A Translational Bioinformatics Approach to Parsing and Mapping ISCN Karyotypes: A Computational Cytogenetic Analysis of Chronic Lymphocytic Leukemia (CLL)

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Zachary Benjamin Abrams, B.S.

* * * * *

Biomedical Sciences Graduate Program

The Ohio State University

2016

Dissertation Committee:

Philip R.O. Payne PhD, Advisor

Kevin Coombes PhD

Amy Johnson PhD

Albert M. Lai PhD

Kun Huang PhD
ABSTRACT

Translational Bioinformatics is the field of study pertaining to the interpretation, analysis, and storage of large volumes of biomedical data for the purpose of improving human health. This thesis takes a translational bioinformatics approach through the large-scale analysis of karyotype data. Karyotyping, the practice of visually examining and recording chromosomal abnormalities, is commonly used to diagnose and treat disease. Karyotypes are written in a special language known as the International System for Human Cytogenetic Nomenclature (ISCN). Analyzing these karyotypes is currently done in a manual, non-computational manner due to the structure of the ISCN. The ISCN is generally considered not computationally tractable and as such precludes the potential of these genomic data from being fully realized. In response, this thesis presents the development of a cytogenetic platform (the Loss-Gain-Fusion model) that allows the transformation of human-readable ISCN karyotypes into a machine-readable model for computational analysis. This platform then utilizes text based cytogenetic data to create a structured binary karyotype language.

Based on this computer readable language, several analyses are performed to demonstrate the potential of these data. First, the LGF model was
applied to the Mitelman database (a publically-available karyotype database) to distinguish different diseases; in the process, we discerned which algorithms performed the best on the LGF data format. Second, an analysis was conducted to find potentially missed cytogenetic aberrations that recur in chronic lymphocytic leukemia from clinical data at the Ohio State university. Third, triplets containing drug, gene, and disease information were generated via a computational pipeline that connected various public drug-gene interaction data sources to identify potential drug repurposing hypotheses. The research presented here has detailed a novel approach to analyzing cytogenetic data.
DEDICATION

For Sarah E. Wyatt with many thanks.
ACKNOWLEDGMENTS

I would like to thank the CLL Research Consortium at the University of California, San Diego, for financial support (NCI - 2P01CA81534-12A1), as well as the National Library of Medicine for their fellowship support (T15LM011270) while conducting this research. I am greatly indebted to my dissertation committee members (Philip Payne [chair], Kevin Coombes, Kun Huang, Amy Johnson, and Albert Lai) for their invaluable assistance.

Great respect and thanks also goes to the patients who took part in the OSU 0025 Clinical Trial. Additionally, I would like to thank the research lab of John Byrd for their help and support; Lynne Abruzzo and Nyla A. Heerema for their work in validating the LGF model; Arkaprava Pattanayak, William (D'Arcy) Kenworthy, Natarajan Muthusamy, and, John Byrd, Philip Payne and Kevin Coombes for their help in the theoretical development of this work. I would also like to extend my sincerest gratitude to several fellow trainees: Andrea L. Peabody and Satyajeet Raje provided insightful research input; Kelly Regan, Tasneem Motiwala, Deborah Lin, and Bobbie Kite also provided important support as members of Philip Payne’s research lab at OSU. I would also like to thank Marjorie Kelley for her research assistance, as well as James Gentry and Judith Buster for their invaluable assistance during my doctoral studies.
Major aspects of this dissertation were published in 2015 with co-authors Andrea Peabody, Nyla Heerema, and Philip Payne under the title, “Text Mining and Data Modeling of Karyotypes to aid in Drug Repurposing Efforts”, in *Studies in Health Technology and Informatics* 216:1037. Additionally, several aspects of this thesis were presented at conferences as papers: 1) “The Use of Cytogenetic Data to Enable Drug Repurposing Studies” with co-author Philip Payne, at the 2015 AMIA Joint Summits TBI CRI conference in San Francisco; 2) “Using Karyotype Data for Automated Disease Classification: A Proof of Concept Demonstration” with co-authors Satyajeet Raje, Kevin Coombes, and Philip Payne at the 2016 TBI CRI conference in San Francisco; and 3) “Discovering novel genetic aberrations for leukemia using karyotype data” with co-authors Satyajeet Raje, Kevin Coombes, and Philip Payne also at the 2016 TBI CRI conference.

On a personal level, I would like to thank my lifelong friends, Daniel Shusta, Victoria Morgan, Aisha Stern, and Beth Riordan for their continuous support and encouragement. Finally, a special thanks goes to my parents for their unwavering support.
VITA

June 2007…………………………..Athens High School

June 2012………………………………B.S. Microbiology, Ohio University

September 2012 - Present…………Ph.D. Biomedical Sciences – Computational Biology & Bioinformatics,

January 2014 – Present………………NLM Graduate Fellow, Department of Biomedical Informatics, The Ohio State University (T15LM011270)

PUBLICATIONS


FIELDS OF STUDY

Major Field: Biomedical Sciences

Secondary Field: Translational Bioinformatics
TABLE OF CONTENTS

Abstract .................................................................................................................. ii
Dedication ............................................................................................................... iv
Acknowledgements ............................................................................................... v
Vita ........................................................................................................................... vii
List of Tables .......................................................................................................... ix
List of Figures ......................................................................................................... x
Chapter 1: Research Background and Thesis Goals ............................................. 1
Chapter 2: LGF Parser and Mapper ...................................................................... 16
Chapter 3: Disease Classification Using the LGF Model .................................... 33
Chapter 4: Clinical Cytogenetics and the LGF Model ........................................ 43
Chapter 5: Drug Repurposing ............................................................................... 57
Chapter 6: Conclusions .......................................................................................... 68
References ............................................................................................................. 76
LIST OF TABLES

Table 2.1. Cytogenetic events and equivalent mappings in the LGF model.................................................................28

Table 2.2. Parsing in mapping errors classified by type of error.........................30

Table 3.1. Diseases with over 1000 karyotype samples in Mitelman..............35

Table 3.2. Weighted average F-Measure for Classifiers at Different Sample Sizes..........................................................36

Table 3.3. Classification results as a confusion matrix..................................38

Table 4.1. Cytogenetic events in the OSU and Mitelman cohort by affected population..........................................................................................................................48
LIST OF FIGURES

Figure 1.1 G Banding pattern of human male chromosomes........................8

Figure 2.1 Steps involved in the conversion of raw karyotype text into the LGF model.................................................................19

Figure 2.2 Example of ISCN text input leading to the parse tree after parsing.................................................................22

Figure 2.3. The ICSN syntax is broken down into a parse tree ....................25

Figure 3.1. Experimental design for automated design classification using karyotype data.................................................................34

Figure 3.2. Classification accuracy of different classifiers for different sample sizes.....................................................................................37

Figure 5.1. Workflow of publically available data sources.................................61

Figure 5.2 Gene-Drug-Disease Triplets..........................................................63

Figure 5.3: Subset of drug disease gene triplet table results.........................64
Chapter 1: Research Background and Thesis Goals

1.1 Introduction

Translational Bioinformatics (TBI) is the scientific field of study pertaining to the interpretation, analysis, and storage of large volumes of biomedical data for the purpose of improving human health. One of the core problems in TBI research is that most biomedical data are “siloded;” in other words, isolated from other potentially useful data sets and data analytics (Payne et al., 2009). Thus a major goal of TBI research involves the explicit act of “de-siloing” data, thus enabling researchers to better analyze and articulate biomedical datasets. In this thesis, a workflow for de-siloing and utilizing karyotype data is demonstrated, thus highlighting the relevance of TBI-oriented research strategies. Such strategies are important in furthering our ability to translate informatics projects into clinically actionable research findings. In this thesis I apply such TBI approaches to chromosome level genomic data in the form of karyotype analysis.

In 1970, Dr. Janet Rowley applied the technique of chromosome banding to a morphologically distinct group of patient cases that we now call chronic myeloid leukemia (CML; Rowley, 1973). She discovered that this disease was characterized by a balanced translocation between the long arm of chromosome 9 at band q34 and the long arm of chromosome 22 at band q11.2 (i.e.,
t(9;22)(q34;q11.2) using ISCN nomenclature). This translocation fused the breakpoint cluster region (BCR) gene on chromosome 22 with the Abelson murine leukemia viral oncogene homolog 1 (ABL1) gene on chromosome 9, resulting in a chimeric fusion protein with aberrant tyrosine kinase activity. The identification of this translocation by karyotypic analysis and its subsequent molecular genetic characterization led to the development one of the first targeted therapy for a human cancer, the tyrosine kinase inhibitor (TKI) imatinib mesylate, and revolutionized cancer treatment (Novartis Pharma, 2003).

It is important to contextualize Rowley’s (1973) finding within the technology historically available at the time. During her work in the early 1970s, personal computers, the use of the Internet, packaged statistical software programs, high throughput computing, and medical database management, were all well in the future. Her historic finding was based on the analysis of specific cases all presenting the same symptoms, and her demonstration that cancer had a genetic component was landmark. Since then, cytogeneticists have identified the basic karyotypic abnormalities that characterize a wide variety of hematologic malignancies (leukemias and lymphomas) as well as a number of solid tumors (e.g., Ewing sarcoma); simultaneously the secondary use of medical data for research purposes has exploded as computer technologies advanced.

The secondary use of medical data today is vitally important for the advancement of medical research (Safran et al., 2007). Secondary data use involves the selection and analysis of data for a research purpose other than the
reason for which those data were originally collected. Medical data from patients are routinely collected to help improve patient quality-of-life and treatment; these data are not typically collected with research in mind (deLusignan and vanWeel, 2005). However, many of these pieces of medical data, such as blood pressure, diagnosis code, and karyotype, contain information that could lead to advancements in medical science if framed within the broader context of medical research. Arguably, the majority of medical data collected by physicians and other medical personnel were not collected within the framework of an explicit research design (Rosenbloom et al., 2011). The secondary use of these data thus has enormous potential in medical research and the key to unlocking this potential is to create an analytic model that understands the complexity of secondary data usage. The following section reviews the various methods of secondary use in medical research and specifically considers the secondary use of karyotype data.

1.2 Using Karyotype Data for Secondary Medical Research

Secondary use of medical data presents numerous challenges that must be overcome in order to perform accurate biomedical research (deLusignan and vanWeel, 2005). The first challenge is that the form in which the data are recorded is often incompatible with the computational data format required to conduct research (Rosenbloom et al., 2011). Therefore, the information first has to be transformed into a different format while still preserving the medical
knowledge inherent within the data. An example of this would be karyotype data that are kept in a text string format called the International System for Human Cytogenetic Nomenclature (ISCN; Shaffer et al., 2013). These text strings are not machine-readable. For knowledge to be extracted for research purposes, this text must first be parsed into a machine-readable format (Hiller et al., 2005). This transformation is nontrivial and, depending on the data type, may prevent some forms of information from being used secondarily.

Another major problem in the secondary use of data is the lack of centralization of data as it relates to individual patients (Weiner, 2011). That is, the data needed by a researcher may be dispersed in various places. This is a common problem when observing electronic health records (EHR) data (Weiner, 2011). Since a patient’s EHR is not a complete record due to the fragmentation of the file, a patient's past medical history is often not complete when mining the EHR. This commonly occurs when a patient is going from one clinic to another, with these clinics not being within the same network. The consequence of this disarticulated information means that time points for diagnosis codes and other information that is temporally sensitive are not obtainable unless those different institutions are involved in a data sharing agreement, and this is highly unlikely.

