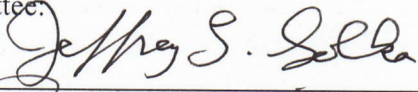
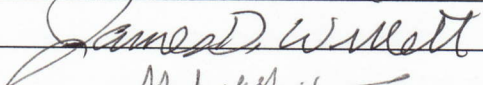


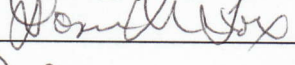
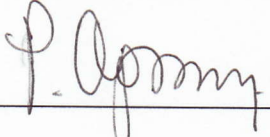


MULTIPLE KERNEL LEARNING FOR GENE PRIORITIZATION, CLUSTERING,
AND FUNCTIONAL ENRICHMENT ANALYSIS

by

David H. Millis
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics and Computational Biology

Committee:

	Dr. Jeffrey L. Solka, Dissertation Director
	Dr. James D. Willett, Dissertation Chair
	Dr. Lakshmi K. Matukumalli, Committee Member
	Dr. James D. Willett, Department Chair
	Dr. Donna Fox, Associate Dean, Student Affairs & Special Programs, College of Science
	Dr. Peggy Agouris, Interim Dean, College of Science
Date: <u>April 29, 2014</u>	Spring Semester 2014 George Mason University Fairfax, VA

Multiple Kernel Learning for Gene Prioritization, Clustering, and Functional Enrichment
Analysis

A Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at George Mason University

by

David H. Millis
Doctor of Medicine
Howard University College of Medicine, 1983

Director: Jeffrey L. Solka, Professor
Department of Bioinformatics and Computational Biology

Spring Semester 2014
George Mason University
Fairfax, VA



This work is licensed under a [creative commons attribution-noncommercial 3.0 unported license](https://creativecommons.org/licenses/by-nc/3.0/).

DEDICATION

In memory of my parents,

Ena Grace Millis

Clovis Bolton Millis

You believed that I could do anything.

ACKNOWLEDGMENTS

I would like first to express gratitude to my dissertation committee. Dr. Jeffrey Solka served as dissertation director, helped provide order and focus to this research effort, and patiently reviewed multiple document drafts. Dr. Lakshmi Matukumalli allowed access to data generated by the USDA Bovine Functional Genomics Laboratory, and gave wise, practical advice on the dissertation completion process. Dr. James Willett served as dissertation chair, offered biological insights that helped move the research forward, and provided validation and support for the usefulness of the ideas developed over the course of this project.

Many thanks to Dr. Ronald Kostoff, who arranged funding for a summer internship with The Mitre Corporation and provided helpful insights on text mining and literature-based discovery.

I worked at the Thomas B. Finan Center, a psychiatric hospital in Cumberland, Maryland, while completing this dissertation. To the leadership, staff, and patients of the Finan Center: You have taught me much about suffering, survival, resilience, community, and the challenges of being human. For all you have given me, I will always be grateful.

To the many music teachers and church musicians with whom I worked over the years, in New York, California, and Maryland: I learned a lot from you about persistence, about being unafraid to make mistakes, and about honing one's craft by always going back to the fundamentals. Everything that you taught me about artistic growth and craftsmanship is reflected in these pages.

To my sisters Dianne Millis and Kathy Brown, brother Michael Millis, and brothers-in-law Eric Brown and Peter Brown: Each of you has had a role in making this accomplishment a reality. Thank you for your love and support throughout this long process. Spending time with you during the holidays always reminds me of the things in life that are truly important.

To my relatives across the country and around the world, in New York, New Jersey, Connecticut, Georgia, Florida, California, England, and Jamaica: Although we have been far apart geographically, we remain close in spirit. I am honored to make this accomplishment a part of our family history.

Finally, to my nephews Marcus Brown and Jordan Brown: Find something that you like doing and that the world needs to get done, and have fun working hard at it. The world needs you. Everything you need to change the world, you already possess. You are perfect as you are.

David H. Millis, M.D., Ph.D.
Gainesville, Virginia
April 29, 2014

TABLE OF CONTENTS

	Page
List of Tables	x
List of Figures	xi
List of Equations	xii
List of Abbreviations	xiii
Abstract	xiv
Chapter One: Introduction	1
Research Problem: Gene Prioritization	1
Research Hypothesis	5
Research Contributions	5
Chapter Two: Ensemble Methods for Classification	7
Classification Strategies	7
Ensemble Classifiers	8
Structure of Base Classifiers.....	9
Design of Base Classifier Training Sets	10
Combining Classifier Outputs	12
Chapter Three: Kernel Methods and Multiple Kernel Learning.....	14
Kernel Methods	14
Support Vector Machines.....	15
Combining Kernels	20
Multiple Kernel Learning.....	20
Chapter Four: Text Processing Methods for Measuring Gene Similarity	25
Gene Similarity Based on Shared Abstracts	26
Gene Similarity Based on Co-Occurrence of Gene Names in Abstracts	26
Gene Similarity Based on Cosine Similarity of Free Text in Abstracts.....	27
Chapter Five: Classifier Design	29
Classifier Design: Overview	29

Data Sources.....	30
The Gene Ontology (GO) Database	30
The KEGG Pathway Database	31
The REACTOME Pathway Database.....	31
The STRING Protein-Protein Interaction (PPI) Database.....	31
PubMed: Shared Identifiers	32
PubMed: Named Entity Recognition.....	32
PubMed: Cosine Similarity of Abstracts.....	33
MicroRNA Target Prediction Algorithms	34
Risperidone Differential Expression Data.....	34
Data Preprocessing.....	35
SVM Step	36
MKL Step.....	36
Classifier Diversity.....	37
Measuring Classifier Performance	37
Selecting Best MKL Classifier.....	38
Clustering Methods	38
Functional Enrichment Analysis	39
Chapter Six: Project #1: microRNA-Gene Interactions.....	41
Background: Biology of microRNAs and microRNA Target Prediction	41
MicroRNAs and Gene Expression	42
MicroRNA Target Prediction.....	44
Computational microRNA Target Prediction Methods	44
miRanda.....	45
PITA	45
RNAhybrid	46
DIANA-microT	46
PicTar.....	46
TargetScan.....	47
mirTarget2	47
Methods.....	47
Kernel Matrices for Classifier Training and Testing.....	48

Kernel Matrices for Labeling Unlabeled Genes	50
Results	50
Micro-RNA Gene Interactions: Classifier Performance	50
Micro-RNA Gene Interactions: Biological Plausibility	52
MicroRNA-Gene Interactions: Functional Annotation Analysis	57
MicroRNA-Gene Interactions: Using Target Rankings to Identify Genes with Possible Relationship to Horn Development.....	58
Chapter Seven: Project #2: Neurotransmitter-Gene Interactions.....	63
Background: Melatonin and Serotonin	63
Methods.....	64
Results	65
Serotonin-Gene Interactions: Classifier Performance	65
Serotonin-Gene Interactions: Biological Plausibility of Highly Ranked Unknown Interaction Partners.....	66
Serotonin: Clustering and Functional Enrichment Analysis	68
Melatonin-Gene Interactions: Classifier Performance	70
Melatonin-Gene Interactions: Biological Plausibility of Highly Ranked Unknown Interaction Partners.....	72
Melatonin: Clusters and Functional Enrichment Analysis	74
Chapter Eight: Project #3: Drug-Gene Interactions.....	78
Background: Risperidone.....	78
Methods.....	79
Gene Expression Data Set	79
Identification of Examples for Classifier Training.....	80
Results	81
Risperidone-Gene Interactions: Classifier Performance	81
Risperidone-Gene Interactions: Biological Plausibility of Highly Ranked Unknown Interaction Partners.....	83
Risperidone-Gene Interactions: Clustering and Functional Enrichment Analysis....	86
Chapter Nine: Conclusions	91
Contributions.....	91
Limitations and Areas for Future Work	93
References.....	95

LIST OF TABLES

Table	Page
Table 1. Kernel functions used in this dissertation.	19
Table 2. Best mir1-gene base classifiers.	50
Table 3. Ten highest-diversity mir1-gene ensemble classifiers out of 225 classifiers with AUROC of 0.75.	51
Table 4. Group 1: microRNA interaction partners related to regulation of transcription and DNA replication.	53
Table 5. Group 2: microRNA interaction partners related to glycolipid and lipid metabolism.	54
Table 6. Group 3: microRNA interaction partners with other functions.	56
Table 7. Most highly enriched functions for the highest-ranked interaction partners shared by the 14 microRNA studied.	57
Table 8. Predicted microRNA targets in region 0.8MB to 2.8MB on Chromosome 1. ...	59
Table 9. Genes on Chr1, 0.8MB to 2.8MB, predicted as microRNA targets.	61
Table 10. Best serotonin-gene base classifiers.	65
Table 11. Serotonin-gene ensemble classifiers with AUROC score of 0.9839.	66
Table 12. Previously unlabeled genes in the core gene set that were ranked highest as potential interaction partners for serotonin.	66
Table 13. Functional annotations for clusters of genes ranked highly as interaction partners for serotonin.	69
Table 14. Best melatonin-gene base classifiers.	71
Table 15. Melatonin-gene ensemble classifiers with AUROC score of 0.9222.	71
Table 16. Previously unlabeled genes in the core gene set that were ranked highest as potential interaction partners for melatonin.	72
Table 17. Functional annotations for clusters of genes ranked highly as interaction partners for melatonin.	75
Table 18. Best risperidone-gene base classifiers.	81
Table 19. Risperidone-gene ensemble classifiers with AUROC score of 0.8571.	82
Table 20. Previously unlabeled genes in the core gene set that were ranked highest as potential interaction partners for risperidone.	83
Table 21. Functional annotations for clusters of genes ranked highly as interaction partners for risperidone.	86
Table 22. Genes in Cluster 14, one of the clusters of potential interaction partners for risperidone.	88

LIST OF FIGURES

Figure	Page
Figure 1. Graphical representation of a support vector machine classifier. Positive training examples are shown as white circles. Negative training examples are shown as black circles. A new instance to be classified is shown as a gray circle.....	17
Figure 2. Sonnenburg's multiple kernel learning algorithm (Sonnenburg et al., 2006)....	24

LIST OF EQUATIONS

Equation	Page
Equation 1. The solution to this convex quadratically constrained quadratic program forms the basis of a support vector machine.....	18
Equation 2. Convex quadratically constrained quadratic program representing a conic combination of kernel matrices (G. R. G. Lanckriet et al., 2004).	21
Equation 3. Equivalent dual formulation of Equation 2 (Sonnenburg et al., 2006).....	21
Equation 4. Reorganization of Equation 3 by rearrangement of terms (Sonnenburg et al., 2006).	22
Equation 5. Conversion of Equation 4 to a semi-infinite linear program (Sonnenburg et al., 2006).	22

LIST OF ABBREVIATIONS

3' untranslated region	3'UTR
area under receiver operating characteristic.....	AUROC
biological process (Gene Ontology annotation category)	BP
Chromosome 1	Chr1
Gene Ontology	GO
hidden Markov model.....	HMM
Kyoto Encyclopedia of Genes and Genomes.....	KEGG
microRNA.....	miRNA
multiple kernel learning	MKL
named entity recognition.....	NER
Probabilistic Identification of Combinations of Target Sites	PicTar
Probability of Interaction by Target Accessibility.....	PITA
protein-protein interaction	PPI
PubMed identifier	PMID
quadratically constrained quadratic program.....	QCQP
receiver operating characteristic	ROC
Search Tool for Interactions of Chemicals	STITCH
Search Tool for the Retrieval of Interacting Genes/Proteins	STRING
semi-infinite linear program	SILP
support vector machine	SVM
term frequency-inverse document frequency.....	tf-idf
with respect to	w.r.t.

ABSTRACT

MULTIPLE KERNEL LEARNING FOR GENE PRIORITIZATION, CLUSTERING, AND FUNCTIONAL ENRICHMENT ANALYSIS

David H. Millis, M.D., Ph.D.

George Mason University, 2014

Dissertation Director: Dr. Jeffrey L. Solka

Gene prioritization is the process of ranking a list of candidate genes such that the genes that are most likely involved in a biological process of interest receive the highest rankings. In a supervised learning approach to gene prioritization, candidate genes are ranked in terms of their degree of similarity to genes that have already been shown to be involved in the process of interest. Gene prioritization thus can be cast as a classification task, in which a training set of genes and data associated with those genes is used to train a classifier to assign rankings to unknown genes, based on their degree of similarity to the training genes. This thesis describes the use of kernel methods, and particularly a method known as multiple kernel learning, for combining information from multiple data sources for purposes of gene prioritization. Multiple kernel learning facilitates the incorporation of heterogeneous data types into the assessment of similarity among genes. In addition, the rows of the kernel matrix can be repurposed as feature vectors. We apply clustering

methods to these vectors to partition the gene list into related groups. We then perform functional enrichment analysis on the gene clusters to identify biological functions that are significantly represented in each gene cluster. We thus are able to use a single data structure, namely a kernel matrix representing similarities among genes based on multiple information sources, as the basis for three common types of bioinformatics analysis: gene prioritization, gene clustering, and functional annotation analysis of gene lists. This research contributes to the exploration of methods for extracting useful biological insights from the continually expanding knowledge base of biological data.

CHAPTER ONE: INTRODUCTION

Research Problem: Gene Prioritization

A common type of bioinformatics analysis begins with the identification of some biological entity of interest, and concludes with the generation of a list of candidate interaction partners that may be worthy of further study for their relationship to the entity of interest. The biological entity of interest could be a physical structure, such as a nucleic acid, small molecule, or pharmaceutical agent. It also could be a descriptive construct, such as a disease or phenotype. The candidate interaction partners often are genes, and this dissertation will focus on the analysis of lists of candidate genes.

Gene prioritization is the process of ranking a list of candidate genes such that the genes that are most likely involved in a biological process of interest receive the highest rankings. In a supervised learning approach to gene prioritization, candidate genes are ranked in terms of their degree of similarity to genes that have already been shown to be involved in the process of interest. Gene prioritization thus can be cast as a classification task, in which a training set of genes and data associated with those genes is used to train a classifier to assign rankings to unknown genes, based on their degree of similarity to the training genes.

Once a list of candidate genes is generated, the researcher is left with the task of making a biological interpretation of the list. If the candidate genes are given a numerical score in the course of the analysis, the candidates with the highest scores might be

considered to be the most important. However, simply selecting the highest-scoring genes from the front of the list might result in discarding useful information that is embedded in the rest of the list. If the bioinformatics analysis generated a list of several hundred or several thousand candidate genes, interpretation of the entire list can be difficult.

An additional problem is that the candidate genes that are included in the list may differ depending on the data sources used in the bioinformatics analysis. This necessitates the selection of a strategy for generating a single consensus list of candidate genes for further analysis.

One approach might be to run the analysis multiple times, each time using a different information source for identifying potential candidate genes. This will result in multiple different lists of candidate genes. The researcher would then have to devise a strategy for combining the lists. The optimal strategy for doing this might not be obvious. The different analyses may rank the genes in different orders, using different scales of measurement.

A different approach might be to combine the data sources and run the bioinformatics analysis just once, to create a single list of candidate genes. A potential obstacle with this approach is that the different bioinformatics data sources may use different formats for data representation, making it difficult to incorporate the data sources into a single representation for bioinformatics analysis.

This dissertation will focus on the second approach: combining the diverse data sources into a single data structure, and then performing an analysis of the combined data to generate the candidate gene list. We will examine the use of kernel methods, and in

particular multiple kernel learning (MKL) using support vector machines, to address many of the difficulties involved in the generation and interpretation of candidate gene lists.

Kernel methods are a machine learning approach that has gained much interest in the bioinformatics research community. An attractive aspect of kernel methods is that a collection of data points gathered on a set of genes can be expressed in terms of pairwise relationships among the genes, instead of in the data format of the original data collection method. If the set of data points collected on one gene can be represented as a vector, then the relationship between two genes can be represented as some function of the dot product of the two vectors.

Any function that can be expressed in terms of the dot product of vectors can be used to transform the vectors to a higher-dimensional space. When used in this way, the function is called a kernel function. We can construct a matrix consisting of the application of the kernel function to all pairs of vectors in the data set. This matrix, called a kernel matrix, maps the entire data set of vectors to a higher-dimensional feature space. This mapping can sometimes make it easier to computationally identify patterns in the data set that might not be easy to recognize in the original vector space.

In many bioinformatics studies, there may be several diverse data sources that can provide data on the set of genes. For this situation, each data source can be used to generate a separate kernel matrix. This provides multiple views on the set of genes, with each view based on a different source of data on the genes. An attractive aspect of kernel methods is that the kernel matrix generated from each data source describes the genes in

terms of pairwise relationships among the genes, instead of in the data format of the original data collection method. With the various original data formats abstracted away, it becomes more tractable to perform mathematical operations that combine the different data sources (G. R. Lanckriet, Deng, Cristianini, Jordan, & Noble, 2004).

We will demonstrate that multiple kernel learning can provide the following benefits to the analysis of candidate gene lists:

1. Heterogeneous bioinformatics data sources can be combined into a single data representation, using the mathematics of kernel combination.

2. The confidence level calculated for each gene by the MKL classifier, which is typically translated into a binary label, can also be used for creating an ordered ranking of the genes. A higher confidence level indicates a higher degree of relatedness to the molecule of interest.

3. Each row of the kernel matrix can be treated as a vector encapsulating a profile of a single gene. Clustering the kernel rows provides a way of breaking the gene list down into smaller lists of similar genes.

4. For each cluster, we can calculate the average confidence level of all genes in the cluster. This allows us to rank the clusters, and identify the clusters composed of genes that are most highly related to the molecule of interest.

5. For each cluster, we can perform functional enrichment analysis to identify Gene Ontology terms that are significant for that cluster. This allows for a finer-grained functional analysis than submitting the entire list of candidate genes for enrichment

analysis. The functions identified for the highest-ranking clusters may be the more important biological functions of the molecule of interest.

Research Hypothesis

The research hypothesis for this dissertation is as follows:

“Multiple kernel learning enhances the bioinformatics analysis of candidate gene lists by [1] allowing integration of heterogeneous data sources into the analysis, [2] providing a prioritized ranking of genes in the list, [3] facilitating clustering of the genes using the kernel matrix, and [4] simplifying functional enrichment analysis by clustering the original gene list into meaningful sub-lists.”

Research Contributions

This dissertation demonstrates the benefits of kernel methods for addressing several issues that arise in the analysis of biological data: integration of diverse data types for problem solving, gene prioritization to add order to a long list of candidate genes, and separation of a long gene list into a set of smaller, meaningful lists to facilitate further analysis of the gene list. In this research, kernel methods are applied to combine multiple data sources to rank a gene list. The kernel matrix is used as a similarity matrix to partition the gene list into meaningful sub-lists. The priority rankings within clusters are used to identify the most biologically significant clusters. We use functional enrichment analysis on the clusters with the most highly-ranked genes to identify the most significant functions of the most significant clusters.

Essentially, this approach starts with an unordered list of genes related to a target, and generates a prioritized ranking that provides structure to the list. This structuring not

only identifies the genes that should be selected for further study, but also allows generation of hypotheses on the most significant biological functions of the target.

The construction and analysis of a candidate gene list is a common bioinformatics problem. The methods described here are applicable to a wide range of problem domains that involve prioritizing a list of candidate genes.

CHAPTER TWO: ENSEMBLE METHODS FOR CLASSIFICATION

Classification Strategies

In supervised learning, a training set of instances that have already been assigned correct labels is used by a classification algorithm as a gold standard from which a model of the mapping from instance features to appropriate labels is derived. Each label represents a class of instances, where instances within a class have similar sets of features, and instances within a class have sets of features that are different from the features of instances in the other classes. The objective of a supervised machine learning algorithm is to use the labeled training set to develop a decision model that a classifier can use to assign labels to new instances, where the features are known but the correct labels are not known.

A classification problem may involve any number of features, and any number of correct classes. Restriction of the features used as inputs to the machine learning process can reduce the computational complexity of the learning task, particularly when the number of features is very large. However, visual inspection of the data set may not provide clues about the features that contribute most to the class assignment. Training and testing classifiers on a training data set can help identify the most useful features for the classification task. A classifier is trained and tested on various different subsets of the training data set, where each subset leaves out one or more of the features from the

original training set. The performance of the classifier is measured after each round of training and testing. The differences in performance of the various versions of the classifier give an indication of how much the availability of a given feature or combination of features contributes to the ability of the classifier to assign correct class labels.

The selection of the best machine learning algorithm for a classification problem can be quite challenging. Factors that may influence the performance of a particular machine learning algorithm include the shape of the decision boundary that separates instances in different classes from one another. However, the shape of the decision boundary is typically unknown. Identifying the best machine learning algorithm for a classification problem can be attempted by training and testing classifiers using different learning algorithms on the same training set (Duda, Hart, & Stork, 2000).

Ensemble Classifiers

Because the selection of a single best classifier for a classification problem can be difficult, we might consider running several different classifiers and then pooling their results. The objective of this approach is to help mitigate any weaknesses in the individual classifiers, and possibly to generate a labeling of instances that is better than could be provided by any of the individual classifiers. Terms used to describe a group of classifiers include ensemble systems (Dietterich, 2000; Drucker, Cortes, Jackel, LeCun, & Vapnik, 1994), multiple classifier systems (Fumera & Roli, 2005; Ho, Hull, & Srihari, 1994; Re & Valentini, 2010; Woods, Kegelmeyer, & Bowyer, 1997; L. Xu, Krzyzak, &

Suen, 1992), committee of classifiers (Drucker et al., 1994; Hady & Schwenker, 2010), and mixture of experts (Jacobs, Jordan, Nowlan, & Hinton, 1991; Jordan & Jacobs, 1994; Khalili, 2010; Ng & McLachlan, 2007). The term *base classifiers* is often used to denote the individual classifiers that compose a classifier ensemble (L. Kuncheva, 2004).

