Predicting Cancer Susceptibility from Single-Nucleotide Polymorphism Data: A Case Study in Multiple Myeloma

Michael Waddell
University of Wisconsin Department of Computer Sciences Madison, Wisconsin, 53706
mwaddell@biostat.wisc.edu

David Page
University of Wisconsin Department of Biostatistics and Medical Informatics Department of Computer Sciences Madison, Wisconsin, 53706
page@biostat.wisc.edu

Fenghuang Zhan
University of Arkansas for Medical Sciences Donna D. and Donald M. Lambert Laboratory of Myeloma Genetics Little Rock, Arkansas 72205
zhanfenghuang@uams.edu

Bart Barlogie
University of Arkansas for Medical Sciences Myeloma Institute for Research and Therapy Little Rock, Arkansas 72205
barlogiebart@uams.edu

John Shaughnessy, Jr.
University of Arkansas for Medical Sciences Donna D. and Donald M. Lambert Laboratory of Myeloma Genetics Little Rock, Arkansas 72205
shaughnessyjohn@uams.edu

ABSTRACT

This paper asks whether susceptibility to early-onset (diagnosis before age 40) of a particularly deadly form of cancer, Multiple Myeloma, can be predicted from single-nucleotide polymorphism (SNP) profiles with an accuracy greater than chance. Specifically, given SNP profiles for 80 Multiple Myeloma patients – of which we believe 40 to have high susceptibility and 40 to have lower susceptibility – we train a support vector machine (SVM) to predict age at diagnosis. We chose SVMs for this task because they are well suited to deal with interactions among features and redundant features. The accuracy of the trained SVM estimated by leave-one-out cross-validation is 71%, significantly greater than random guessing. This result is particularly encouraging since only 3000 SNPs were used in profiling, whereas several million SNPs are known.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Miscellaneous

BIOKDD '05, August 2005, Chicago, Illinois, USA

Keywords

supervised machine learning, support vector machines, single-nucleotide polymorphism, multiple myeloma

1. INTRODUCTION

A significant contribution to the genetic variation among individuals is the cumulative effect of a number of discrete, single-base changes in the human genome that are relatively easy to detect. These single positions of variation in DNA are called single nucleotide polymorphisms, or SNPs. While it is presently infeasible to obtain the sequence of all the DNA of a patient, it is feasible to quickly measure that patient's SNP pattern – the particular DNA bases present at a large number of these SNP positions [15].

Our case study employs support vector machines (SVMs) to analyze this new and promising form of genetic data. The authors present lessons for machine learning throughout the paper. Some biological terminology is necessarily used. Critical terms are defined for general machine learning (ML) readers; undefined terms are not critical to understand the ML lessons, but are used as needed to clarify issues for computational biology readers.

One promise of SNP data is that this data may make it possible to identify markers for genetic predisposition to disease. In addition to providing patients with information about their risk for disease, such markers may give researchers insight into the genes involved in a disease process and hence into proteins that may serve as targets for novel pharmaceutical therapies. In order to find such markers, the traditional approaches are to use linkage analysis and

association studies [17].

Linkage analysis requires obtaining data on families with known pedigrees and disease histories. This requirement can make accurate linkage analysis difficult since many family members – including previous generations – are unavailable for genetic testing. Also, since the results of linkage analysis studies often come from a small number of families, they may not be generalizable to the rest of the population. Association studies do not require known family pedigrees. However, they do require a number of "candidate genes" that are suspected to be important in the disease process of interest. Thus, this method relies on the quality of the candidate genes, which are chosen based upon prior knowledge about the disease.

Both of these traditional approaches have been very successful when dealing with simple Mendelian or near-Mendelian disorders, but fail when attempting to identify disorders controlled by quantitative trait loci (QTL) [17]. QTL are genes, each of modest effect, whose combined effects cause a particular complex, continuous trait [5]. To deal with the complexities that QTL bring to this task, we will use an ML algorithm that is well suited to tasks involving interactions and redundant features.