Third, there is a large inconsistency in how different institutions and practitioners record medical codes (Weiner, 2011; Botsis et al., 2010). ICD-9 diagnosis codes, for example, might have differences in use at different institutions or by different individuals. This creates bias and variability when trying
to use these codes in research. To analytically use such medical codes with any integrity, it is important to attempt to minimize these recording biases, such as looking at code usage within a single system/institution.

Finally, medical record data are often incomplete (Botsis et al., 2010; Jollis et al., 1993). Rarely are all possible data points filled out for an individual patient since that information is costly both in terms of time to fill out forms and actual expense to conduct tests required to discern that specific information. Also, clinical workflows are very extensive and thus time restrictive; consequently nurses and doctors are going to collect only the information that is immediately relevant to help the individual at that moment in time. This makes perfect sense since their job is to help the individual and not to perform research, but this does provide a broad and recurrent problem for later researchers who attempt to use this information in a secondary manner.

In conclusion, the secondary use of medical information is the backbone of clinical informatics research but data modeling for secondary use needs to be performed on an individual problem basis to create a best-fit solution that addresses and answers each medical problem. When considering how to model data for secondary use, it is critical to unpack and de-silo the data through the construction of a research design that is targeting specifically for the data that will be used.

Some forms of medical data are available to researchers for secondary research purposes. Karyotype data represent biological information that is
amenable to secondary research use. Although there are complications, when thoughtfully addressed, karyotype data can be well utilized secondarily for medical research. Any remodeling however must maintain the biological integrity of the data. Prior machine-learning approaches using secondary karyotype data have been attempted but achieved only partial and minimal success (Kamada et al. 1983; Hashimoto et al. 1989), reflected by the fact that the results of these efforts were effectively useless in diagnosing diseases and treating patients. In part, the failure of these prior efforts gets at the core of the significance of a TBI framework; that is, they failed to properly unpackage and de-silo the karyotype data, thus failing to maintain the integrity of the karyotype data in its fullest sense. This reflects the importance of a TBI-oriented research framework (Payne et al., 2009).

1.3 Background to Karyotype Analysis

An individual’s genotype influences their clinical phenotype and disease incidence, severity, and progression. One of the premises of precision medicine, a key element of translational bioinformatics wherein the individual is viewed as the primary unit of analysis and observation, is the use of genomic information to optimize an individual’s likelihood of achieving and maintaining health through prevention and/or targeted therapy. Karyotyping, the practice of visually examining and recording chromosomal abnormalities, is one of the earliest and most common genotyping techniques. The utility of karyotyping for diagnosis and
prognosis has been demonstrated in multiple disease states, and such information is regularly and widely used for clinical decision-making (Rowley, 1973; Swerdlow et al., 2008). For example, conventional karyotyping is part of the standard-of-care for hematologic malignancies (Swerdlow et al., 2008). As a result, large karyotypic databases are widely available and well populated.

One publicly available data set of this type, the Mitelman database (Mitelman et al., 2015), has been curated from the literature for over forty years. Mitelman contains >65,000 karyotypes, nearly 70% of which are of hematologic malignancies. However, current clinical use of karyotypic data is limited to only those patterns that are visually apparent to cytogeneticists. Cytogeneticists record karyotype information in a standardized notation, the International System for Human Cytogenetic Nomenclature (ISCN; Shaffer et al., 2013). ISCN is meant to be readable by humans, but is often difficult to interpret systematically due to the volume, complexity, and variability of the information contained therein. As a result, potentially clinically relevant patterns hidden within the long and complex karyotypes remain unused.

Conventional karyotypic analysis is performed on dividing cells in metaphase, the stage in the cell cycle after the chromosomes have duplicated and are maximally condensed. Chromosomes are routinely treated with trypsin and stained with Giemsa (i.e., G-banding) that produces a characteristic pattern of light and dark bands along each chromosome shown in Figure 1.1 (Singh, R. J. 1993), (Bates, 2011). On G-banding, 22 pairs of autosomes and 1 pair of sex
chromosomes can be differentiated by size, position of the centromere (which divides each chromosome into a short (p) and a long (q) arm), and unique banding pattern (which subdivides each arm into regions, bands, and sub-bands). In general, each chromosome contains between 2000 and 4000 genes, and each band contains between 125 and 200 genes. G-banding allows the detection of numerical (gains and losses of whole chromosomes) and structural (partial chromosomal gains and deletions, translocations, inversions) changes at a resolution of 5 megabases (www.coriell.org).

![G Banding patterns of human male chromosomes](image)

**Fig. 1.1 G Banding patterns of human male chromosomes**

Karyotypes of malignant cells often contain three or more cytogenetic abnormalities. This is especially true for karyotypes obtained using modern methods to stimulate mitosis (Heerema *et al.*, 2010) (chronic lymphocytic
leukemia [CLL], plasma cell myeloma, and other lymphoid malignancies). Thus, while conventional karyotypic analysis has allowed us to identify many basic karyotypic abnormalities, a full comparative analysis of cytogenetic abnormalities defies human visual inspection.

Cytogenetic data have been used clinically for decades; many institutions have databases containing tens of thousands of karyotypes. For example, the National Cancer Institute (NCI) hosts the aforementioned Mitelman database (Mitelman et al., 2015), containing more than 65,000 karyotypes collected since 1971. However, little of the information embedded in karyotypic databases has been exploited clinically since karyotypes are stored as text using a standard human interpretable notation, the International System for Human Cytogenetic Nomenclature (ISCN; Shaffer et al., 2013). Several versions of the ISCN standard have been introduced over the decades. Following the release of the earliest version of ISCN in 1971, the standard was revised in 1981, 1985, 1991, 1995, 2005, 2009, and 2013. Many of these revisions contained significant changes from the previous standards, thereby compounding the difficulty of translation across and between cytogenetic data sets.

The first major attempts at transforming karyotype data for secondary use were by Kamada et al. (1983) and Hashimoto et al. (1989). In both cases a programmatic parser was created to break down ISCN karyotypes to discern whether an entire chromosome had been lost or gained. However neither group could look within an aberration to discern information relating to the bands; they
were only able to discern whole chromosomal gains or losses. This is a major limitation since only a fraction of the total amount of knowledge within the karyotype could be utilized and the information that was being utilized was the easiest to discern from human reading of the karyotype.

The KaryoReader (Liang, 2004), the next major attempt to secondarily use karyotypes, sought the breakdown of genetic aberrations at the band level to discern gains and losses. It was methodologically based on machine-learning algorithms that were incapable of reliably identifying complex aberrations, thus resulting in misclassification.

CyDAS (Liang, 2004; Hiller et al., 2005) is the most current system that can be used to parse karyotype data for secondary use. Unlike KaryoReader, this system is able to break down more complex aberrations to the band level. It also is able to accurately represent whole chromosomal losses and gains. The research team behind CyDAS also produced a very useful web interface that allows for the use and testing of their tool by the public, an important feature in the free distribution of information and knowledge within science. The limitation of the system comes from the fact that it is based on machine-learning algorithms similar to those in KaryoReader but containing a different flaw; i.e., there are misclassifications of unusual aberration that leads to misinformation in the output. This is testable using their web service. Part of this flaw is that there is no control for the ISCN version that regularly changes. Consequently, CyDAS only observes 440 cytogenetic features compared to the 910 currently identified ISCN
regions. This is an artifact of the lack of version control that artificially reduces the number of observable features.

Another reason for the misclassifications lies in the fact that the order of cytogenetic aberrations does not matter when writing the karyotype but is taken into account by the machine-learning system utilized in the CyDAS workflow. This, combined with the lack of updates since 2004, means that the system currently is not desirable for researchers who wish to utilize karyotype data for secondary use.

Thus, these prior research scholars identified an important topic for secondary analysis and explored the use of karyotype data. However, all their methods were too limited in scope for contemporary research. The overarching aim of this thesis is to address this problem and advance, in an actionable manner, the analysis and secondary use of karyotype data.

1.4 Thesis Goals

In response to these challenges and opportunities relative to the re-use of cytogenetic data, this thesis is organized into four specific research and development goals. The primary goal of this thesis is to develop a computational cytogenetic model that can translate the text-based ISCN karyotypes into a machine-readable language that can subsequently be analyzed. This will require parsing and mapping the ISCN karyotypes into a biologically based model that then can be validated. This goal is accomplished in chapter 2, with the
development of the LGF (Loss, Gain, Fusion) model that converts ISCN-compliant karyotypes (compliant with all ISCN versions) into a machine-readable format comprised of binary vectors (ordered lists of 0’s and 1’s) that represent the biological phenomena of a loss, gain, or fusion (LGF) at the resolution of cytogenetic bands (Abrams and Payne, 2015).

The second goal of this thesis is to assess whether the LGF representations of karyotype data as a binary vector can be analyzed using machine learning algorithms to distinguish different diseases and to discern which algorithms perform the best on this form of data. This goal is accomplished and presented in chapter 3. Machine learning tests were conducted to distinguish sub-classes of acute myeloid leukemia (AML). Employing the Mitelman database (Mitelman et al., 2015) as the primary source of karyotype data, analyses were conducted to verify the utility of the data and validate the computability of the representation format. These experiments provide a “template” demonstrating the utility of the LGF representation to provide actionable knowledge for clinical and translational applications. The four methods used to classify the CLL and AML samples were (1) Support Vector Machine (SVM), (2) Random Forest (RF), (3) Naïve Bayes (NB) and (4) Multi-Label Logistic Regression (MLR). These algorithms were implemented using the WEKA tool (Hall et al., 2009). Experiments were performed on different sample sizes: 100, 200, 500, 750, and 1000. Each algorithm was run for all the different sample sizes to determine which method yielded the best performance as well as what sample size had the
highest classification efficiency. A ten-fold cross validation scheme was used for evaluation.

The third thesis goal is a clinical application test of the model, presented in chapter 4. It is important to note that in order to meet this specific goal, CLL was employed as a model disease to evaluate if the LGF model representation allows us to find currently known, gold standard, and novel cytogenetic abnormalities (Döhner et al., 1997; Döhner et al., 2000; Garcia-Marco et al., 1996). CLL is the most common leukemia in the Western hemisphere (Swerdlow et al., 2008). In 2013, there were 15,680 new cases of CLL and 4,580 CLL related deaths in the United States. For patients with advanced CLL, conventional doses of chemotherapy are not curative. Although the median survival for CLL patients is 8 to 12 years, individual variation is significant, ranging from several months to a normal life expectancy (Swerdlow et al., 2008). Oncologists currently diagnose and may treat CLL patients primarily based on clinical phenotype information with a few biomarkers, such as the cytogenetic technique of fluorescence in situ hybridization (FISH) and IGVH mutation status, used in prognosis. Data driven analyses of CLL have led to major breakthroughs in staging the disease that have increased our understanding of CLL progression (Binet et al., 1977; Binet et al., 1981; Rai et al., 1975; Rai et al., 1987; Wang et al., 2004). However, it is still critical to continue to increase our understanding of CLL.

At OSU, there has been a long history of research collaboration and innovation in the study of CLL. OSU scholars were the first team in the U.S. to
develop methods to stimulate CLL cells with CpG oligodeoxynucleotide (Heerema et al., 2010; Muthusamy et al., 2011). Emblematic of such activities, the cytogenetic database at OSU contains more than 30,000 karyotypes from patients with CLL, including samples collected over time from at least 1827 patients. These data are extremely useful in terms of understanding disease progression and response to therapy. These data will be employed to verify/validate the analysis tools developed as the primary thesis aim. This will be a first step in terms of demonstrating its promise as a means of supporting precision medicine in the cancer domain and contributing to the optimization of diagnostics and therapeutics that are tailored to an individual’s unique genetic makeup.