The design of an ensemble classifier requires consideration of three key questions (Polikar, 2006):

1. Will the base classifiers be of the same type (i.e. all using the same machine learning algorithm) or different types?
2. What data will be used to train each of the base classifiers?
3. How will the outputs of the base classifiers be combined?

Structure of Base Classifiers

The base classifiers in an ensemble can be all of the same type, such as all decision trees, all neural networks, or all support vector machines. If the set of available training instances is very large, it may be possible to train multiple classifiers by simply dividing the training instances into multiple subsets. Each base classifier could then be trained on one of the training subsets. A strategy would then be required to combine the outputs of the base classifiers into the ensemble's final output.

An ensemble of multiple base classifiers of the same type might be designed if it is difficult to train a single classifier to cover the entire feature space. If the feature space can be divided into smaller partitions, it may be possible to train individual classifiers to

distinguish instances based on the smaller feature sets. In this way, each base classifier becomes an expert on how a small number of features relate to the class labels. The outputs of the base classifiers can then be combined into the consensus set of labels for the ensemble (Giacinto, Roli, & Didaci, 2003).

Instead of dividing the training set, all training examples can be used for each base classifier, while changing the training parameters. For example, an ensemble of classifiers might be composed of a set of neural networks, with each using different initial edge weights and different numbers of layers (El-Melegy & Ahmed, 2007).

The third strategy for introducing diversity in the ensemble is to include base classifiers that use different classification algorithms. This strategy can incorporate results from classifiers based on different architectures, such as k -nearest neighbor, decision tree, and support vector machine, into a single ensemble (Aszfalg, Kriegel, Pryakhin, & Schubert, 2007).

Design of Base Classifier Training Sets

In many classification problems, the number of features is large, but the number of available training instances is small. For these situations, dividing the training instances into smaller non-overlapping groups may not provide a large enough number of instances for any of the base classifiers to build useful models. Several resampling methods have been developed to allow the creation of multiple training sets by allowing the inclusion of some instances in more than one training set. This reuse of instances

across training sets allows each base classifier to be exposed to a different training set, while providing each base classifier with a large enough number of instances for model generation (Breiman, 1996).

Training base classifiers with the same learning algorithm but different training sets means that the classification model learned by each classifier will be slightly different. The more different the models learned by the base classifiers, the higher the diversity of the ensemble (L. I. Kuncheva & Whitaker, 2003). Diversity is a useful characteristic of a classifier ensemble. In an ideal diversification of an ensemble, each base classifier generally performs well, while each makes errors on different types of instances. The hope is that the classifier outputs can then be combined such that the overall error of the ensemble is less than that of the individual base classifiers.

Creation of different training sets typically involves drawing instances randomly from the original full set of training examples. If the original training set is small, dividing the original training set into equal partitions may result in subsets that are too small for the base classifiers to build adequate models. One alternative is to select instances randomly with replacement, known as *bootstrap aggregation* or *bagging* (Breiman, 1996). An alternative is the jackknife or *k-fold data split*, in which the data set is divided into k partitions without replacement, and then each classifier is trained on $k-1$ partitions (Efron, 1982; Polikar, 2006). In both cases, some instances will be presented to more than one base classifier during training, but each base classifier is exposed to a

unique set of training instances. The base classifiers in the ensemble may all use the same classification algorithm, but their performance is likely to differ because of the differences in the training sets used.

Combining Classifier Outputs

Methods for combining classifier outputs can be categorized in two ways: as supervised vs. unsupervised methods, and as methods that combine categorical classifier outputs vs. methods that combine numeric classifier outputs (Polikar, 2006). In the supervised case, a separate training algorithm is used to determine optimal weights for combining classifier outputs. The unsupervised case includes rules such as majority voting. Combination methods based on numeric classifier outputs operate on values computed by the base classifiers for each instance in the data set. When functioning as a solo classifier, each base classifier would apply some rule to each of these values to determine how each instance will be labeled. In ensemble mode, some operation is performed on a set of values generated for each instance before determining how that instance will be labeled. Combination methods based on categorical classifier outputs allow each base classifier to both complete its calculations and apply its own rule to assign a label to each instance. The ensemble then performs some operation on the labels assigned to each instance by the base classifiers to determine the final set of labels that will be generated by the ensemble.

In the following chapter, we introduce multiple kernel learning, the classifier approach applied in this dissertation. This is a supervised classification method. The base classifiers are individual support vector machines. After training of several base

classifiers, the best-performing base classifiers are selected. The kernel matrices underlying these classifiers are combined to generate a single ensemble classifier. We can then evaluate the performance of the ensemble classifier and compare its performance to that of the base classifiers.

CHAPTER THREE: KERNEL METHODS AND MULTIPLE KERNEL LEARNING

Kernel Methods

A kernel function is any function that can be represented as the dot product of vectors. Given a matrix of $n \times m$ data points, in which the n rows represent instances and the m columns represent features, we can construct a kernel matrix of dimension $n \times n$ in which each cell is computed by application of the kernel function to two of the n row vectors. The kernel matrix thus serves as a similarity matrix consisting of pairwise comparisons of each instance to all the other instances, using the kernel function as the measure of similarity. For the examples in this dissertation, each data set is comprised of a single type of data collected on a list of genes, where each row vector consists of several features related to a single gene. Application of a kernel function to all pairs of row vectors generates a kernel matrix representing pairwise similarity of all genes in the gene list to all other genes in the list

There are several advantages to the use of kernel methods in the analysis of data from heterogeneous sources of genomic data. The creation of a separate kernel matrix for each available data source provides different views or perspectives on the genes and on their relationships to each other. While the original data sources may store data in a wide variety of formats, all kernel matrices represent the relationships between genes as pairwise similarities. Abstracting out the original data formats makes it easier to apply

straightforward mathematical operations to compare and combine the data sources.

Additionally, the mapping of a original data to a higher-dimensional space may reveal patterns that are easier to identify than in the original vector representation of the data source.

Support Vector Machines

The support vector machine (SVM) is a kernel method that has gained wide adoption in machine learning applications. A typical application area for support vector machines is binary classification. The support vector machine approach applies an optimization algorithm to identify a boundary that provides maximal separation between labeled instances in a training set of data. Identifying a boundary that separates the training items into two different classes can be made easier by first applying a function that maps the items into an alternative space, known as the kernel space. Items in a test set can then be classified by mapping them to the same kernel space, and noting their position relative to the hyperplane. Items that map to positions on the same side of the hyperplane as the positive examples are assigned to the positive class, and items that map to the same side as the negative examples are assigned to the negative class. An item in the test set can also be ranked based on its distance from the hyperplane. If the magnitude of the distance is small, then the instance from the test set maps to a position close to the hyperplane. We might consider such an instance to be only slightly representative of its assigned class, since it lies fairly close to training examples in the opposite class. If the magnitude of the distance is large, then the instance from the test set maps to a position closer to examples of its assigned class, and farther from the examples of the opposite

class. Such an instance might be considered to be more strongly representative of its assigned class. Distance from the hyperplane may also be useful in identifying unusual instances. For example, an instance that lies more than three standard deviations from the hyperplane might be classified as an outlier.

The support vector machine (SVM) is a binary classification algorithm that has gained wide adoption in machine learning applications. The algorithm was first introduced in 1992 (Boser, Guyon, & Vapnik, 1992), and is based on principles from statistical learning theory (Vapnik, 1999). This learning method takes as input a set of known, labeled instances, represented as vectors, and applies an optimization algorithm to identify a function that characterizes the relationship between the vector values and the labels. The function defines a boundary, known as a hyperplane, which provides the best possible separation between instances of two different classes. The function identified through examination of labeled instances can then be used to classify new instances whose labels are unknown. The new instances are classified by mapping them to the same kernel space as the training instances, and noting their position relative to the hyperplane. An item in the test set can also be ranked based on its distance from the hyperplane. Items on the positive side of the hyperplane are classified as members of the positive class, while items on the negative side of the hyperplane are classified as members of the negative class. Items very close to the hyperplane may be ambiguous and difficult to classify, possessing a set of features that is intermediate between the positive class and the negative class. Items that are very distant from the hyperplane in either the

positive or negative direction may be worth considering as possible outliers. A graphical representation of a support vector machine classifier is depicted in Figure 1.

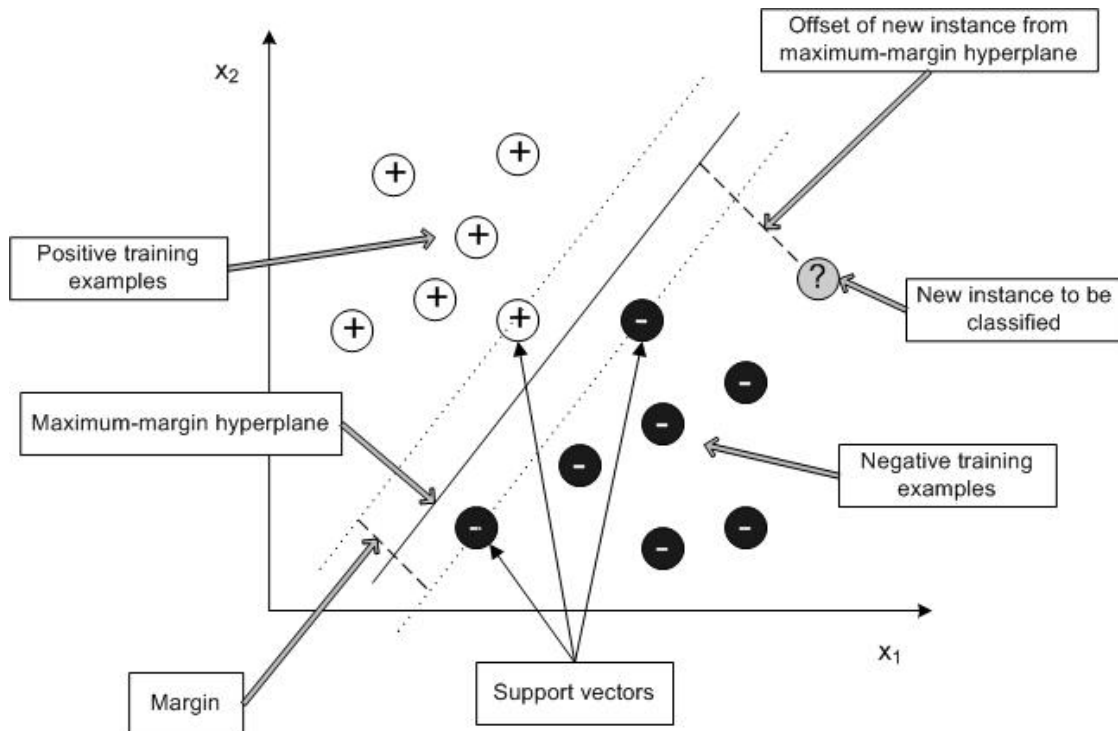


Figure 1. Graphical representation of a support vector machine classifier. Positive training examples are shown as white circles. Negative training examples are shown as black circles. A new instance to be classified is shown as a gray circle.

The set of training instances consists of N pairs composed of vectors \mathbf{x}_i and labels y_i , where $i = 1, \dots, N$, $\mathbf{x}_i \in \mathbb{R}^N$, and $y \in \{1, -1\}^N$. A support vector machine requires a solution to a convex optimization problem, and can be cast as the convex quadratically constrained quadratic program (QCQP) shown in Equation 1 (Cortes & Vapnik, 1995):

Equation 1. The solution to this convex quadratically constrained quadratic program forms the basis of a support vector machine.

$$\begin{aligned}
 \min \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^N \xi_i \\
 \text{w.r.t.} \quad & \mathbf{w} \in \mathbb{R}^N, \xi \in \mathbb{R}^N, b \in \mathbb{R}, \\
 \text{subject to} \quad & y_i \langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle + b \geq 1 - \xi_i \\
 & \text{and } \xi_i \geq 0, \forall_i = 1, \dots, N
 \end{aligned}$$

In this formulation of the optimization problem, the ξ_i are error values, C is a pre-defined cost parameter for the error term, representing a trade-off between model simplicity and classification error (Gonen & Alpaydin, 2011), and b is an offset parameter estimated from the training examples (Poggio, Mukherjee, Rifkin, Rakhlin, & Verri, 2001).

We would like to use some function Φ to map the training vectors to a higher-dimensional space, where the separation between instances of the two classes might be more readily observable than in the original feature space. However, explicit mapping of features from the original feature space to the higher-dimensional space may be computationally expensive. Machine learning methods based on kernels typically apply the so-called “kernel trick” (Aizerman, Braverman, & Rozoner, 1964) to avoid this mapping. Instead of mapping each of the training vectors individually to the higher-dimensional space, we apply a kernel function to all pairwise combinations of the training vectors. A kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ is any function that can be represented in terms of the dot product between two vectors: $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$. Application of the kernel function to two vectors in the original feature space returns the dot product of the images of the two vectors in the higher-dimensional feature space. We can then compute the

linear separation of the dot-product images, without an explicit higher-dimensional mapping (Scholkopf & Smola, 2001).

The optimization algorithm does not identify the best kernel function to use for the domain of the classification problem. The investigator must choose from among a large number of possible kernel functions. Some experimentation might be required to identify a kernel function that provides a useful mapping of the original data into an alternative feature space. For this dissertation, we experimented with three commonly-used kernel functions (Karatzoglou, Smola, Hornik, & Zeileis, 2004). The equations for these functions are presented in Table 1.

Table 1. Kernel functions used in this dissertation.

Name	Kernel Function
Linear:	$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
Polynomial:	$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^T \mathbf{x}_j + r)^d, \gamma > 0$
Gaussian:	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \ \mathbf{x}_i - \mathbf{x}_j\ ^2), \gamma > 0$

The kernel parameters γ , r , and d also are not selected by the optimization algorithm. Identification of a useful set of kernel parameters typically requires some experimentation by the investigator.

Combining Kernels

There are several approaches to combining kernels. Kernels can be combined using simple addition of corresponding kernel matrix cells. In an equal weighting approach, all kernels contribute equally to the combined kernel. The kernels also can be weighted such that the kernels contribute in different proportions to the combined kernel. In a heuristic weighting approach, the weights are chosen in advance by the investigator. This approach assumes that the investigator has some a priori knowledge that suggests that certain types of knowledge should be weighted more heavily in the classification process. An alternative approach to selecting kernel weights is known as multiple kernel learning (MKL). In this approach, the kernel weights are not selected by the investigator, but are chosen computationally. An optimization process for selection of kernel weights identifies the weights that maximize the separation of the positive and negative classes in the combined kernel space.

Multiple Kernel Learning

The support vector machine algorithm operates on a single kernel matrix, representing information from a single data source. However, biological problems are often better characterized by incorporating data from more than one source of data. This has led to an exploration of methods for combining multiple kernel matrices into a single classifier.

Lanckriet discussed using a conic combination of kernel matrices for classification (G. R. G. Lanckriet, De Bie, Cristianini, Jordan, & Noble, 2004). This can

be cast as the convex quadratically constrained quadratic program (QCQP) shown in Equation 2, where D_k is the dimensionality of the k^{th} feature space.

Equation 2. Convex quadratically constrained quadratic program representing a conic combination of kernel matrices (G. R. G. Lanckriet et al., 2004).

$$\begin{aligned}
\min \quad & \frac{1}{2} \left(\sum_{k=1}^K \|\mathbf{w}_k\|_2 \right)^2 + C \sum_{i=1}^N \xi_i \\
\text{w.r.t.} \quad & \mathbf{w}_k \in \mathbb{R}^{D_k}, \xi \in \mathbb{R}^N, \mathbf{b} \in \mathbb{R}, \\
\text{s.t.} \quad & y_i \left(\sum_{k=1}^K \langle \mathbf{w}_k, \Phi_k(\mathbf{x}_i) \rangle + b \right) \geq 1 - \xi_i \\
& \text{and } \xi_i \geq 0, \forall_i = 1, \dots, N
\end{aligned}$$

Bach notes that the QCQP algorithm is feasible only for small problems, incorporating a small number of features and a small number of kernels (Bach, Lanckriet, & Jordan, 2004). Sonnenburg demonstrates that, instead of solving this optimization problem directly, the Lagrangian dual function allows rewriting of the problem as the equivalent dual formulation shown in Equation 3 (Sonnenburg, Ratsch, Schafer, & Scholkopf, 2006):

Equation 3. Equivalent dual formulation of Equation 2 (Sonnenburg et al., 2006).

$$\begin{aligned}
\min \quad & \gamma - \sum_{i=1}^N \alpha_i \\
\text{w.r.t.} \quad & \gamma \in \mathbb{R}, \alpha \in \mathbb{R}^N \\
\text{s.t.} \quad & \mathbf{0} \leq \alpha \leq \mathbf{1}C, \sum_{i=1}^N \alpha_i y_i = 0 \\
& \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{k}_k(\mathbf{x}_i, \mathbf{x}_j) \leq \gamma, \forall_k = 1, \dots, K
\end{aligned}$$

Here α is a vector of dual variables corresponding to each separation constraint.

Rearranging terms, and substituting $\frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{k}_k(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^N \alpha_i = S_k(\alpha)$ to

represent the separation constraints, Sonnenburg rewrites the dual formulation as shown in Equation 4 (Sonnenburg et al., 2006):

Equation 4. Reorganization of Equation 3 by rearrangement of terms (Sonnenburg et al., 2006).

$$\begin{aligned}
& \min && \gamma \\
& \text{w.r.t.} && \gamma \in \mathbb{R}, \alpha \in \mathbb{R}^N \\
& \text{s.t.} && \mathbf{0} \leq \alpha \leq \mathbf{1}C, \sum_{i=1}^N \alpha_i y_i = 0 \\
& && S_k(\alpha) \leq \gamma, \forall_k = 1, \dots, K
\end{aligned}$$

Sonnenburg converts the dual to the semi-infinite linear program (SILP) presenting in Equation 5 (Sonnenburg et al., 2006):

Equation 5. Conversion of Equation 4 to a semi-infinite linear program (Sonnenburg et al., 2006).

$$\begin{aligned}
& \max && \theta \\
& \text{w.r.t.} && \theta \in \mathbb{R}, \beta \in \mathbb{R}^K \\
& \text{s.t.} && \mathbf{0} \leq \beta, \sum_{k=1}^K \beta_k = 1 \text{ and } \sum_{k=1}^K \beta_k S_k(\alpha) \geq \theta \\
& && \text{for all } \alpha \in \mathbb{R}^N \text{ with } \mathbf{0} \leq \alpha \leq C\mathbf{1} \text{ and } \sum_{i=1}^N y_i \alpha_i = 0
\end{aligned}$$

Here the β_k values are kernel weights for the K kernels.

The SILP formulation of the multiple kernel learning problem has lower computational complexity than the QCQP formulation. This version of the optimization problem can be solved using a general-purpose linear programming (LP) solver and a standard support vector machine implementation (Sonnenburg et al., 2006).

Sonnenburg's multiple kernel learning algorithm, depicted in Figure 2, divides the problem into an inner subproblem and an outer subproblem. The outer loop, delimited by Line 2 and Line 9, determines the optimal β for a fixed α , using a general-purpose linear optimizer. The inner loop, Lines 6 through 8, is the dual optimization problem for the single kernel case for fixed β . The outer loop is solved using the results of the inner loop as input, while the inner loop uses the results of the outer loop as input. The process continues until a convergence criterion is met (Sonnenburg et al., 2006).

<p>1 $S^0 = 1, \theta^1 = -\infty, \beta_k^1 = \frac{1}{k}$ for $k = 1, \dots, K$</p> <p>2 for $t = 1, 2, \dots$ do</p> <p>3 Compute $\alpha^t = \operatorname{argmin}_{\alpha \in C} \sum_{k=1}^K \beta_k^t S_k(\alpha)$ by single kernel algorithm with $k = \sum_{k=1}^K \beta_k^t \mathbf{k}_k$</p> <p>4 $S^t = \sum_{k=1}^K \beta_k^t S_k^t$, where $S_k^t = S_k(\alpha^t)$</p> <p>5 if $\left 1 - \frac{S^t}{\theta^t} \right \leq \varepsilon_{MKL}$ then break</p> <p>6 $(\beta^{t+1}, \theta^{t+1}) = \operatorname{argmax} \theta$</p>

7	w.r.t.	$\beta \in \mathbb{R}^K, \theta \in \mathbb{R}$
8	s.t.	$0 \leq \beta, \sum_{k=1}^K \beta_k = 1, \text{ and } \sum_{k=1}^K \beta_k S_k^r \geq \theta \text{ for } r = 1, \dots, t$
9	end for	

Figure 2. Sonnenburg's multiple kernel learning algorithm (Sonnenburg et al., 2006).

Sonnenburg's multiple kernel learning algorithm is implemented in the Shogun machine language toolbox, with interfaces written in Python, Java, C#, Matlab, Octave, R, Lua, and Ruby (Sonnenburg et al., 2010). For this dissertation, the Python interface was used to incorporate the multiple kernel learning algorithm into each of the three gene prioritization problems described.

