5-2017

Graph-Theoretical Tools for the Analysis of Complex Networks

Ronald Dewayne Hagan
University of Tennessee, Knoxville, rhagan@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Ronald Dewayne Hagan entitled "Graph-Theoretical Tools for the Analysis of Complex Networks." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Computer Science.

Michael A. Langston, Major Professor

We have read this dissertation and recommend its acceptance:

Bruce MacLennan, Charles Collins, Jian Huang

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Graph-Theoretical Tools for the Analysis of Complex Networks

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Ronald Dewayne Hagan
May 2017

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Michael A. Langston, for his patience and guidance. My time as his student has provided me with more opportunities than I would have ever believed possible. I would also like to thank the members of my dissertation committee: Drs. Charles Collins, Jian Huang, and Bruce MacLennan. I have been fortunate to work with several talented students on Dr. Langston's research team whose friendship I will always value: Carissa Bleker, Stephen Grady, Clarence Jackson, Allan Lu, Sudhir Naswa, Charles Phillips, Gary Rogers, and Kai Wang. I would like to thank my friends in the Department of Mathematics at UT for their support as I worked towards completing my studies while serving as a Lecturer. I also extend my appreciation to the many great professors that I have had the honor to study under throughout my journey from the completion of my bachelor's degree at Carson Newman College, to my master's degree in the Math department at UT, to now. In particular, I would like to remember Dr. William Wade, whom we lost earlier this year. He was without a doubt one of the greatest educators and minds I have ever known. Finally, my profound thanks go to my family. My wife Cindy, sons Brett and Isaac, brother James, sister-in-law Ada, new nephew Dax, mother Linda, father Ronald, and grandmother Ethel. Without their support and encouragement, none of this would have been possible.

ABSTRACT

We are currently experiencing an explosive growth in data collection technology that threatens to dwarf the commensurate gains in computational power predicted by Moore's Law. At the same time, researchers across numerous domain sciences are finding success using network models to represent their data. Graph algorithms are then applied to study the topological structure and tease out latent relationships between variables. Unfortunately, the problems of interest, such as finding dense subgraphs, are often the most difficult to solve from a computational point of view. Together, these issues motivate the need for novel algorithmic techniques in the study of graphs derived from large, complex, data sources. This dissertation describes the development and application of graph theoretic tools for the study of complex networks. Algorithms are presented that leverage efficient, exact solutions to difficult combinatorial problems for epigenetic biomarker detection and disease subtyping based on gene expression signatures. Extensive testing on publicly available data is presented supporting the efficacy of these approaches. To address efficient algorithm design, a study of the two core tenets of fixed parameter tractability (branching and kernelization) is considered in the context of a parallel implementation of vertex cover. Results of testing on a wide variety of graphs derived from both real and synthetic data are presented. It is shown that the relative success of kernelization versus branching is found to be largely dependent on the degree distribution of the graph. Throughout, an emphasis is placed upon the practicality of resulting implementations to advance the limits of effective computation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE INTRODUCTION AND BACKGROUND

Recent years have witnessed an explosive growth in the amount of raw data available for researchers. Access to such vast stores of data provides exciting opportunities for new discovery, while the very size of the datasets of interest creates ever evolving computational challenges. Even though growth in computing power has continued to observe Moore's Law and double roughly every two years, the gains are being offset, and in many cases dwarfed, by a corresponding growth in data size. In fact, it was estimated in 2013 that 90% of the data available in the world had been generated in the two preceding years alone [1].

At the same time, the study of complex networks such as biological, transportation, social, and communication networks has become ubiquitous across the domain sciences. For example, network models have been used in the elucidation of putative gene networks [2, 3], the study of protein-protein interactions [4], examining the spread of influence through social networks [5], and studying the organization of the human brain [6]. Central to the understanding of such networks are their topological structures. As such, the study of complex networks is intimately interwoven with graph theory and graph algorithms.

In the study of complex networks, we must deal with two distinct hurdles. First, we are faced with the ever-increasing rate of growth of available data sets. The second, and perhaps more insidious challenge, is that the problems we often wish to solve are among the most difficult from a computational point of view. For example, the problem of finding the maximum clique in a graph, its largest fully connected subgraph, is in fact *NP*-complete. Overcoming these difficulties will require the use of novel algorithmic techniques such as Fixed Parameter Tractability (FPT) and the development of scalable parallel solutions.

The primary focus of this dissertation is the development of novel graph-theoretical approaches utilizing exact solutions for classical *NP*-complete problems to analyze the structure of complex networks. The specific settings considered in chapters two and three are from the biological sciences, but the applications transfer easily to other domains. To address the computational challenges the quest for such exact solutions impose, chapter four investigates the relative importance of the two key components of a parallel FPT implementation of vertex cover – kernelization and branching. Finally, chapter five summarizes the results and main contributions, as well as suggesting some avenues for further research.

Data Types

Gene Expression

The history of modern DNA microarrays traces back to the colony hybridization method of Grunstein and Hogness in 1975 [7]. In the roughly four decades since, microarray technologies have proven to play a central role in advancing biological research. Microarrays measure gene expression levels using chips with a set of hybridization probes designed to target and bond to a specific sequence of mRNA. The technology has advanced to the point that platforms such as the Affymetrix Exon 1.0 ST array are capable of exon level resolution of expression with approximately 1.4 million probesets comprised of over 5 million individual probes. For an excellent review of the history, types, and applications of DNA microarrays see [8].

Epigenetic Data

When the Human Genome Project undertook its mission to map the entire DNA sequence of the human genome in 1990, it carried the hope of transforming our understanding of

biology. In 2001 it was even chronicled in an episode of NOVA on PBS entitled "Cracking the Code of Life [9]." While certainly representing a great leap forward in our fundamental knowledge of genetics, it has become clear since its completion in 2003 that there are mechanisms at play in the actual expression of genes that go far beyond the physical arrangement of the underlying genetic code. These discoveries have led to the establishment of a new branch of research – epigenetics. This fledgling field was defined in 1990 by Robin Holiday as "the study of the mechanisms of temporal and spatial control of gene activity during the development of complex organisms [10]." In recent years, however, there has been some disagreement over whether the definition of an epigenetic trait should be limited to those that are heritable [11].

A variety of epigenetic mechanisms have been discovered. Two of the major classes of epigenetic modifications are DNA methylation and histone modification.

DNA methylation generally occurs when a methyl group is added at the 5' position of the cytosine ring, transforming the cytosine to 5-methylcytosine. Usually this occurs at CpG dinucleotides, although non-CpG methylation has been seen to occur more frequently in specific contexts such as neural development and in embryonic stem cells [12]. The process is believed to be regulated by DNA methyltransferases including DNMT1, DMNT3a, and DMNT3b. DNMT1 works to maintain methylation patterns by recognizing and copying them to the unmethylated daughter strands during DNA replication. DNMT3a and DMNT3b are thought to be responsible for *de novo* methylation events. Mutations in the DNMT3b gene have been found to be responsible for ICF (Immunodeficiency, centromeric instability, facial anomalies) syndrome [13], while mutations to any of DNMT1, DNMT3a, or DNMT3b have been found to be embryonically lethal in mice [14, 15].

In humans, some 70% of CpG dinucleotides throughout the genome are methylated [16]. At the same time, there are genomic regions with a heavy concentration of CpG content

that can be found in the promoter regions of many genes. The cytosines in these CpG rich regions, termed CpG islands, tend normally to be unmethylated with exceptions in the context of the inactive X chromosome [17] and imprinted genes [18, 19]. Aberrant methylation patterns have been found to play a role in many diseases. In particular, it has been shown to play a dual role in many forms of cancer through both a pattern of global hypomethylation, allowing aberrant overexpression and ensuing oncogenesis, together with hypermethylation of CpG islands in the promoter regions of tumor suppressor genes, leading to their silencing [20-22].

Histone acetylation occurs when an acetyl group is added to the NH3+ group on Lysine. The process takes place at the N-terminal of histone tails. Acetylation and deacetylation are generally catalyzed by histone acetyltransferase (HAT) or histone deacetylase (HDAC) enzymes respectively. Acetylation acts to transform the overall positive charge of the histone tail to neutral, weakening the binding of the nucleosomal components and making the DNA more accessible to transcriptional agents. Thus, hyperacetylation is positively correlated with actively transcribed genes.

The lysine and arginine residues of the histone tails can be methylated, but it is most commonly observed on the lysine residues of the tails of H3 and H4. Lysine can be mono-, di-, or trimethylated with a methyl group replacing a hydrogen of its NH3+ group. Arginine has a free NH2 and NH2+ group and can be mono- or dimethylated. A demethylation of arginine can occur on a single group, or with an asymmetric methylation of each group.

Although DNA methylation and histone modifications are perhaps the most studied and well known epigenetic mechanisms, a vast amount of ongoing research is also being invested in other post-translational modifications such as phosphorylation, ubiquitination, and somoylation.

Other Amenable Data Types

While the applications we consider in the next two chapters are drawn from the biological domain, our methods are based on abstract models. As such, they can be applied to virtually any type of data that admits such a model. The only requirements are that there must be some entities that can be represented by nodes and we must be able to calculate some similarity metric between them. Data can be continuous like the biological data already considered, or categorical. For example, other work at the Langston Lab has used graph algorithms to examine categorical data from the health disparities domain [23].

Translation to Graphs

Creation of Graphs

In seeking to analyze vast networks, we model the interactions as a graph so that we can apply graph algorithms to better our understanding of the latent relationships. Nodes can then be used to represent entities of interest, be they genes in a gene network, proteins in a protein-protein interaction network, locations in a transportation network, or people in a social network. In order to model the relationships between the actors represented by the nodes, we must have some concept of how to place edges between them. In some contexts, this is as straightforward as connecting “related” nodes, such as placing edges between nodes in a social network if the people they represent are acquaintances or between nodes in a road network if the locations they represent are connected by a road. In others, we need a more mathematical or technical idea of a similarity measure such as correlation coefficients or Jaccard similarity.

In the context of gene expression networks, we often build graphs with nodes representing individual genes. Pearson correlation coefficients are then calculated for each pair of genes taken across the measures for each sample. An edge is then placed

connecting the nodes if the correlation level (in absolute value) exceeds some thresholding value.

Thresholding

There are a variety of methods that can be employed to select a suitable threshold value to use. Often times, we may use methods based on experimentation. For example, when building graphs for a paraclique analysis as in [24, 25], we may build a series of graphs that gradually reduce the threshold until we encounter an inflection point in the number of paracliques produced. Another approach is to use the spectral method introduced in [26], which uses the eigenstructure of the adjacency matrix.

Relevant Graph Theory and Algorithms

Notations

A graph is an abstract representation of a network consisting of a set of vertices and a set of edges that connect pairs of vertices. Formally, a graph $G(V, E)$ consists of a vertex set V and an edge set E containing ordered pairs of vertices from V such that $(u, v) \in E$ indicates the presence of an edge between u and v. A graph is said to be simple if it contains no self-loops or multiple edges. That is to say, there is no vertex with an edge back to itself, and there is at most one edge connecting any particular pair of vertices. A graph can be weighted by adding weight properties to its vertices, its edges, or both. In this dissertation, all graphs considered are simple, finite, undirected and unweighted unless explicitly noted otherwise.

A pair of vertices u and v are said to be adjacent if $(u, v) \in E$. The size, or order, of a graph is taken to be its number of vertices, $|V|$. The degree of a vertex v is the number of vertices to which v is adjacent, or equivalently, the number of edges with v as an endpoint. The neighborhood of a vertex v in G, denoted by $N_G(v)$ or, when the graph is clear in context simply by $N(v)$, is the set of vertices that are adjacent to v. A subgraph of a graph $G(V, E)$ is a graph $G'(V', E')$ such that $V' \subseteq V$ and $E' \subseteq E$, with the requirement that if $(u', v') \in E'$, both u' and v' belong to V'. A subset of vertices can be used to identify a subgraph of G called an induced subgraph. An induced subgraph contains the inducing vertices as its vertex set, and all the edges between those vertices that were present in G. The complement of a graph G, denoted $\bar{G}$, is formed by removing the existing edges, and adding those that were not originally present. Examples of an induced subgraph and the complement of a graph can be seen in figure 1.

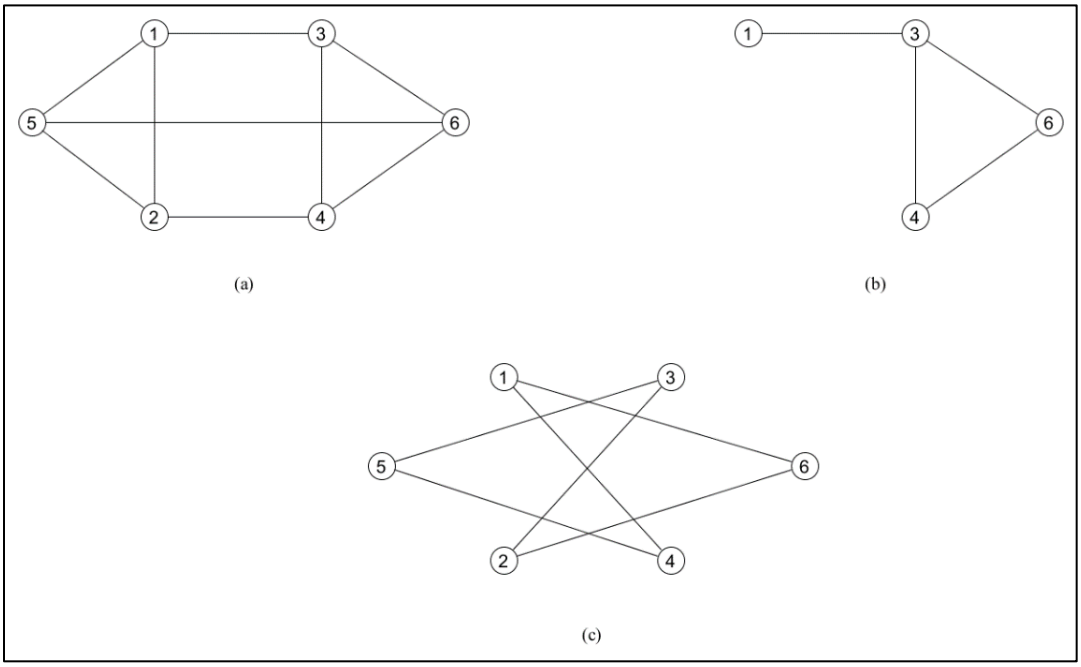

Figure 1. Induced subgraphs and graph complements. Suppose G is the graph shown in (a). Then (b) is its subgraph induced by the set of vertices {1,3,4,6}, and (c) is $\bar{G}$, its complement.

The Clique Problem and the Paraclique Algorithm

A clique in a graph is a fully connected subgraph. In other words, a clique is simply a subgraph containing all of its possible edges. The usual notation for a clique with n vertices is K_n. An example of a K_5 can be seen in figure 2.

The identification of cliques has found a wide variety of applications, from identifying putative gene pathways [27] to detecting collusive trading patterns in the stock market [28]. The maximum clique problem is one of the most studied in graph theory. The decision version, simply deciding if a graph contains a clique of a given input size, is one of the original 21 NP-complete problems listed in Karp's seminal work [29].

In practice, focusing solely on searching for cliques often proves too stringent a requirement. The presence of noise in real world data often leads to missing edges in our graphs. As the loss of even a single edge destroys a clique, such noise results in a higher than desirable number of false negatives when examining the data for signal.

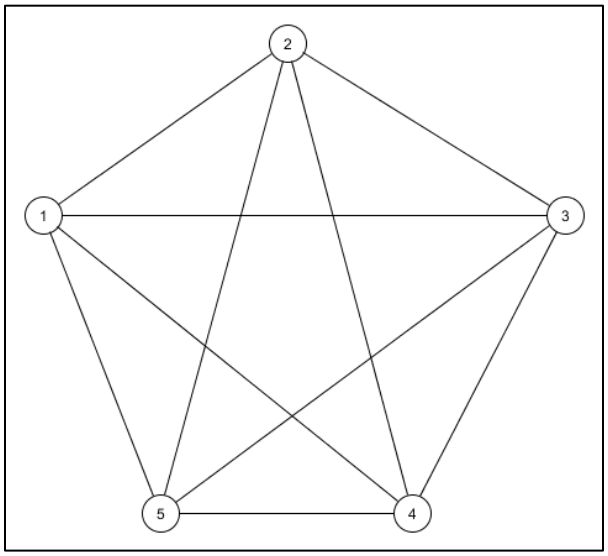

Figure 2. K_5, a clique on 5 vertices.

In order to provide more robust resistance to the effects of noise, the paraclique algorithm was introduced in [25]. The basic idea is that we start with a maximum clique and grow it to a paraclique by successively adding in, or *glomming onto*, vertices that are adjacent to all but some acceptable number of vertices present at the current stage. The number of edges allowed to be missed at each iteration is governed by a constant called the glom term. An example of a paraclique can be seen in figure 3.

The paraclique algorithm has proven to be highly effective, and shown to outperform other popular clustering techniques in terms of density and ontological enrichment [30]. Although most of the work involving the algorithm has to this point focused on its application, theoretical work can be found in [25, 31].

The Dominating Set Problem

We can speak of both edge-dominating and vertex-dominating sets for a graph *G*. In this dissertation, by a dominating set, we will mean a vertex-dominating set, that is, a subset *S* of the vertex set *V* such that for every *v* in *V*, either *v* is in *S* or at least one of its neighbors is in *S*. In particular, we will be interested in a variant of Dominating Set, namely Red-Blue Dominating Set, in which the nodes of a graph are colored either red or blue and we seek the smallest set of red vertices that dominate all the blue vertices (or vice versa). See figure 4 for an example.

Fixed Parameter Tractability

In the past few decades, fixed parameter tractability has emerged as a powerful approach to the design and implementation of practical algorithms for solving problems once

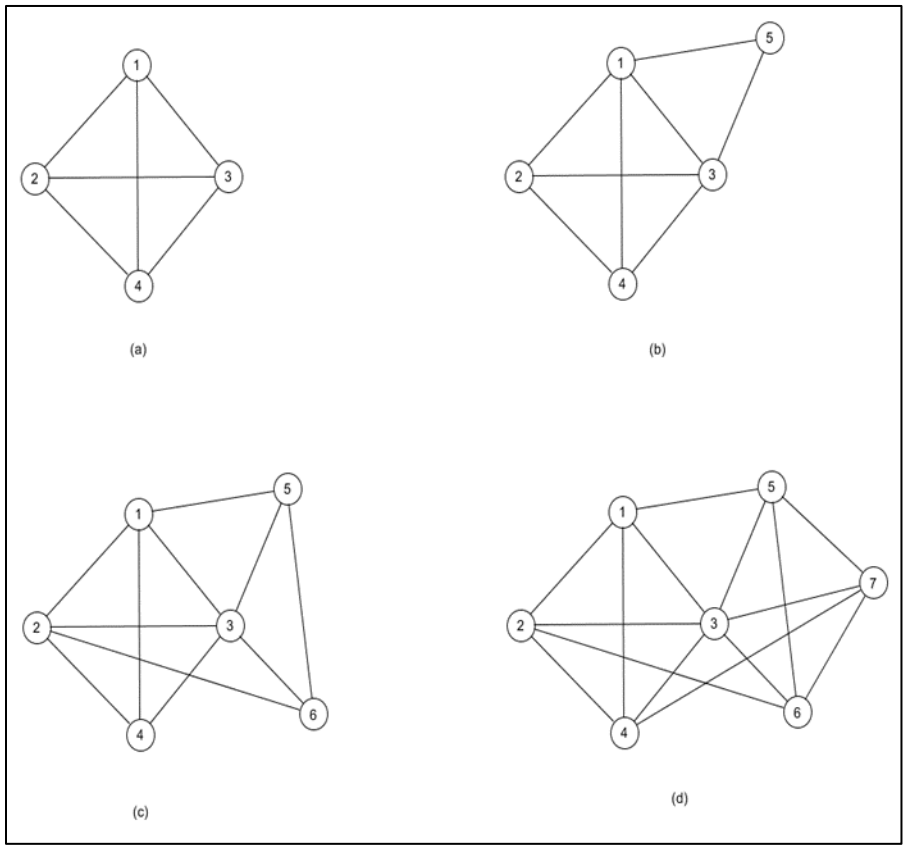

Figure 3. An example of a paraclique. This example was produced with glom term g=2 from a maximum clique of size four {1,2,3,4} (a) and progressively adding vertices 5 (b), 6 (c), and 7 (d).

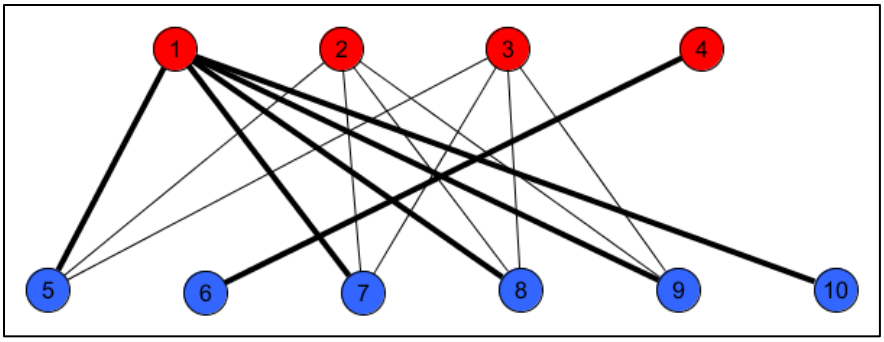

Figure 4. An example of red-blue dominating set. A minimum red-blue dominating set is formed by the set of vertices {1,4}. Note that any such set must contain vertex 4 as it is the only one to cover vertex 6. The covering edges as highlighted for clarity.

deemed hopelessly intractable. The origins of FPT trace back at least to the foundational work of Fellows and Langston in the area of well-quasi-orderings and nonconstructive tools for proving polynomial time computability [32-34]. Around the same time, Robertson and Seymour proved the graph minor theorem, showing that undirected graphs are well-quasi-ordered under the graph minor relationship [35]. These results provided the impetus for the establishment of a new area of research, which would be termed Fixed Parameter Tractability, in the seminal work by Downey and Fellows [36].

