ONTOLOGY BASED DATA INTEGRATION (OBDI): A PIPELINE TO INTEGRATE AND MODEL HIGH DIMENSIONAL DATA

by

Sharanya Raghunath

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

December 2014

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Sharanya Raghunath**

has been approved by the following supervisory committee members:

Lewis J Frey	, Chair	**04/21/2014** Date Approved
Nicola J Camp	, Member	**04/21/2014** Date Approved
Karen Eilbeck	, Member	**04/21/2014** Date Approved
Elena Y Enioutina	, Member	**04/21/2014** Date Approved
Alun Thomas	, Member	**04/21/2014** Date Approved

and by **Wendy W Chapman** , Chair/Dean of

the Department/College/School of **Biomedical Informatics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Gene expression data repositories provide large and ever increasing data for secondary use by translational informatics methods. For example, Gene Expression Omnibus (GEO) houses over 37,000 experiments with the goal of supporting further research. To use these published results in a larger meta-analysis, consolidation of the data are needed; however, the data are largely unstructured, thus hindering data integration efforts. Here, I propose the use of a novel pipeline, Ontology Based Data Integration (OBDI), which uses an ontological approach to combine the samples across multiple GEO experiments. The ODBI pipeline uses machine learning algorithms that permit researchers to consolidate and analyze data across GEO experiments.

Here, I demonstrate how using an ontological approach to integrate samples across experiments can be used to explore the immune response at a molecular level. As part of this process, a Web Ontology Language (OWL) was developed for each data platform used. OWL serves as a core component in successfully processing different sample types. Immunological experiments from GEO were consolidated to evaluate this methodology. The experiments included samples analyzed on expression arrays, BeadChips, and sequencing technologies. The integration of a complex biological system and the incorporation of different biological data types will validate the potential of OBDI.

The nature of biological data is highly dimensional. OBDI incorporates tools and techniques that can handle the analysis of various biological data. The machine learning analysis performed within the OBDI pipeline successfully evaluated the newly annotated experiments and provides insights that can be further explored.

The OBDI pipeline can help researchers annotate experiments using ontologies and analyze the annotated experiments. To successfully build the pipeline, ontologies served as the backbone of integrating samples from GEO Series records into machine learning experiments using ML-Flex. By using the OBDI pipeline, researchers can access the uncurated experiments from GEO (GEO Data Series) and annotate the data using the terms in the ontologies. This mechanism allows for the organization of data sets in relationship to new experiments independent of GEO's GDS curation process. The OBDI system allows ontologies to grow organically around a cluster of experiments. These experiments are then further analyzed in ML-Flex using machine learning algorithms. The curated experiments are analyzed in silico and the computational analyses are supported by the OBDI ontological system.

To Aniket, Mom, and Dad

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

My journey at the Department of Biomedical Informatics started in Fall 2008. I was fairly confident that I wanted to work in the subdomain of bioinformatics; however, finding a project was challenging. I moved to Utah from Georgia, with a degree in genetics and the programing background of a bench researcher. Despite my varied background, I was determined to be a part of a growing field, and biomedical informatics seemed to be a perfect fit. The cohesive and collaborative nature of this department has made me cherish my learning experience. Having supportive and patient mentors, helpful colleagues have added to my growth as a graduate student.

I would like to thank my wonderful spouse, Aniket Surdi. His support and pragmatic solutions have helped overcome various challenges faced during the course of my degree. His ability to listen patiently through my problems has put me at ease during several occasions. Aniket's cooperation and support has helped me keep a balanced outlook throughout my graduate career.

I would like to thank my parents who have helped me achieve my personal goals. My parents have made personal and financial sacrifices in order to provide me with a wonderful education and help me build strong professional values. My father has taught me how to face difficult situations with strength and composure. My mother has always guided me through my problems by listening and advising only when

appropriate. Their belief in my abilities have helped me triumph over the hurdles faced during my graduate degree.

Next, I would like to acknowledge and thank my committee chair, Dr. Lewis Frey. He has made himself available on a weekly basis to guide me through my dissertation process. Dr. Frey was also open to collaborating with Dr. Elena Enoutina in order to explore an immunological problem from an informatics point of view. His ability to keep an open mind has played a key role in my ability to form a well-integrated informatics project. In the course of my dissertation, Dr. Frey has challenged me to think of innovative solutions to tackle problems. He has always promoted open dialogue and has played an important role in helping me achieve my goals.

I would like to appreciate the endless support of my committee members: Drs. Nicola Camp, Elena Enoutina, Karen Eilbeck, and Alun Thomas. They have offered me advice while executing and designed an integrated project that explores the complex domains of immunology, informatics, ontologies, and statistics.

Barry Moore from Dr. Eilbeck's lab helped me understand the process required to perform RNA-Sequencing analysis. I would like to thank Sean Igo, who has helped me master my programing skills that allowed me to build my pipeline successfully. Finally, I would like to thank the endless support provided by my friends at the department of biomedical informatics.

1. INTRODUCTION

Performing differential gene expression analysis provides insight into the biomolecular mechanisms that play a role in cellular process that are explored in both laboratory and clinical research. Gene expression data from microarray experiments and sequencing experiments are available in public repositories. These repositories allow researchers to upload their experiments, making the raw data, the metadata and the associated journal publications available to the scientific community. There is tremendous potential for novel discoveries that can be found by integrating studies from public repositories and performing meta-analysis on the integrated data sets. Current efforts are focused on increasing the data stored to public repositories; however, the efforts to make data usable across experiments is lacking. With an exponential increase of data being generated, there is a need to improve the ability to combine data in order to identify novel findings that were not possible with smaller data sets or were not the focus of original experiment.

When integrating data, it is important to take into account the biological and/or the clinical complexity associated to specific samples being combined. Although repositories store the biomedical metadata associated to each study, there is a lack of standardized vocabulary that is used to describe metadata; hence, parsing the metadata are not sufficient for tackling biological complexity during data integration. The variability in metadata occurs since repository submissions come from different

laboratories and users are not provided with a standardized way to enter biological information [1-3]. An ontological approach is one way of handling metadata inconsistencies. Ontologies provide a standardized way of storing the metadata, and making the information machine readable [4].

Once the data are integrated, the meta-analysis of the newly integrated set can lead to new insights. Performing differential expression analysis on integrated data sets is an effective way of exploring how a set of genes, are regulated differently across a given set of variables. Genes can be further clustered based on their biological responses they share to treatments, diseases, and different time points. The set of variable can be classified using a data driven approach. Machine learning algorithm can be used to classify variables or cluster genes into different groups. Machine learning focuses on prediction that is based on previous knowledge or on new information learned from a subset of the curated data set. Analyzing integrated data sets can lead to a more comprehensive model because integration across studies increases the number of variables that are being assessed.

In this study, a novel pipeline is proposed that integrates samples from a public repository while maintaining consistent representations to generate novel data sets. The analysis of these data sets can lead to generating hypotheses with targeted features that can be tested in a laboratory setting. The ontological representations in this pipeline can be re-used, as more samples are available in the public repositories. The consistent representation that can be achieved through ontologies allows for modeling high dimensional data to find new insights that generate experiments using a hypothesis driven approach. The analytical protocol is stored in ontological representations,

allowing users to modify the protocol within the ontologies. This unique feature in the proposed pipeline can help reproduce the results in this study. To validate the pipeline, gene expression data from laboratory experiments are integrated and used to generate hypotheses that explore a specific biomolecular mechanism. The results generated from the newly integrated data sets can be translated to the bench in order to help further biomedical research. The pipeline provides an example of how research in translational science can transform an in silico model to aid hypothesis generation that may be applied at the bench.

2. DISSERTATION AIMS AND OBJECTIVES

This dissertation proposes and implements an ontology dependent pipeline to extend the field of data integration. This pipeline, called Ontology Based Data Integration (OBDI), helps solve major hurdles that occur when integrating data across Gene Expression Omnibus (GEO) [5-7] experiments. The information in experiments is incorporated into ontologies; therefore, aiding the evolving nature of new submission and meta-analysis being performed. New integrated data sets created through OBDI can provide new information and potential targets. The novel OBDI pipeline will provide a means to increase sample size and offers mechanisms to build a more comprehensive view of knowledge domains. OBDI extends biological information using ontologies to facilitate the exploration of complex biological spaces by annotating novel experiments.

Using ontologies, the OBDI pipeline is built such that: [1] data are downloaded from GEO, [2] reorganized into biologically relevant machine learning experiments that [3] exist outside the GEO framework, and [4] finally, these novel experiments are analyzed using machine learning algorithms. By integrating the ontologies, workflow, and the results, a comprehensive model is generated. With the methodologies incorporated in OBDI a comprehensive model can be replicated, examined, modified, and extended with additional knowledge.

The primary motivation of this project was to generate a comprehensive model representation that is supported by the OBDI pipeline. The methodologies in OBDI are replicable and adding more GEO Series records can increase sample size in the annotated

experiments. Other laboratories can replicate the described methodology in order to successfully integrate GEO experiments, and thoroughly evaluate a biological space from a data driven approach. The OBDI pipeline will allow informaticians to curate new experiments from preexisting GEO experiments thus adding new knowledge that may help move research at the bench.

Aims

The components of the OBDI pipeline combine high throughput samples that have never been integrated in previous experiments, allowing users to generate a hypothesis driven immunological model. This integrated model will support the bench researcher in exploring how the immune system interacts at a molecular level. Based on the motivation and the objectives described in the previous sections, the following research aims were evaluated.

- Aim 1: Develop an Ontology Based Data Integration (OBDI) pipeline using the Java programming language for preprocessing, integrating and performing meta-analysis on high throughput data from GEO
- Aim 2: Evaluate four annotated experiments, generated by integrating GEO experiments, and perform machine learning analysis as implemented in ML-Flex
- Aim 3: Extend machine learning results in AIM 2 by generating a well-integrated model for bench researchers by incorporating prior knowledge that is gathered using literature review and from databases that contain biological pathway information

- Aim 4: Generalize predictive model from Experiment 1 to other GEO experiments that explore the immune system response in order to create a more comprehensive model that could be applied in a laboratory setting

The purpose of Aim 1 is to build a pipeline that allows the integration of GEO experiments and various tools to analyze high throughput data. Ontologies are developed to unify the vocabulary used to describe GEO experiments. Each annotated experiment is associated with the respective ontology and the various elements of the experiment are stored as OWL entities. Multiple data sets are related to generate curated experiments constrained by specified prior knowledge encoded within the ontology. Integrating samples using ontologies allows researchers to not only replicate this approach, but also extend upon existing ontologies. More biological and clinical information from GEO can be added to the ontologies; therefore reducing the barrier to curate newer data sets. Reasoning over the ontologies promotes the integration of samples and analysis of curated experiments. Using ontologies within the OBDI pipeline allows maintaining consistency as more samples are added to GEO and other public repositories. By building the ODBI pipeline, I will [1] resolve some complexities around data integration and [2] successfully build in silico experiments, outside the GEO framework, that can be analyzed using machine learning algorithms.

In Aim 2 machine learning algorithms are implemented to perform analysis on four experiments. For the analysis component, ML-Flex is incorporated within OBDI. To perform in silico analysis, ML-Flex is provided with a configuration file that contains analytical protocols. Information needed to perform in silico analysis is stored as annotation properties within the ontology related to the specific experiments. Using

ontological representations to store the analysis component allows users to reproduce the results, change parameters, and maintain consistency when adding more samples to the curated experiments.

Aim 3 includes biological information retrieved from different sources that extends the in silico analysis by incorporating biological information. The genes in the predictive model are grouped together based on preexisting knowledge of biological pathways. This knowledge is gathered by various sources; however, the pathway information primarily comes from the knowledge stored in the Kyoto Encyclopedia of Genes and Genomes (KEGG). A heatmap is used to display the comparative model and the expression of a subset of individual genes is depicted with boxplots. Visualizing the results of the machine learning analysis provides insight into how certain immune system response genes are differentially expressed. By integrating samples that were previously analyzed in silos, I am able to generate results that were overlooked due to the lack of combing similar samples into a single experiment. Finding biological relevance in newly combined experiments depicts how OBDI can generate novel insights to be validated at the bench.

In Aim 4, I show how OBDI is used to generate a comprehensive hypothesis that can be tested using data from other cell types and clinical experiments. Focusing the analysis on samples related to a specific research domain validates the OBDI pipeline. Multiple in silico experiments are analyzed to explore the biomolecular interactions in the field of cancer immunotherapy. To generate a comprehensive model, queries in GEO had to be setup in a logical manner. Incorporating more experiment from GEO that are relevant to the field of cancer immunotherapy was possible because OBDI supports

hypothesis generation through ontological representations and meta-analysis of integrated experiments. By incorporating more experiments from GEO, I am also able to design OBDI as a flexible framework that incorporates other data types, such as, sequencing data from NGS experiments. In AIM 4, OBDI is used to derive models from in silico analysis and also extends integration methods by incorporating other data types.

The goal of creating the OBDI pipeline is to study a complex environment such as immune system response in order to generate new knowledge and help bench researchers design targeted experiments. With the use of ontologies, samples across GEO experiments are annotated to generate new experiments that support research in translational science. Experiments analyzed in this dissertation using OBDI support research at the bench and may lead to new findings that can help improve immune system response in cancer immunotherapy.

3. BACKGROUND

Biomedical data can be successfully stored in online repositories; however, the backlog of integrated data sets impedes the research in translational science. The expanding sources of data offer new opportunities for discoveries and validation of previous research. Researchers can archive their experimental data samples by submitting them to a data repository. In addition to the raw data, a researcher can include publications and metadata associated with the experiment. Submitted data can be kept private until the experimental results are published in a manuscript, after which the data are made available for public use [5]. Making experimental data publically available via open repositories allows for furthering research by querying submissions and reusing the data.

Some of the commonly used data repositories are: ArrayExpress, Sequence Read Archive (SRA), and Gene Expression Omnibus (GEO) [5-7]. ArrayExpress is a database housed at the European Bioinformatics Institute (EBI) that stores gene expression microarray data and high throughput sequencing data [6]. SRA is managed by the National Center for Biotechnology Information (NCBI) and primarily stores raw Next Generation Sequencing (NGS) data [5]. GEO is housed at NCBI and contains a vast range of data, including, gene expression, SNP arrays, protein arrays and NGS [7]. The curators of these three databases collaborate with each other to share data structures and protocols for archiving raw expression and sequencing data.

This study focuses on retrieving and organizing high-throughput data across different experiments, such that, it allows new development in the field of basic science and medicine. Samples integrated in this study are constrained to a single repository, GEO.

Gene Expression Omnibus (GEO)

GEO is a large database that stores over 1 million samples from high throughput experiments. When submitting data to GEO, researchers can use several formats to enter information related to their experiment. The submission format options include, spreadsheets, Simple Omnibus Format in Text (SOFT) and Extensible Markup Language (XML). Spreadsheets work best when researchers want to quickly describe their study. The spreadsheet contains the metadata and the spreadsheet is bundled with the raw data files for submission. This submission method is recommended for most users. If the data are already in the GEO database, the SOFT or XML format is recommended for submission. SOFT is a line based plain text format that can be readily generated from common database applications, such as MySQL [7, 8].

In GEO the submitted data are stored as a GEO Series (GSE) record and are assigned a GSE accession number. GEO Series records may contain samples that are analyzed on different platforms and vary in their experimental properties. The staff at GEO reassembles the original submissions into curated GEO DataSets that are given a GEO DataSet (GDS) accession number. Samples identified by a single GDS number are analyzed on the same platform, and share similar array elements. This makes organization, normalization, and processing equivalent across samples within a DataSet. Within a DataSet or a Series record, the platform information and samples are assigned

different identifiers (GPLxxx and GSMxxx, respectively). Both GSE and GDS are searchable using the GEO interface; only the GDS can be used to perform GEO's advanced data display and analysis tools. Using these tools for the curated DataSets, a researcher can query gene names, visualize charts and perform clustering analysis. Meta-analysis within GEO includes differential expression analysis performed using GEO2R. To identify genes that have a similar response pattern, hierarchical clustering can be used within GEO. Due to the large volume of data submissions and the inability to reassemble all Series records, there is a backlog of converting original submissions into curated DataSets. For instance, in Figure 3.1 only 64,919 samples are curated into GEO DataSets [7], leaving 992,559 uncombined samples that cannot be analyzed using GEO based tools (GEO2R, Hierarchical Clustering).

Once the experiments are submitted, the MIAME (Minimum Information About a Microarray Experiment) Notation in Markup Language (MINiML) formatted files can be downloaded to access metadata. MINiML is a data exchange format used in GEO that captures the minimum information required when describing a high throughput experiment, in addition to any other information supplied by the submitter [7]. To analyze high throughput data at a genomic level different manufactures, such as Affymetrix, and Illumina, have developed competing technologies [8]. In expression analysis, individual assays differ in the type of probes being used (cDNA, oligonucleotide size, probe I.D, etc.), the hybridization methodology (specific versus nonspecific) and the labeling method (direct fluorescence, indirect antibody, etc.). The data generated from current microarray technologies are comparable, especially when mining for genes that are differentially expressed. When comparing gene information between two platforms it

is important to be aware of the differences between probe annotations [9]. When integrating samples across experiments, it is important to keep a record of the platform assay used to analyze the samples [1]. The metadata in GEO has XML tags that contain platform information.

Each sample in GEO has associated metadata that stores specific information about the samples, including how the samples were treated in the laboratory setting to generate biological outcomes. The associated metadata plays a crucial role in the integration of data across studies. GEO DataSets and Data Series contain metadata stored in parsable XML format. Specific XML tags that are relevant while integrating data across experiments are retrieved and stored in consistent representations. However, the content within the XML tags are user defined and do not require users to follow consistent vocabulary. Drop down menus with standardize laboratory terms are not provided, causing increased variability in the content encompassed between XML tags. For example, the content of the XML tag, *Title*, requires users to define the treatment used on their samples in the laboratory experiment. For Sample A, the user may define a control sample as "B305 immature DCs without IFN alpha treatment." For Sample B, the user may define a control sample as "iDC_6h." Simply parsing the content for the XML tag, *Title*, would not provide information that Sample A and Sample B are controls and that they contain a population of untreated cells. Similar to the above examples, content between XML tags are stored using lengthy phrases that are not machine readable, making data integration an arduous task.

Due to the inconsistencies described in this section, integrating across GEO experiments is challenging. Although the GEO staff combines Series Data into curated

DataSets, only 6% of the samples have been combined, creating a huge backlog. The methods in this dissertation show how combing previously overlooked samples can generate models that are supported by in silico analysis.

Analysis Pipeline: Gene Pattern

Gene Pattern is a web service tool that is hosted by the Broad Server. Gene Pattern can be used to analyze different types of genomic data including differential expression analysis on data generated from microarray experiments and NGS [10]. There are modules within Gene Pattern that interact with GEO in order to acquire samples from GDS and GSE experiments. The interaction between Gene Pattern and GEO makes data acquisition a simpler process.

Once data are available in Gene Pattern, analysis pipeline can be created using the Pipeline Designer. The Gene Pattern Pipeline Designer allows users to create workflows that can start and finish analyses without breakpoints between each processing component. The different analysis modules in Gene Pattern can be connected to show the flow of data through the pipeline. The output from one module can serve as an input to the following module. The Pipeline Designer can be used to track analysis of each module and retrieve individual results when applicable [10]. The Pipeline Designer is useful when performing concurrent processing; however, to successfully integrate GEO experiments, samples must be annotated. A consistent way to annotate samples or represent metadata does not exist in Gene Pattern. The ontological representations used in this dissertation solve the missing annotation component in Gene Pattern. Through the use of ontologies, GEO samples can be annotated with the corresponding biological of clinical information.

Although Gene Pattern does not have a methodology to create standardized representations, there are several modules that aid in preprocessing and analyzing high throughput data. Samples from high throughput data can be classified or clustered using machine learning algorithms. Classification methods can be used to create a model, which can be used to predict and classify samples with unknown class variables. Similar to clustering methods in GEO, Gene Pattern has modules that perform clustering analysis by grouping genes that share similar expression patterns. The use of machine learning classification techniques can generate predictive models that may provide additional insight into experimentally generated data. Machine learning algorithms can be used to analyze various data sets and classify gene expression data based on different phenotypes.

Data Analysis: Techniques and Tools

Machine learning analysis to evaluate differential expression can be performed using tools such as Weka, R Statistical Package, and ML-Flex. Weka is software with an extensive collection of algorithms where analysis can be performed using classification and clustering methods. R is primarily a statistical tool that can be used to perform machine learning analysis and statistical tests. R can be used to create heatmaps, which is the standard way of visualizing results generated from differential expression analysis. ML-Flex is a tool that implements machine learning algorithms, independently and algorithms from other tools, such as, Weka, and R.

ML-Flex

ML-Flex is an open source tool that is a wrapper to a suite of third party machine learning software, such as, Weka, Orange, R Statistical Software, and C5.0 Decision Trees [11-14]. Analysis in ML-Flex can be used to aid biomedical research by generating predictive models for biological and clinical data. Different classification techniques can be compared within a single ML-Flex experiment. ML-Flex allows computationally intensive algorithms to be performed in parallel. It is a command line tool written in the Java programming language that organizes machine learning analyses into an experiment-based framework. In addition to analyzing data, ML-Flex keeps track of the various settings used in a particular experiment [15].

To run analysis in ML-Flex, an experiment file must be created. This file must contain the following information in order to successfully analyze the data: location of input data, classification algorithm being used, and cross validation methods. Other relevant fields can also be entered, such as feature selection variables, algorithms to perform feature selection, samples used for training and samples used in the validation set [15]. See Appendix A.

Machine Learning Algorithms Used to Analyze High-Throughput Data

A machine learning approach allows for the use of experiment driven analyses that may provide new insight on samples that may not be aggregated in GEO or other repositories. Using machine learning methods can help separate genes that are differentially expressed. Machine learning algorithms can help identify genes that have an increased level of gene expression, compared to genes that have a decreased level of expression [16].

Microarray results contain millions of probe IDs; therefore, the classification algorithms used need to be capable of handling a large number of features (i.e., probes). Classification algorithms help identify whether samples within a newly observed population belong to observations that have been previously made. Three major classification algorithms used are discussed below.

Naïve Bayes is a classification algorithm that relies on the assumption that each attribute is independent of the other and it helps estimate the conditional probabilities of predicted classes in a given experiment. The performance goal of a Naïve Bayes classifier is to accurately predict the class of test instances in which the training instances contain class information. One limitation of using Naïve Bayes is that it assumes independence between each attribute, which is not usually satisfied by the use of gene expression data. Although the independence assumption is typically violated when analyzing gene expression data, the Naïve Bayes is robust if the rank order is maintained between classes [17]. Naïve Bayes is a simple classifier that has shown to perform well when combined with feature selection methods [17-19].

The Decision Tree algorithm is a graph that uses a branching method to create a predictive model. The goal of this learning method is to create a predictive model based on several input variables [12, 20]. Decision Tree chooses a feature that partitions the training data and then partitions the remaining data recursively until no further partitions can be made. Finally, the tree is pruned to mitigate overfitting [12, 21].

Support Vector Machine (SVM) is a linear classification algorithm that separates two subsets of data with the largest margin by constructing a high dimensional boundary, commonly referred as a hyperplane. SVM can be used to handle high dimensional data,

such as genomic data used to analyze biological samples [19, 22]. SVM works well in dealing with large feature sets and is able to successfully identify outliers. SVM is capable of using prior knowledge information in terms of training data to make distinctions between two different instances [23, 24].

A significant problem with high throughput data are that the numbers of features (genes) greatly exceeds the number of instances (samples). To manage this problem, feature selection can be used to reduce the number of features. The ReliefF algorithm has been successfully used as a feature subset selection method [25, 26]. The algorithm does not assume independence between features and selects instances by giving more weight to features that can distinguish between classes. This allows for the estimation of attributes according to how their values differentiate between samples that are close to each other [25]. The ReliefF algorithm is not limited to two class experiments and is able to handle noisy data. The performing feature selection has successfully been applied to microarray data, leading to the selection of informative cancer genes [27].

In a study performed by Wang et al. [27], feature selection methods were compared in three cancer related data sets: ALL/AML leukemia, MLL leukemia, and colon cancer. Along with ReliefF three additional feature selection methods were used to compare the data sets: Information gain, Gain ratio, and X^2 statistic. Information gain is a type of feature selection that can be used to define a set of attributes that accurately build a predictive model. Gain ratio is a modification of information gain and reduces its bias by taking the size of branches into account using a Decision Tree algorithm. The X^2 statistic method evaluates each feature based on the X^2 statistic of each feature compared to the classes. The results from this study show that the ReliefF algorithm performed

better than other methods when analyzing the ALL/AML Leukemia data [27]. In another study performed by Kononenko et al. [28], ReliefF, along with other versions of Relief, were used to analyze artificial and primary tumor samples. This study clearly showed that Relief-F was better at estimating attributes even if the data are noisy or incomplete [28].

