϵ determines the level of expression fluctuation or change relative to the average of the corresponding expression pattern, τ_α controls the level of average expression, and τ_ϕ controls the level of dispersion, with $\tau_\phi = 0$ corresponding to a Poisson model. The algorithm used, described in algorithm 3.1, takes as input the vector $(G, I, J, K, \tau_\mu, \tau_\epsilon, \tau_\alpha, \tau_\phi, \{\delta_{ki}\}, \epsilon_\mu, \epsilon_{\sigma^2}, \alpha_\mu, \alpha_{\sigma^2}, \phi_\alpha, \phi_\beta)$ and outputs the $G \times I \times J$ matrix N, with N_{gij} sampled appropriately from $f(N_{gij} \mid \lambda_{gij}, \phi_g)$.

G	Total number of genes whose expression is being modeled
I	Number of distinct treatments or expression experiments
J	Number of replicates for each expression experiment
K	Number of distinct expression patterns
τ_μ	Controls the level of change in gene expression across experiments
τ_ϵ	Controls the level of expression fluctuation
τ_α	Controls the level of average gene expression
τ_ϕ	Controls the level of dispersion in the experiments
δ_{ki}	Determines the type of expression change for expression pattern k in experiment i
ϵ_μ	The mean of the normal distribution ϵ_{gi} is sampled from
ϵ_{σ^2}	The variance of the normal distribution ϵ_{gu} is sampled from
α_μ	The mean of the normal distribution α_g is sampled from
α_{σ^2}	The variance of the normal distribution α_g is sampled from
ϕ_α	The shape of the gamma distribution ϕ_g is sampled from
ϕ_β	The rate of the gamma distribution ϕ_g is sampled from
Z_{gk}	An indicator variable representing whether or not gene g has expression pattern k
μ_{ki}	The average expression for pattern k in experiment i
ϵ_{gi}	The fluctuation of the expression of gene g in experiment i
β_{gi}	The variance of expression level of gene g caused by experiment i
α_g	The overall mean expression for gene g
ϕ_g	The overall dispersion of gene g
s_{gij}	The normalization factor for gene g in replicate j of expression experiment i
λ	The mean of the negative binomial distribution to be sampled from
r	The r parameter for the negative binomial distribution to be sampled from
p	The p parameter for the negative binomial distribution to be sampled from.
N_{gij}	The number of reads assigned to gene g in replicate j of experiment i

Table 3.1: Definition of parameters and variables in algorithm 3.1

Algorithm 3.1 An algorithm to simulate RNA-Seq data

Require: $(G, I, J, K, \tau_\mu, \tau_\epsilon, \tau_\alpha, \tau_\phi, \{\delta_{ki}\}, \epsilon_\mu, \epsilon_{\sigma^2}, \alpha_\mu, \alpha_{\sigma^2}, \phi_\alpha, \phi_\beta)$
Ensure: $\{N_{gij}\} \sim f(N_{gij} \mid \lambda_{gij}, \phi_g)$
  $\{Z_{gk}\} \leftarrow \text{Mu}\left(1, \{\frac{1}{K}\}_{i=1,...,K}\right)$
  $\boldsymbol{\mu} \leftarrow \tau_\mu \boldsymbol{\delta}$
  $\{\epsilon_{gi}\} \leftarrow \tau_\mu \tau_\epsilon \mathcal{N}(\epsilon_\mu, \epsilon_{\sigma^2})$
  $\boldsymbol{\beta} \leftarrow (\boldsymbol{Z} \times \boldsymbol{\mu}) + \epsilon$
  $\boldsymbol{\alpha}_g \leftarrow \tau_\alpha \mathcal{N}(\alpha_\mu, \alpha_{\sigma^2})$
  $\boldsymbol{\phi}_g \leftarrow \tau_\phi \text{Gam}(\phi_\alpha, \phi_\beta)$
  $\{s_{gij}\} \leftarrow \mathcal{N}(0, 1)$
  for $g = 1, ..., G$ **do**
    for $i = 1, ..., I$ **do**
      for $j = 1, ..., J$ **do**
        $\lambda \leftarrow \exp(\alpha_g + \beta_{gi} + s_{gij})$
        $r \leftarrow \phi_g$
        $p \leftarrow \frac{\phi_g}{\phi_g + \lambda}$
        $N_{gij} \leftarrow \text{NB}(r, p)$
      end for
    end for
  end for

3.2.2 Sample Simulations

Using an implementation of algorithm 3.1 it is possible to generate sample RNA-Seq data from a variety of experiments, with an arbitrary number of genes, experiments, replicates, and distinct expression patterns. This data can be high dimensional, and thus difficult to visualize. However, by reducing dimensionality, it is possible to visually and more intuitively understand the workings of the algorithms and especially the consequence of altering the control parameters. Sample results for a toy dataset generated using the parameters listed in table 3.2 are shown in figure 3-1. In these figures, the color of the data points represent the original expression pattern, and the axes are $\beta_1 \times \beta_2 \times \beta_3$, which is the variance in expression levels by gene across experiments. Because there are three experiments, this can be visualized in three dimensions. Note that, as τ_ϵ, increases, the clusters of expression patterns becomes less discernible.

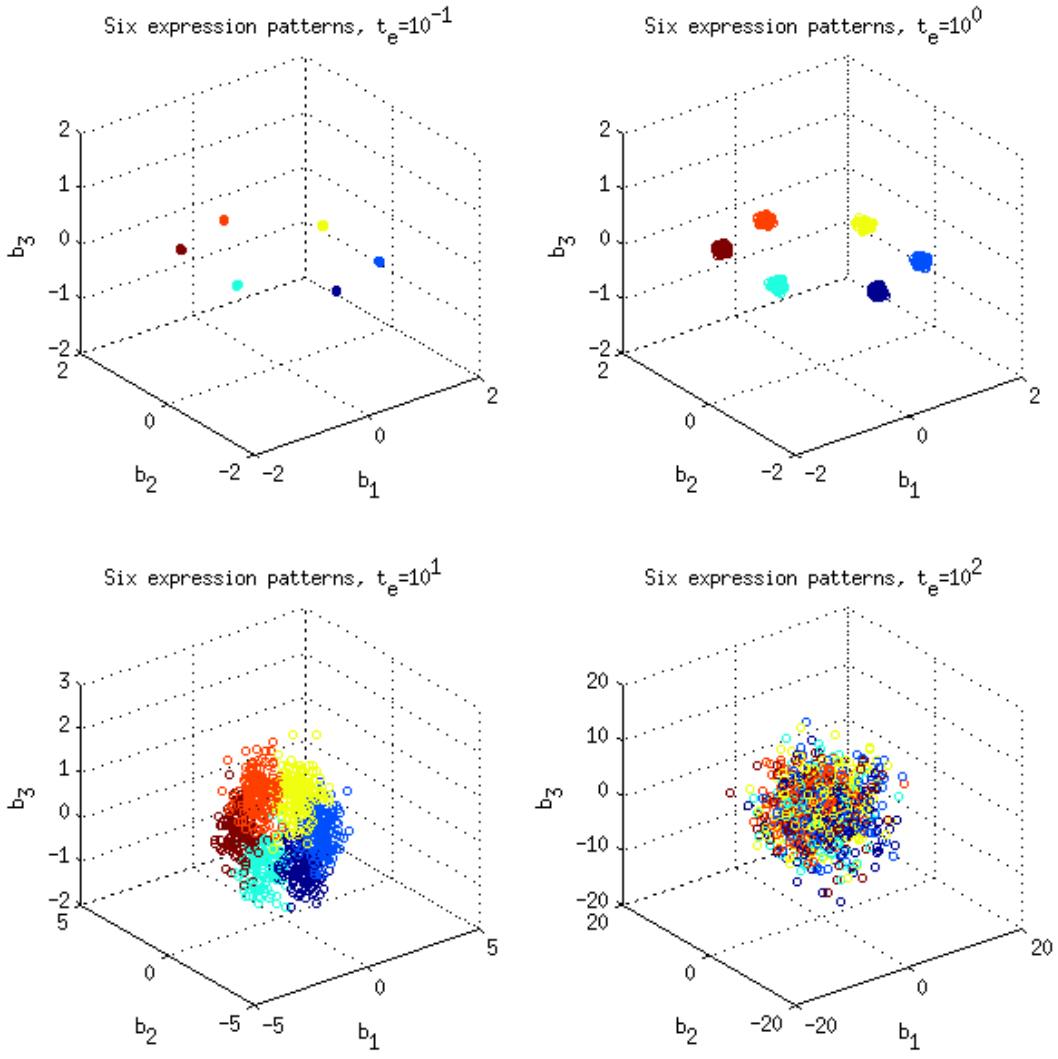

Figure 3-1: Simulated RNA-Seq data from algorithm 3.1, using the parameters from table 3.2. Here, τ_ϵ is varied from $10^{-1}, 10^0, 10^1$, and 10^2. As is intuitive, an increase in τ_ϵ results in increased experiment-independent variability. Note the axes values.

