Machine Learning Reveals Pan-Cancer Biomarker

Jesse Michel

Mentored by Andrew Matteson

1 Personal

1.1 Research Background

In my junior year, I came to a new high school (Massachusetts Academy of Mathematics and Science) that required that I complete a science fair project. I knew that I wanted to pursue a math-heavy topic because I have enjoyed math my whole life. My preliminary ideas included combining math with physics in order to optimize a paper airplane design and perhaps study different folding techniques as well as researching how one might harness the surface tension of water as a possible mode of transport for heavy objects. However, in the end I decided to pursue the most creative option by choosing a topic in pure mathematics.

After reading an article by Dr. James Tanton that alluded to the concept of non-integer bases, I decided to research the construction and properties of Base 1.5 and its relation to a classic problem in number theory called the Collatz Conjecture. I did not have an advisor, but despite that, I came upon some interesting and beautiful math. At the state science fair, I met one of the judges (not mine), Andrew Matteson. He offered to advise me in a future project, and I gladly accepted.

After several science fairs, I decided that in my senior year I wanted to pursue and present research in a new field of study that was still mathematically based. I initially considered physics and theoretical computer science, but came upon bioinformatics and was fascinated.

1.2 What is Bioinformatics?

Bioinformatics is a field that draws from mathematics, computer science, and engineering to develop biological understanding [26]. Bioinformatics uses many techniques and analyses to identify the biological mechanisms that underlie biological data. Bioinformatic analysis begins with data such as sequences of DNA, structural information about a protein, or measures of gene expression. Much of this data is available online in publicly accessible repositories. Using these repositories, researchers can apply various machine learning techniques to high-quality data without incurring the cost of generating the data themselves.

Jesse Michel
jmmichel@mit.edu
Please contact me for complete results.

1.3 Interest in Bioinformatics

Bioinformatics intrigues me because I respect the potential of biology to help people directly and I appreciate the mathematical aspects of data analysis. Bioinformatics also appeals to me greatly because much of the research uses freely available tools and resources. This allows anyone with an interest and a little brain power to conduct research. I appreciate the ideology of this type of bioinformatics research: a community of researchers working together to gain insights into freely available data sets.

1.4 Learning About Bioinformatics

Obtaining an understanding and overview of biology and machine learning was a critical step before analyzing specific data sets and looking for biologically important features. I obtained background in bioinformatics from a massively open online course on Machine Learning from Stanford University. Topics included linear regression, gradient descent, neural networks, system design, support vector machines, and unsupervised learning algorithms (Ng, 2014). Programming exercises and quizzes were completed to reinforce the material and provide understanding of algorithms, although libraries are more practical for most uses. I used MATLAB to implement the algorithms in the course. Separately, I also received mentoring in molecular cell biology to complement the machine learning. To obtain an understanding of breast cancer, I studied topics such as histone modification, transcriptional and post-transcriptional control of gene expression, and signaling pathways that control gene activity (Lodish, 2014).

1.5 Location of Research

I conducted most of my research at my home on my laptop. I did not work with any lab or research program. I met with Dr. Matteson, my advisor, on a bimonthly basis.

1.6 Some Advice

I think that research is mostly a personal challenge, the brain against the problem, and as a result it is difficult to give meaningful advice. Despite this I will suggest having as much "structured-unstructured" research time as possible. What I mean by this is to spend some time pursuing your research (that's the structured aspect) and to make sure that during that time you are free to explore beyond your specific area of research. I would strongly suggest not having a time limit on your work. Just start browsing and enjoy!

I often learned from searching an unfamiliar term on Wikipedia and then searching all of the terms I did not recognize on that page. This unstructured exploration helped me to gain a perspective on bioinformatics as a field.

2 The Research

2.1 Background

2.1.1 Overview

Machine Learning is a type of artificial intelligence that allows computers to recognize patterns and to generate predictions on data sets without being explicitly programmed. Attractor metagene learning is a machine learning algorithm specifically designed for large gene expression data sets. Gene expression data sets provide a measurement that indicates a future phenotype or observable characteristic. In this work, gene expression data are used in conjunction with pharmacological profiling data (information about how well drugs work) to provide insight into which groups of people will potentially respond best to a specific chemotherapy drug. We discovered a biomarker or measurable trait that allows the prediction of biological behavior that provides insight into what drug should be selected to treat a particular patient with a given type of cancer.