Ontologies

GEO has a process of integrating samples into curated DataSets; however, there is a backlog of transforming original submissions into DataSets. Developing ontologies can aid the process of integrating related samples that have been overlooked by GEO. A consistent standard using ontologies can not only help data integration but also help perform analysis on the newly curated data sets. The ontologies developed for this dissertation were used to organize multiple GSE experiments from GEO into novel experiments that can be analyzed in silico. Using ontological components, various attributes of an experiment, such as, platform information, sample IDs, clinical or biological relevance are represented.

Ontologies are a standardized way of representing knowledge in a particular domain. They are composed of consistent representations that can be used to structurally organize various entities extracted from the metadata. Ontologies can be shared with other researchers to improve the consistency of knowledge in a given domain [4, 29]. Storing metadata from GEO as an ontological entity makes the information associated with each GEO experiment computationally tractable. Repositories, like GEO, provide metadata information regarding biological experiments; however, specialized knowledge is required to parse the metadata and organize the raw data in a logical manner. To help

resolve the complexity of data integration, ontologies can be developed to store metadata information that describe experimental properties of GEO Series records as ontology components that are machine readable. Once the knowledge domain is organized in relationship to classes and properties, logical reasoning can be performed upon the representations.

Ontologies are used in the field of biomedical informatics to solve different problems such as: unifying vocabulary across knowledge domains, consolidating and supporting common data formats, creating inferences on asserted ontologies, and driving natural language processing [4, 29, 30]. Ontologies can be created as reference ontologies and as application ontologies. Reference ontologies focus on theoretical knowledge that demonstrate a larger knowledge domain, whereas application ontologies focus on a smaller knowledge domain and solving a specific problem using the smaller knowledge representation. Reference ontologies such as the Gene Ontology (GO) and the Foundational Model of Anatomy (FMA) have unified the language surrounding their respective domain [31, 32]. They have defined the terms and relationships between terms and provide a hierarchical basis for annotation of data. Application ontologies can often use portions of a reference to represent a knowledge domain [30]. For instance, the Experimental Factor Ontology (EFO) supports the analysis and data handling of various experimental variables stored in EBI (European Bioinformatics Institute) databases [4].
In this dissertation, application ontologies are built to provide a structured way of relating complex concepts to create a cluster of experiments that are newly generated by using the ontologies.

Ontology Development

Ontologies can be created using a Graphical User Interface (GUI), such as, Protégé, OBO (Open Biomedical Ontologies)-edit, etc. [33-35]. A GUI allows users to create ontologies using visual descriptors versus a text-based interface. Protégé has an intuitive GUI that allows the users to edit and develop ontologies. The ontology files can be imported and exported in various formats: Web Ontology Language (OWL), Extensible Markup Language (XML), and Resource Description Framework (RDF) [33]. Another GUI tool commonly used to create ontologies is OBO-Edit. The ontologies created using OBO-Edit contain the OBO format that is composed of stanzas describing the various entities of an ontology [35, 36]. The ontologies in this dissertation are created in OWL, which can easily be parsed using the Java programming language.

To successfully build ontologies using GEO experiments, the metadata for each GEO experiment needs to be parsed, extracting the content from specific XML tags. Using the extracted content, a base ontology is created using four ontological components: classes, individuals, object properties, and annotation properties. Next, relationships between appropriate entities are constructed using *is-a* relationships and user defined object properties. The use of annotation properties helps store the descriptive entry related to how each sample is treated in the laboratory. Logical definitions may be added to the components in order to accurately define the knowledge space within the ontology. Building the base ontology helps reduce the amount of implicit information. A reasoner is a complex algorithm used in ontologies to generate logical consequences from a set of asserted terms in the base ontology [37]. HermiT is the most commonly used reasoner to check for inconsistencies within ontologies, to infer

relationships between OWL entities, and to classify various OWL classes [38]. The detail on how the ontologies are created is described in Chapter 5, (OBDI, A Novel Pipeline).

Incorporating Prior Knowledge Information

When analyzing biological data, applying prior knowledge information about a biological domain plays a crucial role in assessing a newly developed predictive model. To decipher complex biological systems, it is important to incorporate the diverse data types along with preexisting knowledge of well-understood biological systems [39]. Selecting features based on domain specific, prior knowledge can have a positive effect on the performance of a model. When analyzing high dimensional data, such as microarray, incorporating prior knowledge can positively affect the performance of the machine learning algorithms [40].

Domain specific prior knowledge information helps evaluate the machine learning results in order to determine biological meaning [39]. It has already been established that feature selection is an important method to reduce the number of probe IDs to a manageable group. In order to assess domain specific knowledge, further investigation must be performed on the features selected using machine learning analysis [40]. Research performed at the bench often requires preexisting knowledge in order to make inferences about the current investigation [41].

Clustering Genes via Known Biological Pathways

Several approaches can be used that cluster genes into biological relevant sets such as KEGG and Protein ANalysis THrough Evolutionary Relationships (PANTHER),

and Database for Annotation, Visualization and Integrated Discovery (DAVID) [42]. The KEGG database contains extensive information about biological pathways [43]. The biological pathways in KEGG represent a network of knowledge that contains biomolecular interactions. KEGG can be used to query specific genes, and explore how these genes play a role in specific biological pathways. The KEGG database contains the most comprehensive information about biological pathways across different species [43]. The PANTHER classification system also consists of a large number of pathways that can be used to cluster a set of features (genes or probe IDs). The advantage of using PANTHER is that genes can be passed in a batch file and PANTHER associates the genes to corresponding pathways. PANTHER also supports various ID formats, which allows the application of a wide range of high throughput analyses [42]. DAVID bioinformatics consists of a collection of analytical tools and biological knowledge base targeted at extracting information from large lists, including genes, proteins, and probe ids. DAVID also had visualization capabilities that help visualize genes on KEGG pathway maps [44].

Biological relevance can be examined using other methodologies that do not involve the interaction of biomolecular components. GoMiner is a tool that uses GO to identify meaningful biological information in genomic data. GoMiner supports classification of genes according to biological process, cellular components, and molecular functions. When assessing microarray data with GoMiner, the user can determine whether specific genes are upregulated or downregulated. GoMiner also links specific genes to external biomolecular pathways [45].

Incorporation Prior Knowledge Using Ontologies

With the use of ontologies, biological or clinical knowledge can be communicated among researchers working in the same domain knowledge. Ontologies help share newly discovered knowledge within a research community[41]. Preexisting ontologies can be used to build larger ontologies, thereby extending the domain knowledge of the original ontology. Furthermore, biological data stored in databases require the incorporation of additional domain specific knowledge that may be required for analysis and interpretation of the data [41]. In the field of biomedical research, the cell ontology is commonly used to incorporate a structured vocabulary of various cell types. The plant, animal, fugal, and prokaryotic kingdoms are included in this ontology. The cell ontology also includes cells in their native state and those that are experimentally modified. Within the ontologies created in this study, the cell ontology is extended to support the different in vitro cell types that are retrieved from the metadata of specific GEO experiments. The structured vocabulary of cell types, along with newly added information, helps facilitate the interoperability among databases that house high throughput data [46, 47].

The tools described above can be used to successfully incorporate domain specific prior knowledge; thus, checking for biological relevance in the machine learning results. By incorporating domain specific knowledge, meaningful models for bench researchers can be generated [42, 43, 45].

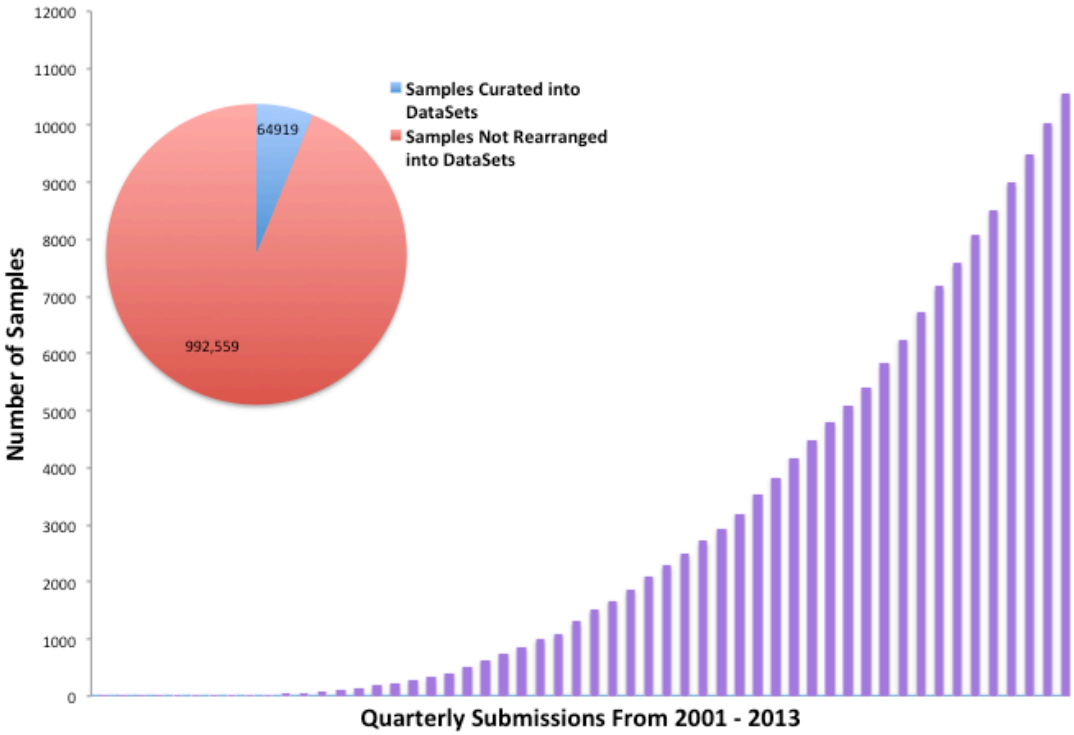

Figure 3.1: The bar graph depicts the number of samples that are being added quarterly from 2001 through 2013. The pie chart shows the number of samples that are curated into GEO DataSets.

4. OVERVIEW OF CANCER IMMUNOTHERAPY

Cancer immunotherapy is a treatment method that harnesses a patient's immune system cells to treat malignancies. The primary goal of cancer immunotherapy is to stimulate the patient's immune system response in order to fight tumor cells. Immunotherapy vaccines can either work by boosting the general immune system response or by training the immune system to attack specific cancer cells [48]. The field of immunology contains several cell types and pathways that interact to generate an effective immune system response. To focus on a single problem in the field, an ontology driven approach is used to represent the domain knowledge and expand upon the cell ontology. The ontology driven approach also allows for the communication of knowledge amongst researchers [41], allowing for the development of cancer immunotherapeutics. Currently, the success of cancer immunotherapy is as an adjuvant to other cancer therapeutics. Certain chemotherapeutics appear to enhance the effect of cancer vaccines, by increasing the T cell mediating response against tumors [49]. Immunotherapy treatments are designed to empower the patient's immune system, creating a prolonged antitumor response. Although efficacy of immunotherapy vaccines has improved, there remain challenges in developing successful clinical assays to monitor immune responses in patients [50]. Cancer immunotherapy is an active field of research; hence, investigating unexplored GEO experiments can lead to potential discoveries.

Among the various immunotherapy methods, Dendritic Cell (DC) vaccines are a newly emerging form of cancer vaccines. DC vaccines are meant to harness the body's immune system response in order to fight tumor cells by initiating a CD8+ antitumor T cell response [51]. Due to the recent advancements in cancer therapeutics, DC based vaccines have shown promising results for initiating an antitumor immune response in melanoma, prostate cancer, glioblastoma, and non-Hodgkin lymphoma [51-53]. DCs are immune system cells that play a role in recognizing, processing and presenting antigens to T cells. To create a DC-based cancer therapy, monocytes are harvested from a patient and stimulated in the laboratory to produce DCs that phagocytose the patient's tumor in vitro. The DCs are injected back into the patient where they generate a strong antitumor immune response [51].

Although DC vaccines have shown promising results when used with other cancer treatments, there are certain roadblocks that may cause DC vaccines to become ineffective. DCs are the most potent Antigen Presenting Cells (APCs); however, to successfully initiate an antitumor T cell response DCs must be in their mature state. APCs efficiently process antigens using phagocytosis or endocytosis. Once the antigen is processed, Major Histocompatibility Complex (MHC) class proteins and the processed antigen are displayed on the APC surface. Additional cell surface molecules, called Cluster of Differentiation (CD), CD80, CD83, or CD86, are also present on the APC surface. The surface markers CD80 and CD86 are part of the B7 family of membrane surface markers. The presentation of these surface markers allows APCs to effectively interact with T cells [54, 55]. Figure 4.1 is a simple adaptation of the interaction of DCs and T cells in a tumor microenvironment. The figure shows how the surface markers on

a mature DC interact with T cell surface marker to initiate a T Cell response [55].

In immunotherapy vaccines, mature DCs have been shown to prime and boost the antigen-specific T cell response in cancer patients. There are several treatments that can be added to successfully mature DCs; however, research in this field has shown that maturation of DCs also induces regulatory T cells (Tregs). When induced, Tregs have been shown to suppress the activities of effector T cells and DCs. The ability to produce an effective antitumor response is the goal of cancer immunotherapy; unfortunately the induction of Tregs renders the treatment ineffective [53, 56, 57].

DC vaccines offer an effective and potentially nontoxic treatment option for cancer patients. DC vaccines have proven to be effective in clinical trials and they have been successfully used with other cancer therapies [51, 53]. To explore the data in this growing field of research, an ontology driven approach can be used. This approach will allow researchers to annotate data, expand on the existing domain knowledge, and share newly discovered knowledge with the community [29, 41].

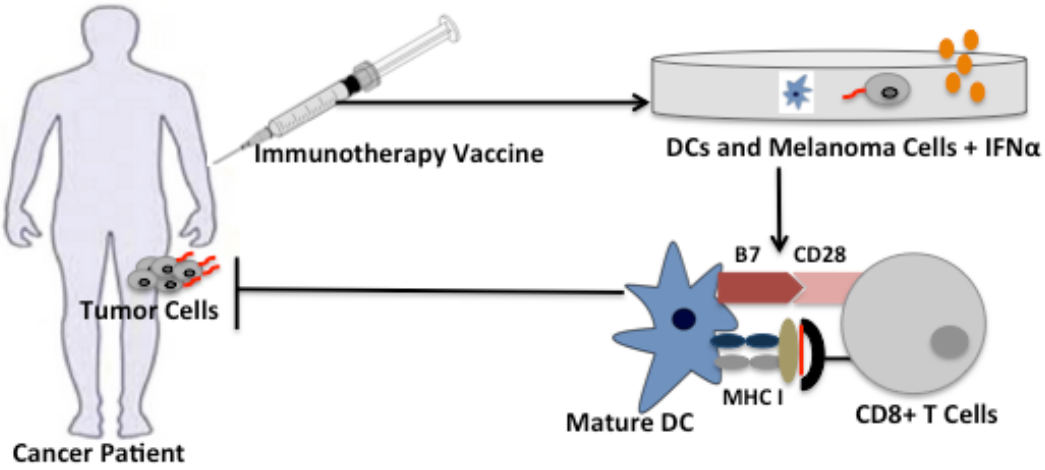

Figure 4.1: The immunotherapy vaccine is generated ex vivo, in the presence of tumor antigen loaded DCs and maturation stimuli (IFNα). The DCs uptake tumor antigens and present it to CD8+ T cells coupled with MHC I. This leads to the proliferation of CD8+ T cells and the initiation of a T cell mediated immune response. *Note: B7 represents CD80/CD86 complex

5. OBDI, A NOVEL PIPELINE

The OBDI pipeline incorporates various processes that are required to prepare integrated experiments that can be analyzed in silico. OBDI pipeline integrates informatics tools that are used independently in the field (see Figure 5.1). The ability to use these tools in a single pipeline allows users to integrate samples that were overlooked by previous curation efforts. OBDI was developed to support the reuse of data by annotating both GSD and GSE experiments from GEO experiments such that novel combinations of experiments can be analyzed using the pipeline. In Figure 5.1 the five components of OBDI are depicted. A general overview of each component is provided.

The OBDI pipeline is written in the Java programming language and it contains five components: 1) Build Ontology, 2) Acquire Data, 3) Organize Data, 4) Process Data, and 5) Analyze Data.

To execute each experiment in OBDI a configuration file is required. Since data integration and analysis is a multistep process, a configuration file allows users to track the input of different parameters that are required to successfully execute OBDI. The configuration file can be created in any text editor. The various parameters of the configuration file are provided in Appendix B.

The first component, Build Ontology, requires users to build the base ontology. The basic framework is provided; however, using OBDI, users must add the specific

GEO samples that will be integrated to create novel experiments. Once the base ontology is built, the next four components (i.e., Acquire Data, Organize Data, Process Data, and Analyze Data) are executed consecutively within OBDI. In the second component, Acquire Data, data are directly acquired from GEO by providing the http or ftp link for the data associated to each GEO DataSet sample. There is no limit on the number of GEO data that can be acquired. The third component, Organize Data, requires the organization of a single experiment from annotated data downloaded during the Acquire Data step. Currently, OBDI, supports Affymetrix Microarray Chips, Illumina Bead Chip, and RNA-Sequencing (RNA-Seq) experiments. More platforms can reasonably be integrated according to the user's requirements. The obtained samples are organized by reasoning over the base ontology that was created by the Build Ontology step. The reasoned ontology creates a directory structure and a mapping file for the annotated machine learning experiments. The directory contains subfolders that are associated to the specific machine learning conditions. The mapping file associates each sample to the specific machine learning condition, allowing the samples to be stored in their appropriate directory. The Process Data step requires preprocessing of the annotated experiments. The appropriate modules, from Gene Pattern, required for each gene expression platform is implemented in the Process Data component of OBDI. During the last step, Analyze Data, the normalized data are analyzed using ML-Flex. The output from this analysis allows users to interpret the results for biomedical purposes.

OBDI is a command line pipeline that must contain the following input parameters: Gene Pattern Username, Gene Pattern Password, OBDI High-Throughput

Module (Affymetrix, Illumina, RNA-Seq), and the Configuration File. The Build Ontology component can be executed separately. This is recommended because individual GEO Series record samples and metadata information is automatically inserted into the base ontology. In following sections, the OBDI pipeline components are described in more detail.

Build Ontology

In the Build Ontology step of the OBDI pipeline, the following steps are required: building base ontology components, creating an indexing file using the HermiT reasoner, and automatically creating a local directory structure that stores the annotated machine learning experiments.

Building Base Components of the OWL File

Ontologies in this dissertation are developed using the OWL syntax. Different fields of each GEO experiment are described using the following owl components: classes, individuals, and properties. Classes represent the main building blocks in an OWL ontology, and classes can include OWL individuals as instances [33]. There are three classes created to represent the components from a GEO experiment: *condition, sample*, and *treatment*. A fourth class, *cell*, is imported from the cell ontology [46, 47]. *Treatment* contains information about the different stimuli that were used to treat biological samples in a GEO experiment. A *treatment* may have two of more subclasses that refer to control and experimental stimuli. The imported *cell* class contains a subclass, *cell in vitro*, to store cells or other specimens that are manipulated in the laboratory for research purposes [46]. To link the relationships between OWL

components, object properties are used. The object properties along with the domain and the range are defined in Table 5.1.

Creating A Directory Structure in the User's Local Drive

Ontologies developed using the OWL syntax can be parsed by implementing libraries available in the Java programming language [58]. I used the OWL Application Programming Interface (API) to reason over the base ontology. By using the HermiT, reasoner within the OBDI pipeline, classes that did not satisfy the user-defined condition in the base ontology are printed to the console. This allows the user to check for any inconsistencies that may occur between the OWL entities. Once the ontology is reasoned without inconsistencies, the user can access the inferred OWL entities [59].

Using the reasoned OWL file, a directory structure mimicking the *ML-Experiment* OWL class can be created. Each directory structure contains the name of the machine learning experiment with subfolders that contain the conditions associated with that experiment. To build the directory structure, a single OWL class is parsed. The OWL class *ML-Experiment* and all it subclasses are parsed out to create the directory structure. The subclasses for each *ML-Experiment* are represented by the machine learning condition associated to the specific *ML-Experiment*. An example of how each ontology is used to create the file structure in the user's local system, depicting the integrated samples is discussed in each experiment chapter.

Creating an Indexing File Using the Reasoned Ontology

GEO provides metadata files that store experiment information in a structured XML format. However, the standardized vocabulary to describe each field of an

experiment does not exist. This makes it difficult to organize samples in a logical and meaningful manner. Ontologies were developed to simplify the complexity of data integration across GEO experiments.

The development of ontologies provides a means of integrating samples across GEO experiments. The key to organizing the samples is based on the ontology developed to annotate across related GEO experiments. Creating the directory structure and generating the master-indexing file occur in the same Java module. An indexing file is created using the reasoned OWL file. An indexing file allows users to place raw samples from GEO experiments into an appropriate *ML-Experiment* subfolder, which is represented by the appropriate machine learning condition (See Appendix C for details).

Acquire Data

The methods needed to acquire data from the GEO repository are similar. Each GEO experiment entry contains a link to download the samples from that experiment. The user must specify this link in the OBDI configuration file. The downloaded data are stored temporarily in the output directory. As the samples are acquired from GEO, the metadata files are also downloaded and processed with the pipeline.

Extracting Metadata

The process to extract metadata information from GEO data is similar across experiments. Each samples in GEO has an XML file [7] that contains the associated metadata. The metadata file for each experiment contains several fields that guide the development of various OWL entities. The XML tag retrieval was written in native Java using XPath expressions [60]. XPath is a query language that allows for easy

navigation through the tree structure of XML documents. By using path expressions, XPath can be used to select specific nodes from XML files.

The following XML fields are retrieved for developing the ontology: Sample iid, the Title field for the sample, Platform iid, and the Title field for the particular platform. To maintain clarity in the document, the Title fields will be referred to as Sample Title and Platform Title. The Sample iid is an alphanumeric identification assigned to each sample in an experiment. A single sample in GEO is comprised expression values or sequencing data for a given set of genes. Each Sample iid is associated to a Platform iid and a Sample Title. The Platform iid is also an alphanumeric identification that is associated to the Platform Title, which provides information about the specific high throughput technique used to analyze the sample. The Sample Title provides information about how a particular sample was treated in the laboratory before performing the high throughput analysis. The retrieval of these XML tags serves as a key component in building the base ontology.

Organize Data

The OBDI pipeline requires that data modules are developed for each type of data platform. High throughput analysis techniques measure different biological samples at varied coverage [61]. Due to the variability in biological data formats, the preprocessing for each platform is handled independently in OBDI. The samples used in this have been analyzed on three different platforms: microarray experiments performed using Affymetrix, differential expression analysis or RNA-Seq samples, and experiments performed using the Illumina Beadchip.

For the GEO experiments analyzed on Affymetrix the raw data are downloaded in tar format. The first step to reveal the raw data is to extract the tar file. Individual samples in the experiment are compressed; hence, each sample file is unzipped to reveal the raw data. The intensity files in Affymetrix experiments are stored as cel files. Although the downstream processing of RNA-Seq data are different from expression array data, similar preprocessing method is used to extract raw data files. The RNA-Seq samples processed in OBDI were sequenced using the Illumina Genome Analyzer. Data for each sample are stored as Sequence Alignment Map (SAM) or a Binary Alignment Map (BAM) file. A SAM file is tab-delimited data that contain aligned sequences. Each SAM file has 11 mandatory fields that appear in the same order. The BAM format is the machine readable version that stores the data in binary form [62]. The Illumina Expression Beadchip data are acquired using Gene Pattern. Once the data are available locally, each sample is separated as a column and stored as an individual file. In all three modules, the samples are organized using the ontology and the metadata XML file.

When generating the ontology, the platform information from each metadata file is extracted. Within the XML metadata file, the platform information is associated to each sample in the GEO Series record. In most cases, this information is specified in the Platform Title of the XML file. This is information is added to the ontology as an annotation property to the respective sample. This provides users the ability to distinguish between platforms within the base ontology. Once a single platform is chosen, it is added to the name of the curated experiment. Samples that are not part of the experiment may be deleted or left in the ontology for future use. The organization

method is generalized due to the ontology; however, it is a key component because the preprocessing methods vary among platforms and data types.

When combing data across GEO experiments there are several samples; therefore, the reasoner is used to add samples to the appropriate *condition*. Using the asserted ontology and the inferred links that are established by the reasoner, researchers can create similar machine learning experiments using other GEO experiments. When the reasoner is applied, the sample individuals denoted with a GSM ID are inferred as members of the appropriate *condition.* Based on the equivalency classes defined in each *ML-Experiment* subclass, the GSM IDs are inferred into the specific machine learning experiment. This allows users to build in silico experiments outside of GEO.

Process Data

Once the samples are organized into a structured directory system, the samples can be used to perform machine learning analysis. However, before performing meta-analysis, the integrated data sets must be normalized and processed into the appropriate file format. The goal of this project is to not only integrate GEO experiments, but also integrate bioinformatics tools into a single pipeline. To perform some of the preprocessing analysis, Gene Pattern is used [10].

