CPSC 68 Final Report: Predicting Population Membership using Single Nucleotide Polymorphisms

Jessica Jowdy JJOWDY1@SWARTHMORE.EDU
Swarthmore College, 500 College Ave, Swarthmore, PA 19081 USA

Brooke Kelsey BKELSEY1@SWARTHMORE.EDU
Swarthmore College, 500 College Ave, Swarthmore, PA 19081 USA

Abstract

In this paper we present *SNPClassify*, a modified ID3 approach for classifying individuals into global populations by comparing their genetic variations, referred to as single nucleotide polymorphisms (SNPs). These differences date back to the divergence of the earliest human species. Evolutionary forces, such as gene flow and genetic drift, have resulted in the spread of genetic differences to specific locations around around world. Our implementations utilize the different frequencies at which various regional populations express the dominant and recessive alleles of each SNP. We first use a large population data set to train an ID3 decision tree that determines the population to which a particular individual belongs. We then use patient-specific genotype information to create multiple decision trees that predict whether an individual belongs to a particular population. We test the accuracy of the models by performing both internal and external validation. Our results reveal that the decision trees created with these implementations are appropriate models for the data and accurately predict population membership.

1. Introduction

1.1. Modeling Human Evolution Patterns

One of the most debated issues in paleoanthropology is the evolutionary origin of (*homo sapiens*), the current species of humans. Two main theories have arisen from these debates: (1) the 'Multiregional Continuity Model', which suggests that regional populations underwent convergent evolution into modern human beings; and (2) the 'Out of Africa Model', which suggests that human beings originated from Africa and simply migrated to other parts of the world. There is evidence to support both hypotheses, and while we cannot know for certain which theory is true, we can say that geographic distance played a major role in the development of distinct human populations. Over the last thousands of years, genetic drift, gene flow, and natural selection (as well as other natural phenomena) have contributed to these regional differences in human beings (Edwards, 2012).

1.2. Documenting Variation with SNPs

Modern technology significantly improved the ability to study the evolutionary relationship between various populations of humans. Advanced studies in human genome sequencing have revealed that roughly 99.9% of every healthy human genome is identical (Network, 2003). Studies have placed significant emphasis on determining the 0.1% of the genome that differs. A single nucleotide polymorphism (SNP) is defined as a DNA sequence variation occurring commonly within a population in which a single nucleotide in the genome differs between organisms of the same species. In the human genome, SNPs occur once every 300 nucleotides. They can be found both on coding and non-coding regions of DNA. Their pervasiveness in a population is affected by factors such as natural selection, genetic recombination, and mutation rates (Norrgard & Schultz, 2008).

We performed two methods of classification, both of which used SNPs to predict a patient's population membership. These methods utilized the ID3 algorithm to create a decision tree. In our first method, the training set consisted of global allele counts for each population. In our second method, the training set consisted of individual patient data from each population.

CPSC 68 Bioinformatics Final Report Proceedings, Swarthmore College, Fall 2014.

2. Related Work

2.1. ID3 Adaptations

The ID3 algorithm has been modified time and again in attempts to improve various perceived shortcomings. (John et al., 1994) presents a new method for feature selection that avoids selecting irrelevant features. Other works such as (Cios & Liu, 1992) modify ID3 to make it more suitable for a neural network decision pattern. It converts the hierarchical tree structure of ID3 into hidden layers. This models the steps an artificial nervous system would take between the input and output to continuously arrive at various decisions. (Cendrowska, 1987) concludes that the fundamental problem with ID3 is its production of decision trees. An alternative algorithm, known as the PRISM algorithm, was introduced as a method of arriving at modular rules and avoiding the decision tree structure. It is arguable that these types of works are not so much modifying ID3 as they are using it as a basis for their own algorithms.