2.1.2 Core Concepts

Biomarkers are measurable traits that allow predictions of biological behavior. Disease-related prognostic and diagnostic biomarkers provide information about likely patient outcomes. Prognostic biomarkers predict the likely progression of a disease. For example, a prognostic biomarker can predict how long a patient will survive or how long before cancer progresses to a more advanced state. Diagnostic biomarkers indicate the probable effect of a particular treatment of a disease with a specific drug [2]. Diagnostic biomarkers offer information that helps decide between two otherwise comparable treatments. This information can help make them more valuable than prognostic biomarkers in cancer treatment. Treating only a specific subgroup of a population, identified using a biomarker, can drastically improve the prognosis of a patient and ensure that the benefits of a drug are sufficient to justify exposing a patient to the side-effects of a drug.

Bioinformatic approaches have enabled many advances in understanding cancer and choosing the best course of treatment. Cancerous cells have molecular features that mark them as different from normal tissue. Common characteristics of cancer include dividing uncontrollably and invading surrounding tissues [6]. Cancerous cells proliferate and metastasize through dysregulation of signaling pathways, such as constitutive activation of the Ras/MEK pathway that drives many of the characteristics of cancer [18]. The goal of drug research is to identify cellular components, targets, that are affected by chemicals that can be introduced to improve the outcome of a disease [4]. Drug research in cancer often entails finding potential targets and therapeutic interventions, but identifying the subset of patients that will benefit from treatment with a specific drug is equally important [26]. Machine learning techniques in bioinformatics can provide a greater understanding of diseases, providing tools to discover diagnostic biomarkers [6].

2.1.3 The Math Driving the Model

Metafeatures or attractor metagene learning is an iterative algorithm that finds groups of related features and averages across the related features. Metafeatures is the name of the MATLAB function often referred to as attractor metagene learning. Metafeatures uses mutual information as the measure of similarity. Metafeatures use of mutual information makes it particularly useful in biology because relationships between genes are captured more clearly by this metric than by linear metrics. Metafeatures is useful in biology because averaging over co-expressed genes better captures the biological phenomena within gene expression datasets. For the jth gene, in the ith iteration of the metafeatures function, we generate a weight for the $j+1$th iteration. The current metagene is M_i, when the correlation between M_i and G_j is less than zero then:

$$J(M_i, G_j) = 0$$

where G_j is the expression of the ith gene, and M_i is the metagene in the current iteration. For genes that have a positive correlation with the current correlation,

$$J(M_i, G_j) = I(M_i, G_j)^\alpha$$

where $\alpha > 0$ and is used to modify the weighting to vary the number of metagenes generated. The mutual information between M_i and G_j is $0 \leq I(M_i, G_j) \geq 1$. An estimate of the metagene is given by:

$$M_{i+1} = J(M_i, G_j)G_j$$

$J(M_i, G_j)$ is the "weight" of the gene used in an average. The weight for the jth gene in the ith iteration is referred to as w_{ji}. The vector of all weights at the ith iteration is W_i. For a given tolerance, the algorithm iterates until

$$\|W_i - W_{i-1}\| < tolerance$$

The resulting metagenes tend to be weighted averages of co-expressed genes. These averages are proxies for broader molecular features rather than measures of a single gene alone. The weighted average of gene expression is a robust measure of molecular phenotypes because many genes fill the same or similar purposes because of biological redundancy [6].

2.1.4 Previous Approaches

Biological data sets often have many features, making them complicated and challenging to analyze. In this work, the data sets considered have gene expression measurements for every human gene: about 20,000 features in total. High dimensionality causes several problems with data analysis, including visualization of the data. Common approaches are either to reduce the dimensionality of the data using algorithms like principal component analysis (PCA), or to identify groups in data using clustering algorithms such as k-means or attractor metagene learning.

PCA identifies linear combinations of features that explain the maximal variance of the data. It does this by using linear algebra techniques to decompose the data into a combination of matrix multiplications. This decomposition allows us to identify a lower rank matrix that minimizes the sum of squared distances between the true data and the low rank approximation [22].

K-means is an iterative algorithm that partitions data into a given number of disjoint subsets. Initially, k-means randomly selects cluster centroids. The algorithm iterates over two steps: a cluster-label step that labels each of the points as belonging to the cluster with the nearest centroid, and a centroid update step that calculates the best centroid for a group given the data in that group. K-means is not guaranteed to find the globally best division of data into groups, but multiple random initializations often produce a "good-enough" grouping of data [21].

PCA and k-means are both unsupervised algorithms (meaning that they search for patterns in unlabeled data) that facilitate the visualization of complex data. Feature reduction is an important element of biological data mining. Attractor metagene learning is also an unsupervised learning algorithm, but is more relevant to this problem because it more accurately models the underlying biological processes [5].