For Affymetrix expression data, the Robust Multi-array Average (RMA) normalization technique is considered the standard [63]. When analyzing Illumina Expression Beadchip data from GEO, samples are already normalized using quantile normalization [63]. Gene Pattern generates a tab-delimited file for all three modules. This generalized file format can be converted to the Attribute-Relation File Format (ARFF) format.

In order to use the OBDI pipeline in conjunction with Gene Pattern, users need to create an account with a Gene Pattern username and password. This is a simple two-minute process that provides each user with an account that allows them to access Gene Pattern modules and all the analysis performed on the Gene Pattern server. The username and password are passed into the OBDI configuration file so that users are not interrupted by prompts for this information (See Appendix D). Gene Pattern modules are implemented into the pipeline and no external use of Gene Pattern is required during analysis.

Analyze Data

The final step involves analyzing the data from the annotated experiment using ML-Flex. The commonly used fields in the ML-Flex experiment file are defined in all the ontologies as annotation properties. Reasoning over the base ontology generates the experiment files required to execute the analysis in ML-Flex. Until this point, the ontologies were accessed using the OWL API [59]; however, for this step XPath is used to write a generalized method that creates ML-Flex experiment files for each machine learning experiment. The annotation properties are organized under a different *XML namespace*, which allows for easy parsing using XPath. A user can define XML *namespaces* user to avoid conflict between elements, while mixing different documents. Although the annotation properties are defined in the same ontology, it is important to keep the ML-Flex experiment file fields in a different namespace because they address a different step in the pipeline. The namespace assigned to the annotation properties is: http://bmi.utah.edu/ML-Flex. This allows for easy and error free extraction of the annotation properties and the values [15, 33].

When more than one algorithm is used, ML-Flex performs ensemble-learning [15, 64]. When only one algorithm is specified, the ensemble-learning method will not be performed [64]. This unique feature of ML-flex is used to compare across different algorithms. There are seven ensemble-learning methods implemented in ML-Flex:

1. Majority Vote tallies the number of times a data instance was predicted for a class and favors the class that receives the most counts. If multiple classes receive the same vote, majority vote will choose a class at random.
2. Weighted Vote emphasizes single predictions that are considered the most informative and places higher emphasis on those individual predictions that perform the best.
3. Select Best makes a prediction based on the individual prediction that received the best Area Under the Curve (AUC) in nested cross validation.
4. Max Probability examines the probability for each class across individual predictions, and the class with the highest individual probability is selected.
5. Mean Probability examines the probabilities for each class across individual predictions and the class with the highest average probability is selected.
6. Weighted Mean Probability is a combination of the Weighted Vote method and the Mean Probability method. Predictions are calculated similar to the Mean Probability method but weight is assigned to each individual probability.

7. Finally the Stacked method uses the probabilities from individual predictions and trains a second-level classification algorithm to make cumulative predictions based on those values. The Decision Trees algorithm is set as default for the second-level predictions.

After the analysis is complete, an output folder for each experiment is created.

This folder contains a summary of results, which can be viewed on a browser. These results can also be parsed in the output directory. When comparing different classifiers in ML-Flex, results are aggregated into a single table, which makes it easy to compare the analysis among different algorithms. Statistical measures, such as AUC, accuracy, error rate, and recall is also summarized in a table. The advantage of using ML-Flex is that the postanalysis results are summarized allowing the user to evaluate the performance of the classification algorithms. Users can further process the machine learning results for biological relevance outside the OBDI pipeline.

Table 5.1: There are six object properties that are used to relate OWL entities. It is optional to specify domain and range restrictions to an object property. Domain and range allow users to specific what entities should and should not be related. Defining a domain and a range can increase the reasoning power of ontologies.

Object Property	**Domain**	**Range**
containsCellType		
hasCondition	*sample*	*condition*
hasSample	*ML-Experiment*	*sample*
hasTreatment	*sample*	*treatment*
inExperiment	*condition*	*ML-Experiment*
isTreatmentOf		

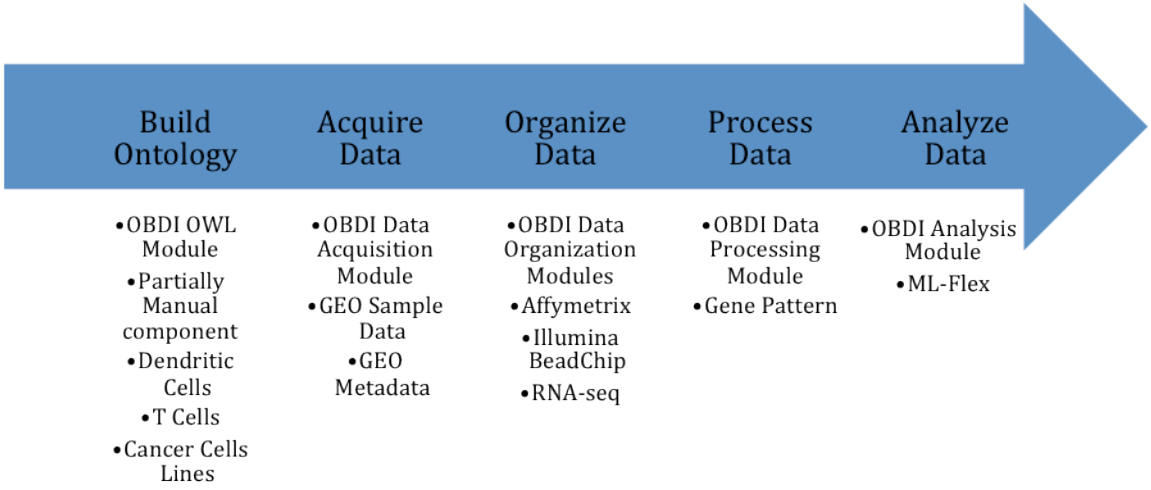

Figure 5.1: This figure shows the generalized diagram of the OBDI pipeline. It contains five components that allow for successful execution of each experiment.

6. PIPELINE USE-CASES: CANCER IMMUNOTHERAPY EXPERIMENTS

Cell ontologies contain extensive information about cell types in their native state. In the OBDI pipeline, the biological knowledge in the cell ontology is extended by adding information about cell lines cultured in a laboratory setting. To successfully integrate data across GEO Series records, it was important to incorporate prior knowledge information in relation to various cell types. Incorporating the cell ontology supports expanding the OBDI framework to other biomedical domains. The information within the cell ontology can be extended to include in vitro cell types that are cultured in the lab for gene expression and RNA-Seq analysis.

The OBDI pipeline helps generate hypotheses and also informs existing hypotheses by adding new information that is generated from analyzing newly curated experiments. To analyze cancer immunotherapy as a use-case with the OBDI pipeline, four use-case experiments are evaluated by performing meta-analysis. Each experiment is curated using ontological representations. Experiment 1 consists of information about DCs and treatments related to maturing DCs. This experiment served as a staring point to explore the cancer immunotherapy space. Since the downstream activation of T cells and the induction of Tregs depend on the maturation of DCs, it was important to characterize these cells, in silico, successfully. Based on the results from Experiment 1, the subsequent experiments were curated to explore the tumor microenvironment in silico.

Experiment 2 explores how various T cell subtypes are classified. The ontological representation for Experiment 2 deals with treatments that lead to T cells differentiating into different subtypes. Biomolecular components relevant to T cells are explored to see if new information from analyzing Experiment 2 can help enhance the results in Experiment 1. To complete the model exploring a tumor microenvironment, it is important to analyze differentially expression in a specific cancer cell lines. In Experiment 3, samples were consolidated so that IFNα, a specific treatment used to mature DCs, is also used to evaluate gene expression in cancer cell lines. Finally, in Experiment 4 OBDI is extended to include other sequencing data making OBDI inclusive of other data types. By relating a specific treatment to DCs and to a clinical therapy setting makes OBDI an effective tool that promotes translational research.

7. EXPERIMENT 1: CHARACTERIZATION OF DENDRITIC CELL MATURATION

To validate the OBDI pipeline, the field of cancer immunotherapy is explored because of the complexity of the data set and the potential benefits from combining data across experiments. Since there are several methodologies used to generate cancer immunotherapy treatments, I focus on acquiring data that allow the exploration of the immune system response in DC based vaccines. For an overview of topics related to cancer immunotherapy see the chapter titled, Overview of Cancer Immunotherapy.

Mature DCs are the most effective Antigen Presenting Cells (APCs) for initiating a T cell mediated immune response. Normally DC cells are in an immature state in the mucosal membranes or the epithelial layer of the skin, but when DCs capture pathogenic or inflammatory stimuli, they mature, migrate to the lymph nodes and present antigens to T cells [65]. There are surface markers that have upregulated expression on mature DCs. The GEO samples in Experiment 1 assessed maturity based on three Cluster of Differentiation (CD) markers: CD 80, CD 83, and CD 86, all of which are standard markers used in the field of immunology to assess mature DCs [65]. Figure 7.1 displays how a DC can be matured and how a mature DC interacts with CD8+ T cells to initiate a immune system response.

The weakness of DC vaccines is an ineffective antitumor immune response resulting from the induction of Tregs (T regulatory cells) that result in a tolerance to the

tumor antigen being presented by the mature DCs. This tolerance to tumor antigen is an unwanted effect that causes immunotherapy vaccines to become ineffective [51, 57].

Induced Tregs act as a barrier to prevent the effector functions required to kill tumor cells [56, 57]. There are several known mechanisms by which DCs induce Tregs. In recent studies, researchers have shown indoleamine 2,3-dioxygenase (INDO) is one mechanism responsible for induction of forkhead box P3 (FoxP3)+ Tregs. FoxP3+ induced Tregs are generated by DCs and this induction is caused by multiple factors including INDO, retinoic acid, Vitamin D, and Transforming Growth Factor-beta (TGF-β).

Experiment 1 uses the OBDI pipeline to combine data from four GEO experiments that are publically available. The objective of Experiment 1 is to successfully characterize the maturation of DCs across the 11 treatments that occur across the four GEO studies. Since the four GEO Series records used in Experiment 1 have not been curated by GEO, a secondary analysis of these GEO samples may provide important information of how mature DCs play a role in immune system response during cancer immunotherapy.

Methods

Build Ontology

The ontology is built around experiments measuring the maturation of DCs. Peripheral Blood Mononuclear Cell *PBMC* is added as a subclass to *cell in vitro*. Using the OWL classes, *treatment* and *cell in vitro*, machine learning conditions are established. These conditions play a role in classifying individual instances (samples) during the machine learning analysis. Unless otherwise mentioned, OWL entities are associated

with *is-a* relationships. The next several steps describe the detailed framework of how the base ontology for Experiment 1 is created:

1. The cell ontology is imported into the OWL file created for Experiment 1.
2. Within the cell ontology there are two OWL classes, *cell* and *cellular_component*. Under *cell*, two subclasses differentiate between native cell and cells generated in vitro [46, 66].
3. The OWL class *PBMC* is added as a subclass to *cell in vitro*. PMBCs are obtained from blood banks for experimental purposes. They are inclusive of different types of immune system cells. Since this experiment deals with DCs, *MDDC* (Monocyte Derived Dendritic Cells) is added as a subclass to *PMBC*. Various DC subtypes are added based on surface marker expression to differentiate between mature and immature subtypes [66]. The subtypes are disjointed from each other.
4. The next three OWL classes are unique to the ontologies generated in OBDI: *treatment*, *condition*, and *sample*.
5. The *treatment* contains a subtype *DCTreatment*. The two types of DC treatments used in this experiment are control treatments and maturation treatments. These subclasses have individuals associated to them as OWL members. The members of the *ControlTreatment* subclass are: *untreated* and *isotype*. The members of the *MaturationTreatment* subclass are: *Poly (I: C), anti-FcgRIIb, CD40L, galectin, INF-alpha, inflammatory_cytokines, Lipopolysaccharide (LPS), LPS-gamma*, and *schuler*.

6. The subclasses defined under the condition OWL class for Experiment 1 are *mature* and *immature*. The two conditions are disjointed from each other. The equivalency for each condition is established by adding the appropriate cell type and treatment subclasses that would come together to establish the machine learning condition OWL class. Object properties, *containsCellType* and *hasTreatment* are used to associate the cell type and the treatment, respectively.
7. *Sample* is the final class that completes the base ontology. The machine learning experiment, *ML-Experiment*, is a subclass of *sample*. Experiment 1 annotated from this ontology is a subclass of *ML-Experiment. DCMaturationU133Plus* contains equivalency classes associated by the *mature* or *immature* OWL entities.
8. The final step is automated by implementing the OWL API [59] into the pipeline. In this step all samples from the GEO Series records are added as OWL *individuals* and associated to *sample* as *members*. The Sample iids that are parsed out of the metadata XML file are added as an instance to the *sample* OWL *class*. There are different annotation fields that can be added to individuals. The Sample Title parsed out of the metadata file is added as a RDF comment under Annotation Properties to each corresponding Sample iid individual. The information from the Title Text field is summarized into a treatment term and stored as an individual. Each *sample* member is associated to a particular cell type by the *containsCellType object properties*. Finally, *hasTreatment* object properties are used to link the *sample* members to the *treatment* members.

The above-defined logic forms the framework of the base ontologies used in the OBDI pipeline. The asserted ontology for the DC maturation experiment is depicted in

Figure 7.2.

Acquire Data

I consolidate 50 samples from the following four GEO experiments: Fulcher et al., Dohnal et al., Ebstein et al., and Dhodapkar et al (see Figure 7.3). Across the four experiments, the researchers generated Monocyte Derived Dendritic Cells (MDDCs) by using a cocktail of GM-CSF and IL4. Fulcher et al. identify expression differences between Galectin-treated DCs and LPS-treated DCs compared with untreated DCs. Dohnal et al. examine differential gene expression of a specific set of genes when LPS in the presence of IFNγ triggers DC maturation. Ebstein et al. focused their study on a set of 1200 ubiquitin-proteasome system (UPS)-related genes, and found differences in UPS gene regulation between DCs matured with infectious stimuli (LPS and Poly(I:C)) versus DCs matured under T cell stimulatory or inflammatory conditions. The analyses performed by Dhodapkar et al. compared anti-FcγRIIB, inflammatory cytokines, or IFNα versus untreated and Isotype control [52, 67-69].

Organize Data

The four GEO Series records were analyzed on the Affymetrix U133 2.0 Plus microarray platform. The data organization method using the ontology is already described in the "Organize Data" section of the general methods used in the OBDI pipeline. Once the four GEO experiments were downloaded into a temporary location, the first step involved extracting the tar file. The samples were compressed in gzip file format. Unzipping these filed extracted the raw Affymetrix intensity files in cel format.

Figure 7.4 shows an example of how the *DCMaturationU133Plus* experiment looks when the directory structure is automatically created in the user's local drive.

The ontology serves severs as the core component to integrate samples across GEO experiments and to organize novel machine learning experiments on a user's home directory. This is a powerful and efficient way to integrating GEO Series records and to store the samples in an organized directory structure.

Figure 7.5 shows the four GEO experiments from where the samples are acquired. The curated experiment using the OBDI pipeline contains 50 samples. The 50-sample data set was divided into a training (n=12) and a hold-out testing set (n=38). The training included all LPS [3] (n=6) samples and one to two randomly selected untreated [1] samples from each GEO experiment (n=6). Thus, the 12 sample training data contained a set of 6 control samples and 6 samples of DCs matured with LPS. The remaining 38 samples are used as hold-out samples.

Process Data

For microarray data analyzed in Affymetrix, RMA method is used to normalize the data across the combined samples in each machine learning experiment [63]. RMA normalization contains the following steps: background adjustment, quintile adjustment, and finally a summarization step [63]. The raw Affymetrix cel files are zipped and normalized using the ExpressionFileCreator module on the Gene Pattern server [10]. When normalizing Affymetrix data, a Gene Pattern clm annotation file is required. This annotated file allows for accurate replacement of GSM IDs with the appropriate machine learning condition associated to that sample in the ontology. The Gene pattern clm file can be created using the reasoned ontology. It is similar to the master-indexing file;

however, the spacing is tab-delimited. Each Gene Pattern clm file contains three columns: GSM ID, user defined machine learning condition, and the Title Text annotation for each GSM ID. Using the clm file, the integrated GEO samples can be normalized [10, 63].

The normalized data produced by the Gene Pattern module, ExpressionFileCreator, is stored as a tab-delimited file called a Gene Pattern gct file. This is the standard file format used by Gene Pattern where samples are represented in columns and the probe IDs are represented as rows. The final file formatting happens outside the Gene Pattern server, where the tab-delimited file is converted to Attribute-Relation File Format (ARFF) format for machine learning analysis.

A generalized Java method is written within the pipeline to convert the normalized file generated by Gene Pattern into ARFF, which is used when running machine learning analysis in Weka [11]. Although Weka is not directly used in the pipeline, ML-Flex implements various algorithms that are used in Weka. In an ARFF file, all the attributes (probe IDs or gene names) are listed in the beginning. Next, the corresponding data value is listed for each attribute in a single line. At the end of each line, the machine learning class variable is listed [11].

Analyze Data

Microarray data contain an extensive number of features; however, each feature is not relevant to the study of immunology. The ReliefF ranking method is used on the training set (n=12) to select and rank features that successfully differentiate among classes [18, 27, 70]. The features are selected by adding top ranked features and stopping when the accuracy fell below 100%.

The Naïve Bayes classifier, a machine learning algorithm in ML-Flex, is used to characterize the maturation of DCs. The classifier is trained using the training set. The Naïve Bayes classifier uses all selected attributes and treats them as independent of one another. Leave One Out Cross Validation (LOOCV) is used on the training set to get an estimate of the generalized error.

Results

Generating Inferred Ontology

Once the base entities are added to the OWL file, the HermiT 1.3.6 reasoner in Protégé is used. The framework of the ontology provides a structured way of storing GEO related elements and the user-interpreted machine learning conditions. Using the base ontology, I have successfully created a way to store metadata across GEO experiments in machine readable format. The next step is to add GEO samples to suitable *conditions* and, hereby, adding those samples to the *ML-Experiment* that contains the specific *conditions* as an equivalent class.

To build new *ML-Experiments* users can manipulate equivalency classes associated by *condition.* The OBDI methodology extends the experiment driven architecture of ML-Flex by enhancing it with ontologies that specifies the experiments with necessary and sufficient conditions. Generating the inferred models from the base ontology plays a crucial role in annotating samples into novel experiments. By relying upon the logical definitions and the reasoner, the user is able to integrate various GEO samples in order to create annotated experiments that can be analyzed in silico. These newly annotated experiments are represented under the OWL *class* labeled *ML-*

Experiment. Figure 7.4 shows how GEO samples, represented as OWL individuals, are correctly positioned as members to respective OWL *condition* classes.

Characterizing DC Maturation Using Naïve Bayes in ML-Flex

Based on the stopping criteria, the ReliefF ranking method applied to the training set resulted in 65 probes out of a total 54,675 probes IDs that were differentially expressed between immature and mature DCs. Based on the results from the training data, I checked if the 65-feature classifier generalized to the 38-sample hold out test set consisting of both mature and immature DCs. The classifier generalized across the 38 hold samples that included 11 treatments with a hold out test accuracy of 100%. All hold out untreated and Isotype samples were classified as immature; all hold out IFNα, CD40L, anti-FcγRIIB, Schuler, Cytokines, Galectin, Poly(I:C), and LPS/IFNγ samples were classified as mature. Figure 7.7 depicts a screenshot of the ML-Flex analysis where the performance metrics of the analysis are summarized. The performance metrics show that the number of correctly classified test instances was 38 out of a total of 38.

Probe Level Analysis of DC Maturation Across the Four GEO Experiments

The heatmap in Figure 7.8 displays the expression patterns of the 65 probe IDs. The probe IDs are clustered according to biological function and pathway information obtained from the KEGG database [43]. The probe IDs that did not associate with interferon regulatory and inducible genes, Nuclear Factor-Kappa Beta (NF-κB) pathway, chemokine signaling pathway, kynurenine pathway, or cell adhesion molecules and ECM Interaction were clustered as miscellaneous.

Table 7.1 shows the total mean for all probe IDs of specific clusters in each sample. Using two-tailed independent-samples t-tests, I compared the mean expression patterns for probe IDs in the biological function and pathway clusters for the untreated condition against the means of probes in clusters for all other conditions.

The table shows the treatments in order of increasing expression intensity from left to right. Isotype serves as a negative control that does not mature DCs. No significant difference was observed between the untreated samples and samples treated with Isotype. It is important to note that IFNα successfully matures DCs as measured by CD86 mRNA expression; however, the treatment produces significant changes only in the interferon regulatory and miscellaneous clusters. The biomarker genes INDO, CD274, and CD44 are of particular interest because of their influence on FOXP3+ Treg cells. In Fig, I use a box plot to compare INDO expression across treatments.

Discussion

The 65-probe panel clearly demonstrates differences between untreated immature DC expression and LPS mature DC expression. The Isotype, negative control, follows the untreated expression pattern, while the LPS/IFNγ treatment has gene expression patterns similar to LPS treatment. The distinguishing features between classical LPS-matured DCs and immature DCs were expected. The ReliefF feature selection technique ranks the probes based on distinctive differences between untreated and LPS-treated DCs [25, 27] and the Naïve Bayesian classifier determines the threshold for the number of probes needed to correctly classify the 12 training samples.

Surprisingly, while the remaining eight maturation treatments display patterns ranging between untreated and classical LPS-treated DCs, the Naïve Bayesian classifier,

using the 65 probes, correctly classified all mature hold-out samples as mature. The classifier did not overfit the data as demonstrated by its ability to correctly predict the maturity or immaturity of nineteen hold-out samples in Dhodapkar while being trained on only one untreated sample from Dohdapkar (see Figure 7.5).

When DCs are treated with IFNα as a maturation treatment, NFκB pathway genes are downregulated and interferon-related genes have high expression. NFκB2 and NFκBIA are downregulated in DCs treated with IFNα; therefore, these genes are differentially expressed across DC maturation treatments. The Naïve Bayesian classifier identifies IFNα as mature partially due to the high expression values of interferon regulatory and inducible genes (Table 7.2). Interferon regulatory factor 7 and factor 9 (IRF7, IRF9) mediate IFNα signaling, and the interferon-induced proteins IFI44L, IFI6, IFIT5, IFIT3, GBP1, and MX1 make up the maximum total mean expression value for the cluster (i.e., 5045) across all the treatments.

Three biomarkers in my panel, INDO, CD274, and CD44, have been shown to increase the induction of Treg cells [71-77]. Recent studies have shown that INDO mediates the induction of Tregs through the alteration of TGF-β secretion via tryptophan metabolism and the kynurenine pathway [57]. INDO is an enzyme in the tryptophan pathway, as is KYNU, and both catalyze tryptophan metabolites. These metabolites increase secretion of TGF-β from DCs, which induces FoxP3+ in CD4+ T cells and drives native T cells to become induced Tregs [57, 73]. The overexpression of INDO has been shown to induce Tregs rendering immunotherapy vaccines to become ineffective [71, 73]. INDO and KYNU are downregulated in DCs that are matured with IFNα, making the treatment a potential intervention to prevent the induction of Tregs while

maturing DCs.

The gene CD274 is the ligand to the cell surface membrane protein PD-1, which is expressed on the surface of active T cells. It has been shown to suppress host immunity in T-cell lymphoproliferative disorders [75] by promoting the induction of Tregs. CD274 is downregulated in the IFNα treatment of DCs but upregulated in other treatments. This is consistent with my hypothesis that IFNα-treated DCs should have a reduced level of molecules that play a role in DC based induction of Tregs. Bollyky et al. show how hyaluronan cross-linked CD44 (downregulated in IFNα treatment) enhances production of TGF-β, which promotes FoxP3+ expression and induces Treg function. Hence, CD44 is similar to INDO in that it can influence Treg induction through TGF-β [57, 76].

My work provides a gene probe panel to further explore IFNα maturation of DCs in relationship to other maturation treatments. The low expression of INDO, CD274, and CD44 in IFNα-treated DCs suggests that IFNα may activate cytotoxic T cells and reduce the levels of Tregs. I have identified potential biomarker targets and mechanisms to investigate why IFNα treatments show a measurable immune response in cancer patients. Using the 65-probe panel, I identified two mechanisms that suggest ways to increase the immune response for better outcomes: maturation of DCs and suppression of Treg induction. The results from Experiment 1 generate a hypothesis of how IFNα can result in the downregulation of INDO, affecting the production of TGFβ downstream; therefore, inhibiting the induction of Tregs to the tumor microenvironment. This hypothesis can be translated and tested at the bench, with the hope of generating more effective DC-based vaccines.

Table 7.1: The total mean of all probe IDs involved in specific clusters for each sample. The number of observations (n) and standard deviation (SD) are listed within the parentheses using the following format (n, SD). A series of t-tests are conducted across each row by comparing the untreated condition to each of the other conditions. Asterisks indicate significant differences ($p < 0.05$). The results show significantly greater expression values for at least two clusters for each treatment other than the negative control Isotype. Note: when calculating the mean expression of the clusters, NFKBIA is in both the NF-Kappa-B and Chemokine Signaling clusters. The gene ICAM1 is in both the Cell Adhesion Molecules and NF-Kappa-B clusters.