One goal of FPT is to provide a more fine-grained approach to classifying the difficulty of a problem than that taken in classical complexity theory. Not all *NP*-complete problems are created equally after all, and perhaps even more importantly, some "difficult" problems may be solved very quickly in practice if we are only interested in solutions of a particular size. The central idea is that if there is some parameter k in the problem that, when held fixed, will allow the problem to be solved in time with any super-exponential dependence being only in k, then the problem is in the class FPT. In other words, a problem is FPT if it can be solved in time $O(f(k)n^c)$ where n is the problem size, k is an input parameter, and c is a constant independent of both k and n.

There are two core tenets to Fixed Parameter Tractability – kernelization and branching. Kernelization refers to a process of reducing the problem size to a compute kernel whose size depends only on the fixed parameter *k*. This reduction must also be doable in polynomial time with regards to the problem size. In [37] it is shown that a problem being kernelizable is in fact equivalent to it being FPT.

Once the problem has been reduced to a compute-intensive kernel, it remains to explore the resulting reduced search space. Various branching techniques aim to explore this space as efficiently as possible. Whereas kernelization techniques aim to reduce the impact of the original problem size in algorithm runtimes, improved branching aims to reduce the remaining, often exponential, dependence on the size of the parameter *k*.

CHAPTER TWO NOVEL ALGORITHMIC DEVELOPMENT: EPIGENETIC BIOMARKER DISCOVERY

In this chapter, we investigate the use of a combination of statistical scoring methods with a red-blue dominating set filter for the elucidation of novel methylation biomarkers. Parts of the material in this section have appeared previously in preliminary form in posters or in papers currently submitted for publication. These will be cited as such wherever appropriate.

Introduction and Overview of Methods

Previous work in the Langston Lab has produced innovative graph-theoretical tools for mining gene co-expression data for transcriptomic biomarkers. In this study, we investigate extensions and applications of these tools to the analysis of differential DNA methylation data. Sites are scored with a statistical metric that gauges their potential effectiveness for separating case data from control. Results are then filtered using red blue dominating set. Inter-sample scores are calculated for sites passing this filter using a custom scoring function that, based on site quality, favors homogeneous case-case and control-control pairs over case-control pairs. This general line of research can be traced back to the seminal toolchain first employed in [39].

Site Merit Scores

To begin, each methylation site is assigned a merit score by means of the following function:

$$score(site\ i) = |\mu_i(case) - \mu_i(control)| - \alpha|\sigma_i(case) + \sigma_i(control)|$$

Where μ_i and σ_i are the mean and standard deviations, respectively, of the indicated sample group and α is a constant used as a tuning factor with $0 < \alpha \leq 1$. We start with $\alpha = 1$ and adjust it downward as necessary until we are able to identify sites with positive merit scores.

Inter-sample Discrimination Scores

Ultimately the goal is to be able to identify sites that are capable of providing a clean separation between case and control. To that end, we next calculate inter-sample scores. The score comparing sample i and sample j is assigned via:

$$\sum score(site\ k) \cdot \left(1 - \left|methylation\_value_{ik} - methylation\_value_{jk}\right|\right).$$

This metric is designed in such a way so as to favor homogeneous over heterogeneous sample pairs. Thus, case-case pairs and control-control matched pairs will tend to receive higher scores than mismatched case-control pairs. Presented with an unknown sample, the user can then compute its scores against the training set and classify it based on the group to which it most closely aligns.

Dominating Set Filter

Our scoring function has the potential to return a large number of sites with positive merit scores. In practice, we would like to be able to winnow these sites down to those with the best potential as discriminatory markers. In such situations, we apply a filter based on red-blue dominating set.

We first construct a bipartite graph with sites as red vertices forming one partite set and samples as blue vertices in the other. For each site, we calculate the p-value of its observed methylation level for each sample. This p-value is calculated in the distribution of the levels at that site across all the group samples of the same type, be it case or control. A site is said to cover a sample and an edge is added between them in the graph if the p-value is greater than .05. This culls from the tails of the distributions and leaves us with observed methylation values that are in some sense “normal” for the sample within its group at each site.

Unfortunately, a straight application of minimum dominating set might sacrifice sites with high merit scores for those with lower discriminatory power based solely on the size of the returned set. To guard against this, we begin with the top scoring site and iteratively add the next highest scoring until the set built up forms a dominating set. We then take a minimum set from among this collection.

In order to visualize the effectiveness of our reduced set in discriminating between case and control samples, we examine the distribution of the inter-sample scores. Presented with an unknown sample, we could then compute its scores against the training set and classify it based on which group it most closely aligns. That is, if it scores highly against case, with lower scores against control, it would be classified as case and vice-versa.

Application: Biomarker Discovery in Human Disease

Parts of this section have appeared in preliminary form in the following research poster: “*Methylation Biomarker Discovery in Age-Related Diseases*”, Ronald D. Hagan, Michael A. Langston, Keystone Symposia Conference on Epigenetic Aging and Aging Related Diseases, May 1-May 5, 2016, Santa Fe, New Mexico.

Age is the single greatest risk known for developing a host of diseases including cancer, dementia and osteoarthritis. Recent research has shown that epigenetic mechanisms too seem to play a key role in these afflictions. Aberrant DNA methylation patterns in particular have been associated with nearly all forms of human cancer. These discoveries provide a compelling impetus for the development of methods for the discovery of novel methylation biomarkers capable of differentiating between healthy and diseased states. Such markers could then potentially be used in screening and diagnosis, and as guides for the selection of therapeutic targets for DNA-demethylating agents.

We applied our method to eight sets of publicly available data obtained from GEO, the Gene Expression Omnibus. These sets were chosen to span a variety of diseases, have a relatively large number of case and control samples, and to come from a common platform. All the sets are from the Illumina Infinium HumanMethylation450 BeadChip array, often referred to as the Illumina 450k methylation array. The data sets are summarized in Table 1.

Osteoarthritis

According to the Arthritis Foundation, osteoarthritis is the most common chronic condition of the joints. Sometimes called degenerative joint disease, it has no specific cause, but is influenced by several factors including age, occupation, obesity, injury and overuse. As well as having a known genetic component, several studies have been conducted that point to epigenetic mechanisms such as DNA methylation [40, 41] and histone modifications [42, 43]. The GEO series GSE63695 consists of methylation data from chondrocyte DNA samples drawn from the hip cartilage of 23 patients with osteoarthritis, knee cartilage of 73 osteoarthritis patients, and 21 hip samples from healthy controls. For the purpose of this study, we discarded the data from the knee cartilage in order to avoid possible confounding issues due to a mixture of tissue types.

Table 1. An overview of our methylation datasets. All sets consist of publicly available data obtained from GEO – the Gene Expression Omnibus.

GEO Series Number	Disease	Case Samples	Control Samples
GSE63695	Osteoarthritis	23	21
GSE66695	Breast Cancer	80	40
GSE54503	Liver Cancer	66	66
GSE61107	Schizophrenia	24	24
GSE52588	Down Syndrome	29	29
GSE40360	Multiple Sclerosis	28	19
GSE62003	Dementia in T2D	18	18
GSE75679	Sjogren's Syndrome	24	24

With $\alpha = 1$, we identified 777 sites with positive merit scores for separation. The top scoring sites mapped to the genes ALX4, ANK1, and ARNT2. Differential expression of, or differential methylation in the promoter regions for, each of these genes has been indicated in the literature as playing a role in the development of osteoarthritis [44-46].

Our dominating set filter identified a set of three methylation sites that covered all samples. As seen in Figure 5, the homogeneous sample pairs clearly cluster toward the high end of the inter-sample scores giving a nice separation and a classifier built to train on the data using our three sites should perform rather well.

In order to verify our hypothesis, we used the Multi-layer Perceptron model available in the scikit-learn python package to construct a fully connected artificial neural net with three hidden layers as illustrated in Figure 6. Using our three sites as the feature set, we then conduct five-fold cross validation testing on the data set. For this Osteoarthritis data, we obtained a mean accuracy of 0.89 indicating that our selected sites are indeed effective for the purpose of classification.

Breast Cancer

Breast cancer in women accounts for one in ten of all new cancers diagnosed worldwide annually [47]. As with all cancers, DNA methylation is known to play a large role in its progression. A host of studies have been undertaken in efforts to improve our understanding of that relationship. For example, see [48-50]. Series GSE66695 consists of methylation data drawn from 40 normal and 80 breast cancer tissue samples.

In the breast cancer data, our scoring metric produced 12,107 sites with positive merit scores for $\alpha = 1$. The top scoring sites mapped to ZFP106, MXRA7, and the tumor suppressor gene ST7. MXRA7 has been found to be differentially expressed in a number

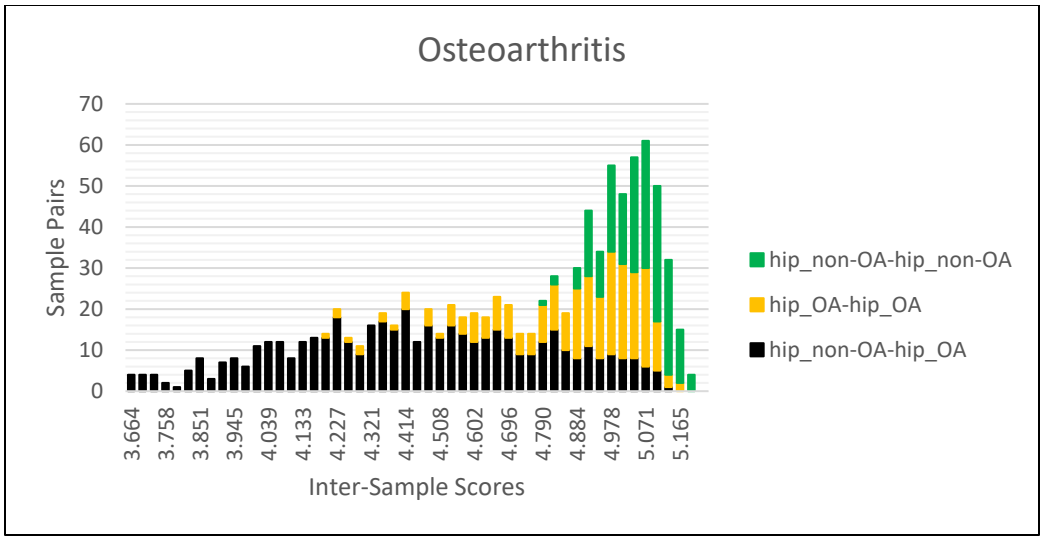

Figure 5. Distribution of Osteoarthritis inter-sample scores.

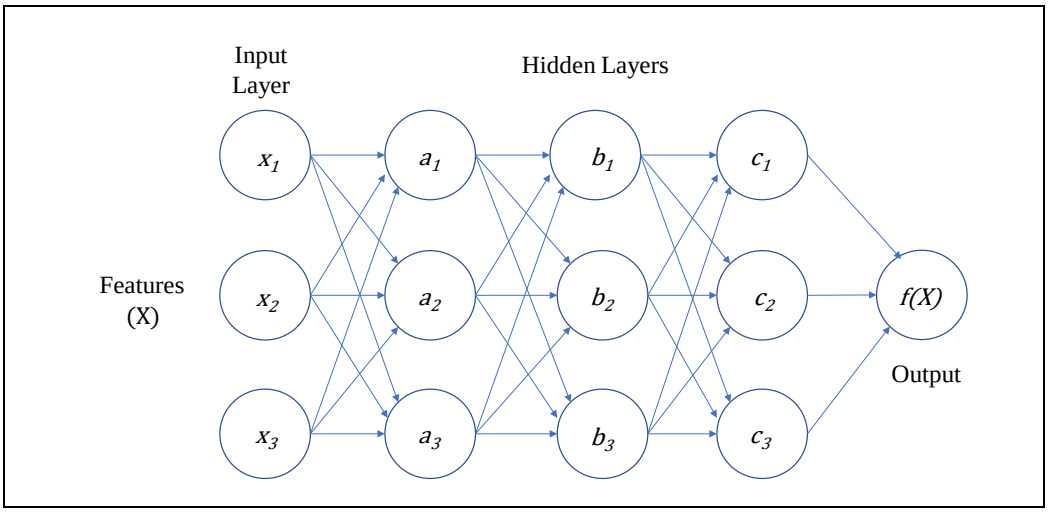

Figure 6. A fully connected neural network with three inputs and three hidden layers. The input layer consists of a set of neurons $\{x_1, x_2, x_3\}$ representing the input features. Each neuron in the hidden layers applies a weighted linear transform, followed by a nonlinear activation function to the values it receives and feeds the result forward through the network. From the results of the final hidden layer, the output layer calculates a result classifying the original input.

of cancers [51]. We were able to uncover a dominating set consisting of five sites separating the data. The distribution of inter-sample scores shows a near total separation of the homogeneous and heterogeneous sample pairs, lending strong evidence to support the utility of our five sites as biomarkers for breast cancer. See Figure 7.

Once again we went a step further, seeking to validate our method using an artificial neural net. Using scikit-learn once again with the same configuration as for the osteoarthritis data, we used the five sites identified by our dominating set reduction as features and conducted five-fold cross validation. In this case, we achieved a mean accuracy of 0.96, strongly supporting our conclusions and also providing good anecdotal evidence that the better the visual separation in the distribution of inter-sample scores, the better the performance of the identified sites as features for classification.

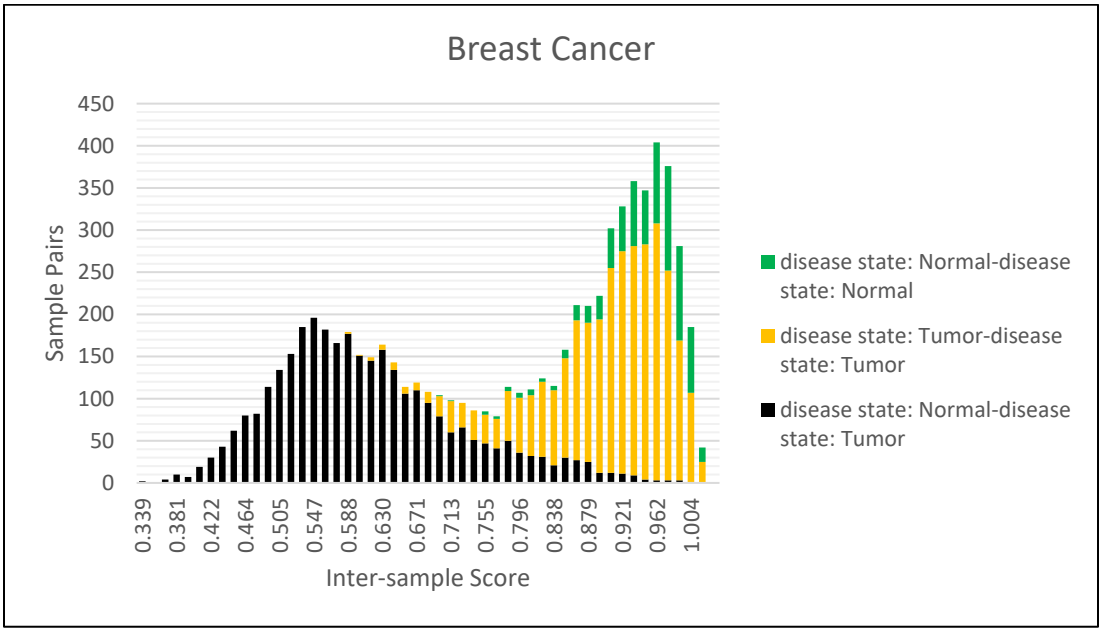

Figure 7. Distribution of Breast Cancer inter-sample scores.

Liver Cancer

The most common type of primary liver cancer is hepatocellular carcinoma or HCC. It ranks as the fifth most common type of cancer globally and is responsible for the third most deaths due to cancers. Despite its high global rankings, the distribution of cases is strongly centred in sub-Saharan Africa and Eastern Asia with China accounting for more than 50% of all cases worldwide [52]. GSE54503 is made up of methylation data drawn from 66 pairs of hepatocellular carcinoma (HCC) liver tumors and adjacent non-tumor tissues.

With $\alpha = 1$, our scoring produced a set of 30,576 methylation sites having positive merit scores. Dominating set filtering produced a set of four probes covering all samples. These sites map to the genes KCNQ2, C1orf70, GRASP, and PTPRN2. All four can be found in the literature as being involved in HCC, see [53-55]. As can be seen in Figure 8, the distribution of inter-sample scores again provides a nearly ideal separation between like and mixed sample pairs.

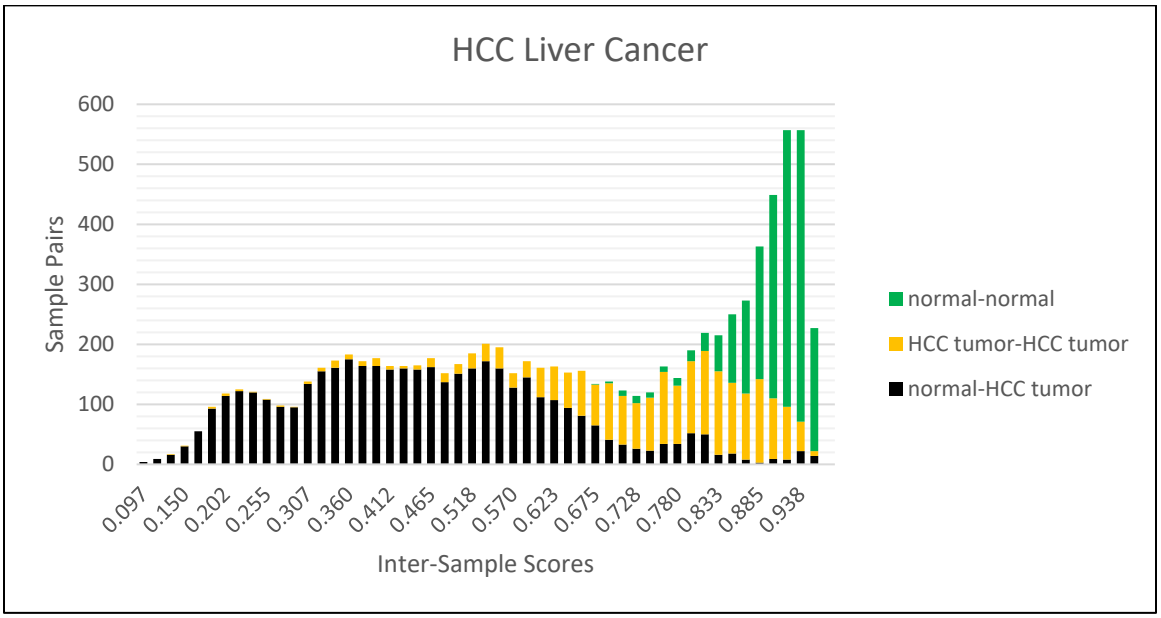

Figure 8. Distribution of HCC inter-sample scores.

Schizophrenia

GSE61107 comes from a genome-wide methylation analysis of brain tissue in schizophrenia patients [56]. It is made up of data drawn from frontal cortex post-mortem tissue from 24 individuals diagnosed with schizophrenia and 24 controls. The tissue samples themselves were provided by the Human Brain and Spinal Fluid Resource Centre.

With $\alpha = 1$, only four sites were identified with positive merit scores. Three of these sites mapped to TNRC6C, ZNF787, and HOXA13, while the fourth mapped to an intragenic region on chromosome 6. HOXA13 appears repeatedly as a potential biomarker for schizophrenia in the literature. See for example [57-59]. While the separation we obtain in this case is not to the level observed with the cancer datasets, we still observe a marked upshift in the distribution of homogeneous inter-sample scores as can be seen in Figure 9.

Down Syndrome

GEO series GSE52588 arises from a study aiming to identify a DNA methylation signature of Down Syndrome in whole blood cells [60]. Data comes from whole peripheral blood samples taken from 29 subjects affected by Down Syndrome as well as 29 matched samples taken from healthy familial controls (either mothers or unaffected siblings).