The fourth and final goal is a multi-modeling test of the LFG Model via an exploration of drug repurposing, and this is presented in chapter 5. By combining and analyzing publically available data sets (Mitelman, The Drug Gene Interaction Database, National Center for Biotechnology Information, National Cancer Institute, Reactome, and Kyoto Encyclopedia of Genes and Genomes) through the LGF model platform, it is possible to discern gene-drug-disease triplets for cases where the gene was over expressed due to a gain and the drugs were targeted suppressors of the gained gene. A list of 68,543 drug-gene-disease triplets from a total of 69 diseases is investigated. It is anticipated that this information will prove useful to researchers in the domain of pharmaceutical
drug repurposing and in the treatment of these conditions, and provides a further
demonstration of the clinical value and potential of the LGF model.

From a broader view, the specific goals of the thesis are embedded within
the TBI framework of integrating and analyzing previously siloed biomedical
datasets utilizing new computational approaches. Once our biomedical data -- in
this case karyotype data -- have been computationally translated and thus de-
siloed, new computational analyses can be applied in a novel manner. This is
critical both as an example of TBI scientific design that enables such
transformative medical research and show how this research can help improve
patient quality of life and well-being.
Chapter 2: LGF Parser and Mapper

2.1 Introduction

Karyotypes are written descriptions of chromosomes. Karyotyping tests detect chromosomal structural and numerical defects, which can serve as genetic indicators of disorders or diseases. These tests are readily available relative to more specialized sequencing technologies and are thus a more accessible means of assessing patient's genetics for diagnosis and/or therapeutic planning purposes. It has long been observed that the chromosomal aberrations detected in these tests can be used as a diagnostic tool in medicine since many diseases have a pattern of chromosomal aberrations (Swerdlow et al., 2008). Prior attempts to parse and map karyotype data have not been fully successful, as described in Chapter 1. The major gap in these prior attempts was to biologically map cytogenetic events to their biological function.

The International System for Human Cytogenetic Nomenclature (ISCN) (Shaffer et al., 2013) is a domain-specific language that records these chromosomal defects, which have been encoded through visual inspection of chromosomes under a microscope. These data, however, are difficult to analyze using existing computational methods by virtue of their syntactic variability, information density, and potential for human error.
Karyotyping is a common laboratory procedure to identify cytogenetic aberrations (Shuman, 2000). Karyotype data contains valuable clinical information about cytogenetic markers for specific diseases, especially different tumor types in cancer (Chomczynski and Sacchi, 1987). Currently, visual and digital karyotype data are used in the cancer domain (Sambrook et al., 1989). The discovery of the Philadelphia chromosome, a cytogenetic translocation that is strongly associated with chronic myelogenous leukemia (CML; Bradtke et al., 2003), is just one such example that illustrates the potential value of karyotype data in identifying disease progression and informing potential treatment courses, even at very early stages.

In addition, karyotype data have specific advantages over sequencing data (such as targeted, exome, or whole genome sequencing). Karyotype data are commonly available clinical data and in addition, there are publically available large-scale historical data sets with clinical correlates that can be leveraged for analysis. Thus, the karyotype data represent an opportunity with great potential to provide rich clinically important knowledge.

However, only a few significant cytogenetic aberrations that have critical clinical impact have been successfully identified. This is primarily due to the fact that karyotype data are, generally, stored in a text based syntactic format that is not innately computer-readable. Such syntactic ISCN representation of karyotype data, although semi-structured, is too complex to be directly and intuitively processed by humans in a timely manner. Computational tools and
analysis methods that can effectively extract knowledge from the karyotype data cannot be applied in its current syntactic form directly. As a result, a majority of the karyotype data remains uninterpreted and, as a consequence, is underutilized in medical research and clinical practice.

This chapter presents a computational (Cytogenetic Platform) system that transforms ISCN-encoded karyotypes into machine-readable constructs through a process of parsing and mapping into a biological model. This generates an automated pipeline that extracts biologically important information from text-based karyotype data in a form that can be mined. The Loss-Gain-Fusion (LGF) biological model represents the aberrations in the karyotype based on the biological event (loss, gain, or fusion) occurring at that location in the chromosome as opposed to focusing on the cause of that particular biological event (e.g., deletion, translocation). This places sole computational emphasis on the underlying biological result or event itself. This allows researchers to compare across biological events in the analysis of similar functional outcomes in patients.

2.2 Workflow

The Cytogenetic Platform system is a workflow pipeline used to input human-readable cytogenetic data in the form of ISCN karyotypes and output machine-readable representations of those karyotype data. Figure 2.1 shows the workflow of the Cytogenetic parser/mapper system. Karyotype data
were acquired from a patient's tumor by selecting tumor cells that were in mitosis. The researcher then mapped the chromosomes into chromosomal pairs and arranged them in sequential order. Then the patient's aberrations were noted in the ISCN format and an ISCN karyotype was written for the patient based on the number of chromosomes and aberrations in the patient's karyotype (Bates, 2011). These ISCN karyotypes were our input data; as previously mentioned, the largest public repository for karyotype data is the Mitelman database.

Before analysis, two pre-processing steps were performed on the data set. First, karyotypes were removed that were listed as incomplete, denoted by the abbreviation 'inc.' Second, karyotypes with a
'?' were removed since they did not contain sufficient information to generate an accurate assessment of the chromosomal structure of the patient. After this preprocessing, only complete processed karyotypes remained in our data set. No FISH data was employed since they do not represent the entirety of the patient karyotype; rather they represent only whether the patient had a specific aberration.

2.3 Methods

This section describes the LGF model including the computational workflow for transforming ISCN karyotypes as well as the parser and mapper accuracy through experimental evaluation. The core hypothesis tested in this chapter is whether the LGF model accurately represents karyotype data within the Mitelman dataset. As a test of this hypothesis, we measured the accuracy of the LGF model by categorizing error logs reported by the LGF model pipeline.

2.3.1 The Loss – Gain – Fusion (LGF) Model

The design of the LGF model was guided by the three major events known to cause deleterious oncogenic effects at the chromosomal level. The first event is the loss of a gene, which can disrupt the cell cycle or other critical pathways. Deletion is an example of a loss event on the chromosomal level. Deletion of an entire section of chromosomal DNA represents a loss of DNA; thus the genes in that region are not present on that chromosome. The second
event is the gain of an additional copy of the gene. This can cause additional gene products to be created which could disrupt the balance of pathway regulation in a comparable manner to the loss event. As with a loss event, a gain event alters one component of a pathway and thus can disrupt that pathway's function. Duplication represents an example of a gain event, wherein a segment of chromosomal DNA is represented twice. If the fully intact gene is present within the duplicated region, that cell may express more products than will a non-mutated cell. The third event is a fusion event that can produce a chimeric oncogene that disrupts the cell cycle in some capacity. A fusion event requires a breakage event so there could also be a loss of function at fusion points as well as the possibility of a chimeric oncogene. The chimeric gene is formed from multiple genes; this can occur due to a fusion event. It is possible for chimeric genes to alter cell cycle regulation or other pathways that, when disrupted, can lead to a loss of regulation that leads to oncogenic development.

Utilizing the nomenclature found in the ISCN language for writing karyotypes, a set of parser rules was created that facilitated breaking down the text form of the karyotype into a tree structure. By traversing the tree and pulling out the tokenized elements representing the cytogenetic events (e.g. translocation and inversion: and the location, e.g. 17p13), the event and location could be reassembled into pairs (Figure 2.2). This tree is then recursively walked to identify event and location information from the leaves of the tree. These data elements are then mapped to the Loss Gain Fusion (LGF) binary vector.
A mapping language was then developed that assigned each possible cytogenetic event to one of three overall biological events -- loss, gain or fusion -- based on where the event occurred. This provided the ability to map each location and any potential event that could occur at that location. The LGF model represents the karyotype information in a standardized binary vector representation (Figure 2.2). There are 910 cytogenetic locations (at the cytogenetic band granularity) within this model. There can be a loss, gain or
fusion at each location. Thus, the binary vector for each karyotype is represented using 2,730 bits, representing 910 cytogenetic band regions, for each of loss, gain and fusion event at that location. Each bit in the model starts with the alphabet indicating type of event, viz. “L” for loss, “G” for gain and “F” for fusion, followed by the chromosome and the band at which the event occurs. Each observed “event” is indicated by setting that bit to “1” and the rest are “0” by default. For example, a 1 at G7p13 (G + 7 + p13) represents “gain at chromosome 7 at band p13”. The resulting binary vector is of fixed size and can be read computationally in a variety of different applications.

2.3.2 Parsing

The first step in translating the semi-structured language ISCN text into a computer-readable binary vector involves parsing. We began by understanding the nomenclature and grammar of the semi-structured ISCN language for writing karyotypes. A standard set of rules is applied to write karyotypes utilizing a set of commas and semicolons to denote separation of elements within the text. Utilizing this grammar to generate the parser production rules that allowed us to break down the text form of the karyotype into a tree structure.

All code used in the parser and mapper can be found at git@code.bmi.osumc.edu:kbase/CytoBMI.git. To transform ISCN-encoded karyotypes (in syntactic format) into a machine-interpretable format, several parsers were generated using ANTLR (ANother Tool for Language Recognition;
In this paper we use ANTLRv2.0. Specifically, a Context-Free Grammar (CFG), encoded using the Extended Backus-Naur Form (BNF), was used to write ISCN-specific production rules. This approach was based on the observation that any ISCN karyotype was describable in its entirety using a recursive walk of its morphological composition employing a minimal number of CFGs.

An example of how an ISCN aberration is tokenized into the parse tree is shown in Figure 2.3. Each “element” in the karyotype is separated by commas. By default the first element is the total number of chromosomes and the second is the sex chromosomes, in this case “XX” indicating female. Following are the elements representing independent cytogenetic aberration events. Let us look at the example of the first aberration in the example karyotype, the del(5)(q12;q34). We can use the grammar of ISCN to further break this down into sub-elements. In this case there are three, as indicated by the parenthesis and the prefix text. The prefix is set as the ID recognizing the cytogenetic event that the element indicates, “del” meaning a deletion in this case. The remaining parentheses indicate the location of the event. This first parenthesis indicates the chromosome, whereas the second indicates the specific bands where the cytogenetic aberration occurred. A semi-colon “;” separates the start and end of each band. In our example, the deletion event occurs in chromosome 5 from q12 to q34. The set of parser production rules is written to recognize all possible patterns for each aberration in the karyotype data, shown in Table 2.1, starting
with a fundamental rule for a karyotype and incrementally adding further rules for increased complexity.

Thus, the parsing process produces a tokenized computationally readable “parse tree” form of the karyotype (Figure 2.3). The complex karyotype text structure is morphologically deconstructed into constituent phrases as separate elements represented by each main branch of the tree. The parser separates each aberration event into an ID and its location consisting of the chromosome, the arm of that chromosome, and the band of that chromosome. Later, the mapper interprets these events.