	Miscellaneous	**Cell Adhesion Molecules and ECM Interaction**	**Kynurenine Pathway**	**Chemokine Signaling Pathway**	**NF-Kappa-B Pathway**	**Interferon Regulatory and Inducible Genes**
untreated	537 (385, 1135)	1051 (55, 1067)	1049 (33,734)	1137 (55, 1719)	886 (121,1162)	381 (88,342)
Isotype	601 (175, 1426)	1735 (25, 1644)	1431 (15, 944)	1274 (25, 1579)	929 (55,1036)	336 (40, 229)
IFNα	1098* (105, 1837)	2111 (15, 1997)	1302 (9,872)	1320 (15, 1403)	1109 (33, 922)	5045* (24, 3779)
CD40L	1971* (105, 3212)	3982* (15, 3896)	3227* (9, 2374)	7162* (15, 3153)	2715* (33, 2441)	778* (24, 473)
Schuler	2088* (105, 3367)	4233* (15, 4529)	5645* (9, 3677)	4583* (15, 3866)	3283* (33, 2838)	615* (24, 477)
Inflam Cytokines	2258* (140, 4067)	4128* (20, 3784)	9459* (12, 5087)	2635 (20, 4235)	3577* (44, 3094)	463 (32, 556)
anti- FcγRIIB	1986* (175, 3654)	3822* (25, 3531)	5166* (15, 2538)	6730* (25, 3979)	2993* (55, 2750)	1569* (40, 1066)
Galectin	2637* (105, 3905)	4783* (15, 3822)	5687* (9, 1309)	10326* (15, 3651)	3634* (33, 3506)	2811* (24, 1535)
Poly (I: C)	2330* (105, 3204)	4101* (15, 4140)	6761* (9, 4096)	11375* (15, 4043)	3576* (33, 2564)	5982* (24, 4119)
LPS/IFNγ	3461* (140, 5109)	4806* (20, 3641)	9510* (12, 7030)	13919* (20, 2909)	5412* (44, 4353)	4280* (32, 4236)
LPS	2954* (210, 4065)	5227* (30, 4378)	7371* (18, 2765)	11615* (30, 4357)	4041* (66, 3541)	4086* (48, 2548)

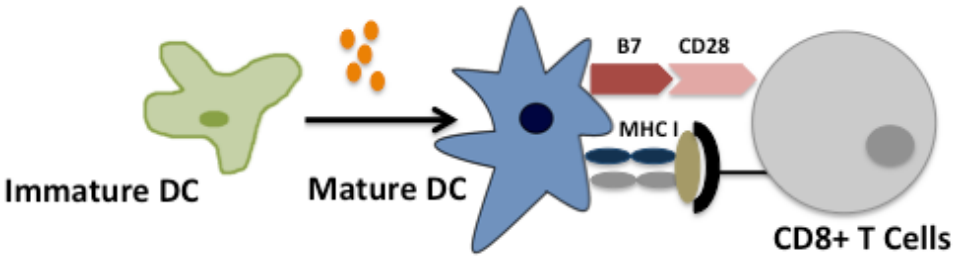

Figure 7.1**:** In the presence of stimuli (example: IFNα) immature DCs are matured. Mature DCs represent certain surface markers that interact with T cell markers to generate a T cell based immune system response. In this example, the B7 surface marker represents the CD80/CD86 complex. Along with these costimulatory molecules, the MHC I is expressed on the DC surface.

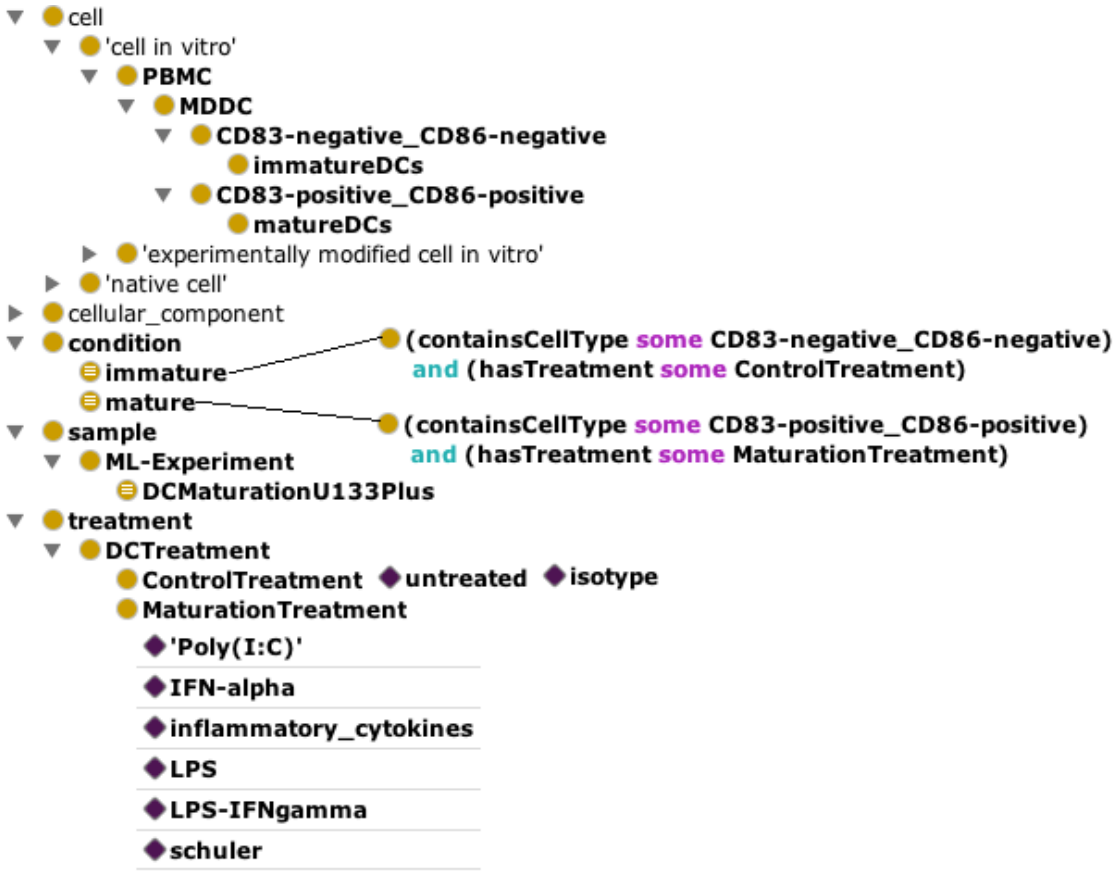

Figure 7.2: In this image, the major components of the asserted ontology are depicted. Using equivalency rules associated to *cell in vitro* and *treatment* classes generates the condition classes. For this experiment, there are two conditions asserted on the 50 integrated GEO samples. Each sample is asserted as a member under the sample class (Shown in Figure 7.4). The specific treatments are asserted under *ControlTreatment* and *MaturationTreatment*, respectively. Members are identified by purple triangles. The images can also be visualized using Onto Graph with the Protégé interface. The PG ETI Sova plug-in can be used to visualize the inferred ontology and the inferred individuals

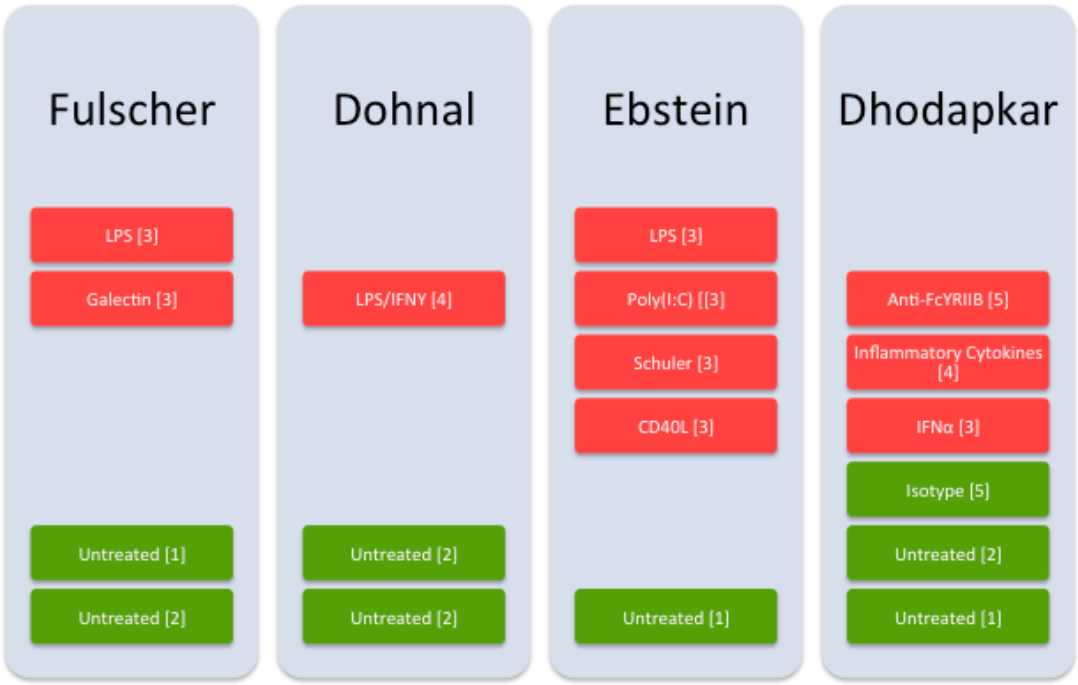

Figure 7.3: This image displays the four GEO experiment that are integrated to generate a novel OBDI data set to explore DC maturation across 11 treatments. Treatments in green are control treatments and treatments in red are maturation treatments. The numbers in parenthesis represent the number of samples in each treatment for the specific study.

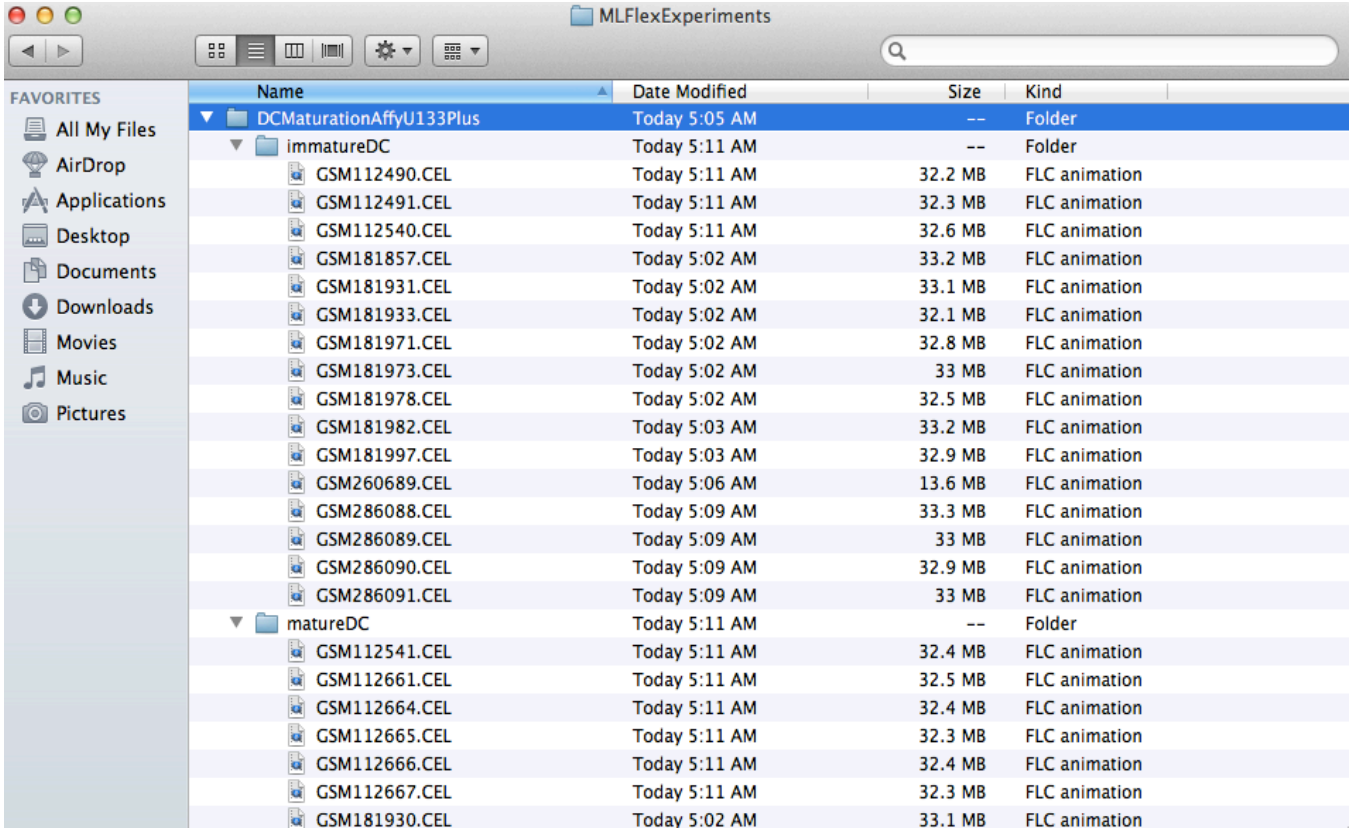

Figure 7.4: Directory structure mimicking the *ML-Experiment* OWL class. The machine learning conditions are the subdirectory, *immatureDC* and *matureDC*. The raw samples files, with the GSM ID, are organized into the appropriate condition folder using the reasoner OWL file.

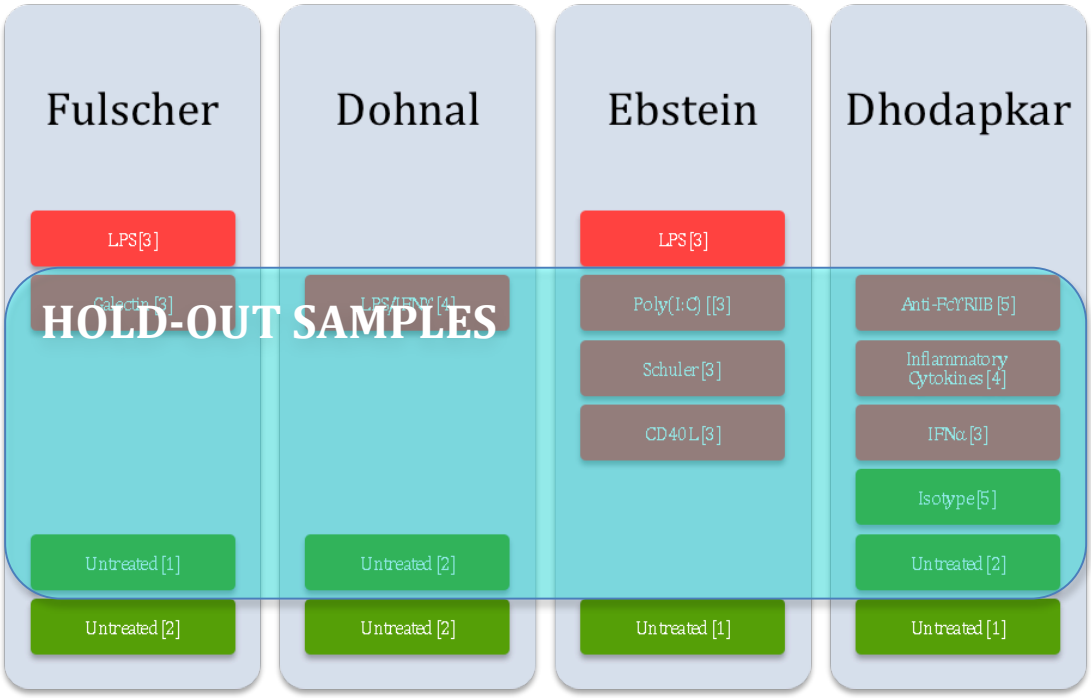

Figure 7.5: The samples highlighted in red are treatments used to mature DCs. The samples highlighted in green contain samples of immature DCs. The numbers of samples are listed in brackets. The highlighted blue box represents the combination of hold out samples used in the training set.

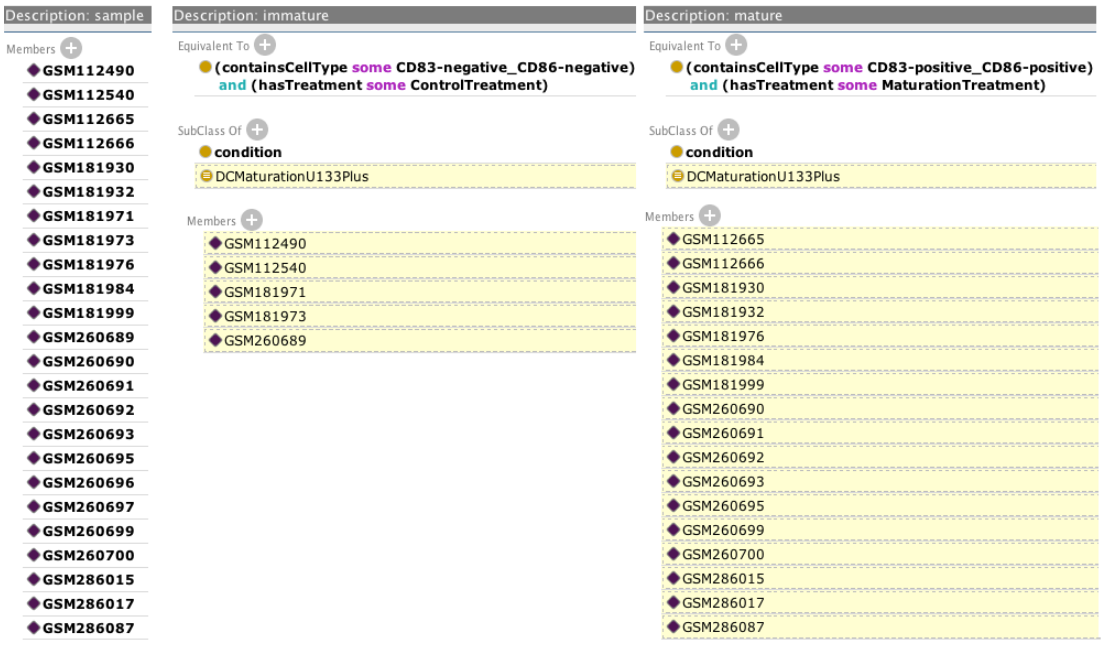

Figure 7.6: This figure is an illustration of the inferred model that is established after executing the reasoner. Samples denoted by an alphanumeric GSM ID are asserted as *members* to the *Sample* class but they are not associated to specific conditions or the experiment (*DCMaturationU133Plus*) until the reasoner is executed. The entities highlighted in yellow are reasoned as members to associated conditions.

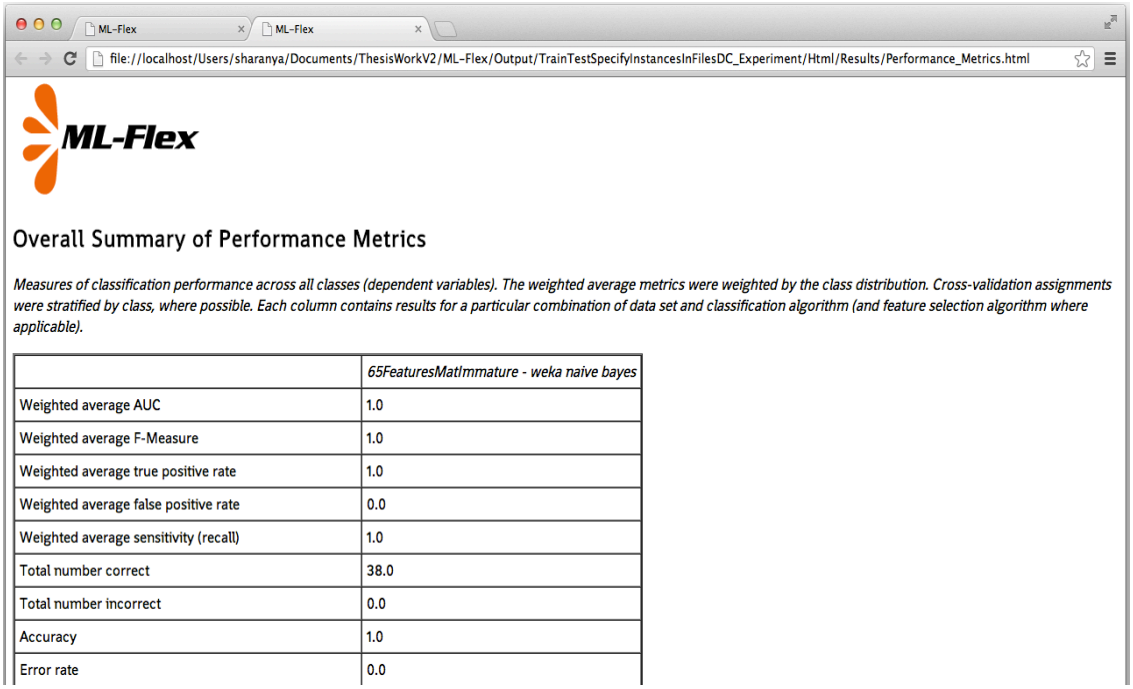

Figure 7.7**:** Snapshot of summary results in ML-Flex for aggregated machine learning experiments that classify the maturation of DCs.

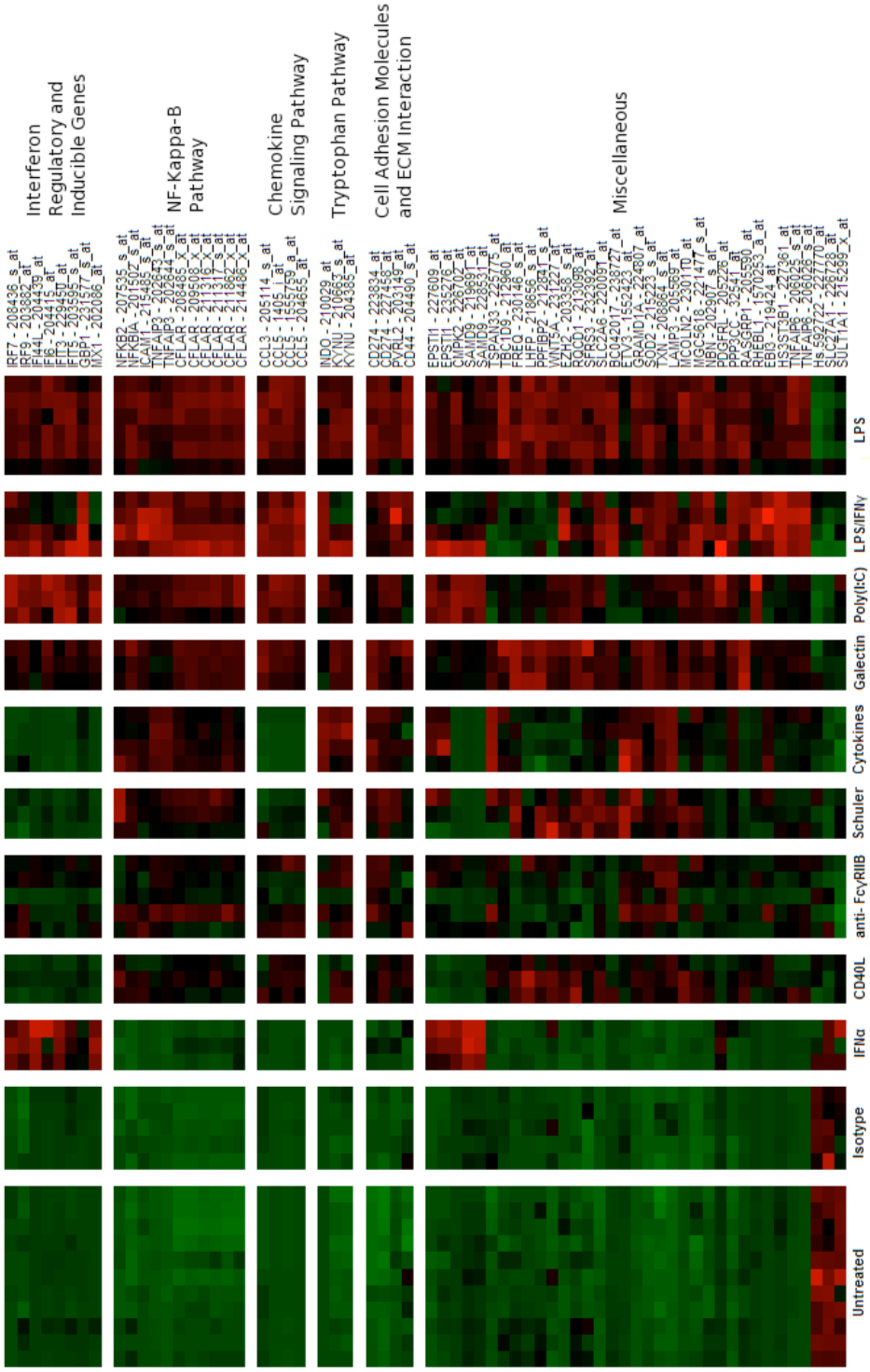

Figure 7.8: The x-axis lists the nine treatments, along with the untreated and LPS treated samples. The 50 columns of samples are grouped by treatment. The y-axis lists the probe IDs and corresponding genes. The probe IDs and genes are clustered according to biological function and pathway according to KEGG.