While the cause of Down Syndrome is established as a trisomy of chromosome 21, the underlying mechanisms influencing the variety of physical and mental manifestations of the phenotype are largely unknown. Perhaps unsurprisingly, recent research has begun to

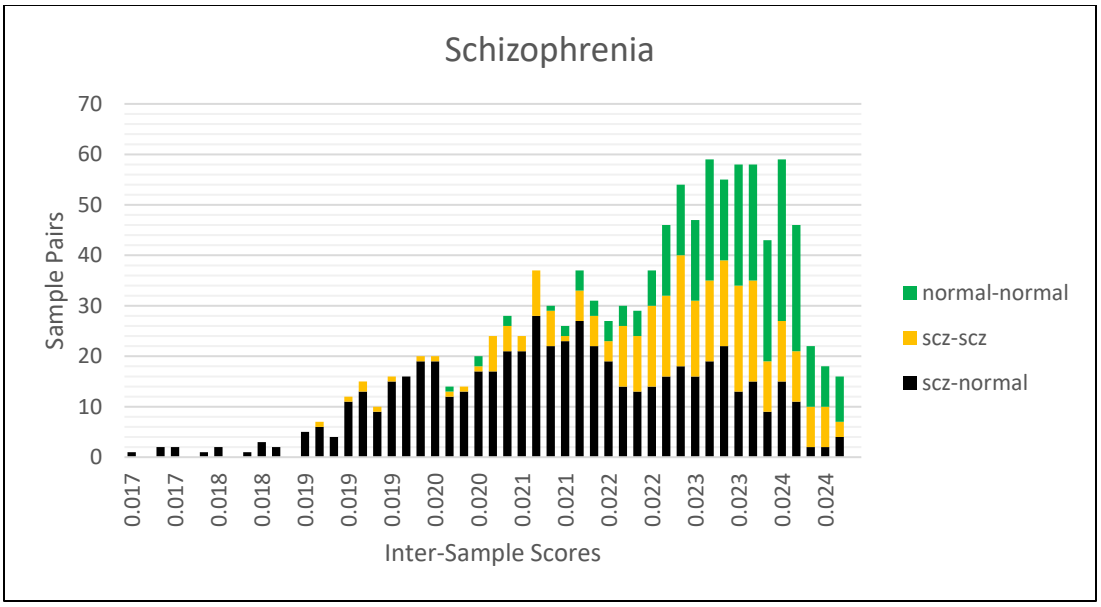

Figure 9. Distribution of Schizophrenia inter-sample scores.

show a strong link to global patterns of aberrant methylation [61]. It has even been suggested that methylation biomarkers could form the basis for non-invasive prenatal screening [62].

With $\alpha = 1$, we initially obtained 2897 sites with positive merit scores. Interestingly, only 42 of these sites map to chromosome 21. As can be seen in figure 10, these sites produce a perfect separation between like and unlike sample pairs.

Multiple Sclerosis

GSE40360 originated from a study seeking to identify differences in methylation patterns in pathology-free regions of brain tissue in persons affected by multiple sclerosis [63]. Drawn from brain bank samples of normal appearing white matter dissected from the frontal lobe, it consists of post-mortem samples from 28 multiple sclerosis patients as well as 19 healthy controls. This particular set turned out to be quite dirty, with numerous

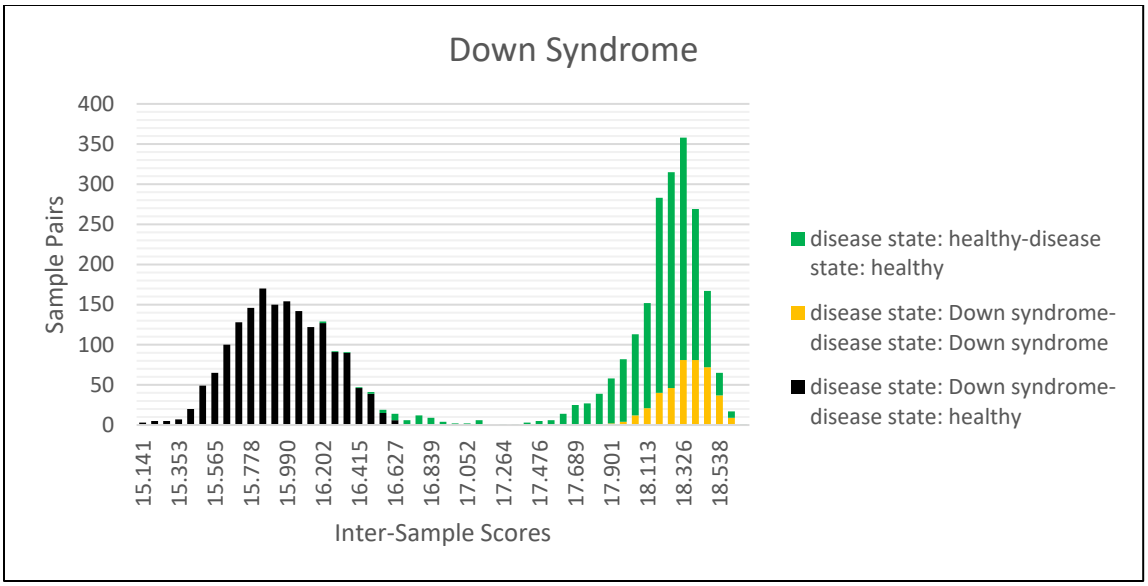

Figure 10. Distribution of Down Syndrome inter-sample scores.

missing values. As such, an initial preprocessing step was required. We chose to discard records for all probes missing entries for a sample, leaving us with data for 460,421 probes.

This is our first dataset that returned no positive scores for a tuning factor of 1. Reducing to $\alpha = 0.9$, we obtained a set of seven sites with positive merit scores. As can be seen in figure 11, we start to see a decreased separation in the distribution commiserate with the need to lower α. Notice however, that the homogeneous sample pairs still produce scores that fall primarily in the top third of the distribution.

Dementia in Type 2 Diabetes

Along with the primary challenges associated with their disease, those suffering from Type 2 Diabetes are at elevated risks for a host of other maladies including Alzheimer's disease. GSE62003 is taken from a study looking for methylation signatures

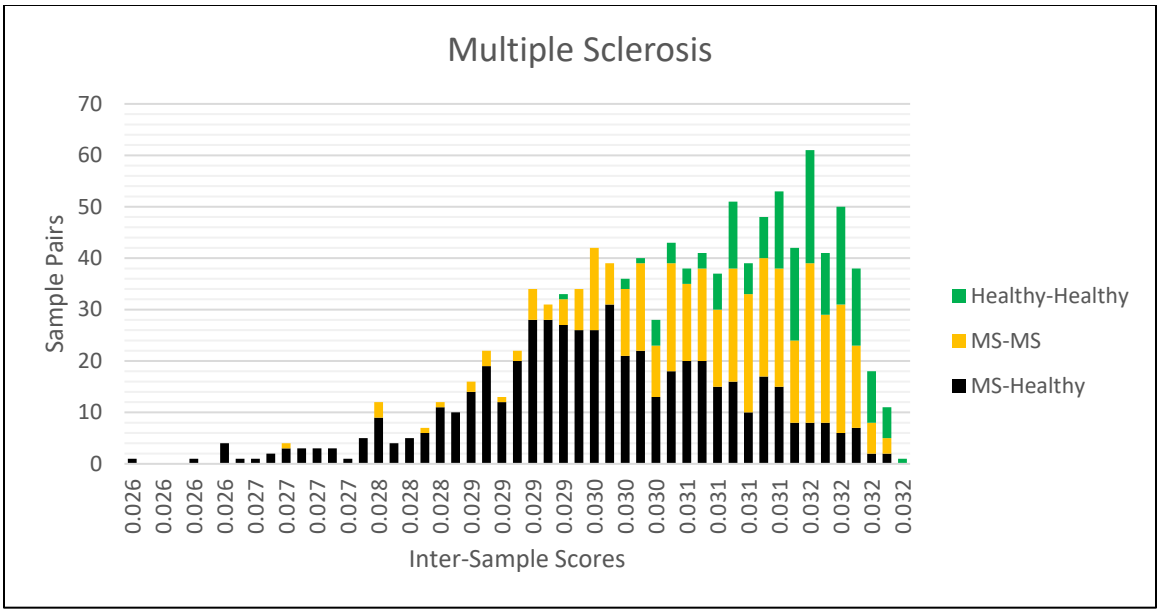

Figure 11. Distribution of Multiple Sclerosis inter-sample scores.

differentiating between risk for developing dementia in elderly T2D patients [64]. It consists of data taken from peripheral blood samples drawn from 18 individuals with T2D who developed pre-symptomatic dementia within 18 months of undergoing a baseline assessment and 18 controls who maintained normal cognitive function. Samples are matched based on age, sex and education.

Initially we were unable to produce sites with a positive merit score for separation. Reducing the tuning factor to $\alpha = 0.8$ produced 14 such sites. A look at the distribution of inter-sample scores in figure 12 again shows a pattern that, while providing a less ideal separation than observed in some previous cases, clearly segregates the majority of homogeneous sample pairs into the top third.

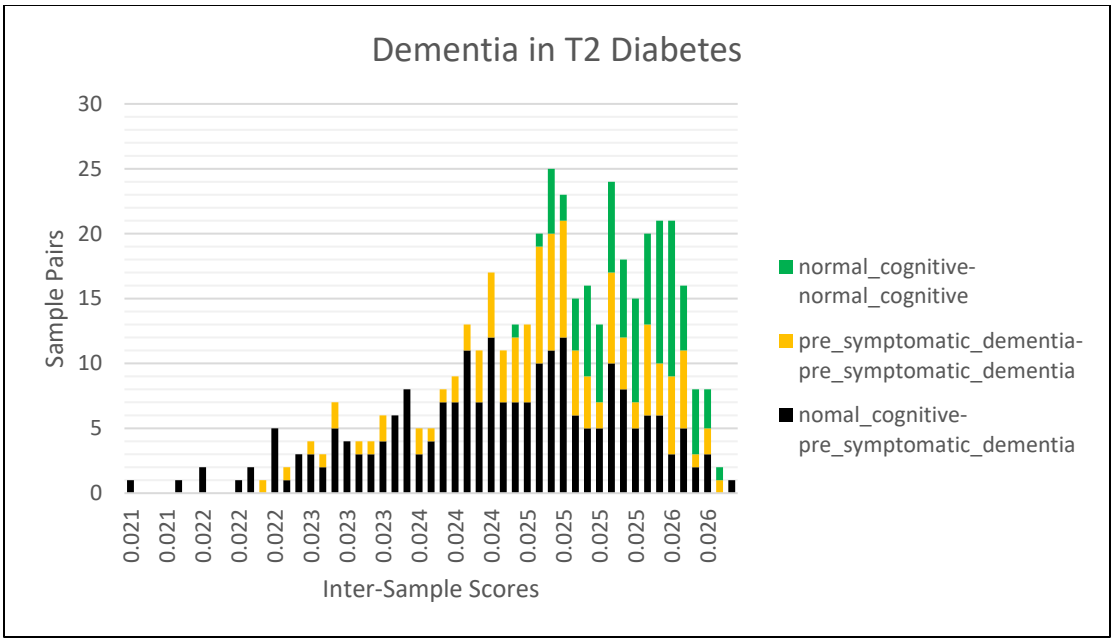

Figure 12. Distribution of inter-sample scores for dementia in Type 2 Diabetes.

Sjogren's Syndrome

Sjogren's syndrome is a systemic autoimmune disorder that affects the entire body, but generally presents initially by attacking the moisture producing glands of the eyes and mouth. Among the many possible complications for those suffering from primary Sjogren's syndrome (pSS) is background chronic fatigue. In [65], the authors study methylation patterns associated with such concurrent fatigue. The associated dataset, GSE75679, consists of methylation data measured in whole blood samples from 24 pSS patients with high fatigue and 24 identified as low fatigue.

Once again, the initial run identified no sites having positive merit scores. Dropping the tuning factor to $\alpha = 0.7$ produced eight sites. As can be seen in figure 13, in this case we observed a distinctive comingling of the homogeneous and heterogeneous inter-sample

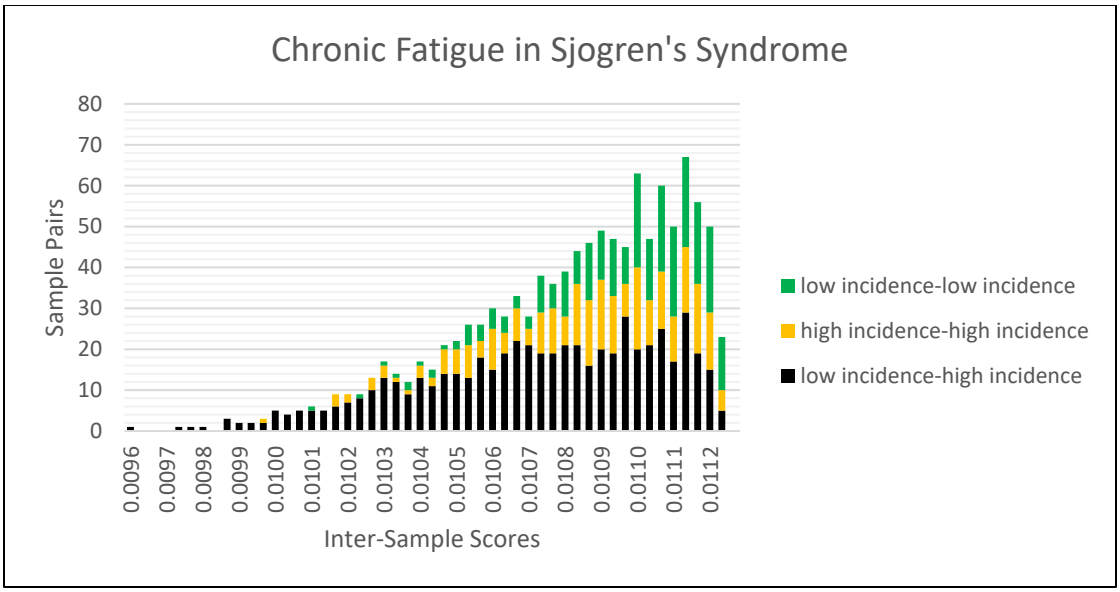

Figure 13. Distribution of inter-sample scores for chronic fatigue in Sjogren's.

scores. This leads us to believe that methylation based biomarkers might not be feasible for identification of susceptibility to chronic fatigue in Sjogren's Syndrome, at least not when drawn from peripheral whole blood.

Summary and Discussion

Ongoing research by the biological community has established that epigenetic mechanisms play a central role in the development of a variety of human diseases, including most known forms of cancer. In this study, we examined the extension of a method previously developed for biomarker discovery in gene expression data to a similar role for use with DNA methylation data. Using a combination of statistical scoring functions and graph-theoretical based filtering methods, our methods identify a set of methylation sites that are putatively best suited to distinguish between case and control samples.

We applied our methods to eight publicly available data sets obtained from the Gene Expression Omnibus that together represent a broad cross-section of human disease. Each set consists of both case and associated control samples. Our results show that our methods have great promise in successfully identifying sets of putative biomarkers. Beyond visual validation via observable separations in the distributions of inter-sample scores, we also tested two of our resultant sets for efficacy as feature sets in the construction of artificial neural network models. These models show the capacity for a high degree of fidelity. Using five-fold cross validation we obtained mean accuracies of 0.96 and 0.89 for models for breast cancer and osteoarthritis respectively.

Application: Obesity Related Biomarkers for Adipose Tissue Differentiation

Most of this section has appeared previously in either a joint paper submitted for publication: "*Genome-wide DNA Methylation Analysis Reveals Loci that Distinguish Different Types of Adipose Tissue in Obese Individuals*", Donia Macartney-Coxson, Miles C. Benton, Ray Blick, Richard S. Stibs, Ronald D. Hagan, Michael A. Langston or in preliminary form in the poster: *"Adipose Tissue DNA Methylation Markers Associated with Weight-loss and Tissue Specificity"*, Donia Macartney-Coxson, Miles Benton, Alice Johnstone, Richard Stubbs, Ronald D. Hagan, Michael A. Langston, Keystone Symposia Conference on Epigenetic Programming and Inheritance, April 6 – April 10, 2014, Boston, Massachusetts.

My contributions include running experiments, enhancing our biomarker toolchain, preparing figures, and writing up the results.

Background

In humans, white adipose tissue consists of two main types, subcutaneous and visceral (including omentum), that are distributed throughout the body in distinct depots. Mitochondria rich brown adipose tissue depots, on the other hand, were once thought to be found only in early development. Recent studies, however, have found higher than expected levels of such depots in adults [66-68]. White adipose tissue can also undergo a process of 'browning' or 'beiging' [69, 70]. Each of these types of adipose tissues have distinct structural and biochemical properties [71-73], and both body fat distribution and function influence metabolic risk [74-81]. Various studies have been undertaken investigating the different developmental origins of subcutaneous and visceral adipose tissue based on gene expression [73, 82-84], with recent evidence suggesting a mesothelial origin for visceral adipose [85]. In addition, studies have begun to focus on the role of DNA methylation in the differentiation and development of the various types of adipose tissue [86, 87].

My co-authors previously performed DNA methylation analyses of paired subcutaneous and omental adipose from obese individuals undergoing gastric bypass, seeking to identify within-tissue differences before and after significant weight-loss (defined as a subject losing more than 27% of their initial weight) [88]. In the current study, we compare methylation levels between subcutaneous adipose and omentum tissue in an effort to identify methylation biomarkers suitable for differentiating between the tissue types both before and after weight-loss.

Identification of Tissue Differentiation Markers

Applying our tools to the pre-operative samples produced 4900 CpG sites, mapping to 2309 genes, with positive merit scores for separating adipose tissue types and passing initial filtering. As can be seen in figure 14, the distribution of inter-sample scores using all 4900 sites produces an appealing, but not quite complete, separation of the tissues. For the post-operative group, we identified 8624 CpG sites, mapping to 4066 genes, such sites having positive merit scores. Once again, the identified sites produce an attractive separation of tissue types, see figure 15. Of the sites identified in the two groups of samples, 1022 produced positive merit scores for separation at both time points (pre- and post-operative).

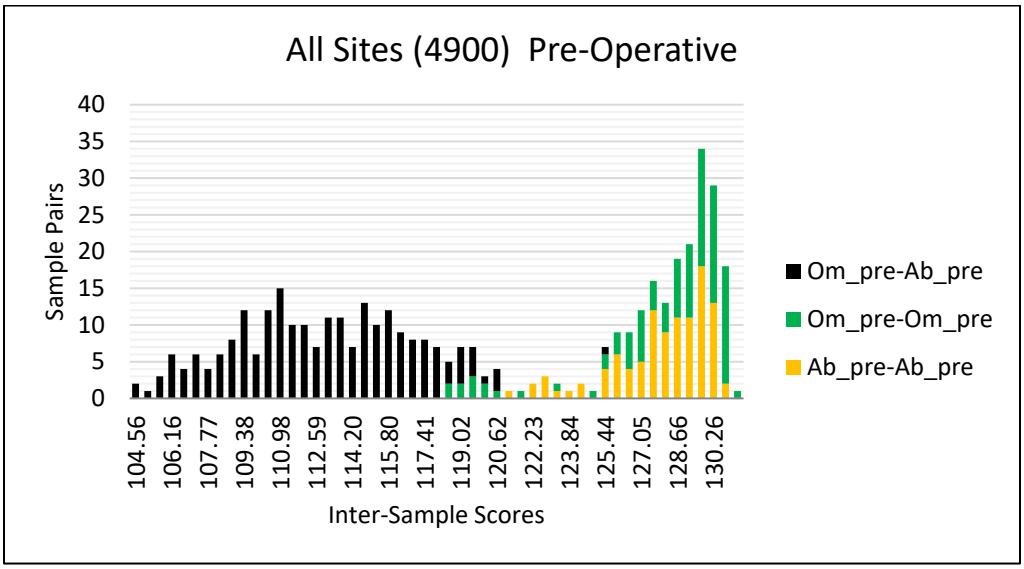

Figure 14. Distribution of inter-sample scores for pre-operative tissue differentiation using all 4900 sites with positive merit scores. Here and in the following figures legends, omental tissue is denoted Om while subcutaneous adipose tissue is denoted by Ab.

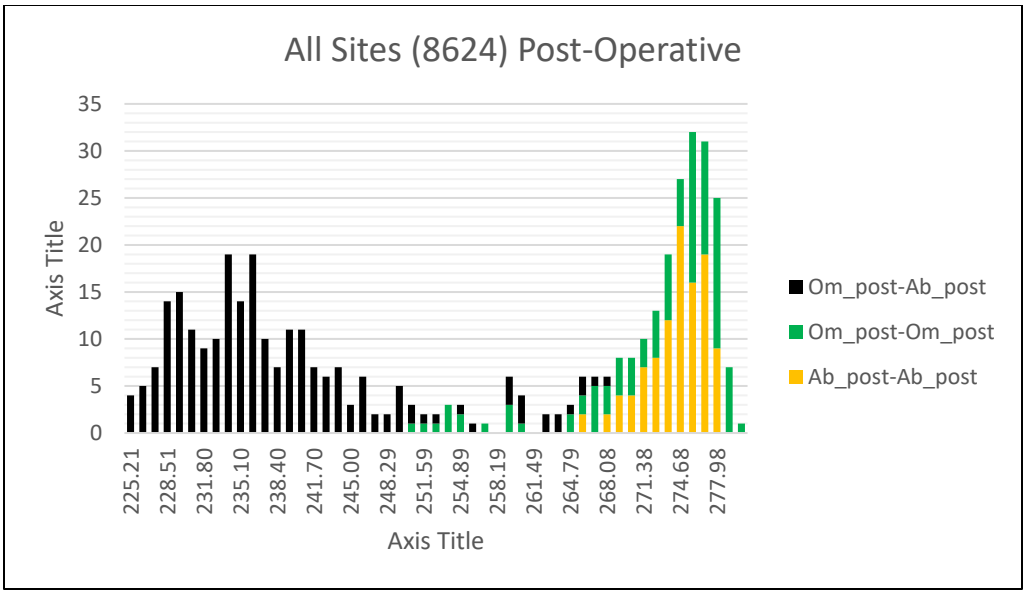

Figure 15. Distribution of inter-sample scores for post-operative tissue differentiation using all 8624 sites with positive merit scores

Our initial analysis considered all sites with positive merit scores. As an ideal biomarker panel would include the fewest markers needed for clean separation between case and control groups, we examined the discriminatory power of a reduced number of sites at each time point. We found that using only the 10 highest ranked sites was sufficient to obtain good separation between the tissue types for both the before and after weight-loss groups, see figure 16 and 17, respectively. We then reduced to the top scoring site for each group. While a single site did not perform as well in the pre-operative group, we found that the top site provided full discriminatory power in the post-operative samples. See figures 18 and 19. An overview of the ten highest ranked sites at each time point can be seen in Table 2.