2.2. SNPs and Health

The primary goal of documenting SNPs is to be able to describe all possible genetic variations between the genomes of two humans. Patterns of allele expression are used to measure how global populations relate to each other. One prominent application is the study of how different populations react to various diseases. Studies such as (Cheetham et al., 2013) have found that up to 80% of cancer-related SNP mutations fall within long non-coding RNA sequences. Within cancerous cells, the mutation is typically more drastic than the changing of a single SNP. Regions prone to cancerous mutation are referred to as CNVs or CNAs (Copy-Number Variations/Alliterations); within a single CNA, thousands of SNPs can mutate or be deleted. (Chiang et al., 2008) presents an algorithm for best determining new CNAs through parallel sequencing, due to the fact that continued reliance on microarray development is becoming insufficient. Works such as (Wolf et al., 2004; Temam et al., 2007; Leary et al., 2010) are only a few of many dedicated to improving sequence technologies available for personalized cancer detection. The correlation between SNP genetic profiles and cancer risk is so strong that it prompted the creation of CaSNP (Cao et al., 2010), a database entirely dedicated to matching SNP profiles with those found in numerous cancers. By comparing gene expression profiles in non-coding regions of normal and cancerous tissues, works such as (Reiche et al., 2014) have contributed to the relatively recent ability to test for genetic predisposition to breast cancer.

2.3. SNPs and Human Evolution

Many recent studies share the goal of using SNPs to identify a person's population of origin using only their genetic data and an ID3-based decision tree algorithm. (Choudhury et al., 2014) has found several interesting patterns of SNP behavior between populations using data from the 1000 Genomes database (Consortium et al., 2010). The NCBI (NCBI, 2014) and HapMap Project (Gibbs et al., 2003) provide additional SNP population data. The human genome contains hundreds of thousands of SNPs, and studies such as (Zhou & Wang, 2007) are working to determine the most effective and relevant SNPs to perform this classification. Many population-specific studies such as (Hsieh et al., 2014; Kodaman et al., 2013; Yamaguchi-Kabata et al., 2008) have also been done with SNP data. These types of studies have important ties to areas such as genetic ancestry and disease behavior.

3. Methods

We executed two implementations of *SNPClassify*. The underlying classification algorithm for both Method 3.3 and Method 3.4 was the ID3 decision tree algorithm, presented in Algorithm 1.

3.1. Input Data

The input data for the first method (Method 3.3), was downloaded from the 1000 Genomes browser (Consortium et al., 2010) available through the NCBI. Within the 1000 Genomes browser, the user specifies the chromosome and position range for which they would like to download the SNP data. We chose positions 98152280-99152033 on the first chromosome and downloaded the aggregate population data for all fourteen available populations[1]. To build our test set, we selected two patients from each of the populations, one male and one female, and downloaded their individual SNP data. The aggregate population data came in the form of dominant and alternate allele counts for each population (as well as a global total). The patient data came in the form of genotype likelihoods for each SNP. These were simply the probabilities that a patient was dominant-dominant, dominant-recessive, or recessive-recessive at the particular SNP location.

The input data for the second method (Method 3.4), was obtained through the HapMap project (Gibbs et al., 2003).[2]

[1] Several related works used the entire genome as possible features. We were limited in computing power and could not find any evidence to support any one chromosome range being most informative, so our choice was arbitrary.

[2] Both 1000 Genomes and HapMap use the same 14 population identifiers: ASW - Americans of African Descent in Southwestern U.S.; CEU - Utah Residents; CHB - Han Chinese in Beijing; CHS - Southern Han Chinese; CLM - Columbians from Medellin,

We were able to download patient-specific data for three different populations (*YRI, CEU, JPT*) for all of the SNPs located on the second chromosome. Rather than using genotype likelihoods, this data set provided us with the actual allele type of each individual patient. We used cross-validation to split the data into training sets and test sets. The training sets were used to create the decision trees, while the test sets were used to evaluate the accuracy of the model.