CHAPTER FOUR: TEXT PROCESSING METHODS FOR MEASURING GENE SIMILARITY

Online literature databases provide rich sources of information about genes stored as free text (Lars Juhl Jensen, Saric, & Bork, 2006; Peng & Zhang, 2007). The texts can be analyzed in different ways to generate measures of similarity between pairs of genes. In this chapter, we will consider two sources of textual information about genes. Both sources are maintained by the National Library of Medicine.

The PubMed database is organized around research articles, such that each record stores information about one article. Each article is given a unique identifier, known as the PubMed identifier, or PMID. Fields in the database include information such as the title of the article, date of publication, author names, keywords assigned to the article, and the actual sentences comprising the free text of the article abstract. In most cases, a PMID can be associated with both a title and an abstract. However, the title field, abstract field, or both title and abstract fields are empty for several PMIDs.

The Entrez Gene database is organized around genes, such that each record stores information about one gene. Each gene is assigned to a single Entrez Gene identifier. One field in the database stores a list of PubMed identifiers related to that gene. This field may list a single PMID, multiple PMIDs, or may be empty.

By combining data in the Entrez Gene and PubMed databases, we can associate a gene with both a list of PubMed identifiers and a list of free-text sentences. We now

describe three methods by which two genes can be compared to each other using data from the Entrez Gene and PubMed databases.

Gene Similarity Based on Shared Abstracts

The simplest method for comparing two genes is to retrieve from Entrez Gene the list of PubMed identifiers associated with each gene. We then simply count the number of PubMed identifiers that appear in both lists. By this method, two genes are assigned a score of zero if they have no PubMed identifiers in common, or an integer score representing the number of PubMed identifiers they share. This method is computationally very straightforward to implement, since it is based entirely on retrieving annotations from a single field of the Entrez Gene database.

Gene Similarity Based on Co-Occurrence of Gene Names in Abstracts

This method of similarity assessment is based on the premise that related genes will likely occur together in the same sentences in research abstracts. Several steps are required for the processing of abstracts. First, the text of the abstract is separated into sentences. Second, each sentence is examined for linguistic cues to find phrases likely to represent the names of genes. Third, since different phrases might be used to represent the same gene, a thesaurus is used to translate the found phrases to the Entrez Gene identifier for the corresponding gene. Finally, for each pair of genes, we count the number of sentences in which the genes occur together.

The identification of textual references to a specific type of object is known as named entity recognition (Leser & Hakenberg, 2005). A simple string search can be used to identify objects that have only a small number of commonly used names. However,

biological molecules often have multiple different names and abbreviations that have arisen through different nomenclature systems. The recognition of gene names in free text is a difficult task because there is no single standardized system of nomenclature for genes. A single gene may be referenced by a large number of noun phrases, abbreviations, and numerical identifiers in different texts (Leser & Hakenberg, 2005; Torii, Hu, Wu, & Liu, 2009). In addition, some genes have been assigned names that are identical to common words in the English language.

The three main approaches to biological named entity recognition are association-rule methods involving application of hand-crafted rules to recognize characteristic phrases (Chang, Schutze, & Altman, 2004; Hanisch, Fluck, Mevissen, & Zimmer, 2003), dictionary lookup methods based on an archived list of entity names (Egorov, Yuryev, & Daraselia, 2004; Koike, Niwa, & Takagi, 2005; Kou, Cohen, & Murphy, 2005), and machine learning methods in which the system is trained on a tagged corpus (L. Smith et al., 2008; Yeh, Morgan, Colosimo, & Hirschman, 2005). Language models are a widely used machine learning approach that involves encoding patterns of word use that are more often used in text discussing genes. A phrase will be marked as a likely reference to a gene only if the word context surrounding the phrase is typically seen in other texts that describe genes. The creation of a language model requires training on a set of documents in which known references to genes have been labeled.

Gene Similarity Based on Cosine Similarity of Free Text in Abstracts

Text processing methods based on analyzing the structure of natural-language text are computationally intensive and difficult to implement, since they require the design of

programs that attempt to model some of the subtleties of the human understanding and interpretation of language (McDonald, Chen, Su, & Marshall, 2004; Saric, Jensen, Ouzounova, Rojas, & Bork, 2006). Several approaches based on the statistical analysis of words in texts have been studied (Jenssen, Laegreid, Komorowski, & Hovig, 2001; Raychaudhuri & Altman, 2003). A vector-space model is a statistical approach that converts documents into vectors of word counts (Berry, Drmac, & Jessup, 1999). In the vector space, the degree of similarity between documents can be calculated as the cosine of the angle between the corresponding vectors.

CHAPTER FIVE: CLASSIFIER DESIGN

Classifier Design: Overview

This dissertation discusses the application of multiple kernel learning to the gene prioritization task in three biological domains: prediction of microRNA-gene relationships, prediction of neurotransmitter-gene relationships, and prediction of drug-gene relationships. These three biological problems will be discussed in more detail in Chapters 5, 6, and 7. For the following discussion, we will use the general term “entity” to refer to a biological object (microRNA, neurotransmitter, or drug) that is being explored for its potential to interact with genes in a gene list.

For the three domains, lists of known entity-gene interactions was identified by querying publicly-available databases: the miRecords microRNA target prediction database for microRNA-gene interactions, the STITCH chemical-protein interactions database for neurotransmitter-gene interactions, and the PharmGKB pharmacogenomics database for drug-gene interactions. In each case, 70% of the known interactions was used for classifier training, with 30% set aside for testing.

The objective of training classifiers on known entity-gene interactions is to then apply the classifiers to previously unknown or unlabeled entity-gene interactions. The classifier results can then be used to rank the unlabeled genes in order of likelihood of an interaction with the entity of interest.

To simplify our analysis, we imposed some limitations on the population of genes that we would attempt to classify. We focused on human genes, and further limited the set of genes we will study to 3217 human genes which appear in all of the data sources used for classifier construction. This helped reduce the problem of trying to combine information sources that vary greatly in their coverage of the entire set of human genes. Throughout this document, we will refer to this list of 3217 genes as our list of core genes.

For each entity of interest, several support vector machine classifiers were created, with each classifier incorporating data from a single data source. We will denote classifiers trained on a single data source as “base classifiers.” We tested multiple sets of parameters to train and test multiple base classifiers for each data source available, for each entity. The best-performing base classifiers, as measured by area under the ROC curve, were then combined to generate multiple kernel learning classifiers. The best-performing MKL classifiers were then used to assign labels to all the genes in the list of 3217 core genes.

Data Sources

The three problem domains differ in the data sources available for classifier construction. The following curated bioinformatics databases were consulted for all three domains:

The Gene Ontology (GO) Database

The Gene Ontology database includes annotations for genes along three dimensions: cellular compartment, molecular function, and biological function. We

focused on the biological function annotations for genes. Each gene may be annotated with multiple biological function terms. For a list of g genes, we construct a $g \times g$ feature matrix in which each matrix cell represents the number of GO biological process annotations shared by gene g_i and gene g_j . In our discussion, we refer to this data source as GO.

The KEGG Pathway Database

The KEGG pathway database includes annotations for the biological pathways in which a gene participates. For a list of g genes, we construct a $g \times g$ feature matrix in which each matrix cell represents the number of KEGG pathway annotations shared by gene g_i and gene g_j . In our discussion, we refer to this data source as KEGG.

The REACTOME Pathway Database

The REACTOME pathway database includes annotations for the biological pathways in which a gene participates. It has some similarities to the KEGG database, but with some differences in coverage. For a list of g genes, we construct a $g \times g$ feature matrix in which each matrix cell represents the number of REACTOME pathway annotations shared by gene g_i and gene g_j . In our discussion, we refer to this data source as REACTOME.

The STRING Protein-Protein Interaction (PPI) Database

Protein-protein interactions were extracted from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), a database providing information about known and predicted interactions among the proteins coded by genes (Lars J. Jensen et al., 2009; Szklarczyk et al., 2011). The database provides a score for protein-protein interactions based on eight criteria: results of high-throughput experiments, co-expression data,

inference based on homology, evolutionary conservation, gene fusion evidence, physical closeness of protein-coding genes on the genome, text mining analysis of PubMed abstracts, and retrieval of interactions from other curated databases. For a list of g genes, we construct a $g \times g$ feature matrix in which each matrix cell represents the score in the STRING database representing the strength of evidence for an interaction between gene g_i and gene g_j . In our discussion, we refer to this data source as PPI.

In addition to the above curated databases, several data sources were constructed through text mining analysis of PubMed abstracts. For each of the 3217 genes in our list of core genes, we queried the Entrez Gene database for the PubMed identifiers associated with that gene. We then retrieved the PubMed abstracts linked to those identifiers. This provided a database of 207,129 PubMed abstracts. These abstracts were processed to generate five different data sources:

PubMed: Shared Identifiers

This data source is based on a simple count of the number of PubMed identifiers shared by genes. For a list of g genes, we construct a $g \times g$ feature matrix in which each matrix cell represents the number of PubMed identifiers shared by gene g_i and gene g_j in the Entrez Gene database. In our discussion, we refer to this data source as PMID.

PubMed: Named Entity Recognition

This data source is constructed using named entity recognition methods. Each abstract was analyzed using a language model based on a hidden Markov model (HMM), trained on the Genetag corpus, and implemented in the Lingpipe suite of statistical

natural language processing tools. The model tags phrases based on the likelihood that a phrase represents a reference to some gene. The HMM cannot determine what specific gene is being referenced by the phrase. The Biothesaurus ontology, maintained by the Protein Information Resource, was used to translate text phrases to the corresponding Entrez Gene identifiers. This process allowed resolution of synonymous phrases to the same Entrez Gene identifier. For any two genes, we can then count the number of abstracts that include references to both genes. For a list of g genes, we construct a $g \times g$ feature matrix in which each matrix cell represents the number of PubMed abstracts that include references to both gene g_i and gene g_j . In our discussion, we refer to this data source as NER.

PubMed: Cosine Similarity of Abstracts

We constructed three data sources, designated as COS1, COS2, and COS3, by calculating the cosine similarity of PubMed abstracts. The data sources differ based on the minimum value used for term frequency-inverse document frequency (tf-idf). This resulted in data sources that treated different sets of words as being uninformative, thus using different-sized vocabularies in the word counts. All abstracts were processed by removal of stopwords, stemming, and removal of whitespace. We then identified the global vocabulary of words used in all the documents. A vector of word counts was created for each abstract. We then calculated the tf-idf value for each term. Any term with a tf-idf value below a set minimum was dropped from the vocabulary. The COS1 data source uses a minimum tf-idf value of 0.1, resulting in a vocabulary size of 295,882 words. COS2 uses a minimum tf-idf value of 0.75, resulting in a vocabulary size of

279,817. COS3 uses a minimum tf-idf value of 1.5, resulting in a vocabulary size of 207,042. At this point, each data source has an associated document-term matrix.

The document-term matrix was used to generate a set of gene-term vectors. If a gene in the Entrez Gene database was associated with only one PubMed identifier, then the corresponding row of the document-term matrix was used as the gene-term vector for that gene. If a gene was associated with multiple PubMed identifiers, then a single gene-term vector was created through a linear combination of vectors, i.e. by adding corresponding term counts in all relevant rows in the document-term matrix.

Once gene-term vectors were generated for all genes, we can calculate the cosine similarity between the vectors. For a list of g genes, we construct a $g \times g$ feature matrix in which each matrix cell represents the cosine similarity between gene g_i and gene g_j .

MicroRNA Target Prediction Algorithms

Three data sources, MIRANDA, PITA, and RNAHYBRID were used only for the microRNA-gene relationship domain. Each data source represents the results from running one microRNA target prediction algorithm for a single microRNA sequence and a list of potential target genes. Details on the design of these data sources are found in Chapter 6.

Risperidone Differential Expression Data

One data source was used only for the drug-gene relationship domain. This data source, which we designate as RISPERIDONE, consists of differential expression values for human neuroblastoma cells exposed to risperidone, compared to control cells not exposed to risperidone. The expression data was generated in the lab of Mas et al. at the

University of Barcelona (Mas, Gasso, Bernardo, & Lafuente, 2013). The expression values were downloaded from the GEO Omnibus database (Barrett & Edgar, 2006). Genes are compared pairwise based on the magnitude of difference between differential expression values.

Data Preprocessing

Once a list of genes is selected to serve either as a labeled list for classifier training or as an unlabeled list to be classified, the data related to each gene is extracted from one or more data sources. The data from each data source is stored in a separate matrix, with each matrix row representing one gene and each matrix column representing one field from the data source. The columns can be conceptualized as features that characterize the genes. Each raw data matrix goes through several processing steps. In cases where the number of features is small, as in the microRNA target prediction scores (2 or 3 features recorded in each data source), the raw training data is normalized to zero mean and unit variance. The scaling parameters used for the normalization are saved, so that the same scaling can be applied to the test data and to the unlabeled data.

In many cases, the number of features is large. This is particularly true when the raw data consists of pairwise counts or comparisons between each gene and all other genes. Each of the 3217 genes in the study population functions as a feature in the data matrix. We felt that it would be useful to reduce the number of features considered, in order to limit the possible effect of noise in the data matrix and focus on features that are informative in distinguishing the genes from one another. We elected to apply principle components analysis on the training data. We retain enough of the principle components

to account for 95% of the variance in the data. These principle components are then used as a reduced set of features for the training data. We save the scaling parameters and loadings matrix used to reduce the training data, and apply the same scaling parameters and loadings matrix to the test data and to the unlabeled data.

SVM Step

For each problem domain, multiple SVM classifiers were trained for each data source. For each data source, we tested three kernel functions: linear, Gaussian, and polynomial. The linear kernel function (same as dot product) produces a single support vector machine. With the Gaussian kernel function, we varied the weight parameter to generate several SVMs. With the polynomial kernel function, we varied the degree parameter to generate several SVMs. Each classifier was trained on a training set of data, and tested on a test set. The area under the ROC curve was used to evaluate the classifiers. We considered the best kernel function for each data source to be the one that generated the SVM with the highest AUROC value.

MKL Step

For each data source, we select the kernel matrix that generated the SVM with the highest AUROC. We only considered data sources that can generate an SVM with an AUROC of at least 0.5, since a lower AUROC suggests that the classifier is performing at a level worse than random classification. We then ran the multiple kernel learning procedure on all combinations of the selected kernels. For each run, we allowed the MKL algorithm to select the optimal weight for combining the kernels. We saved the AUROC values for all the MKL classifiers.

Classifier Diversity

In many cases, several MKL classifiers share the top position for highest AUROC value. To break ties, we favored more diverse classifiers. The concept of classifier diversity is found in the literature on ensemble classifiers. If a classifier uses only one data source, it may have some limitations in situations where coverage of the unlabeled data by that data source has gaps. On the other hand, if a classifier uses many data sources, but assigns a very high weight to one kernel and low weights to the others, this is not much different from using just one kernel. We used the following heuristic for scoring classifiers for diversity. First, we take the average of the weights assigned to each kernel. This tends to penalize classifiers that include many low-weight kernels. We then multiply the average by the number of kernels with weights greater than 0.3. This rewards classifiers that include larger numbers of high-weight kernels.

Measuring Classifier Performance

Once a combined kernel is constructed, it can be incorporated into a support vector machine classifier. This allows multiple information sources to be incorporated into a classification system. The performance of the classifier can be tested using a set of known examples, with the class labels removed. The classifier is run to assign its own labels to the test cases. We can then compare how closely the labels assigned by the classifier match the actual class labels.

Examining the kernel weights derived by a MKL classifier can provide information on the relative usefulness of a data source in the classification process. If the classifier chose very low weights for certain kernels during the weight optimization

process, then those data sources did not contribute much to the ability of the classifier to separate the examples in the positive and negative classes.

For the following studies, overall classifier performance is evaluated by calculating the area under the receiver operating characteristic curve, or ROC curve. This value is defined as the area under a curve plotting sensitivity against 1-specificity, using a sequential series of threshold values as the separator between the positive and negative classes (Fawcett, 2006).

Selecting Best MKL Classifier

To select the classifier that will be used to label the unlabeled data, we first identify the MKL classifiers with the highest AUROC. To break ties among the highest-AUROC classifiers, we select the classifier with the highest diversity score. This classifier can then be used to label the unlabeled data.

Once the classification of unlabeled instances is complete, we save the combined kernel matrix. This structure can be considered a similarity matrix that uses multiple sources of data to compare instances to each other. The rows of this matrix can be used as vectors to cluster the unlabeled genes based on their similarity to each of the other genes.

Clustering Methods

Once data from all available sources is integrated into a combined kernel matrix, each matrix row represents a set of comparisons of that gene to all the other genes in the data set. Applying a classification algorithm allows us to score each gene based on the amount of similarity to genes in the training set. We can then sort the gene list and focus our attention on the genes with the highest similarity scores. However, beyond

identifying the genes that are most similar to genes in the training set, another level of analysis of the gene list is to identify groups of genes that are similar to each other. We collect the highest-scoring genes in the gene list, and then cluster these genes. We expect that genes with similar functions will tend to cluster together.

In our research, we identify groups of related genes by clustering the kernel matrix rows. Many clustering algorithms, such as k-means clustering, require the investigator to begin the analysis by providing the number of clusters to be extracted from the data (Hastie, Tibshirani, & Friedman, 2001). Our approach is to use model-based clustering, in which the optimal number of clusters is determined computationally (Fraley & Raftery, 2002). We use the implementation of model-based clustering developed at the University of Washington and implemented as a package for the R programming language (Fraley, Raftery, Murphy, & Scrucca, 2012).

Functional Enrichment Analysis

Given a list of genes, we can retrieve a list of annotations for each gene from an annotation database such as the Gene Ontology. For our research, we focus on the biological process (BP) annotations in the Gene Ontology. An annotation may be assigned to only one gene, or to many genes. Of the annotations assigned to many genes, some are common annotations that would occur frequently even in a randomly selected set of genes. We would like to identify the biological process annotations that occur in the gene list at a rate higher than would be expected by chance. These annotations are designated as being highly enriched for the gene list. For the task of functional annotation enrichment analysis, we use the Bioconductor package GOstats, which tests the number

of occurrences of an annotation against a hypergeometric distribution to identify annotations that are highly enriched in a gene list (Falcon & Gentleman, 2007).

CHAPTER SIX: PROJECT #1: MICRORNA-GENE INTERACTIONS

This set of studies involves the ranking of genes based on the likelihood of an interaction with a microRNA. The biological question for this project is to identify likely gene targets for microRNAs that are differentially expressed in horn tissue in cattle. For cattle farmers, handling polled cattle (cattle without horns) is safer than handling horned cattle. Several researchers have proposed possible roles for microRNAs in bone and keratin development (He, Eberhart, & Postlethwait, 2009; Itoh, Nozawa, & Akao, 2009; Li et al., 2008; Lin, Kong, Bai, Luan, & Liu, 2009; Price, 2010; Rogler et al., 2009). A better understanding of the gene targets of microRNAs that might be related to horn biogenesis may help to clarify the genomic mechanisms underlying horn development, and would be useful from the perspective of cattle breeding.

Background: Biology of microRNAs and microRNA Target Prediction

MicroRNAs (microRNAs) are short ribonucleotide sequences that are involved in post-transcriptional regulation (Bartel, 2009). A microRNA can alter the expression of a targeted gene indirectly, by binding to the 3' untranslated region (3' UTR) of the messenger RNA (mRNA) transcribed by the gene. The microRNA is incorporated into the RNA induced silencing complex (RISC). If the degree of complementarity of the microRNA to the mRNA is high, the mRNA is cleaved by the RISC. In cases of lower complementarity, translation of the mRNA is repressed. Translational repression is the

more common mode for regulation of gene expression by a microRNA. The identification of large numbers of microRNAs and their presence in a wide variety of cell types has sparked many research efforts to understand the roles that microRNAs play in the life of a cell.

MicroRNAs and Gene Expression

Several factors are believed to be relevant to the propensity of a microRNA to target a gene. These include: [1] the closeness of the alignment of the microRNA sequence to the 3'UTR sequence; [2] the number of sites along the 3'UTR sequence to which the microRNA can be aligned; [3] the calculated thermodynamic stability of the microRNA:mRNA complex; and [4] the degree of evolutionary conservation of the potential binding site within the 3'UTR.

[1] Alignment of microRNA and mRNA Sequences

The higher the degree of complementarity of the microRNA sequence and the 3'UTR sequence, the more likely the microRNA may target the gene. In plants, perfect complementarity is typically observed between the microRNA and target sequences. In animals, perfect complementarity is rare. This complicates the process of microRNA target prediction for animal genomes, since the alignment scoring must allow for some mismatches in the alignment.

[2] Number of Potential Target Sites

While the sequence of the mature microRNA molecule is typically small, around 22 nt, the length of a 3'UTR sequence can vary greatly. A long 3'UTR sequence may include multiple locations to which the same microRNA sequence can be aligned, and also may provide sites to which different microRNAs can align. Several target prediction methods include a count of the number of sites on a gene to which a microRNA could bind as a factor in assessing the likelihood that the microRNA targets the gene. For these methods, a large number of potential binding sites is considered evidence in favor of targeting of the gene by the microRNA.

[3] Thermodynamics of microRNA:mRNA Binding

Several target prediction programs include an estimation of the thermodynamic stability of the microRNA:mRNA duplex, since the microRNA must remain bound long enough to either initiate degradation of the mRNA or repress translation. In addition, some methods consider how accessible the binding site on the mRNA would be to the microRNA, and estimate the amount of free energy that would be required to unfold the mRNA enough to make the binding site accessible.