Validation of Tissue Differentiation Markers

We performed a technical validation of our observations using pyrosequencing of robust tissue discriminators from each analysis, i.e. the top 10 ranking sites from the before

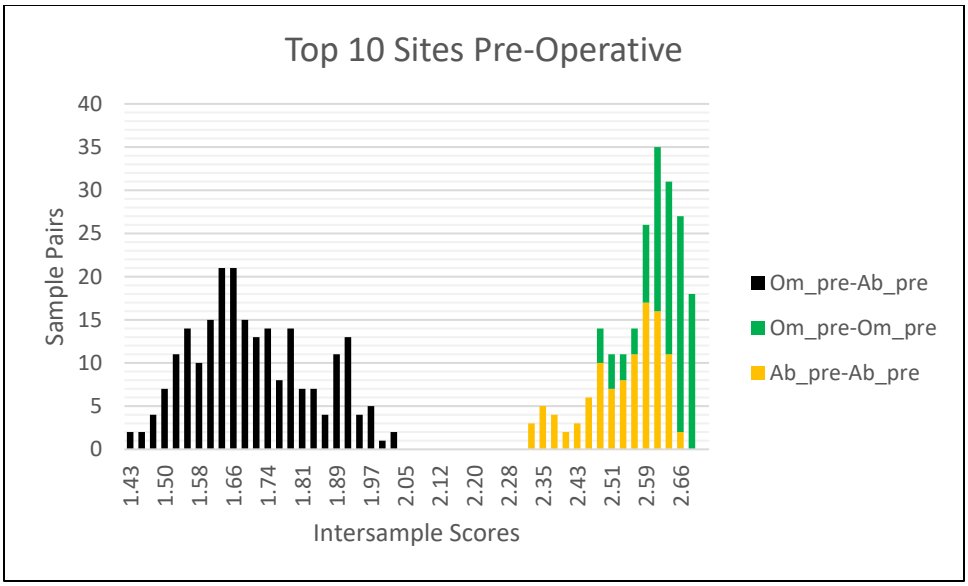

Figure 16. Distribution of inter-sample scores for pre-operative tissue differentiation using the ten sites with highest merit scores.

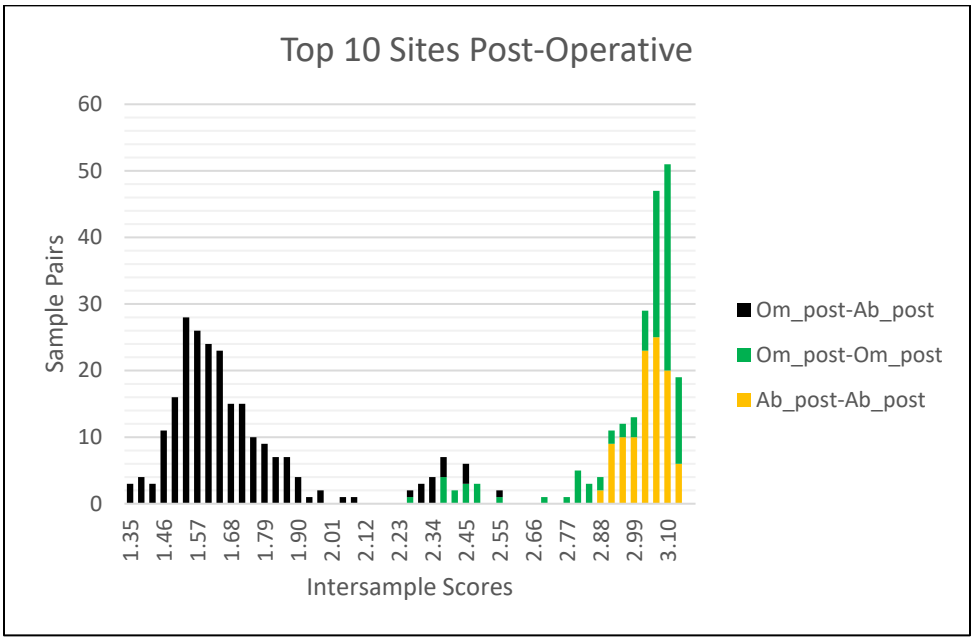

Figure 17. Distribution of inter-sample scores for post-operative tissue differentiation using the ten sites with highest merit scores.

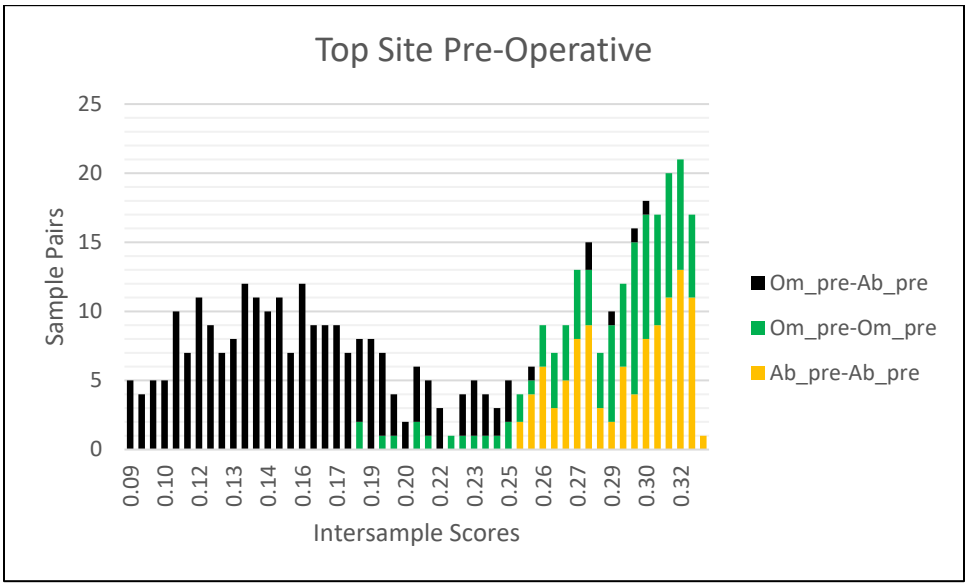

Figure 18. Distribution of inter-sample scores for pre-operative tissue differentiation using the single site with the highest merit score.

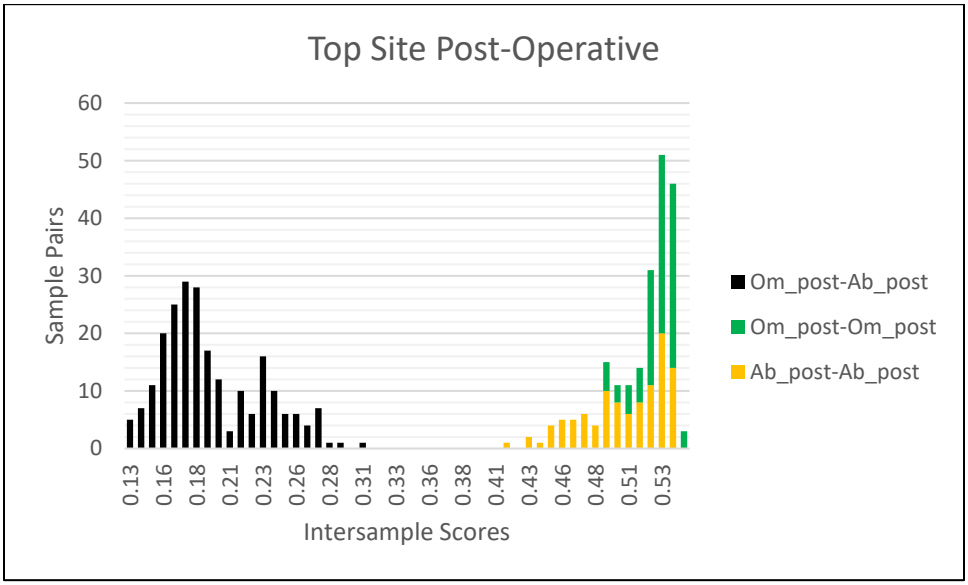

Figure 19. Distribution of inter-sample scores for post-operative tissue differentiation using the single site with the highest merit score.

Table 2. Top 10 CpG sites ranked by merit score.

Illumina probe ID	Mean beta AB (+/- standard deviation)	Mean beta OM (+/- standard deviation)	UCSC Gene Name	CpG site chromosome and position[+]
Before weight-loss analysis				
cg02245004	0.12 (+/- 0.08)	0.64 (+/- 0.11)	Intergenic	15:76634887
cg03923561	0.45 (+/- 0.07)	0.04 (+/- 0.02)	*HOXC4*	12:54447220
cg22747076	0.42 (+/- 0.09	0.02 (+/- 0.02)	*HOXC4*	12:54447873
cg09400037	0.42 (+/- 0.06)	0.82 (+/- 0.07)	Intergenic	16:84822801
cg24376776	0.03 (+/- 0.01)	0.33 (+/- 0.03)	Intergenic	10:101297245
cg11797364	0.59 (+/- 0.1)	0.97 (+/- 0.02)	Intergenic	6:436969
cg17496661	0.65 (+/- 0.08)	0.99 (+/- 0.01)	Intergenic	5:3326343
cg09720701	0.37 (+/- 0.05)	0.06 (+/- 0.01)	*HOXC4*	12:54447283
cg02264990	0.35 (+/- 0.06)	0.03 (+/- 0.02)	*HOXC4*	12:54447243
cg01524853	0.43 (+/- 0.05)	0.12 (+/- 0.03)	*HOXC4*	12:54447807
After weight-loss analysis				
cg00838040	0.34 (+/- 0.07)	0.97 (+/- 0.03)	*ATP2C2*	16:84446919
cg21917524	0.31 (+/- 0.07)	0.81 (+/- 0.04)	Intergenic	11:74200334
cg24145118	0.84 (+/- 0.07)	0.31 (+/- 0.16)	Intergenic	10:2777041
cg12984729	0.09 (+/- 0.03)	0.59 (+/- 0.16)	*ISL2*	15:76633817
cg12437821	0.46 (+/- 0.04)	0.05 (+/- 0.09)	Intergenic	12:114852027
cg01184975	0.53 (+/- 0.05)	0.10 (+/- 0.11)	Intergenic	12:114852091
cg20291855	0.03 (+/- 0.03)	0.44 (+/- 0.11)	Intergenic	4:13524143
cg25365014	0.79 (+/- 0.05)	0.41 (+/- 0.05)	Intergenic	5:2727713
cg16297011	0.07 (+/- 0.04)	0.48 (+/- 0.11)	Intergenic	4:13539023
cg21982455	0.56 (+/- 0.04)	0.19 (+/- 0.07)	Intergenic	5:2757976

weight-loss and the top site from the after weight-loss analyses. This revealed an excellent agreement between the two methylation assays; before weight-loss (R^2 = 0.91-0.98, P = 1.2x10^{-15}–8.3x10^{-25}), and the single site after weight-loss (R^2 = 0.97, P = 4.9 x 10^{-14}).

Next, we sought to test the performance of our candidate biomarkers in additional, independent samples. DNA was available for a further 15 individuals before weight-loss and 13 of these 15 individuals after weight-loss. Principal component analysis (PCA) revealed a strong agreement between the discovery and validation samples for the before weight-loss comparison, with no significant variation observed between genders in the validation samples as can be seen in figure 20.

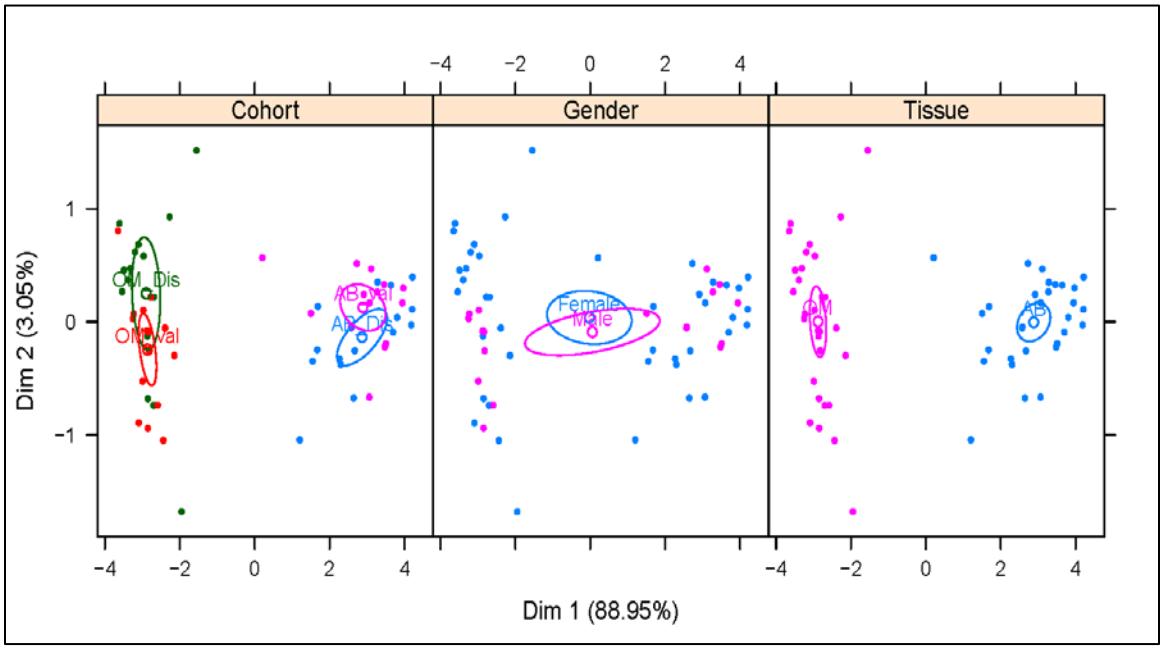

Figure 20. PCA analysis of the top 10 CpG sites from the before weight-loss tissue samples. Analysis was performed on pyrosequence data. Tissue type is indicated by OM (omentum) and subcutaneous abdominal adipose (AB), with Dis indicating the discovery and Val the validation samples.

We were able to obtain publicly available Illumina 450K data for a number of adipose samples. Samples accessible through MARMAL-AID [89] included samples from 6 lean male individuals with both subcutaneous and omental adipose tissue collected ≤ 12h post-mortem (GSE48472 [90]), 14 visceral adipose samples from severely obese men, 7 of which had metabolic syndrome (GSE54776 [91]), as well as 6 paired subcutaneous adipose and gluteal adipose samples from lean females (GSE47513 [86]). Illumina 450K data was also available for 642 subcutaneous adipose samples from individuals with an average BMI of 26.7 (and a standard deviation of 4.9) [92] from the MuTHER (multiple tissue human expression resource) project [93]. In examining the methylation profile of all 11 strong discriminators from the before and after weight-loss groups in these publicly available datasets, we observed good agreement within a tissue type irrespective of gender and/or obesity phenotype.

In general, omental samples showed a tighter distribution of methylation for a given probe than the subcutaneous adipose samples, and for the majority of probes a strongly hyper-methylated (mean methylation beta >0.8, cg00838040, cg17496661, cg17496661) or hypo-methylated (mean methylation beta <0.2, cg09720701, cg02264990, cg22747076, cg03923561, cg22747076) phenotype. Given that the biomarker analysis was trained on data from subcutaneous and omental adipose it is interesting to note that the samples of subcutaneous gluteal adipose tissue from lean females showed a methylation profile similar to that of the subcutaneous abdominal adipose samples and clearly distinct from the omental samples.

Summary and Discussion

In this study, we applied our methylation biomarker tools to isolate CpG sites capable of differentiating between white adipose tissue in human subjects both before and after undergoing gastric bypass surgery. We identified a large set of sites with positive merit

scores that we filtered to reveal 10 CpG sites before weight-loss and a single site after weight-loss that strongly separated the tissue types. Additional confidence in our results is provided by an excellent overlap between the pyrosequence methylation profiles of the 11 CpG sites in the discovery and validation cohorts and from publicly available subcutaneous and omental adipose Illumina 450K data from lean, overweight and obese individuals.

While there is no clinical utility in a marker to differentiate subcutaneous and omental adipose, this study provides further support for the potential of DNA methylation as a biomarker. Combined with the promising results of our work with disease data described in the previous section, we believe future work to translate our combinatorial approach to the detection of clinically applicable DNA methylation biomarkers is warranted, and may have particular merit for situations in which robust differentiators are still urgently required. Furthermore, because epigenetic mechanisms are dynamic and as such potentially reversible, these types of analyses may also highlight innovative new avenues for clinical treatment. Such analyses should take account of practicability concerns such as ease of sample/tissue collection, and whether there is a biological argument for potential DNA methylation differences between sample groups.

CHAPTER THREE NOVEL ALGORITHMIC DEVELOPMENT: SUBTYPING

In this chapter, we turn to gene co-expression data and the development of a method based on the paraclique algorithm to identify unknown subtypes in human disease. A portion of the material in this section appear in [94]. My contributions to this paper include algorithm development, data aggregation, and experimentation and document preparation.

Introduction

The ability to identify disease subtypes accurately and efficiently is a central pursuit in the drive to individualized medicine. In the case of cancer patients, knowledge of a genetic propensity towards chemo-resistance or towards early response can be used to tailor treatment to be more, or less, aggressive. Early responders, for example, might require less aggressive treatment, mitigating the long-term risks of adverse effects of radiation therapy or chemotherapy associated toxicity. For example, studies have identified gene expression signatures for chemo-resistance in both acute myeloid leukemia and breast cancer [95, 96].

While techniques grounded in graph theory have been used to great effect in the pursuit of genetic biomarkers for human disease and the discovery of novel gene networks, little work has been done in trying to extend such tools to subtyping. The existing research has primarily centered on the use of basic statistical or machine learning clustering methods such as k-means, latent variable models, or mixture models [97-99].

Clique-centric methods have often been used for the discovery and modeling of coherent networks. Unfortunately, clique finders are inherently prone to high false negative rates.

Indeed, an entire clique will be lost if even a single edge is missing. The paraclique algorithm was introduced to address the difficulties presented by signal loss due to noise in the data. While paracliques can be constructed in different ways, the basic idea is to take a maximum clique as a core, then expand it to a paraclique by adding vertices adjacent to all but some allowable number g of its vertices.

Here, we present a method for subtyping based on the paraclique algorithm. The remainder of this chapter is organized as follows. An overview of our method and the data sets initially used for experimentation are provided in the next section. We then discuss the initial validation testing and examine additional steps taken to verify the biological relevance of our results. Finally, we provide avenues for continued research.

Methods and Data

In an effort to limit the effects of confounding factors, we use an initial filtering step. FDR corrected p-values for differential expression of genes between case and control sets are calculated, and only those with p-values less than 0.1 are retained. The idea being that this technique will limit our attention to those genes of interest only in the case group. After all, we are not interested in subgroups based on age, ethnicity, or hair color. The p-values are calculated using the EntropyExplorer R package [100]. After filtering, the data tables are transposed and we calculate the pairwise Pearson correlation coefficients between samples across the expression levels. Thresholding the correlation matrix, we then create an unweighted graph with vertices representing individual samples and edges between vertices if the corresponding samples are correlated above the thresholding level (in absolute value). Once the graph is built, we run the paraclique algorithm to extract dense subgraphs of patients representing putative subtypes in the case samples. Finally, we separate the groups, return the tables to their original configuration, and calculate differential expression across subtypes. Here, the level of differential expression is taken

to be simply the difference in the mean expression levels between the groups. An overview of our method can be found in Figure 21.

We initially applied or methods to 12 sets of publicly available gene expression data obtained from the Gene Expression Omnibus (GEO). The data sets were selected to provide a wide cross-section of human disease and to have information on both a case and control group for initial filtering. Table 3 gives an overview of the datasets used.

Initial Validation and Discussion

Our investigation into the effectiveness of our proposed method was focused on two guiding questions. First, is the method capable of reliably identifying putative subtypes? Second, is there biological evidence confirming or at least supporting these subtypes as being biologically relevant to the associated disease?

We found that the answer to the first question is unequivocally yes. As summarized in Table 4, we found putative subtypes in 10 of our datasets. In the other two cases, we hypothesize that our method fails due to limiting conditions present in the data. First, the sizes of the datasets are relatively small (only 10 samples in both cases) and may not give a true representation of the underlying populations. Secondly, the disease in question might not have meaningful subtypes based on differential gene expression.