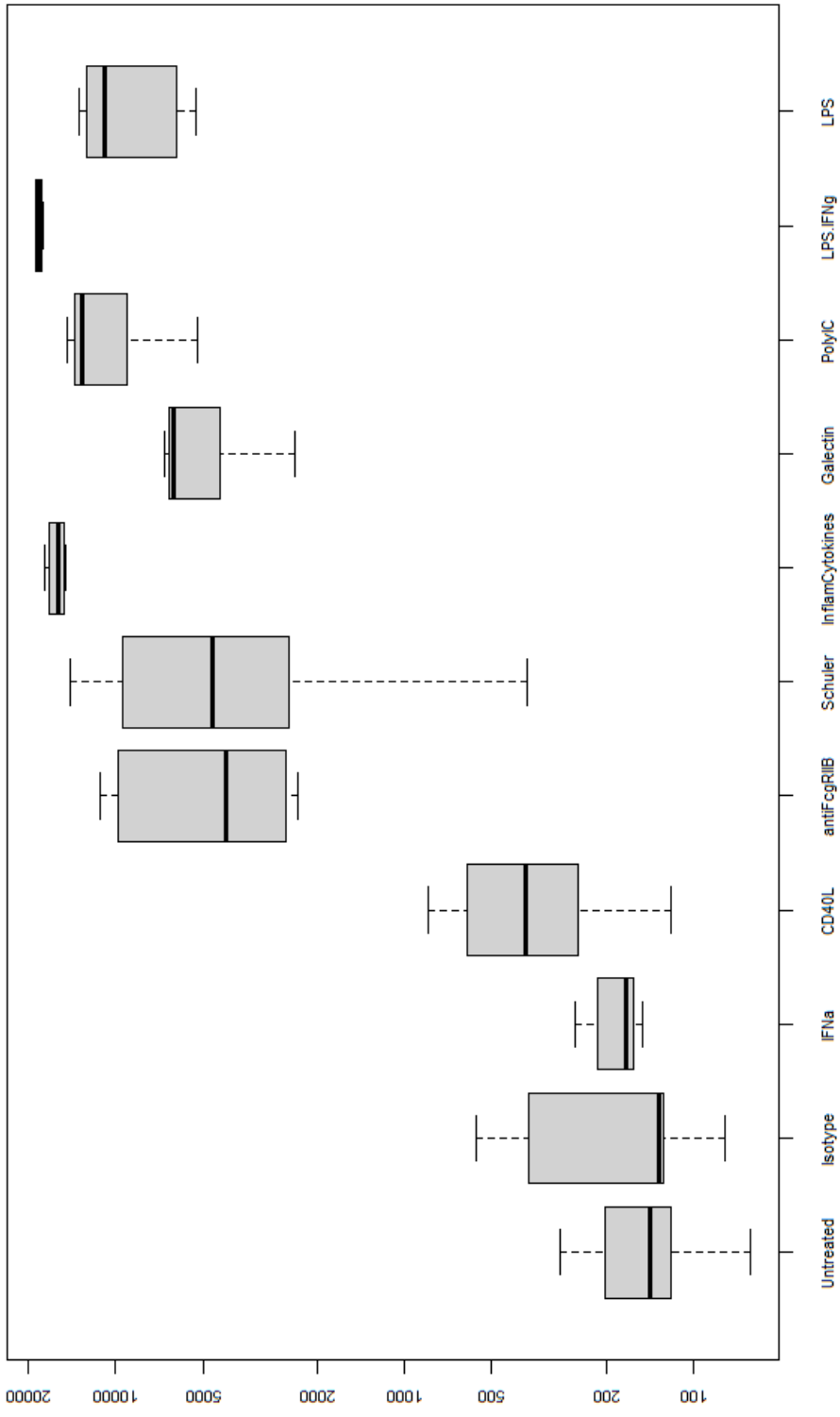

Figure 7.9: This image compares the expression value of INDO across various treatments. Since INDO plays an important role in Treg induction, it is crucial to focus on treatments that successfully mature DCs but generate poor INDO expression.

8. EXPERIMENT 2: CLASSIFICATION OF T CELL SUBTYPES

In Experiment 2, GEO was queried for data that explored the differential expression of genes across T cell subtypes. Generating a single panel where T cell subtypes are compared may help understand how Tregs are differentially expressed from other T cell subtypes. Naturally occurring Tregs have immunosuppressive properties. Under normal conditions, they play a key role in maintaining tolerance to self-antigens. Tregs have very similar expression patterns to that of other T cell subtypes, i.e. CD4+ [56]. What differentiates them from other T cell markers is the expression of the transcription factor, FoxP3 [56, 78]. Treg activity is crucial in maintaining self-tolerance; however, this suppressive activity becomes counterproductive during an immune response against tumor cells [79].

In cancer immunotherapy clinical trials, the use of DC based vaccines has shown promising results. The effectiveness of immunotherapy lies in the fact that treatments are based on inducing, enhancing or suppressing an immune response. The maturation of DCs is an important factor for generating immunotherapy vaccines; however, the suppressive activity of Tregs impedes the effector function of immune system cells. It is imperative to maintain low levels of Tregs when using an immunotherapy vaccine in order to achieve an enhanced immune response. A better understanding of how maturation treatments affects Tregs may help identifying potential biomarkers that can be targeted to suppress the induction of Tregs [56]. Therefore, it has become clear that to

generate an effective immunotherapy vaccine, induction of Tregs to the tumor site must be suppressed [51, 56].

The various mechanisms by which mature DCs induce Tregs have been assessed in detail. Since Tregs share similar expression patterns to that of other T cell subtypes, in Experiment 2, I decided to explore how Tregs were differential expressed in comparison to other T cell subtypes. For Experiment 2, I focused on the classification of T cell subtypes. To do this, more than one GEO experiment must be integrated, creating another example of how an ontological representation can be used to curate a DataSet outside GEO. Ninety-six samples from three GEO Series records were consolidated. The experiments contain conventional T cells (Tconvs: Th0, Th1, and Th2), naïve T cells, natural Tregs (nTregs), iTregs, and effector T cells. In the first study by Prots et al. [78], researchers assessed the development of Tregs and compared the gene expression of induced Tregs (iTregs) to that of naturally occurring Tregs (nTregs). The second study by Geffers et al. [80] also assessed the gene expression of iTregs when treated with different activation agents. The third study by Lund et al. [81], evaluated the differential expression of Th1 and Th2 cells in the presence of TGFβ [78, 80, 81]. The purpose of combing these three GEO studies was to see if comparing T cell subtypes would provide a better understanding on how Tregs are differentially expressed in comparison to other T cells.

Methods

The data integration component to combine T Cell samples is similar to the methods in Experiment 1. The components that are unique to Experiment 2 are explained in the following sections.

Build Ontology

An ontology is created for each experiment to reflect the different parameters used to annotate each experiment; however, the framework for the ontologies is generalized. Each ontology uses the same set of OWL *classes* and OWL *object properties*. The cell ontology is imported and each ontology is provided a framework that includes prior knowledge. However, the prior knowledge information used to build the subclasses varies according to cell types, treatments, and machine learning conditions.

1. Steps 1 thru 4, and 8 are same as Experiment 1. Since the framework is generalized, specific methods to disjoin classes and adding samples as instances are not repeated.
2. The OWL class *PBMC* is added as a subclass to *cell in vitro*. Since this experiment deals with T cells, *purifiedTCell* is added as a subclass to *PMBC*. The T cell subtypes are defined under the *purifiedTCell* class using surface markers. Under the *CD4-positive_CD25-negative* subclass, *undifferentiated T cells, Th0, Th1, Th2* are defined. Under the *CD4-positive_CD25-postive, effector T cells*, *nTregs* and *iTregs* are defined.
3. The *treatment* contains a subtype *TCellTreatment*. The T cell treatments used in Experiment 2 are control treatments and treatments used to generate specific T cell subtypes. These subclasses have individuals associated to them as OWL members. The *Control* subclass has one member, *untreated*. The *inducedTregTreatment* also has one member, *IL4*. The *naturalTregTreatment* has two members: *antiCD3+antiCD28* and *antiCD3+TR66+IL2*. The members of the *TconvsTreatment* subclass are the following treatments: *antiCD3+antiCD28,*

antiCD3+antiCD28+IL12, antiCD3+antiCD28+IL4, antiCD3+TR66+IL2 and *TGFβ*.

4. The subclasses defined under the *condition* OWL class for Experiment 2 are *conventionalTcells, naïveTcells, Teffs,* and *Tregs*.
5. *Sample* is the final class that completes the base ontology. Experiment 2 annotated from this ontology is a subclass of *ML-Experiment. TCellClassificationAffyU133A* contains equivalency classes associated by all four conditions separated with or limitation.

Similar to Experiment 1, the asserted ontology for the *TCellClassificationAffyU133A* experiment is displayed in Figure 8.1.

Acquire Data

The GEO Series records in Experiment 2 are analyzed in a different Affymetrix Platform; however, they are stored in the same format as Experiment 1. The GEO Series records for Experiment 2 are: GSE24634, GSE13017, and GSE2770 (see Figure 8.2).

Organize Data

The steps to organize the data do not change for Experiment 2. The only observed change was that GSE13017 and GSE2770 contained samples analyzed in different Affymetrix platforms. Since this was handled when parsing the XML file and building the ontology, multiple platforms did not alter the methods. Figure 8.3 is an example of how the samples are organized in the user's local directory.

Process Data

Same as Experiment 1.

Analyze Data

Three machine learning classification algorithms, SVM, Naïve Bayes and Decision Tree, were compared to see which is able to successfully differentiate between the four T cell subtypes. For each classification method I also perform LOOCV to estimate the generalization error.

Using the Decision Tree algorithm, features are selected using forward selection. The ReliefF Ranking method is used to rank features that best aid in informing the classification accuracy. Based on the classification algorithm that performs the best, the data that contain the reduced feature set are analyzed. Once the optimum accuracy is reached, a cut off threshold is set.

Results

Inferred Ontology

Similar to Experiment 1, the inferred ontology is generated using the HermiT reasoner. Due to scaling purposes, only a few samples for each condition are shown in Figure 8.4.

Analyzing Complete Data Sets In Experiment 2

Using ML-Flex

In Experiment 2, three machine learning algorithms were compared in order to accurately classify the T cell subtypes. The 96 samples contain four different types of T cells: naïve T cells, conventional T cells, Tregs and effector T cells. There are 22,283

variables that denoted by a probe ID. The conventional T cell subset includes Th0, Th1, and Th2 cells; there are 58 samples in that category. There are 11 naïve T cell samples, 16 samples of Tregs, and 11 samples of effector T cells. The confusion matrix is used to check whether each machine learning class was correctly predicted. When classes were not accurately predicted, the confusion matrix allows for the identification of which samples were misclassified. In this *ML-Experiment*, predictions can be made using the different machine learning algorithms, and this is accomplished by using the ensemble-learning methods implemented in ML-Flex [64].

Next, the seven ensemble-learning methods were compared to see which ensemble method was able to accurately classify the T cell subtypes (described in the OBDI, A Novel Pipeline chapter). In this experiment, three machine learning algorithms are specified in the settings; hence, ensemble learners will aggregate across the three learners. The ensemble-learning method in ML-Flex combines individual predictions to generate a single prediction. For this analysis I first evaluate the results based on the individual algorithms. A detailed summary of each analysis is stored in the result folder and can be accessed using a web browser.

When all the features (22,283) are included in the analysis, SVM and Select Best Ensemble Learners perform better than the other machine learning algorithms implemented in the analysis. Figure 8.5 shows the classification accuracy of the machine learning algorithms used to successfully classify T cell subtypes across three GEO experiments.

Probe Level Analysis of T Cell Subtype Classification Across Three GEO Experiments

The SVM algorithm was used to select a subset of features that are differentially expressed in five T cell subtypes: naïve T Cells, undifferentiated T Cells, Tconvs, Tregs and Teffs. Feature selection is performed using SVM, Naïve Bayes and Decision Tree. The optimum classification accuracy is achieved by using fewer attributes when SVM is used to perform forward selection.

The differential expression of the 123 features does not change vastly between naïve T cells and undifferentiated T cells. When comparing the differential expression of genes across two Treg cell subtypes, nTregs and iTregs, the expression of all genes are upregulated in iTregs; although the expression patterns between the two subtypes are very similar. The Treg subtypes in this analysis included nTregs and iTregs, both of which express the CD25 cell surface marker. However, iTregs are induced peripherally, outside the thymus, and unlike nTregs, do not necessarily require the costimulation of CD28 for their development and function [82]. Furthermore, the expression patterns between Teffs and iTregs largely overlap, thus it is difficult to determine differential expression between the two subtypes. Both Teffs and iTregs express CD25 markers on the surface; however, iTregs are treated with IL4 express the FOXP3 transcription factor. [78].

There are several mechanisms that play a role in the induction of Tregs by DCs. These mechanisms have been previously discussed in Experiment 1. Vitamin D is one of the four mechanisms by which DCs induce Tregs. The active form of Vitamin D, 1a, 25-dihydroxyvitamin D3 (VD3), inhibits the maturation and the differentiation of DCs by

downregulating key costimulatory molecules, such as CD80, CD 40, and CD86. The receptor of VD3, 1α**,** 25-dihydroxyvitamin D3 Receptor (VDR), is expressed on many immune system cells including T cells. After activation, VDR is present in CD8+ and CD4+ T cells [83, 84]. Given that the binding of VD3 and VDR may play a role in the induction of Tregs, I compared the expression of VDR across T cell subtypes in Figure 8.6. Through the meta-analysis performed in Experiment 2, it is evident that the mean expression of VDR is higher in iTregs treated with IL4 and Teffs. The binding of VD3 to VDR promotes the expression of FoxP3 which is characteristic of iTregs [85].

Figure 8.6 shows the expression of VDR across T cell subtypes. A two-tailed t-test is performed to check for the significance of VDR expression between naïve T cells and CD 25+ T cells. When compared to naïve T Cells, the expression of VDR in Teffs (p-value = 2.8e-6), and iTregs treated with IL4 (p-value = 9.0e-7) are significantly higher. However, when naïve T cells are compared to nTregs treated with IL2, the expression VDR does not change significantly (p-values = 0.41).

Interaction of Kynurenine and AHR

The elevated expression of INDO in Experiment 1 led to exploring the kynurenine pathway and how it may affect other cells in the tumor microenvironment. In the kynurenine pathway, Tryptophan metabolized by INDO and kynurenine is the first metabolite of this pathway. A further literature review was done to check if kynurenine plays a role during antitumor immune system response. The metabolism of tryptophan and the generation of kynurenine in the environment is related to the proliferation of iTregs. This is mediated by the interaction of kynurenine and the Aryl Hydrocarbon Receptor (AHR). The binding of kynurenine to AHR in T cells leads to the differentiation

of CD25+ FoxP3+ Tregs [86, 87]. However, the lack of AHR in T cells prevents the interaction of AHR and kynurenine; therefore, preventing the generation of Tregs. In the presence of TGF-β, FoxP3+ Tregs are induced to suppress function of CD4+ and CD8+ effector T cells [86]. Although AHR is not part of the feature list, in Figure 8.7 I explore the expression of AHR across the T cell subtypes in the integrated OBDI data set.

Discussion

To continue exploring the hypothesis that surrounds elements in a tumor microenvironment, I combined samples that allowed the exploration of how T cell subtypes may be differential expressed. The results from Experiment 1 directed a hypothesis where DCs have an increased expression of INDO, thus metabolizing tryptophan and inducing Tregs by promoting the production of TGF-β. The results in Experiment 1 also lead to other markers that play a role in DC based induction of Tregs, such as, CD274.

The samples that were combined in Experiment 2 were based on a targeted hypothesis generated from the results in Experiment 1. When the combined samples from Experiments 2 were analyzed in silico, VDR was present as part of the 123 feature selected list. The results show that there are high levels of VDR expressed in Tregs treated with IL4 as compared to naïve T cells and Tconvs.

I also explore the expression of AHR across T cell subtypes. AHR plays a role in inducing Tregs by interacting with a specific ligand, kynurenine, the first metabolites generated by the breakdown of tryptophan by INDO. It is observed that CD4+ Th cells treated with TGF-β have a higher mean expression of AHR compared to iTregs treated with and IL4 or nTregs treated with IL2. Further investigations can be tested in a

laboratory environment where IL2 or IL4 is added as a secondary treatment to reduce the expression of AHR in T cells; therefore, inhibiting the interaction of kynurenine and AHR.

It is evident that the metabolism of tryptophan mediated by INDO through the kynurenine pathway plays a role in negatively regulating the immune system response in a tumor microenvironment [86-88]. Using OBDI to integrate data across GEO experiments allowed us to generate and investigate this hypothesis using in silico analysis. The expression of AHR in T cells treated with TGF-β shows that these cells may have a higher affinity to bind to kynurenine and induce FoxP3+ iTregs.

Figure 8.1: This figure shows the asserted ontology that helps integrate three GEO experiments that help classify various T cell subtypes. The equivalency for the conditions, Teffs and Tregs, are not shown above. All conditions are generated using properties associated to cell type and treatment. *PBMC: Peripheral Blood Mononuclear Cell.

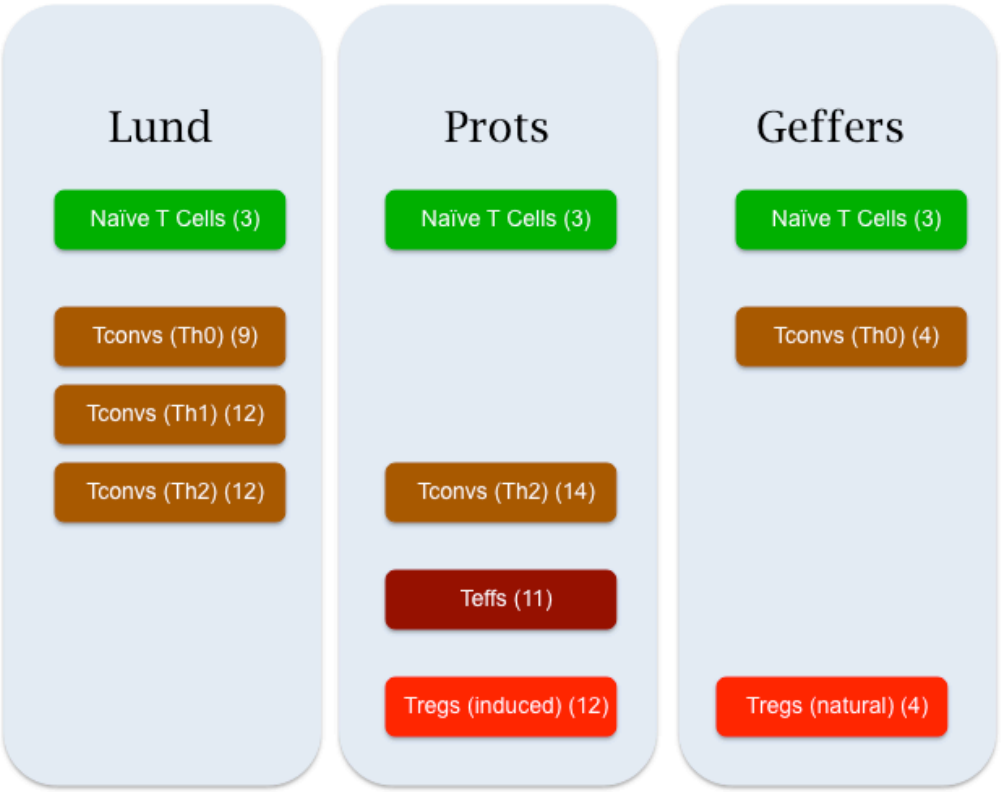

Figure 8.2: The samples highlighted in green are native T cells that do not contain any stimulants. The samples highlighted in brown are Tconvs treated with specific stimulants that allow for differentiation of T cells. The samples highlighted in red represent natural and induced Tregs. Finally, effector T cells (Teffs) represent a population of cells that are CD25+ but are not of regulatory function. LOOCV is used to analyze the samples; therefore, a hold-out test set (similar to Experiment 1) is not provided.

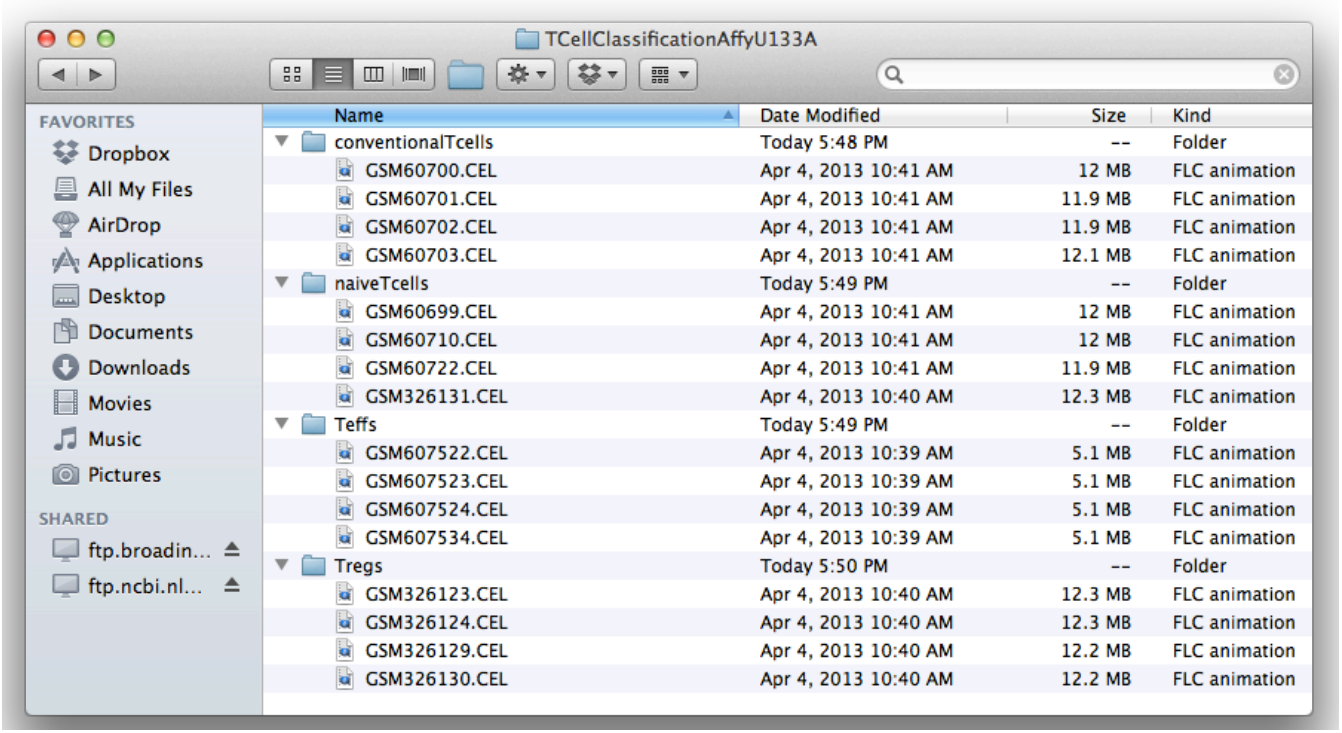

Figure 8.3: The image is a screenshot of directory structure created for Experiment 2. Only four samples per machine learning condition are shown for scaling purposes.