**Figure 2.3. The ISCN syntax is broken down into a parse tree**
2.3.4 Mapping

Mapping the parsed karyotypes into our machine-readable model required the creation of a domain-specific language used in the mapper function. This mapping language allowed chromosomal aberrations to be rewritten and then redefined into the LGF model. The ISCN abbreviation for the aberration was written first, followed by the number of elements that would be pulled out from that aberration. The elements of the operation consisted of the chromosome, the arm of that chromosome, and the band of that chromosome. A colon was used to distinguish biological events. Then, starting with element zero, each element was assigned to their respective part of the biological model. The biological elements were separated by a comma ",," and the different symbols represented how that element should be classified. Examples of this language are shown in Table 2.1. This was based on whether that element was lost, gained or involved in a fusion event. Thus, ISCN karyotypes were rewritten to purely represent the biological function at each location. Although antecedent events, such as translocation and deletion, were not preserved in the model, the biological effect of those events was represented allowing for comparison across aberrations with similar biological effects. Each aberration had its own unique classifier in this domain-specific language. Following the use of the domain-specific language to map the karyotypes into the LGF model, we were left with a matrix in which each row represented a karyotype and each column represented a potential cytogenetic band.
The tree generated from the parsing is traversed to extract the individual tokenized elements representing the cytogenetic events, e.g. translocation, deletion, etc., and their respective cytogenetic locations. This allows the chromosomal aberration and the affected band to be isolated and mapped into the LGF model. The mapper reassembles the event and location into pairs and determines which of the binary bits in the LGF model need to be set to 1, as illustrated in Figure 2.3.

To exemplify, let us consider del(5)(q12;q34). The event is a deletion as noted by the letters and “del” at the beginning. We know that this is specifically an interstitial deletion because there is only one chromosome that is affected, namely only two cytogenetic bands follow chromosome 5 thus, indicating that the deletion event occurred between bands 5q12 and 5q34. So we know that precisely those areas of the chromosome are lost. Note that other deletion events may only have one cytogenetic location and some have more than two as identified by the mapper in Table 2.1. To model the above example of del(5)(q12;q34) in the LGF model, we note that cytogenetically for such a deletion the start and end location are sites of “loss” and “fusions” and that all locations in between the two fuse points are sites of “loss”. This informs the binary LGF model to set bits representing F5q12 and F5q34 to 1 for the fusion and loss and at the same time set all loss vectors between these points to 1 indicating the loss at these locations.
<table>
<thead>
<tr>
<th>Event</th>
<th>LGF Mapping</th>
<th>ISCN Example</th>
<th>LGF model translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additional material of unknown origin</td>
<td>add1:F</td>
<td>add(1)(q13)</td>
<td>F1q13 set to 1</td>
</tr>
<tr>
<td>Deletion until the telomere</td>
<td>del1:L</td>
<td>del(1)(q13)</td>
<td>L1q13 till the end of 1q arm set to 1</td>
</tr>
<tr>
<td>Interstitial deletion</td>
<td>del2:L1F,F</td>
<td>del(1)(q13;q22)</td>
<td>L1q13 till L1q22 set to 1, F1q13 and F1q22 set to 1</td>
</tr>
<tr>
<td>Translocation</td>
<td>t2:F,F</td>
<td>t(2;3)(q13;p12)</td>
<td>F2p13 and F3p12 set to 1</td>
</tr>
<tr>
<td>Dicentric chromosome</td>
<td>dic2:LF,LF</td>
<td>dic(2;3)(q13;p12)</td>
<td>L2q13 till the end of 2q arm set to 1. L3q12 till the end of 3q arm set to 1. F2q13 and F2q12 set to 1</td>
</tr>
<tr>
<td>Duplicated chromosome region</td>
<td>dup2:G1F,F</td>
<td>dup(1)(q12;q22)</td>
<td>G1q12 till 1q22 set to 1. F1q12 and F1q22 set to 1</td>
</tr>
<tr>
<td>Homogeneously staining region</td>
<td>hsr1:F</td>
<td>hsr(1)(q12)</td>
<td>F1q12 set to 1</td>
</tr>
<tr>
<td>Inversion</td>
<td>inv2:F,F</td>
<td>inv(3)(p25q11)</td>
<td>F3p25 and 3q11 set to 1</td>
</tr>
<tr>
<td>Ring chromosome</td>
<td>r2:LF,LF</td>
<td>r(7)(p22q11)</td>
<td>L7q22 till end of 7p arm set to 1. L7q11 till end of 7q arm set to 1. F7p22 and F7q11 set to 1</td>
</tr>
<tr>
<td>Telomeric association</td>
<td>tas2:F,F</td>
<td>tas(6;19)(p25;p13)</td>
<td>F6p25 and F19p13 set to 1</td>
</tr>
<tr>
<td>Insertion</td>
<td>ins2:F,F</td>
<td>ins(9;22)(q34;q11)</td>
<td>F9q34 and F22q11 set to 1</td>
</tr>
<tr>
<td>Tricentric chromosome</td>
<td>trc3:LF,LF</td>
<td>trc(9;9;20)(p11;p13;q11)</td>
<td>F9p11, F9p13 and F20q11 set to 1. L9p11 till end of 9p arm set to 1. L9p13 till end of 9q arm set to 1. L20q11 till end of 20q arm set to 1.</td>
</tr>
<tr>
<td>Isochromosome</td>
<td>i1:GFL!</td>
<td>i(9)(q10)</td>
<td>G9q10 till end of 9q arm set to 1. F9q10 set to 1. L9p10 till end of 9p arm set to 1.</td>
</tr>
</tbody>
</table>

**Table 2.1. Cytogenetic events and equivalent mappings in the LGF model**

Thus, the ISCN karyotypes are rewritten to purely represent the biological result at each location in the LGF model using the workflow shown in Figure 2.1. Although antecedent events, such as translocation and deletion, were not preserved in the model, the biological effect of those events was represented, allowing for comparison across aberrations with similar biological effects.
2.4 Results

The LGF model was applied to the Mitelman collection of 66,362 cancer-related karyotypes in ISCN notation. We successfully parsed 98% and mapped 87% of the entries. During a subsequent validation that examined why some entries could not be parsed or mapped, four groups of errors were identified shown in Table 2.2. First, uncertain karyotypes were 20% of the total (13,400) representing 66% of the errors. In ISCN, uncertain karyotypes contained question marks to indicate that some part of the karyotype could not be fully ascertained; these were intentionally designated parser errors. Second, set of errors were mapping errors that represented 8.2% of the total (5,482) that were 27% of the errors. These were syntactically valid ISCN karyotypes that could not be mapped to the current LGF model. Third were syntax errors that were 1.6% of the total or 1,055 cases, representing 5.2% of the errors. These occurred usually because the entry contains characters such as dashes or parentheses in locations not permitted by any ISCN standard. Finally, there was an “other” error category consisting of 0.4% of the total, or 280 cases making up 1.4% of the errors. This category contained all other failures and included situations in which large derivative chromosomes were not translated.
The uncertain karyotype errors, which are 66% of the total error, were intentional, in terms of being designed to fail by the parser because the karyotype information was incomplete. Removing these from the total error, the actual parsing accuracy is relatively high - 93% succeeded. Consequently, the vast majority of errors (27%) were produced by the mapping function. This is because the mapper function is a set of rules and if it comes across an aberration and lacks a rule for it the mapper fails the entire karyotype not simply that specific aberration. This means a karyotype with five aberrations, four of which can be mapped, but one that fails mapping will be discarded as completely failed. This was designed so that any parsed and mapped karyotype was completely represented as opposed to partially represented. This stringent design choice allows for a more complete analysis of karyotypes since we are only observing karyotypes in their full form. In the future more production rules can be added to the mapper by observing which aberrations fail mapping. This will drive the model to perform better and more completely.

<table>
<thead>
<tr>
<th>Error Group</th>
<th>Number of Karyotypes</th>
<th>Percent of total karyotypes</th>
<th>Percent of errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncertain karyotype</td>
<td>13,400</td>
<td>20%</td>
<td>66%</td>
</tr>
<tr>
<td>Mapping</td>
<td>5,482</td>
<td>8.2%</td>
<td>27%</td>
</tr>
<tr>
<td>Syntax</td>
<td>1,055</td>
<td>1.6%</td>
<td>5.2%</td>
</tr>
<tr>
<td>Other</td>
<td>280</td>
<td>0.4%</td>
<td>1.4%</td>
</tr>
</tbody>
</table>

Table 2.2. Parsing in mapping errors classified by type of error.
In sum, the LGF model was able to accurately parse and map 87% of the 66,362 karyotypes in the Mitelman dataset. This shows that the LGF model can accurately parse and map the vast majority of inputted karyotype data. This is particularly important given the large sample size of input karyotypes, demonstrating that the LGF model can manage large biomedical datasets, a major goal of TBI-oriented research.

2.5 Conclusion

A critical aspect to the system’s success is that it is a rule-based system on both the parsing and mapping steps. The parsing system is completely done using production rules taken from the ISCN handbook. This means that as long as the inputted karyotype was properly written, the system is able to accurately tokenize and parse the relevant information in a rules-based manner. This is also true of the mapping step, wherein we took all biological abnormalities and wrote out a rule to map the biological result of the abnormality. These rules were written with the help of, and reviewed by, domain experts to ensure that these biological mapping rules preserved the biological result of the cytogenetic event. By doing this we prevented misclassifications of cytogenetic abnormalities as have been noted in machine learning-based attempts to solve this problem. By using an extensive rule-based mapper, we guaranteed that even complex cytogenetic aberrations were properly biologically preserved in the LGF model. Because prior attempts at solving this problem did not accommodate large scale
parsing and mapping of karyotype data and many of the systems are no longer available, it is not practically possible to do a direct comparison between our system and previous attempts. Such a comparison would also be unfair to the prior attempts since computer technology has improved considerably since those attempts and the ISCN rules have also been expanded over that time.
Chapter 3: Disease Classification Using the LGF Model

3.1 Introduction

Disease classification is the process of identifying a disease state based on medical data (e.g., Liu et al., 2008). Karyotype data are a readily available form of genomic medical data. Utilizing the LGF karyotype representation as a binary vector, a disease classification system could be built to utilize this previously-un-wieldy data.

The goals of this chapter are (1) to assess whether the LGF representations of karyotype data as a binary vector can be analyzed using machine learning algorithms to distinguish different diseases and (2) to discern which algorithms perform the best on this form of data. To meet these goals, we performed a set of benchmark tests. By using karyotype data obtained from eight diseases in the Mitelman database (as described in Chapter 2), we tested a set of machine learning algorithms to assess each method’s performance.

The results of this analysis help us to identify which machine learning algorithm would be a best fit for a disease classification system based on the LGF model. Such a system would be clinically useful as a decision support system, allowing clinicians to see whether their patient has a similar karyotype to any past patients and what those patients were diagnosed with (Arocha et al.,
To this end, we examined four different forms of acute myeloid leukemia (AML) and assessed how well machine learning algorithms were able to differentiate different forms of AML.

3.2 Data

The data in this analysis were taken from the Mitelman database and processed, as noted in Chapter 2. Eight different diseases had over 1,000 parsed and mapped karyotypes, and these diseases were selected for testing. Using this as the primary source of karyotype data, two analyses were conducted to verify the utility of the data. These experiments provide a “template” demonstrating the utility of the LGF representation to provide actionable knowledge for clinical and translational applications.

3.3 Automated Disease Classification

This analysis was conducted to test whether the LGF model data could be leveraged for clinical applications. The experimental design is shown in Figure 3.1.

![Figure 3.1. Experimental design for automated disease classification using karyotype data](image)

<table>
<thead>
<tr>
<th>Data Sample Size</th>
<th>Chosen Classifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Top 8 disease classes with over 1000 samples each chosen for experiment</td>
<td>• Four classification methods were chosen –</td>
</tr>
</tbody>
</table>
| • Sample sizes: 100, 200, 500, 750, 1000 from each class | 1. Support Vector Machine (SVM)
2. Random Forest (RF)
3. Naïve Bayes (NB)
4. Linear Logistic Regression (LLR) |

Evaluation using ten fold cross-validation scheme
Within the Mitelman database there were eight diseases that had more than 1,000 samples shown in Table 3.1. Benchmarking tests were performed utilizing these eight diseases.