Parameter	Value	Parameter	Value
G	1000	ϵ_μ	0
I	3	ϵ_{σ^2}	$(1/5)^2$
J	2	α_μ	4
K	6	ϵ_{σ^2}	$(1/5)^2$
τ_μ	1	α_{σ^2}	1
τ_ϵ	$\{10^i \mid i = -1, 0, 1, 2\}$	ϕ_α	3/4
τ_α	1	ϕ_β	2
τ_ϕ	1	$\boldsymbol{\delta}$	$\begin{bmatrix} -1 & 0 & 1 & -1 & 0 & 1 \\ 0 & 1 & -1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 0 & 1 & -1 \end{bmatrix}^T$

Table 3.2: Parameters used in the implementation of algorithm 3.1 to generate a the toy simulated RNA-Seq dataset visualized in figure 3-1.

3.3 Limitations of Model and Simulation

While the model described and implemented holds true to true biological datasets in many ways, including incorporating normalization constants on a gene-experiment level and modeling differential expression rather than mean expression, it falls far short from mimicking a true RNA-Seq dataset for several reasons. While much can be learned from analyzing a small, known set of clusters, it must be noted that this behavior cannot simply be immediately extrapolated to true biological datasets.

One of the faults of this model is that it assigns clusters of roughly uniform size. Because cluster assignments are sampled, on an individual basis, from a uniform distribution, the expected cluster sizes are equal. This is a far cry from true biological datasets, where clusters representing core metabolic pathways are often orders of magnitude larger than the average cluster, and the variance in size between clusters of interest will likely be very large [41]. However, this functionality could be added to the model by sampling cluster distributions from a different, more uneven distribution.

Another problem with this simulator is that the expected cluster shapes are also equivalent. While simulations and visualizations show that cluster shape certainly does vary substantially, and while nice defined clusters make appealing data to test clustering algorithms on, the biological reality is again very different. Cluster shapes

from true biological datasets can be expected to be widely disparate in shape [41], and thus more difficult to cluster correctly.

A serious issue with the simulator, which extends to the function and economics of RNA-Seq itself, is a question of dimensionality. Discovering clusters of genes that vary their expression together becomes an easier task as the number of sample points, I in the parametrization of the sampling algorithm, increases. However, for various reasons, including the cost of RNA-Seq analysis, true RNA-Seq data often has low dimensions in I and very high dimensions in G, making the clustering an increasingly difficult task as the parameters increasingly become more realistic.

There are, of course, many other areas in which my simulated RNA-Seq data differs from the biological reality. However, again, beginning with a small, more easily digestible dataset and then moving towards real biology is a sensible way to approach RNA-Seq data simulation. The next set of issues lies in the implementation of the clustering algorithms, where ever-complex biology is again simplified and abstracted in several ways, discussed in section 4.5

Chapter 4

Algorithms for Cluster Analysis of RNA-Seq data

4.1 The General Clustering Problem

Cluster analysis, or clustering, is perhaps the most important problem in unsupervised statistical analysis algorithms. The general clustering problem is to create a partitioning on a set of data points such that those data points that share a partition are more similar, given some definition of similarity, than those data points in other partitions.

Because there are many different definitions of what constitutes a cluster, there is no one algorithm to solve—clustering is, then, a multi-objective optimization problem, and the correct clustering algorithm and parameter settings depend on the dataset under analysis.

However, given these caveats, a general cluster analysis proceeds with the definition of some metric of distance between two data points in some sub-space, $d(x, y)$, and then an attempt to find a k-partitioning $S^* = \{S_1, S_2, ..., S_k\}$ on n data points $x_1, ..., x_n$ in that space such that the sum of the distances between the points in any cluster k,

$$\sum_{x_i, x_j \in S_k} d(x_i, x_j),$$

is minimized while the distance between any two clusters k and k',

$$\sum_{x_i \in S_k, x_j \in S_{k'}} d(x_i, x_j),$$

is maximized. While the exact objective varies from formulation to formulation, this intuition is consistent throughout.

It is often helpful to perform cluster analysis on RNA-Seq data, as mentioned earlier. Now that I have developed an algorithm and program to simulate RNA-Seq data, I turn to the description of several important clustering algorithms and their implementation on this simulated RNA-Seq data.

4.2 Lloyd's Algorithm

Lloyd's algorithm was first described by Stuart Lloyd in 1957 as a tool for pulse-code modulation, and was first published in 1982 [27]. Known less-specifically as "the" k-means algorithm, Lloyd's algorithm aims to partition n data points into k clusters so that each data point belongs to the cluster with the closest mean. Because this problem is NP-hard [10], there exist heuristic algorithms that are often converge quickly to a local optimum. Lloyd's algorithm is similar in some ways to *expectation maximization* algorithms, as they both proceed through an arbitrary number of cycles, calculating new parameters and determining new clusters in each cycle.

4.2.1 Standard Lloyd's Algorithm

To outline Lloyd's algorithm, assume again a set of observations $X = \{x_1, x_2, ..., x_n\}$ where $x_i \in \mathbb{R}^d$, as well as an integer k representing the number of clusters to be found. Lloyd's algorithm seeks to find a partition $S^* = \{S_1, S_2, ..., S_k\}$ with corresponding means $\mu_1, \mu_2, ..., \mu_k$ such that

$$S^* = \underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2,$$

where $||\bullet||^2$ represents Euclidean distance[1]. Lloyd's algorithm is a heuristic for the minimization of this objective function, and is described in algorithm 4.1, from Lloyd (1982).

Algorithm 4.1 Lloyd's algorithm

Require: $(X = (x_1, x_2, ..., x_n), k, T)$ where $x_i \in \mathbb{R}^d,\ k, T \in \mathbb{Z}$
Ensure: $S^* \approx \underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$
  Sample initial means $\mu_1, \mu_2, ..., \mu_k$ at random
  for $t = 1, ..., T$ **do**
    for $i = 1, ..., k$ **do**
      $S_i^{(t)} = \left\{ x_j \middle| \left|\left| x_j - \mu_i^{(t)} \right|\right|^2 \leq \left|\left| x_i - \mu_r^{(t)} \right|\right|^2 \ \forall 1 \leq r \leq k \right\}$
    end for
    for $i = 1, ..., k$ **do**
      $\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$
    end for
  end for

4.2.2 Lloyd's Algorithm for RNA-Seq Data

However, when working with RNA-Seq data, it is necessary to pre-process the data before using Lloyd's algorithm. There are several reasons for this. First, the basic Lloyd's algorithm uses Euclidean distance, although it can be easily modified for use with most distance metrics, which is not appropriate for use with exponentially distributed data. Second, in seeking genes that modify their expression together and

[1]It is interesting to note that, in any optimal solution S^*, $\mu_j = \frac{1}{||S_j||} \sum_{x_i \in S_j} x_i$, the mean of the points in cluster j. Therefore, μ_j may be removed from the objective function in the following manner [10]. Let X and Y be chosen independent and identically distributed samples from the cluster S_j. Then, because $\mathrm{E}\left[||X - Y||^2\right] = 2\mathrm{E}\left[||X - \mathrm{E}[X]||^2\right]$,

$$\sum_{x_j \in S_i} ||x_j - \mu_i||^2 = \frac{1}{2 \mid S_i \mid} \sum_{x_j, x_k \in S_i} ||x_j - x_k||^2$$

and the objective function can be rewritten as

$$\sum_{i=1}^{k} \frac{1}{2 \mid C_i \mid} \sum_{x_j, x_k \in S_i} ||x_j - x_k||^2 .$$

thus are likely involved in similar functions or networks, it is desirable to cluster genes by their change in expression, β_g, and not by their overall expression, as is implied in either N or $\alpha_g + \beta_g$. Third, the introduction of normalization constants, which are commonly used with RNA-Seq data [35], represented in algorithm 3.1 by the matrix $\{s_{gij}\}$, should be accounted for.