3.2. Data Structures

To handle the large amount of input data in an organized way, we created multiple data structures for both implementations of *SNPClassify*: *SNP*, *Country*, *Patient*, *Node* and *Leaf*. Some of these classes were applicable to both methods, while some were relevant to only one specific method.

Each *SNP* object stored the chromosome and position numbers of the given SNP, as well as the unique ID assigned to it. Additionally, the *SNP* objects created for Method 3.3 stored fifteen *Country* objects for the global population and fourteen individual populations defined by the 1000 Genomes browser. Each *Country* object stored the four allele counts given in the aggregate population data.

A *Patient* object stored the patient's ID and a dictionary of SNP expressions for each of the positions in the range specified. For Method 3.3, the dictionary stored genotype frequencies converted to base-10 probability values (originally given as log probabilities). In the case of Method 3.4, it stored the actual allele types expressed for each patient. Method 3.4 also maintained a variable for the true population label of the patient.

Node and *Leaf* objects were created to help render the decision tree produced by the ID3 algorithm. The *Node* contained a *SNP* object with the maximum information gain and the value at which its split. In Method 3.3, the splitting value was a genotype likelihood threshold, while in Method 3.4, the splitting value was the dominant-dominant allele type at the node. Additionally, the Node class stored pointers to the left and right children for later traversal of the tree, and in the case of Method 3.4, a parent pointer was also stored. The *Leaf* class was a simplification of the Node class. These objects contained populations label, which classified a patient when traversing the tree.

3.3. One Tree, All Populations

Algorithm 1 ID3

```
Input: examples $E$, features $F$
if $|E| == 0$ then
    return $None$
end if
if $|F| == 0$ then
    return $leaf$ with majority vote
end if
for each attribute $A$ in $F$ do
    Calculate $infoGain$
end for
$A_{best}$ = max $infoGain$
newNode($A_{best}$)
$F_{new}$ = $F - A_{best}$
for each value $v$ of $A_{best}$ do
    $E_v$ = examples where $A_{best} = v$
    newNode.child[$v$] = ID3($E_v$, $F_{new}$)
end for
```

3.3.1. Decision Tree Design

In this approach to building the ID3 decision tree, the algorithm was applied by treating each of the possible populations as examples (leaves) and individual SNPs as features. The initial decision tree was created by adapting the information gain equations to use the dominant and recessive ratios of each SNP. The internal nodes represented the SNPs providing the highest information gain for the set of populations, with the root node providing the maximum information gain. In addition, a threshold was set at each node to establish a means for partitioning the populations using the recessive allele ratios calculated from the aggregated allele counts. Because the 1000 Genomes patient data was in the form of continuous probabilities rather than discrete dominant or recessive allele expressions, a second threshold was set at each SNP using the ratio of the global recessive count to the total count. Patients in the test group were classified by traversing the tree according to the value of their genotype frequencies in relation to the genotype likelihood threshold at the current node.

3.3.2. Information Gain

Equation 1 presents our application of the ID3 information gain mechanism to fit the goals of the method. $H(A)$ represents the entropy of a particular SNP, and $H(A|C)$ represents the entropy of a particular SNP given that we know the population being studied. As each SNP has a dominant and alternate (recessive) allele expression, the entropy $H(A)$ represents the predictability of the SNP being expressed as its dominant or recessive value.

$$infoGain = H(A) - H(A|C) \tag{1}$$

Columbia; FIN - Finnish in Finland; GBR - British in England and Scotland; IBR - Iberian Population in Spain; JPT - Japanese in Tokyo, Japan; LWK - Luhya in Webuye, Kenya; MXL - Mexican Descendants in Los Angeles, U.S.; PUR - Puerto Ricans from Pureto Rico; TSI - Toscani in Italia; YRI - Yoruba in Ibadan, Nigeria.