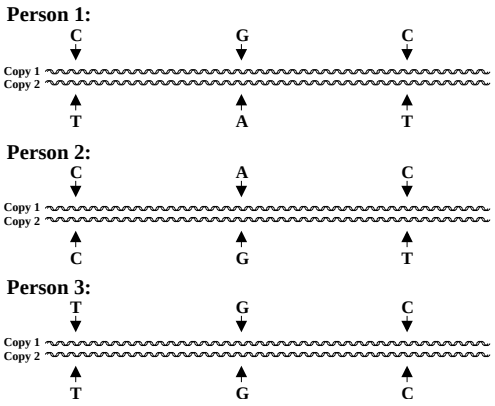
First, we will divide the data points into two classes. Next, we will use an ML or statistical modeling algorithm to construct a classifier, or model, based upon all of the SNP data that were collected. The accuracy of the model at predicting the class (e.g., susceptible vs. not susceptible) will then be estimated using cross-validation. If the accuracy of the model is significantly better than chance, one may then study this model to gain insight into the disease. We have chosen not to employ candidate genes, like in an association study, because little is known about the genetics of Myeloma and its epidemiology. The hypothesis is that if there is an association between Myeloma and a particular gene, then a SNP in the haplotype block [4] containing that gene will be discovered in the present study. Given the general lack of knowledge about the etiology of this disease, we believe that using a candidate gene approach would put unreasonable bias on the analysis and, in the end, may fail and eventually cost more than doing a global search for associations.

This same general methodology has been employed in numerous cancer studies using microarray data [1, 6, 16, 18, 23]. A major advantage of using SNP data over microarray data to study genetic predisposition is that, unlike microarray data, a person's SNP pattern is unlikely to change over time. Loosely stated, the SNP pattern collected from a person with a disease is likely to be the same pattern that would have been collected from that person at birth or early in life. Thus, we can use SNP data from patients at any stage of their life and at any stage of their disease progression.

Single-nucleotide polymorphisms are extremely stable over evolutionary time [11]. Furthermore, relative to microsatellite polymorphisms, which are susceptible to mutations during the aging process [20], SNPs are much more stable and hence are unlikely to change over the lifetime of an individual [3]. The DNA used to perform our study is derived from peripheral blood mononuclear cells, which should be a mixture of cells whose germline DNA has no over-representation of any given clone containing any specific mutation. Thus, it is highly unlikely that the SNPs discovered in this study to be associated with the age of onset of Multiple Myeloma would be related to a SNP that tends to be mutated as a person ages. As a result of these arguments, SNP data has the potential to provide more insight into genetic predisposition to Multiple Myeloma, as well as many other diseases, than does microarray data.

A second major advantage of using SNP data is that the data can be collected from any tissue in the body. With microarray data, the mRNA samples for cancer patients are taken from tumor tissue (e.g., from the colon), and the mRNA samples for healthy donors are taken from healthy tissue of the same type (e.g., colon again). SNP data, on the other hand, is not taken directly from tumor samples, but from any tissue in the body. The benefit of this is that, in addition to being faster to obtain, SNP data is also easier to obtain since less invasive procedures can be used. On the other hand, when using SNP data, we do not expect to have predictors of as high accuracy as we get with microarray data. This is because microarray data is taken directly from the tumor tissue. Since gene expression is greatly altered in cancer, it is possible to obtain highly-accurate predictive models for cancer vs. normal. While such models may provide insight into the disease itself, they do not provide information on genetic predisposition. When working with SNP data, we expect to gain more information about a person's genetic predisposition to a disease than we would gain from microarray data; however, we do not expect to have predictors of as high accuracy as we get with microarray data.

Despite these advantages, SNP data does present three major challenges for our approach. The first challenge of SNP data is that there are now well over 1.8 million SNPs known [22], but measuring them all is typically cost-prohibitive. Hence, in contrast to microarray data where measurements are recorded for a substantial fraction of the known genes, SNP data contains measurements for only a small fraction of the known SNPs – typically a few thousand. Therefore, it is quite possible that, for a given classification task, the features that would allow for highly accurate prediction will be missing. Second, missing values are more common in SNP data than in microarray data. This must be taken into consideration when choosing a learning algorithm, since some methods are more capable of handling missing data than others. Third, and perhaps most interesting, SNP data is "unphased." Figure 1 illustrates this issue. The result of SNP data being unphased is that this additional, and potentially highly informative, phase information is not available for model building. Algorithms for haplotyping, or determining this phasing information, exist, but their solutions are not guaranteed to be correct. Also, these algorithms typically require additional data on related individuals and a large number of individuals relative to the number of SNPs [12]. Thus, one may approach this phasing problem either by estimating the phase information and accepting the consequences of incorrect estimates, or by working with the data in its unphased form. Because of the inaccuracies inherent in haplotyping and lack of additional data, we have elected to work with the data in its unphased form. We believe that this decision will not adversely affect our modeling algorithm since our research uses a relatively sparse coverage of the genome. Thus, adjacent SNPs are not linked strongly enough for phasing information to be informative. In future studies with a denser SNP coverage, this information would be potentially more useful.



(a) The true phased SNP patterns for persons 1, 2 and 3.