[4] Evolutionary Conservation of Target Sequences

The conservation across species of a sequence within the 3'UTR may indicate that the sequence has some useful biological function and is being preserved over evolutionary time. The binding of a microRNA to a conserved gene sequence is more likely to reflect a biologically meaningful relationship between the microRNA and the

gene compared to microRNA binding to a sequence that is not retained through evolution. For this reason, many target prediction programs give a higher score to potential targets that have evolutionarily conserved target sequences.

MicroRNA Target Prediction

One possible approach to understanding the functional roles of a microRNA is to examine the functional roles of the genes targeted by the microRNA. However, a microRNA can have multiple targets, and determination of all the targets of a microRNA is a difficult task. Several computational approaches have been developed to use microRNA and mRNA sequence features to predict which mRNAs will be targeted by a microRNA (Enright et al., 2003; Griffiths-Jones, Saini, van Dongen, & Enright, 2008; Maziere & Enright, 2007). Databases such as miRBase have been developed to archive sequence information about microRNAs, and to store lists of computationally predicted microRNA targets (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006). Verification of the accuracy of computational target prediction approaches requires the availability of databases of experimentally verified microRNA targets, such as the TarBase database (Sethupathy, Corda, & Hatzigeorgiou, 2006); (Papadopoulos, Reczko, Simossis, Sethupathy, & Hatzigeorgiou, 2009).

Computational microRNA Target Prediction Methods

Multiple computational methods have been developed for predicting whether a particular microRNA will target a gene. To date, no one method has been demonstrated to consistently outperform all others on the target prediction task. The relative importance of the factors influencing gene targeting by a microRNA is not clear.

Therefore, different prediction methods may incorporate different factors into the target prediction task, and may weight these factors differently. This section will discuss several well-documented computational microRNA target prediction methods.

miRanda

The miRanda algorithm (Enright et al., 2003) uses a dynamic programming approach to compute a weighted sum of scores for base pair matches and mismatches. Matches in positions 2-8 of the 5' end of the miRNA (the seed region) and in the 3' region of the miRNA are given greater weight. The free energy of the miRNA-mRNA duplex is estimated using the Vienna RNA folding package (Hofacker, 2003). The PhastCons conservation score is used as a measure of the degree of conservation of the genome sequence to which the miRNA is aligned (Siepel et al., 2005). Thus, the miRanda algorithm uses a combination of sequence alignment, secondary structure prediction, and sequence conservation information to generate a score for the target.

PITA

PITA (Probability of Interaction by Target Accessibility) takes into account the accessibility of potential target sites when calculating the likelihood that a microRNA will form a complex with that site (Kertesz, Iovino, Unnerstall, Gaul, & Segal, 2007). The algorithm first uses sequence alignment to identify possible target sites. It then calculates the free energy that would be required to unfold the target mRNA enough to allow binding between the microRNA and the mRNA. Evolutionary conservation of the target sequence is not considered by the PITA algorithm.

RNAhybrid

The RNAhybrid algorithm is a variation of Zuker's algorithm for RNA secondary structure prediction (Zuker & Stiegler, 1981). Instead of predicting the secondary structure for a single RNA sequence, RNAhybrid extends the algorithm to determine the most energetically favorable hybridization between two RNA sequences. RNAhybrid uses a dynamic programming method to calculate the minimum free energy of hybridization for all possible start positions in the microRNA and the target gene sequence, with some allowance for the possibility of stretches of unpaired nucleotides in either sequence. The RNAhybrid algorithm has been incorporated into a program for microRNA target prediction, and is available freely online (Kruger & Rehmsmeier, 2006).

DIANA-microT

The DIANA-microT algorithm considers both sequence alignment and evolutionary conservation in scoring potential microRNA targets (Maragkakis et al., 2009). A weighted sum of scores for each potential target site on the 3'UTR is calculated to produce a total score for the target gene.

PicTar

PicTar (Probabilistic Identification of Combinations of Target Sites) uses sequence alignment, free energy of the miRNA-mRNA complex, and conservation of the target sequence to score the likelihood that a microRNA targets a gene (Krek et al., 2005). In addition, PicTar assigns a higher score to 3'UTR sequences that can be aligned simultaneously to multiple microRNAs. This is based on the notion that microRNAs may sometimes act cooperatively to regulate target genes.

TargetScan

TargetScan and TargetScanS are related programs that score potential microRNA targets primarily on base-pairing criteria. TargetScanS also considers primary sequence features, based on the observation of an overrepresentation of adenosines flanking the mRNA sequences complementary to the mature microRNA sequence (Lewis, Burge, & Bartel, 2005).

mirTarget2

The mirTarget2 system is different from the other target prediction methods discussed here in that it uses mRNA expression data, rather than sequence data, to predict microRNA targets (X. Wang & El Naqa, 2008). A support vector machine approach is used to train the system to recognize patterns of gene downregulation in microarray data that are correlated with microRNA targeting. This approach can potentially identify microRNA targeting when the microRNA achieves gene regulation through cleavage of the mRNA. A potential limitation of this approach is that microRNAs more commonly affect gene regulation by translational repression than by degradation of the mRNA. Measurement of changes in levels of protein synthesis, rather than mRNA expression, would provide a better indication of microRNA targeting when repression is the mechanism used for modulation of gene expression (Selbach et al., 2008).

Methods

MicroRNAs were selected for this study based on review of data generated at the USDA Bovine Functional Genomics Laboratory, consisting of microRNA expression counts for horn and poll tissues. The microRNAs of interest for this study include 14 microRNAs: four known from previous research to be involved in bone development

(miR-21, miR-214, miR-133, and miR-135), and ten microRNAs that are overexpressed in the USDA data set but whose functions are not as well characterized (miR-106, miR-145, miR-193, miR-195, miR-22, miR-29, miR-423, miR-497, miR-660, and miR-93).

To create support vector machine classifiers to rank genes as interaction partners with microRNAs, a set of known interactions with human microRNA hsa-miR-1 was used to provide positive and negative examples. We chose hsa-miR-1 because the relatively long history of research on this microRNA has generated a larger number of both positive and negative training examples than is available for other microRNAs.

The plan for the study was to combine data from multiple sources to rank our 3217 core genes as potential interaction partners for the fourteen microRNAs selected for analysis. The data sources used included data from archival bioinformatics databases, free-text abstracts, and the results of running microRNA target prediction algorithms. The creation of feature matrices based on archival databases and free-text abstracts was described in Chapter Five. Here we focus on describing the construction of feature matrices based on microRNA target prediction algorithms.

Kernel Matrices for Classifier Training and Testing

For classifier training and testing, the miRanda, PITA, and RNAhybrid algorithms were run using the mature microRNA sequence for miR-1 and human 3'UTR sequences. A list of known gene targets for miR-1 was retrieved from the TarBase database. This database lists both experimentally demonstrated targets (positive examples) and genes that have been experimentally demonstrated not to be targeted by miR-1 (negative examples). From the list of known positive and negative examples, we selected those

genes that were also represented in the archival information sources. This would allow creation of combined kernel matrices that included data from all the data sources. The final lists of positive and negative examples were divided into a training set and test set.

We ran the miRanda, PITA, and RNAhybrid algorithms using the miR-1 mature sequence and the 3'UTR sequences of the genes in the training and test sets. While many microRNA target prediction programs have been developed, we selected three programs because they are easily accessible online as downloadable software packages for use on a local computer, and they have been widely used in other research on microRNA target prediction. The human 3'UTR sequences were downloaded from the Ensembl database (Flicek et al., 2012). The database includes some very short 3'UTR sequences, some as short as one or two nucleotides. Such short 3'UTR sequences caused program crashes when submitted to computational target prediction programs to attempt alignment with mature microRNA sequences. We eliminated from consideration 3'UTR sequences shorter than 25 nt. The miR-1 mature microRNA sequence was retrieved from miRBase (Griffiths-Jones et al., 2008).

For PITA and RNAhybrid, the feature matrices included two features: the number of hits, and the free energy of hybridization. For miRanda, the feature matrices included three features: the number of hits, the free energy of hybridization, and a composite score that gives points for sequence alignment, evolutionary conservation of sequence, and free energy of hybridization. Each of the feature matrices was converted to a kernel matrix, as described in Chapter 5. The kernel matrix that generated the best-performing classifier was selected for incorporation into a set of multiple kernel learning classifiers.

Kernel Matrices for Labeling Unlabeled Genes

After training and testing the MKL classifier, the next step was to use the classifier to rank each of the 3217 core genes as a possible target for each of the fourteen microRNAs identified as differentially expressed in bovine horn tissue. The mature sequences for the microRNAs were downloaded from miRBase. The bovine 3'UTR sequences were downloaded from the UCSC Genome Browser (Fujita et al., 2011). For each of the microRNAs, we ran the miRanda, PITA, and RNAhybrid algorithms against the 3'UTR sequences for the core genes. The results were assembled into feature matrices. The feature matrices were scaled using the same scaling parameters as the feature matrices used for classifier training and testing. The feature matrices were converted to kernel matrices, using the same kernel functions as were used in the best-performing base classifiers identified during classifier testing. We then used the best-performing MKL classifier to assign a numeric score to each of the core genes.

Results

Micro-RNA Gene Interactions: Classifier Performance

Table 2 shows the performance, as measured by area under the ROC curve, for base classifiers trained to recognize genes as potential interaction partners for miRNA-1. We note that the best-performing base classifier was an SVM trained on the cos1 free-text data, using a linear kernel function, with a resulting AUROC of 0.7321.

Table 2. Best mir1-gene base classifiers.

Abbreviations: linear = linear kernel function; poly = polynomial kernel function; deg = degree parameter for polynomial kernel function; w = weight parameter for Gaussian kernel function.

Data Source	Kernel Function	AUROC
-------------	-----------------	-------

cos1	linear	0.7321
ppi	linear	0.6786
pmid	linear	0.6786
kegg	poly, deg=10	0.6250
cos2	linear	0.6250
reactome	poly, deg=2	0.5893
pita	gaussian, w= 0.1	0.5714
rnahybrid	gaussian, w=0.1	0.5714
go	poly, deg=3	0.5714
cos3	linear	0.5714
ner	linear	0.5357
miranda	gaussian, w=0.0	0.5000

Among the classifier ensembles, several combined classifiers achieved an AUROC score of 0.75. This demonstrates that combining data sources can result in ensemble classifiers that perform better than any of the base classifiers.

There are 225 combined kernels that achieved an AUROC score of 0.75. In Table 3 we show just the classifiers with diversity scores of at least 1.3.

Table 3. Ten highest-diversity mir1-gene ensemble classifiers out of 225 classifiers with AUROC of 0.75.

Kernels	Weights	AUROC	Diversity
ppi, ner, cos1, cos2	0.8002, 0.3059, 0.3699, 0.3595	0.75	1.8356
ppi, ner, cos1	0.8372, 0.3176, 0.4452	0.75	1.6000
ppi, cos1, cos2	0.8495, 0.3774, 0.3687	0.75	1.5956

ppi, ner, cos2	0.8474, 0.3166, 0.4263	0.75	1.5902
ppi, ner, pmid	0.8664, 0.3024, 0.3974	0.75	1.5662
kegg, ppi, ner, cos1, cos2	0.0007, 0.8003, 0.3059, 0.3699, 0.3594	0.75	1.4689
pita, ppi, ner, cos1, cos2	0.0006, 0.8003, 0.3059, 0.3699, 0.3594	0.75	1.4689
rnahybrid, ppi, ner, cos1, cos2	0.0006, 0.8003, 0.3059, 0.3699, 0.3594	0.75	1.4688
go, ppi, ner, cos1, cos2	0.0006, 0.8003, 0.3059, 0.3699, 0.3594	0.75	1.4688
reactome, ppi, ner, cos1, cos2	0.0006, 0.8003, 0.3059, 0.3699, 0.3594	0.75	1.4688
ppi, cos1	0.8901, 0.4558	0.75	1.3459
ppi, cos2	0.8982, 0.4396	0.75	1.3378

Among the classifier ensembles that achieved an AUROC score of 0.75, the ensemble with the highest diversity score was based on data sources ‘ppi’ (protein-protein interaction data), ‘ner’ (named entity recognition in free text), ‘cos1’ (similarity of free-text abstracts using tf-idf of 0.1), and ‘cos2’ (similarity of free-text abstracts using tf-idf of 0.75). This classifier was applied for classifying and ranking the set of unclassified genes.

Micro-RNA Gene Interactions: Biological Plausibility

Many of the microRNAs share targets. For each of the 14, we looked at the 35 highest-ranked targets, and then looked at targets shared among these lists. This gave 16 shared targets among the most highly-ranked targets. We have separated these genes into

three categories: regulation of transcription and DNA replication; lipid and glycolipid metabolism; and other functions.

Group 1: Regulation of Transcription and DNA Replication

The genes in this group are listed in Table 4. This set of interaction partners is biologically plausible because it reflects a known general role for microRNAs in the regulation of transcription.

Table 4. Group 1: microRNA interaction partners related to regulation of transcription and DNA replication. Gene descriptions are summarized from the GeneCards database (Safran et al., 2010).

entrez	symbol	name	description
10274	STAG1	stromal antigen 1	member of SCC3 family; expressed in nucleus; encodes a component of cohesion involved in sister chromatid cohesion during DNA replication
54464	XRN1	5'-3' exoribonuclease 1	involved in mRNA degradation, meiosis, telomere maintenance, microtubule assembly; may act as a tumor suppressor protein in osteogenic sarcoma
5984	RFC4	replication factor C (activator 1) 4, 37kDa	accessory proteins required for elongation of primed DNA templates by DNA polymerase delta and DNA polymerase epsilon
7029	TFDP2	transcription factor Dp2; E2F dimerization partner 2	member of transcription factor DP family; involved in transcriptional activation of cell cycle regulated genes; involved in both cell proliferation and apoptosis

Group 2: Glycolipid and Lipid Metabolism

The genes in this group are listed in Table 5. Several studies indicate a role for microRNAs in glucose and lipid metabolism (Vickers, Sethupathy, Baran-Gale, & Remaley, 2013). Much research attention has been given to miR-21, miR-33, miR-122, miR-125, miR-370, and miR-758. These microRNAs have been suggested as possible biomarkers for disease progression and response to therapy in dyslipidemias, and as potential therapeutic targets (Fernández-Hernando, Suárez, Rayner, & Moore, 2011; Flowers, Froelicher, & Aouizerat, 2013 Poy, 2007 #1397). One of these, miR-21, is included in our analysis. The gene ABCA1 codes for an ATP binding cassette protein which is important in humans in transport of lipids, regulation of cholesterol in peripheral cells, and etiologically implicated in the development of atherosclerosis (Brunham, Singaraja, & Hayden, 2006). One gene suggested as a possible interaction partners for our list of microRNAs is ABCG1, a gene with a protein product that is also in the ATP binding cassette class of proteins. Thus, proposing a set of interaction partners involved in lipid metabolism appears to have some biological plausibility.

Table 5. Group 2: microRNA interaction partners related to glycolipid and lipid metabolism. Gene descriptions are summarized from the GeneCards database (Safran et al., 2010).

entrez	symbol	name	description
56894	AGPAT3	1-acylglycerol-3-phosphate O-acyltransferase 3	converts lysophosphatidic acid into phosphatidic acid; second step in de novo phospholipid biosynthetic pathway
10402	ST3GAL6	ST3 beta-galactoside alpha-2,3-sialyltransferase 6	member of sialyltransferase family; enzymes that transfer sialic acid to sialylated glycolipids or glycoproteins
30849	PIK3R4	phosphoinositide-3-kinase, regulatory	PIK3s are lipid kinases involved in several cell functions: proliferation,

		subunit 4	cell survival, degranulation, vesicular trafficking, cell migration
9619	ABCG1	ATP-binding cassette, sub-family G (White), member 1	ABC proteins transport molecules across extra- and intra-cellular membranes. This protein is a member of the White subfamily; involved in macrophage cholesterol and phospholipid transport; may regulate cellular lipid homeostasis
4047	LSS	lanosterol synthase; 2,3-oxidosqualene-lanosterol cyclase	member of terpene cyclase/mutase family; catalyzes first step in biosynthesis of cholesterol, steroid hormones, and vitamin D

Group 3: Other Functions

The genes in this group are listed in Table 6. Of the genes in this group, two are of particular interest: CLDN1, which codes for a claudin, and TNFSF10, which codes for a tumor necrosis factor.

Several claudins has been found to be regulated by microRNAs. Proteins coded by the gene claudin-14 are involved in calcium reabsorption in the kidney, This gene is suppressed by two microRNAs, miR-9 and miR-374 (Gong et al., 2012). In the retinal epithelium, miR-204 and miR-211 induce the expression of claudins 10, 16, and 19, which appear to be important in maintenance of an intact epithelium (F. E. Wang et al., 2010). Thus, a gene coding for a claudin protein is a biologically plausible microRNA interaction partner.

Several microRNAs are now known to interact with genes coding for proteins in the tumor necrosis factor class. MicroRNA-23a mediates the regulation of osteoblast apoptosis (Dong, Cui, Jiang, & Sun, 2013) and endothelial cell injury (Ruan, Xu, Li,

Yuan, & Dai, 2012) by TNF-alpha. TNF-alpha induces the expression of miR-18a in rheumatoid arthritis synovial fibroblasts, contributing to cartilage destruction and chronic joint inflammation (Trenkmann et al., 2013). In acute liver failure, miR-1187 mediates apoptosis of hepatocytes by TNF-alpha (Yu et al., 2012). In head and neck squamous cell carcinoma, miR-375 modulates apoptosis induced by TNF-alpha (J. Wang et al., 2013). The large number of studies indicating interactions between microRNAs and tumor necrosis factor support the biological plausibility of the relationship proposed in our analysis.

Table 6. Group 3: microRNA interaction partners with other functions.
Gene descriptions are summarized from the GeneCards database (Safran et al., 2010).

entrez	symbol	name	description
9076	CLDN1	claudin 1	claudin is an integral membrane protein; component of tight junction strands
26061	HACL1	2-hydroxyacyl-coA lyase 1	catalyzes a carbon-carbon cleavage reaction
6747	SSR3	signal sequence receptor, gamma (translocon-associated protein gamma)	glycosylated endoplasmic reticulum membrane receptor; associated with protein translocation across the ER membrane
8743	TNFSF10	tumor necrosis factor (ligand) superfamily, member 10	a cytokine in the tumor necrosis factor ligand family; induces apoptosis in transformed and tumor cells; does not appear to kill normal cells

MicroRNA-Gene Interactions: Functional Annotation Analysis

Table 7 shows the most highly enriched functions for the highest-ranked interaction partners shared by the fourteen microRNAs studied. While the microRNAs included in this analysis are differentially expressed in horn tissue, the functional enrichment analysis for the shared gene targets does not reveal functions that are specific to horn development. However, we note a preponderance of functions related to signaling and responses to various biochemical entities. This is consistent with research findings indicating roles for microRNAs in modulating a wide range of biochemical processes (Ameres & Zamore, 2013; H. Dong et al., 2013; Van Wynsberghe, Chan, Slack, & Pasquinelli, 2011).

Table 7. Most highly enriched functions for the highest-ranked interaction partners shared by the 14 microRNA studied.

goid	pval	description
GO:0044281	3.17E-05	small molecule metabolic process
GO:0007195	9.46E-05	adenylate cyclase-inhibiting dopamine receptor signaling pathway
GO:0006112	1.70E-04	energy reserve metabolic process
GO:0032870	1.80E-04	cellular response to hormone stimulus
GO:0071375	2.11E-04	cellular response to peptide hormone stimulus
GO:1901653	2.11E-04	cellular response to peptide
GO:0050852	2.42E-04	T cell receptor signaling pathway
GO:0044030	2.63E-04	regulation of DNA methylation
GO:0071495	3.38E-04	cellular response to endogenous stimulus

GO:0009719	3.54E-04	response to endogenous stimulus
GO:0007191	4.22E-04	adenylate cyclase-activating dopamine receptor signaling pathway
GO:0007166	4.27E-04	cell surface receptor signaling pathway
GO:0044710	4.59E-04	single-organism metabolic process
GO:0050851	5.74E-04	antigen receptor-mediated signaling pathway
GO:0002429	7.48E-04	immune response-activating cell surface receptor signaling pathway
GO:0043434	7.53E-04	response to peptide hormone stimulus
GO:0009725	8.23E-04	response to hormone stimulus
GO:1901652	8.58E-04	response to peptide
GO:0048015	8.83E-04	phosphatidylinositol-mediated signaling
GO:0048017	8.83E-04	inositol lipid-mediated signaling
GO:0071417	8.96E-04	cellular response to organic nitrogen

MicroRNA-Gene Interactions: Using Target Rankings to Identify Genes with Possible Relationship to Horn Development

In bovine genomics, an animal born without horns is denoted as possessing the polled phenotype. Identification of genes related to the polled phenotype may be helpful in planning the selective breeding of animals in order to favor generation of offspring that have no horns. Previous research on bovine horn development has identified a region between the 0.8MB and 2.8MB positions on Chromosome 1 (Chr1) that has been implicated in the polled phenotype (Allais-Bonnet et al., 2013). We examined our data on

predicted microRNA targets to see whether the microRNAs that are differentially expressed in horn tissue target any of the genes located in this region.

For each of the differentially expressed microRNAs, we retrieved the Chr1 coordinates for each predicted target, and the score assigned by our ensemble classifier.

The results are shown in Table 8.