<table>
<thead>
<tr>
<th>Disease name in Mitelman</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute_lymphoblastic_leukemia_lymphoblastic_lymphoma</td>
<td>6417</td>
</tr>
<tr>
<td>Acute_myeloid_leukemia_NOS</td>
<td>3117</td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>2345</td>
</tr>
<tr>
<td>Acute_myeloblastic_leukemia_with_maturation_FAB_type_M2</td>
<td>2156</td>
</tr>
<tr>
<td>Chronic_lymphocytic_leukemia</td>
<td>2051</td>
</tr>
<tr>
<td>Chronic_myeloid_leukemia-t-9-22</td>
<td>1856</td>
</tr>
<tr>
<td>Acute_myelomonocytic_leukemia_FAB_type_M4</td>
<td>1694</td>
</tr>
<tr>
<td>Acute_promyelocytic_leukemia_FAB_type_M3</td>
<td>1099</td>
</tr>
</tbody>
</table>

**Table 3.1. Diseases with over 1000 karyotype samples in Mitelman**

The four machine learning methods used in benchmarking classification tests were (1) Support Vector Machine (SVM), (2) Random Forest (RF), (3) Naïve Bayes (NB) and (4) Multi-Label Logistic Regression (MLR). These four methods were selected because they each represent a different underlying mathematical model for achieving the same goal of feature-based pattern recognition and sample classification. SVM is a supervised learning model based on regression analysis. RF is a forest-based decision-tree algorithm that generates a large number of decision-trees and then selects the highest performing tree from the forest. NB is a probabilistic independence model based
on Bayesian classification. MLR is a way of using logistic regression to classify data with multiple data labels. These algorithms were implemented using the WEKA tool (Hall et al., 2009).

The analyses were performed on five different heuristic karyotype sample sizes: 100, 200, 500, 750, and 1,000 per disease type. Each algorithm was run for all the different sample sizes to determine which method yielded the highest F measure per sample size. A ten-fold cross validation scheme was used for evaluation (Simon et al., 2011). The results of these experiments are shown in Table 3.2

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>SVM</th>
<th>RF</th>
<th>NB</th>
<th>MLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.383</td>
<td>0.582</td>
<td>0.471</td>
<td>0.605</td>
</tr>
<tr>
<td>200</td>
<td>0.383</td>
<td>0.588</td>
<td>0.458</td>
<td>0.604</td>
</tr>
<tr>
<td>500</td>
<td>0.537</td>
<td>0.553</td>
<td>0.501</td>
<td>0.627</td>
</tr>
<tr>
<td>750</td>
<td>0.567</td>
<td>0.643</td>
<td>0.517</td>
<td>0.653</td>
</tr>
<tr>
<td>1000</td>
<td>0.578</td>
<td>0.635</td>
<td>0.504</td>
<td>0.641</td>
</tr>
</tbody>
</table>

Table 3.2. Weighted average F-Measure for Classifiers at Different Sample Sizes

These tests showed that a multi-label logistic regression at a sample size of 750 was the slightly more effective performer by weighted F-measure, or the harmonic mean of precision and recall, as shown in Figure 3.2. For context if there was no disease specific information contained in the karyotypes in the
classification would be completely random which would yield and F measure of .125, or 1/8th correct at random because there are eight classes. Multi-label logistic regression with the sample size of 750 was chosen to perform additional analyses because it had the highest F-measure across all sample sizes.

![Figure 3.2. Classification accuracy of different classifiers for different sample sizes](image)

3.4 Case Study of Acute Myeloid Leukemia (AML)

To test a positive control case, the classification between acute promyelocytic leukemia (APL) and acute myeloblastic leukemia type M2 (AML-M2) was observed. APL is largely identified by a specific cytogenetic event, t(15;17)(q22;q12) (Kotiah et al., 2013). When comparing these two diseases,
there was a high correct classified instance of 95%. This is not surprising because 72% of APL cases contained fusions at 15p and 53% of cases had a fusion at 17p. This is in contrast to the AML-M2 cases that only had 1.0% of cases with this translocation. The AML-M2 class also contained a common translocation present in 38% of AML-M2 cases, translocation 21q-8q (Nishii et al., 2003). The percentage of APL cases that contained this translocation was less than 1%. When these two translocations are the only features used in classification, the correct classification percentage drops to 83%. This demonstrates that there are other cytogenetic features that play a key role in distinguishing between these two subclasses of AML.

There are four subclasses of AML that have over 1,000 karyotypes logged in Mitelman. All four of these classes were compared to each other using multi-label logistic regression with the sample size of 750 doing tenfold cross validation that produced the confusion matrix shown in Table 3.3.

<table>
<thead>
<tr>
<th>Classified Class</th>
<th>Target</th>
<th>Disease label</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 506</td>
<td>B 134</td>
<td>C 93</td>
</tr>
<tr>
<td>281</td>
<td>376</td>
<td>87</td>
</tr>
<tr>
<td>291</td>
<td>79</td>
<td>371</td>
</tr>
<tr>
<td>53</td>
<td>11</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 3.3. Classification results as a confusion matrix
APL, Group D, is the highest correctly classified group; this is expected due to the fact that APL is defined by its cytogenetic karyotype. The other three groups have much lower correct classification percentages. This is important since, if there was no use of the cytogenetic data as a classifier, Groups A, B, and C would each have approximately one-third of the 750 samples; i.e., a random distribution across these three groups. This however was not the case, as approximately 50% in Groups B and C and 67% in Group A were correctly classified. This analysis again demonstrates that the binary representation of karyotype data does indeed yield disease specific information that can be found using data science methods.

3.5 Results

The above machine-learning based experiments collectively validate the LGF model by showing that the model’s binary representation accurately allows conversion of the ISCN’s formatted karyotype data into a clinically actionable format. By doing so, they further indicate that there is an abundance of biological information embedded in the karyotype data.

This observation – that karyotype data contain information that is disease-specific, thus allowing for data-driven classification to be conducted utilizing machine-learning methodology – is a critical point of discussion. This observation is supported by the fact that there are a small number of cytogenetic abnormalities that are clinically used in disease classification. The reason for the
small number of clinically used aberrations is primarily due to the formatting of karyotype data in the ISCN text string rather than the utility of the data itself. Thus, the usage was limited to human visual inspection. The data in its original form were not compatible for any automated applications. However, once the data become machine-readable, the constraints of the data storage format and the limitations presented by human visual inspection were removed. The entirety of the karyotype’s features can be accessed and utilized in several clinical applications, such as a data-driven classification scheme described above.

This research also demonstrates that there are more cytogenetic aberrations that can be used in disease classification than are currently employed. This was shown with the AML-M2 versus APL classification experiments. It is also important to recognize that the machine-learning models performed this disease classification with a completely data-driven method. Thus previously unknown cytogenetic aberrations that may play a key role in disease classification and thus disease progression can now be identified. This could open the door for future research by giving disease domain specialists greater insights into potential genetic factors related to their disease of interest.

3.6 Limitations and Future Work

One limitation of this study is that we used data from the Mitelman database. Although this database represents the largest number of publically available karyotype data, it has a bias towards rare aberration events. This bias
is because published karyotypes are more likely to be rare events due to the pressure of publishing something novel. This means that using it to assess population level aberration data will be biased towards more complex karyotypes. However, this limitation is not severe in our study since we were assessing methodologies that are agnostic to karyotype content. In the future, disease classification systems will be built using clinical populations to help address this limitation.

We also limited this preliminary research to “out-of-the-box” WEKA implementations of the four methods described previously. These methods cover a broad reach of machine-learning methodologies as they relate to the classification of binary data. We plan to pursue two specific directions as part of the future research for the classification task. The first is experimenting with other methods and parameters under different machine-learning classifiers. The goal is to identify the right classifier based on the results from the out-of-the-box solutions. The second direction is experimenting with multi-label classification, meaning each karyotype can be assigned multiple class labels with varying confidence scores. The intuition is that for the automated disease classification task this can lead to simultaneous risk assessment of multiple diseases for clinical decision support. Both these pursuits would enable the automated disease classifier to become an effective instrument in the precision medicine toolbox.
3.7 Conclusions

Automated disease classification utilizing machine learning systems on binary karyotypic data yields promising results. This makes sense since binary features are commonly used in machine learning systems (Speed and Tavare, 2011). As demonstrated in this chapter, machine learning algorithms are capable of distinguishing between similar conditions in a reliable form. This both shows that computational analyses can be performed on LGF model data as it is represented in a binary format as well as the fact that the data contain important genomic information. In the future other forms of genomic as well as phenotypic data could be added into an automated disease classification system to increase data coverage as well as increase performance of the model.

The present research also has great potential as a hypothesis-generating model. By taking large scale karyotype data and processing it through machine learning workflows, cytogeneticists can determine data-driven features within the karyotype that distinguish between different clinical conditions. This both increases the utility of the karyotype data and highlights the research benefits of hypothesis-generating systems for use in translational biomedical research.
Chapter 4: Clinical Cytogenetics and the LGF Model

4.1 Introduction

This chapter explores the potential of the LGF model platform to discover new recurrent cytogenetic aberrations in CLL. The importance of this chapter is to highlight the use of karyotypes in retrospective data mining studies to find previously unrecognized novel cytogenetic aberrations that recur in CLL. This is especially relevant following the advent of stimulating CLL karyotypes (Heerema et al., 2010) since stimulated karyotypes display a wider range of aberrations than non-stimulated karyotypes. It also is shown that hematological malignancies other than CLL, as well as lymphomas, could be studied using this methodology and platform.

In this chapter I specifically use known recurrent cytogenetic aberrations, such as Trisomy 12 and the deletion of 13q14, as positive controls to test the accuracy of the LGF model and its ability to represent the karyotype data. This serves as a method of validation since we can compare the rates of cytogenetic aberrations in our patient cohort with the percentages of patients who expressed these aberrations from previous studies. Since we are looking at a large patient population from The Ohio State University Wexner Medical Center, we would expect that this patient population should show these various aberrations at approximately the same percentages as the broader patient population. Thus,
this chapter not only serves as further validation that the LGF model can accurately represent and capture ISCN karyotype data, but also as an example of the new discoveries that can be generated through the use of this model.

4.2 Data

As we have shown in previous chapters, the Mitelman database represents an important publically available dataset of karyotype information. However, there are several qualifiers that limit the use of this database in some applications. One element that is important to understand is the nature of the cohort of patients present within that database. Since all of the karyotypes in Mitelman are there because they have been published, there is a predisposition for karyotypes to be more atypical than what is commonly found within that disease. This is useful when trying to find unusual aberrations within a disease; however it is problematic when trying to assess recurrent cytogenetic abnormalities within a disease population. For this reason, although it is useful to apply Mitelman data for other research purposes, when studying a disease population, we do not want this bias in favor of rare events in our study.

Another potential issue with the Mitelman dataset, in this case specifically in its application in CLL research, is the advent of stimulated karyotypes. Stimulated karyotypes are taken from cells that were chemically stimulated to induce cell growth (Heerema,N.A. et al. 2010). This is needed to culture CLL cancer cells to obtain their karyotype. Before stimulation, the vast majority of CLL
karyotypes appeared normal; i.e., they lack any cytogenetic abnormalities. This is largely due to the fact that CLL cells are hard to culture and so often the cells being observed are the patient’s normal background cells rather than the CLL cancer cells (Heerema, N.A. et al. 2010). Since we do not know which karyotypes in Mitelman were taken from stimulated versus non-stimulated cells, this uncertainty introduces a bias.