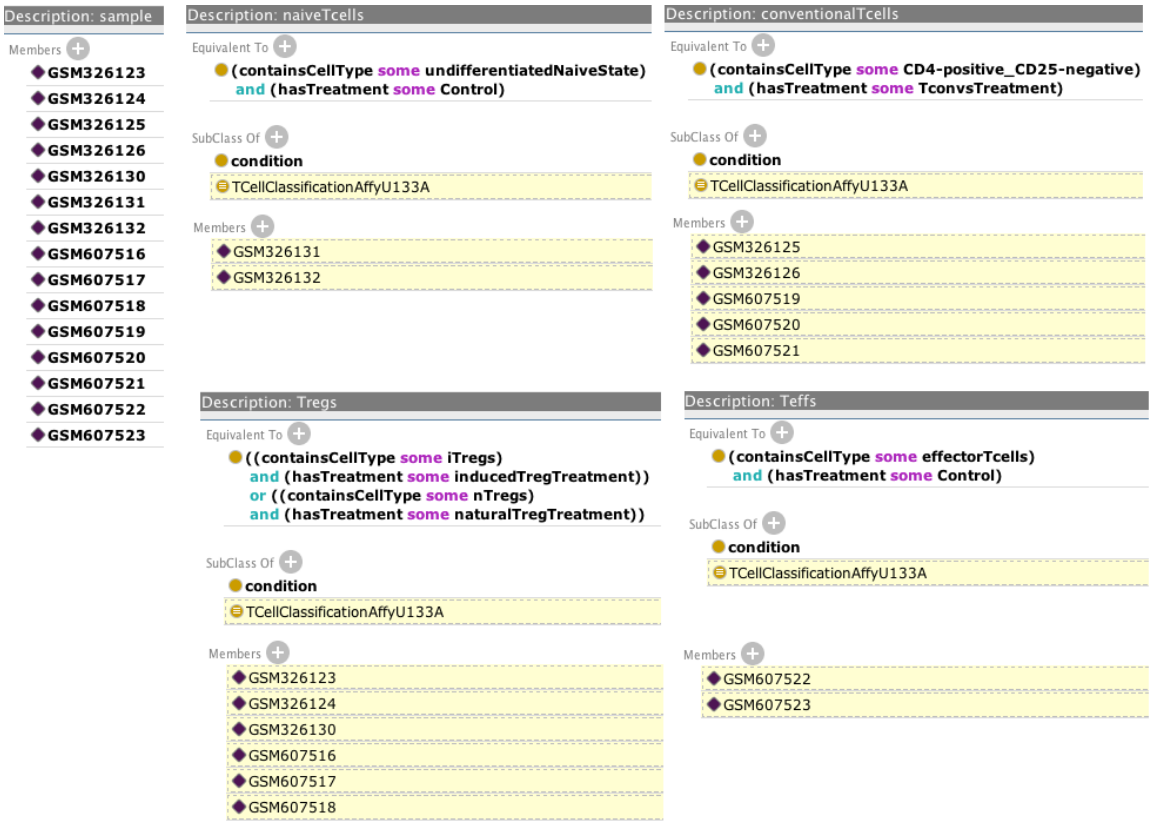

Figure 8.4: Inferred ontology generated for Experiment 2. On the left, 15 samples are shown before they were reasoned into specific conditions. This is done across 96 samples that are not shown in this image.

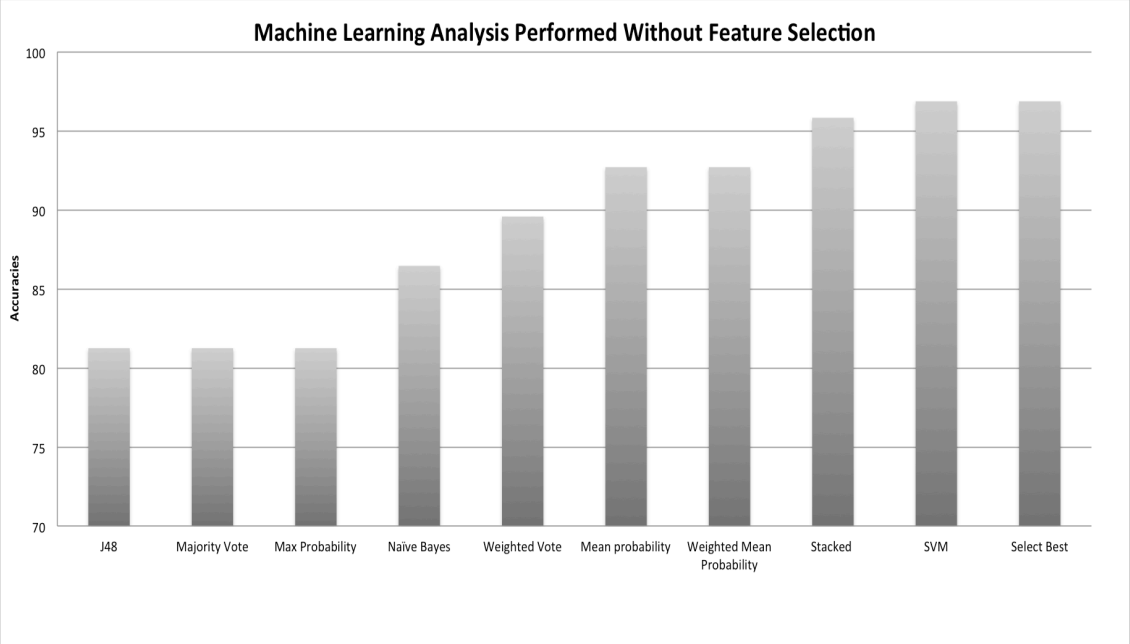

Figure 8.5: ML-Flex makes it easy to compare the performance of different algorithms. Each machine learning algorithm is separately assessed to evaluate how the algorithms perform in predicting T cell subtypes. The performances of the ensemble methods are plotted in the same graph. The accuracies are plotted in the graph above.

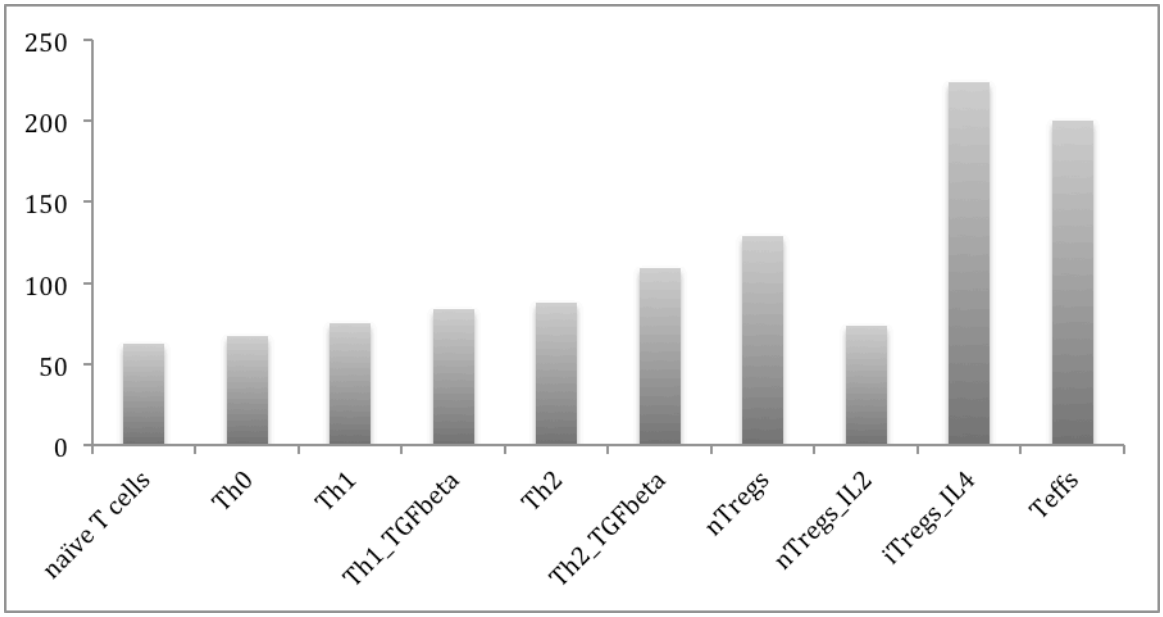

Figure 8.6: This image explores the expression of VD3 across T cell subtypes. The expression of VD3 is higher in samples that express CD25+ on the cell surface. These include samples of nTregs, iTregs treated with IL4, and Teffs. A two-tailed t-test is performed to check for significance.

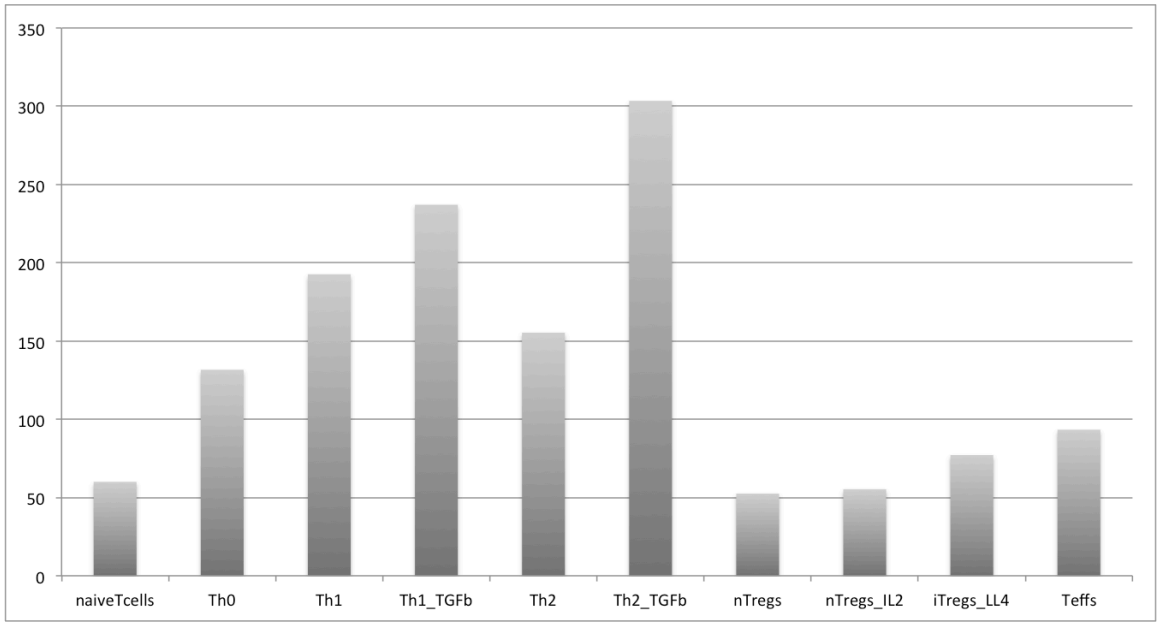

Figure 8.7: The average expression of AHR is plotted across T cell subtypes. It is noted that T cells treated with TGF-β have a higher expression of AHR versus T cells treated with IL2 and IL4.

9. EXPERIMENT 3: CHARACTERIZATION OF CANCER CELL LINES

To further evaluate the components of a tumor microenvironment using OBDI, I included samples from comparing the gene expression across cancer cell lines. These samples were analyzed on a different gene expression platform; therefore, generalizing OBDI to other microarray platforms. Based on the meta-analysis in Experiment 1 it was evident that INDO was differentially expressed in DC treated with different maturation stimuli. The metabolism of tryptophan through the kynurenine pathway plays a role in inducing the Tregs; thus, inhibiting the maturation of DCs and CD 8+ T cells. This causes the DC-based vaccine to be ineffective. However, when comparing the maturation treatments in Experiment 1, it was noted that IFNα successfully matured DCs but the expression of INDO was downregulated in DCs treated with IFNα. This led to setting up the hypothesis for Experiment 3.

Relating a specific DC maturation treatment, IFNα, to clinical research where cancer cell lines are treated with IFNα at different time points, makes OBDI an effective tool to analyze data in translational research. It has been documented that the use of IFNα in conjunction with chemotherapy and cancer vaccines overcomes tumor-induced immunosuppression by improving the outcome of immunotherapy [89]. Using IFNα as an adjuvant therapy has been show to improve disease free survival in patients with high-risk cutaneous melanoma [90]. IFNα is a type I interferon that is mainly produced by

macrophages. In patients with high-risk melanoma, IFNα is the only adjuvant therapy that is currently approved. IFNα had antitumor effects in preclinical and in clinical models; however, the mechanistic approach of how IFNα treatment results in an antitumor response is not well understood [91-93].

Methods

To better understand how IFNα plays a role in antitumor response, meta-analysis was performed on 10 cancer cell lines treated with IFNα. Exploring the immune system response across cellular and disease states gives a more comprehensive understanding of immune system response in a tumor microenvironment. For Experiment 3, I queried GEO for specific experiments where cancer cells were treated with IFNα2a. To explore how genes may be differential expressed in cancer cells treated with IFNα2a, a GEO experiment where four cancer cell lines were treated with IFNα at different time points was analyzed through the OBDI pipeline.

Build Ontology

1. Steps 1-4, and 8 are same as Experiment 1. Since the framework is generalized, specific methods to disjoin classes and adding samples as instances are not repeated.
2. The OWL *class cancer_cell_lines* is added as a subclass to *cell in vitro*. Since Experiment 3 deals with the differential expression across melanoma cell lines, the cell type information is acquired from the paper associated to the GEO Series record [94]. Subclasses to *cancer_cell_lines* are *colon, melanoma, lung,* and *pancreas.* The specific cell lines are defined under each subclass (See Figure 9.1).

3. The *treatment* contains a subtype *CancerCellLineTreatment*. The treatments used in Experiment 3 are *control* and *experimentalTreatment.* These subclasses have individuals associated to them as OWL *members.* The *control* subclass has one *member*, *control_medium*. The *experimentalTreatment* has two *members: IFNα-2a_24Hr* and *IFNα-2a_4H*.
4. The subclasses defined under the condition OWL *class* for Experiment 3 are based on the tissue of origin: *colonControl, colonIFNα-2a*, *endothelialControl*, *endothelialIFNα-2a*, *lungControl*, *lungIFNα-2a*, *pancreasControl*, and *pancreasIFNα-2a*.
5. *Sample* is the final class that completes the base ontology. *CancerCellLineComparisonIlluminaBeadchip* contains equivalency classes associated by all eight conditions separated with an "or." The asserted ontology is displayed in Figure 9.1.

Acquire Data

The third platform used to validate the OBDI methodology is the Illumina Expression BeachChip. Data from a single melanoma experiment are preprocessed and organized for machine learning analysis [95]. The GEO Series record used in Experiment 3 is GSE 21158 (see Figure 9.2).

Organize Data

Organizing the melanoma GEO samples vary since the experiment was analyzed using the Illumina Expression BeadChip. The previously described Affymetrix platform uses a glass or silicone chip where the probes are attached; however, the Illumina

technology uses microscopic beads that are associated to a specific probe. To further explore the immunological space of cancer immunotherapy, I chose to focus on expression array samples where different melanoma cell lines are treated with Interferon-alpha (IFN-alpha). There are 10 different melanoma cell lines that are treated with 10 U/ml IFN-alpha for 4 hour and 24 hours, respectively. Each cell line has associated control samples that were not treated with IFN-alpha.

Process Data

The samples retrieved from GEO experiment, GSE21158, were normalized using quantile normalization [63]. The data were downloaded to the pipeline by interfacing with Gene Pattern. Once data were available locally, each sample was separated into columns and stored as an individual file. The samples are organized using the ontology and the metadata XML file.

The sample for each experiment is formatted differently in GEO because of the platform specification used to generate the high throughput data. In order to perform meta-analysis for the Illumina BeadChip experiment, the data are directly imported into a Gene Pattern module. Unlike the samples analyzed from the Affymetrix chip, individual samples from the Illumina BeadChip GEO experiment cannot be downloaded. The GEOImporter module downloads the experiment from GEO and the file is temporarily stored in the Gene Pattern file format [10]. When using Gene Pattern to process samples, an annotation file, like the clm file, can be used to annotate the GSM IDs with appropriate machine learning conditions. However, The GEOImporter module does not account for an annotation file, like the clm file. To account for this, I incorporated an annotation file into the pipeline to convert GSM IDs into corresponding machine learning

condition. The annotation is created using the reasoned OWL file for the experiment labeled *CancerCellLineComparisonIlluminaBeadchip*. Although the GEOImporter module does not allow for the replacement of GSM IDs with machine learning conditions, it simplifies data processing and acquisition by creating a gct file that contains all samples from the melanoma experiment, GSE20156. The same method is used to covert Gene Pattern gct files into ARFF files. Once the ARFF file is generated, meta-analysis can be performed using ML-Flex.

Analyze Data

Same as previously described (see Experiment 2).

Results

Inferred Ontology

Inferred ontology serves as a backbone for allocating appropriate samples into their respective machine learning conditions. This is done in a standardized way since the various parameters of a GEO Series record are store as OWL entities. Figure 9.3 displays how the reasoner can be used to infer the samples into the appropriate conditions.

ML-Flex Results

Machine learning analyses similar to Experiment 2 are performed using the cancer cell line data. The full data set was used to analyze 90 cancer samples using individual machine learning algorithms and ensemble learners. Without performing feature selection, the ensemble learners and SVM perform better than Naïve Bayes or Decision Tree. Figure 9.4 displays the accuracy for each algorithm when no feature

selection is used on the 90 samples. The meta-analysis for Experiment 3 did not yield high classification accuracy, and the feature list contained an extensive list of genes. In the ensemble learning methods, mean probability and weighted mean probability performs poorly with the highest classification accuracy (64.4%). SVM classifier also performs poorly with an accuracy of 63.3%.

Probe Level Analysis of Cancer Cell Lines

Feature selection was performed using the ReliefF ranking method using three different classification algorithms: Naïve Bayes, Decision Tree, and SVM. The SVM algorithm was used to perform forward selection. The feature list was reduced to 1304 attributes and the accuracy for all three algorithms increased to 80%.

The extensive feature list was searched for any genes that may be related to the mechanisms involved in the induction of Tregs, revealing kynureninase (KYNU) as part of the feature list. KYNU is part of the kynurenine pathway that breaks down kynurenine during tryptophan metabolism. Figure 9.5 depicts the expression of KYNU compared in melanoma cell lines and across treatment time points. The results indicate that when melanoma cell lines are treated with IFNα2a from 4 hours, the expression of KYNU is significantly higher (p-value = 1.47E-4) than samples treated with IFNα2a from 24 hours. The expression pattern of KYNU in melanoma cells treated with IFNα2a for 24 hours is similar to untreated melanoma cell lines (control).

Discussion

From the analysis performed in Experiment 1, it was evident that INDO plays a role in inhibiting the effectiveness of cancer immunotherapy. INDO is also differentially

expression in DCs treated with different maturation stimuli. To promote translational research in the field, cancer cell lines treated with IFN α 2a are analyzed using OBDI. IFNα2a is similar to the IFNα treatment used to mature DCs, while downregulating the expression INDO. These findings were supported by the integrated analysis made possible by using OBDI.

To overcome tumor-induced immunosuppression, IFNα can be used as an inducer of DCs in cancer vaccines [89]. Injecting patients with IFNα-matured DCs in conjunction with chemotherapy is designed to suppress Tregs, while creating an environment for DCs to take up antigens and present them to T cells in lymph nodes. Patients treated with adjuvant IFNα-2b who had a measurable autoimmune response had higher probability of relapse-free survival, as well as a higher probability of overall survival [92]. The suppression of Tregs is important in the effectiveness of DC based vaccines; however, an enhanced CD8+ T cell response has been observed in stage IV melanoma patients vaccinated with melanoma-associated peptides near local lymph nodes, in conjunction with adjuvant IFNα injections [91].

Kynurenine is metabolized by INDO in order to drive the metabolism of tryptophan. The in silico analysis in this experiment generated an extensive feature list where, KYNU, an enzyme that breaks down kynurenine, was differentially expressed in melanoma cell lines treated with IFNα2a for 4 hours and melanoma cell lines treated with IFNα2a for 24 hours. KYNU plays a role in the further breakdown of kynurenine into 3-Hydroxyanthranilic Acid (3-HAA). A treatment of 3-HAA has shown to drive the production of TGF-β; therefore, inducing Tregs mediated by the production of TGF-β [57, 96]. 3-HAA also plays a role in impeding the antitumor immune system response by

inhibiting the antigen dependent proliferation of CD8+ T cells [96]. When melanoma cell lines are treated with IFNα2a a longer time point (24 hours), the expression of KYNU is reduces. Lower expression of KYNU can affect the production of 3-HAA and thereby reducing the production of TGF-β in the tumor microenvironment. A reduced amount of TGF-β may inhibit the proliferation of Tregs.

The in silico findings from analyzing cancer cell lines can lead to designing laboratory experiments to further explore the role of tryptophan metabolism via the kynurenine pathway in immune system response [91, 93, 96, 97].

Figure 9.1: The asserted ontology for the Cancer Cell Line Experiment has a condition associated to each cell line. The equivalency condition for melanoma cell lines are defined using the specific cell lines defined under *cell in vitro* and using the treatments used on each cell line.

Figure 9.2: In Experiment 3, there are 90 samples that can be classified into four different cell lines. Each cell line is treated with IFNa2a for 4 hours and 24 hours, respectively.

Figure 9.3: The inferred ontology displays how samples are added to specific conditions based on their defined equivalency rules.

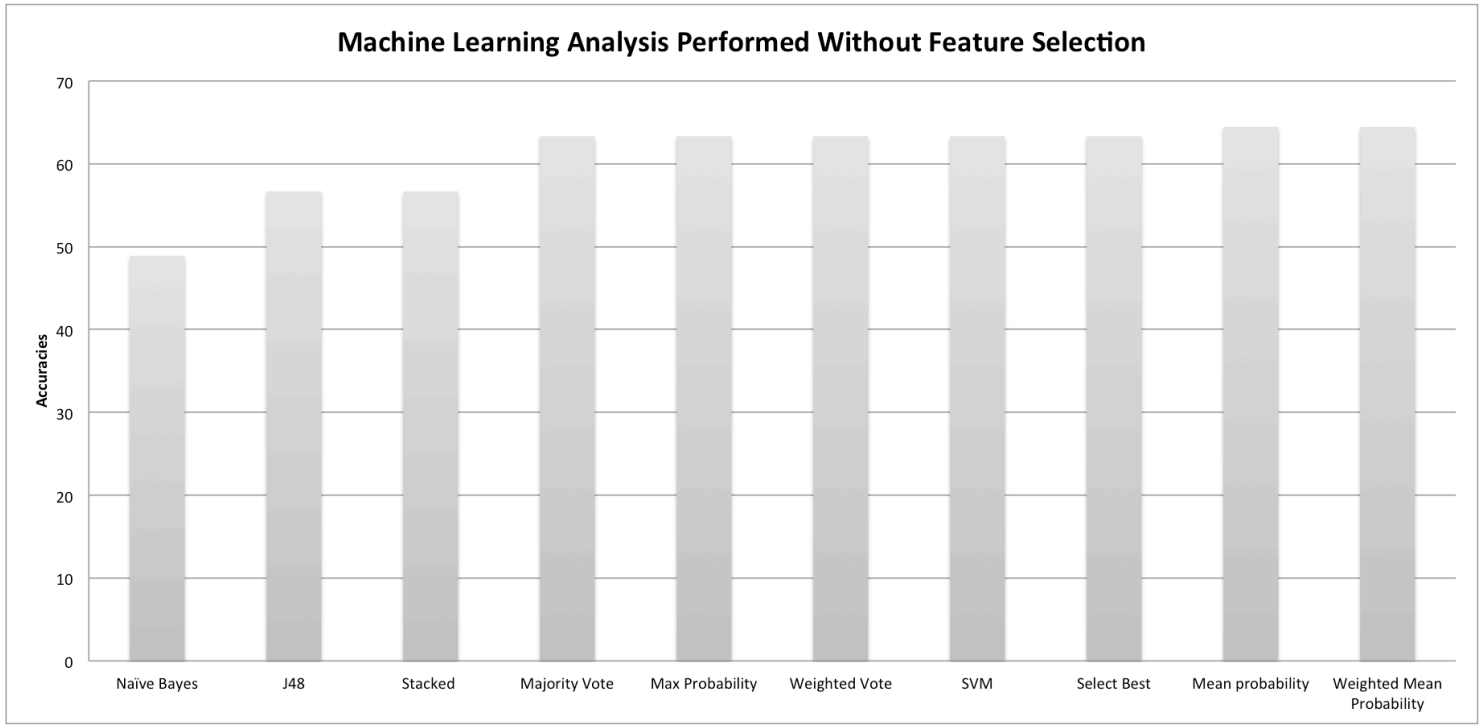

Figure 9.4: The image displays the accuracies for the machine learning classifiers and the ensemble learners.