Table 8. Predicted microRNA targets in region 0.8MB to 2.8MB on Chromosome 1.

mir	entrez	symbol	score	start	end	strand
mir21	3455	IFNAR2	0.518454	1593290	1627127	-
mir21	3460	IFNGR2	0.624291	1376156	1408234	-
mir214	3460	IFNGR2	0.624291	1376156	1408234	-
mir133	3460	IFNGR2	0.624291	1376156	1408234	-
mir133	3455	IFNAR2	0.518454	1593290	1627127	-
mir135	3460	IFNGR2	0.624291	1376156	1408234	-
mir106	3460	IFNGR2	0.624291	1376156	1408234	-
mir106	3455	IFNAR2	0.518454	1593290	1627127	-
mir145	539	ATP5O	0.541558	922635	929993	+
mir145	3460	IFNGR2	0.624291	1376156	1408234	-
mir193	3460	IFNGR2	0.624291	1376156	1408234	-
mir193	3455	IFNAR2	0.518454	1593290	1627127	-
mir195	3455	IFNAR2	0.518454	1593290	1627127	-
mir22	3460	IFNGR2	0.624291	1376156	1408234	-

mir22	3455	IFNAR2	0.518454	1593290	1627127	-
mir29	3460	IFNGR2	0.624291	1376156	1408234	-
mir423	3460	IFNGR2	0.624291	1376156	1408234	-
mir423	3455	IFNAR2	0.518454	1593290	1627127	-
mir497	3455	IFNAR2	0.518454	1593290	1627127	-
mir660	3455	IFNAR2	0.518454	1593290	1627127	-
mir660	3460	IFNGR2	0.624291	1376156	1408234	-
mir93	3460	IFNGR2	0.624291	1376156	1408234	-

We note that all of the predicted microRNA targets in the specified Chr1 region have been assigned positive scores by the ensemble classifier. Three genes in the specified Chr1 region, ATP50, IFNAR2, and IFNGR2, are predicted as targets for one or more of the differentially expressed microRNAs. IFNGR2 is a predicted target for twelve of the fourteen microRNAs, IFNAR2 is a predicted target for nine microRNAs, and ATP50 is a predicted target for one microRNA. These three genes are described in Table 9.

Table 9. Genes on Chr1, 0.8MB to 2.8MB, predicted as microRNA targets.

entrez	symbol	name	description
539	ATP5O	ATP synthase, H+ transporting, mitochondrialF1 complex, O subunit	a component of the F-type ATPase found in the mitochondrial matrix; part of connector linking the catalytic core and and membrane proton channel of the ATPase
3455	IFNAR2	interferon alpha, beta, and omega receptor 2	associates with IFNAR1 to form the type I interferon receptor; involved in signal transduction through interaction with Janus protein kinases
3460	IFNGR2	interferon gamma receptor 2	associates with IFNGR1 to form the gamma interferon receptor; involved in signal transduction through interaction with Janus protein kinases

Each of the genes listed in Table 9 has been suggested in previous research literature as a possible marker for the polled phenotype in cattle. A recent study shows a high correlation between the polled phenotype and a single nucleotide polymorphism in IFGNR2, based on sequencing and association studies (Glatzer et al., 2013). This is the gene in the region of interest predicted to be targeted by the largest number of differentially expressed microRNAs.

The differential expression in horn tissue of multiple microRNAs that target these genes is consistent with a possible role for microRNAs in regulating the expression of genes that may determine whether offspring display the horn phenotype or the polled phenotype. The targeting of two interferon receptors by many of the microRNAs suggests

that signal transduction through the Janus protein kinases may have a significant role in the expression of the horned vs. polled phenotype.

This study provides further evidence that the MKL ensemble classifier method can help to identify biologically meaningful relationships between microRNAs and genes. For several of the microRNAs, the classifier assigned positive labels to genes on Chromosome 1 that have been previously implicated in determining the polled phenotype. These are biologically plausible relationships which could be explored further experimentally through linkage disequilibrium evaluation.

CHAPTER SEVEN: PROJECT #2: NEUROTRANSMITTER-GENE INTERACTIONS

This set of studies involves the ranking of genes based on the likelihood of an interaction with a chemical entity. The two chemicals examined are the neurotransmitters serotonin and melatonin. The STITCH database provides information about known interactions between chemicals and proteins. A wide range of chemical entities is covered by the database, including neurotransmitters. Here we use interactions with proteins as a proxy for interactions with genes coding for the proteins.

Background: Melatonin and Serotonin

For this set of experiments, we focus on identifying genes that interact with the molecules melatonin and serotonin. This is a useful problem to explore because melatonin and serotonin are part of the same biosynthetic pathway, but have different biological functions. Serotonin is a precursor to melatonin. Serotonin functions primarily as a neurotransmitter, modulating interactions between neurons at the level of the synapse. Serotonin also has some hormonal capabilities in that it modulates functions in the hypothalamic-pituitary-adrenal axis, though the mechanisms behind this activity are unclear. Melatonin functions as a hormone, regulator of circadian rhythms, and modulator of apoptosis. Ongoing biological research is uncovering the mechanisms by which each of these functions is controlled by interactions among genes. If our information extraction method retrieves valid relationships between a biological molecule

and the genes with which it interacts, then we should be able to recover known functions of the molecule by analysis of the list of genes predicted to interact with the molecule. In our case, once we have extracted a gene list of proposed interaction partners for melatonin and a separate gene list of interaction partners for serotonin, we should be able to verify that the gene lists are different and represent different sets of biological functions. In addition, the ranking of genes as potential interaction partners for melatonin and serotonin may help to uncover possible mechanisms of action that previously have not been recognized.

Methods

For this study, we only used kernel matrices based on data from the archival bioinformatics data sources. For classifier training, we retrieved information from the STITCH chemical database on genes known to interact with melatonin and serotonin. We used the STITCH score assigned by the database curators and stored in the database for each neurotransmitter-gene pair to select positive and negative examples for training. The score represents a level of confidence that a chemical and a gene in the database interact, based on multiple information sources, including experimental data, mining of free-text abstracts, and information retrieved from other curated databases. We selected neurotransmitter-gene pairs with a score above 0.85 to serve as positive examples, and neurotransmitter-gene pairs with a score below 0.5 to serve as negative examples. The preparation of data from our set of archival data sources for representation as kernel matrices is described in Chapter 5.

Results

Serotonin-Gene Interactions: Classifier Performance

Table 10 shows the performance, as measured by area under the ROC curve, for base classifiers trained to recognize genes as potential interaction partners for serotonin. We note that the best-performing base classifier was an SVM trained on the ppi data, using a linear kernel function, with a resulting AUROC of 0.9666.

Table 10. Best serotonin-gene base classifiers.

Abbreviations: linear = linear kernel function; poly = polynomial kernel function; deg = degree parameter for polynomial kernel function; w = weight parameter for Gaussian kernel function.

Data Source	Kernel Function	AUROC
ppi	linear	0.9666
reactome	poly, deg=10	0.9091
go	linear	0.8988
cos2	linear	0.8977
cos1	linear	0.8896
pmid	linear	0.8724
cos3	linear	0.8643
kegg	linear	0.8287
ner	linear	0.7561

Among the classifier ensembles, several combined classifiers achieved an AUROC score of 0.9839. This demonstrates that combining data sources can result in classifiers that perform better than any of the base classifiers. The best ensemble classifiers are described in Table 11.

Table 11. Serotonin-gene ensemble classifiers with AUROC score of 0.9839. This AUROC score is higher than that of any of the base classifiers. The ensemble classifiers are presented in order of decreasing diversity.

Kernels	Weights	AUROC	Diversity
ppi, ner	0.9513, 0.3082	0.9839	1.2595
reactome, ppi, ner	0.0014, 0.9513, 0.3082	0.9839	0.8406
kegg, ppi	0.2722, 0.9622	0.9839	0.6172
kegg, ppi, ner	0.2520, 0.9335, 0.2552	0.9839	0.4802
kegg, reactome, ppi	0.2722, 0.0015, 0.9622	0.9839	0.4120
kegg, reactome, ppi, ner	0.2520, 0.0013, 0.9335, 0.2551	0.9839	0.3605

Among the classifier ensembles that achieved an AUROC score of 0.9839, the ensemble with the highest diversity score was based on data sources 'ppi' (protein-protein interaction data) and 'ner' (named entity recognition in free text). This classifier was selected for classifying and ranking the set of unclassified genes.

Serotonin-Gene Interactions: Biological Plausibility of Highly Ranked Unknown Interaction Partners

The ppi-ner classifier was run to assign a numeric score to each of the unlabeled genes in our core gene set. The genes were then ranked based on their scores. The ten highest-ranked genes are listed in Table 12.

Table 12. Previously unlabeled genes in the core gene set that were ranked highest as potential interaction partners for serotonin. The ppi-ner classifier was used to rank the genes.

entrez	symbol	name
5565	PRKAB2	5'-AMP-activated protein kinase, beta-2 subunit

245972	ATP6V0D2	ATPase, H ⁺ transporting, lysosomal 38kDa, V0 subunit D, isoform 2
3985	LIMK2	LIM domain kinase 2
7534	YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide
6093	ROCK1	Rho-associated coiled-coil containing protein kinase 1
54106	TLR9	toll-like receptor 9
635	BHMT	betaine-homocysteine S-methyltransferase
8890	EIF2B4	eukaryotic translation initiation factor 2B, subunit 4 delta
50628	GEMIN4	GEM (nuclear organelle) associated protein 4
226	ALDOA	fructose-1,6-bisphosphate aldolase A

While the neurotransmitter serotonin has a well-known role as a neurotransmitter, a review of the highly-ranked unclassified genes suggests a possible role for serotonin in muscle metabolism. The gene PRKAB2 codes for a regulatory subunit of the AMP-activated protein kinase, which is highly expressed in skeletal muscle (Lee-Young et al., 2009; Sanchez et al., 2012). The proteins coded by LIMK2 and ROCK1 participate in a phosphorylation pathway that contributes to the reorganization of the actin cytoskeleton (Dai et al., 2006; Heng et al., 2012). While actin is a significant component of all eukaryotic cells, it is particularly abundant in muscle cells (Gerthoffer, 2005; Kee, Gunning, & Hardeman, 2009; Koubassova & Tsaturyan, 2011; Lehman & Morgan, 2012; J. Wang, Zohar, & McCulloch, 2006). The gene ALDOA codes for the protein aldolase-A, which is highly expressed in skeletal muscle. Deficiency in aldolase-A is associated with hemolytic anemia and myopathy (Dai et al., 2006; Heng et al., 2012).

The possibility of a role for serotonin in muscle metabolism is supported by recent studies in which the expression of tryptophan hydroxylase 1 (TPH1), the enzyme that catalyzes the rate limiting step in serotonin synthesis, was found to be increased in skeletal muscle in mice (Chandran et al., 2012). The article describing this research was not included in the set of articles used to construct the classifiers, since it did not have a PubMed identifier assigned at the time that texts were retrieved from PubMed to build the classifiers. Thus, a proposed relationship between serotonin and muscle, suggested by the high ranking of muscle-related genes by the ensemble classifier, turns out to be a biologically plausible relationship, based on recent experimental studies.

Serotonin: Clustering and Functional Enrichment Analysis

In reviewing the functional enrichment of the gene clusters, it might be reasonable to expect that the largest clusters would be enriched for functions related to chemical response and signaling. Since genes typically act in networks rather than alone, a cluster consisting of a single gene may represent an outlier that the clustering algorithm could not relate to any other genes. Medium-sized clusters may represent less well characterized but potentially important functions. Table 13 depicts lists of highly enriched functional annotations for clusters of genes ranked highly as potential interaction partners for serotonin. In this analysis, we note several large clusters functionally enriched for terms related to signaling and chemical response. One cluster of 69 genes is enriched for terms related to wound healing, coagulation, and stress response. This is consistent with a role for platelet-derived serotonin in wound healing. Thus, the

functional annotation analysis of predicted gene targets is able to reveal different categories of function for the chemical entity of interest.

Table 13. Functional annotations for clusters of genes ranked highly as interaction partners for serotonin. Model-based clustering, using the mclust R package, was applied to the rows of the kernel matrix on which the combined classifier is based, with each matrix row representing one gene. The annotation lists were generated using the GOstats R package, with the list of entrez IDs in one cluster serving as the input to GOstats.

cluster	# of genes	functions
clust_4	291	G-protein coupled receptor signaling pathway; cellular response to stimulus; cell surface receptor signaling pathway; response to stimulus; G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger; signal transduction; signaling; single organism signaling; response to chemical stimulus; cell communication
clust_3	256	response to chemical stimulus; cell surface receptor signaling pathway; response to stimulus; cellular response to stimulus; G-protein coupled receptor signaling pathway; small molecule metabolic process; cellular response to chemical stimulus; regulation of biological quality; cellular calcium ion homeostasis; phospholipase C-activating G-protein coupled receptor signaling pathway
clust_5	192	G-protein coupled receptor signaling pathway; G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger; adenylate cyclase-modulating G-protein coupled receptor signaling pathway; system process; response to stimulus; cell surface receptor signaling pathway; adenylate cyclase-inhibiting G-protein coupled receptor signaling pathway; cell-cell signaling; adenylate cyclase-activating G-protein coupled receptor signaling pathway; response to chemical stimulus
clust_2	172	response to chemical stimulus; cell surface receptor signaling pathway; cellular response to chemical stimulus; cellular response to stimulus; response to stimulus; response to stress; signaling; single organism signaling; signal transduction; cell communication

clust_1	69	response to wounding; response to chemical stimulus; response to stress; wound healing; blood coagulation; coagulation; hemostasis; response to external stimulus; regulation of body fluid levels; cellular component movement
clust_11	10	negative regulation of hydrolase activity; basophil chemotaxis; negative regulation of cysteine-type endopeptidase activity involved in apoptotic process; negative regulation of cysteine-type endopeptidase activity
clust_10	2	regulation of vitamin D biosynthetic process; regulation of vitamin metabolic process; vitamin D biosynthetic process; regulation of calcidiol 1-monooxygenase activity; fat-soluble vitamin biosynthetic process; vitamin D metabolic process; vitamin biosynthetic process; fat-soluble vitamin metabolic process; regulation of lipid storage; regulation of steroid biosynthetic process
clust_9	1	coenzyme A biosynthetic process; pantothenate metabolic process; nucleoside bisphosphate biosynthetic process; ribonucleoside bisphosphate biosynthetic process; purine nucleoside bisphosphate biosynthetic process
clust_14	1	glycogen cell differentiation involved in embryonic placenta development; negative regulation of fatty acid beta-oxidation; regulation of G1/S transition checkpoint; negative regulation of plasma membrane long-chain fatty acid transport; negative regulation of fatty acid oxidation; regulation of plasma membrane long-chain fatty acid transport; response to UV-A; activation-induced cell death of T cells; plasma membrane long-chain fatty acid transport; peripheral nervous system myelin maintenance

Melatonin-Gene Interactions: Classifier Performance

Table 14 shows the performance, as measured by area under the ROC curve, for base classifiers trained to recognize genes as potential interaction partners for melatonin. We note that the best-performing base classifier was an SVM trained on the ppi data, using a linear kernel function, with a resulting AUROC of 0.9194.

Table 14. Best melatonin-gene base classifiers.

Abbreviations: linear = linear kernel function; poly = polynomial kernel function; deg = degree parameter for polynomial kernel function.

Data Source	Kernel Function	AUROC
ppi	linear	0.9194
reactome	linear	0.9083
pmid	linear	0.8056
go	linear	0.7944
cos2	linear	0.7556
cos1	linear	0.7389
ner	linear	0.6917
cos3	linear	0.6778
kegg	poly, deg = 10	0.6333

Among the classifier ensembles, two combined classifiers achieved AUROC scores of 0.9222, which is higher than the AUROC achieved by any of the individual classifiers. This demonstrates that combining data sources can result in classifiers that perform better than any of the base classifiers. The two best ensemble classifiers are described in Table 15.

Table 15. Melatonin-gene ensemble classifiers with AUROC score of 0.9222

This AUROC score is higher than that of any of the base classifiers

The ensemble classifiers are presented in order of decreasing diversity.

Kernels	Weights	AUROC	Diversity
go, reactome, ppi, ner, pmid,	0.3424, 0.2196, 0.7427,	0.9222	1.4813

cos3	0.3375, 0.3118, 0.2678		
go, kegg, reactome, ppi, ner, pmid, cos3	0.3424, 0.0002, 0.2196, 0.7427, 0.3375, 0.3118, 0.2678	0.9222	1.2698

Among the classifier ensembles that achieved an AUROC score of 0.9222, the ensemble with the highest diversity score was based on six data sources: 'go' (Gene Ontology annotations), 'reactome' (REACTOME pathway data), 'ppi' (protein-protein interaction data), 'ner' (named entity recognition in free text), 'pmid' (shared PubMed identifiers), and 'cos3' (similarity of free-text abstracts). This classifier was selected for classifying and ranking the set of unclassified genes.

Melatonin-Gene Interactions: Biological Plausibility of Highly Ranked Unknown Interaction Partners

The go-reactome-ppi-ner-pmid-cos3 classifier was run to assign a numeric score to each of the unlabeled genes in our core gene set. The genes were then ranked based on their scores. The ten highest-ranked genes are listed in Table 16.

Table 16. Previously unlabeled genes in the core gene set that were ranked highest as potential interaction partners for melatonin. The go-reactome-ppi-ner-pmid-cos3 classifier was used to rank the genes.

entrez	symbol	name
1176	AP3S1	adaptor-related protein complex 3, sigma 1 subunit
84265	POLR3GL	polymerase (RNA) III (DNA directed) polypeptide G (32kD)-like

9184	BUB3	bub3 budding uninhibited by benzimidazoles 3 homolog
347733	TUBB2B	tubulin beta-2b chain
5296	PIK3R2	phosphatidylinositol 3-kinase, regulatory subunit, polypeptide 2
3918	LAMC2	laminin gamma 2
23705	CADM1	cell adhesion molecule 1
80025	PANK2	pantothenate kinase 2
27127	SMC1B	structural maintenance of chromosomes protein 1b
5289	PIK3C3	phosphatidylinositol 3-kinase, catalytic subunit type 3

Melatonin is a hormone with a well-characterized role in the maintenance of circadian rhythms (Cagnacci, Elliott, & Yen, 1992). It is also known to have strong antioxidant properties (Rodriguez et al., 2004).

A review of the unclassified genes highly ranked as potential interaction partners for melatonin reveals several genes involved in neurogenesis and cell migration. AP3S1 is part of the AP-3 complex, which is required for packaging proteins into vesicles for delivery to nerve terminals (Hirst, Bright, Rous, & Robinson, 1999; Simpson, Peden, Christopoulou, & Robinson, 1997). PIK3R2 and PIK3C3 are phosphatidylinositol kinases. These enzymes phosphorylate phosphatidylinositol to create second messengers involved in growth signaling pathways, which are the basis of cellular functions such as cell migration and proliferation (Gout et al., 1992). LAMC2 codes for a laminin, a class

of proteins involved in cell adhesion, differentiation, migration, signaling, neurite growth, and metastasis (Korang, Christiano, Uitto, & Mauviel, 1995; S. C. Smith et al., 2009). Studies on CADM1 reveal a possible role as a synaptic cell adhesion molecule driving synapse assembly, and possible involvement in neuronal migration, axon growth, and neuronal differentiation (Michels et al., 2008; Moiseeva, Leyland, & Bradding, 2012; Zhiling et al., 2008). Mutations in the gene PANK2 are associated with pantothenate kinase-associated neurodegeneration (Brunetti et al., 2012; Gatto, Etcheverry, Converso, Bidinost, & Rosa, 2010).

This review of the highest-ranked unclassified genes suggests the possibility of a role for melatonin in neuronal cell differentiation and migration. Such a role is supported by recent laboratory research indicating the involvement of melatonin in neurogenesis (Chern, Liao, Wang, & Shen, 2012; Ramirez-Rodriguez, Ortiz-Lopez, Dominguez-Alonso, Benitez-King, & Kempermann, 2011; Ramirez-Rodriguez, Vega-Rivera, Benitez-King, Castro-Garcia, & Ortiz-Lopez, 2012; Sarlak, Jenwitheesuk, Chetsawang, & Govitrapong, 2013). Thus, a proposed relationship between melatonin and neurogenesis, suggested by the high ranking of genes involved in neuronal cell differentiation and migration by the ensemble classifier, turns out to be a biologically plausible relationship, based on recent experimental studies.

Melatonin: Clusters and Functional Enrichment Analysis

Table 17 depicts lists of highly enriched functional annotations for clusters of genes ranked highly as potential interaction partners for melatonin. The clusters of genes related to melatonin are fairly similar in size. As in the functional enrichment analysis for

serotonin, the largest clusters of interaction partners for melatonin are highly enriched for function terms related to cell signaling and responses to chemical stimuli. An interesting difference in the functional annotations for melatonin-related genes is the large number of terms related to organonitrogen compound metabolism, and to symbiosis and inter-species processes. Both may be explained by emerging knowledge about the relationships between melatonin and mitochondria. Endosymbiotic theory suggests that mitochondria evolved from symbiotic alpha-proteobacteria (Burger & Lang, 2003; Richards & Archibald, 2011; Wallace, 2009). Mitochondria are significant sources of free radical generation. Melatonin is a strong free radical scavenger and antioxidant, and it has been theorized that synthesis of melatonin originated to protect mitochondria from oxidative and nitrosative stress (Acuña Castroviejo et al., 2011; Tan et al., 2013). Thus, our clustering and functional enrichment analysis reveals an important functional difference between genes that interact with serotonin and those that interact with melatonin.