The second question is considerably more difficult to answer. In order to address it, we followed a two-prong approach in examining the top differentially expressed genes between subgroups. To begin, we calculated the GO enrichments of the top 100 genes

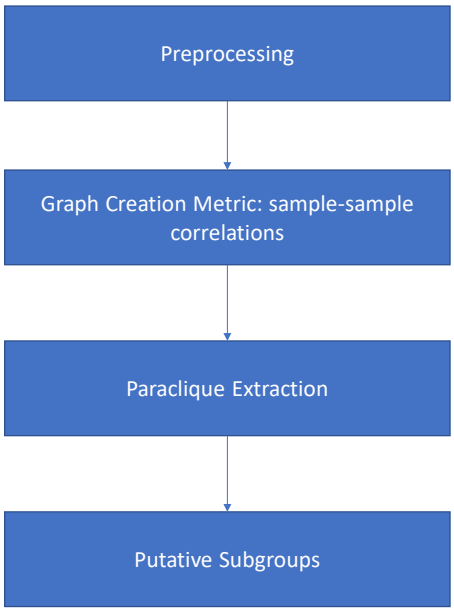

Figure 21. Suptyping method overview. A greatly simplified view of our subtyping workflow. Gene expression data is preprocessed with an initial filtering step. A graph is created using the pairwise sample correlation scores taken across the expression scores. Paracliques are then extracted that represent putative subtypes.

Table 3. Subtyping datasets. An overview of the datasets used in this study.

Disease	GEO Accession	Probes	Probes After Filtering	Case	Control
Asthma	GSE4302	54675	2322	42	28
Breast Cancer	GSE10810	18382	11531	31	27
CLL	GSE8835	22283	1338	24	12
Colorectal Cancer	GSE9348	54675	22968	70	12
Lung Cancer	GSE7670	22283	7458	27	27
Multiple Sclerosis	GDS3920	54674	9844	14	15
Pancreatic Cancer	GDS4102	54613	23711	36	16
Parkinson’s	GSE20141	54674	6625	10	8
Prostate Cancer	GSE6919	12625	1531	61	63
Psoriasis	GSE13355	54675	29407	58	58
Schizophrenia	GSE17612	54675	4250	28	23
T2 Diabetes	GSE20966	61294	93	10	10

Table 4. Subgroups identified. Summary of the number and sizes of putative subgroups identified by our methods in the test data.

Disease	Number of Subgroups Identified	Size of Subgroups Identified
Asthma	2	32,8
Breast Cancer	2	22,5
CLL	2	4,18
Colorectal Cancer	2	63,5
Lung Cancer	2	21,5
Multiple Sclerosis	2	11,3
Pancreatic Cancer	2	31,5
Parkinson’s	1	8
Prostate Cancer	2	56,3
Psoriasis	2	49,5
Schizophrenia	2	19,6
T2 Diabetes	1	9

and the associated enrichment p-value. The results, along with the associated enriched GO categories, are summarized in Table 5. We next performed a literature search to check the lists of the top genes for involvement in known subtypes. We were able to find strong empirical evidence to support having successfully identified known groups in four of our datasets: Asthma, Breast Cancer, CLL, and Colorectal Cancer. Before we go further, let us look at these four cases in more detail.

Asthma

The incidence of asthma in the U.S has been on the rise for two decades. It is currently estimated that 9.6% of children under 18 are asthmatic, with the risks for some groups based on ethnicity (particularly African American and Puerto Rican) and stratification by lower socioeconomic status rising as high as 16% [101].

The GEO series GSE4302 is derived from a study designed to identify genes associated with response in asthmatics to treatment with corticosteroids [102]. It consists of expression data derived from epithelial airway brushings taken from 42 asthmatics and a control group consisting of 28 healthy subjects and 16 smokers. For our analysis, we used only the healthy subjects as control, discarding the smokers. Data is derived from a microarray analysis using the Affymetrix Human Genome U133 Plus 2.0 Array.

Our initial filtering step reduced the original 54,676 probes to a set of 2322 having FDR corrected p-values < 0.1 for differential expression between the case and control groups. With the reduced set of variables and a threshold of 0.93, our method produced three putative subgroups of size 31,8, and 3. The 100 most differentially expressed genes between the groups included CLCA1, periostin, and ovalbumin. All three of these genes were reported in [103] to be markers of a Th2-high endotype of asthma.

Table 5. GO enrichment. The GO term category with the lowest p-value for the enrichment of the 100 most differentially expressed genes across identified subgroups.

Data Set	**Category**	**p-value**
Asthma GSE4302	Oxireductase	1.1E-4
Breast Cancer GSE10810	Secreted	1.0E-13
CLL GSE8835	Mhc ii	2.4E-15
Colorectal Cancer GSE9348	Translational elongation	2.8E-28
Lung Cancer GSE7670	Secreted	7.7E-10
Multiple Sclerosis GDS3920	Translational elongation	1.9E-34
Pancreatic Cancer GDS4102	Signal	4.59E-15
Prostate Cancer GSE6919	Translational elongation	4.92E-46
Psoriasis GSE13355	Immune response	3.5E-15
Schizophrenia GSE17612	Organelle membrane	5.24E-4

Breast Cancer

Genetic factors have long been known to play a significant role in breast cancer. In families that have had at least 4 cases of breast cancer, studies have shown the majority of cases can be linked to mutations in either BRCA1 or BRCA2 genes [104, 105]. In addition, breast cancer has a variety of known subtypes that significantly impact prognosis and treatment. For example, tumors negative for estrogen receptors, progesterone receptors, and the expression of HER2 are indicative of triple-negative breast cancer, a subtype identified with higher risk of recurrence and 5-year mortality [106].

GSE10810 comes from a study aiming to investigate more fully the links between gene expression and phenotypic differences in breast cancer [107]. It consists of gene expression data for 31 tumor samples and a control set of 27 healthy tissue samples. The platform used for the study was again the Affymetrix Human Genome U133 Plus 2.0 Array, although only data for 18,382 probes was provided.

The number of probes was reduced to 11,531 with false discovery rate adjusted p-values for differential expression between the case and control groups < 0.1. Using a threshold of 0.8, our tools produced 2 putative subgroups of size 22 and 5. The 100 most differentially expressed genes between these subgroups include SLC39A6, S100a4, AGR3, Cd24, and epcam. All of these genes have been reported in the literature as being markers for different phenotypes of breast cancer [108-112].

Chronic Lymphocytic Leukemia

Chronic Lymphocytic Leukemia (CLL) is one of the most common types of leukemia with pathogenesis presenting as an overproduction of neoplastic B cells in the

bloodstream. The median age at diagnosis is 65, and more commonly affects males [113]. It typically presents with a slow progression, with patients living out a normal life expectancy. In some cases, however, it can be aggressive, with death occurring less than 5 years after onset.

The dataset GSE8835 was provided by a study with the aim of examining the effects of CLL on the expression levels in peripheral blood T cells [114]. It is made up of 24 CD4 cell samples from CLL patients and a control group of 12 CD4 cell samples from healthy, age matched donors. The Affymetrix Human Genome U133A Array with 22,283 probes was used.

Filtering reduced the initial probes to a set of 1338 having p-values less than 0.1 (again FDR corrected). A threshold of 0.8 produced two subgroups of size 4 and 18. The most differentially expressed genes across the two groups included ZAP-70, previously identified as the best discriminator of Ig-mutated and Ig-unmutated CLL [115].

Colorectal Cancer

The incidence of colorectal cancer (CRC) has been in decline since the mid 1980's [116]. Despite this significant drop in prevalence, it still annually accounts for both the third highest number of new cases of cancer, and the third highest number of cancer deaths [117]. As in breast cancer, there are known hereditary links to CRC. For example, a mutation of the gene APC is responsible for two syndromes, Familial Adenomatous Polyposis and Hereditary Nonpolyposis Colorectal Cancer, that each carry a significant increase in the risk of developing CRC [118].

GSE9348 was derived from a study aiming to find a gene expression signature for cases of early stage CRC that are prone to metastasis [119]. It consists of gene expression data

derived from tumors of 70 patients and tissue from 12 healthy controls. The samples are age and ethnicity matched. The study utilized the Affymetrix U133 Plus 2 array.

Our filtering reduced the 54675 probes in the original set to 22968 with FDR corrected p-values < 0.1 for differential expression between the case and control groups. With a threshold of 0.87, our tools produced two paracliques representing putative subgroups in the tumor samples of size 63 and 5. The list of 100 genes most differentially expressed between these two groups include Cd24, identified as a prognostic marker for CRC [120] as well as OLFM4, indicated in as a marker for tumor differentiation and progression [121, 122].

Further Validation: Testing with Known Subtypes

While our initial validation efforts indicate that our methods indeed have the potential to identify both known and novel subtypes based on biologically relevant genetic signatures, the lack of an established ground truth is a distinct limitation for our testing. In order to address this point, we identified two additional sets of publicly available data on GEO that included labeling for membership into distinct subtypes. The goal of course being to apply our tools in a search for evidence that we could identify subgroups appropriately stratifying the samples.

Gastric Cancer

The first set is GSE35809, consisting of gene expression data from 70 primary gastric tumors used as a validation set for testing of a classifier for tumor subtypes [123-125]. The samples are identified in the data as belonging to one of three subtypes: proliferative, invasive, or metabolic. The number of samples for each type in the original data are 29, 26, and 15 respectively.

As this set does not contain accompanying control data, we forgo the initial filtering phase. Applying the remainder of our toolchain and using a threshold of 0.955, we identified two subgroups that, while both contained a mixture of metabolic tumors, nearly perfectly segregated the invasive and proliferative types. See Table 6.

Non-Small Cell Lung Cancer

Non-small cell lung cancer, or NSCLC, has two major known subtypes – adenocarcinoma (AC), and squamous cell carcinoma (SCC). GSE10245 is made up of gene expression data for 40 adenocarcinoma, and 18 squamous cell carcinoma NSCLC tumors and was provided from a study examining differential expression between the two types [126].

Table 6. Known subtype results. A breakdown of the composition of the subtypes obtained from the cases with included subtype labeling. In both cases, we obtained putative subgraphs that cleanly segregated the samples according to subtype.

Gastric Cancer			**NSCLC**			
	Paraclique Sizes			**Paraclique Sizes**		
	29	16		26	12	8
Subtype			**Subtype**			
proliferative	1	12	AC	23	0	8
invasive	19	1	SCC	3	12	0
metabolic	9	3				

Once again, the set lacks control data, so the initial filtering must be skipped. For this dataset, our tools, with a threshold of 0.94, identifies 3 putative subgroups that show a remarkable separation between the subtypes. Referring once again to Table 6, we see that two of the three groups are completely homogeneous, while the third consists of 23 AC tumors with a crossover of only 3 SCCs.

Subtyping Summary

In this chapter, we have described a method based on the paraclique algorithm to identify putative subtypes separating samples based on signatures in their gene expression profiles. We applied our methods to a variety of publicly available data starting with 12 sets obtained from the Gene Expression Omnibus. Of those 12, our tools identified putative subtypes for 10. We sought to validate the relevance of our findings by reviewing the literature and examining the GO enrichment for biological relevance of genes differentially expressed across our potential subtypes. We also performed additional testing with two sets data containing phenotypic information for known subtypes, also obtained from GEO. The results of our testing indicate a strong potential for our approach to be highly effective in the discovery of novel subtypes.

CHAPTER FOUR CRITICAL ALGORITHMIC ANALYSIS: KERNELIZATION VERSUS BRANCHING

Most of this chapter appeared previously in [127]. My contributions include implementing the full degree 2 rule, collecting testbed graphs, running experiments and writing the paper.

In this chapter, we investigate the relative significance of kernelization versus branching for parallel FPT implementations. Using the well-known vertex cover problem as a familiar example, we build and experiment with a testbed of five different classes of difficult graphs. For some, we find that kernelization alone obviates the need for parallelism. For others, we show that kernelization and branching work in synergy to produce efficient implementations. And yet for others, kernelization fails completely, leaving branching to solve the entire problem. Structural graph properties are studied in an effort to explicate this trichotomy. The *NP*-completeness of vertex cover makes scalability an extreme challenge.

Introduction

Fixed-Parameter Tractability (FPT) has become a popular and powerful technique for dealing with the recalcitrance of *NP*-completeness. An amenable problem is FPT if it has an algorithm that runs in $O(f(k)n^c)$ time, where n is the problem size, k is the input parameter, and c is a constant independent of n and k. Representative citations include [36] for theoretical development, [128] for previous work on parallel implementations, and [129] for historical perspective.

Vertex cover is probably the best known and most widely studied FPT problem. In its usual decision formulation, we are given a simple undirected graph G of order n and an integer k, and asked whether G contains a set S of at most k vertices so that every edge in G has at least one endpoint in S. Minimum vertex cover and its complementary dual, maximum clique, are highly appreciated for both their prominence in complexity theory and their wealth of practical applications.

Two tenets of FPT are *kernelization*, in which an input of size n is reduced to a compute core whose size depends only on k, and *branching*, whereby an efficient tree structure is used to explore the solution space. It would not be a gross oversimplification to say that kernelization is generally fast, while branching is not. That some part of the solution process is slow and exhaustive should not come as a surprise. After all, we are trying to solve enormous *NP*-complete problems exactly. On the other hand, the relative speed of kernelization should not belie its importance. Only by reducing the problem size dramatically can we hope for runtimes polynomial in n.

In recent work [130], it has been shown that one can often obtain excellent parallel speedup on large-scale biological graphs, as primarily derived from transcriptomic data. These sorts of graphs tend to be relatively sparse, more or less scale free, and contain many highly overlapping cliques of various sizes. It was found that kernelization and branching tend to work together quite well for these types of inputs. Methods were in fact often interleaved [131].

In this chapter, we study these and other input domains in an effort to elucidate the relative significance of kernelization versus branching and the need for parallel FPT computation. To this end, we use reasonably straightforward sequential and parallel FPT vertex cover implementations and examine performance on large graphs of five general classes: physical infrastructure, social interaction, high-throughput biological, pseudo-random, and regularly structured. Despite their notoriety, we find that infrastructure and

social graphs tend to succumb to mere kernelization. Biological graphs, meanwhile, benefit from both kernelization and branching. On the other hand, regularly structured and, to a lesser extent, pseudo-random graphs tend to be resistant to kernelization. Thus they benefit from highly efficient branching, parallelized implementations and effective load balancing strategies.

In the next two sections, we detail important relevant features of kernelization and branching. Section 4 contains descriptions and discussions of the various sources and sorts of domain data we use in this study. We also present sample timings for each. In Sections 5 and 6 we analyze these results and discuss directions for future study.

Kernelization Rules

Let us briefly review standard vertex cover kernelization reductions. The easiest to apply are the low degree rule, the high degree rule, and the degree two rule. It is noteworthy that the high degree rule alone ensures an $O(k^2)$ kernel. Given G, n and k, we iteratively apply these rules, at each stage creating a new graph G' with $n' \leq n$ and $k' \leq k$.

The Low Degree Rule: Any isolated vertex cannot cover any edges and may be removed, reducing n' by 1. In the case of a vertex of degree one, we can cover no fewer edges by discarding it and removing its parent and putting it in the cover. This reduces n' by 2 and k' by 1.

The High Degree Rule: If a vertex has degree greater than k then it must be included in an acceptable cover. Otherwise there would be at least k+1 edges, which could only be covered by including all its neighbors. Removing such a vertex and putting it in the cover reduces n' by 1 and k' by 1.

The Degree Two Rule: This rule is considerably more complicated. There are two separate cases to consider. Suppose u has degree two, with neighbors v and w.

Case 1: v and w are neighbors. In this case, at least two of u, v, and w must be included in any satisfying cover in order to account for the edges of the triangle they form. Including u would cover only two edges, while v and w both cover at least two and possibly more. Thus, we are best served by removing all three vertices and including v and w in the cover. This reduces n' by 3 and k' by 2.

Case 2: v and w are not neighbors. In this case, we form G' by folding u, v, and w into a new vertex, u', whose neighborhood consists of the union of the other neighbors of v and w. It turns out that G' contains a vertex cover of size $k' = k - 1$ if and only if G contains a vertex cover of size k. See Figure 22 for an example. (For completeness, we note that there are actually two options. If the cover of G' contains u', then v and w will be in the cover for G while u may be discarded. If the cover of G' does not contain u', then u will be in the cover for G while v and w may be discarded. With either option, n' is reduced by 2 and k' is reduced by 1. The specific option to choose is only relevant in backtracking, when solving the search not the decision version of the problem.)

After applying these rules, we have reduced our original question of the existence of a cover for G of size no more than k to the search for a cover of G' of size less than or equal to k'.

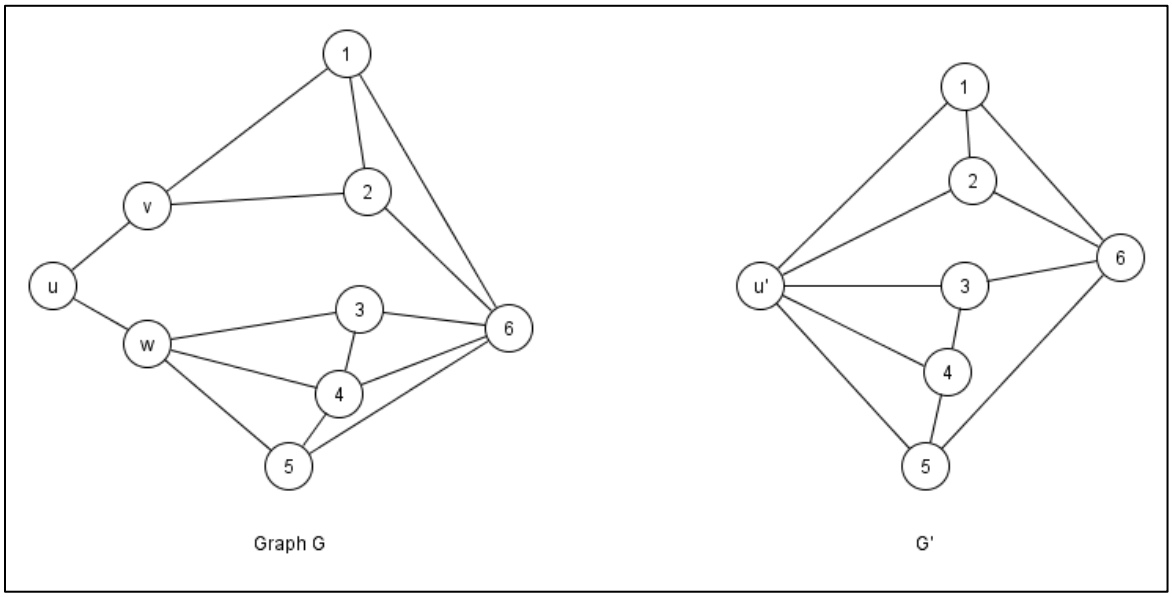

Figure 22. Kernelization. An example of Case 2 of the degree two rule. Vertices *u*, *v*, and *w* are folded into *u′*.

Branching Strategies

After kernelization, we are left with the computationally imposing task of exploring a search space whose size is exponential in k. The best current theoretical running time can be found in [132]. There, a limit of $O(1.2738^k + kn)$ is achieved, but at the cost of excruciating branching techniques, some of which may actually work against us on average. Thus, we employ instead a relatively simple search tree strategy to traverse the possible covers. At each level in the tree, we choose the vertex, v, of highest degree and branch in two directions. For the left branch, we add v to the candidate cover. For the right branch, we discard v and place all of v's neighbors in the cover. This straightforward but easily parallelized technique is illustrated in Figure 23.

The search along each branch proceeds until either k vertices have been added to the candidate solution or all edges are covered. At that point, the validity of the cover is checked. If a satisfying cover has indeed been found, then all search paths are terminated and we return a "yes" decision. On the other hand, if the candidate cover is not a satisfying cover, then branching must continue until we find a solution or exhaust the search space. At each branching stage, we include at least one more vertex in the candidate solution. Thus, this process results in a search tree of depth bounded by k. Kernelization times are generally insignificant. Branching tends to be exhaustive, however, and so it is the usual focus for parallel FPT speedups.

Each branching path reduces graph size and complexity. It is therefore sometimes possible to re-kernelize at branch points. We use this technique, known as interleaving [133], in our implementations in an effort to reduce overall run times and enhance parallel speedups.