Equation 2 presents our method for calculating the initial entropy of each SNP using the global count information provided by the 1000 Genomes Project. A_d represents the count for the dominant allele, and A_r represents the count for the recessive allele.

$$H(A) = -(P(A_d)\log_2(A_d) + P(A_r)\log_2(A_r)) \quad (2)$$

The entropy of each SNP given a particular population was computed using Equation 3. Here, the examples were the fourteen different populations of which a particular patient could be a member. Because all healthy variations between human genomes occur in the SNPs, population ratios of dominant and recessive allele counts should vary enough to differentiate them purely based on allele expression frequencies. The conditional entropy was summed across all populations, as a SNP could be informative for partitioning some subgroups while providing little to no partitioning information for others.

$$H(A|C) = \sum_{v \epsilon C} P(A|C=v)H(A|C=v) \quad (3)$$

At each feature selection in Algorithm 1, the *infoGain* was calculated for every SNP that had not yet been chosen as a partitioning node in the tree. While this required a significant amount of computation time when a large number of SNPs were considered, it ensured that the most informative SNP for the population group was selected each time. This limited the size of the tree without needing to add an additional pruning component. Given the difficulty of working with the data, a method that controlled the size of the tree while also avoiding the need to extract an additional tuning set was highly beneficial.

3.3.3. Validation Method

Due to the enormity of the data set, we reduced the project to classifying individuals from three different populations *(YRI, CEU, JPT)*. To validate our model, we began by loading patient data for individuals within these populations from the 1000 Genomes Project interactive browser. The original population data set was compared to the patient data set to ensure that only SNPs that were common among both were considered. Each individual was represented by an object of the *Patient* class, which stored the patient identification number and a dictionary with *(SNP, genotype)* key-value pairs. The accuracy of the decision tree was tested using the patient data set as the test set. The patient set consisted of two individuals, male and female, from each of the three populations. Each patient's data was used to traverse down the tree, choosing child branches based on the patient's genotype likelihoods in comparison to the threshold set at each node. The traversal ended at the first leaf node encountered, which represented the tree's prediction of the patient's population of origin. These results are presented in 4.1.

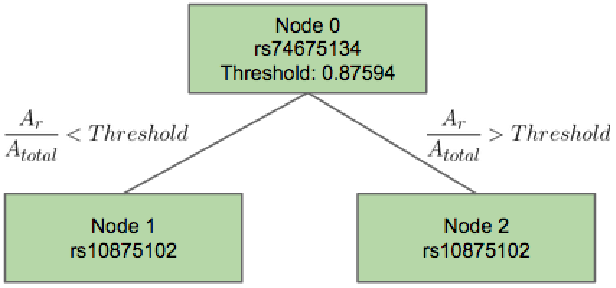
Node 0
rs74675134
Threshold: 0.87594
$\frac{A_r}{A_{total}} < Threshold$
$\frac{A_r}{A_{total}} > Threshold$
Node 1
rs10875102
Node 2
rs10875102

Figure 1. Example of a 3-*Node* tree in Method 3.3 using a threshold for dominant frequency to partition the populations. Populations with a recessive frequency higher than the threshold were sent down the right branch, while populations under the threshold were sent to the left.

3.4. One Tree, One Population

3.4.1. Decision Tree Design

In this method, we utilized patient-specific data sets from the HapMap Project database to construct decision trees for each population. These decision trees made binary predictions as to whether a particular individual belonged to the population in question. Due to the enormity of the data sets, we only classified individuals from three different populations *(YRI, CEU, JPT)*. The data collected for each patient included every SNP located on the second chromosome. Upon extracting the data, the patients were split into training sets and test sets. The initial inputs to the algorithm were the patients in the training set data (*E*) and the SNPs common among the three populations (*F*). The SNP that maximized information gain was chosen as the new node and added to the list of nodes in the tree. The examples were distributed to the right and left children of this node based on their allele value for the chosen SNP. As illustrated in Figure 3.4.3, those patients who were dominant-dominant for the given SNP were sent to the left child, while patients with any other allele combination were sent to the right child. The algorithm was then called recursively on the children until either *E* was empty or there were no SNPs left in *F* that could effectively split the data. In the latter case, a leaf was created whose value was the majority label among the patients left in *E*.