Table 17. Functional annotations for clusters of genes ranked highly as interaction partners for melatonin. Model-based clustering, using the mclust R package, was applied to the rows of the kernel matrix on which the combined classifier is based, with each matrix row representing one gene. The annotation lists were generated using the GOstats R package, with the list of entrez IDs in one cluster serving as the input to GOstats.

cluster	# of genes	functions
clust_3	185	small molecule metabolic process; single-organism metabolic process; cellular process; response to stress; organonitrogen compound metabolic process; carbohydrate derivative biosynthetic process; phosphorus metabolic process; response to chemical stimulus; cellular metabolic process; phosphate-containing compound metabolic process
clust_5	148	G-protein coupled receptor signaling pathway; cell surface receptor signaling pathway; G-protein coupled receptor

		signaling pathway, coupled to cyclic nucleotide second messenger; small molecule metabolic process; defense response; metal ion homeostasis; organonitrogen compound metabolic process; single-organism metabolic process; cation homeostasis; regulation of biological quality
clust_4	143	cell-cell signaling; response to stimulus; response to chemical stimulus; small molecule metabolic process; regulation of biological quality; organonitrogen compound metabolic process; response to oxygen-containing compound; positive regulation of cell proliferation; response to organic substance; response to wounding
clust_6	135	single-organism metabolic process; phosphorus metabolic process; phosphate-containing compound metabolic process; small molecule metabolic process; multi-organism process; regulation of cellular protein metabolic process; cellular protein metabolic process; regulation of protein metabolic process; regulation of protein modification process; regulation of biological quality
clust_8	107	small molecule metabolic process; single-organism metabolic process; response to chemical stimulus; cellular amino acid catabolic process; immune response; organonitrogen compound metabolic process; organonitrogen compound catabolic process; response to cytokine stimulus; phosphorus metabolic process; cellular response to cytokine stimulus
clust_2	96	cellular response to stimulus; signal transduction; cell surface receptor signaling pathway; response to endogenous stimulus; response to organic nitrogen; signaling; single organism signaling; response to nitrogen compound; cell communication; response to chemical stimulus
clust_7	80	G-protein coupled receptor signaling pathway; response to chemical stimulus; response to external stimulus; chemotaxis; taxis; G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger; response to stimulus; inflammatory response; metal ion homeostasis; single-multicellular organism process
clust_1	69	response to stress; mRNA metabolic process; symbiosis, encompassing mutualism through parasitism; interspecies interaction between organisms; multi-organism process; RNA splicing, via transesterification reactions with bulged adenosine as nucleophile; mRNA splicing, via spliceosome; antigen processing and presentation; RNA splicing, via

		transesterification reactions; antigen processing and presentation of exogenous peptide antigen
clust_9	35	translational termination; nuclear-transcribed mRNA catabolic process, nonsense-mediated decay; SRP-dependent cotranslational protein targeting to membrane; cotranslational protein targeting to membrane; protein targeting to ER; establishment of protein localization to endoplasmic reticulum; protein localization to endoplasmic reticulum; viral genome expression; viral transcription; translational elongation

CHAPTER EIGHT: PROJECT #3: DRUG-GENE INTERACTIONS

This set of studies involves the ranking of genes based on the likelihood of an interaction with a pharmaceutical agent. The agent examined is risperidone, an atypical antipsychotic medication approved for use in the treatment of both schizophrenia and bipolar disorder.

Background: Risperidone

Despite wide adoption of the atypical or second-generation antipsychotics for treatment of schizophrenia and bipolar disorder, the mechanism of action of these agents is not fully understood. All antipsychotic medications have some affinity for dopaminergic receptors, but they vary in the degree of affinity for different types of dopamine receptors. They also vary in their affinity for serotonergic, cholinergic, and histaminic receptors. Variability in receptor binding is felt to have a role in the different side effect profiles of antipsychotic medications. The second-generation antipsychotic medications, such as risperidone, clozapine, olanzapine, quetiapine, ziprasidone, and aripiprazole, are felt to have a somewhat lower risk of neuromuscular side effects such as tardive dyskinesia in comparison to first-generation antipsychotic medications, such as haloperidol, fluphenazine, and chlorpromazine. However, the second generation antipsychotics have been found to increase the risk for adult-onset diabetes and lipid abnormalities for many patients.

While both first-generation and second-generation antipsychotics are effective in controlling psychotic symptoms, their mechanisms of action remain unclear. This makes it difficult to predict which medication will be the most effective for any particular patient. There has been much interest in exploring both possible genomic mechanisms underlying drug effectiveness, and genotypic markers that might predict medication effectiveness or sensitivity to particular side effects for an individual patient.

In the following study, we examine a gene expression data set comparing cells exposed to risperidone and control cells. We integrate the expression data with data from other bioinformatics information sources in order to rank genes in our core set of genes as possible interaction partners for risperidone. We then focus attention on highly-ranked genes that have not previously been reported as interaction partners for risperidone, and examine their functions. This type of exploratory analysis can help to suggest previously unrecognized mechanisms for either the therapeutic action or adverse effects of a drug.

Methods

Gene Expression Data Set

The classifiers used in this study incorporated data from the archival bioinformatics data sources, plus one data source based on gene expression data. This data source, which we designate as RISPÉRIDONE, consists of differential expression values for human neuroblastoma cells exposed to risperidone, compared to control cells not exposed to risperidone. The expression data was generated in the lab of Mas et al. at the University of Barcelona (Mas et al., 2013).

The gene expression data was downloaded from the GEO Omnibus database (Barrett & Edgar, 2006). The data set includes, for each gene, a single value for differential expression, expressed as the \log_2 of the fold change (\log_2FC) in expression values between the risperidone-exposed cells and the control cells. To create a feature matrix, a $g \times g$ matrix for our 3217 core genes is created such that each matrix cell is calculated as the difference in magnitude between the \log_2FC for gene1 and the \log_2FC for gene 2. To limit noise in the feature matrix, principle components analysis was done to reduce this to a 3217 x 32 matrix, using as features the 32 principle components that account for 95% of the variability in the data. This feature matrix was then transformed into a set of kernel matrices for incorporation into several support vector machine classifiers, as discussed in Chapter 5.

Identification of Examples for Classifier Training

The PharmGKB pharmacogenomics database provides information curated from the research literature about known interactions between pharmaceutical agents and genes. This database was used to identify a list of genes known to interact with risperidone. The list of known interaction partners was used as a set of positive examples for classifier training. Identifying a set of negative examples was more challenging, since the database stores information about known interactions but not about confirmations of the absence of an interaction. For our studies, the set of negative examples was chosen randomly, with one constraint on the population of genes used for the random selection. We created a list of all drugs in PharmGKB that are used to treat psychiatric illnesses. For each drug, we queried PharmGKB to retrieve a list of genes known to interact with

that drug. We combined these gene lists to generate a single list of genes that are known to interact with at least one psychotropic agent. These genes were subtracted from the total population of candidate genes. The remaining genes, at the present time, have not been found to interact with any psychotropic agent. These were felt to represent likely negative examples of interaction partners for risperidone. From this list, we randomly selected a list of genes equal in size to the list of known positive examples. These genes were used as negative examples for classifier training.

Results

Risperidone-Gene Interactions: Classifier Performance

Table 18 shows the performance, as measured by area under the ROC curve, for base classifiers trained to recognize genes as potential interaction partners for the drug risperidone. We note that the highest AUROC score was 0.8571. This score was achieved by two classifiers. The first was trained on the KEGG pathway data, and was based on a SVM using a polynomial kernel function. The second was trained on the PMID database of shared PubMed identifiers, and was based on a SVM using a Gaussian kernel function.

Table 18. Best risperidone-gene base classifiers.

Abbreviations: linear = linear kernel function; poly = polynomial kernel function; deg = degree parameter for polynomial kernel function; w = weight parameter for Gaussian kernel function..

Data Source	Kernel Function	AUROC
kegg	poly, deg=10	0.8571
pmid	gaussian, w=2.1	0.8571
go	poly, deg=5	0.7857
ppi	gaussian, w=5.8	0.7857

cos1	linear	0.7857
reactome	poly, deg=2	0.7143
cos2	linear	0.7143
cos3	linear	0.7143
risperidone	poly, deg=2	0.6429
ner	gaussian, w= 0.1	0.6429

Among the combined classifier ensembles, none achieved an AUROC score higher than 0.8571. For this problem domain, combining the data sources did not result in an improvement in classifier performance. The best ensemble classifiers are described in Table 19.

Table 19. Risperidone-gene ensemble classifiers with AUROC score of 0.8571. This AUROC score is the same as that of the best-performing base classifiers. The ensemble classifiers are presented in order of decreasing diversity.

Kernels	Weights	AUROC	Diversity
kegg, ppi, ner	0.6433, 0.5414, 0.5414	0.8571	1.7260
kegg, ppi, pmid	0.6433, 0.5414, 0.5414	0.8571	1.7260
kegg, ner, pmid	0.6433, 0.5414, 0.5414	0.8571	1.7260
kegg, ppi	0.7466, 0.6653	0.8571	1.4119
kegg, ner	0.7466, 0.6653	0.8571	1.4119
kegg, pmid	0.7466, 0.6653	0.8571	1.4119
kegg	1.0000	0.8571	1.0000
pmid	1.000	0.8571	1.0000

Among the classifier ensembles that achieved an AUROC score of 0.8571, the highest diversity score was 1.7260. This score was reached by three different classifier ensembles. In this case, the diversity score did not break the tie among the highest-performing classifiers. We randomly selected the ensemble based on data sources ‘kegg’ (KEGG pathway data), ‘ppi’ (protein-protein interaction data), and ‘ner’ (named entity recognition in free text). This classifier was used for classifying and ranking the set of unclassified genes.

Risperidone-Gene Interactions: Biological Plausibility of Highly Ranked Unknown Interaction Partners

The kegg-ppi-ner classifier was run to assign a numeric score to each of the unlabeled genes in our core gene set. The genes were then ranked based on their scores. The ten highest-ranked genes are listed in Table 20.

Table 20. Previously unlabeled genes in the core gene set that were ranked highest as potential interaction partners for risperidone. The kegg-ppi-ner classifier was used to rank the genes.

entrez	symbol	name
93010	B3GNT7	beta-1,3-N-acetylglucosaminyltransferase 7
10914	PAPOLA	poly(a) polymerase, alpha
126541	OR10H4	olfactory receptor, family 10, subfamily h, member 4
51302	CYP39A1	cytochrome P450, family 39, subfamily A, polypeptide 1
8021	NUP214	nucleoporin 214kDa
130399	ACVR1C	activin A receptor kinase
10250	SSRM1	serine/arginine repetitive matrix protein 1
10478	SLC25A17	solute carrier family 25 (mitochondrial carrier), member 17

2796	GNRH1	gonadotropin-releasing hormone 1
4361	MRE11A	meiotic recombination 11 homolog A

One of the highly ranked unclassified genes, CYP39A1, is a member of the cytochrome P450 family of enzymes, known to have significant involvement in drug metabolism (Wrighton & Stevens, 1992). This enzyme is a plausible interaction partner for risperidone, as it has already been established that risperidone undergoes metabolism through other members of the cytochrome P450 family, notably CYP2D6 and CYP3A4 (Berecz et al., 2004).

The relationship between risperidone and GNRH can be traced through known biological mechanisms. Dopamine normally suppresses prolactin secretion by the pituitary gland. Blockade of dopamine receptors by risperidone results in increased prolactin secretion. A rise in prolactin levels inhibits GNRH. This results in biological changes that have been reported as side effects of risperidone, including decreased menstruation and spontaneous lactation in females, and gynecomastia in men (Byerly et al., 2006; Liu-Seifert, Kinon, Tennant, Sniadecki, & Volavka, 2009; Mendhekar & Andrade, 2005; Nakonezny, Byerly, & Rush, 2007; Roke, Buitelaar, Boot, Tenback, & van Harten, 2012).

The list of unclassified genes highly ranked as potential interaction partners for risperidone reveals several genes involved in cell signaling and receptor mechanisms. SLC25A17 is a member of the mitochondrial transporter family of proteins. These proteins are potassium-chloride cotransporters that affect the cell potential by lowering

the intracellular chloride concentration. SLC25A17 is the only member of the mitochondrial transporter family known to be localized in the membranes of peroxisomes, which are cellular organelles involved in fatty acid metabolism (Agrimi, Russo, Scarcia, & Palmieri, 2012). Previous studies indicate changes in fatty acid metabolism associated with both schizophrenia and mood disorders (Hamazaki, Hamazaki, & Inadera, 2013; Iwayama et al., 2010; Maekawa, Owada, & Yoshikawa, 2011; Ramos-Loyo et al., 2013). ACVR1C is a receptor for the TGF β family of signaling molecules. When bound, these receptors phosphorylate cytoplasmic SMAD transcription factors, which move to the nucleus and can interact directly with DNA or with other transcription factors (Bondestam et al., 2001). PubMed searches do not reveal any previously established relationships between ACVR1C and either schizophrenia or risperidone. An exploration of additional mechanisms for the action of risperidone might begin with these highly-ranked potential interaction partners.

The appearance of SLC25A17 as a potential interaction partner is interesting to note. Two other genes in the class of potassium-chloride cotransporters, SLC1A1 and SLC6A4, appear in the PharmGKB database as known interaction partners for risperidone (Kwon et al., 2009; Llerena, Berecz, Penas-Lledo, Suveges, & Farinas, 2013; Lopez-Rodriguez et al., 2013). In addition, another gene in the same class, SLC18A1, was one of eight genes differentially expressed in the human neuroblastoma cells exposed to risperidone (Mas et al., 2013). Thus, we identified a potentially important gene class that was identified by gene expression experiments, even though our ensemble classifier did not include the kernel matrix based on the expression data. These findings together

support the potential usefulness of further exploration of the potential role of this gene class in the mechanism of action of risperidone.

Risperidone-Gene Interactions: Clustering and Functional Enrichment Analysis

The kernel matrix rows were clustered using model-based clustering as implemented in the R package *mclust* (Fraley et al., 2012). Functional annotation analysis was then performed on the gene list in each cluster, using the R/Bioconductor package *GOstats* (Falcon & Gentleman, 2007). Table 21 depicts lists of highly enriched functional annotations for clusters of genes ranked highly as potential interaction partners for risperidone. The clustering of the combined kernel matrix rows placed a large number of genes in a single cluster (Cluster 4). There are several very small clusters, several containing only a single gene. Here we offer some comments on Cluster 14, the largest of the small clusters, with seven members. We found this group of genes to be interesting because the clustering algorithm grouped them together and segregated them from the genes in the largest cluster. The genes in Cluster 14 are listed in Table 22.

Table 21. Functional annotations for clusters of genes ranked highly as interaction partners for risperidone. Model-based clustering, using the mclust R package, was applied to the rows of the kernel matrix on which the combined classifier is based, with each matrix row representing one gene. The annotation lists were generated using the GOstats R package, with the list of entrez IDs in one cluster serving as the input to GOstats.

cluster	# of genes	functions
Cluster 4	976	response to stimulus; response to chemical stimulus; response to stress; cellular response to stimulus; cell surface receptor signaling pathway; regulation of biological quality; single-organism cellular process; signaling; single organism

		signaling; signal transduction
Cluster 14	7	integrin-mediated signaling pathway; cell adhesion; biological adhesion; cell-substrate adhesion; cell-substrate junction assembly; leukocyte migration; cell migration; hemidesmosome assembly; cell motility; localization of cell
Cluster 2	4	response to mycotoxin; B cell selection; B cell negative selection; establishment or maintenance of transmembrane electrochemical gradient; post-embryonic camera-type eye morphogenesis
Cluster 12	2	cell-matrix adhesion; cell-substrate adhesion
Cluster 1	1	nucleotide-excision repair, DNA damage recognition
Cluster 5	1	regulation of blood vessel remodeling; positive regulation of protein kinase C signaling cascade; regulation of protein kinase C signaling cascade; lymphangiogenesis
Cluster 6	1	D-aspartate transport; D-aspartate import; D-amino acid transport; C4-dicarboxylate transport; aspartate transport; L-glutamate import; L-amino acid import; amino acid import
Cluster 7	1	glycogen cell differentiation involved in embryonic placenta development; negative regulation of fatty acid beta-oxidation; regulation of G1/S transition checkpoint; negative regulation of plasma membrane long-chain fatty acid transport; negative regulation of fatty acid oxidation; regulation of plasma membrane long-chain fatty acid transport; response to UV-A; activation-induced cell death of T cells; plasma membrane long-chain fatty acid transport; peripheral nervous system myelin maintenance
Cluster 8	1	aromatic amino acid transport
Cluster 10	1	protein neddylation
Cluster 11	1	ISG15-protein conjugation; histone H2B ubiquitination
Cluster 13	1	hemidesmosome assembly
Cluster 15	1	ganglioside catabolic process; glycosphingolipid catabolic process; neuromuscular process controlling posture; glycolipid catabolic process; keratan sulfate catabolic process; ganglioside metabolic process; chondroitin sulfate catabolic process; hyaluronan catabolic process

Table 22. Genes in Cluster 14, one of the clusters of potential interaction partners for risperidone.

entrez	symbol	name
7058	THBS2	thrombospondin 2
3678	ITGA5	integrin, alpha 5
3695	ITGB7	integrin, beta 7
3655	ITGA6	integrin, alpha 6
3685	ITGAV	integrin, alpha V
3694	ITGB6	integrin, beta 6
3915	LAMC1	laminin, gamma 1

Integrins are integral membrane proteins, involved in adhesion and cell-surface mediated signal transduction (Arcangeli & Becchetti, 2010; Campbell & Humphries, 2011). There is a plausible relationship between integrins and risperidone. Patients studied after a first episode of schizophrenia were found to have increased expression of the integrin receptor alpha(IIb)beta(IIIa) (Walsh et al., 2002). Other studies have shown that patients diagnosed with schizophrenia show higher platelet aggregation than healthy individuals, and antipsychotic medications, including risperidone, reduce platelet aggregation (De Clerck, Somers, Mannaert, Greenspan, & Eerdeken, 2004; Dietrich-Muszalska & Olas, 2009). Peptides designed to bind with integrins have been shown to inhibit platelet aggregation (Silverman, Kariolis, & Cochran, 2011). It is possible that the reduction in platelet aggregation related to risperidone could be caused by an interaction between risperidone and one or more of the integrins. Interestingly, a PubMed search for

articles discussing both risperidone and integrins yields zero results. Thus, this relationship would not be revealed by a simple literature search.

Thrombospondin-2, another gene included in this cluster, is a glycoprotein known to have a role in angiogenesis (Bornstein, Kyriakides, Yang, Armstrong, & Birk, 2000) and inhibition of tumor growth (Hawighorst et al., 2001). It also has been demonstrated that thrombospondin-2 activity is required in megakaryocytes for normal platelet formation and function (Kyriakides et al., 2003). The link between thrombospondin-2 and platelet function, and between the integrins and platelet function, supports the biological plausibility of clustering gene THBS2 with the integrins.

Laminins are a constituent of basement membranes, and are widely expressed in the nervous system (Yang, Ma, Liu, & Lee, 2011). They are involved in cell adhesion, differentiation, migration, signaling, neurite outgrowth, and metastasis. Laminin expression has been found to be decreased in the parieto-occipital cortex in patients with schizophrenia, depression, and bipolar disorder, and increased in the prefrontal cortex in patients with schizophrenia (Laifenfeld, Karry, Klein, & Ben-Shachar, 2005). In post-mortem studies of depressed patients, antidepressant treatment was found to reverse the decrease in laminin expression in the parieto-occipital cortex (Laifenfeld et al., 2005). The positive effects of antidepressants on mood may be modulated through the mechanism of neuronal plasticity (Castrén & Hen, 2013). There is also strong evidence that disruption in neuronal plasticity is a key mechanism in the etiology of schizophrenia (Hasan, Falkai, & Wobrock, 2013; Meyer-Lindenberg & Tost, 2013; Voineskos, Rogasch, Rajji, Fitzgerald, & Daskalakis, 2013). Studies also suggest that effects on

neuronal plasticity may be a component of the action of atypical antipsychotic medications (Calabrese et al., 2013; Fumagalli et al., 2012; Fumagalli et al., 2003; Molteni, Calabrese, Racagni, Fumagalli, & Riva, 2009). While our analysis suggests the gene for the laminin gamma-1 chain as a possible interaction partner for risperidone, gene sequencing studies have revealed mutations for genes coding for the laminin alpha-1 chain (Girard et al., 2011) and the laminin alpha-2 chain (B. Xu et al., 2012) in patients with schizophrenia. In summary, these studies support the biological plausibility of a relationship between risperidone and genes that code for laminins.

CHAPTER NINE: CONCLUSIONS

Contributions

The research hypothesis stated in Chapter 1 is as follows:

“Multiple kernel learning enhances the bioinformatics analysis of candidate gene lists by [1] allowing integration of heterogeneous data sources into the analysis, [2] providing a prioritized ranking of genes in the list, [3] facilitating clustering of the genes using the kernel matrix, and [4] simplifying functional enrichment analysis by clustering the original gene list into meaningful sub-lists.”

We believe that all of the elements of this research hypothesis have been demonstrated.

[1] We have shown that the multiple kernel learning approach allows integration of bioinformatics data sources in various data formats into the gene prioritization task. Data sources that we were able to utilize include Gene Ontology annotations, KEGG pathway annotations, REACTOME pathway annotations, protein-protein interaction data from the STRING database, free-text abstracts from the PubMed database, results from microRNA target prediction algorithms, and gene expression data. While other studies have explored the use of kernel methods for gene prioritization, our studies are unique in using the kernel weights calculated by the multiple kernel learning algorithm both to identify the most useful data sources and to maximize the diversity of base classifiers in

the final classifier ensemble while limiting the incorporation of base classifiers that do not contribute significantly to classifier performance.