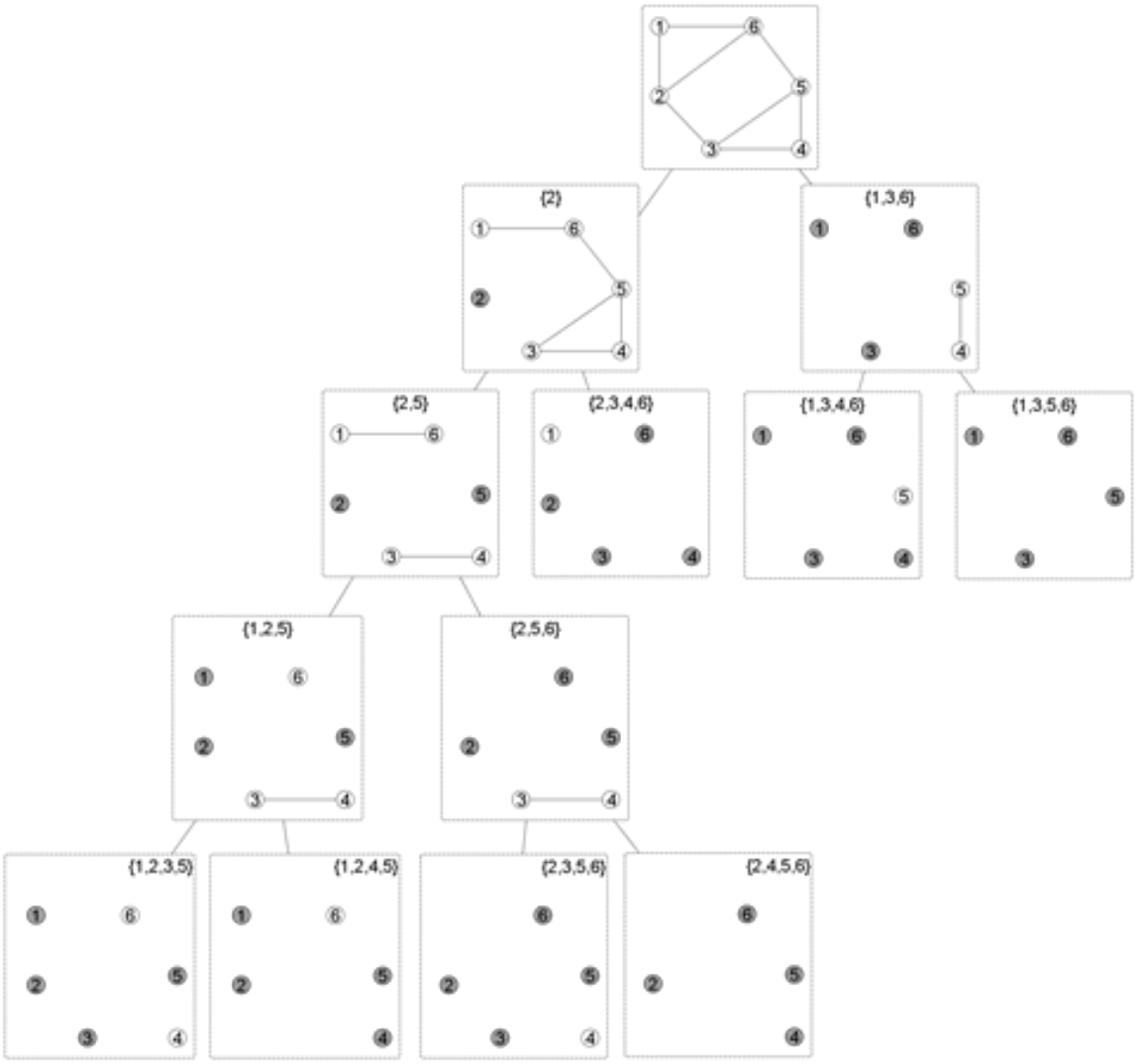

Figure 23. Branching. Dark vertices are placed into a candidate cover.

For parallel implementations, we perform search tree decomposition in MPI, using a master-slave paradigm. Each processor is initially assigned an independent branch to search. Without some form of load balancing, however, a processor may starve should it finish its branch early. Thus, we employ dynamic load balancing to keep all processors busy. Our approach designates a highest degree branch as a donor. When any branch completes, it receives additional work from the donor. Should the donor itself terminate early, a new donor is selected again based on highest degree. For a thorough description of the approach and an analysis of its effectiveness, see [128, 130].

Data and Experiments

Large-scale experiments were conducted on DOE's Hopper supercomputer, named after the remarkable computing pioneer Admiral Grace Murray Hopper. Part of the National Energy Research Scientific Computing (NERSC) Center, the Hopper platform was ranked 16th fastest in the world according to the TOP500 list for June, 2012. It is a CrayXE6 system rated at a peak speed of 1.28 PetaFLOPS, which it attains with 6,384 nodes made up of two 2.1 GHz twelve-core AMD 'MagnyCours' processors, for a total of 153,216 compute cores. Hopper is maintained at the Lawrence Berkeley National Laboratory.

We sought to test graphs from a broad variety of application domains. Our interest focused on finding graphs that are enormous and/or difficult enough to pose a challenge, even for our streamlined vertex cover FPT implementations. The graphs we selected fall into five general categories: physical infrastructure, social interaction, high-throughput biological, pseudo-random, and regularly structured. Three representative graphs were selected from each class. For each such graph, we computed the minimum vertex cover in the complement, and employed a simple binary search to determine the optimum parameter value. To identify the hardest instances, we report timings for the largest "no"

parameter value. (Had we used instead the smallest “yes” value, runs could have encountered a satisfying solution early, biasing the results, whereas a “no” instance must exhaust the entire search tree.) Every graph was sequentially kernelized. As expected, no bottleneck was encountered during this process. Branching was performed sequentially and, when sequential times were sizable, in parallel. For standard comparisons, all parallel runs were timed on 24 cores.

Physical Infrastructure Graphs

We were able to obtain large-scale connectivity information on road systems, airport networks and power grids. The Road Graph comes to us from California. Its vertices denote intersections and destinations. Its edges represent connections between these sites. The Airport Graph is based on direct flight connectivity in the United States. The Power Graph was obtained from the high voltage power grid for the western states. All three sets of data were obtained from the Stanford Large Network Dataset Collection.

Both the Road Graph and the Airport Graph produce very small kernels, which complete branching in under a hundredth of a second. The Power Graph was solved completely through kernelization, requiring no branching whatsoever. We note that our physical infrastructure graphs have modest clique sizes and low average degree, and many of their cliques are disjoint. Therefore, the high degree rule applies to many vertices in their complements. The graphs are rapidly decomposed almost entirely. This sort of algorithmic behavior highlights the power of kernelization, and eliminates the need for heavyweight computations during branching. See Table 7.

Table 7. Computational experience with large physical infrastructure graphs. Run times are reported in seconds. The graph derived from power networks was solved completely during kernelization, and thus needed no branching at all.

Graph	Road	Airport	Power
|V|	30000	1858	4941
|E|	87628	17215	6594
Density	0.00019	0.00998	0.00054
Kernel Size	147	126	7
Clique Size	3	56	6
Sequential Branching	0.00363	0.00598	unneeded

Social Interaction Graphs

We obtained interaction graphs to cover three more or less unrelated social areas. The Enron Graph represents emails during the company's financial crisis. The arXiv Graph denotes a collaboration network, mainly for high-energy physics, whose edges link electrical preprint co-authorship. Finally, the Wiki-Vote Graph depicts votes for Wikipedia moderators. As before, all three sets of data were obtained from the Stanford collection.

As with the physical infrastructure graphs, social graphs kernelize so well that they can be solved easily by branching sequentially. There is no need for parallelism. In fact, the arXiv graph was solved completely through kernelization alone. It seems that low average degree again leads to a very effective application of the high degree rule in the complement. See Table 8.

Table 8. Computational experience with large social interaction graphs. Run times are reported in seconds. The graph derived from electronic preprints was solved completely during kernelization, and thus needed no branching at all.

Graph	Enron	arXiv	Wiki-Vote
|V|	36692	12008	7115
|E|	183831	118521	100656
Density	0.00027	0.00164	0.00398
Kernel Size	2276	240	2226
Clique Size	20	239	17
Sequential Branching	118.3	unneeded	111.1

High-Throughput Biological Graphs

We tried to focus on markedly different sorts of biological graphs, and ended up employing data about human proteomics, methylation and transcriptomics. PPI was created from a curated protein-protein interaction database. Colitis.98 was created using genomic methylation data from a study on Ulcerative Colitis while LRMS.8 was created from gene expression microarray data concerning Low-Risk Myelodysplastic Syndrome. Both data sets are publicly available on the Gene Expression Omnibus (GEO). The associated datasets are GSE27899 and GSE41130 respectively. For Colitis.98 a weighted graph was first created using methylation sites as vertices with edges weighted by p-values of the Pearson correlation between the methylation levels. The same was done for LRMS.8 using genes and gene expression levels. In each case, the graph was then converted to an unweighted graph by removing all edges whose weights fall below a certain threshold ($p = .98$ and $p = .8$ respectively).

As with our physical infrastructure and social interaction experiences, kernelization nearly solved PPI. The other two graphs, Colitis.98 and LRMS.8, were each left with a

fairly formidable kernel. Although effective parallel vertex cover algorithms have been tailored to such graphs in previous work [134, 135], it is notable here that, like in the cases of the Enron and Wiki-Vote graphs, the kernels produced were such that sequential branching finished in a matter of a very few minutes without the need for parallelism. See Table 9.

Pseudo-Random Graphs

We created a pair of pseudo-random graphs using standard random graph models, the Erdös-Rényi model [136] and the Watts-Strogatz model [137]. The Erdös-Rényi graph, ER_5k, was constructed using an edge density of 0.1. Each potential edge was either present or absent based on this probability. The Watts-Strogatz graph, WS_800, began with a regular ring lattice, where every vertex had degree 300. One endpoint of each edge was then randomly rewired with probability 0.2. We tried to use the Barabási-Albert preferential attachment model [138] for a third pseudo-random graph. Unfortunately, this simple model never yields a clique of size greater than x+1, where x is the number of edges connecting each new node to the existing graph. For instance, when x=2, we obtain a graph with a maximum clique of size 3. Such graphs and their complements are solved very quickly by standard clique and vertex cover implementations. We therefore chose to create the graph Norm_900 by choosing normally distributed random degrees for the vertices and then setting edges with uniform probability.

Table 9. Computational experience with large high-throughput biological graphs. Run times are reported in seconds.

Graph	PPI	Colitis.98	LRMS.8
|V|	9314	12755	14977
|E|	39473	107205	178377
Density	0.00091	0.00132	0.00159
Kernel Size	344	2040	1795
Clique Size	13	19	30
Sequential Branching	0.75154	264.87	229.99

The structure of these pseudo-random graphs seems to be such that some reasonable modicum of speedup is possible with our balanced tree search. In the case of ER_5k, speedup seems to have been limited by its low density. Both the Erdös-Rényi and the Watts-Strogatz models produce graphs whose degree distribution is roughly normal, with mean centered at the average degree of the graph. We will say more about this issue in the next section. See Table 10.

Regularly Structured Graphs

We loosely define regularly structured graphs to be those with a narrow degree range. The limiting case is of course truly regular graphs, whose vertices all have the same degree. In fact two of the graphs we tested, 5x50_C and 90_Cycle_C, are truly regular. The third, a subgraph of a Hamming graph, is highly but not truly regular, with vertex degrees in the range [247,270]. 5x50_C is the complement of the graph obtained by arranging 50 5-cliques in a circle and connecting the corresponding adjacent clique

Table 10. Computational experience with large pseudo-random graphs. Run times are reported in seconds.

Graph	ER_5k	WS_800	Norm_900
|V|	5000	800	900
|E|	1250505	120000	158221
Density	0.10006	0.37547	0.39110
Kernel Size	5000	800	900
Clique Size	7	26	20
Sequential Branching	7635	8894.45	13027.7
Parallel Branching	4260	646.5	844.9

vertices. 90_Cycle_C was constructed by starting with an even cycle, connecting its opposite vertices, then taking its complement. That is, beginning with a cycle of length n, with vertices labeled 0,1,2,...,n-1, add edges (0,n/2), (1, n/2 + 1), (2, n/2 + 2), ... , (n/2-1, n-1). Figure 24 illustrates this notion with an example for n = 10.

Ham_9_4 was obtained as a 350-node subgraph of a standard Hamming graph, constructed using parameters n=9 and d=4. In such a graph, vertices are viewed as binary vectors of length n, and an edge is present if and only if the Hamming distance between the two vectors is greater than or equal to d.

Because low and high degree vertices are absent, regularly structured graphs are notoriously difficult, and generally show no benefit whatsoever from kernelization. Moreover, such graphs tend to have topological symmetry and fill many branches of the

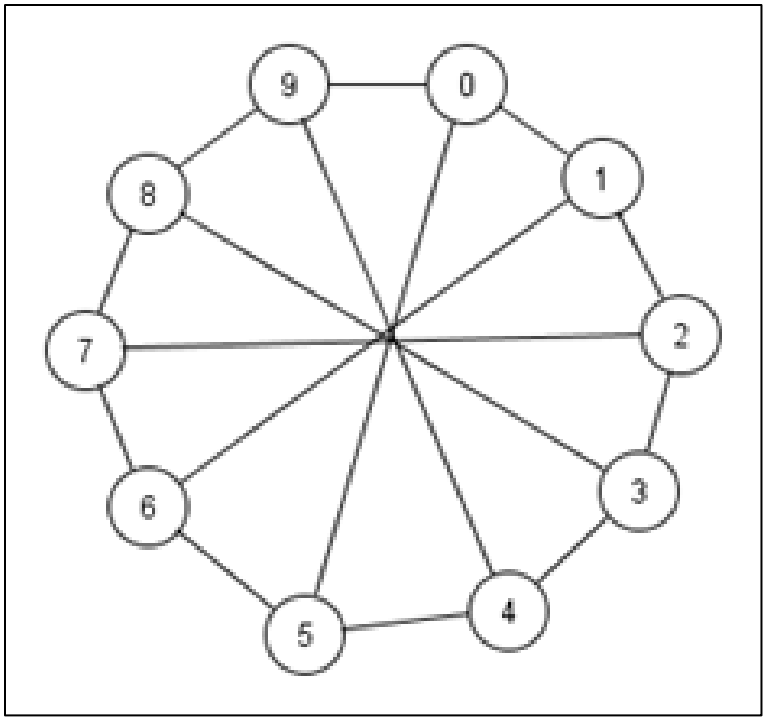

Figure 24. A regularly structured graph. 90_Cycle_C is the complement of this graph of order 90.

search tree with near-optimal solutions. Thus, many branches must be fully explored. Ironically, however, the very topological symmetry that makes such instances difficult makes them amenable to parallelization. In this study, regularly structured graphs produced the best speedup we see.

We emphasize that this efficiency is relative. These are not linear-time systolic problems that easily yield linear speedups. Instead, these are *NP*-compete problems. Their kernels require exhaustive search. They generally incur several penalties with respect to communication overhead, non-uniform memory access and so forth. Thus, we must temper our enthusiasm and lower our expectations. Nevertheless, the efficiencies we see are respectable for problems of this ilk. See Table 11.

Table 11. Computational experience with large regularly structured graphs. Run times are reported in seconds.

Graph	5x50_C	90_Cycle_C	Ham_9_4
|V|	250	180	350
|E|	6500	15390	44842
Density	0.75155	0.95531	0.73421
Kernel Size	250	180	350
Clique Size	5	90	18
Sequential Branching	369	3802	15686
Parallel Branching	23.75	236	1011.1

Parallel Utilization

There seem to be at least two major factors influencing the relative effectiveness of our parallel branching strategy: average vertex degree and vertex degree distribution. If the average vertex degree in the complement is very high, then each time a vertex is added to a possible cover in a left branch, its many neighbors will be added to a possible cover in the right branch. This may result in a highly unbalanced search tree, and thus poor parallel speedup. In Table 4, for example, we see that the density of ER_5k is 0.10006. Therefore the average degree in its complement is very high, and we witnessed only a modest speedup. In order to produce a balanced search tree without excessive load balancing, it seems from these experiments that the complement must exhibit a relatively tight vertex degree distribution about the mean, with the mean itself relatively low. This is precisely the case with the regularly structured graphs and with Norm_900.

We also note that graph size will naturally limit the scalability we can expect to see. On Ham_9_4, for example, we saw speedups up to 48 cores that then leveled out quickly. On the other hand, we were able to scale up to 720 cores before leveling out on the larger Norm_900 graph. See Figures 25 and 26.

Summary

We have studied the significance of kernelization versus branching when solving the classic vertex cover decision problem on large graphs from five different application domains. For some, we found that kernelization alone sufficed. For others, we found that kernelization and branching could work together to find solutions quickly. And yet for others, we discovered that kernelization fails completely, leaving branching to solve the entire problem.

An important hallmark of problems amenable to kernelization appears to be a complement whose degree structure follows an inverse power law distribution. Such a graph may benefit from reductions by the high degree rule, while its compute kernel may have a similar distribution, which in turn increases the effectiveness of interleaving. Our physical infrastructure, social communication, and high throughput biological graphs all have this type of structure. See, for example, Figure 27. On the other hand, graphs that do not kernelize were either pseudo-regular or those whose complements' degree distributions clustered tightly about a mean value low enough to avoid the high degree rule. See Figure 28.

Turning now to branching, graphs benefiting most from our parallel strategy were those that had a relatively tight vertex degree distribution about a relatively low mean. It remains to be seen whether different branching strategies will behave in markedly

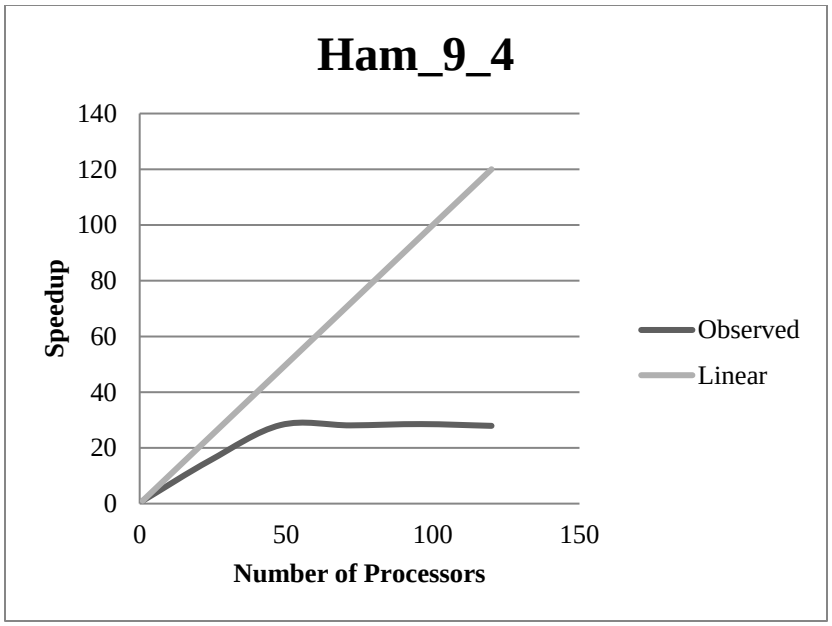

Figure 25. Speedup for Ham_9_4. Relatively small graph size limits scalability.

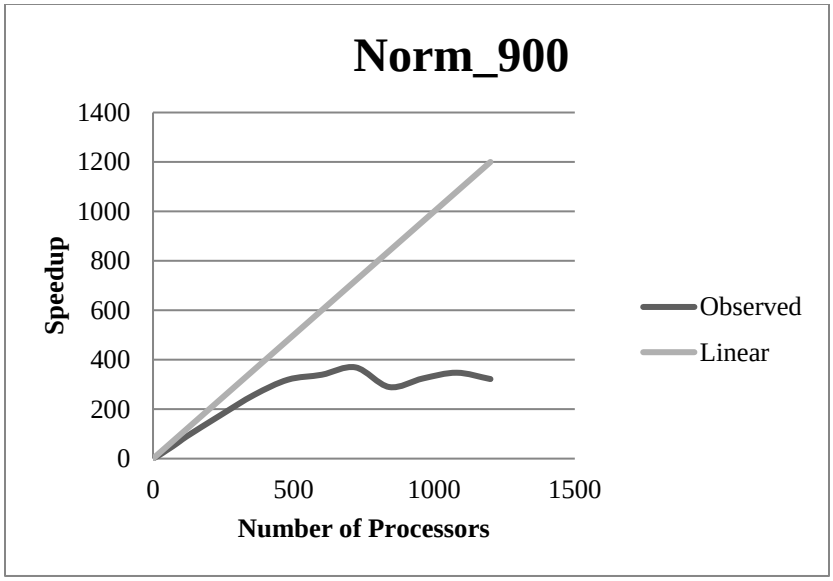

Figure 26. Speedup for Norm_900. Larger graphs provide enhanced opportunities for scalability.

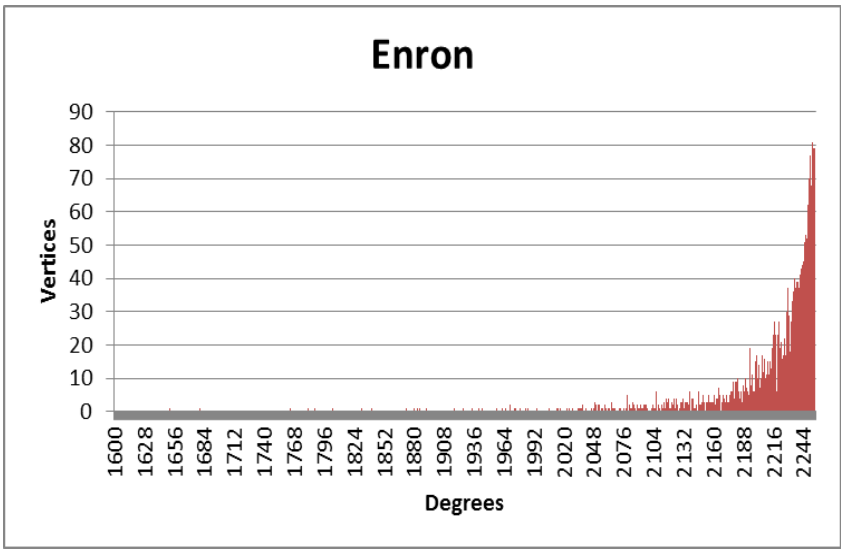

Figure 27. Degree distribution of Enron's complement. Kernelization and sequential branching proved most effective on graphs whose complements show an inverse power law degree distribution.