For these reasons, I have slightly modified Lloyd's algorithm to allow it to be used with RNA-Seq data. The modified version proceeds as follows and is outlined in detail in algorithm 4.2. The count data contained within the matrix N is log-normalized, and the normalization constants within S are subtracted. This data is then thresholded at 10^{-6}, and the replicates are averaged. The mean expression for each gene is found, and subtracted from all averaged treatments. The data points are then normalized to have zero mean and unit variance. Any data points with standard deviation less than 10^{-6} are removed. Finally, a version of Lloyd's algorithm is run on this data. However, herein the distance metric used, $d(r, s)$, is one minus the sample correlation between points, or

$$d(r,s) = 1 - \frac{(x_r - \bar{x}_r)(x_s - \bar{x}_s)'}{\sqrt{(x_r - \bar{x}_r)(x_r - \bar{x}_r)'}\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}} \tag{4.1}$$

The cluster centroids, c_i for $i = 1, ..., K$, are defined as the component-wise mean of all the data points in the cluster.

4.2.3 Implementation Using Simulated Data

Algorithm 4.2 was implemented. To visualize the results of running this modified Lloyd's algorithm on simulated RNA-Seq data, it is helpful once again to limit our number of treatments to three, and then visualize our results in three dimensions. Running this implementation on simulated data generated using the parameters in table 3.2, I obtain the results shown in figure 4-1. Here, the data points are shown graphed by their true β values, which in a real experiment would not be known, but are helpful here. The colors of the data points are assigned by the program.

Algorithm 4.2 Modified Lloyd's algorithm for RNA-Seq data

Require: (N, S, k, T)
Ensure: $S^* \approx \underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in S_i} d(x_j, c_i)$
  if $N'_{gij} \leq 0$ **then**
    $N'_{gij} = 10^{-6}$
  end if
  for $g = 1, ..., G$ **do**
    for $i = 1, ..., I$ **do**
      $D_{gi} = \frac{1}{J} \sum_{j=1}^{J} N'_{gij}$
    end for
  end for
  for $i = 1, ..., I$ **do**
    $D_{:i} = D_{:i} - \frac{1}{I} \sum_{i=1}^{I} D_{:i}$
  end for
  for $g = 1, ..., G$ **do**
    $\mu_g = \frac{1}{I} \sum_{j=1}^{I} D_{dj}$
    $\sigma_g = \sqrt{\frac{1}{I} \sum_{k=1}^{I} (D_{g,i} - \mu_g)^2}$
    if $\sigma_g \leq 10^{-6}$ **then**
      Remove row $D_{g:}$ from D
    end if
    $D_{g:} = D_{g:} - \mu_g$
    $G_{g:} = D_{g:} / \sigma_g$
  end for
  for $t = 1, ..., T$ **do**
    for $i = 1, ..., k$ **do**
      $S_i^{(t)} = \left\{ x_j \middle| d(x_j, c_i) \leq d(x_j, c_r) \ \forall 1 \leq r \leq k \right\}$
    end for
    for $i = 1, ..., k$ **do**
      $c_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$
    end for
  end for

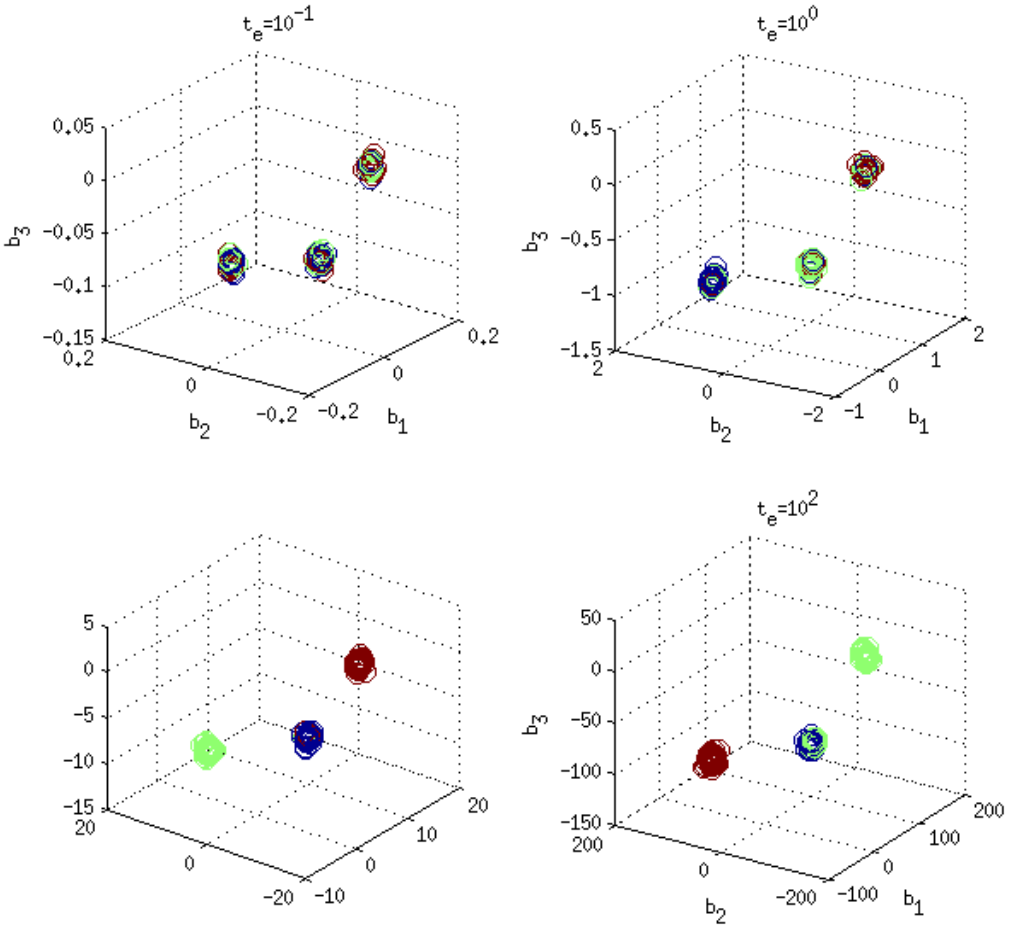

Figure 4-1: Algorithm 4.2 run on simulated RNA-Seq data with varying values for τ_ϵ. As the distance between the centroids of the simulated clusters is increased, the clustering algorithm becomes more accurate.

4.3 Agglomerative Hierarchical Clustering

Eisen (1998) was one of the first scientists to apply hierarchical clustering to analyze expression data [12]. *Agglomerative*, as opposed to *divisive*, clustering is the process of partitioning data into sets by starting with a single element, and aggregating them into clusters. Hierarchical clustering algorithms result in a *dendrogram*, or rooted tree, which can be sorted according to some criteria and are often displayed alongside a gene expression heat map. Perhaps the most commonly used agglomerative hierarchical clustering algorithm is the unweighted pair group method with arithmetic mean, or

UPGMA.

4.3.1 Unweighted Pair Group Method with Arithmetic Mean

UPGMA is a simple hierarchical clustering algorithm that is very commonly used in bioinformatics. UPGMA assumes a constant rate of evolution[2], which is often cited as a drawback of using this algorithm. There exist many variations of UPGMA that are used in bioinformatics, including *single-linkage clustering*, *average-linkage clustering*, *complete-linkage clustering*, *weighted pair group average*, *within-groups clustering*, and *Ward's method*. These algorithms vary in their complexity, and each works better with some datasets than with others. For example, complete-linkage clustering is efficient at finding compact, uniformly sized clusters, while the weighted pair-group average should be used when cluster sizes are expected to be greatly uneven.