[2] Binary classification algorithms such as support vector machines typically provide output in the form as true/false labels for instances. This can generate two long, unordered lists of labeled instances. Generating new insights from unordered gene lists can be difficult. By recovering the distance of each newly labeled instance from the maximally separating hyperplane, we can order the instances in terms of their strength of membership in the positive or negative class.

[3] We have not seen other examples in the research literature of using the rows of a combined kernel matrix to cluster instances. We have found that clustering the kernel matrix rows provides an additional avenue for analysis of a gene list. While the prioritized list of labeled instances allows identification of genes that are most similar to the training instances, clustering allows identification of groups of genes that are similar to each other.

[4] Clustering of the genes by the kernel matrix rows provides small sub-lists of genes that can be submitted for functional enrichment analysis. We see this as preferable to submitting the entire list of positively labeled genes for enrichment analysis. The list of all positive examples may include functions that are not enriched for the entire list of positive examples, since that list includes a large number of genes representing a wide variety of functions. Our analysis revealed some gene clusters that could be distinguished from other clusters by their unique functional enrichment. The functions identified for the gene clusters provides a basis for proposing functions for the entity that interacts with

those genes. We generated hypotheses for functions and mechanisms of action for the neurotransmitters serotonin and melatonin, several microRNAs, and the drug risperidone, based on functional enrichment analysis of their interaction partners. We were able to find results in the research literature that supports the biological plausibility of these hypotheses. We noted several examples in which the research support for a function suggested by our enrichment analysis appears in recent research papers that had not yet been written at the time we completed the construction of our classifiers.

Limitations and Areas for Future Work

The research described here demonstrates a kernel-based machine learning approach to gene prioritization that integrates data from multiple sources. One limitation of this approach is shared with all supervised machine learning methods, which is the requirement of a set of examples with known labels for classifier training. For experimental data in which there are no known examples for training, exploration of the data set using unsupervised learning methods might help to clarify the structure of the data set, uncovering groups of related instances and features that are most useful in separating the instances into groups. Instances that emerge as the clearest prototype examples from each group could then be selected as examples for classifier training.

The archival bioinformatics data sources vary in their coverage of genes. We limited our list of core genes to genes that had entries in all of the data sources. This reliance on prior knowledge limits the number of genes that can be included in the prioritized list of interaction partners. A possible improvement might be some method for imputation of missing database entries. Any such method brings the risk of introducing

inaccuracies that could defeat the advantages of using archival data, if the number of imputed database entries is large in comparison to the number of curated entries.

We did not do a formal evaluation of methods for gene name entity recognition, but instead selected a method that has been tested in other well-documented research. However, there is much ongoing research in gene name entity recognition. More detailed testing of available methods may help us to identify approaches that would perform better on the recognition of gene names in the free text abstracts.

In the design of both the base SVM classifiers and the combined MKL classifiers, there are many parameters that could be optimized. We selected a small number of parameters for optimization, and used default values for other parameters. A more rigorous approach to parameter testing will help determine which parameters have the greatest influence on the performance of either the base classifiers or the combined classifiers.

REFERENCES

- Acuña Castroviejo, D., López, L. C., Escames, G., López, A., García, J. A., & Reiter, R. J. (2011). Melatonin-mitochondria interplay in health and disease. *Current Topics In Medicinal Chemistry*, 11(2), 221-240.
- Agrimi, G., Russo, A., Scarcia, P., & Palmieri, F. (2012). The human gene SLC25A17 encodes a peroxisomal transporter of coenzyme A, FAD and NAD⁺. *Biochem J*, 443(1), 241-247. doi: 10.1042/BJ20111420
- Aizerman, A., Braverman, E. M., & Rozoner, L. I. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25, 821-837. doi: citeulike-article-id:431797
- Allais-Bonnet, A., Grohs, C., Medugorac, I., Krebs, S., Djari, A., Graf, A., . . . Capitan, A. (2013). Novel Insights into the Bovine Polled Phenotype and Horn Ontogenesis in *Bovidae*. *PLoS One*, 8(5), e63512. doi: 10.1371/journal.pone.0063512
- Ameres, S. L., & Zamore, P. D. (2013). Diversifying microRNA sequence and function. *Nat Rev Mol Cell Biol*, 14(8), 475-488. doi: 10.1038/nrm3611
- Arcangeli, A., & Becchetti, A. (2010). Integrin structure and functional relation with ion channels. *Adv Exp Med Biol*, 674, 1-7.
- Aszfalg, J., Kriegel, H.-P., Pryakhin, A., & Schubert, M. (2007). *Multi-represented classification based on confidence estimation*. Paper presented at the Proceedings of the 11th Pacific-Asia conference on Advances in knowledge discovery and data mining, Nanjing, China.
- Bach, F. R., Lanckriet, G. R. G., & Jordan, M. I. (2004). *Multiple kernel learning, conic duality, and the SMO algorithm*. Paper presented at the Proceedings of the twenty-first international conference on Machine learning, Banff, Alberta, Canada.
- Barrett, T., & Edgar, R. (2006). Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*, 411, 352-369. doi: 10.1016/S0076-6879(06)11019-8

- Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, 136(2), 215-233. doi: 10.1016/j.cell.2009.01.002
- Berecz, R., Dorado, P., De La Rubia, A., Caceres, M. C., Degrell, I., & A, L. L. (2004). The role of cytochrome P450 enzymes in the metabolism of risperidone and its clinical relevance for drug interactions. *Curr Drug Targets*, 5(6), 573-579.
- Berry, M. W., Drmac, Z., & Jessup, E. R. (1999). Matrices, vector spaces, and information retrieval. *SIAM Review*, 41(2), 335-362.
- Bondestam, J., Huotari, M. A., Moren, A., Ustinov, J., Kaivo-Oja, N., Kallio, J., . . . Ritvos, O. (2001). cDNA cloning, expression studies and chromosome mapping of human type I serine/threonine kinase receptor ALK7 (ACVR1C). *Cytogenet Cell Genet*, 95(3-4), 157-162. doi: 59339
- Bornstein, P., Kyriakides, T. R., Yang, Z., Armstrong, L. C., & Birk, D. E. (2000). Thrombospondin 2 Modulates Collagen Fibrillogenesis and Angiogenesis. *J Investig Dermatol Symp Proc*, 5(1), 61-66.
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers*. Paper presented at the Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, USA.
- Breiman, L. (1996). Bagging predictors. *Mach. Learn.*, 24(2), 123-140. doi: 10.1023/a:1018054314350
- Brunetti, D., Dusi, S., Morbin, M., Uggetti, A., Moda, F., D'Amato, I., . . . Tiranti, V. (2012). Pantothenate kinase-associated neurodegeneration: altered mitochondria membrane potential and defective respiration in Pank2 knock-out mouse model. *Hum Mol Genet*, 21(24), 5294-5305. doi: 10.1093/hmg/dds380
- Brunham, L. R., Singaraja, R. R., & Hayden, M. R. (2006). Variations on a Gene: Rare and Common Variants in ABCA1 and Their Impact on HDL Cholesterol Levels and Atherosclerosis. *Annual Review of Nutrition*, 26(1), 105-129. doi: doi:10.1146/annurev.nutr.26.061505.111214
- Burger, G., & Lang, B. F. (2003). Parallels in genome evolution in mitochondria and bacterial symbionts. *IUBMB Life*, 55(4-5), 205-212. doi: 10.1080/1521654031000137380
- Byerly, M. J., Nakonezny, P. A., Bettcher, B. M., Carmody, T., Fisher, R., & Rush, A. J. (2006). Sexual dysfunction associated with second-generation antipsychotics in outpatients with schizophrenia or schizoaffective disorder: an empirical

- evaluation of olanzapine, risperidone, and quetiapine. *Schizophr Res*, 86(1-3), 244-250. doi: 10.1016/j.schres.2006.04.005
- Cagnacci, A., Elliott, J. A., & Yen, S. S. (1992). Melatonin: a major regulator of the circadian rhythm of core temperature in humans. *J Clin Endocrinol Metab*, 75(2), 447-452. doi: 10.1210/jcem.75.2.1639946
- Calabrese, F., Luoni, A., Guidotti, G., Racagni, G., Fumagalli, F., & Riva, M. A. (2013). Modulation of neuronal plasticity following chronic concomitant administration of the novel antipsychotic lurasidone with the mood stabilizer valproic acid. *Psychopharmacology (Berl)*, 226(1), 101-112. doi: 10.1007/s00213-012-2900-0
- Campbell, I. D., & Humphries, M. J. (2011). Integrin structure, activation, and interactions. *Cold Spring Harb Perspect Biol*, 3(3). doi: 10.1101/cshperspect.a004994
- Castrén, E., & Hen, R. (2013). Neuronal plasticity and antidepressant actions. *Trends in Neurosciences*, 36(5), 259-267. doi: <http://dx.doi.org/10.1016/j.tins.2012.12.010>
- Chandran, S., Guo, T., Tolliver, T., Chen, W., Murphy, D., & McPherron, A. (2012). Effects of serotonin on skeletal muscle growth. *BMC Proceedings*, 6(Suppl 3), O3.
- Chang, J. T., Schutze, H., & Altman, R. B. (2004). GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*, 20(2), 216-225.
- Chern, C. M., Liao, J. F., Wang, Y. H., & Shen, Y. C. (2012). Melatonin ameliorates neural function by promoting endogenous neurogenesis through the MT2 melatonin receptor in ischemic-stroke mice. *Free Radic Biol Med*, 52(9), 1634-1647. doi: 10.1016/j.freeradbiomed.2012.01.030
- Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Mach. Learn.*, 20(3), 273-297. doi: 10.1023/a:1022627411411
- Dai, Y. P., Bongalon, S., Tian, H., Parks, S. D., Mutafova-Yambolieva, V. N., & Yamboliev, I. A. (2006). Upregulation of profilin, cofilin-2 and LIMK2 in cultured pulmonary artery smooth muscle cells and in pulmonary arteries of monocrotaline-treated rats. *Vascul Pharmacol*, 44(5), 275-282. doi: 10.1016/j.vph.2005.11.008
- De Clerck, F., Somers, Y., Mannaert, E., Greenspan, A., & Eerdeken, M. (2004). In vitro effects of risperidone and 9-hydroxy-risperidone on human platelet function, plasma coagulation, and fibrinolysis. *Clin Ther*, 26(8), 1261-1273.

- Dietrich-Muszalska, A., & Olas, B. (2009). The changes of aggregability of blood platelets in schizophrenia. *World J Biol Psychiatry*, *10*(2), 171-176.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), *Multiple Classifier Systems* (Vol. 1857, pp. 1-15). Berlin: Springer-Verlag Berlin.
- Dong, H., Lei, J., Ding, L., Wen, Y., Ju, H., & Zhang, X. (2013). MicroRNA: Function, Detection, and Bioanalysis. *Chem Rev.* doi: 10.1021/cr300362f
- Dong, J., Cui, X., Jiang, Z., & Sun, J. (2013). MicroRNA-23a modulates tumor necrosis factor-alpha-induced osteoblasts apoptosis by directly targeting Fas. *J Cell Biochem*, n/a-n/a. doi: 10.1002/jcb.24622
- Drucker, H., Cortes, C., Jackel, L. D., LeCun, Y., & Vapnik, V. (1994). Boosting and Other Ensemble Methods. *Neural Computation*, *6*(6), 1289-1301. doi: doi:10.1162/neco.1994.6.6.1289
- Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification* (Second ed.). New York: Wiley-Interscience.
- Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.
- Egorov, S., Yuryev, A., & Daraselia, N. (2004). A simple and practical dictionary-based approach for identification of proteins in Medline abstracts. *J Am Med Inform Assoc*, *11*(3), 174-178. doi: 10.1197/jamia.M1453
- El-Melegy, M., & Ahmed, S. (2007). Neural Networks in Multiple Classifier Systems for Remote-Sensing Image Classification. In M. Nachttegael, D. Van der Weken, E. Kerre & W. Philips (Eds.), *Soft Computing in Image Processing* (Vol. 210, pp. 65-94): Springer Berlin / Heidelberg.
- Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. S. (2003). MicroRNA targets in Drosophila. *Genome Biol*, *5*(1), R1. doi: 10.1186/gb-2003-5-1-r1
- Falcon, S., & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, *23*(2), 257-258. doi: 10.1093/bioinformatics/btl567
- Fawcett, T. (2006). An introduction to ROC analysis. *ROC Analysis in Pattern Recognition*, *27*(8), 861-874. doi: citeulike-article-id:820297

- Fernández-Hernando, C., Suárez, Y., Rayner, K. J., & Moore, K. J. (2011). MicroRNAs in lipid metabolism. *Current Opinion in Lipidology*, 22(2), 86-92. doi: 10.1097/MOL.1090b1013e3283428d3283429d.
- Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., . . . Searle, S. M. (2012). Ensembl 2012. *Nucleic Acids Res*, 40(Database issue), D84-90. doi: 10.1093/nar/gkr991
- Flowers, E., Froelicher, E. S., & Aouizerat, B. E. (2013). MicroRNA regulation of lipid metabolism. *Metabolism: clinical and experimental*, 62(1), 12-20.
- Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611-631.
- Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012). mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation (Technical Report No. 597): Department of Statistics, University of Washington.
- Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., . . . Kent, W. J. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Research*. doi: 10.1093/nar/gkq963
- Fumagalli, F., Calabrese, F., Luoni, A., Bolis, F., Racagni, G., & Riva, M. A. (2012). Modulation of BDNF expression by repeated treatment with the novel antipsychotic lurasidone under basal condition and in response to acute stress. *Int J Neuropsychopharmacol*, 15(2), 235-246. doi: 10.1017/S1461145711000150
- Fumagalli, F., Molteni, R., Roceri, M., Bedogni, F., Santero, R., Fossati, C., . . . Riva, M. A. (2003). Effect of antipsychotic drugs on brain-derived neurotrophic factor expression under reduced N-methyl-D-aspartate receptor activity. *J Neurosci Res*, 72(5), 622-628. doi: 10.1002/jnr.10609
- Fumera, G., & Roli, F. (2005). A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Trans Pattern Anal Mach Intell*, 27(6), 942-956. doi: 10.1109/TPAMI.2005.109
- Gatto, E., Etcheverry, J. L., Converso, D. P., Bidinost, C., & Rosa, A. (2010). Pantothenate kinase-associated neurodegeneration: novel mutations in the PANK2 gene in an Argentinean young woman. *Mov Disord*, 25(13), 2262-2264. doi: 10.1002/mds.23063

- Gerthoffer, W. T. (2005). Actin cytoskeletal dynamics in smooth muscle contraction. *Can J Physiol Pharmacol*, 83(10), 851-856. doi: 10.1139/y05-088
- Giacinto, G., Roli, F., & Didaci, L. (2003). Fusion of multiple classifiers for intrusion detection in computer networks. *Pattern Recogn. Lett.*, 24(12), 1795-1803. doi: 10.1016/s0167-8655(03)00004-7
- Girard, S. L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., . . . Rouleau, G. A. (2011). Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet*, 43(9), 860-863. doi: <http://www.nature.com/ng/journal/v43/n9/abs/ng.886.html#supplementary-information>
- Glatzer, S., Merten, N. J., Dierks, C., Wöhlke, A., Philipp, U., & Distl, O. (2013). A Single Nucleotide Polymorphism within the *Interferon Gamma Receptor 2* Gene Perfectly Coincides with Polledness in Holstein Cattle. *PLoS One*, 8(6), e67992. doi: 10.1371/journal.pone.0067992
- Gonen, M., & Alpaydin, E. (2011). Multiple Kernel Learning Algorithms. *J. Mach. Learn. Res.*, 12, 2211-2268.
- Gong, Y., Renigunta, V., Himmerkus, N., Zhang, J., Renigunta, A., Bleich, M., & Hou, J. (2012). Claudin-14 regulates renal Ca⁺⁺ transport in response to CaSR signalling via a novel microRNA pathway. *EMBO J*, 31(8), 1999-2012. doi: http://www.nature.com/emboj/journal/v31/n8/supinfo/emboj201249a_S1.html
- Gout, I., Dhand, R., Panayotou, G., Fry, M. J., Hiles, I., Otsu, M., & Waterfield, M. D. (1992). Expression and characterization of the p85 subunit of the phosphatidylinositol 3-kinase complex and a related p85 beta protein by using the baculovirus expression system. *Biochem J*, 288 (Pt 2), 395-405.
- Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucl. Acids Res.*, 34(suppl_1), D140-144. doi: 10.1093/nar/gkj112
- Griffiths-Jones, S., Saini, H. K., van Dongen, S., & Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucl. Acids Res.*, 36(suppl_1), D154-158. doi: 10.1093/nar/gkm952
- Hady, M. F. A., & Schwenker, F. (2010). Combining Committee-Based Semi-Supervised Learning and Active Learning. *Journal of Computer Science and Technology*, 25(4), 681-698. doi: 10.1007/s11390-010-9357-6

- Hamazaki, K., Hamazaki, T., & Inadera, H. (2013). Abnormalities in the fatty acid composition of the postmortem entorhinal cortex of patients with schizophrenia, bipolar disorder, and major depressive disorder. *Psychiatry Res*, *210*(1), 346-350. doi: 10.1016/j.psychres.2013.05.006
- Hanisch, D., Fluck, J., Mevissen, H. T., & Zimmer, R. (2003). Playing biology's name game: identifying protein names in scientific text. *Pac Symp Biocomput*, 403-414.
- Hasan, A., Falkai, P., & Wobrock, T. (2013). Transcranial brain stimulation in schizophrenia: targeting cortical excitability, connectivity and plasticity. *Curr Med Chem*, *20*(3), 405-413.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, New York: Springer.
- Hawighorst, T., Velasco, P., Streit, M., Hong, Y.-K., Kyriakides, T. R., Brown, L. F., . . . Detmar, M. (2001). Thrombospondin-2 plays a protective role in multistep carcinogenesis: a novel host anti-tumor defense mechanism. *EMBO J*, *20*(11), 2631-2640.
- He, X., Eberhart, J. K., & Postlethwait, J. H. (2009). MicroRNAs and micromanaging the skeleton in disease, development and evolution. *Journal of Cellular and Molecular Medicine*, *13*(4), 606-618. doi: 10.1111/j.1582-4934.2009.00696.x
- Heng, Y. W., Lim, H. H., Mina, T., Utomo, P., Zhong, S., Lim, C. T., & Koh, C. G. (2012). TPPP acts downstream of RhoA-ROCK-LIMK2 to regulate astral microtubule organization and spindle orientation. *J Cell Sci*, *125*(Pt 6), 1579-1590. doi: 10.1242/jcs.096818
- Hirst, J., Bright, N. A., Rous, B., & Robinson, M. S. (1999). Characterization of a fourth adaptor-related protein complex. *Mol Biol Cell*, *10*(8), 2787-2802.
- Ho, T. K., Hull, J. J., & Srihari, S. N. (1994). Decision combination in multiple classifier systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *16*(1), 66-75.
- Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, *31*(13), 3429-3431. doi: 10.1093/nar/gkg599
- Itoh, T., Nozawa, Y., & Akao, Y. (2009). MicroRNA-141 and -200a are involved in bone morphogenetic protein-2-induced mouse pre-osteoblast differentiation by targeting distal-less homeobox 5. *J Biol Chem*, *284*(29), 19272-19279. doi: M109.014001 [pii]; 10.1074/jbc.M109.014001

- Iwayama, Y., Hattori, E., Maekawa, M., Yamada, K., Toyota, T., Ohnishi, T., . . . Yoshikawa, T. (2010). Association analyses between brain-expressed fatty-acid binding protein (FABP) genes and schizophrenia and bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet*, *153B*(2), 484-493. doi: 10.1002/ajmg.b.31004
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive Mixtures of Local Experts. *Neural Computation*, *3*(1), 79-87. doi: doi:10.1162/neco.1991.3.1.79
- Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., . . . von Mering, C. (2009). STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, *37*(suppl 1), D412-D416. doi: 10.1093/nar/gkn760
- Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, *7*, 119-129.
- Jenssen, T.-K., Laegreid, A., Komorowski, J., & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, *28*(1), 21-28.
- Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, *6*(2), 181-214. doi: doi:10.1162/neco.1994.6.2.181
- Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, *11*(i09).
- Kee, A. J., Gunning, P. W., & Hardeman, E. C. (2009). Diverse roles of the actin cytoskeleton in striated muscle. *J Muscle Res Cell Motil*, *30*(5-6), 187-197. doi: 10.1007/s10974-009-9193-x
- Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., & Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat Genet*, *39*(10), 1278 - 1284.
- Khalili, A. (2010). New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics-Revue Canadienne De Statistique*, *38*(4), 519-539.
- Koike, A., Niwa, Y., & Takagi, T. (2005). Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, *21*(7), 1227-1236. doi: 10.1093/bioinformatics/bti084

- Korang, K., Christiano, A. M., Uitto, J., & Mauviel, A. (1995). Differential cytokine modulation of the genes LAMA3, LAMB3, and LAMC2, encoding the constitutive polypeptides, alpha 3, beta 3, and gamma 2, of human laminin 5 in epidermal keratinocytes. *FEBS Lett*, 368(3), 556-558.
- Kou, Z., Cohen, W. W., & Murphy, R. F. (2005). High-recall protein entity recognition using a dictionary. *Bioinformatics*, 21 Suppl 1, i266-273. doi: 10.1093/bioinformatics/bti1006
- Koubassova, N. A., & Tsaturyan, A. K. (2011). Molecular mechanism of actin-myosin motor in muscle. *Biochemistry (Mosc)*, 76(13), 1484-1506. doi: 10.1134/S0006297911130086
- Krek, A., Grun, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., . . . Rajewsky, N. (2005). Combinatorial microRNA target predictions. *Nat Genet*, 37(5), 495-500. doi: 10.1038/ng1536
- Kruger, J., & Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res*, 34(Web Server issue), W451-454. doi: 10.1093/nar/gkl243
- Kuncheva, L. (2004). *Combining Pattern Classifiers: Methods and Algorithms*: Wiley-Interscience.
- Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, 51(2), 181-207. doi: 10.1023/a:1022859003006
- Kwon, J. S., Joo, Y. H., Nam, H. J., Lim, M., Cho, E. Y., Jung, M. H., . . . Hong, K. S. (2009). Association of the glutamate transporter gene SLC1A1 with atypical antipsychotics-induced obsessive-compulsive symptoms. *Arch Gen Psychiatry*, 66(11), 1233-1241. doi: 10.1001/archgenpsychiatry.2009.155
- Kyriakides, T. R., Rojnuckarin, P., Reidy, M. A., Hankenson, K. D., Papayannopoulou, T., Kaushansky, K., & Bornstein, P. (2003). Megakaryocytes require thrombospondin-2 for normal platelet formation and function. *Blood*, 101(10), 3915-3923. doi: 10.1182/blood.V101.10.3915
- Laifenfeld, D., Karry, R., Klein, E., & Ben-Shachar, D. (2005). Alterations in cell adhesion molecule L1 and functionally related genes in major depression: A postmortem study. *Biological Psychiatry*, 57(7), 716-725.