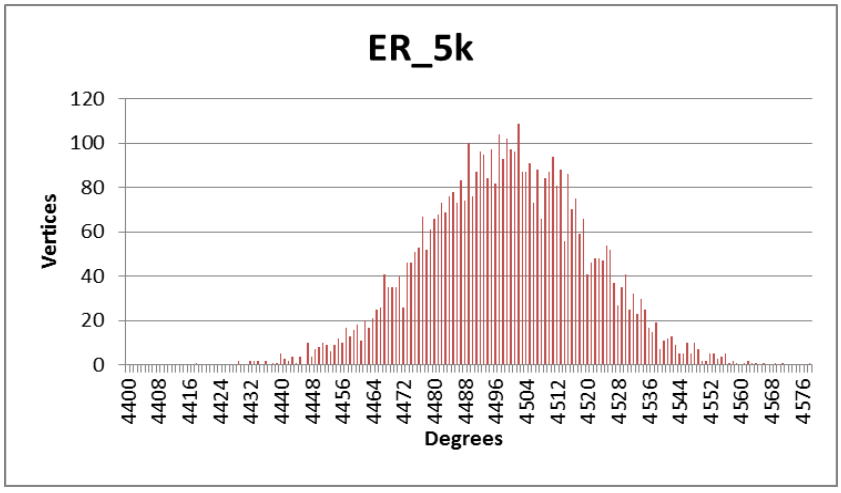

Figure 28. Degree distribution of ER_5k. A normalesque distribution combined with a sufficiently low mean reduces the benefits of kernelization.

different ways. In previous work, it has been shown that parallel vertex cover can be tailored to transcriptomic and other specific classes of graphs [128, 130, 135]. For a given branching strategy, it might prove fruitful to trace through numerous search trees in an attempt to identify cases for which the strategy is ideal and those for which it fails.

CHAPTER FIVE CONCLUSIONS AND DIRECTIONS FOR FURTHER RESEARCH

We have examined both the application and efficiency of graph-theoretical algorithms to the study of complex networks. We considered the use of exact solvers for two *NP*-complete problems, primarily dominating set and clique, and their utility in solving difficult biological problems. This work required novel approaches to algorithm development, as well as implementation and testing on a wide variety of publicly available data. We also detailed our efforts to analyze the interplay between the two critical components of fixed parameter tractable implementations. In this final section, we review the primary contributions of this dissertation and avenues for further research.

Summary of Contributions

The effort has focused on the development of combinatorial tools for the analysis of complex networks. Its primary contributions are the adaptation of a graph-based toolchain for biomarker discovery to DNA methylation data, with improvements to the method, a novel method for identifying subtypes of complex disease from gene expression data using the paraclique algorithm, and an empirical study of the relative significance of kernelization compared to branching on different graph classes for the vertex cover problem on parallel platforms. Although the application focus in the development of these algorithms was mainly on microbiology, their foundation is set in the mathematical abstraction of graph theory. As such, their applications transfer easily to other domains. Together, these results provide algorithmic advances for the study of complex networks as well as insight into the design of efficient algorithms through the use of fixed parameter tractability.

Originally, the graph-based toolchain for biomarker discovery was designed for gene expression data. Our contributions expanded and improved it in several key ways:

- Its adaptation to methylation data required scaling to a ten-fold increase in the number of variables.
- We introduced a tuning factor to address initial runs not identifying sites with positive merit scores.
- We employed an exact rather than an approximate dominating set filter.
- We added a neural net classifier to the toolchain.

We then tested the improved method on eight data sets obtained from GEO representing a variety of complex human disease. With the exception of chronic fatigue associated with Sjogren's syndrome, our method performed well in all cases. For instance, in breast cancer data, it identified a set of only five methylation sites that the neural net classifier used to discriminate between tumor and healthy tissue samples with an accuracy of 0.96. We note that our method is of course reliant on the data provided. Methylation biomarkers may be observable in one type of tissue, but not others. For example, the inability to identify markers in the Sjogren's syndrome data may indicate the absence of a link between associated chronic fatigue and methylation levels, or only that such markers are not carried in peripheral blood.

The method we developed for using the paraclique algorithm to identify putative subgroups in gene expression data has several key features:

- It reduces the effects of confounding factors by introducing an initial filtering step when accompanying control data is available.
- It transposes the correlation matrices typically used in a gene network analysis, computing sample-sample correlations to which a threshold is applied to build a graph.
- It then applies the paraclique algorithm to identify dense subgraphs representing potential subtypes.

We applied the method to twelve publicly available data sets obtained from GEO representing a range of human disease. Our method identified putative subtypes in ten of those cases. In the other two cases, we feel that either there is no discernable molecular subtype associated with the disease, or perhaps more likely, the available data set was too small to allow its discovery. To help validate the putative subtypes produced by our method, we conducted a literature review for the involvement of genes differentially expressed between paracliques having known biological relevance. We also conducted GO enrichment to determine whether the method differentiated between known molecular subtypes. Despite our efforts at validation, we recognize that this work is somewhat limited by a lack of ground truth. Further investigation into the biological role of the genes driving the separation of our putative subtypes is needed.

To investigate the relative significance of kernelization compared to branching in a parallel FPT version of a vertex cover solver, we implemented several reduction rules, including the low degree, high degree and full degree rules. We performed extensive testing on a variety of large graphs with different structural characteristics drawn from five classes. The tests were ran on the Hopper machine at the NERSC facility, at the time the 16th fastest supercomputer for academic research in the world. Our testing revealed wide disparities between classes of graphs. Graphs whose degree structure follow a power law or inverse power law distribution can be almost completely solved with kernelization, while pseudo-regular graphs gain little benefit from kernelization, and must rely entirely on branching. The parallel speedup obtained by our algorithm relies greatly on the structure of the input graph. Potential steps to reduce this dependence and improve scalability include investigating alternative branching strategies and improving load balancing.

Possible New Directions for Future Work

The work presented in this dissertation represents a stepping stone in the path to harnessing the power of exact solutions to *NP*-complete graph problems for the practical exploration of large, complex networks. As such, it leaves open a variety of avenues for further research.

The work on methylation biomarker discovery could potentially form the basis for the analysis of other types of epigenetic data. Given the success of using our dominating set filter for feature selection, future work is warranted to investigate potential synergies between classical graph algorithms and traditional machine learning approaches. Additionally, work remains to be done improving the current state of the art implementations of dominating set.

The subtyping work can also be applied to other types of data rather than gene expression, such as epigenetic data. In particular, it would be interesting to see how well the approach adapts to categorical data. An open area is the development of automatic validation metrics to judge the "goodness" of putative subtypes.

Finally, the work on FPT vertex cover in chapter four opens several paths forward. Perhaps most promising would be the development of structural metrics that predict scalability and performance of various vertex cover/clique finders. With such metrics in hand, one could then design a targeted solver that would select the best approach based on the input graph. As we observed, satisfactory parallel speedup in our current implementation often depends on the structure of the input graph. As such, much work remains in developing parallel graph algorithms for problems such as vertex cover that require highly randomized memory access.

LIST OF REFERENCES

[1] Data, B. for better or worse: 90% of world's data generated over last two years. *SCIENCE DAILY, May*, 22 (2013).
[2] Akutsu, T., Miyano, S. and Kuhara, S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pacific symposium on biocomputing*, Vol. 4. 1999.
[3] Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. and Miyano, S. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, 19, suppl 2 (2003), ii227-ii236.
[4] Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A. and Chinnaiyan, A. M. Probabilistic model of the human protein-protein interaction network. *Nature biotechnology*, 23, 8 (2005), 951-959.
[5] Kempe, D., Kleinberg, J. and Tardos, É. Maximizing the spread of influence through a social network. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2003.
[6] Sporns, O. Contributions and challenges for network models in cognitive neuroscience. *Nat Neurosci*, 17, 5 (05//print 2014), 652-660.
[7] Grunstein, M. and Hogness, D. S. Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. *Proceedings of the National Academy of Sciences*, 72, 10 (1975), 3961-3965.
[8] Bumgarner, R. Overview of DNA microarrays: types, applications, and their future. *Current protocols in molecular biology* (2013), 22.21. 21-22.21. 11.
[9] Krulwich, R. and LANDER, E. *Cracking the Code of Life*. Public Broadcasting Service, City, 2001.
[10] Holliday, R. DNA methylation and epigenetic inheritance. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 326, 1235 (1990), 329-338.
[11] Russo, V. E., Martienssen, R. A. and Riggs, A. D. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, 1996.

[12] Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P. and Jaenisch, R. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences*, 97, 10 (2000), 5237-5242.
[13] Hansen, R. S., Wijmenga, C., Luo, P., Stanek, A. M., Canfield, T. K., Weemaes, C. M. and Gartler, S. M. The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. *Proceedings of the National Academy of Sciences*, 96, 25 (1999), 14412-14417.
[14] Okano, M., Bell, D. W., Haber, D. A. and Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99, 3 (1999), 247-257.
[15] Li, E., Bestor, T. H. and Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69, 6 (1992), 915-926.
[16] Strichman-Almashanu, L. Z., Lee, R. S., Onyango, P. O., Perlman, E., Flam, F., Frieman, M. B. and Feinberg, A. P. A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome research*, 12, 4 (2002), 543-554.
[17] Yen, P. H., Patel, P., Chinault, A. C., Mohandas, T. and Shapiro, L. J. Differential methylation of hypoxanthine phosphoribosyltransferase genes on active and inactive human X chromosomes. *Proceedings of the National Academy of Sciences*, 81, 6 (1984), 1759-1763.
[18] Razin, A. and Cedar, H. DNA methylation and genomic imprinting. *Cell*, 77, 4 (1994), 473-476.
[19] Barlow, D. P. Gametic imprinting in mammals. *Science*, 270, 5242 (1995), 1610.
[20] Jones, P. A. and Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nature reviews genetics*, 3, 6 (2002), 415-428.
[21] Herman, J. G. and Baylin, S. B. Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine*, 349, 21 (2003), 2042-2054.

[22] Ehrlich, M. DNA methylation in cancer: too much, but also too little. *Oncogene*, 21, 35 (2002), 5400.
[23] Langston, M. A., Levine, R. S., Kilbourne, B. J., Rogers, G. L., Kershenbaum, A. D., Baktash, S. H., Coughlin, S. S., Saxton, A. M., Agboto, V. K. and Hood, D. B. Scalable combinatorial tools for health disparities research. *International journal of environmental research and public health*, 11, 10 (2014), 10419-10443.
[24] Wolen, A. R., Phillips, C. A., Langston, M. A., Putman, A. H., Vorster, P. J., Bruce, N. A., York, T. P., Williams, R. W. and Miles, M. F. Genetic dissection of acute ethanol responsive gene networks in prefrontal cortex: functional and mechanistic implications. *PloS one*, 7, 4 (2012), e33575.
[25] Chesler, E. J. and Langston, M. A. Combinatorial genetic regulatory network analysis tools for high throughput transcriptomic data. *Systems Biology and Regulatory Genomics*. Springer, Berlin Heidelberg, 2007, 150-165.
[26] Perkins, A. D. and Langston, M. A. Threshold Selection in Gene Co-Expression Networks Using Spectral Graph Theory Techniques. *BMC Bioinformatics*, 10 (2009).
[27] Voy, B. H., Scharff, J. A., Perkins, A. D., Saxton, A. M., Borate, B., Chesler, E. J., Branstetter, L. K. and Langston, M. A. Extracting gene networks for low dose radiation using graph theoretical algorithms. *PLoS Computational Biology*, 2, 7 (2006), e89.
[28] Wang, J., Zhou, S. and Guan, J. Detecting potential collusive cliques in futures markets based on trading behaviors from real data. *Neurocomputing*, 92 (2012), 44-53.
[29] Karp, R. Reducibility among combinatorial problems. *Complexity of computer computations*, Springer US, 1972, 85-103.
[30] Jay, J., Eblen, J., Zhang, Y., Benson, M., Perkins, A., Saxton, A., Voy, B., Chesler, E. and Langston, M. A systematic comparison of genome-scale clustering algorithms. *BMC Bioinformatics*, 13, Suppl 10 (2012), S7.
[31] Hagan, R. D., Langston, M. A. and Wang, K. Lower bounds on paraclique density. *Discrete Applied Mathematics*, 204 (5/11/ 2016), 208-212.
[32] Fellows, M. R. and Langston, M. A. Nonconstructive Tools for Proving Polynomial-Time Decidability. *Journal of the ACM*, 35 (1988), 727-739.

[33] Fellows, M. R. and Langston, M. A. Nonconstructive Advances in Polynomial-Time Complexity. *Information Processing Letters*, 26 (1987), 157-162.
[34] Fellows, M. R. and Langston, M. A. On well-partial-order theory and its application to combinatorial problems of VLSI design. *SIAM Journal on Discrete Mathematics*, 5, 1 (1992), 117-126.
[35] Robertson, N. and Seymour, P. D. Graph Minors. XX. Wagner's conjecture. *Journal of Combinatorial Theory, Series B*, 92, 2 (2004), 325-357.
[36] Downey, R. G. and Fellows, M. R. *Parameterized Complexity*. Springer, New York, 1999.
[37] Downey, R. G., Fellows, M. R. and Stege, U. *Parameterized complexity: A framework for systematically confronting computational intractability*. City, 1999.
[38] Langston, M. A., Lan, L., Peng, X., Baldwin, N. E., Symons, C. T., Zhang, B. and Snoddy, J. R. *A combinatorial approach to the analysis of differential gene expression data: the use of graph algorithms for disease prediction and screening*. Kluwer Academic Publishers, City, 2005.
[39] Langston, M. A., Lin, L., Peng, X., Baldwin, N. E., Symons, C. T., Zhang, B. and Snoddy, J. R. A combinatorial approach to the analysis of differential gene expression data. *Methods of Microarray Data Analysis IV*, 4 (2004), 223.
[40] Reynard, L. N., Bui, C., Canty-Laird, E. G., Young, D. A. and Loughlin, J. Expression of the osteoarthritis-associated gene GDF5 is modulated epigenetically by DNA methylation. *Human molecular genetics*, 20, 17 (2011), 3450-3460.
[41] Iliopoulos, D., Malizos, K. N. and Tsezou, A. Epigenetic regulation of leptin affects MMP-13 expression in osteoarthritic chondrocytes: possible molecular target for osteoarthritis therapeutic intervention. *Annals of the rheumatic diseases*, 66, 12 (2007), 1616-1621.
[42] Barter, M., Bui, C. and Young, D. Epigenetic mechanisms in cartilage and osteoarthritis: DNA methylation, histone modifications and microRNAs. *Osteoarthritis and cartilage*, 20, 5 (2012), 339-349.

[43] El Mansouri, F. E., Chabane, N., Zayed, N., Kapoor, M., Benderdour, M., Martel-Pelletier, J., Pelletier, J. P., Duval, N. and Fahmi, H. Contribution of H3K4 methylation by SET-1A to interleukin-1–induced cyclooxygenase 2 and inducible nitric oxide synthase expression in human osteoarthritis chondrocytes. *Arthritis & Rheumatology*, 63, 1 (2011), 168-179.
[44] Aref-Eshghi, E., Zhang, Y., Liu, M., Harper, P. E., Martin, G., Furey, A., Green, R., Sun, G., Rahman, P. and Zhai, G. Genome-wide DNA methylation study of hip and knee cartilage reveals embryonic organ and skeletal system morphogenesis as major pathways involved in osteoarthritis. *BMC musculoskeletal disorders*, 16, 1 (2015), 287.
[45] Chen, H.-C. *Genetics and Biomarkers of Osteoarthritis and Joint Hypermobility*. Duke University, 2009.
[46] Saito, T., Fukai, A., Mabuchi, A., Ikeda, T., Yano, F., Ohba, S., Nishida, N., Akune, T., Yoshimura, N. and Nakagawa, T. Transcriptional regulation of endochondral ossification by HIF-2 [alpha] during skeletal growth and osteoarthritis development. *Nature medicine*, 16, 6 (2010), 678-686.
[47] Ferlay, J., Héry, C., Autier, P. and Sankaranarayanan, R. Global burden of breast cancer. *Breast cancer epidemiology*. Springer New York, 2010.
[48] Huang, T. H.-M., Perry, M. R. and Laux, D. E. Methylation profiling of CpG islands in human breast cancer cells. *Human molecular genetics*, 8, 3 (1999), 459-470.
[49] Dobrovic, A. and Simpfendorfer, D. Methylation of the BRCA1 gene in sporadic breast cancer. *Cancer Research*, 57, 16 (1997), 3347-3350.
[50] Ottaviano, Y. L., Issa, J.-P., Parl, F. F., Smith, H. S., Baylin, S. B. and Davidson, N. E. Methylation of the estrogen receptor gene CpG island marks loss of estrogen receptor expression in human breast cancer cells. *Cancer Research*, 54, 10 (1994), 2552-2555.
[51] Pihur, V., Datta, S. and Datta, S. Finding common genes in multiple cancer types through meta–analysis of microarray experiments: A rank aggregation approach. *Genomics*, 92, 6 (2008), 400-403.
[52] El–Serag, H. B. and Rudolph, K. L. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology*, 132, 7 (2007), 2557-2576.

[53] Shen, J., LeFave, C., Sirosh, I., Siegel, A. B., Tycko, B. and Santella, R. M. Integrative epigenomic and genomic filtering for methylation markers in hepatocellular carcinomas. *BMC medical genomics*, 8, 1 (2015), 28.
[54] Shen, J., Wang, S., Zhang, Y.-J., Wu, H.-C., Kibriya, M. G., Jasmine, F., Ahsan, H., Wu, D. P., Siegel, A. B. and Remotti, H. Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. *Epigenetics*, 8, 1 (2013), 34-43.
[55] Yamada, N., Yasui, K., Dohi, O., Gen, Y., Tomie, A., Kitaichi, T., Iwai, N., Mitsuyoshi, H., Sumida, Y. and Moriguchi, M. Genome-wide DNA methylation analysis in hepatocellular carcinoma. *Oncology reports*, 35, 4 (2016), 2228-2236.
[56] Wockner, L., Noble, E., Lawford, B., Young, R. M., Morris, C., Whitehall, V. and Voisey, J. Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients. *Translational psychiatry*, 4, 1 (2014), e339.
[57] Vawter, M. P., Ferran, E., Galke, B., Cooper, K., Bunney, W. E. and Byerley, W. Microarray screening of lymphocyte gene expression differences in a multiplex schizophrenia pedigree. *Schizophrenia research*, 67, 1 (2004), 41-52.
[58] Pickard, B. S. Schizophrenia biomarkers: translating the descriptive into the diagnostic. *Journal of Psychopharmacology*, 29, 2 (2015), 138-143.
[59] Mamdani, F., Martin, M. V., Lencz, T., Rollins, B., Robinson, D. G., Moon, E. A., Malhotra, A. K. and Vawter, M. P. Coding and noncoding gene expression biomarkers in mood disorders and schizophrenia. *Disease markers*, 35, 1 (2013), 11-21.
[60] Bacalini, M. G., Gentilini, D., Boattini, A., Giampieri, E., Pirazzini, C., Giuliani, C., Fontanesi, E., Scurti, M., Remondini, D. and Capri, M. Identification of a DNA methylation signature in blood cells from persons with Down Syndrome. *Aging (Albany NY)*, 7, 2 (2014), 82-96.
[61] Jones, M. J., Farré, P., McEwen, L. M., MacIsaac, J. L., Watt, K., Neumann, S. M., Emberly, E., Cynader, M. S., Virji-Babul, N. and Kobor, M. S. Distinct DNA methylation patterns of cognitive impairment and trisomy 21 in Down syndrome. *BMC medical genomics*, 6, 1 (2013), 58.

[62] Old, R. W., Crea, F., Puszyk, W. and Hultén, M. A. Candidate epigenetic biomarkers for non-invasive prenatal diagnosis of Down syndrome. *Reproductive biomedicine online*, 15, 2 (2007), 227-235.
[63] Huynh, J. L., Garg, P., Thin, T. H., Yoo, S., Dutta, R., Trapp, B. D., Haroutunian, V., Zhu, J., Donovan, M. J. and Sharp, A. J. Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains. *Nature neuroscience*, 17, 1 (2014), 121-130.
[64] Lunnon, K., Smith, R. G., Cooper, I., Greenbaum, L., Mill, J. and Beeri, M. S. Blood methylomic signatures of presymptomatic dementia in elderly subjects with type 2 diabetes mellitus. *Neurobiology of aging*, 36, 3 (2015), 1600. e1601-1600. e1604.
[65] Norheim, K. B., Imgenberg-Kreuz, J., Jonsdottir, K., Janssen, E. A., Syvänen, A.-C., Sandling, J. K., Nordmark, G. and Omdal, R. Epigenome-wide DNA methylation patterns associated with fatigue in primary Sjögren's syndrome. *Rheumatology*, 55, 6 (2016), 1074-1082.
[66] Cypess, A. M., Lehman, S., Williams, G., Tal, I., Rodman, D., Goldfine, A. B., Kuo, F. C., Palmer, E. L., Tseng, Y.-H., Doria, A., Kolodny, G. M. and Kahn, C. R. Identification and Importance of Brown Adipose Tissue in Adult Humans. *New England Journal of Medicine*, 360, 15 (2009), 1509-1517.
[67] Saito, M., Okamatsu-Ogura, Y., Matsushita, M., Watanabe, K., Yoneshiro, T., Nio-Kobayashi, J., Iwanaga, T., Miyagawa, M., Kameya, T., Nakada, K., Kawai, Y. and Tsujisaki, M. High incidence of metabolically active brown adipose tissue in healthy adult humans: effects of cold exposure and adiposity. *Diabetes*, 58, 7 (Jul 2009), 1526-1531.
[68] Virtanen, K. A., Lidell, M. E., Orava, J., Heglind, M., Westergren, R., Niemi, T., Taittonen, M., Laine, J., Savisto, N.-J., Enerbäck, S. and Nuutila, P. Functional Brown Adipose Tissue in Healthy Adults. *New England Journal of Medicine*, 360, 15 (2009), 1518-1525.