2.1.5 The Data

Several sources of rich gene expression data sets are available online [22]. These data sets variously quantify relative gene expression for many patients, samples, cancer types, or experimental conditions, depending on the specific goals underlying the generation of the data. Correlations between gene expression and mortality rate can be used to identify likely targets for therapeutic intervention [23].

2.1.6 Data from the Cancer Cell Line Encyclopedia (CCLE) Analyzed with Metafeatures

The attractor metagene algorithm developed by Cheng et al. models trends in gene expression data [5][6]. This algorithm was created specifically for use with biological data. The CCLE dataset was also deliberately made available for predictive modeling of cell sensitivity to specific anticancer drugs [12]. Attractor metagene learning has not previously been applied to the CCLE data. This application provides an opportunity for connecting molecular phenotypes with outcomes of pharmacological intervention for many important drugs. All code used to generate the results presented was written in MATLAB.

2.2 Illustrations

Below is the graph used to identify the lymphatic-character biomarker.

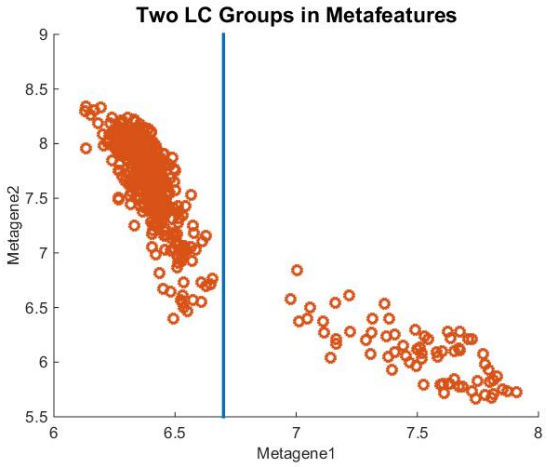

Figure 1: Two Lymphatic-Character Groups. Metagenes 1 and 2 were generated by a robust version of metafeatures, which iterates through n genes with weights that are all initialized to zero except the ith weight, which is set to one. The value of α that generated the desired number of metagenes for $n = 5000$ was 1.2. The first and second metagenes were graphed against each other, displaying two groups. A value of 6.7 for the first metagene was chosen to be the threshold because it effectively separated the two groups. This is represented by the vertical blue line on the graph.

2.2.1 Identification of the Lymphatic-Character Metagene

Cheng et al. reported lymphatic-character as a prognostic biomarker, however, we found a similar signature that was a diagnostic biomarker [5]. Cheng et al. demonstrated the significance of the lymphatic character as an important prognostic biomarker [5] [6]. Of the genes listed as the hundred most heavily weighted genes that comprise the metagene, many of them were common to the metagene discovered from our analysis of the CCLE data set using the metafeatures algorithm. As a result, we identified the metagene as the lymphatic-character biomarker. Cheng et al. demonstrated that a high-lymphatic-character biomarker is prognostically protective on a data set of patients that were treated with chemotherapy drugs [5] [6]. The diagnostic result that the tumors with high lymphatic character are often more responsive to drugs explains why patients with high-lymphatic-character tumors have a better prognosis: drugs are more effective for that group. Our result more closely illustrates the mechanisms that result in a better prognosis for the patients with high-lymphatic-character tumors. Furthermore, the gene signature for the high-lymphatic-character metagene provides novel diagnostic information, which should improve prognosis accordingly.

By applying attractor metagene learning to the CCLE data, we found several metagenes. The most prominent metagene found had CD53, IKZF1, LOC100506779, and ARHGAP30 up-regulated. This is closely related to the metagene found in Cheng et al. that was associated with the expression of lymphocyte differentiation factors [6]. These two groups happen to have a dramatic difference in their response to the 24 drugs provided in the CCLE in-vitro pharmacological profiles. Generally, the high-lymphatic-character cells are more responsive to drugs [5].

2.3 Implications of The Research

2.3.1 An Overview of Results

Dovitinib, topotecan, and L-685458 are only three of many drugs that show clear separation between the low- and high- lymphatic-character groups for our biomarker. The insight

provided by the graphs of dovitinib helps explain why the recent study of dovitinib versus sorafinib failed and how it could be redesigned in the future. The biomarker separates the graphs of topotecan into two groups, one of which is far more responsive to treatment than the other. This sizable subset would reap more benefits from treatment, while the remaining group may have improved quality-of-life-adjusted survival by avoiding the toxicities associated with treatment using topotecan. L-685458 could be extremely effective at treating this same subset of patients. These three drugs with profound separation between the low- and high-lymphatic-character groups demonstrate that this biomarker provides meaningful information for diagnosing patients.