	SNP 1		SNP 2		SNP 3		Class
Person 1	C	T	A	G	C	T	Diseased
Person 2	C	C	A	G	C	T	Healthy
Person 3	T	T	G	G	C	C	Diseased

(b) The unphased SNP data for persons 1, 2 and 3.

Figure 1: In a SNP data file (b), each example, or data point, corresponds to a single person. The features, or variables, used to describe the person are the SNPs. A SNP position on one copy of a chromosome typically can take one of two values; for example, SNP 1 can be either C or T. But because every person has two copies of chromosomes 1 through 22, most SNP features can take one of three values. For example, the feature labeled SNP 1 can be either heterozygous CT as for Person 1, homozygous CC as for Person 2, or homozygous TT as for Person 3. If both SNP 2 and SNP 3 are on the same chromosome, then they can be arranged either as for Person 1 or for Person 2. Although these 2 arrangements are distinct, they lead to the same SNP pattern. The process of determining which of these two cases holds is called *phasing* or *haplotyping*. Data for which the haplotypes are not known is said to be *unphased*.

Phasing, or haplotypes, are potentially informative because within a haplotype block there is very little, if any, meiotic recombination. Thus, the linkage of SNPs within a given haplotype block will remain unchanged over time. Once the haplotype map is established, it will be feasible to use a single SNP to define a haplotype block just as well as if one used all the SNPs within that block. It is estimated that there are approximately 600,000 haplotype blocks (there are currently some 300,000 defined) representing the millions of SNPs in the human genome [21]. These haplotype blocks may eventually be used to define the entire human genotype. When this occurs, haplotypes (defined by a single SNP) that are found to be linked to a disease could be searched for candidate genes and mutations within candidate genes. This will eliminate the guesswork that is inherent in the current candidate-based approaches which rely on an investigator's best guess or hunch.

This paper discusses the application of SVMs to SNP data in order to study genetic predisposition to Multiple Myeloma. Multiple Myeloma is a cancer of antibody secreting plasma cells that grow and expand in the bone marrow. Although Multiple Myeloma is hypoproliferative (the cancer cells replicate at a relatively low rate), the disease is incurable and usually progresses rapidly after diagnosis – with bone demineralization, renal failure, anemia, and secondary infections resulting from immunosuppression as common causes of mortality [19].

Multiple Myeloma occurs with relatively high frequency in adults over 70 (0.035% of the US population aged 70+) compared with younger adults (0.002% of the US population aged 30–54)[1]. We hypothesize that those who are diagnosed with Multiple Myeloma at a young age (under 40) have a genetic predisposition to the disease. If this is the case, then it may be possible to see differences in SNP patterns between Multiple Myeloma patients diagnosed before the age of 40 (predisposed) and those diagnosed after the age of 70 (not predisposed), and we can use these differences to gain insight into the disease. If this hypothesis is false, then it should not be possible to predict "predisposed" vs. "not predisposed" with accuracy significantly better than chance.

2. METHODOLOGY

Our data set[2] consists of unphased SNP data for 80 patients, based on 3000 SNPs, taking the form shown in Figure 1(b). The class values are "predisposed" and "not predisposed" as described at the end of Section 1. The 40 "predisposed" patients were diagnosed with Multiple Myeloma before age 40, while the 40 "not predisposed" patients were diagnosed after age 70. High molecular weight DNA was produced from peripheral blood lymphocytes from patients with Multiple Myeloma using conventional methods. DNA was subsequently sent to Orchid Biosciences^{TM}. SNP genotyping was performed using a proprietary SNP-IT^{TM} primer-extension technology. SNP-IT primer extension is a method of isolating the precise location of the site of a suspected SNP and utilizing the inherent accuracy of DNA polymerase to determine the allele type or the absence of that SNP. In order to conduct SNP-IT primer extension, a DNA primer (SNP-IT Primer) is hybridized to the sample DNA one base position short of the suspected SNP site. DNA polymerase is then added and it inserts the appropriate complementary terminating base at the suspected SNP location. Detection of the single base extension is accomplished by conventional methods. The result is a direct read-out method of detecting SNPs that creates a simple binary "bit" of genetic information. The SNPcode system couples SNP-IT genotyping technology with the Affymetrix GenFlex^{TM} platform to create a versatile, high-density SNP scoring system. In the assay, multiplex PCR is followed by solution phase SNP-IT primer extension. The SNP-IT products are then hybridized to the GenFlex chip – the sorting mechanism for the multiplexed reactions [14]. In the present study, 3000 SNPs were investigated on 80 patients. The SNPs were not selected based on

[1]Source: http://seer.cancer.gov

[2]The new SNP data set is available online from the authors at http://lambertlab.uams.edu/publicdata.htm.

prior knowledge of genetic disposition to Multiple Myeloma; rather, the SNPs were selected to give good overall coverage of the human genome. SNPs were chosen so that they would be evenly spaced at approximately every 1 megabase across the human genome. A denser coverage would be desirable but was cost-prohibitive.