- Lanckriet, G. R., Deng, M., Cristianini, N., Jordan, M. I., & Noble, W. S. (2004). Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput*, 300-311.
- Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I., & Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16), 2626-2635. doi: 10.1093/bioinformatics/bth294
- Lee-Young, R. S., Griffee, S. R., Lynes, S. E., Bracy, D. P., Ayala, J. E., McGuinness, O. P., & Wasserman, D. H. (2009). Skeletal muscle AMP-activated protein kinase is essential for the metabolic response to exercise in vivo. *J Biol Chem*, 284(36), 23925-23934. doi: 10.1074/jbc.M109.021048
- Lehman, W., & Morgan, K. G. (2012). Structure and dynamics of the actin-based smooth muscle contractile and cytoskeletal apparatus. *J Muscle Res Cell Motil*, 33(6), 461-469. doi: 10.1007/s10974-012-9283-z
- Leser, U., & Hakenberg, J. (2005). What makes a gene name? Named entity recognition in the biomedical literature. *Brief Bioinform*, 6(4), 357-369. doi: 10.1093/bib/6.4.357
- Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, 120(1), 15-20. doi: DOI: 10.1016/j.cell.2004.12.035
- Li, Z., Hassan, M. Q., Volinia, S., van Wijnen, A. J., Stein, J. L., Croce, C. M., . . . Stein, G. S. (2008). A microRNA signature for a BMP2-induced osteoblast lineage commitment program. *Proceedings of the National Academy of Sciences*, 105(37), 13906-13911. doi: 10.1073/pnas.0804438105
- Lin, E. A., Kong, L., Bai, X. H., Luan, Y., & Liu, C. J. (2009). miR-199a, a bone morphogenic protein 2-responsive MicroRNA, regulates chondrogenesis via direct targeting to Smad1. *J Biol Chem*, 284(17), 11326-11335. doi: M807709200 [pii]; 10.1074/jbc.M807709200
- Liu-Seifert, H., Kinon, B. J., Tennant, C. J., Sniadecki, J., & Volavka, J. (2009). Sexual dysfunction in patients with schizophrenia treated with conventional antipsychotics or risperidone. *Neuropsychiatr Dis Treat*, 5, 47-54.
- Llerena, A., Berez, R., Penas-Lledo, E., Suveges, A., & Farinas, H. (2013). Pharmacogenetics of clinical response to risperidone. *Pharmacogenomics*, 14(2), 177-194. doi: 10.2217/pgs.12.201

- Lopez-Rodriguez, R., Cabaleiro, T., Ochoa, D., Roman, M., Borobia, A. M., Carcas, A. J., . . . Abad-Santos, F. (2013). Pharmacodynamic genetic variants related to antipsychotic adverse reactions in healthy volunteers. *Pharmacogenomics*, *14*(10), 1203-1214. doi: 10.2217/pgs.13.106
- Maekawa, M., Owada, Y., & Yoshikawa, T. (2011). Role of polyunsaturated fatty acids and fatty acid binding protein in the pathogenesis of schizophrenia. *Curr Pharm Des*, *17*(2), 168-175.
- Maragkakis, M., Reczko, M., Simossis, V. A., Alexiou, P., Papadopoulos, G. L., Dalamagas, T., . . . Hatzigeorgiou, A. G. (2009). DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucl. Acids Res.*, *37*(suppl_2), W273-276. doi: 10.1093/nar/gkp292
- Mas, S., Gasso, P., Bernardo, M., & Lafuente, A. (2013). Functional analysis of gene expression in risperidone treated cells provide new insights in molecular mechanism and new candidate genes for pharmacogenetic studies. *Eur Neuropsychopharmacol*, *23*(4), 329-337. doi: 10.1016/j.euroneuro.2012.04.016
- Maziere, P., & Enright, A. J. (2007). Prediction of microRNA targets. *Drug Discov Today*, *12*(11-12), 452-458. doi: 10.1016/j.drudis.2007.04.002
- McDonald, D. M., Chen, H., Su, H., & Marshall, B. B. (2004). Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. *Bioinformatics*, *20*(18), 3370-3378. doi: 10.1093/bioinformatics/bth409
- Mendhekar, D. N., & Andrade, C. R. (2005). Unilateral gynecomastia induced by risperidone in a geriatric male patient. *Indian J Med Sci*, *59*(8), 361-362.
- Meyer-Lindenberg, A., & Tost, H. (2013). Neuroimaging and plasticity in schizophrenia. *Restor Neurol Neurosci*. doi: 10.3233/RNN-139014
- Michels, E., Hoebeeck, J., De Preter, K., Schramm, A., Brichard, B., De Paepe, A., . . . Speleman, F. (2008). CADM1 is a strong neuroblastoma candidate gene that maps within a 3.72 Mb critical region of loss on 11q23. *BMC Cancer*, *8*, 173. doi: 10.1186/1471-2407-8-173
- Moiseeva, E. P., Leyland, M. L., & Bradding, P. (2012). CADM1 isoforms differentially regulate human mast cell survival and homotypic adhesion. *Cell Mol Life Sci*, *69*(16), 2751-2764. doi: 10.1007/s00018-012-0948-y
- Molteni, R., Calabrese, F., Racagni, G., Fumagalli, F., & Riva, M. A. (2009). Antipsychotic drug actions on gene modulation and signaling mechanisms. *Pharmacol Ther*, *124*(1), 74-85. doi: 10.1016/j.pharmthera.2009.06.001

- Nakonezny, P. A., Byerly, M. J., & Rush, A. J. (2007). The relationship between serum prolactin level and sexual functioning among male outpatients with schizophrenia or schizoaffective disorder: a randomized double-blind trial of risperidone vs. quetiapine. *J Sex Marital Ther*, *33*(3), 203-216. doi: 10.1080/00926230701267829
- Ng, S. K., & McLachlan, G. J. (2007). Extension of mixture-of-experts networks for binary classification of hierarchical data. *Artificial Intelligence in Medicine*, *41*(1), 57-67. doi: 10.1016/j.artmed.2007.06.001
- Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P., & Hatzigeorgiou, A. G. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res*, *37*(Database issue), D155-158. doi: 10.1093/nar/gkn809
- Peng, Y., & Zhang, X. (2007). Integrative data mining in systems biology: from text to network mining. *Artif Intell Med*, *41*(2), 83-86. doi: 10.1016/j.artmed.2007.08.001
- Poggio, T., Mukherjee, S., Rifkin, R., Rakhlin, A., & Verri, A. (2001). b. *CBCL Paper #198/AI Memo #2001-011*. Cambridge, MA: Massachusetts Institute of Technology.
- Polikar, R. (2006). Ensemble Based Systems in Decision Making. *IEEE Circuits and Systems Magazine*, *6*(3), 21-45. doi: citeulike-article-id:2820275
- Price, S. (2010). Bone: novel microRNA expressed in osteoblasts promotes bone formation. *Nat Rev Rheumatol*, *6*(2), 64.
- Ramirez-Rodriguez, G., Ortiz-Lopez, L., Dominguez-Alonso, A., Benitez-King, G. A., & Kempermann, G. (2011). Chronic treatment with melatonin stimulates dendrite maturation and complexity in adult hippocampal neurogenesis of mice. *J Pineal Res*, *50*(1), 29-37. doi: 10.1111/j.1600-079X.2010.00802.x
- Ramirez-Rodriguez, G., Vega-Rivera, N. M., Benitez-King, G., Castro-Garcia, M., & Ortiz-Lopez, L. (2012). Melatonin supplementation delays the decline of adult hippocampal neurogenesis during normal aging of mice. *Neurosci Lett*, *530*(1), 53-58. doi: 10.1016/j.neulet.2012.09.045
- Ramos-Loyo, J., Medina-Hernandez, V., Estarron-Espinosa, M., Canales-Aguirre, A., Gomez-Pinedo, U., & Cerdan-Sanchez, L. F. (2013). Sex differences in lipid peroxidation and fatty acid levels in recent onset schizophrenia. *Prog Neuropsychopharmacol Biol Psychiatry*, *44*, 154-161. doi: 10.1016/j.pnpbp.2013.02.007

- Raychaudhuri, S., & Altman, R. B. (2003). A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, *19*(3), 396-401. doi: 10.1093/bioinformatics/btg002
- Re, M., & Valentini, G. (2010). Noise tolerance of multiple classifier systems in data integration-based gene function prediction. *J Integr Bioinform*, *7*(3). doi: 10.2390/biecoll-jib-2010-139
- Richards, T. A., & Archibald, J. M. (2011). Cell evolution: gene transfer agents and the origin of mitochondria. *Curr Biol*, *21*(3), R112-114. doi: 10.1016/j.cub.2010.12.036
- Rodriguez, C., Mayo, J. C., Sainz, R. M., Antolin, I., Herrera, F., Martin, V., & Reiter, R. J. (2004). Regulation of antioxidant enzymes: a significant role for melatonin. *J Pineal Res*, *36*(1), 1-9.
- Rogler, C. E., Levoci, L., Ader, T., Massimi, A., Tchaikovskaya, T., Norel, R., & Rogler, L. E. (2009). MicroRNA-23b cluster microRNAs regulate transforming growth factor-beta/bone morphogenetic protein signaling and liver stem cell differentiation by targeting Smads. *Hepatology*, *50*(2), 575-584. doi: 10.1002/hep.22982
- Roke, Y., Buitelaar, J. K., Boot, A. M., Tenback, D., & van Harten, P. N. (2012). Risk of hyperprolactinemia and sexual side effects in males 10-20 years old diagnosed with autism spectrum disorders or disruptive behavior disorder and treated with risperidone. *J Child Adolesc Psychopharmacol*, *22*(6), 432-439. doi: 10.1089/cap.2011.0109
- Ruan, W., Xu, J.-m., Li, S.-b., Yuan, L.-q., & Dai, R.-p. (2012). Effects of down-regulation of microRNA-23a on TNF- α -induced endothelial cell apoptosis through caspase-dependent pathways. *Cardiovascular Research*, *93*(4), 623-632. doi: 10.1093/cvr/cvr290
- Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., . . . Lancet, D. (2010). GeneCards Version 3: the human gene integrator. *Database (Oxford)*, *2010*, baq020. doi: 10.1093/database/baq020
- Sanchez, A. M., Candau, R. B., Csibi, A., Pagano, A. F., Raibon, A., & Bernardi, H. (2012). The role of AMP-activated protein kinase in the coordination of skeletal muscle turnover and energy homeostasis. *Am J Physiol Cell Physiol*, *303*(5), C475-485. doi: 10.1152/ajpcell.00125.2012

- Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I., & Bork, P. (2006). Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*, 22(6), 645-650. doi: 10.1093/bioinformatics/bti597
- Sarlak, G., Jenwitheesuk, A., Chetsawang, B., & Govitrapong, P. (2013). Effects of Melatonin on Nervous System Aging: Neurogenesis and Neurodegeneration. *J Pharmacol Sci*.
- Scholkopf, B., & Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*: MIT Press.
- Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., & Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209), 58-63. doi: 10.1038/nature07228
- Sethupathy, P., Corda, B., & Hatzigeorgiou, A. G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12(2), 192-197. doi: 10.1261/rna.2239606
- Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., . . . Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8), 1034-1050. doi: 10.1101/gr.3715005
- Silverman, A. P., Kariolis, M. S., & Cochran, J. R. (2011). Cystine-knot peptides engineered with specificities for alpha(I**II**)beta(3) or alpha(I**II**)beta(3) and alpha(v)beta(3) integrins are potent inhibitors of platelet aggregation. *J Mol Recognit*, 24(1), 127-135. doi: 10.1002/jmr.1036
- Simpson, F., Peden, A. A., Christopoulou, L., & Robinson, M. S. (1997). Characterization of the adaptor-related protein complex, AP-3. *J Cell Biol*, 137(4), 835-845.
- Smith, L., Tanabe, L., Ando, R., Kuo, C.-J., Chung, I. F., Hsu, C.-N., . . . Wilbur, W. J. (2008). Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(Suppl 2), S2. doi: 10.1186/gb-2008-9-s2-s2
- Smith, S. C., Nicholson, B., Nitz, M., Frierson, H. F., Jr., Smolkin, M., Hampton, G., . . . Theodorescu, D. (2009). Profiling bladder cancer organ site-specific metastasis identifies LAMC2 as a novel biomarker of hematogenous dissemination. *Am J Pathol*, 174(2), 371-379. doi: 10.2353/ajpath.2009.080538

- Sonnenburg, S., Ratsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., . . . Franc, V. (2010). The SHOGUN Machine Learning Toolbox. *J. Mach. Learn. Res.*, *11*, 1799-1802.
- Sonnenburg, S., Ratsch, G., Schafer, C., & Scholkopf, B. (2006). Large Scale Multiple Kernel Learning. *J. Mach. Learn. Res.*, *7*, 1531-1565.
- Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguetz, P., . . . Mering, C. v. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, *39*(suppl 1), D561-D568. doi: 10.1093/nar/gkq973
- Tan, D. X., Manchester, L. C., Liu, X., Rosales-Corral, S. A., Acuna-Castroviejo, D., & Reiter, R. J. (2013). Mitochondria and chloroplasts as the original sites of melatonin synthesis: a hypothesis related to melatonin's primary function and evolution in eukaryotes. *J Pineal Res*, *54*(2), 127-138. doi: 10.1111/jpi.12026
- Torii, M., Hu, Z., Wu, C. H., & Liu, H. (2009). BioTagger-GM: a gene/protein name recognition system. *J Am Med Inform Assoc*, *16*(2), 247-255. doi: 10.1197/jamia.M2844
- Trenkmann, M., Brock, M., Gay, R. E., Michel, B. A., Gay, S., & Huber, L. C. (2013). Tumor Necrosis Factor α -Induced MicroRNA-18a Activates Rheumatoid Arthritis Synovial Fibroblasts Through a Feedback Loop in NF- κ B Signaling. *Arthritis & Rheumatism*, *65*(4), 916-927. doi: 10.1002/art.37834
- Van Wynsberghe, P. M., Chan, S. P., Slack, F. J., & Pasquinelli, A. E. (2011). Analysis of microRNA expression and function. *Methods Cell Biol*, *106*, 219-252. doi: 10.1016/B978-0-12-544172-8.00008-6
- Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Trans Neural Netw*, *10*(5), 988-999. doi: 10.1109/72.788640
- Vickers, K. C., Sethupathy, P., Baran-Gale, J., & Remaley, A. T. (2013). Complexity of microRNA function and the role of isomiRs in lipid homeostasis. *J Lipid Res*, *54*(5), 1182-1191. doi: 10.1194/jlr.R034801
- Voineskos, D., Rogasch, N. C., Rajji, T. K., Fitzgerald, P. B., & Daskalakis, Z. J. (2013). A review of evidence linking disrupted neural plasticity to schizophrenia. *Can J Psychiatry*, *58*(2), 86-92.
- Wallace, D. C. (2009). Mitochondria, bioenergetics, and the epigenome in eukaryotic and human evolution. *Cold Spring Harb Symp Quant Biol*, *74*, 383-393. doi: 10.1101/sqb.2009.74.031

- Walsh, M. T., Ryan, M., Hillmann, A., Condren, R., Kenny, D., Dinan, T., & Thakore, J. H. (2002). Elevated expression of integrin alpha(IIb) beta(IIIa) in drug-naive, first-episode schizophrenic patients. *Biol Psychiatry*, *52*(9), 874-879.
- Wang, F. E., Zhang, C., Maminishkis, A., Dong, L., Zhi, C., Li, R., . . . Miller, S. S. (2010). MicroRNA-204/211 alters epithelial physiology. *The FASEB Journal*, *24*(5), 1552-1571. doi: 10.1096/fj.08-125856
- Wang, J., Huang, H., Wang, C., Liu, X., Hu, F., & Liu, M. (2013). MicroRNA-375 sensitizes tumour necrosis factor-alpha (TNF-alpha)-induced apoptosis in head and neck squamous cell carcinoma in vitro. *Int J Oral Maxillofac Surg*, *42*(8), 949-955. doi: 10.1016/j.ijom.2013.04.016
- Wang, J., Zohar, R., & McCulloch, C. A. (2006). Multiple roles of alpha-smooth muscle actin in mechanotransduction. *Exp Cell Res*, *312*(3), 205-214. doi: 10.1016/j.yexcr.2005.11.004
- Wang, X., & El Naqa, I. M. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, *24*(3), 325-332. doi: 10.1093/bioinformatics/btm595
- Woods, K., Kegelmeyer, W. P., Jr., & Bowyer, K. (1997). Combination of multiple classifiers using local accuracy estimates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *19*(4), 405-410.
- Wrighton, S. A., & Stevens, J. C. (1992). The human hepatic cytochromes P450 involved in drug metabolism. *Crit Rev Toxicol*, *22*(1), 1-21. doi: 10.3109/10408449209145319
- Xu, B., Ionita-Laza, I., Roos, J. L., Boone, B., Woodrick, S., Sun, Y., . . . Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet*, *44*(12), 1365-1369. doi: <http://www.nature.com/ng/journal/v44/n12/abs/ng.2446.html#supplementary-information>
- Xu, L., Krzyzak, A., & Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *Systems, Man and Cybernetics, IEEE Transactions on*, *22*(3), 418-435.
- Yang, Y. C., Ma, Y. L., Liu, W. T., & Lee, E. H. Y. (2011). Laminin-[beta]1 Impairs Spatial Learning through Inhibition of ERK/MAPK and SGK1 Signaling. *Neuropsychopharmacology*, *36*(12), 2571-2586. doi: <http://www.nature.com/npp/journal/v36/n12/supinfo/npp2011148s1.html>

- Yeh, A., Morgan, A., Colosimo, M., & Hirschman, L. (2005). BioCreAtIvE Task 1A: gene mention finding evaluation. *BMC Bioinformatics*, 6(Suppl 1), S2.
- Yu, D. S., An, F. M., Gong, B. D., Xiang, X. G., Lin, L. Y., Wang, H., & Xie, Q. (2012). The regulatory role of microRNA-1187 in TNF-alpha-mediated hepatocyte apoptosis in acute liver failure. *Int J Mol Med*, 29(4), 663-668. doi: 10.3892/ijmm.2012.888
- Zhiling, Y., Fujita, E., Tanabe, Y., Yamagata, T., Momoi, T., & Momoi, M. Y. (2008). Mutations in the gene encoding CADM1 are associated with autism spectrum disorder. *Biochem Biophys Res Commun*, 377(3), 926-929. doi: 10.1016/j.bbrc.2008.10.107
- Zuker, M., & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1), 133-148.

BIOGRAPHY

David H. Millis was born in Brooklyn, New York, and attended Stuyvesant High School in New York City. He received a Bachelor of Science degree in biology and psychology from Yale University in 1979 and a Doctor of Medicine degree from Howard University in 1983. He completed an internship in surgery at University of Illinois in Chicago, Illinois in 1984, a residency in psychiatry at State University of New York Health Science Center in Brooklyn, New York in 1988, and a fellowship in administrative psychiatry at University of Maryland in Baltimore, Maryland in 1995. He achieved certification by the American Board of Psychiatry and Neurology in 1993. He completed a Master of Science in Medical Information Sciences at Stanford University in 1992, a Master of Business Administration with a concentration in Information Systems at University of Maryland College Park in 2000, and a Master of Science in Biotechnology Studies with a concentration in Bioinformatics at University of Maryland University College in 2006. He achieved certification by the American Society of Clinical Psychopharmacology in 2012. He received the degree of Doctor of Philosophy in Bioinformatics and Computational Biology from George Mason University in 2014.

Dr. Millis currently serves as the Clinical Director of the Thomas B. Finan Center, a psychiatric hospital in Cumberland, Maryland. The Finan Center is operated by the Department of Health and Mental Hygiene, and provides mental health services to residents of the State of Maryland who suffer from severe and persistent mental illness.