[69] Bi, P., Shan, T., Liu, W., Yue, F., Yang, X., Liang, X. R., Wang, J., Li, J., Carlesso, N., Liu, X. and Kuang, S. Inhibition of Notch signaling promotes browning of white adipose tissue and ameliorates obesity. *Nat Med*, 20, 8 (Aug 2014), 911-918.
[70] Nedergaard, J. and Cannon, B. The browning of white adipose tissue: some burning issues. *Cell Metab*, 20, 3 (Sep 2 2014), 396-407.
[71] Wajchenberg, B. L. Subcutaneous and visceral adipose tissue: their relation to the metabolic syndrome. *Endocr Rev*, 21, 6 (Dec 2000), 697-738.
[72] Wajchenberg, B. L., Giannella-Neto, D., da Silva, M. E. and Santos, R. F. Depot-specific hormonal characteristics of subcutaneous and visceral adipose tissue and their relation to the metabolic syndrome. *Hormone and metabolic research = Hormon- und Stoffwechselforschung = Hormones et metabolisme*, 34, 11-12 (Nov-Dec 2002), 616-621.
[73] Berry, D. C., Stenesen, D., Zeve, D. and Graff, J. M. The developmental origins of adipose tissue. *Development*, 140, 19 (Oct 2013), 3939-3949.
[74] Jensen, M. D. Role of body fat distribution and the metabolic complications of obesity. *The Journal of clinical endocrinology and metabolism*, 93, 11 Suppl 1 (Nov 2008), S57-63.
[75] Manolopoulos, K. N., Karpe, F. and Frayn, K. N. Gluteofemoral body fat as a determinant of metabolic health. *International journal of obesity*, 34, 6 (Jun 2010), 949-959.
[76] Snijder, M. B., Visser, M., Dekker, J. M., Goodpaster, B. H., Harris, T. B., Kritchevsky, S. B., De Rekeneire, N., Kanaya, A. M., Newman, A. B., Tylavsky, F. A. and Seidell, J. C. Low subcutaneous thigh fat is a risk factor for unfavourable glucose and lipid levels, independently of high abdominal fat. The Health ABC Study. *Diabetologia*, 48, 2 (Feb 2005), 301-308.
[77] Bluher, M. The distinction of metabolically 'healthy' from 'unhealthy' obese individuals. *Current opinion in lipidology*, 21, 1 (Feb 2010), 38-43.
[78] Canoy, D., Boekholdt, S. M., Wareham, N., Luben, R., Welch, A., Bingham, S., Buchan, I., Day, N. and Khaw, K. T. Body fat distribution and risk of coronary heart disease in men and women in the European Prospective Investigation Into Cancer and

Nutrition in Norfolk cohort: a population-based prospective study. *Circulation*, 116, 25 (Dec 18 2007), 2933-2943.
[79] Yusuf, S., Hawken, S., Ounpuu, S., Bautista, L., Franzosi, M. G., Commerford, P., Lang, C. C., Rumboldt, Z., Onen, C. L., Lisheng, L., Tanomsup, S., Wangai, P., Jr., Razak, F., Sharma, A. M. and Anand, S. S. Obesity and the risk of myocardial infarction in 27,000 participants from 52 countries: a case-control study. *Lancet*, 366, 9497 (Nov 5 2005), 1640-1649.
[80] Tchkonia, T., Thomou, T., Zhu, Y., Karagiannides, I., Pothoulakis, C., Jensen, M. D. and Kirkland, J. L. Mechanisms and metabolic implications of regional differences among fat depots. *Cell Metab*, 17, 5 (May 7 2013), 644-656.
[81] Lee, M. J., Wu, Y. and Fried, S. K. Adipose tissue heterogeneity: implication of depot differences in adipose tissue for obesity complications. *Molecular aspects of medicine*, 34, 1 (Feb 2013), 1-11.
[82] Macotela, Y., Emanuelli, B., Mori, M. A., Gesta, S., Schulz, T. J., Tseng, Y. H. and Kahn, C. R. Intrinsic differences in adipocyte precursor cells from different white fat depots. *Diabetes*, 61, 7 (Jul 2012), 1691-1699.
[83] Tchkonia, T., Lenburg, M., Thomou, T., Giorgadze, N., Frampton, G., Pirtskhalava, T., Cartwright, A., Cartwright, M., Flanagan, J., Karagiannides, I., Gerry, N., Forse, R. A., Tchoukalova, Y., Jensen, M. D., Pothoulakis, C. and Kirkland, J. L. Identification of depot-specific human fat cell progenitors through distinct expression profiles and developmental gene patterns. *American journal of physiology. Endocrinology and metabolism*, 292, 1 (Jan 2007), E298-307.
[84] Yamamoto, Y., Gesta, S., Lee, K. Y., Tran, T. T., Saadatirad, P. and Kahn, C. R. Adipose depots possess unique developmental gene signatures. *Obesity*, 18, 5 (May 2010), 872-878.
[85] Chau, Y. Y., Bandiera, R., Serrels, A., Martinez-Estrada, O. M., Qing, W., Lee, M., Slight, J., Thornburn, A., Berry, R., McHaffie, S., Stimson, R. H., Walker, B. R., Chapuli, R. M., Schedl, A. and Hastie, N. Visceral and subcutaneous fat have different

origins and evidence supports a mesothelial source. *Nature cell biology*, 16, 4 (Apr 2014), 367-375.
[86] Gehrke, S., Brueckner, B., Schepky, A., Klein, J., Iwen, A., Bosch, T. C., Wenck, H., Winnefeld, M. and Hagemann, S. Epigenetic regulation of depot-specific gene expression in adipose tissue. *PloS one*, 8, 12 (2013), e82516.
[87] Keller, M., Hopp, L., Liu, X., Wohland, T., Rohde, K., Cancello, R., Klös, M., Bacos, K., Kern, M., Eichelmann, F., Dietrich, A., Schön, M. R., Gärtner, D., Lohmann, T., Dreßler, M., Stumvoll, M., Kovacs, P., DiBlasio, A.-M., Ling, C., Binder, H., Blüher, M. and Böttcher, Y. Genome-wide DNA promoter methylation and transcriptome analysis in human adipose tissue unravels novel candidate genes for obesity. *Molecular Metabolism*, Article in Press (2016).
[88] Benton, M. C., Johnstone, A., Eccles, D., Harmon, B., Hayes, M. T., Lea, R. A., Griffiths, L., Hoffman, E. P., Stubbs, R. S. and Macartney-Coxson, D. An analysis of DNA methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss. *Genome biology*, 16, 1 (2015), 8.
[89] Lowe, R. and Rakyan, V. K. Marmal-aid--a database for Infinium HumanMethylation450. *BMC Bioinformatics*, 14 (2013), 359.
[90] Slieker, R. C., Bos, S. D., Goeman, J. J., Bovee, J. V., Talens, R. P., van der Breggen, R., Suchiman, H. E., Lameijer, E. W., Putter, H., van den Akker, E. B., Zhang, Y., Jukema, J. W., Slagboom, P. E., Meulenbelt, I. and Heijmans, B. T. Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics Chromatin*, 6, 1 (2013), 26.
[91] Guenard, F., Tchernof, A., Deshaies, Y., Perusse, L., Biron, S., Lescelleur, O., Biertho, L., Marceau, S. and Vohl, M. C. Differential methylation in visceral adipose tissue of obese men discordant for metabolic disturbances. *Physiological genomics*, 46, 6 (Mar 15 2014), 216-222.
[92] Dick, K. J., Nelson, C. P., Tsaprouni, L., Sandling, J. K., Aissi, D., Wahl, S., Meduri, E., Morange, P. E., Gagnon, F., Grallert, H., Waldenberger, M., Peters, A., Erdmann, J., Hengstenberg, C., Cambien, F., Goodall, A. H., Ouwehand, W. H.,

Schunkert, H., Thompson, J. R., Spector, T. D., Gieger, C., Tregouet, D. A., Deloukas, P. and Samani, N. J. DNA methylation and body-mass index: a genome-wide analysis. *Lancet* (Mar 12 2014).

[93] Grundberg, E., Meduri, E., Sandling, J. K., Hedman, A. K., Keildson, S., Buil, A., Busche, S., Yuan, W., Nisbet, J., Sekowska, M., Wilk, A., Barrett, A., Small, K. S., Ge, B., Caron, M., Shin, S. Y., Multiple Tissue Human Expression Resource, C., Lathrop, M., Dermitzakis, E. T., McCarthy, M. I., Spector, T. D., Bell, J. T. and Deloukas, P. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet*, 93, 5 (Nov 7 2013), 876-890.

[94] Hagan, R. D., Phillips, C. A., Rhodes, B. J. and Langston, M. A. Compound Analytics: Templates for Integrating Graph Algorithms and Machine Learning. *Proceedings of 31st International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, in press.

[95] Leith, C. P., Kopecky, K. J., Godwin, J., McConnell, T., Slovak, M. L., Chen, I.-M., Head, D. R., Appelbaum, F. R. and Willman, C. L. Acute myeloid leukemia in the elderly: assessment of multidrug resistance (MDR1) and cytogenetics distinguishes biologic subgroups with remarkably distinct responses to standard chemotherapy. A Southwest Oncology Group study. *Blood*, 89, 9 (1997), 3323-3329.

[96] Balko, J. M., Cook, R. S., Vaught, D. B., Kuba, M. G., Miller, T. W., Bhola, N. E., Sanders, M. E., Granja-Ingram, N. M., Smith, J. J. and Meszoely, I. M. Profiling of residual breast cancers after neoadjuvant chemotherapy identifies DUSP4 deficiency as a mechanism of drug resistance. *Nature medicine*, 18, 7 (2012), 1052-1059.

[97] Shen, R., Olshen, A. B. and Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25, 22 (2009), 2906-2912.

[98] Ambrogi, F., Biganzoli, E., Querzoli, P., Ferretti, S., Boracchi, P., Alberti, S., Marubini, E. and Nenci, I. Molecular subtyping of breast cancer from traditional tumor

marker profiles using parallel clustering methods. *Clinical Cancer Research*, 12, 3 (2006), 781-790.
[99] Wessman, J., Paunio, T., Tuulio-Henriksson, A., Koivisto, M., Partonen, T., Suvisaari, J., Turunen, J. A., Wedenoja, J., Hennah, W. and Pietiläinen, O. P. Mixture model clustering of phenotype features reveals evidence for association of DTNBP1 to a specific subtype of schizophrenia. *Biological psychiatry*, 66, 11 (2009), 990-996.
[100] Wang, K., Phillips, C. A., Saxton, A. M. and Langston, M. A. EntropyExplorer: an R package for computing and comparing differential Shannon entropy, differential coefficient of variation and differential expression. *BMC Research Notes*, 8 (12/30 2015), 832.
[101] Bope, E. T. and Kellerman, R. D. *Conn's Current Therapy 2016*. Elsevier Health Sciences, 2015.
[102] Woodruff, P. G., Boushey, H. A., Dolganov, G. M., Barker, C. S., Yang, Y. H., Donnelly, S., Ellwanger, A., Sidhu, S. S., Dao-Pick, T. P. and Pantoja, C. Genome-wide profiling identifies epithelial cell genes associated with asthma and with treatment response to corticosteroids. *Proceedings of the National Academy of Sciences*, 104, 40 (2007), 15858-15863.
[103] G. Woodruff, P. Subtypes of asthma defined by epithelial cell expression of messenger RNA and microRNA. *Annals of the American Thoracic Society*, 10, Supplement (2013), S186-S189.
[104] Ford, D., Easton, D., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D., Weber, B., Lenoir, G. and Chang-Claude, J. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *The American Journal of Human Genetics*, 62, 3 (1998), 676-689.
[105] Easton, D., Bishop, D., Ford, D. and Crockford, G. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *American journal of human genetics*, 52, 4 (1993), 678.
[106] Dent, R., Trudeau, M., Pritchard, K. I., Hanna, W. M., Kahn, H. K., Sawka, C. A., Lickley, L. A., Rawlinson, E., Sun, P. and Narod, S. A. Triple-negative breast cancer:

clinical features and patterns of recurrence. *Clinical cancer research*, 13, 15 (2007), 4429-4434.
[107] Pedraza, V., Gomez-Capilla, J. A., Escaramis, G., Gomez, C., Torné, P., Rivera, J. M., Gil, A., Araque, P., Olea, N. and Estivill, X. Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness. *Cancer*, 116, 2 (2010), 486-496.
[108] Srour, N., Reymond, M. A. and Steinert, R. Lost in translation? A systematic database of gene expression in breast cancer. *Pathobiology*, 75, 2 (2008), 112-118.
[109] de Silva Rudland, S., Martin, L., Roshanlall, C., Winstanley, J., Leinster, S., Platt-Higgins, A., Carroll, J., West, C., Barraclough, R. and Rudland, P. Association of S100A4 and osteopontin with specific prognostic factors and survival of patients with minimally invasive breast cancer. *Clinical Cancer Research*, 12, 4 (2006), 1192-1200.
[110] King, E. R., Tung, C. S., Tsang, Y. T., Zu, Z., Lok, G. T., Deavers, M. T., Malpica, A., Wolf, J. K., Lu, K. H. and Birrer, M. J. The anterior gradient homolog 3 (AGR3) gene is associated with differentiation and survival in ovarian cancer. *The American journal of surgical pathology*, 35, 6 (2011), 904.
[111] Ricardo, S., Vieira, A. F., Gerhard, R., Leitão, D., Pinto, R., Cameselle-Teijeiro, J. F., Milanezi, F., Schmitt, F. and Paredes, J. Breast cancer stem cell markers CD44, CD24 and ALDH1: expression distribution within intrinsic molecular subtype. *Journal of clinical pathology* (2011), jcp. 2011.090456.
[112] Yamashita, T., Forgues, M., Wang, W., Kim, J. W., Ye, Q., Jia, H., Budhu, A., Zanetti, K. A., Chen, Y. and Qin, L.-X. EpCAM and α-fetoprotein expression defines novel prognostic subtypes of hepatocellular carcinoma. *Cancer research*, 68, 5 (2008), 1451-1461.
[113] Rozman, C. and Montserrat, E. Chronic lymphocytic leukemia. *New England Journal of Medicine*, 333, 16 (1995), 1052-1057.
[114] Görgün, G., Holderried, T. A., Zahrieh, D., Neuberg, D. and Gribben, J. G. Chronic lymphocytic leukemia cells induce changes in gene expression of CD4 and CD8 T cells. *The Journal of clinical investigation*, 115, 7 (2005), 1797-1805.

[115] Wiestner, A., Rosenwald, A., Barry, T. S., Wright, G., Davis, R. E., Henrickson, S. E., Zhao, H., Ibbotson, R. E., Orchard, J. A. and Davis, Z. ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. *Blood*, 101, 12 (2003), 4944-4951.
[116] Edwards, B. K., Ward, E., Kohler, B. A., Eheman, C., Zauber, A. G., Anderson, R. N., Jemal, A., Schymura, M. J., Lansdorp-Vogelaar, I. and Seeff, L. C. Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. *Cancer*, 116, 3 (2010), 544-573.
[117] Siegel, R. L., Miller, K. D. and Jemal, A. Cancer statistics, 2016. *CA: a cancer journal for clinicians*, 66, 1 (2016), 7-30.
[118] Kinzler, K. W. and Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell*, 87, 2 (1996), 159-170.
[119] Hong, Y., Downey, T., Eu, K. W., Koh, P. K. and Cheah, P. Y. A 'metastasis-prone'signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. *Clinical & experimental metastasis*, 27, 2 (2010), 83-90.
[120] Belov, L., Zhou, J. and Christopherson, R. I. Cell surface markers in colorectal cancer prognosis. *International journal of molecular sciences*, 12, 1 (2010), 78-113.
[121] Besson, D., Pavageau, A.-H., Valo, I., Bourreau, A., Bélanger, A., Eymerit-Morin, C., Moulière, A., Chassevent, A., Boisdron-Celle, M. and Morel, A. A quantitative proteomic approach of the different stages of colorectal cancer establishes OLFM4 as a new nonmetastatic tumor marker. *Molecular & Cellular Proteomics*, 10, 12 (2011), M111. 009712.
[122] Huang, M.-Y., Wang, H.-M., Chang, H.-J., Hsiao, C.-P., Wang, J.-Y. and Lin, S.-R. Overexpression of S100B, TM4SF4, and OLFM4 genes is correlated with liver metastasis in Taiwanese colorectal cancer patients. *DNA and cell biology*, 31, 1 (2012), 43-49.

[123] Chia, N.-Y., Deng, N., Das, K., Huang, D., Hu, L., Zhu, Y., Lim, K. H., Lee, M.-H., Wu, J. and Sam, X. X. Regulatory crosstalk between lineage-survival oncogenes KLF5, GATA4 and GATA6 cooperatively promotes gastric cancer development. *Gut*, 64, 5 (2015), 707-719.
[124] Lei, Z., Tan, I. B., Das, K., Deng, N., Zouridis, H., Pattison, S., Chua, C., Feng, Z., Guan, Y. K. and Ooi, C. H. Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. *Gastroenterology*, 145, 3 (2013), 554-565.
[125] Wu, Y., Grabsch, H., Ivanova, T., Tan, I. B., Murray, J., Ooi, C. H., Wright, A. I., West, N. P., Hutchins, G. G. and Wu, J. Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer. *Gut*, 62, 8 (2013), 1100-1111.
[126] Kuner, R., Muley, T., Meister, M., Ruschhaupt, M., Buness, A., Xu, E. C., Schnabel, P., Warth, A., Poustka, A. and Sültmann, H. Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung cancer*, 63, 1 (2009), 32-38.
[127] Hagan, R. D., Phillips, C. A., Wang, K., Rogers, G. L., Loway, C. and Langston, M. A. On the relative significance of kernelization versus branching for parallel FPT implementations.*Proceedings, International Conference on Parallel and Distributed Computing and Networks*, 2013.
[128] Abu-Khzam, F. N., Langston, M. A., Shanbhag, P. and Symons, C. T. Scalable Parallel Algorithms for FPT Problems. *Algorithmica*, 45 (2006), 269-284.
[129] Langston, M. A. Fixed-Parameter Tractability, A Prehistory. *Lecture Notes in Computer Science*, 7370 (2012), 3-16.
[130] Weerapurage, D. P., Eblen, J. D., Rogers, G. L. and Langston, M. A. Parallel Vertex Cover: A Case Study in Dynamic Load Balancing. *Proceedings, Australasian Symposium on Parallel and Distributed Computing*, 118 (2011), 25-32.
[131] Niedermeier, R. *Invitation to Fixed-Parameter Algorithms*. Oxford University Press, 2006.

[132] Chen, J., Kanj, I. and Xia, G. Improved Parameterized Upper Bounds for Vertex Cover. *International Symposium on Mathematical Foundations of Computer Science.* Springer Berlin / Heidelberg, 2006.
[133] Niedermeier, R. and Rossmanith, P. A General Method to Speed Up Fixed-Parameter-Tractable Algorithms. *Information Processing Letters*, 73, 3-4 (2000), 125-129.
[134] Abu-Khzam, F. N., Fellows, M. R., Langston, M. A. and Suters, W. H. Crown Structures for Vertex Cover Kernelization. *Theory of Computing Systems*, 41 (2007), 411-430.
[135] Langston, M. A., Perkins, A. D., Saxton, A. M., Scharff, J. A. and Voy, B. H. Innovative Computational Methods for Transcriptomic Data Analysis: A Case Study in the Use of FPT for Practical Algorithm Design and Implementation. *The Computer Journal*, 51 (2008), 26-38.
[136] Erdös, P. and Rényi, A. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 5 (1960), 17-61.
[137] Watts, D. J. and Strogatz, S. H. Collective Dynamics of `Small-World' Networks. *Nature*, 393 (1998), 440–442.
[138] Barabási, A. L. and Albert, R. Emergence of scaling in random networks. *Science*, 286 (1999), 509-512.

VITA

Ronald Hagan was born and raised in Morristown Tennessee. After graduating from Morristown East High School, he briefly attended the University of Tennessee before leaving and marrying his lovely wife, Cynthia Dawn Carpenter, in 1991. His eldest son Brett was born the following year. Over the next few years he worked as an upholsterer and fabric cutter, first at England's in Tazewell, TN and then at Berkline in Morristown. While working graveyard shifts at Berkline, he returned to school, finishing a bachelor's degree in Mathematics at Carson Newman College in Jefferson City, TN in 1996. Upon graduation, he went to the University of Tennessee to pursue a master's degree in Mathematics, finishing under the direction of Dr. Carl Wagner in 2000. He then accepted a position at his alma mater, Morristown East, teaching mathematics for two years. In 2002, after the birth of his second son, Isaac, he returned to the Mathematics department at the University of Tennessee, serving first as a Lecturer, then Senior Lecturer until 2016. Currently he is a Senior Research Engineer at BAE Systems. Since 2010, he has been working on the completion of a PhD in Computer Science at the University of Tennessee under the direction of Dr. Michael A. Langston.