We employed the approach of linear SVMs as our chosen modeling algorithm. We chose SVMs for this task because they are well suited to deal with interactions among features and redundant features. In particular, we used the algorithm SVM^{light} [9][3]. Because SVMs assume that all features are numerical, we needed to convert the discrete features from Figure 1(b) into continuous features. We will now present a brief review of SVM technology to help our readers understand the motivation behind our particular method of converting SNP features into numerical values.

In its simplest form, a support vector machine is an algorithm that attempts to find a linear separator between the data points of two classes, as Figure 2 illustrates. SVMs seek to maximize the margin, or the separation between the two classes, in order to improve the chance of making accurate predictions on future data. Maximizing the margin can be viewed as an optimization task solvable using linear or quadratic programming techniques. Of course, in practice there may be no good linear separator of the data. Support vector machines based on kernel functions can efficiently produce separators that are non-linear [2]. Nevertheless, the output of a linear SVM is easier to understand and glean insights from; effectively, features that get large coefficients in the function of the linear separator are more important than those that get small coefficients. In addition, linear SVMs have given better results than other kernel-based SVMs in several studies of microarray data, including our prior work with Multiple Myeloma. Therefore, for the present work we use linear SVMs. Experimenting with SNP data using other kernel functions is a direction for future work.

Each SNP feature in our data set takes one of three possible non-numerical values – either heterozygous or one of two homozygous settings (see Figure 1) – but SVMs require numerical features. Therefore we convert the three possible values for a SNP feature to the values -1, 0 and +1, where 0 represents heterozygous. We arbitrarily choose one homozygous case to set to -1 and the other to set to +1. As we see in Figure 3, when using this method with a linear SVM, it will be impossible to model the case where heterozygosity for a particular SNP is indicative of one class while homozygosity is indicative of the other, since it is not possible to separate 0 from both -1 and 1 with a single line. For example, it is not possible to say that either CC or TT is indicative of "predisposed" while CT is indicative of "not predisposed." Nevertheless, it is possible to distinguish having no copies of C from having at least one copy, or to distinguish having two copies of C from having zero or one copies (Figure 3).

Discriminating based upon the presence or absence of a single base appears to be more biologically relevant than discriminating solely based upon the presence or absence of homozygosity. In order for a heterozygous feature to not predispose cancer, whereas either of the two homozygous states do, the gene products of either allelic variant would be deleterious in sufficient quantities, but in the case of heterozygosity, neither would be present in sufficient quantities

[3]Publicly available at http://svmlight.joachims.org.

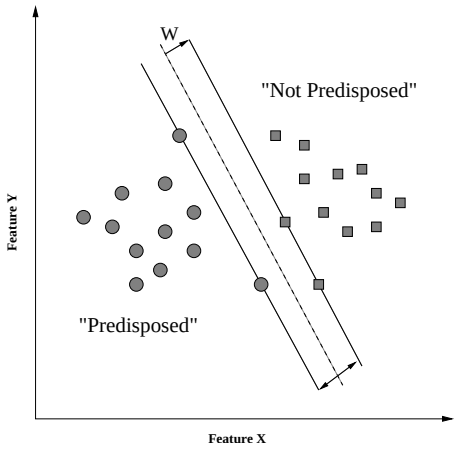


Figure 2: A support vector machine for differentiating between two classes by maximizing the margin, W. This is done in the N-dimensional space defined by N numerical-valued features. In this simple example, there are only two features, X and Y, so $N = 2$. Normally, however, N would be much greater. In a higher-dimensional space, the linear separator is a hyperplane rather than a line.