To avoid these biases, we observed karyotypes taken from patients at the Ohio State University Wexner Medical Center and selected only karyotypes taken since October 2007, the date that the Medical Center started stimulating CLL karyotypes. By using these data, we were able to avoid this bias inherent in Mitelman as well as avoid the problem of observing both stimulated and non-stimulated karyotypes in this study.

For the study presented in this chapter, we observed 1,536 karyotypes taken from 776 individuals with CLL who were treated at The Ohio State University Wexner Medical Center from October 2007 until the present. There were 502 males and 268 females. The ratio of males to females was 1.87:1, or approximately 2:1. This is expected within the broader CLL population since the disease is more common among men; men also have a worse prognosis than women. Six patients were classified as “other” based on gender self-identification and these six were included in the present analysis. In this chapter, all population percentages are calculated based on the number of patients rather than the number of karyotypes since a single patient can have multiple karyotypes.
The CLL cohort database from The Ohio State University also contained some inherent bias since the university is a tertiary care center, meaning the patients have generally sought initial treatment elsewhere, only coming to OSU after they have been diagnosed with a more aggressive form of the disease. One exception to this is when the patient is a resident of the greater Columbus area. Nonetheless, for this reason our patient cohort will have an innate bias towards more aggressive forms of the disease. Thus, while the Mitelman database contained some bias towards rare events, the OSU database contained some bias towards more aggressive cases of CLL. This is noted by the comparative rate of deletion of 17p; it is at 8.6% for this cohort compared to 5.0%, or the average rate of deletion (Puiggros et al., 2014), among the broader population of CLL patients. This same bias in our sample population towards more aggressive disease was also reflected in the fact that the median age at the time of diagnosis in our cohort is 55 years old compared to the average of 71 years old.

4.3 Methods

We began with a patient cohort from the Ohio State University that contained 5,552 karyotypes taken between 2002 and 2014. We then performed some preprocessing on these karyotypes since some of the entries did not contain a karyotype. We removed the following: (1) 33 karyotypes because they were listed as “no dividing cells”, (2) 213 karyotypes because they read “test not performed”, and (3) 93 karyotypes because they were listed as “FISH only”. All of
these were removed because there was no ISCN karyotype to parse and map. In addition, 631 karyotypes were removed since they contained a ‘?’ character. The presence of this character indicated that the cytogeneticist was uncertain of the location of a particular event, which is required to map the karyotype within the LGF model. In the future we may be able to use these karyotypes since our system is able to parse this particular character, but more work needs to be done in order to effectively map events without location information.

After removing these 970 karyotypes, we were left with 4,582 karyotypes. We then proceeded to parse and map all of these karyotypes through the LGF model platform. Of these, we were able to successfully parse and map 4,148 karyotypes, representing a success rate of 92.2%.

We then filtered the 4,148 karyotypes by the date that they were taken. Only ISCN karyotypes that were obtained from October 2007 until the present were selected since OSU started stimulating CLL cells to study karyotypes in October 2007. This left us with a sample of 1,537 parsed and mapped karyotypes taken from stimulated samples. These karyotypes came from a population of 776 patients. We then took this group of patients’ karyotypes and observed which events within the LGF binary model were occurring and at what frequency and percentage of the overall population they were occurring. It is this group of 1,537 karyotypes taken from 776 patients at The Ohio State University Wexner Medical Center that form the basis for the following results.
4.4 Results

In this section the results of the analysis on the 1,537 parsed and mapped, stimulated karyotypes are presented. A summary of results is presented in Table 4.1. This section is divided into known positive recurrent aberrations (4.4.1) and rare or novel discoveries (4.4.2).

<table>
<thead>
<tr>
<th>Cyto Band</th>
<th>OSU</th>
<th>Mitelman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss of 13q14</td>
<td>13.9%</td>
<td>11.0%</td>
</tr>
<tr>
<td>Loss of 17p13</td>
<td>8.6%</td>
<td>6.7%</td>
</tr>
<tr>
<td>Gain of 12</td>
<td>22.3%</td>
<td>34.6%</td>
</tr>
<tr>
<td>Loss of 11q23</td>
<td>16.8%</td>
<td>7.7%</td>
</tr>
<tr>
<td>Fusion of 14q32</td>
<td>8.6%</td>
<td>20.7%</td>
</tr>
<tr>
<td>Loss of 6q21</td>
<td>5.8%</td>
<td>5.4%</td>
</tr>
<tr>
<td>Loss of 8q24</td>
<td>4.2%</td>
<td>1.9%</td>
</tr>
<tr>
<td>Gain of 3q26</td>
<td>2.4%</td>
<td>1.8%</td>
</tr>
<tr>
<td>Loss of 18p11</td>
<td>2.9%</td>
<td>2.5%</td>
</tr>
</tbody>
</table>

Table 4.1. Cytogenetic events in the OSU and Mitelman cohort by affected population

4.4.1 Findings of known clinical cytogenetic abnormalities

There are some cytogenetic aberrations that are known to have clinical significance. Such aberrations can provide positive controls for the findings in the present chapter. We would expect to find these aberrations in our OSU sample population at approximately the same percentage as they are observed in other populations. This comparability of results provides a form of verification that we
are in fact capturing enough data to accurately represent the population. It further supports the conclusion that there are no significant systematic biases based on karyotype content when translated through the LGF model.

From prior research (Döhner et al., 1999), we know that the most important aberrations to observe as positive controls are the following: loss of 13q14, loss of 17p13, Trisomy 12, loss of 11q23, and fusions involving 14q32. In this section we highlight our findings (Table 4.1) relating to these aberrations and then discuss their significance.

The loss of the band 13q14 is the most common chromosomal abnormality observed by FISH in CLL patients (Ouillette et al., 2011). It affects approximately 50% of the CLL population and is normally associated with a good prognosis (Dal Bo et al., 2011). It is biologically important due to the loss of mir-15a and mir-16-1 (Calin et al., 2002), which are microRNAs that bind and suppress mRNA transcripts for the BCL2 gene. BCL2 is anti-apoptotic. When mir-15a and mir-16-1 are not present due to the loss of 13q14, the BCL2 gene is overexpressed and thus becomes oncogenic.

Within our sample population, 52% of patients were positive for the loss of 13q14 as identified by the FISH analysis. However, when processed through the LGF cytogenetic model, only 13.9% of patients had a loss at that location. This indicates that the majority of cases were not detectable using only the ISCN karyotype data compared to utilizing the FISH analysis. This makes sense given that the FISH analysis is able to detect a smaller region (2290 bp) (Rogan et al.,
than is visibly observable by the cytogeneticist (5 mb) (www.coriell.org) while generating the ISCN karyotype.

This raises an interesting question as to the actual number of patients with different genetic aberrations. If some number of patients present a problem at the karyotype level, then potentially a much larger number of patients might have the same issue but it is in a minimally lost or gained region that might only be observable using FISH or other higher resolution techniques.

Trisomy 12 is the most common karyotypic abnormality observed by cytogeneticists in CLL patients, present in approximately 20% of the general CLL population (Puiggros et al., 2014). In our data, between 21.8% and 22.8% of the population gained regions on chromosome 12, depending on which cytogenetic band was observed. The most commonly gained cytogenetic band was 12q24 at 22.8%. This makes it the most common aberration in the entire dataset. Trisomy 12 has been widely studied and there are many ideas as to its effect in CLL. When the entire chromosome is gained, there is a large amount of genetic material to take into consideration, and isolating the minimally gained region of importance is difficult. By observing our patient cohort through the LGF model, it appears as though the 12q24 region may contain the minimally significant gained region since it is differentially gained compared to the other regions on chromosome 12. This means that there are some patients (approximately seven) who experienced a gain of only this region without having complete trisomy, something captured by the LGF representation.
The cytogenetic event of the loss of chromosome 11q23 is another positive control aberration, occurring in the broader population in approximately 20% of CLL cases (Puiggros et al., 2014). We observed the loss of 11q23 in 17% of our sample population. Interestingly, other regions on the 11q arm were lost at a much lower rate than 11q23 in our sample. Losses on 11q24 were at a rate of 5.5% and losses on 11q14 were at a rate of 5.6%. This is biologically important since the 11q23 region contains the ATM gene that helps regulate cell growth. The fact that the LGF model was able to identify the minimally deleted genetic region based on the observation of its occurrence in the overall population is promising since it shows that large scale observational studies of karyotype data reformatted into the LGF model may allow researchers to identify minimally deleted regions at the cytogenetic level.

Another clinically important cytogenetic region previously linked to CLL is 17p13 (e.g., Rossi et al., 2010). This is the site of the TP53 gene, which is a negative regulator of apoptosis. TP53 is the most commonly mutated gene in cancer (Rudenko et al., 2008). A deletion in this region is normally a marker of a bad prognosis in CLL. Approximately 5% of patients are expected to have a loss of 17p13 (Puiggros et al., 2014); however in our cohort the number increased to 8.6%. This may be partially due to this cohort coming from a tertiary care center where we have a bias towards more aggressive diseases, since a loss at 17p13 is often an indicator of a more aggressive case.
Further, translocations involving 14q32 are common in CLL, although the band that is involved in the translocation with 14q32 is not always the same (Cavazzini et al., 2008). Fusion events involving this region occurred in 8.6% of our population and occur at an expected rate of 4% to 9% of the general population of CLL patients (Cavazzini et al., 2008). 14q32 is a genetically important region since it is the site of the IGVH locus. IGVH is the variable region of the immunoglobulin heavy chain and is important to observe whether the patient has a mutation in this region because it identifies treatments that may be more effective for that individual.

Finally, our research indicated that a loss of 6q21 occurred in 5.8% of our patient population. This site has been frequently studied (e.g., Stilgenbauer et al. 1999) and they report a loss of 6q21 occurring in 7% of the population. Although the exact effect of this loss at the gene level is not known, prior research has shown that while this loss is “… associated with a higher white blood cell count and more extensive lymphadenopathy, [there was] no significant difference in treatment-free survival or overall survival” (Stilgenbauer et al., 1999). One important fact is that 6q21 was lost in 5.8% of our population while the rest of the 6q arm was lost in only 4.9% of the population. This suggests that the 6q21 band may hold the minimally deleted region necessary to cause a phenotypic effect.

4.4.2 Finding a novel set of cytogenetic abnormalities

A less common but previously noted cytogenetic region of importance is the 8q24 region (Gunnarsson et al., 2011; Rudenko et al., 2008; Brown et al.,
2012). The LGF model indicated that this specific region is lost in 4.2% of our patients and that the 8p arm is lost in 4.1% of our patients. Prior research has similarly indicated losses of the 8p arm (Puiggros et al., 2014); however, most studies focus on the 8q arm and particularly the gains rather than the losses on this arm. This is because the c-MYC gene is located on the q arm of chromosome 8. The c-MYC gene codes for a nuclear phosphatase, which plays an important role in apoptosis through its regulation of the cell cycle. Given the importance of its oncogenic properties, any alterations of this gene are significant to researchers.

In our patient cohort, however, the 8q arm experienced gains in only 1.3% of patients compared to its observed loss occurring in 4.2% of that same population. This raises some interesting questions, namely whether the loss or gain of this region can potentially disrupt cell cycle regulation regardless of the event. Although losses at this region have been tangentially noted by another researcher (Schweighofer et al., 2013), the focus of their research was not on losses at 8q. Further modeling may need to be done to predict the impact on the cell cycle with a loss of the c-MYC gene but it appears as though the loss of this region occurs at a much greater frequency than the gain of this region.