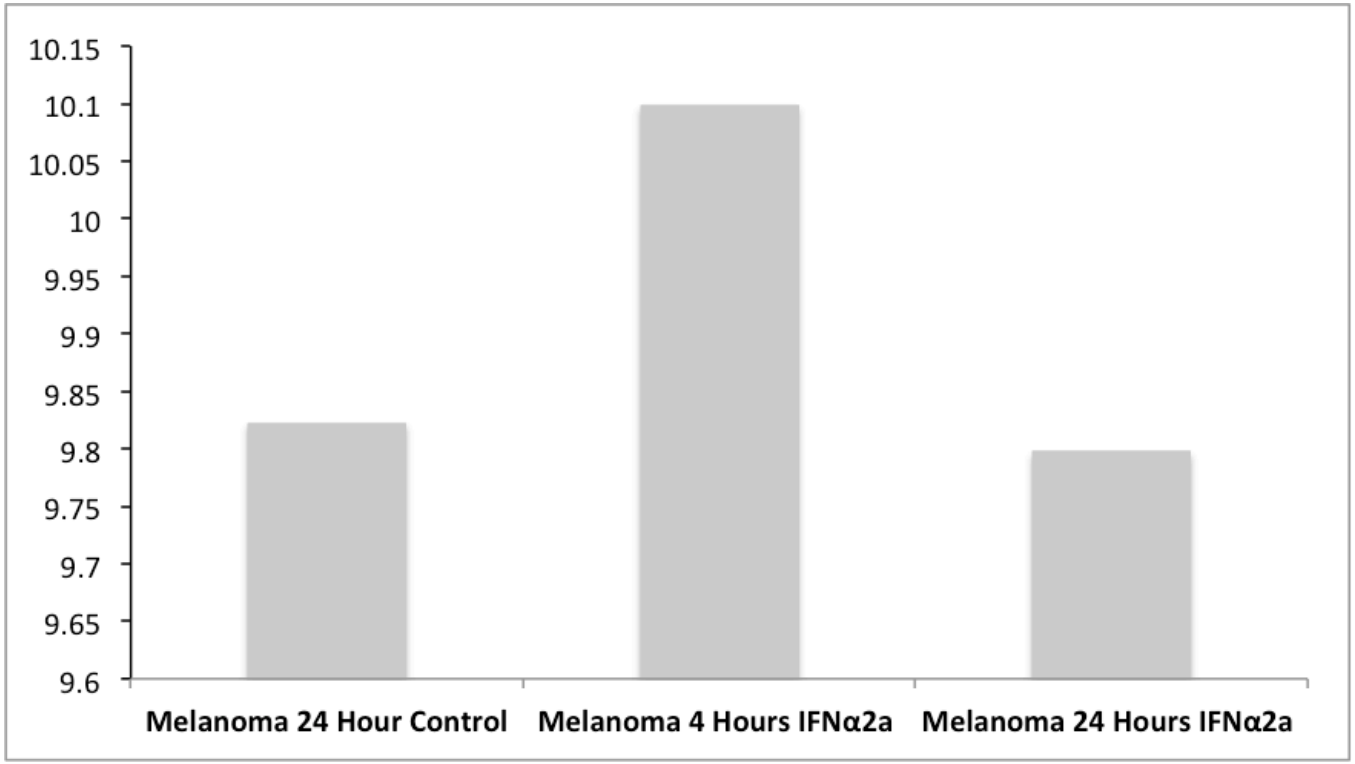

Figure 9.5: The bar graph displays the expression of KYNU across melanoma cell lines. A t-test is performed to compare the gene expression of KYNU in untreated cancer cell lines and cell lines treated with IFNα2a at two different time points, 4 hours and 24 hours.

10. EXPERIMENT 4: RNA-SEQUENCE DATA ANALYSIS

The fourth experiment was chosen to explain how data from RNA-Seq experiments can be successfully analyzed using OBDI. Differential expression analysis is widely performed using microarrays and beachip; however, RNA-Seq is gaining popularity among researchers interested in performing differential expression analysis. RNA-Seq experiments require lesser quantity of RNA to run the laboratory experiments and produce results with higher sensitivity. RNA-Seq analysis can play an important role in discovery-based experiments. In order to make the OBDI pipeline up-to-date, it was important to explore how RNA-Seq could be incorporated within the pipeline.

Methods of genomic and mutation analysis have given us a better understanding of cancer genetics, but they provide limited insight on mRNA-based interpretation during tumorogenesis [98, 99]. Gene fusions are commonly associated with cancers, where two previously separate genes come together. They can occur due to chromosomal activities such as: translocations, insertions, deletions, and inversions [95]. The sequencing of short reads can recapitulate microarray expression predictions and also provide additional information that cannot be obtained by microarray methodologies. Sequencing short reads can provide thorough information about the existence of spliced variants, which occur during a regulated process where parts of an exon may be included or spliced out [100, 101].

RNA sequencing (RNA-Seq) is a high throughput methodology used to sequence the cDNA [102]; whereas, microarrays are tools used to analyze gene expression. Compared to microarray technologies, RNA-Seq has higher sensitivity, requires smaller amounts of RNA and has a larger dynamic range which can contribute to differential expression analysis [100].

Since the OBDI pipeline was developed to integrate and process high throughput data, a GEO experiment, containing a small set of RNA-Seq samples, was used as a proof of principle study to evaluate the RNA-Seq analysis integration within the pipeline. There are several modules in Gene Pattern that help with preprocessing of sequencing. These modules are integrated into the OBDI pipeline allowing the user to analyze RNA-Seq data. For this analysis, I focus on RNA-Seq samples analyzed on the Illumina Genome Analyzer, where the researchers explored the genetic alterations that occur in tumor cells.

Methods

The GEO experiment contains four samples from melanoma cell lines, eight samples from patient melanoma derived short-term cultures, and two samples from leukemia cell lines. These samples were analyzed using the Illumina Genome Analyzer and the reads were aligned using the Burrow-Wheeler Alignment (BWA) tool and the human reference genomes, hg18 [103, 104]. Individual sample files are stored in BAM format [103].

In this experiment, melanoma and leukemia samples are analyzed in silico using OBDI. Melanoma short-term cultures are cell lines created from patient tumors that have undergone few passages outside the patient. A majority of melanoma short-term

cell lines proliferate readily under laboratory conditions [105] and passages refer to the splitting of cultures to allow cells to continue growing. Passages refer to the numerous cell divisions that occur during cell culturing. In Giricz et al, researchers proposed that multiple passages might contribute to subtle genomic modifications that may occur during cell culturing.

RNA-Sequencing technologies are currently being used in conjunction with microarray experiments. The processing of RNA-Seq samples varies significantly; however, the RNA-Sequencing processing module in OBDI can be used to process and analyze RNA-Seq data.

Build Ontology

1. Steps 1-4, and 8 are same as Experiment 1. Since the framework is generalized, specific methods to disjoin classes and adding samples as instances are not repeated.
2. The OWL *class cancer_cell_lines* is added as a subclass to *cell in vitro*. Since this experiment deals with the differential expression across melanoma cell cultures, the cell type information is acquired from the paper associated to the GEO Series record [95]. Subclasses related to *cancer_cell_lines* are: *blood,* and *skin.* Four cell lines are defined under both subclasses (See Figure 10.1).
3. The *treatment* contains a subtype *CancerCellCulturePassage*. The treatments used are based on the number passages involved during cell culturing. Each subclass has one OWL *individual* associated as a *member.* The *controlLeukemia* and *controlMelanoma* subclasses have one *member*, *GreaterThan30*. The *shortTermMelanoma* has one *member, LessThan20*.

4. The subclasses defined under the condition OWL *class* are *LeukemiaNormalPassage, MelanomaNormalPassage*, and *MelanomaShortTerm*.
5. *Sample* is the final class that completes the base ontology. *MelanomaRNA-Seq* contains equivalency classes associated by all three conditions separated with an "or." The asserted ontology is displayed in Figure 10.1.

Acquire Data

To develop this part of the pipeline, the entire GEO experiment file is downloaded and processed to reveal the individual BAM files. The method of preprocessing the files is similar to how the Affymetrix samples from previous experiments were handled. Figure 10.2 displays the samples that are encoded into the ontology and used to perform analysis using OBDI.

Organize Data

Methods are the same as Experiment 1. Instead of handling *cel* files, BAM files are organized.

Process Data

Once the BAM files are extracted, Gene Pattern modules are incorporated into the pipeline. There are several tools available for the analysis of RNA-Seq data. There are modules in Gene Pattern that incorporate some of the major tools used in RNA-Seq analysis: Bowtie, BWA, Cufflinks, and TopHat [10].

Since the data are already in BAM format, Cufflinks was used to generate the Fragment Per Kilobase of exon Million fragments mapped (FPKM) values for each sample [106]. BAM files are already aligned to the reference genome, so Cufflinks can

be used to test for differential expression for RNA-Seq samples. To execute this module, users must provide a Generic Feature Format (GFF) or a Genome Annotation File (GTF). These files aid in the assembly of RNA-Seq samples into transcript reads [107], which allow for the annotation of RNA-Seq samples at the level of gene information [10, 107]. The human GTF files are provided with the OBDI tool; however, if researchers use the RNA-Seq module to analyze other organisms, the GFF files can be accessed via Gene Pattern [10]. The RNA-Seq fragment counts can be used to measure the relative abundance in FPKM values [106].

Finally, Gene Pattern also contains modules that allows for creating gct files from the FPKM values [10]. This provides users with the familiar tab-delimited file that has been generated in the previous experiments. Since there are 14 gct files created for each sample, this module requires preprocessing to combine the samples into a single matrix. Once this achieved, the generalize code converts gct files to ARFF files that can be used. The RNA-Seq samples can now be analyzed in ML-Flex to perform machine learning analysis.

Analyze Data

Unless stated otherwise, all methods in Experiment 4 are identical to Experiment 2.

Results

Inferred Ontology

Similar to the previous experiment, the purpose of the inferred ontology was to accurately associate individual RNA-Seq samples to the correct machine learning

conditions. There were two melanoma cell lines in Experiment 4 that were grouped together under one condition. To reason the ontology accurately, it was important to add both cell line information in the equivalency class of the specific condition. Figure 10.3 shows how the samples are inferred into the right conditions based on the rules defined for the condition class.

ML-Flex Results

Machine learning analysis on the calculated FPKM is plotted in Figure 10.4. Mean probability performs the highest with an accuracy of 76.9% and J48 performs the lowest with an accuracy of 46.2 % classification accuracy. The SVM algorithm is used to perform feature selection along with the ReliefF ranking method. The SVM algorithm performs with an accuracy of 69.2% when no feature selection is performed. Feature selection is performed using SVM along with the ReliefF ranking method. When twelve features are selected, the classification accuracy increases to 84.6%. INDO was not part of the feature selected list; however, I evaluated the FPKM values. The FPKM values for INDO remained 0 across all samples; however, the FPKM value does increase in short-term melanoma samples that underwent 14 passages. In samples labeled GSM506411 and GSM506412, the FPKM values increased to 1.28 and .70, respectively.

Discussion

RNA sequencing analysis plays an important role in providing insight to researchers who are exploring the genetic modifications that occur in a tumor environment. By analyzing RNA-Seq using OBDI and by creating ontological

representations, I am able to make OBDI pipeline flexible to the continuing growth of genomic data. The analysis done using ML-Flex shows that RNA-Seq data can be successfully analyzed using the OBDI pipeline.

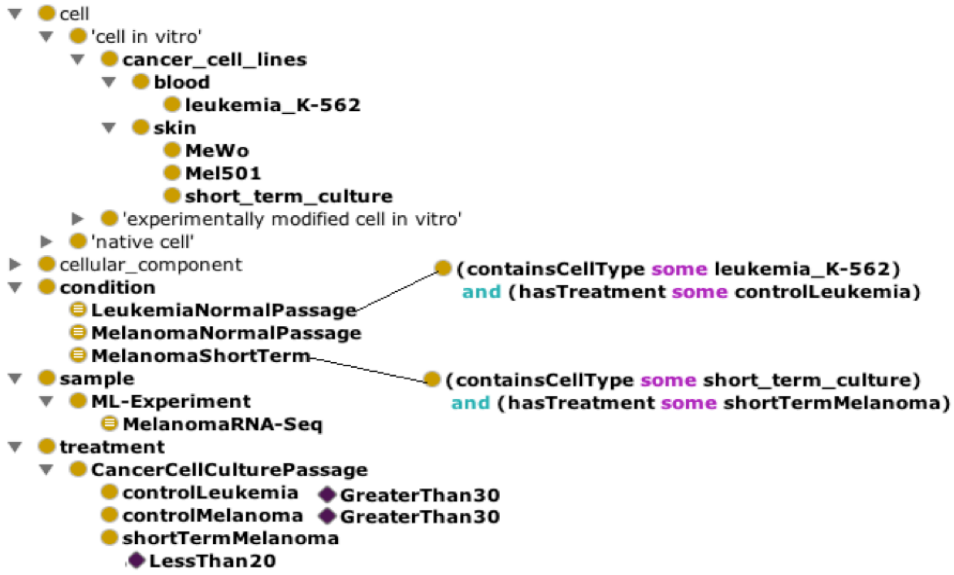

Figure 10.1: This figure shows the asserted ontology that is developed to analyze RNA-Seq data from melanoma cell lines that are cultured for a short term in a laboratory setting. The framework of the ontology is similar to that of Experiment 3. There are 14 samples (denoted by GSM IDs) associated to *MelanomaRNA-Seq*.

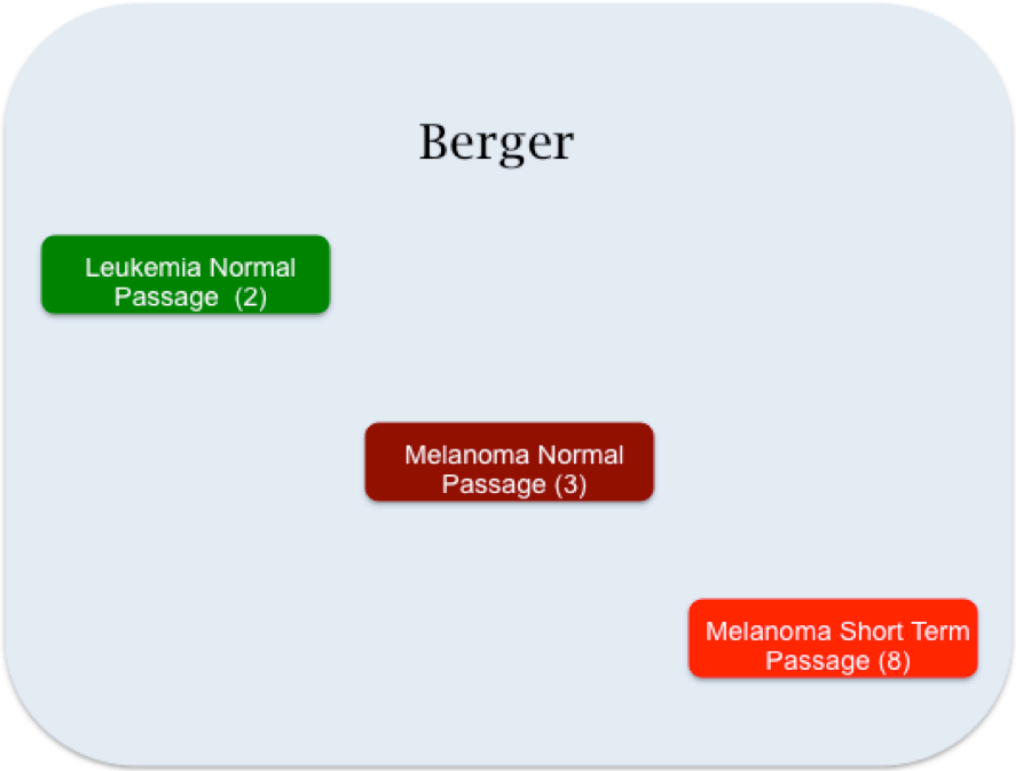

Figure 10.2: Experiment 4 serves as a proof of principle analysis. RNA-Seq samples are classified into three classes that include melanoma and leukemia samples.

Figure 10.3: The inferred ontology used in Experiment 4.

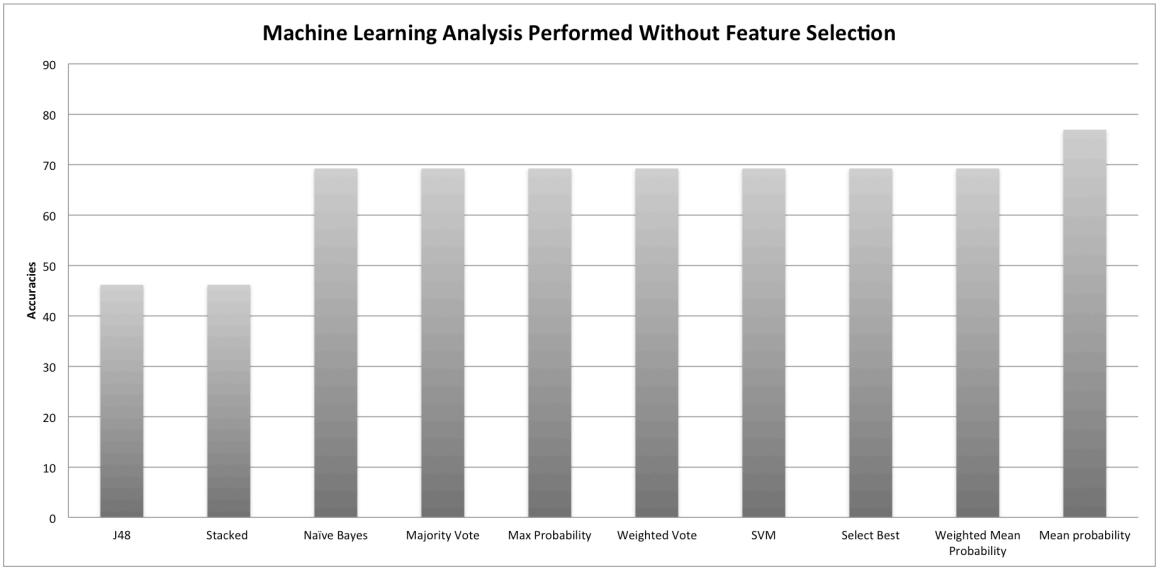

Figure 10.4: Accuracies of machine learning algorithms and ensemble learners.

11. DISCUSSION

OBDI is successfully used to integrate and analyze data across GEO experiments. OBDI is used to mobilize data in studies that exist in silos. Only 6% of samples in GEO are integrated into curated DataSets, allowing the opportunity to generate methods that aid the process of data integration across GEO experiments. OBDI is a pipeline that aids with the process of combing GEO samples; thereby, promoting knowledge discovery by perform analysis on combined studies.

The OBDI pipeline reduces the barrier of data integration by encoding various laboratory elements into consistent representation. Using ontologies to combine GEO samples allows for generating newly curated OBDI data sets. OBDI helps maintain consistency of experiments over time by reasoning over ontologies to preform analysis. Changes to the analysis for each OBDI experiment can be made by directly altering individual ontologies. For instance, the incorporation of ML-Flex allows user to explore a range of algorithms that can be used to perform differential expression analysis on the combined OBDI experiments. This allows users to keep relevant information regarding the in silico analyses in consistent representations, avoiding opportunities for errors and performing robust analysis on the integrated sets.

Adding new samples from GEO can extend the current OBDI experiments. As data exploring the field of cancer immunotherapy increases in GEO, sample information can be added into the ontology; thereby extending the number of samples integrated and analyzed in OBDI. The OBDI pipeline can also be used to explore different biological

questions by changing the domain knowledge of interest. Generalizing the methods in OBDI can help generate testable hypothesis for different biomedical research domains.

Using OBDI to integrate samples that previously existed in silos allowed me to generate a hypothesis that can drive research at the bench. Based on the results from Experiment 1, the expression of INDO across DC maturation treatment allowed me to explore the involvement of the kynurenine pathway during immune system response [97]. The downregulation of INDO in DCs treated with IFNα led to exploring the treatment of cancer cell lines with two time points of IFNα2a. Although these studies were available in GEO, without the use of OBDI, it would be difficult to integrate these samples and create in silico experiments that explored the various elements in a tumor microenvironment. Figure 11.1 summarizes the findings from Experiment 1-3 that may aid in guiding researchers to generate in vitro experiments.

In Figure 11.1, I propose a testable experiment where tryptophan is not fully metabolized; therefore, the metabolites of tryptophan, kynurenine and 3-HAA, are present in lower levels in a tumor microenvironment. This may further hinder the interaction of AHR and kynurenine and also deplete the environmental concentrations of TGF-β. The lack of TGF-β in the environment may prevent the induction of iTregs to the tumor site [57]. Since 3-HAA has shown to inhibit antigen specific proliferation of CD8+ T cells, the downregulation 3-HAA may help in generating an antitumor immune response [96]. Finally, the addition of IFNα in cancer immunotherapy along with chemotherapy treatments may help over tumor-induced immunosuppression and improve clinical outcomes [89]. Using IFNα as an adjuvant therapy has shown to have potent antitumor impact in melanoma patients. Using OBDI I am able to explore

treatment that can be translated to a clinical setting and the IFNα treatment on melanoma patients has proven to be successful in disease free survival [92, 93].

The immediate application of OBDI can be seen at the bench; however, the pipeline can be used to translate hypothesis that can drive experiments in clinical research. OBDI can help researchers create direct experiments because it increases the power of knowledge discovery by exploring the data across multiple studies.

Limitations

Since integrating data across different experiments is a complex process, there are limitations in the current OBDI methodology. Building ontologies is a complex process; however, the OBDI pipeline does add several OWL components directly into the ontology by parsing the GEO metadata. The first limitation of using OBDI is that there is a break between adding OWL elements and generating the final set of results using ML-Flex. OBDI can be executed by using two command line options. The first option allows for adding the various OWL components to the ontology. Once the entities are added, users must add relationships between OWL entities using an ontology editor, like Protégé. This task has been simplified because the metadata are added into the ontologies and will guide users to build the correct OWL relationships. The second command line option allows users to store the combined the data in the local directory, process and run the machine learning analysis on the integrated OBDO experiments. The second limitation is the sample size in each individual experiment. Due to the specificity of the biological problem assessed, finding relevant GEO experiments was a challenging task. When integrating biological data and generating a biological relevant predictive model, it is crucial to choose appropriate GEO

experiments. However, as research in the field progresses, more data from GEO or research laboratories can be integrated using OBDI to create new experiments. The sample size of the integrated OBDI experiments will grow, as more samples are made available.

Future Work

OBDI can be used to incorporate other high throughput data from different repositories. This also allows users to explore data that is being generated in different biological domains. As more samples are incorporated from different repositories, the current modules in OBDI must be expanded to other microarray and genomic platforms.

GEO is the only database that is currently incorporated in OBDI. Other databases store high throughput data that can easily be incorporated into OBDI. The pipeline requires a link associated to the raw data and the supporting metadata.

Relevance to Biomedical Informatics

OBDI offers immediate support at the bench by aiding users to generate testable hypothesis at the bench. The results at the bench can be translated to clinical settings. Combining samples across GEO experiments can be challenging but integrated sets can provide insights that were previously overlooked

In the field of biomedical informatics, translational research is a growing component where researchers are eager to incorporate the findings of a biomedical research directly to patient care. The goal of this project was to provide support for bench researchers to better understand the molecular mechanisms involved during an immune response. I hope that my informatics approach will aid research in the field of

immunotherapy vaccine development and support the development of new insights in the field of translational medicine.

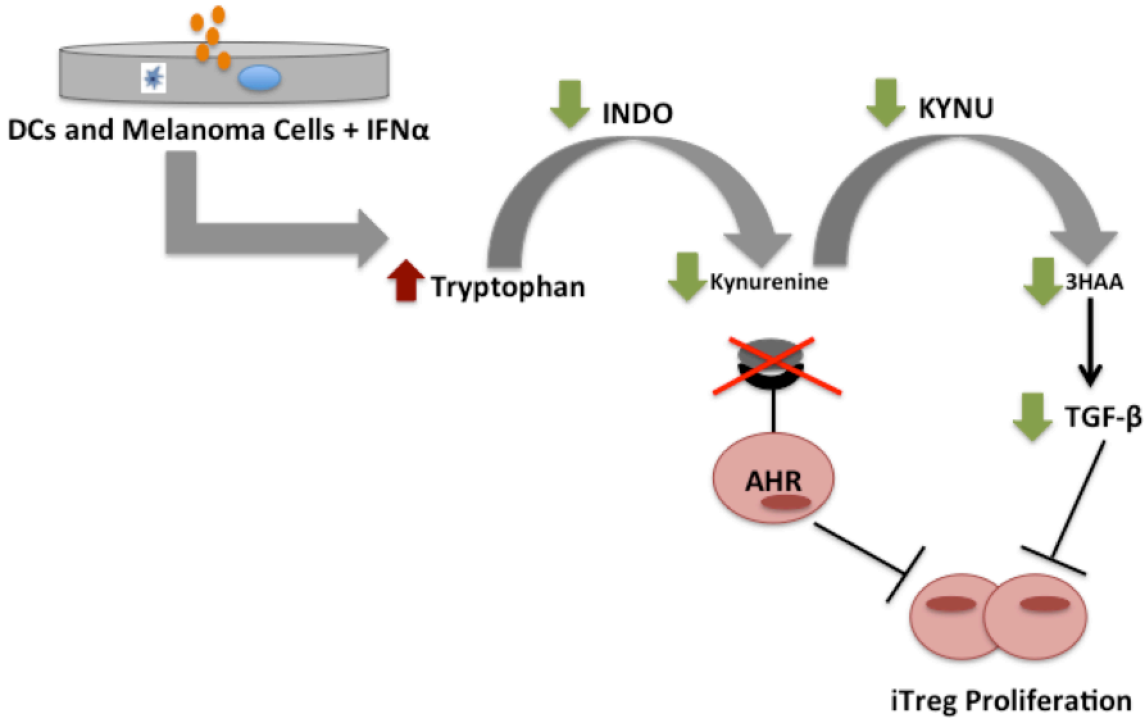

Figure 11.1: This image displays the possible application of IFNα to generate a successfully antitumor immune response. The diagram also displays a hypothesis that can be tested at the bench using in vitro and in vivo experiments.