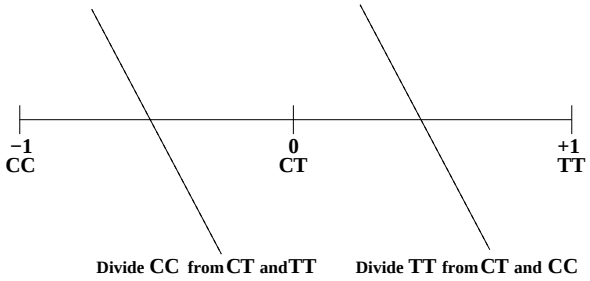


Figure 3: Divisions between feature values that are possible with the -1, 0, +1 encoding of SNP features. Notice that it is not possible to divide both CC (-1) and TT (+1) from CT (0) with a linear SVM.

to cause a negative effect. In this case, regardless of the relative abundance of the two variants, a very large percentage of the population would be homozygous for one allelic variant or the other. Thus this feature would not be very informative and would not be incorporated into our model. In order for a heterozygous feature to predispose cancer, whereas either of the two homozygous states of that feature do not, the gene products of both alleles would need to be present to cause a negative effect. If both allelic variants were common in the general population, then heterozygosity of this feature would be relatively common and would thus not be very informative. If one allelic variant is relatively rare, then a homozygote in this feature will be very rare indeed. If such a rare person were to be found in our non-predisposed group, they would not likely affect our model significantly. Thus, it is very unlikely that the presence or absence of homozygosity would play a significant role in determining predisposition to a specific cancer. This supports our decision to use the absence or presence of a particular

allele when building our model instead. This conclusion is further evidenced by the fact that most known mechanisms of inherited predisposition to cancers are dominant [10].

An alternative encoding that would permit all three possible distinctions between values would be to use two numerical features for each SNP. However, this leads to a doubling of the number of features, and the performance of ML algorithms tends to degrade as the number of features grows relative to the number of examples. Another option, using SVMs based on kernel functions, can efficiently produce separators that are non-linear [2]. Nevertheless, the output of a linear SVM is easier to understand and glean insights from; effectively, features that get large coefficients in the function of the linear separator are more important than those that get small coefficients. In addition, linear SVMs have given better results than other kernel-based SVMs in several studies of microarray data, including our prior work with Multiple Myeloma. Our preliminary studies using kernel functions to create a non-linear separator that *can* separate between the absence and presence of homozygosity have resulted in poorer performance than the simple linear separator. Further experimentation with SNP data using kernel functions is a direction for future work.

A major problem in ML applications is the "curse of dimensionality" – having many more features than examples. SVMs are more robust than some other ML algorithms when faced with high-dimensional data. Nevertheless, as with other ML algorithms, SVMs typically benefit from feature selection. Therefore, before training an SVM on our SNP data, we eliminate 90% of the features. Specifically, we select the top 10% (300) of the features according to information gain as described in the following paragraph. But before discussing the details of this approach, an important methodological point must be made. It is relatively common, though incorrect, to perform feature selection once by looking at the entire data set, and then to run cross-validation to estimate the accuracy of the learning algorithm. The resulting accuracy estimate is typically higher than will be achieved on new data, because the test data for each fold of cross-validation played a role in the initial feature selection process; hence information has "leaked" from the test cases into the training process. To avoid such an over-optimistic accuracy estimate, we repeated the following feature selection process on every fold of cross-validation, using only the training data for that fold. We chose to use cross-validation to assess the accuracy of our model since it is robust to high-dimensional data.

For each SNP feature we compute the information gain of the optimal split point, either between -1 and 0 or between 0 and 1. Information gain is defined as follows. The entropy of a data set is $-p\log_2 p - (1-p)\log_2(1-p)$ where p is the fraction of examples that belong to class "predisposed" (either class could have been used). A split takes one data set and divides it into two data sets: the set of examples for which the SNP feature has a value below the split-point and the set of data points for which the SNP feature has a value above the split-point. The information gain of the split is the entropy of the original data set minus the weighted sum of entropies of the two data sets resulting from the split, where these entropies are weighted by the fraction of data points in each set. The SNP features are then ranked by information gain, and the top-scoring 10% of the features are selected. A natural variant to the preceding procedure would involve making *both* splits, the split between -1 and 0 as well as the split between 0 and +1, dividing the original data set into three instead of two. The entropy and information gain equations extend naturally to this case as well. We chose to use binary splits to rank features because the linear SVM that will use these features will effectively make binary splits for each feature.

3. RESULTS AND DISCUSSION

We tested the approach described in the previous section using leave-one-out cross-validation. The confusion matrix is shown in Table 1. This yields an accuracy estimate of 71%, which is significantly better than random guessing. While this accuracy is not nearly as high as the accuracies we have grown accustomed to seeing for prediction of cancer vs. normal from microarray data, it is nevertheless exciting given that this prediction is based only on SNP data, which does not change once the disease occurs, and given that we had a relatively sparse covering of the genome with only 3000 SNPs.

Table 1: Confusion Matrix. This table shows how the class values predicted by the SVM on the test cases relate to the actual class values. This yields an accuracy estimate of 71%.