The input of the UPGMA algorithm is a distance or similarity matrix D representing the differences between data points, and the output is a dendrogram. At each step in the algorithm, the two clusters that are determined to be most similar are joined. Distance between clusters is defined as the mean distance between elements in each cluster, or, equivalently, the average of all the distances between all pairs of objects. For two clusters X and Y, this is defined as

$$D_{XY} = \frac{1}{|X| + |Y|} \sum_{x \in X} \sum_{y \in Y} D_{xy}$$

. The UPGMA algorithm proceeds as follows. Let n be the number of data points being clustered, and let N be the set of all data points. Create n clusters, and initialize each one with a distinct gene (or species, etc). Iteratively, find the clusters i, j such that $D_{ij} \leq D_{xy}\ \forall x, y \in N$. Form a new cluster by joining these two clusters, and give the branches connecting i and j each a length of $\frac{d(i,j)}{2}$. Compute the distance

[2]This is known as the *molecular clock hypothesis*. This idea is first attributed to Emile Zickerkandl and Linus Pauling who discovered in 1962 that the change in the number of amino acids in hemoglobin between lineages is roughly linearly correlated with time. This observation was based on the use of fossilized data, and was then generalized to say that the rate of evolutionary change of any general, specific protein is roughly constant over time and lineage.

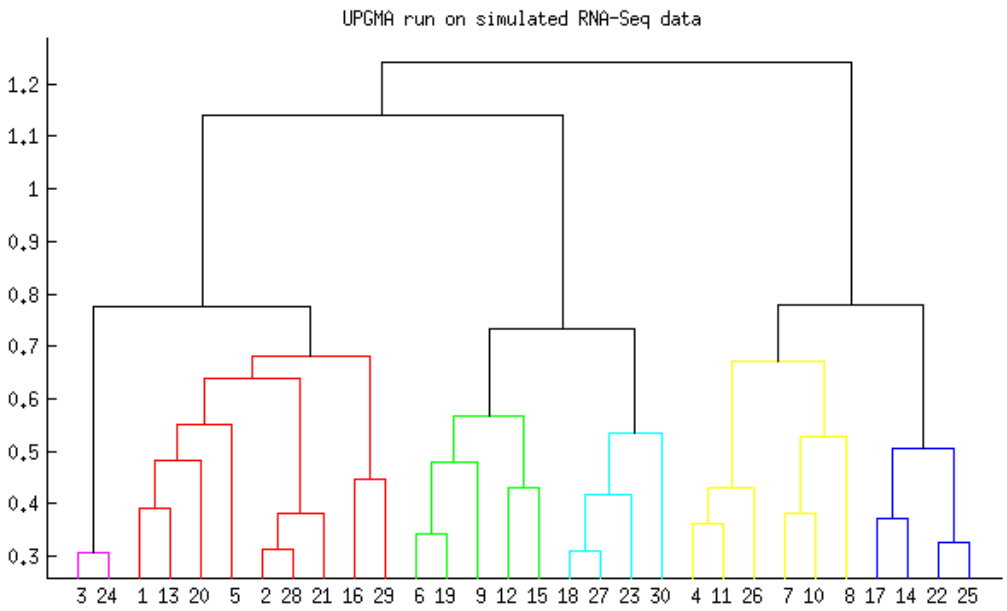

Figure 4-2: The resulting dendrogram from UPGMA run on simulated RNA-Seq data for $\tau_\mu = 1$. The dendrogram has been colored to show the cluster assignments.

from this new cluster to all other clusters in the set $N \setminus \{i, j\}$ by

$$D_{(ij),k} = \frac{|I|}{|I| + |J|} D_{ik} + \frac{|J|}{|I| + |J|} D_{jk}.$$

Delete the rows and columns in D corresponding to i and j and add a column for the new cluster (i, j) as computed above. Continue to iterate in this manner until only one cluster remains.

4.3.2 UPGMA for RNA-Seq Data

It is necessary to modify slightly the UPGMA algorithm to find better results with our simulated RNA-Seq data. As is described in detail in 4.2.2, the count data was log normalized, the normalization constants were subtracted, the data was thresholded, replicates averaged, and the mean expression for each gene subtracted from all of the treatments. The data was then normalized to have zero mean and unit variance, and any data point with a very low variance was removed. A distance matrix D was created using sample correlation, as per equation 4.1. The UPGMA algorithm was

run on D, outputting a dendrogram. Finally, the dendrogram was processed to find the smallest height at which a horizontal cut through the tree would leave k or fewer clusters, for a specified value of k.

4.3.3 Implementation Using Simulated Data

I implemented this above described algorithm and ran it using simulated data. Again, the number of treatments was limited to three, and the visualization is in three dimensions with axes $\beta_1 \times \beta_2 \times \beta_3$, and the colors of the points were assigned by the algorithm. The results are shown in figure 4-3

4.4 Artificial Neural Networks

It has long the case that, while digital computers are far superior in number manipulations, the human brain is capable of seemingly effortlessly solving very difficult computational problems. For example, although years of research and enormous amounts of funding have gone into speech and face recognition, even the best algorithms will struggle to understand a foreign accent in a busy room, or to recognize a face across a dimly-lit room. Human brains, on the other hand, excel at these tasks.

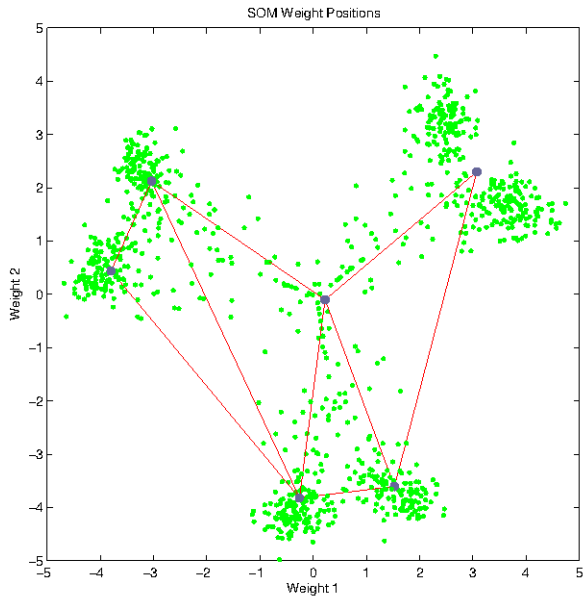

Figure 4-4: The purple dots represent the values of the trained weights for each of the six neurons in the implemented SOM, and the red lines show the inter-neuron connections. The green dots are a scatter plot of the first two principal components of the simulated data. Notice the 'incorrectly' placed neurons in the middle and between the two top-right clusters.

In 1943, a neuron was modeled as a binary switch, receiving input from other neurons and, depending on whether the total weighted value of those inputs exceeded some threshold, outputting an 'active' or 'inactive' response [25]. Later, in 1960, it was shown that these model neurons behave in ways similar to the brain—they are capable of advanced pattern matching, and can sustain function even with a fluctuating number of total neurons. Attempts to simulate this behavior computationally are called *artificial neural networks*, a computational mechanism which seeks to simulate a net of neurons. Interest in artificial neural networks has waxed and waned, and are often criticized for reasons ranging from technical difficulty to the requirement of large amounts of training data. Regardless, they have been used to solve many interesting problems, and are frequently used in bioinformatics to partition gene expression data. Of particular use and interest is the *self-organizing map*.

4.4.1 Self-organizing map

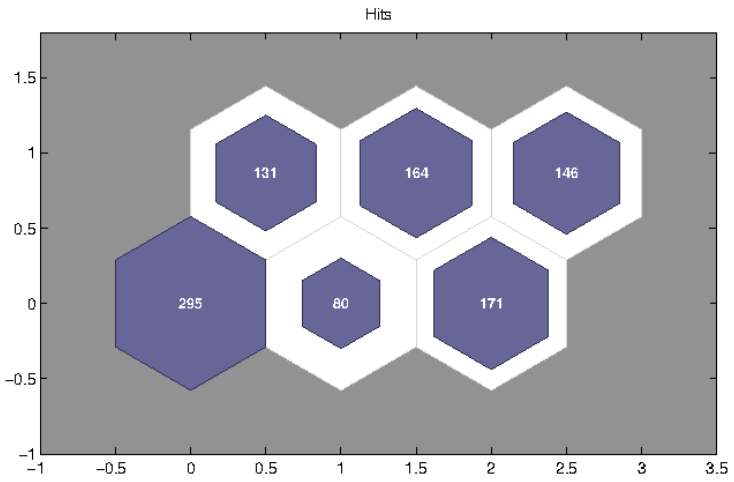

Figure 4-5: Here the topological arrangement for the SOM is visualized, along with the number of data points assigned to each neuron. Notice that the left-most neuron has roughly double the average of the others; this neuron corresponds to the blue cluster in figure 4.4.3.