12. CONCLUSIONS

Only 6% of samples in GEO are manually curated GEO DataSets and the backlog hinders the ability to find new insights from the vast repository of studies. OBDI serves as a solution to combine data across GEO experiments and increase the power of knowledge discovery through high throughput data. The OBDI pipeline incorporates several bioinformatics methods, specifically related to ontologies, data processing and analysis. Ontologies serve as the framework for storing metadata elements and the in silico analysis elements used to conduct machine learning analysis. The OBDI pipeline uses ontologies to annotate and augment experiments from publicly available repositories (e.g., GEO). Using the ontological framework, samples are successfully organized across different GEO experiments. The ontological framework within OBDI allows researchers working in a particular biomedical domain to store high throughput data elements and organize samples across different GEO experiments. The use of ontologies reduces the barrier to integrate new studies into the meta-analysis. Reasoning over ontologies helps maintain consistency of the OBDI experiment structure as more samples are added to current OBDI experiments. Using the OBDI pipeline, researchers can mobilize data that exists in silos to help generate testable hypothesis at the bench.

OBDI's ontological representation promotes the integration of complex data and prior knowledge. The new annotation from this dissertation extends the current biological knowledge in the cancer immunotherapy domain. Using the ontology, four different in

silico machine learning experiments were conducted that explore the mechanism of cancer immunotherapy at a molecular level. Each experiment focuses on different cell types. The OBDI pipeline supports researchers' ability to create predictive models that may lead to new hypotheses. For example, I used OBDI to manage the results from Experiment 1 and generated a hypothesis around the maturation of DCs and mechanism by which mature DCs induce Tregs. When characterizing the maturation of DCs, the expression of two genes, INDO and KYNU, was assessed across DC maturation treatments. It was discovered that the expression of INDO and KYNU is down regulated in DCs treated with IFNα.

These findings are generalized to Experiments 2 and 3 to further explore how the induction of Tregs may play a role in cancer immunotherapy. As extant findings support, INDO plays a role in breaking down tryptophan into the first metabolite of the kynurenine pathway. Kynurenine interacts with AHR in Tregs, thus causing the proliferation of Tregs and suppressing immune system response. The IFNα treatment from Experiment 1 generalizes to Experiment 3, where the expression of cancer cell lines was treated with IFNα2a for 4 and 24 hours. KYNU plays a role in generating 3-HAA for the release of TGFβ into the environment. It was evident that melanoma cell lines treated with IFNα2a for 24 hours have a lower expression of KYNU. The combination of these results generate a testable hypothesis where a downregulation of INDO and KYNU may lower the production or affect certain tryptophan metabolites; thus, inhibiting the suppressive activity of Tregs in a tumor environment. Experiment 4 integrates sequencing data that expands the OBDI pipeline to include data beyond microarray experiments. OBDI can be used to replicate similar hypothesis driven studies in other biological

domains.

Using the OBDI methodology, researchers can generate models and conduct laboratory experiments. Through publically available data (e.g., GEO), I am able to generate integrated models to help bench researchers understand the immune system at a molecular level. By conducting machine learning analysis, I can compare different models and make new discoveries.

APPENDIX A

ML-FLEX PARAMETERS

DATA_PROCESSORS	Location where the ARFF or CSV files is stored
CLASSIFICATION_ALGORITHMS	Classification algorithms used for the specific analysis
FEATURE_SELECTION_ALGORITHMS	Feature selection algorithms used for the analysis
TEST_INSTANCE_IDS	When preforming testing/training, this parameter allows for specifying I.Ds that are used to testing instances
TRAIN_INSTANCE_IDS	When preforming testing/training, this parameter allows for specifying I.Ds that are used to training instances
NUM_INNER_CROSS_VALIDATION_FOLDS	List the number of "outer" folder while performing cross validation (0=LOOCV, 1=Train/Test, n=specify the number of instances).
NUM_OUTER_CROSS_VALIDATION_FOLDS	List the number of "inner" folder while performing cross validation (0=LOOCV, 1=Train/Test, n=specify the number of instances).
NUM_FEATURES_OPTIONS	Specify the appropriate number of features to use in order to generative the most "informative" model.

APPENDIX B

OBDI PARAMETERS

MAIN_OUTPUT_FOLDER	Location where all output files are stored
GEO_DATA_LINKS	GEO links to raw data
GEO_XML_LINKS	GEO links to metadata
OWL_FILE	Location of ontology file
MASTER_INDEX_FILE	Location of master indexing file created by the first part of OBDI
RAW_FILE_FOLDER	Same as MAIN_OUTPUT_FOLDER
ZIPPED_RAW_FILES	Same as MAIN_OUTPUT_FOLDER
GENE_PATTERN_CLM_FILE	Location of CLM file created by the first part of OBDI (Affymetrix Only)
OPTIONAL_FEATURE_LIST	Location of text file containing features to select
GENE_PATTERN_GCT_FILE	Same as MAIN_OUTPUT_FOLDER
WEKA_ARFF_FILE	Same as MAIN_OUTPUT_FOLDER
MLFLEX_JAR_FILE	Location of ML-Flex file
MLFLEX_EXPERIMENT_FILE	Obtained from OWL file

APPENDIX C

MASTER INDEXING FILE FORMAT

The master-indexing file is a text file where each field is separated by a pipe character. The general format of the master-indexing file is as follows:

```
GSM ID|Title Text related to the GSM ID|ML-Experiment Name|Relevant machine learning condition
```

Based on the master-indexing file, a TreeMapping method is created that helps map each sample to the machine learning condition. Since GSMxxx identifies the samples, the master-indexing file makes it easy to place the GEO samples into the appropriate machine learning condition folder.

APPENDIX D

SIGNING UP FOR GENE PATTERN

To create a Gene Pattern account, follow the steps below:

1. Point your internet browser to the following URL to create a Gene Pattern account: http://genepattern.broadinstitute.org/gp/pages/login.jsf
2. Click on the following link: Click to Register
3. Finish entering the required information and make a note of your username and password.

REFERENCES

1. Sujansky W. **Heterogeneous database integration in biomedicine**. *J Biomed Inform*. 2001; 34(4): 285-98.

2. Goble C, Stevens R. **State of the nation in data integration for bioinformatics**. *J Biomed Inform*. 2008; 41(5): 687-93.

3. Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P. **Data integration and genomic medicine**. *Journal of Biomedical Informatics*. 2007; 40(1):5-16.

4. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. **Modeling sample variables with an Experimental Factor Ontology**. *Bioinformatics*. 2010; 26(8):1112-8.

5. Kodama, Y., M. Shumway, and R. Leinonen, **The sequence read archive: explosive growth of sequencing data**. *Nucleic Acids Res*. 2012; 40(D1):D54-6.

6. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, et al. **ArrayExpress--a public database of microarray experiments and gene expression profiles**. *Nucleic Acids Res*. 2007; 35(Database issue):747-50.

7. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. **NCBI GEO: Archive for functional genomics data sets--10 years on**. *Nucleic Acids Res*. 2011; 39(Database issue):1005-10.

8. Anderle P, Duval M, Draghici S, Kuklin A, Littlejohn TG, Medrano JF, et al. **Gene expression databases and data mining**. *Biotechniques*. 2003; Review(34): 36-44.

9. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P. **Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms**. *Nucleic Acids Res.* 2005; 33(18): 5914-23.

10. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. **Gene Pattern**. *Nature Genetics*. 2006; 38(5): 500-501.

11. Witten, I.H. and E. Frank, **Data mining: practical machine learning tools and techniques**. 3rd ed: Morgan Kaufmann; 2011. ISBN-13: 978-0123748560.

12. Quinlan, J.R., **C4.5: programs for machine learning**. Morgan Kaufmann Publishers Inc. 1993; 302. ISBN:1-55860-238-0.

13. R Development Core Team, **R: a language and environment for statistical commuting**. R Foundation for Statistical Computing. Vienna, Austria. 2010. ISBN 3-900051-07-0, URL http://www.R-project.org/.

14. Demšar, J., T. Curk, and A. Erjavec, **Orange: data mining toolbox in Python**. *Journal of Machine Learning Research*. 2013; 14: 2349−2353.

15. Stephen, P. and F. Lewis, *ML-Flex:* **A flexible toolbox for performing classification analyses in parallel**. *Journal of Machine Learning Research*. 2012; 13:555-559.

16. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, Mewes HW. **Gene selection from microarray data for cancer classification--a machine learning approach**. *Comput Biol Chem*. 2005; 29(1):37-46.

17. Langley, P., W. Iba, and K. Thompson. An analysis of bayesian classifiers. Proceedings of the Tenth National Conference on Artificial Intelligence. 1992; San Jose, CA.

18. Fan L, Poh KL, Zhou P. **A sequential feature extraction approach for naïve Bayes classification of microarray data**. *Expert Systems with Applications*. 2009; 36(6):9919-9923.

19. Pirooznia M, Yang JY, Yang MQ, Deng Y. **A comparative study of different machine learning methods on microarray gene expression data**. *BMC Genomics*. 2008; 9(Suppl 1):S13.

20. Quinlan, J., **Improved use of continuous attributes in C4.5.** *Journal of Artificial Intelligence Research*. 1996; 4: 77–90.

21. Netto OP, Nozawa SR, Mitrowsky RAR, Macedo AA, Baranauskas JA. Applying decision trees to gene expression data from DNAMicroarrays: a leukemia case study. Workshop de Informática Médica, 2010. ISSN 2175-2761.

22. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D. **Knowledge-based analysis of microarray gene expression data by using support vector machines**. *Proc Natl Acad Sci USA*. 2000; 97(1):262-267.

23. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. **Knowledge-based analysis of microarray gene expression data by using support vector machines**. *Proc Natl Acad Sci USA*. 2000; 97(1): 262-7.

24. Guyon I, Weston J, Barnhill S, Vapnik V. **Gene selection for cancer classification using support vector machines**. *Machine Learning.* 2002; 46(1-3): 389-422.

25. Robnik-Sikonja M, Kononenko I. **Theoretical and empirical analysis of ReliefF and RReliefF**. *Machine Learning.* 2003; 23-69.

26. Li T, Zhang C, Ogihara M. **A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression**. *Bioinformatics*. 2004; 2429-37.

27. Wang Y, Makedon F. Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data. Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference. 2004.

28. Kononenko, I. Estimating attributes: analysis and extensions of RELIEF. Machine Learning: ECML-94. Lecture Notes in Computer Science. 1994; 171-182.

29. Liu H, Li X, Yoon V, Clarke R. Annotating breast cancer microarray samples using ontologies. AMIA Annu Symp Proc. 2008; 6: 414-8.

30. Menzel C. Reference ontologies -- application ontologies: either/or or both/and? Proceedings of the KI2003 Workshop on Reference Ontologies and Application Ontologies. 2003.

31. Golbreich C, Zhang S, Bodenreider O. **The foundational model of anatomy in OWL: experience and perspectives**. *Web Semant.* 2006; 4(3): 181-195.

32. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. **Gene ontology: tool for the unification of biology. The Gene Ontology consortium**. *Nat Genet*. 2000; 25(1): 25-9.

33. Knublauch H, Fergerson RW, Noy NF, Musen MA. The Protégé OWL Plugin: an open development environment for Semantic Web applications. The Semantic Web-ISWC. 2004; 3298: 229-243.

34. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. **The OBO foundry: coordinated evolution of ontologies to support biomedical data integration**. *Nat Biotechnol*. 2007; 25(11): 1251-5.

35. Day-Richter J, Harris MA, Haendel M, Lewis S. **OBO-Edit--an ontology editor for biologists**. *Bioinformatics.* 2007; 23(16): 2198-200.

36. Tirmizi SH, Aitken S, Moreira DA, Mungall C, Sequeda J, Shah NH, et al. **Mapping between the OBO and OWL ontology languages**. *J Biomed Semantics*. 2011; 2 Suppl 1: S3.

37. Cornet R, Teije A, Keizer N. **Comparison of reasoners for large ontologies in the OWL 2 EL Profile**. *Semantic Web Journal*. 2011; 2(2): 71-87.

38. Shearer R, Motik B, Horrocks I. HermiT: a highly-efficient OWL reasoner. Proc. of the 5th Int. Workshop on OWL: Experiences and Directions. 2008; 26-27.

39. Ideker T, Dutkowski J, Hood L. **In complex biology, prior knowledge is power**. *Cell*. 2011; 144(6): 860-3.

40. Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, et al. **Improving breast cancer survival analysis through competition-based multidimensional modeling**. *PLoS Comput Biol*. 2013; 9(5): e1003047.

41. Stevens R, Goble CA, Bechhofer S. **Ontology-based knowledge representation for bioinformatics**. *Brief Bioinform*. 2000; 1(4): 398-414.

42. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. **PANTHER: a library of protein families and subfamilies indexed by function**. *Genome Res*. 2003; 13(9): 2129-41.

43. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Res*. 2010; 38(Database issue):355-60.

44. Huang da W, Sherman BT, Lempicki RA. **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc*. 2009; 4(1): 44-57.

45. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. **GoMiner: a resource for biological interpretation of genomic and proteomic data**. *Genome Biol*. 2003; 4(4): R28.

46. Bard J, Rhee SY, Ashburner M. **An ontology for cell types**. *Genome Biol*. 2005; 6(2): R21.

47. Meehan TF, Masci AM, Abdulla A, Cowell LG, Blake JA, Mungall CJ, et al. **Logical development of the cell ontology**. *BMC Bioinformatics*. 2011; 12(6).

48. Blattman JN, Greenberg PD. **Cancer immunotherapy: a treatment for the masses**. *Science*. 2004; 305(5681): 200-5.

49. Kirkwood J, Butterfield L, Tarhini A, Zarour H, Kalinski P, Ferrone S. **Immunotherapy of cancer in 2012**. *Cancer Journal for Clinicians*. 2012; 62(5): 309-335.

50. Lesterhuis WJ, Haanen JB, Punt CJ, **Cancer immunotherapy--revisited**. *Nat Rev Drug Discov*. 2011; 10(8): 591-600.

51. Palucka K, Banchereau J. **Dendritic-cell-based therapeutic cancer vaccines**. *Immunity*. 2013; 39(1): 38-48.

52. Ebstein F, Lange N, Urban S, Seifert U, Kruger E, Kloetzel PM. **Maturation of human dendritic cells is accompanied by functional remodelling of the ubiquitin-proteasome system**. *Int J Biochem Cell Biol*. 2009; 41(5): 1205-15.

53. Liau LM, Prins RM, Kiertscher SM, Odesa SK, Kremen TJ, Giovannone AJ, et al. **Dendritic cell vaccination in glioblastoma patients induces systemic and intracranial T-cell responses modulated by the local central nervous system tumor microenvironment**. *Clin Cancer Res*. 2005; 11(15): 5515-25.

54. Abbas A, Lichtman A, Pillai S. **Cellular and molecular immunology**. 2007; 6th Ed: Saunders. 576.

55. Shortman K, Liu YJ. **Mouse and human dendritic cell subtypes**. *Nat Rev Immunol*. 2002; 2(3): 151-61.

56. Cools N, Ponsaerts P, Van Tendeloo VF, Berneman ZN. **Regulatory T cells and human disease**. *Clin Dev Immunol*, 2007; (2007).

57. Kushwah R, Hu J. **Role of dendritic cells in the induction of regulatory T cells**. *Cell Biosci*. 2011; 1(1): 20.

58. Harmelen F, Antoniou G, McGuinness D, **Web ontology language: OWL**. *Handbook on ontologies in information systems*. 2003; 67-92.

59. Horridge M, Bechhofer S. **The OWL API: a java API for working with OWL 2 ontologies**. *Semantic Web Journal*. 2011; 2(1): 11-21.

60. Berglund A, Boag S, Chamberlin D, Fernandez M, Kay M, Robie J, et al. XML Path Language (XPath). W3C. 2010; 2nd Ed.

61. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, Weston AD, et al. ***A data integration methodology for systems biology***. *Proc Natl Acad Sci USA*. 2005; 102(48): 17296-17301.

62. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics*. 2009; 25(16): 2078-2079.

63. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics*. 2003; 4(2): 249-264

64. Dietterich, TG. **Ensemble methods in machine learning**. *Lect Notes Comput Sc*i. 2000; 1857: p.1–15.

65. Wan H, Dupasquier M. **Dendritic cells in vivo and in vitro**. *Cell Mol Immunol.* 2005; 2(1): 28-35.

66. Masci AM, Arighi CN, Diehl AD, Lieberman AE, Mungall C, Scheuermann RH, et al. **An improved ontological representation of dendritic cells as a paradigm for all cell types**. *BMC Bioinformatics*. 2009; 10(70).

67. Fulcher JA, Hashimi ST, Levroney EL, Pang M, Gurney KB, Baum LG, et al. **Galectin-1-matured human monocyte-derived dendritic cells have enhanced migration through extracellular matrix**. *J Immunol.* 2006; 177(1): 216-26.

68. Dohnal AM, Luger R, Paul P, Fuchs D, Felzmann T. **CD40 ligation restores type 1 polarizing capacity in TLR4-activated dendritic cells that have ceased interleukin-12 expression**. *J Cell Mol Med.* 2009; 13(8b): 1741-50.

69. Dhodapkar KM, Banerjee D, Connolly J, Kukreja A, Matayeva E, Veri MC, et al. **Selective blockade of the inhibitory Fcgamma receptor (FcgammaRIIB) in human dendritic cells and monocytes induces a type I interferon response program**. *J Exp Med*. 2007; 204(6): 1359-69.

70. Pirooznia M, Yang JY, Yang MQ, Deng Y, et al. **A comparative study of different machine learning methods on microarray gene expression data**. *BMC Genomics*. 2008; 9 Suppl 1: S13.

71. Lob S, Konigsrainer A. **Is IDO a key enzyme bridging the gap between tumor escape and tolerance induction?** *Langenbecks Arch Surg*. 2008; 393(6): 995-1003.

72. Mellor AL, Baban B, Chandler PR, Manlapat A, Kahler DJ, Munn DH. *Cutting edge:* **CpG oligonucleotides induce splenic CD19+ dendritic cells to acquire potent indoleamine 2,3-dioxygenase-dependent T cell regulatory functions via IFN Type 1 signaling**. *J Immunol.* 2005; 175(9): 5601-5.

73. Sharma MD, Hou DY, Liu Y, Koni PA, Metz R, Chandler P, et al. **Indoleamine 2,3-dioxygenase controls conversion of Foxp3+ Tregs to TH17-like cells in tumor-draining lymph nodes**. *Blood*. 2009; 113(24): 6102-11.

74. Hua D, Sun J, Mao Y, Chen LJ, Wu YY, Zhang XG. **B7-H1 expression is associated with expansion of regulatory T cells in colorectal carcinoma**. *World J Gastroenterol.* 2012; 18(9): 971-8.

75. Wilcox RA, Feldman AL, Wada DA, Yang ZZ, Comfere NI, Dong H, et al. **B7-H1 (PD-L1, CD274) suppresses host immunity in T-cell lymphoproliferative disorders**. *Blood*. 2009; 114(10): 2149-58.

76. Bollyky PL, Falk BA, Long SA, Preisinger A, Braun KR, Wu RP, et al. **CD44 costimulation promotes FoxP3+ regulatory T cell persistence and function via production of IL-2, IL-10, and TGF-beta**. *J Immunol.* 2009; 183(4): 2232-41.

77. Chikamatsu K, Takahashi G, Sakakura K, Ferrone S, Masuyama K. **Immunoregulatory properties of CD44+ cancer stem-like cells in squamous cell carcinoma of the head and neck**. *Head Neck*. 2011; 33(2): 208-15.

78. Prots I, Skapenko A, Lipsky PE, Schulze-Koops H. **Analysis of the transcriptional program of developing induced regulatory T cells**. *PLoS One*. 2011; 6(2): e16913.

79. Wildin R, Smyk-Pearson S, Filipovich A. **Clinical and molecular features of the immunodysregulation, polyendocrinopathy, enteropathy, X linked (IPEX) syndrome**. *J Med Genet.* 2002; 39(8): 537-45.

80. Probst-Kepper M, Geffers R, Kroger A, Viegas N, Erck C, Hecht HJ, et al. **GARP: a key receptor controlling FOXP3 in human regulatory T cells**. *J Cell Mol Med*. 2009; 13(9b): 3343-57.

81. Lund R, Aittokallio T, Nevalainen O, Lahesmaa R. **Identification of novel genes regulated by IL-12, IL-4, or TGF-beta during the early polarization of CD4+ lymphocytes**. *J Immunol*. 2003; 171(10): 5328-36.

82. Piccirillo CA, Shevach EM. **Naturally-occurring CD4+CD25+ immunoregulatory T cells: central players in the arena of peripheral tolerance**. *Semin Immunol*. 2004; 16(2): 81-8.

83. Cantorna MT. **Why do T cells express the vitamin D receptor?** *Ann N Y Acad Sci.* 2011; 1217: 77-82.

84. Wang Y, Zhu J, DeLuca HF. **Where is the vitamin D receptor?** *Arch Biochem Biophys.* 2012; 523(1): 123-33.

85. Kang SW, Kim SH, Lee N, Lee WW, Hwang KA, Shin MS, et al. **1,25-Dihyroxyvitamin D3 promotes FOXP3 expression via binding to vitamin D response elements in its conserved noncoding sequence region**. *J Immunol*. 2012; 188(11): 5276-82.

86. Gandhi R, Kumar D, Burns EJ, Nadeau M, Dake B, Laroni A, et al. **Activation of the aryl hydrocarbon receptor induces human type 1 regulatory T cell-like and Foxp3(+) regulatory T cells**. *Nat Immunol.* 2010; 11(9): 846-53.

87. Pot, C., **Aryl hydrocarbon receptor controls regulatory CD4+ T cell function**. *Swiss Med Wkly*. 2012; 142: w13592.

88. Adams S, Braidy N, Bessede A, Brew BJ, Grant R, Teo C, et al. ***T*he kynurenine pathway in brain tumor pathogenesis**. *Cancer Res.* 2012; 72(22): 5649-57.

89. Moschella F, Proietti E, Capone I, Belardelli F. **Combination strategies for enhancing the efficacy of immunotherapy in cancer patients**. *Ann N Y Acad Sci*. 2010; 1194: 169-78.

90. Gogas H, Ioannovich J, Dafni U, Stavropoulou-Giokas C, Frangia K, Tsoutsos D, et al. **Prognostic significance of autoimmunity during treatment of melanoma with interferon**. *Semin Immunopathol*. 2011; 33(4): 385-91.

91. Di Pucchio T, Pilla L, Capone I, Ferrantini M, Montefiore E, Urbani F, et al. **Immunization of stage IV melanoma patients with Melan-A/MART-1 and gp100 peptides plus IFN-alpha results in the activation of specific CD8(+) T cells and monocyte/dendritic cell precursors**. *Cancer Res*. 2006; 66(9): 4943-51.

92. Gogas H, Ioannovich J, Dafni U, Stavropoulou-Giokas C, Frangia K, Tsoutsos D, et al. **Prognostic significance of autoimmunity during treatment of melanoma with interferon**. *N Engl J Med*. 2006; 354(7): 709-18.

93. Tarhini AA, Gogas H, Kirkwood JM. **IFN-alpha in the treatment of melanoma**. *J Immunol*. 2012; 189(8): 3789-93.

94. Siegrist F, Ebeling M, Certa U. **The small interferon-induced transmembrane genes and proteins**. *J Interferon Cytokine Res*. 2011; 31(1): 183-97.

95. Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, et al. **Integrative analysis of the melanoma transcriptome**. *Genome Res*. 2010; 20(4): 413-27.

96. Weber WP, Feder-Mengus C, Chiarugi A, Rosenthal R, Reschner A, Schumacher R, et al. **Differential effects of the tryptophan metabolite 3-hydroxyanthranilic acid on the proliferation of human CD8+ T cells induced by TCR triggering or homeostatic cytokines**. *Eur J Immunol*. 2006; 36(2): 296-304.

97. Chen Y, Guillemin GJ. **Kynurenine pathway metabolites in humans: disease and healthy states**. *Int J Tryptophan Res*. 2009; 2: 1-19.

98. Stratton MR, Campbell PJ, Futreal PA. **The cancer genome**. *Nature*. 2009; 458(7239): 719-24.

99. Stuart D, Sellers WR. **Linking somatic genetic alterations in cancer to therapeutics**. *Curr Opin Cell Biol*. 2009; 21(2): 304-10.

100. Oshlack A, Robinson MD, Young MD. **From RNA-Seq reads to differential expression results**. *Genome Biol*. 2010; 11(12): 220.

101. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome**. *Science*. 2008; 321(5891): 956-60.

102. Wang Z, Gerstein M, Snyder M. **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet*. 2009; 10(1): 57-63.

103. Li H, Durbin R. **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics*. 2009; 25(14): 1754-60.

104. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. **Initial sequencing and analysis of the human genome**. *Nature*. 2001; 409(6822): 860-921.

105. Giricz O, Lauer JL, Fields GB. **Variability in melanoma metalloproteinase expression profiling**. *J Biomol Tech*. 2010; 21(4): 194-204.

106. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol.* 2010; 28(5): 511-5.

107. Durbin R, Haussler D, Stein L, Lewis S, Krogh A. GFF (General Feature Format) Specifications Document. Welcome Trust Sanger Institute. 2012; 2nd Ed.