3.4.2. Information Gain

Because we implemented a binary decision tree in this method, the classification of each patient was simply whether the patient belonged to the given population. This reduced the complexity of the information gain calculation significantly, as seen in Equation 4.

$$infoGain = H(C) - H(C|A) \tag{4}$$

In this equation, $H(C)$ represents the entropy of a population label, and $H(C|A)$ represents the conditional entropy of a population label given that we know the patient's allele type. As seen in Equation 5, we can calculate $H(C)$ by determining the probabilities of each population label. The set of population labels *C* was simply *(Yes, No)*, indicating whether or not the patient belonged to the particular population. The probabilities used in this equation were calculated using the maximum likelihood estimates of each parameter.

$$H(C) = -\sum_{i=1}^{k} P(C_i) \log_2(C_i) \tag{5}$$

The conditional entropy term in Equation 4 considers how significantly the total entropy is reduced by the introduction of a feature. For each SNP, we had to consider the probability of each possible allele type, as well as the conditional probability of an individual being in a population given each allele type. In this method, we simplified the set of possible alleles *A* to *(Dom, Not-Dom)* to indicate that the individual was either dominant or not dominant for the SNP. As was done in Equation 5, the probabilities and conditional probabilities were calculated using maximum likelihood estimations. Equation 6 illustrates the conditional entropy used in the information gain equation.

$$H(C|A) = \sum_{v \epsilon A} P(C|A = v) H(C|A = v) \tag{6}$$

The information gain was used to determine which SNP in the feature list provided the most information about the data set. Specifically, the calculation determined which SNP best split the data effectively. We refer to this SNP as F_{best}. For each iteration of the ID3 algorithm, we calculated the information gain for all of the SNPs remaining in the feature list to determine F_{best}. If it was the case that none of remaining features in the list provided adequate information gain to be relevant in the classification model, a leaf node was created whose value was the majority population label of the examples left in the list, and the algorithm terminated.

3.4.3. Validation Methods

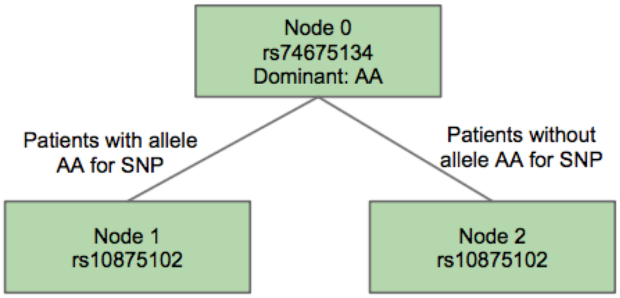
Node 0
rs74675134
Dominant: AA
Patients with allele AA for SNP
Patients without allele AA for SNP
Node 1
rs10875102
Node 2
rs10875102

Figure 2. Example of a subtree generated using Method 3.4. The dominant allele for the most significant SNP acted as the feature selector. Patients who possessed the dominant-dominant allele move to the left child, while patients who possessed any other allele type move to the right child.

The test set was then used to evaluate the accuracy of the model. We created a tree traversal algorithm that allowed each patient in the test set to travel down the decision tree to the appropriate leaf node. The value at the leaf node predicted whether the patient belonged to the give population. However, because we knew the population label for each patient *a priori*, we were able to compute the sensitivity, specificity, and accuracy of each of the trees.

Due to the limited number of patient data available, we implemented 5-Fold Cross Validation on the aggregated patient data set to maximize the amount of data used for training the model. By redistributing the data in this manner, each patient was involved in both the training and the evaluation of the algorithm without overfitting the data and producing overly optimistic results. The true positive, true negative, false positive, and false negative counts were accumulated for each fold. The overall accuracy, specificity, and sensitivity for each population tree were then calculated from these these counts.