The self-organizing map (SOM) induces a mapping of a high-dimensional distribution onto a regular, low-dimensional grid. In doing so, a SOM is capable of converting non-linear statistical relationships between high-dimensional data points into a group of simple, geometric relationships [24]. Because of these properties, SOMs are often used for the visualization of high-dimensional data, and it is this type of dimensionality reduction that makes them useful for cluster analysis. SOMs are frequently used in the clustering of transcriptome data [35].

The learning of a SOM is an iterative process, wherein input vectors are classified according to their grouping in input space. Consider a set of weight vectors $w_i \in \mathbb{R}^n$ and a set of observation vectors $x \in \mathbb{R}^n$. When the training vector $x^{(t)}$ enters the network, the winning neuron i^* and all neurons $i \in N_{i^*}$ are adjusted by

$$w_i^{(t+1)} = w_i^{(t)} + \Phi^{(t)}\alpha^{(t)}(x^{(t)} - w_i^{(t)}),$$

where $0 \leq \alpha(t) \leq 1$ is the learning rate factor, which decreases monotonically with the number of iterations. Also, N_{i^*} defines the *neighborhood function* that contains the indices of all the neurons within radius r of i^*,

$$N_i = \{j \mid d(i,j) \leq r\}.$$

Weight vectors w_i are often randomly initialized, originally because this allowed the self-organizing properties of the SOM to become apparent. However, training time can be decreased by orders of magnitude by initializing the weight vectors with the eigenvectors that correspond to the two principal components of the input data [24].

4.4.2 SOM for RNA-Seq data

To create a SOM suitable for use on the simulated RNA-Seq data, it is first necessary to manipulate our data as was described in section 4.3.2. Following this, a principal component analysis (PCA) is used. A SOM of dimensions appropriate for the number of expected clusters is then created and and trained with the PCA results. Data points are then clustered by finding the nearest neuron to each point in the data set.

4.4.3 Implementation using simulated data

This SOM algorithm was implemented and run with simulated data as well, as described with UPGMA and Lloyd's algorithm. A sample clustering result is shown in figure 4.4.3. Also, a plot of the SOM network over a scatter plot of the first two principal components of the data is shown in figure 4-4. These two figures are interesting because several distinct clusters been clustered together by the SOM. Looking at the graph of SOM weights, we see that, while four of the weights are correctly positioned in the center of the clusters, one is located at the center of the graph, and another between two clusters, thus causing the incorrect clustering. The two-dimensional structure of the implemented SOM is visualized, along with the number of vectors assigned to each neuron, in figure 4.4.1.

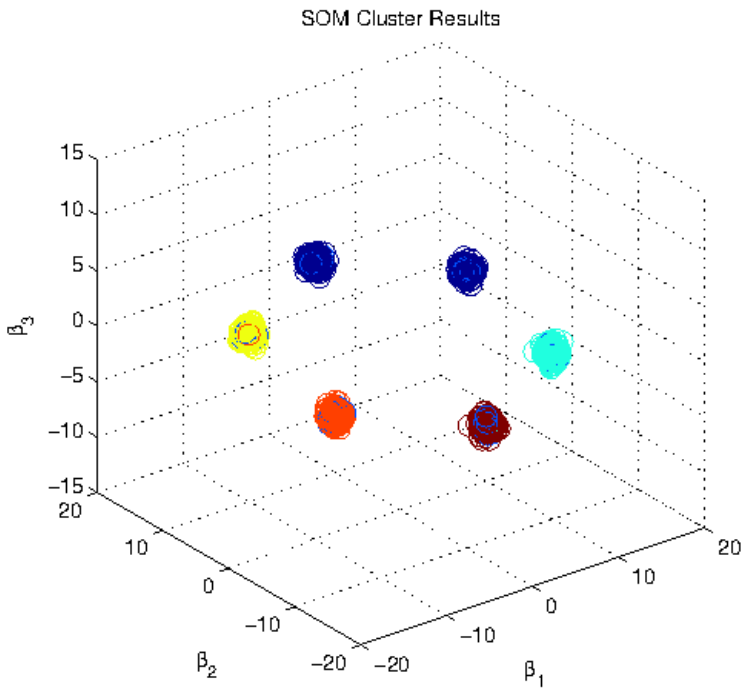

Figure 4-6: The results of running the six-neuron SOM on the simulated data. Notice that two actual clusters have been joined as one, as is implied by the plot of neuron weights in figure 4-4.

4.5 Issues with Discussed Algorithms

The algorithms discussed and implemented were chosen for their ubiquity. These algorithms are often used as an initial test for function, and are used so here. However, as is the case with the described RNA-Seq data simulator, these algorithms also are the result of a trade-off between accuracy and simplicity. There are several key issues with the proposed algorithms when considering their implementation with true RNA-Seq datasets.

Primary among these simplifications is the assumption of knowledge of the true number of clusters, k. In real data the correct choice of k is often unknown—indeed, even defining what would constitute a correct number of clusters is a difficult task. This is a difficult problem because, as is intuitive, simply increasing k without penalty

will *always* result in reduction in clustering error (consider the extreme case when $k = n$, the number of data points, resulting in 0 error). The correct value for k will find some balance between maximum compression of data and maximum accuracy of clustering. If some value for k is not expected, as is assumed in the algorithms outlined above, there are several methods of estimating k.

One such method of choosing the parameter k is maximizing some information criterion, such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC) or the Deviance information criterion (DIC). As the likelihood function for our simulation is known, finding a value for k through this method would be a straightforward process.

Another such method would be using the Silhouette validation metric, which is discussed further in sub-section 5.1.1. Combining this validation metric with an optimization technique to determine the number of clusters that produces the largest Silhouette validation metric is another way to determine an approximation for the correct number of clusters.

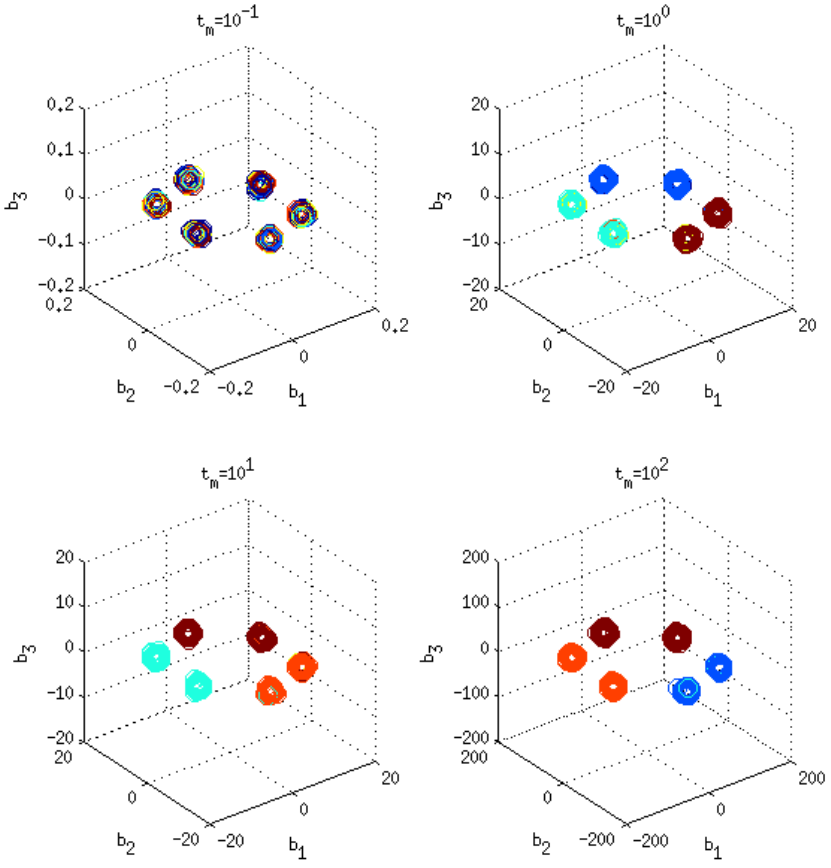

Figure 4-3: UPGMA run on simulated RNA-Seq data with varying values for τ_μ. Again, as expected, as the distance between the centroids of the simulated clusters is increased, the clustering algorithm becomes more accurate.