Another noticeable event is the gain of the chromosome region 3q26. This even occurred in 2.4% of our sample population. The importance of this event may lie in the function of the PIK3CA and BCL6 genes (Brown et al., 2012). PIK3CA mutations have been shown to occur in a variety of cancers and it is
believed to have oncogenic properties (Ma et al., 2000). BCL6 is a zinc finger transcription factor in B cells (Ye et al., 1993). This makes it particularly important in lymphomas and leukemias. It is often mutated in diffuse large B cell lymphoma (Skinnider et al., 1999). For this reason it makes sense that this gene may be important in understanding the progression of CLL.

The loss of 18p11 occurred in 2.9% of our patient population based on the LGF model. This is an interesting observation since the loss of 18p11 has not been highly reported in CLL cases, occurring in 2.5% of CLL cases in Mitelman (see Table 4.1). It has, however, been reported as being associated with a large number of other cancers (Tran et al., 1998). This high association with other cancers is primarily due to the loss of two genes: PTPRZ1, a protein tyrosine phosphatase, and ZBTB14, a c-MYC transcriptional repressor. Both of these genes have been identified as oncogenic in lung cancer, brain cancer, breast cancer, and esophageal squamous cell carcinoma (Karkera et al., 2000; Tran et al., 1998). This implies that these two genes probably have a similar oncogenic effect in these CLL patients.

4.5 Conclusion

In this chapter the utility of the LGF parsing and mapping model to perform retrospective studies that lead to novel discoveries has been highlighted. One important conclusion drawn from this research is that processing such large-scale population karyotype data through the LGF model can allow us to detect
potentially useful FISH probe regions. The loss of 18p11, for example, was observed in our patient population and represents a cytogenetic aberration that has been shown to play a causal role in other cancers but has not been reported in CLL. The identification of this aberration should serve as a target for future FISH probe research.

It is also important to note the disparity between cytogenetic data using the ISCN karyotypes processed through the LGF model and the cytogenetic data obtained using FISH probes. For example, we saw the loss at 13q14 occurring in 13.9% of patients through the LGF model in contrast to 52% of patients using FISH probes. This makes sense since FISH probes can observe a smaller region than is visible by karyotyping. This is important to consider with new discoveries such as 18p11 since only 2.9% of the population showed this loss karyotypically. It is possible that many more patients have the minimally deleted region, detectable only by using FISH. The important conclusion from these results is that we should develop constructive feedback between the results of our LGF cytogenetic model platform and those from FISH probes.

Another example of how the LGF model helps inform more targeted research studies is by looking at the loss of 6q21. It is currently not known why the loss of the 6q arm is genetically important in cases of CLL, but one specific region - 6q21 - is lost at a higher rate than other bands on that arm. It is likely that this band contains the minimally deleted region related to CLL. This type of observation, generated through the LGF platform, can help direct future research
towards a more targeted gene level study of the genes on this band and whether any of them possess oncogenic properties.

Overall this chapter has presented the results of a study of CLL patients in a clinical setting using the LGF model. These results demonstrate the utility of large-scale computational analyses of cytogenetic data by showing that novel cytogenetic findings can be made through retrospective studies of karyotype data. In the future we hope to perform such studies on other leukemias and lymphomas, specifically AML.
Chapter 5: Drug Repurposing

5.1 Introduction

The LGF cytogenetic platform allows the transfer of human-readable ISCN karyotypes to a machine-readable model for computational analysis. This chapter explores the potential of this platform to interact with other databases to produce clinically actionable results in the context of drug repurposing. Utilizing the cytogenetic data from the National Cancer Institute (NCI)-curated Mitelman database, triplets containing a drug, gene, and disease name were generated via a computational pipeline that connected various public drug-gene interaction data sources to identify potential drug repurposing research opportunities.

5.2 Drug Repurposing Background

The druggable genome has been at the forefront of biomedical research since the completion of the human genome project (Hopkins and Groom, 2002; Griffith et al., 2013). The one-drug-one-target concept has been expanded to higher orders of systems biology in order to treat more complex diseases and to curb drug resistance mechanisms commonly observed in targeted drug therapies (Butcher et al., 2004; Holohan et al., 2013). Despite advances in identifying the biomolecular underpinnings of many diseases, progress in developing successful targeted therapies is limited, in part, by the costly and inefficient drug development pipeline. For instance, it is estimated that an average of 1 billion U.S. dollars and 17 years is required to bring a new drug from research and
development to common clinical practice (Adams and Brantner 2006; Morris et al., 2011). Computational drug discovery and drug repurposing methods have gained popularity in recent years and may offer a more efficient means for identifying drug candidates (Hurle et al., 2013). This could mean that drug repurposing enhances precision medicine by allowing for the development of more personally tailored therapeutic strategies. Drug repurposing is also less expensive than current drug development and required less time, which directly helps patients. For these reasons, observing the karyotype as a means to discern genetic alterations that may be candidates for targeted drug repurposing therapeutics has great promise to speed up therapy development and cut the cost of that development for many diseases.

5.3 Drug Repurposing Research Design

Cytogenetic data are currently used to assess which treatment course to prescribe to a patient. For this reason, it makes sense to look at cytogenetic data as a starting point for drug repurposing studies based on genetic data. Such a study, as presented in this chapter, could not be done without a TBI-oriented scientific design and the use of the LGF model. This is the first drug repurposing study utilizing cytogenetic data.

Although there are many potential applications of the loss-gain-fusion (LGF) model, drug repurposing was selected as a use case due to its enormous clinical significance. To that end, all karyotypes in the Mitelman database for
each disease were parsed and mapped, as described in Chapter 2. Each binary-represented karyotype was then separated based on its disease label and the percentage of individuals who had a cytogenetic event at each location was calculated per disease label. Disease labels were based on the morphology label given to the karyotype in the Mitelman database. This provided the percentage of the population that had a cytogenetic abnormality within each disease cohort.

Then, diseases that had fewer than 50 patients were removed. This was heuristically done to eliminate diseases that were underrepresented karyotypically. To narrow the LGF search space, we heuristically only selected gained cytogenetic features that occurred in 20% or more of their disease cohort. This is because we need the gene the downstream-targeted drug therapy suppresses to be present.

The next step was to connect these karyotype data to multiple publicly available datasets. First the selected cytogenetic bands were run through the NIH human genome project to find all genes located on each gained cytogenetic region. This extracted all HUGO [Human Genome Organization] Gene Nomenclature Committee (HGNC) gene names for all genes that were present within one of those cytogenetic band areas. We then converted the HGNC gene name into the ENTREZ gene name and inputted these gene names into the KEGG database to extract pathway information.

We then used the HGNC name to utilize the NCI Biomart system (Durinck et al., 2005) as well as the NCI Nature Curated pathways (Schaefer et al., 2009)
and the Reactome pathway information (Matthews et al., 2009). This gave us a set of pathway information that could be stored by the pathway name given by each of the three systems: KEGG (Kanehisa and Goto, 2000), NCI Nature Curated, and Reactome.

Next, the HGNC gene names were run through the drug-gene interaction database (DGIdb). Searches were conducted to find drugs that were targeted suppressors for the input gene since in all cases these genes may be overexpressed because they are coming from a gained chromosomal region. Through mining this database, we were able to obtain the primary drug name, drug interaction, resource, and the drug accession number. A diagram of this workflow is shown in Figure 5.1.
Figure 5.1: Information workflow for connecting publicly available data sets
After linking all of these data sources, we summarized all of the results in a MySQL database which contained the following field information for all entries: HGNC gene symbol, HGNC gene name, ENTREZ gene ID, drug accession number, primary drug name, drug interaction, drug resource, cytogenetic location, LGF model, NCI nature curated pathway, Reactome pathway, KEGG pathway, disease name, number of patients in disease cohort and finally percentage of population expression within the cohort. By querying this database for paths between gene, disease and drug we were able to generate triplets. This was done for every gene, consisting of the each HGNC gene symbol, the primary drug name and the disease cohort name. Each triplet thus consisted of a disease, a potential over-expressed gene, and a targeted suppressor for that gene.

Since novel drug repurposing possibilities were the focus of this research, a search was then conducted for each disease-drug-gene triplet to see if any prior research had been done on that disease with that drug. To determine this, the United Medical Language Systems (UMLS) for concept unique identifiers (CUI) was used to mine all available PubMed abstracts for references to these triplet co-occurrences. Each gene, drug and disease has its own unique cui code. By re-labeling all diseases, genes and drugs based on their cui and then searching PubMed abstracts for co-occurrences of triplets or pairs of cui within the same PubMed abstract, triplets that had already been studied and published upon could be identified. Searches were run for the presence of at least one abstract that had a co-occurrence of cui codes in all three combinations: gene-
drug, drug-disease and disease-gene. Triplets were selected when the co-occurrence of the disease-drug pair returned no results and where the co-occurrence of the disease-gene and gene-drug pairs returned at least one result each. A diagram illustrating these interactions is shown in Figure 5.2.

After filtering, only triplets with published evidence that the gene and disease were linked but had no published evidence that the drug had ever been used for that disease were chosen. If the gene and disease were linked and the drug suppressed that gene product, then logical congruency suggests that the drug should impact the clinical phenotype of that disease.
5.4 Results

By combining and analyzing publically available data sets (Mitelman, DGIdb, NCBI, NCI, Reactome, and KEGG) through the LGF cytogenetic platform, 68,543 disease-drug-gene triplets were discerned for cases where the gene was overexpressed due to a gain and the drugs were targeted suppressors, subset shown in Figure 5.3. A total of 69 diseases were covered within the triplets. In none of these cases did the literature search return any drug-disease pairings indicating that these drugs had previously been used for these diseases. However, the co-occurrence of the disease and its paired gene occurred at least once in all cases. This strengthens the validity of the relation between the disease and the gene since there is some literature as well as the cytogenetic data to support this linkage. The fact that there were no disease-drug relations within this set indicates that further research into these compounds, drugs, or therapeutics is needed; thus this research provides an ideal springboard for future drug repurposing studies.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>Entrez Gene ID</td>
<td>HGNC Gene</td>
<td>Cytogenetic</td>
<td>Primary Drug</td>
<td>Name</td>
<td>Drug Interact</td>
<td>Percent Population Expression</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Nonneoplastic mesenchymal disorder/lesion</td>
<td>3757 KCMH2</td>
<td>7q26</td>
<td>VERAPAMEL</td>
<td>inhibitor</td>
<td>53.535334</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonneoplastic mesenchymal disorder/lesion</td>
<td>5244 ACHB</td>
<td>7q12</td>
<td>BENZQUINAMIDE</td>
<td>inhibitor</td>
<td>53.535334</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teratoma (mature and immature)</td>
<td>5288 PIK3CG2G</td>
<td>12p121</td>
<td>SODAORETIN</td>
<td>inhibitor</td>
<td>52.499136</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teratoma (mature and immature)</td>
<td>5179 PD33A</td>
<td>12p121</td>
<td>IBUDILAST</td>
<td>inhibitor</td>
<td>52.499136</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teratoma (mature and immature)</td>
<td>775 CACNA1C</td>
<td>12p113</td>
<td>NEUROXIPINE</td>
<td>inhibitor</td>
<td>52.499136</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teratoma (mature and immature)</td>
<td>794 GRIN1B</td>
<td>12p131</td>
<td>BENDROPROD</td>
<td>antagonist</td>
<td>52.499136</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hepatoblastoma</td>
<td>146 ADRA1D</td>
<td>20p111</td>
<td>PROMAZINE</td>
<td>antagonist</td>
<td>44.274798</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hepatoblastoma</td>
<td>2280 FKBP1A</td>
<td>20p111</td>
<td>TACROLIMUS</td>
<td>inhibitor</td>
<td>44.274798</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilm tumor</td>
<td>775 CACNA1C</td>
<td>12p131</td>
<td>ISRADIPINE</td>
<td>inhibitor</td>
<td>38.914027</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retinoblastoma</td>
<td>23408 SR5</td>
<td>6p23</td>
<td>SURAMIN</td>
<td>inhibitor</td>
<td>38.028169</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Embryonal rhabdomyosarcoma</td>
<td>3576 ITGA4</td>
<td>2q311</td>
<td>COP23</td>
<td>antagonist</td>
<td>36.907463</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teratoma (mature and immature)</td>
<td>284 ANGPT1</td>
<td>8q231</td>
<td>AMG 386</td>
<td>inhibitor</td>
<td>34.117647</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malignant melanoma</td>
<td>284 ANGPT1</td>
<td>8q231</td>
<td>AMG 386</td>
<td>inhibitor</td>
<td>34.020619</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic lymphocytic leukemia</td>
<td>352 AVPR1A</td>
<td>12q142</td>
<td>CONIVAPTAN</td>
<td>antagonist</td>
<td>34.017467</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 5.3: Subset of drug disease gene triplet table results**
5.5 Discussion

There are several limitations to this study. PubMed does not contain all potentially relevant literature and thus there are some journals and recent research that may not have been included in the literature search. Additionally, non-significant or negative results may not have been published, thus biasing the sample journal set. Cytogenetic information found in Mitelman was collected from 1971 to the present by a wide array of individuals working in diverse institutions. Since only published karyotypes get entered into Mitelman, there is a tendency to select for more unusual cases that might not be representative of the population. This was the rationale for only selecting aberrations that occurred in 20% or more of each disease population.