2.3.2 Applications of Biomarker to Clinical Decisions

The 24 different chemotherapy drugs examined showed different levels of separation of response for the high- and low- lymphatic-character groups. Six of the drugs showed no separation in response between the two groups, while eight showed some separation and nine showed significant separation. This outcome supports the salience of the results for multiple reasons. If the results had only shown little separation, then the biomarker would provide little useful information. If all the results had shown significant separation, then this diagnostic result may have only prognostic meaning. It would imply that the group identified by the biomarker had a less severe strain of cancer and would be more likely to survive irrespective of treatment. Due to the varied levels of response for different drugs, the lymphatic-character biomarker has diagnostic significance. Furthermore, this biomarker has significance in deciding which drugs to use or combine when there are multiple drugs available for a given set of cancer patients, as is often the case with topotecan and lapatinib. Results like this give clinical insight and have the potential to improve cancer treatment.

2.3.3 Lymphatic Character Biomarker as a Diagnostic Pan-Cancer Pan-Drug Biomarker

The variation in separation between the low- and high- lymphatic-character groups for the 24 chemotherapy drugs and the exemplary results like those of dovitinib, topotecan, and L-685458 demonstrate that this biomarker is meaningful across many drugs with different mechanisms of action ranging from γ-secretase inhibitors to tyrosine kinase inhibitors. The CCLE data set provides both gene expression and in-vitro pharmacological profiles for a variety of cancers, implying that biomarkers discovered using these data may provide meaning for numerous cancers. The lymphatic-character biomarker has the potential for broad application across a wide variety of cancers and drugs.

2.3.4 Future Research

More expansive research for the pan-cancer pan-drug lymphatic-character biomarker is imperative for improving the treatment of patients with a variety of drugs for a variety of cancers. The implications of only three of many drugs with significant separation between high-

and low- lymphatic-character groups are discussed in this paper. There are still multiple drugs whose responses should be explored using both algorithmic techniques and pre-clinical and clinical methods. Methods such as bootstrapping should be used to define more concrete measures of accuracy for the results presented. Furthermore, the properties that make some drugs similarly effective for both groups and others significantly more effective for the high-lymphatic-character groups should be explored. In conclusion, we discovered a meaningful biomarker with immediate diagnostic implications for numerous chemotherapy drugs.

References

[1] Browse Data. (n.d.). Broad-Novartis Cancer Cell Line Encyclopedia. Retrieved August 30, 2014, from http://www.broadinstitute.org/

[2] Carsten O Daub, Ralf Steuer, Joachim Selbig and Sebastian Kloska, Estimating mutual information using B-spline functions - an improved similarity measure for analysing gene expression data. BMC Bioinformatics 2004, 5:118.

[3] Chen, Y., & Chen, F. (2008). Identifying targets for drug discovery using bioinformatics. Expert Opinion on Therapeutic Targets, 12(4), 383-389.

[4] Cheng, Wei-Yi, and D Anastassiou. "Biomolecular events in cancer revealed by attractor metagenes." Center for Computational Biology and Bioinformatics and Department of Electrical Engineering 1 (2012): 1-22. Print.

[5] Cheng Wei-Yei, Ou Yang T-H, Anastassiou D. (2013a) Biomolecular Events in Cancer Revealed by Attractor Metagenes. PLoS Comput Biol 9(2): e1002920. doi:10.1371/journal.pcbi.1002920

[6] Cheng, W., Yang, T. O., & Anastassiou, D. (2013b). Development of a Prognostic Model for Breast Cancer Survival in an Open Challenge Environment. Science Translational Medicine, 5(181), 181ra50-181ra50.

[7] "Comprehensive Cancer Information." National Cancer Institute. National Institutes of Health, n.d. Web. 22 Oct. 2014. http://www.cancer.gov/

[8] Data Matrix. (n.d.). Data Portal. Retrieved August 23, 2014, from https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&diseaseType=BRCA

[9] Daub, C., Steuer, R., Selbig, J., & Kloska, S. (2004). Estimating mutual information using B-spline functions an improved similarity measure for analysing gene expression data. BMC Bioinformatics, 5(118), 1471-2105.