		Predicted: Not predisposed	Predicted: Predisposed
Actual	Not predisposed	31	9
	Predisposed	14	26

To assess the significance of this result, we performed a permutation test. Permutation testing assesses the dependency of a classifier to the specific data set that is was designed for. This method is commonly used in situations where data is limited to give an estimate on the error of a classifier [8]. We performed the permutation test by randomly permuting the labels – "predisposed" and "not predisposed" – among the patients and running the entire cross-validated learning process on this new dataset. This entire procedure was repeated 10,000 times. The accuracy of these 10,000 classifiers very closely fits a normal distribution. The results of this test can be seen in Figure 4 and illustrate that our result of 71% is significant at the $p < 0.05$ level using a two-tailed test of significance. A standard binomial test was also performed and also established significance of the 71% result at the $p < 0.05$ level (two-tailed).

Although SNPs are highly unlikely to change within a single person as that person ages, it is true that certain SNPs will be underrepresented in certain age populations. For instance, a SNP that is associated with a gene responsible for causing a massive heart attack at age 50 will be present in a much higher proportion of 40-year-old patients than of 70-year-old patients. This emphasizes the need for the model that we build to be interpretable so that we can examine the SNPs that the model uses for prediction and determine their potential role in the disease mechanism.

In order to show that our learning algorithm is not basing its model on the age of the patients, we obtained SNP

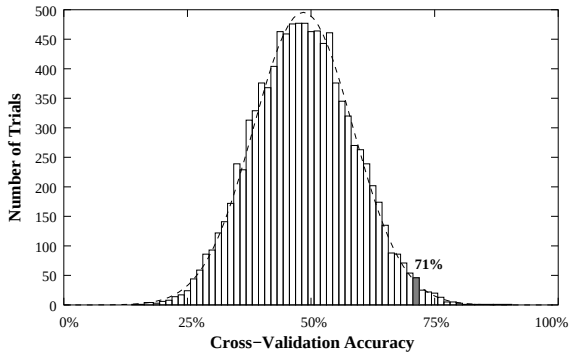


Figure 4: Results of a permutation test to estimate error of the classifier. We performed the permutation test by randomly permuting the labels – "predisposed" and "not predisposed" – among the patients and running the entire cross-validated learning process on this new dataset. This entire procedure was repeated 10,000 times. The accuracy of these 10,000 classifiers very closely fits a normal distribution. The 71% classifier is significant at the $p < 0.05$ level (two-tailed).

data on 28 unrelated persons without Myeloma from the SNP consortium[4]. 14 persons were older than 70 years-of-age and 14 were younger than 40 years-of-age at the time of SNP analysis. For each person, 2911 SNPs were chosen to provide broad genome coverage [13], just as the 3000 SNPs used with our "predisposed" and "not predisposed" patients were. Using the exact same procedure as we used for the "predisposed" and "not predisposed" data, we built a model using SVM^{light} after feature selecting the top 10% of features and using leave-one-out cross validation. The resulting accuracy was 46% and the confusion matrix can be seen in Table 2. Although the 2911 SNPs chosen were a different set of SNPs than the 3000 used with our patients, we believe that this result does provide evidence that the 71% accuracy we are obtaining with our model is unlikely to be from merely predicting age well. Our future work will include obtaining SNP data on persons such as these 28 using the same set of SNPs to further validate this conclusion.

Table 2: Confusion Matrix for Control Data. This table shows how the class values predicted by the SVM on the test cases relate to the actual class values. This yields an accuracy estimate of 46%.

		Predicted: Over 70	Predicted: Under 40
Actual	Over 70	6	8
	Under 40	7	7

From the data in Table 1, we can compute the true positive and false positive rates for our model. The true positive rate is calculated as the fraction of the "predisposed" patients who are correctly classified as "predisposed." The false positive rate is calculated as the fraction of the "not predisposed" patients who are incorrectly classified as "predisposed." Using this method, we see that our model has a true positive rate of 65% and a false positive rate of 22.5%. However, because Myeloma is relatively rare in the general population, a false positive rate of 22.5% would result in a large number of patients being misdiagnosed as "predisposed." This is because our model was built with the naïve assumption that both types of misclassification errors (classifying "predisposed" as "not predisposed" and classifying "not predisposed" as "predisposed") are equally bad. In order to have the freedom to vary the relative misclassification costs of these two types of errors, we have plotted a Receiver Operator Characteristic (ROC) curve. An ROC curve is a standard way of assessing the accuracy of a model at varying degrees of conservativeness. As we see in Figure 5, if we choose a more conservative model that bounds our false positive rate to 5%, we are still able to achieve a true positive rate of 42.5%. This is very encouraging considering the limited data on which this model was based.