There are many applications for this type of model. These knowledge models are proving invaluable in advancing new discoveries in the field of molecular genetics (Payne et al., 2010). One important application is in-silico discovery of unknown or alternative pathways in drug research, particularly since such pathways can dramatically alter patient response (Byrd et al., 2003).

The importance and utility of public datasets with regard to this study cannot be overstated. All information - from Mitelman to PubMed to DGIdb - was publicly available and vital to the completion of the study. This is an important point since it helps show that such large-scale analyses are only possible when researchers and government institutions make such information publicly available.
This also helps highlight why our research group is looking to make the parsing and mapping language publicly available so that other research groups can find innovative ways of utilizing karyotype data.

5.6 Conclusion

Utilizing publicly available datasets, a list of 68,543 drug-gene-disease triplets was generated. Each triplet contained a gene that is up-regulated, a drug that works to suppress or inhibit that up-regulated function, and a disease where the up-regulated gene is an implicated disease agent. This information may play a significant role towards drug repurposing efforts in the 69 diseases covered. It is anticipated that this information will prove useful to researchers in the domain of pharmaceutical drug repurposing and potentially in the treatment of these conditions.

Future research plans involve cross-referencing these drug predictions with other drug repurposing systems such as C-map and LINCS to help narrow down which drug repurposing hypotheses to test (Iorio et al., 2010) This can be done by cross-referencing gene expression profiles and drug response for each drug found in our study to the data in their systems. This would help strengthen the biological rationale behind certain high value repurposing hypotheses, thus paving the way for future validation studies.

This hypothesis-generating system is a representation of how TBI-oriented research operates. Multiple siloed datsets were brought together and linked
using common data models to allow information exchange between each data
set. This data synergy allowed for the model to flow from the cytogenetic band
level to the individual gene level to inhibitory drugs and current literature
knowledge seamlessly. This exemplifies why de-siloing data enables new and
interesting questions to be addressed and novel advancements to be made in
the field of biomedical research.
Chapter 6: Conclusions

6.1. Introduction

The research presented here has detailed a novel approach to analyzing cytogenetic data. In the previous chapters, the LGF model has been described, validated, and applied to both the Mitelman database and The Ohio State University CLL patient database. Further, this thesis has presented several applications of the LGF model that exemplify its utility in medical research. These applications have resulted in advances in disease classification, drug repurposing, and clinical analysis.

In this final chapter, the limitations of the LGF model, general points of discussion relating to cytogenetic research, and future research potential will be considered. The intent is to contextualize the LGF model within the broader setting of medical research.

6.2 Limitations of Research

There are limitations in any research study. One limitation of the LGF model is that the parsing and mapping system used to process karyotypes relies on a rule based mapper. This means that we can only process karyotypes where the mapper function has a rule for that cytogenetic aberration. Thus researchers
must have seen an aberration previously and recorded a rule to classify it for it to be incorporated into the LGF model. Observing which aberrations fail and then writing rules to account for them will lessen this limitation and strengthen the model.

Another limitation of the LGF model is that it is based on ISCN karyotypes. Although karyotypes provide an important source of clinical medical information, more modern technologies such as array-based sequencing have been developed that are more powerful and provide a more granular and detailed understanding of the patient’s genome. This limitation is evident in the fact that, through our LGF model, only 13.9% of the population was identified with a karyotypic loss of 13p14 while 52% of that same population actually had this deletion when observed using FISH probes. This illustrates the limitation of the data that the model is working with.

6.3 General Points of Discussion

In this thesis, many of the core elements of TBI-oriented research have been utilized and expanded. Previously siloed karyotype data can now be freely utilized in a computational manner. This enables a wide array of novel research, none of which would be possible if not for the de-siloing of the karyotype data. Through the LGF model, karyotype data can be more easily analyzed, shared, and stored, all of which are core tenets of TBI-oriented research (Payne et al., 2009).
LGF-represented karyotype data can be more easily analyzed because it is in a shared binary data model that focuses on biological results, not cytogenetic events. These data are also easier to store and share because large-scale datasets can be represented as binary LGF vectors, allowing for easier data management and searching capabilities. This allows researchers to more easily utilize their own data for their own needs, thus enabling biomedical research.

In addition, the significance of this research relates to the ability of the LGF model to enable large-scale karyotype research and thus provide a mechanism to clinically study cancer genomics. The research described in this thesis demonstrates that by transforming the currently under-used karyotypic data into a machine-readable format, researchers can perform large-scale analyses of karyotypic data. Such analytic methods can help clinicians gain a more complete understanding of the connection between a CLL patient’s karyotype, their clinical phenotype, and their responses to therapies, thus representing a significant advance in precision medicine. As such, the research and development described in this thesis is novel and innovative for it enables the representation of karyotypic data and its linkage with clinical data for integrative analyses at a systems level that has not been previously possible.

As a further point of discussion, this thesis research highlights the value of hypothesis generation leading to novel discoveries made from retrospective studies on karyotype data. This leads to new analyses of karyotype data that
may yield new discoveries and further our understanding of diseases, specifically for hematological malignancies. This currently used genomic clinical data has been under-used by the informatics community up until this point and, although other researchers have had marginal success parsing karyotype data, no significant clinical breakthroughs occurred from these endeavors. The research highlighted in this thesis presents a workflow that enables new clinically relevant discoveries to be made by observing karyotype data. This can provide a greater understanding of cancer patient genetics by leveraging the largest amount of genomic data that is commonly clinically collected. This will in turn help foster new research, such as helping to identify cytogenetic locations for new FISH probes. This methodology helps facilitate future novel research to be performed as well as provides novel findings into the cytogenetics of CLL.

6.4 Future Research

In the future, we will use the LGF parser and mapper to assess other leukemia and lymphoma patient populations to attempt to find novel cytogenetic abnormalities present in these populations. Our first target will be AML (acute myeloid leukemia) since karyotype information is often collected on these patients. Additionally, a karyotype database that is a searchable portal to help enable clinical research can also be designed. Such a system, based on Couch DB, can enable karyotypes to be stored in both the ISCN text form and the LGF binary vector form. This would allow researchers and clinicians to more easily
search for patients with specific abnormalities and have that search performed on the binary vector and then return any information the researchers wanted including the ISCN karyotype and/or the binary vector along with any clinical information. This would massively streamline any clinical research looking to see the clinical effects on different patient populations based on karyotypic aberrations.

Such research is commonly performed now but is difficult due to the inherent obstacles of working with large-scale text based ISCN karyotype data. Given that most academic medical institutions have a large quantity of karyotype data, this would enable them to utilize those data for better secondary clinical research as well as allow clinicians the ability to more easily sort through and utilize their own data.

Another future application may be the use of pattern recognition techniques to find means of patient classification. Although it is known that different diseases have observable chromosomal aberration patterns (Jaffe et al., 2001), there may be patterns that cannot be observed using current techniques. The LGF model represents the karyotype data in a format that allows advanced computational pattern recognition analysis. For example, this format makes possible the discovery of karyotype-based biomarkers that may be relevant to clinical decisions and could further our understanding of the biological mechanisms behind diseases and drug development, as discussed in Chapter 4.
In the future, semi-automated predictive disease classification based on karyotypes can be clinically performed. The preliminary results of this work have been presented in Chapter 3. It showed that it is possible to non-randomly distinguish very similar disease states from each other purely based on karyotype data. In the future, we hope to integrate other forms of non-genomic data into such a predictive classifier. Other information could be added to the model to increase classification performance.

The combination of personalized genomic and phenotypic data to be used in a classification system could have broadly-based applications in clinical care, such as a pseudo-decision-support system to aid researchers and clinicians in distinguishing between some similar conditions as well as providing researchers with a data driven mechanism for assessing the genetic and phenotypic differences between conditions.

Drug repurposing represents another future application of this work. As presented in Chapter 5, as a consequence of mining the Mitelman database, a large number of drug-disease-gene triplets have been identified that contain promising drug repurposing hypotheses. Mining this dataset to obtain the higher value hypotheses is the next step of this future research goal since the current work has already identified associations through large-scale literature searches and genotypic and pharmacology data. By adding to other drug repurposing systems such as C-map and LINCS, researchers can identify which triplets in our
dataset are of interest in their research and then perform in vitro and in vivo experiments validation tests to assess the efficacy of the drug for that condition.

Perhaps most importantly, longitudinal analyses of how karyotypes evolve over time may lead to improved understanding of disease progression and remission. Ultimately, these types of personal longitudinal analyses will support the advancement of precision medicine in the cancer domain (Peer, 2014; Rozovski et al., 2014; Shrager and Tenenbaum, 2014) by improving our understanding of the bio-molecular bases for oncogenesis, disease progression, response to treatment, and cancer evolution.

Finally, in the future we plan to build and distribute a production-quality software package to allow other researchers and institutions the ability to transform their karyotype data into the LGF binary vector. This is critical since it enables research discoveries to be made by other groups, thus accelerating the rate of discovery and research.

6.5 Conclusion

Janet Rowley (1973) utilized karyotype data in medical research to help identify one of the first targeted cancer therapy. Today we are still utilizing karyotype data as an important form of clinical genomic data. Currently these data are stored in a researcher-unfriendly format that makes it almost impossible to perform large-scale studies to further our understanding of the genomics of cancer. Through the work discussed in this thesis, a mechanism to parse and
map karyotype data that allows researchers the ability to more easily study karyotype data has been presented. This parsing and mapping system will help enable future cancer research by affording researchers the ability to easily utilize and study karyotype data on a large scale. In this way we hope to help enable current and future researchers and help them make life-changing breakthrough discoveries such as those made by Janet Rowley.
REFERENCES


Nishii, K., et al. "Characteristics of t (8; 21) acute myeloid leukemia (AML) with additional chromosomal abnormality: concomitant trisomy 4 may constitute a distinctive subtype of t (8; 21) AML." *Leukemia* **17.4** (2003): 731-737.


https://www.coriell.org/research-services/cytogenetics/karyotyping