[10] Documentation Center. (n.d.). MATLAB Documentation. Retrieved September 29, 2014, from http://www.mathworks.com/help/matlab/?refresh=true

[11] Erickson-Miller, C. L., May, R. D., Tomaszewski, J., Osborn, B., Murphy, M. J., Page, J. G., et al. (1997). Differential toxicity of camptothecin, topotecan and 9-aminocamptothecin to human, canine, and murine myeloid progenitors (CFU-GM) in vitro. Cancer Chemotherapy and Pharmacology, 39(5), 467-472.

[12] Garraway, L. A., Mesirov, J. P., Gupta, S., Palescandolo, E., Hatton, C., Wang, L., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483(7391), 603-307.

[13] Geyer, P., Mehuyas, M., Perry, R., & Johnson, L. (1982). Regulation of ribosomal protein mRNA content and translation in growth-stimulated mouse fibroblasts. Molecular and Cellular Biology, 2(6), 685-693.

[14] Gillis, B. "Dovitinib Fails as Third-Line Option in Kidney Cancer." OneLive Bringing the Oncology Community Together. OneLive, 3 Oct. 2013. Web. 24 Oct. 2014. http://www.onclive.com/ conference-coverage/ecco-esmo-2013/Dovitinib-Fails-as-Third-Line-Option-in-Kidney-Cancer

[15] "Lapatinib Oral and Topotecan IV Interactions." (2014). Lapatinib Oral and Topotecan IV Drug Interactions. RxList-The Internet Drug Index. Web. 6 Nov. 2014. http://www.rxlist.com/drug-interactions/lapatinib-oral-and-topotecan-iv-interaction.htm

[16] Lheureux, S., S. Krieger, B. Weber, P. Pautier, M. Fabbro, F. Selle, H. Bourgeois, et al. "Expected Benefits of Topotecan Combined with Lapatinib in Recurrent Ovarian Cancer According to Biological Profile: A Phase 2 Trial." International Journal of Gynecological Cancer 22.9 (2012): 1483-8.

[17] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. Pattern Recognition, 36(2), 451-461.

[18] Lodish, H. F. (2008). Molecular Cell Biology (6th ed.). New York: W.H. Freeman.

[19] Mecocci, P., Polidori, M., Cherubini, A., Ingegni, T., Mattioli, P., Catani, M., Beal, M. (2002a). Lymphocyte Oxidative DNA Damage and Plasma Antioxidants in Alzheimer Disease. Archives of Neurology, 59(5), 794-798. Retrieved November 2, 2014, from http://www.ncbi.nlm.nih.gov/pubmed/12020262

[20] Mecocci, P., Polidori, M., Ingegni, T., Catani, M., Mattioli, P., & Cecchetti, R. (2002b). Lymphocyte Oxidative DNA Damage and Plasma Antioxidants in Alzheimer Disease. Archives of Neurology, 13(6), 794-798. Retrieved November 3, 2014, from http://www.ncbi.nlm.nih.gov/pubmed/17944672

[21] Ng, A. (Director) (2014, August 25). Machine Learning. Lecture conducted from Stanford, Stanford.

[22] Ringnar, M. (2008). What is principal component analysis? Nature Biotechnology, 26(3), 303-304. Retrieved August 5, 2014, from http://dx.doi.org/10.1038/nbt0308-303

[23] Salmans, M., Zhao, F., & Anderson, B. (2013). The estrogen-regulated anterior gradient 2 (AGR2) protein in breast cancer: a potential drug target and biomarker. Breast Cancer Research, 15(2), 204-215.

[24] The Human Gene Compendium. (n.d.). GeneCards. Retrieved September 15, 2014, from http://www.genecards.org/

[25] Yao, L., Duan L., Fan M., Wu X (2007).Gamma-secretase inhibitors exerts antitumor activity via down-regulation of Notch and Nuclear factor kappa B in human tongue carcinoma cells. (2007). Oral Diseases, 13(6), 555-63. Retrieved November 2, 2014, from http://www.ncbi.nlm.nih.gov/pubmed/17944672

[26] Wang K, Shrestha R, Wyatt AW, Reddy A, Lehar J, et al. (2014) A Meta-Analysis Approach for Characterizing Pan-Cancer Mechanisms of Drug Sensitivity in Cell Lines. PLoS ONE 9(7): e103050. doi:10.1371/journal.pone.0103050

[27] Zamboni, W., D'Argenio, D., Stewart, C., Farese, A.,MacVittie T., Delauter B. et al. (2001). Pharmacodynamic Course of Neutropenia. Model of Topotecan-induced Time Course of Neutropenia. Clinical Cancer Research, 7, 2301-2308.