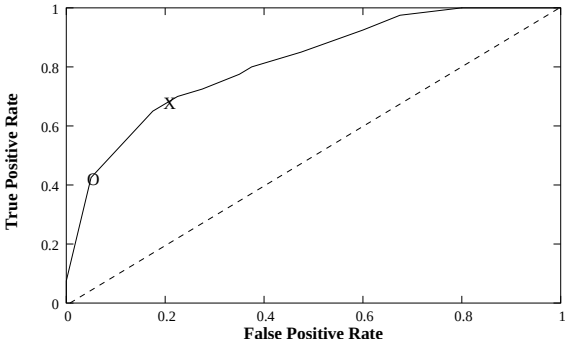


Figure 5: The ROC curve shows that linear SVMs (solid line) perform significantly better than random guessing (dotted line). It also shows the accuracy if we tune the SVM model to bound the false positive rate (since Myeloma is relatively rare in the general population). The point (5%, 42.5%) is noted with an O. The point without tuning (22.5%, 65%) is noted with an X. The true positive rate is calculated as the fraction of the "predisposed" patients who are correctly classified as "predisposed." The false positive rate is calculated as the fraction of the "not predisposed" patients who are incorrectly classified as "predisposed."

From these results we conclude that SNP data does indeed provide predictive ability for cancer susceptibility. That is the primary conclusion of this paper. The next question is whether the resulting SVM model can provide any insight into the disease. Ideally the SVM model would be based on only one or a few SNPs; that is to say, all but a few SNPs would have coefficients of zero in the equation for the separating hyperplane. Unfortunately, the model gives over 150 SNPs with non-zero coefficients. The maximum cross-validation accuracy that can be obtained for this data-set using a single SNP alone (using this SNP as a single voting attribute instead of using an SVM) is 61%, which is obtained using SNP 739514; a SNP on chromosome 4 at a

[4]http://snp.cshl.org

location of 150,853,009 bp from the telomere of the p arm. If we instead use the top 3 SNPs (as determined by information gain) in unweighted majority-voting, we can achieve 72.5% accuracy (using SNPs 739514, 521522, 994532). Investigation of the full list of 150 SNPs is under way, but at this point we cannot claim that the model has provided useful insight into the disease. Although SVMs can accurately model the relative significance of features and their interactions, compared to some other algorithms such as decision trees and naïve Bayesian networks, their models are not easily interpretable.

After finishing analysis of the linear SVM results, we reran our experiments using a few other standard ML algorithms. None of the algorithms that we tried – polynomial SVMs, decision trees (with and without boosting) and naïve Bayesian networks – performed significantly better than chance. Thus, we see that our choice of linear SVMs was a good one for this dataset and that the choice of algorithm can be very important when modeling biological datasets.

The only difference between linear and polynomial SVMs in this model is that polynomial SVMs are able to separate between the absence and presence of homozygosity (see Figure 3) which, as we discussed in Section 2, is not biologically relevant. Thus, it is likely that polynomial SVMs were led astray by irrelevant correlations whereas linear SVMs were not able to be similarly led astray. Like polynomial SVMs, naïve Bayesian networks and decision trees are not well suited to this dataset. Because it appears likely that susceptibility to Myeloma is controlled by QTL and is not a simple Mendelian or near-Mendelian disorder, the feature independence assumption of naïve Bayes is strongly violated in our dataset. Decision trees are not robust with high-dimensional data and may have been led astray like polynomial SVMs since they too can separate absence and presence of homozygosity.

4. ONGOING AND FUTURE RESEARCH

Ongoing and future work is focused in three directions. First, we are cross-tabulating the SNP results with gene expression microarray results for Multiple Myeloma [7]. We are interested in whether any SNPs appear in or near genes that are differentially expressed in Myeloma vs. normal mRNA samples. We have found 11 SNPs that appear within 1Mbp of one of the top 1% informative (by information gain) genes for predicting Myeloma vs. normal from mRNA. We are also interested in whether any SNPs appear in or near genes that are differentially expressed in Myeloma vs. MGUS (a benign form of Myeloma) mRNA samples. We have found 7 SNPs that appear within 1Mbp of one of the top 1% informative (by information gain) genes for predicting Myeloma vs. MGUS from mRNA. We use a tolerance of $\pm$1Mbp for two reasons. First, we see this breadth of deviation in SNP locations when using different information sources, e.g. NCBI and GeneCards. Second, research into haplotype blocks has revealed that large regions of DNA see very little recombination and tend to remain conserved, while recombination is largely isolated to certain "hot spots." Hence a SNP allele could be informative of a gene allele even if the SNP does not occur within the gene but only near it.