4. Evaluation

4.1. One Tree, All Populations

The method presented in 3.3 was evaluated for accuracy using a test set composed of two patients from each of the three populations considered. Unlike the decision trees produced by Method 3.4 that provided a binary classification (concluding the patient was or was not a member of that population), the resulting tree provided trinary classification, with one leaf per population. Thus, it was not possible to clearly define the meaning of a false positive or false negative in the context of the tree. Measures of performance such as specificity and sensitivity could not be applied, but the test set was small enough that the success of each individual classification could easily be determined. Table 4.1 presents the results of testing Method 3.3. The decision tree and traversal method were accurate enough to correctly classify five of the six patients. The one patient that was incorrectly classified was separated from its correct population by a distance of only one leaf. Upon looking at the SNP expression data for that patient, we found that the patient exhibited the dominant allele while most of his population exhibited the recessive allele. This indicates that while our decision tree worked well for the cases where the patient's allele expression matches the majority allele expression for the rest of the population, an atypical patient will diverge on a branch and end up classified as a member of the wrong population. The closer these divergences are to the root SNP, the more severely they affect the overall quality of the classification.

Table 1. Classification Results of Method 3.3

Patient ID	Actual Population	Classification
NA18487	YRI	YRI
NA18489	YRI	JPT
NA06986	CEU	CEU
NA06989	CEU	CEU
NA18940	JPT	JPT
NA18941	JPT	JPT

The method's ability to handle a larger number of populations was tested using simple external validation. The decision tree was built using all fourteen populations, and we evaluated the resulting tree by discovering patterns in how the tree partitioned the populations into leaves. The tree, pictured in Figure 3, was able to group together populations in similar geographic locations. The three leftmost leaves are the three Asian populations: CHB, CHS and JPT. The two rightmost leaves are the two native African populations, LWK and YRI. The central subgroup of AWS, CEU and GBR have likely experienced a noticeable amount of European/North American genetic crossover, so it is possible that these three populations are genetically similar. Overall, the tree utilized SNPs successfully to partition the populations in a way that makes geographical sense.

4.2. One Tree, One Population

Figure 4.2 shows the results of 5-Fold Cross Validation on the data set. Due to the long runtime of the algorithm, we could not use pruning when performing this statistical analysis. The accuracy of our models was relatively high, with each of the three decision trees achieving an accuracy of approximately .80. The sensitivity and specificity values were also high, reaching average values of .75 and .80, respectively. From this, we were able to conclude that the decision trees were appropriate models for determining individual membership in populations.

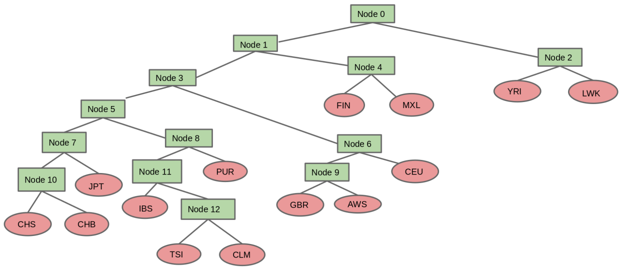

Figure 3. Resulting tree using data taken from a subsection of the first chromosome. From a range of 100,000 positions on the chromosome, the algorithm identified twelve SNPs as the most informative. The 'Node X' labels of the tree represent specific SNPs, referred to by a unique 'rs-x' ID number.

The second method addresses the primary fault of the first. In this method, a patient was not automatically discounted from its true population of origin if it differed from the population majority at one SNP. The patient could take one of multiple possible paths to a leaf, and achieve the same predicted classification.

Another significant benefit to this approach was the ability to clearly define false positives and false negatives, providing concrete measures of performance. Because each decision tree used a binary classification, a false positive was defined as a patient incorrectly classified into the population, and a false negative was defined as a patient that was not classified in the population to which they belonged. This portion of the analysis was not possible in Method 3.3. In order to be classified into a specific population, the patient had to possess a somewhat concrete SNP expression pattern that was shared by the majority of the individuals in their population of origin. If the patient diverged from this pattern at any node in the tree, they would be sent along the wrong branch and ultimately end up classified into a population to which they did not have membership.