Chapter 5

Analysis of cluster performance on simulated data

In a typical experiment, after an appropriate clustering algorithm has been run on the selected features of the data, it is necessary to validate the output of the algorithm. This process is denoted *cluster validation*. As shown in figure 5, cluster validation can lead to refined feature selection, algorithm selection, or to interpretation new knowledge.

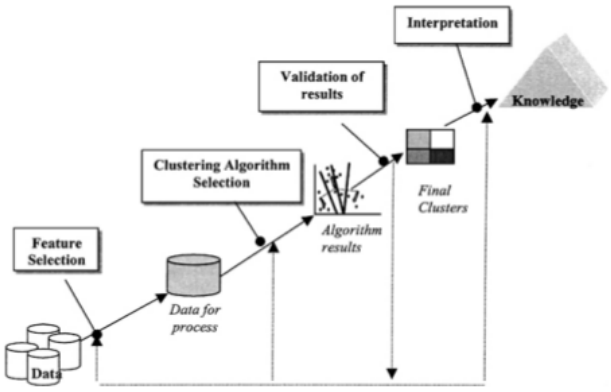

Figure 5-1: The procession of a clustering experiment.

5.1 Quantitative measures of cluster goodness

There exist many different metrics of cluster goodness. In general, these methods are based either on *internal criteria* or on *external criteria*.

A metric based on internal criteria is evaluated solely on the basis of the clustered data. Often, this means assigning the best score results with a high degree of intra-cluster similarity, as measured by some distance metric, and low inter-cluster similarity. The drawback of using internal criteria for cluster validation is that a high score does not necessarily reflect a high level of information retrieval. Further, such metrics are biased towards algorithms that optimize the same criteria—for example, the k-means algorithm is designed so as to maximize cluster distances, so with high probability internal criteria will overrate its results [16]. Well-known validity metrics based on internal criteria include *Silhouette validation*, the *Davies-Bouldin index*, and the *Dunn index*.

Cluster validation based on external criteria evaluates results based on data that was not used in clustering, such as known class labels. Often used external criteria metrics include the *Rand index*, *Jaccard index*, and *mutual information*.

Below, I outline in detail silhouette validation, an internal cluster validation metric, as well as the Rand index and mutual information, two external validation metrics. These metrics will then be used to evaluate the clustering algorithms presented in section 4.1 on simulated data generated using the model from sub-section 3.1.2.

5.1.1 Silhouette validation

The silhouette validation metric was originally described by Peter Rousseeuw in 1986 as a technique for visualizing and evaluating clustering validity [34]. To visualize the cluster goodness, each cluster is represented by a silhouette object, which is based on the tightness and separation of the data. The result is a figure that shows which objects are well-clustered and which objects lie somewhat between clusters. Further, as described by Rousseeuw (1986), the average silhouette width provides a quantitative score of cluster validity.

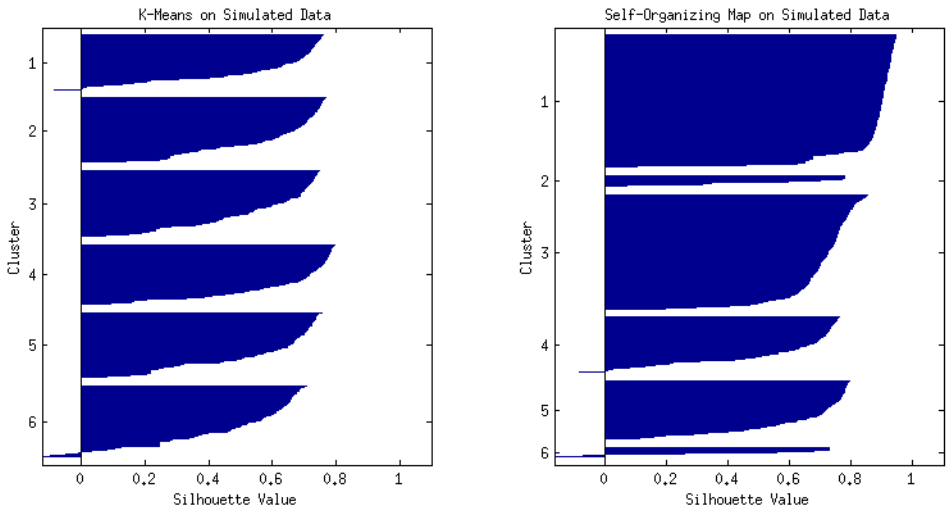

Figure 5-2: Sample silhouette visualizations.

Let a_i be the be the average distance of data point i from all other data points assigned the same cluster, S_i, or $a_i = \frac{1}{|S_i|} \sum_{j \in S_i} d(i,j)$. As always, any appropriate distance function d may be used. Intuitively, a_i measures how well i fits its cluster. Also, let b_i represent the lowest average distance between i and all the data points $j \in S_j$ for all $j \neq i$, or $b_i = \min_{S_j \neq S_i} \frac{1}{|S_j|} \sum_{j \in S_j} d(i,j)$. Clearly, cluster S_j is the cluster that best fits i after S_i. Define s_i on the interval $[-1, 1]$ as

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}.$$

As $|a_i - b_i|$ grows, then, s_i approaches 1. At $s_i = 0$, data points i is on the border between S_i and S_j, and as s_i approaches -1, i is misclassified and would be more appropriately placed within S_j. The overall average silhouette width, $\frac{1}{k} \sum_i s_i$, evaluates the clustering of the entire dataset. See figure 5-2 for sample silhouette visualizations for both the k-means and self-organizing map algorithms run on the same simulated data.

5.1.2 Rand index

The Rand index was first described by William Rand in 1971, and can be understood as the mathematical *accuracy*[1]of the clustering. The Rand index requires knowledge of the correct clustering, and so is an external criteria metric. Calculating the Rand index results in a value in the interval $[0, 1]$, where 1 indicates totally correct clustering and 0 indicates totally incorrect clustering [32].

To find the Rand index, define two clustering over n items, S^1 and S^2. Let a be the number of pairs of data pints that are in the same set in both S^2 and S^2, and b be the number of pairs in different sets in both S^1 and S^2. The Rand index, R, can be quickly calculated by

$$R = \frac{a+b}{\binom{n}{2}}$$

.

To evaluate the goodness of the clustering algorithms in this thesis, I use a modified version of the Rand index, denoted the *adjusted Rand index*. The adjusted Rand index corrects the Rand index for chance clustering [20]. To find the adjusted Rand index as per Hubert (1985), define the contingency matrix $\{n_{ij}\}$ where n_{ij} denotes the number of data points in partition i of S^1 and partition j of S^2, or $n_{ij} = |S_i^1 \cap S_j^2|$. Also, define $a_i = \sum_{j=1}^{k} n_{ij}$ and $b_j = \sum_{i=1}^{k} n_{ij}$, as shown in table 5.1.2. Using $\{n_{ij}\}, a_i$,

	S_1^2	S_2^2	$\cdots$	S_k^2	
S_1^1	n_{11}	n_{12}	$\cdots$	n_{1k}	a_1
S_2^1	n_{21}	n_{22}	$\cdots$	n_{2k}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
S_k^1	n_{k1}	n_{k2}	$\cdots$	n_{kk}	a_k
	b_1	b_2	$\cdots$	b_k	

Table 5.1: Contingency matrix for the adjusted Rand index.