The second direction for ongoing and future work is to further tune the linear SVM algorithm as well as experimenting with other types of SVMs, such as Gaussian kernel SVMs (also available with SVM^{light}, for example), and with other types of modeling algorithms from ML and statistics. The goal of this work is to find a model for predicting predisposition for Myeloma that uses a smaller set of features for classification. This will allow us to gain a better insight into those regions that are important for conferring susceptibility.

Our final direction for future work is to repeat these experiments on a larger pool of participants, and using a denser coverage of SNPs, in order to further validate all of the findings of this study. We plan to do this in the next year or two when a sufficient number of the "predisposed" population (relatively rare) are referred to our center. In addition, we will look at the allele frequencies of the highly predictive SNPs in another similarly aged matched cohort.

5. ACKNOWLEDGMENTS

MW was supported in part by NLM grant 5T15LM007359. DP was supported in part by NSF grant 9987841 and by grants from the University of Wisconsin Medical School, Graduate School, and Comprehensive Cancer Center. JS and BB were supported in part by National Cancer Institute, Bethesda, MD grant CA55819. JS was also supported by a grant from the Fund to Cure Myeloma. We are grateful to Mark Craven, Christina Kendziorski and Jude Shavlik for helpful discussions of methodology and significance tests.

6. REFERENCES

[1] M. P. S. Brown, W. N. Grundy, D. Lin, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *P Natl Acad Sci*, 97(1):262–267, Jan 2000.

[2] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc*, 2:121–167, 1998.

[3] D. Burgner, K. Rockett, H. Ackerman, et al. Haplotypic relationship between SNP and microsatellite markers at the NOS2A locus in two populations. *Genes Immun*, 4(7):506–514, Oct 2003.

[4] S. B. Gabriel, S. F. Schaffner, H. Nguyen, et al. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, Jun 2002.

[5] A. M. Glazier, J. H. Nadeau, and T. J. Aitman. Finding genes that underlie complex traits. *Science*, 298(5602):2345–2349, Dec 2002.

[6] T. R. Golub, D. K. Slonim, P. Tamayo, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct 1999.

[7] J. Hardin, M. Waddell, C. D. Page, et al. Evaluation of multiple models to distinguish closely related forms of disease using DNA microarray data. *Stat Appl Genet Mol*, 3(1), June 2004.

[8] T. Hsing, S. Attoor, and E. Dougherty. Relation between permutation-test p values and classifier error estimates. *Mach Learn*, 52:11–30, 2003.

[9] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

[10] A. G. Knudson, Jr. Genetics of human cancer. *Annual Review of Genetics*, 20:231–251, 1986.

[11] R. Lewis. SNPs as windows on evolution. *The Scientist*, 16(1), Jan 2002.

[12] J. Li and T. Jiang. Efficient inference of haplotypes from genotypes on a pedigree. *J Bioinformat Comput Biol*, 1(1):41–69, Apr 2003.

[13] T. C. Matise, R. Sachidanandam, A. G. Clark, et al. A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *Am J Hum Genet*, 73(2):271–284, Aug 2003.

[14] T. T. Nikiforov, R. B. Rendle, P. Goelet, et al. Genetic bit analysis: a solid phase method for typing single nucleotide polymorphisms. *Nucleic Acids Res*, 22(20):4167–4175, Oct 1994.

[15] M. Phillips and M. Boyce-Jacino. A primer on SNPs - part 1. *Innov Pharm Tech*, 1:54–58, Jan 2001.

[16] M. Ringnér, C. Peterson, and J. Khan. Analyzing array data using supervised methods. *Pharmacogenomics*, 3(3):403–415, May 2002.

[17] N. J. Risch. Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–856, Jun 2000.

[18] A. Rosenwald, G. Wright, W. C. Chan, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New Engl J Med*, 346(25):1937–1947, Jun 2002.

[19] M. V. Seiden and K. C. Anderson. Multiple myeloma. *Curr Opin Oncol*, 6(1):41–49, Jan 1994.

[20] V. V. Symonds and A. M. Lloyd. An analysis of microsatellite loci in arabidopsis thaliana: Mutational dynamics and application. *Genetics*, 165:1475–1488, Nov 2003.

[21] The International HapMap Consortium. The international hapmap project. *Nature*, 426(6968):789–796, Dec 2003.

[22] G. A. Thorisson and L. D. Stein. The SNP consortium website: past, present and future. *Nucleic Acids Res*, 31(1):124–127, Jan 2003.

[23] L. J. Van't Veer, H. Dai, M. J. Van de Vijver, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, Jan 2002.