[1]Accuracy is a statistical metric used in binary classification tests to measure how well a condition is correctly identified or excluded—it is the percentage of correct results. Let the number of true positives, true negatives, false positives, and false negatives be p_t, n_t, p_f, n_f, respectively in a given test. The accuracy, a, of the test is found by $a = \frac{p_t+n_t}{p_t+p_f+n_t+n_f}$.

and b_j, the adjusted Rand index, AR, is calculated by

$$AR = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}.$$

5.1.3 Mutual information

In information theory, mutual information measures the dependence of two random variables. Effectively, mutual information indicates the reduction in uncertainty in one random variable due to the knowledge of another. In cluster validation, mutual information is used to quantify the shared information between the known, true partition and the found partition. A high value for mutual information indicates strong dependence, the presence of more shared information. Consider two continuous random variables, X and Y. Given the joint and marginal probability density functions $p(x), p(y)$ and $p(x, y)$, the mutation information of X and Y is

$$M(X, Y) = \int_Y \int_X p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right).$$

In the case of discrete random variables, the double integral is replaced with a double summation and the marginal and joint probability distribution functions are used [38].

Because $I(X, Y)$ has no upper bound, in this thesis I use *normalized mutual information* from Strehl and Ghosh (2002). The normalized mutual information NMI is defined as

$$NMI(X, Y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}},$$

where $H(X)$ is the entropy of X.

5.2 Analysis

The three metrics of cluster validity discussed above—silhouette validation, Rand index, and normalized mutual information—were implemented. The algorithms from

chapter 4.1 were run on the on simulated RNA-Seq data sampled from the model described in chapter 3.1.2 with various parameter settings, and these metrics were used to analyze the results. For each parameter setting for each algorithm, data was simulated and clustered and goodness of fit using each metric was calculated multiple times, and averages were plotted, along with standard error.

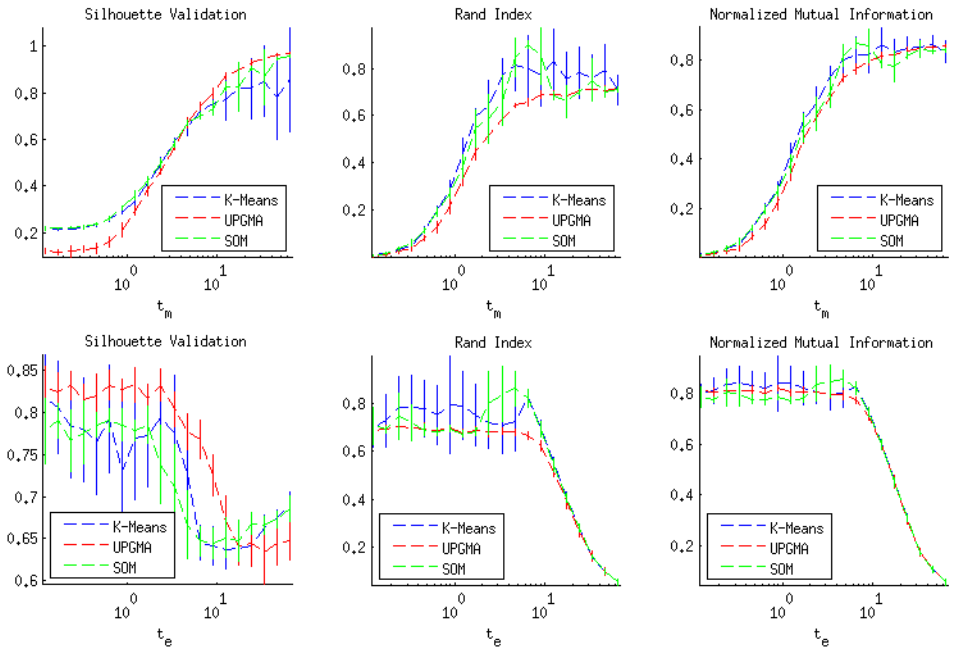

Figure 5-3: Results with four clusters using different clustering algorithms. For each parameter settings, results from 20 data sets were averaged and plotted. The length of each vertical bar represents standard error.

I first performed this analysis on the initial versions of the algorithms outlined above on a toy dataset consisting of $k = 4$ clusters, $G = 1,000$ genes, and $I = J = 3$ experiments and replicates per experiment. I performed this analysis on experiments with varying values for τ_μ, the level of expression change across expression clusters, and τ_ϵ, the level of expression dispersion within a cluster. Each experiment parameterization was repeated 20 times, with results shown in figure 5-3. In these results, each algorithm produced results that were roughly statistically equal, although it seems likely that both SOM and k-means outperformed UPGMA for most experi-

ments. It is also apparent that the Silhouette validation metric is the least predictive of cluster goodness, which makes sense since it does not utilize external information in the form of correct labeling.

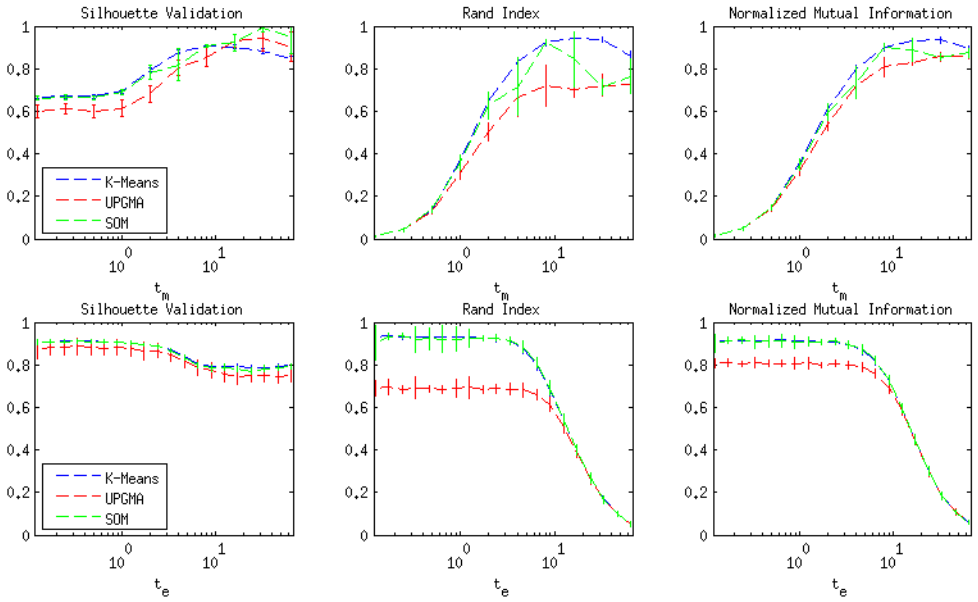

Figure 5-4: Results with six clusters using different clustering algorithms. For each parameterization, 50 results were averaged and plotted. The length of each vertical bar represents standard error.

After these results, I implemented the improvement mentioned in k-means, and re-ran a simulation with $k = 6$ clusters, maintaining the other parameters the same. For this analysis, each experiment was simulated 50 times before results were averaged and plotted. The results are much clearer and have much lower variance, as is shown in figure 5-4. From these experiments, it is clear that the improved version of k-means is strictly better than UPGMA for almost all experiment settings. SOM performs in a manner very close to k-means.

Finally, I simulated a RNA-Seq dataset with the number of clusters k varying from 1 to 80. This simulation was done to check the limits of the clustering algorithms as the complexity of the generated data increased, somewhat more closely approximating a true RNA-Seq dataset. As expected, this drastic increase in the number of clusters

over a small number, $I = 3$, of sample points resulted in a much lower cluster scores, shown in figure 5-5.

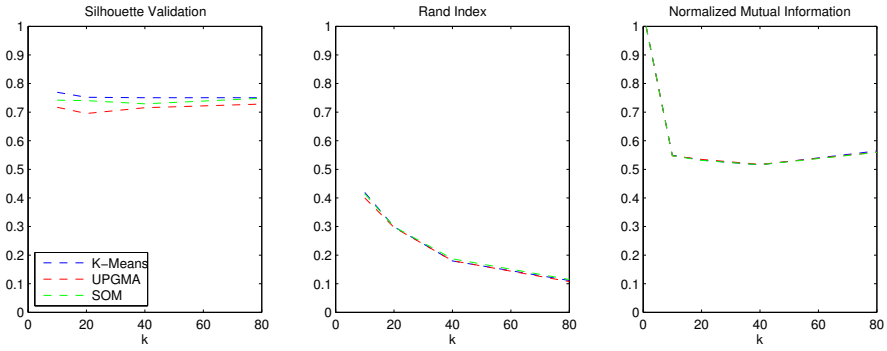

Figure 5-5: Results with $k = 1, ..., 80$ using different clustering algorithms. Note the decrease in cluster score as the complexity of the data increases and becomes unmanageable.

Chapter 6

Conclusion

In conclusion, this work has shown that clustering by change in expression can find co-expressed gene modules in simulated and simplified RNA-Seq data. Also, it has been shown that for most experiment parameterizations, given this simulated RNA-Seq data, an improved version of k-means is the best algorithm amongst those discussed to discovered clusters, given the number of clusters k.

Further, the implemented RNA-Seq experiment simulator, algorithm implementations, and cluster validation metrics form a framework that can currently be used in the simulation and validation of simple RNA-Seq experiments and clustering algorithms. These tools also form a foundation that may be expanding to increase complexity and approach a truer reflection of the true biological structures present in RNA-Seq data.

6.1 Future Direction

The next step in the progression of this research is the validation of the created framework with real-world RNA-Seq data. This validation will take the form of a statistical analysis between simulated RNA-Seq data and true RNA-Seq data, preferably with known gene modules. The model behind this simulator may then be intelligently improved to better simulate true RNA-Seq data.

Another, perphaps parallel, direction is the investigation of other clustering algo-

rithms. Algorithms to be investigated should include standard implementations of support vector machines (SVMs), as well as model and density-based clustering algorithms. Another interesting class of classification algorithms are network-inference based, such as *weighted gene co-expression analysis* [43]. These implementations may be run on the simulated RNA-Seq data, and iteratively validated and improved.

Bibliography

[1] Pratul Agarwal, Christopher Schultz, Aristotle Kalivretenos, Brahma Ghosh, and Sheldon Broedel. Engineering a hyper-catalytic enzyme by photoactivated conformation modulation. *The Journal of Physical Chemistry Letters*, 3:1142, 2012.

[2] S Anders and W Huber. Differential expression analysis for sequence count data. *Nature Precedings*, 11:R106, 2010.

[3] Bruce Babcock. The impact of us biofuel prices on agricultural price levels and volatility. Technical report, Center for Agricultural and Rural Development, 2011.

[4] Steven Brown, Adam Guss, Tatiana Karpinets, Jerry Parks, Nikolai Smolin, Shihui Yang, Miriam Land, Dawn Klingemen, Ashwini Bhandiwad, Miguel Rodriguez, Babu Raman, Xiongjun Shao, Jonathan Mielenz, Jeremy Smith, Martin Keller, and Lee Lynd. Mutant alcohol dehydrogenase leads to improved ethanol tolerance in clostridium thermocellum. *Proceedings of the National Academy of Science*, 108:13752, 2011.

[5] J.H. Bullard, E. Purdom, K.D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11:94, 2010.

[6] Jose Carioca. Biofuels: Problems, challenges and perspectives. *Biotechnology Journal*, 5:260, 2010.

[7] Andrew Carroll and Chris Sumerville. Cellulosic biofuels. *Annual Review of Plant Biology*, 60:165, 2009.

[8] Bruce Dale and David Pimentel. The cost of biofuels. *Chemical & Engineering News*, 85:12, 2007.

[9] Mark Danigole. Biofuels: an alternative to u.s. air force petroleum fuel dependency. Technical report, Center for Strategy and Technology, Air War College, Air University, 2007.

[10] Sanjoy Dasgupta. The hardness of k-means clustering. Technical report, University of California, San Diego, 2008.

[11] Rebecca Davidson, Candice Hansey, Malali Gowda, Kevin Childs, Haining Lin, Brieanna Vaillancourt, Rajandeep Sekhon, Natalia de Leon, Shawn Kaeppler, Ning Jiang, and C. Buell. Utility of rna sequencing for analysis of maize reproductive transcriptome. *The Plant Genome*, 4:191, 2011.

[12] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science*, 95:14863, 1998.

[13] Geraint Evans. International biofuels strategy project. liquid transport biofuels. technology status report, nnfcc 08-017. Published by the NNFCC, April 2008.

[14] J. Goettemoeller and A. Goettemoeller. *Sustainable Ethanol*. Prairie Oak Publishing, 2007.

[15] Government Accountability Office. Biofuels: Doe lacks a strategic approach to coordinate increasing production with infrastructure development and vehible needs. Technical report, U.S. Government Accountability Office, 2007.

[16] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107, 2001.

[17] C. Hammel-Smith, Fang J, M. Powders, and J. Aabakken. Issues associated with the use of higher ethanol blends (e17-e24). Technical report, National Renewable Energy Laboratory, 2002.

[18] Darby Harris and Seth DeBolt. Synthesis, regulation and utilization of lignocellulosic biomass. *Plant Biotechnology Journal*, 8:244, 2010.

[19] S. Horvath, B. Zhang, M. Carlson, K. Lu, S. Zhu, R. Felciano, M. Laurance, W. Zhao, S. Qi, Z. Chen, Y. Lee, A. Scheck, L. Liau, H. Wu, D. Geschwind, P. Febbo, H. Kornblum, T. Cloughesy, S. Nelson, and P. Mischel. Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target. *Proceedings of the National Academy of Science*, 103:17402, 2006.

[20] Lawrence Hubert. Comparing partitions. *Journal of Classification*, 2:193, 1985.

[21] International Energy Agency. World energy outlook, 2006.

[22] Joint BioEnergy Institute. Joint bioenergy institute 2009-2010 report, 2010.

[23] Samir K. Khanal. Bioenergy and biofuel from biowastes and biomass. Technical report, American Society of Civil Engineers, 2010.

[24] Teuvo Kohonen. The self-organizing map. *Neurocomputing*, 21:1, 1998.

[25] Anders Krogh. What are artificial neural networks? *Nature Biotechnology*, 26:195, 2008.

[26] Benjamin Lindner and Jeremy Smith. Oak ridge national laboratories center for molecular biophysics. Published by Oak Ridge National Laboratories., 2011.

[27] Stuart Lloyd. Least squares optimization in pcm. *IEEE Transactions on Information Theory*, 28:129, 1982.

[28] J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, and Y. Gilad. Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18:1509, 2008.

[29] NASDAQ OMX Group. Nasdaq, April 2012. Accessed online on April 23rd, 2012.

[30] Juan Parrondo and Pep Espanol. Criticism of feyman's analysis of the ratchet as an engine. *American Journal of Physics*, 64:1125, 1996.

[31] Nancy Rabalais, R. Turner, Robert Diaz, and Dubravko Justic. Global change and eutrophication of coastal waters. *ICES Journal of Marine Science*, 66:00, 2009.

[32] William Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846, 1971.

[33] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11:R25, 2010.

[34] Peter Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53, 1987.

[35] Yaqing Si, Peng Liu, Pinghua Li, and Thomas Brutnell. Model-based clustering for rna-seq data. In *Joint Statistical Meeting*, 2011.

[36] Jeremy Smith. personal communication, April 2012. Conversations held with the Department of Computational Biophysics at Oak Ridge National Laboratories.

[37] Daniel Sperling. *Two billion cars: driving towards sustainability.* Oxford University Press, 2009.

[38] Alexander Strehl and Joydeep Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583, 2002.

[39] U.S. Energy Information Association. Annual energy report 2010. Accessed online on April 19th., 2011.

[40] Michael Wang. Updated energy and greenhouse gas emission results of fuel ethanol. Center for Transportation Research, June 2009.

[41] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Review Genetics*, 10:57, 2009.

[42] Kellen Winden, Michael Oldham, Karoly Mirnics, Philip Ebert, Christo Swan, Pat Levitt, John Rubenstein, Steve Horvath, and Daniel Geschwind. The organization of the transcriptional network in specific neuronal classes. *Molecular Systems Biology*, 5:291, 2009.

[43] Wei Zhao, Peter Langfeder, Tova Fuller, Jun Dong, Ai Li, and Steve Hovarth. Weighted gene coexpression network analysis: State of the art. *Journal of Pharmaceutical Statistics*, 29:281, 2010.