5. Conclusion

Both applications of the ID3 algorithm resulted in the successful classification of patients of three diverse populations using only SNP data taken from a single chromosome. A significant portion of time was spent modifying the ID3 algorithm to fit the available data sets; as such, we were not able to perform as wide of a variety of tests and adjustments as we would have liked. Method 3.3 was able to classify almost all patients correctly in the three-example tree and produce a logical fourteen-example tree.

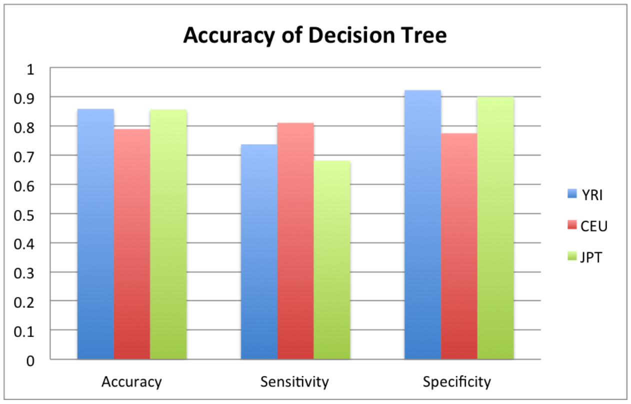

Figure 4. Bar graph of the accuracy, sensitivity, and specificity values for each of the three decision trees. The high accuracy of each of the trees indicate that they are appropriate models for binary population classification.

However, we could not conduct statistical analyses on either of these trees. This would have given us an idea of the scalability of our algorithm design for Method 3.3. We instead focused our efforts on Method 3.4, which provided significantly more tangible results. By creating one tree for each population and limiting the classification to a *(Yes, No)* approach, we were able to both generalize our model and measure the accuracy of our implementation. Use of binary membership allowed us to measure accuracy, sensitivity, and specificity on each individual tree. These results are presented in 4.2, and prove that this method of classification produced the correct predictions on the majority of the test set.

To improve the robustness of Method 3.4's accuracy, we implemented 5-fold cross-validation. We also attempted to implement pruning in order to improve the efficiency of method 3.4, however, this is still a work in progress. Despite this, we were still able to clearly demonstrate two unique and accurate methods of population classification using SNP data.

6. Future Work

We have identified several ways in which these methods could be improved in the near future. To improve Method 3.3, the next logical step would be to obtain more data from 1000 Genomes Project and test the method's performance. The reach of the data could be expanded in a variety of ways, such as the number of populations (the current evaluation considered only three of the fourteen available), the number of patients (only six total were tested), and the chromosome and position ranges (only a section of 100,000 SNPs on the first chromosome were considered when building the tree pictured in Figure 4.1).

To improve Method 3.4, we could finalize our pruning algorithm and allow it to run in conjunction with Five-Fold Cross Validation. By adding pruning to our decision tree features, we would be able to reduce the complexity of the tree and generalize the model for more accurate results. We could also estimate parameters using a Bayesian approach, rather than maximum likelihood estimations. Implementing a strategy such as Laplace estimates would reduce the number of conditions we would have to check when calculating the information gain for each SNP. Also, because we have limited data, we could account for the probabilities of certain events occurring that are not represented in the data set. Finally, we hope to develop a method of aggregating the results of each individual tree to make a final prediction of a patient's population membership. While we question its implementation, the incorporation of this feature would allow decision trees to make much more accurate classifications for examples with *n* possible labels.

7. Project Recap

By far the biggest challenge faced over the course of the project was locating reliable data that fit our needs and processing it into a form that suited our algorithms. Method 3.3 needed to be adjusted multiple times to work with the 1000 Genomes data, as the population data was given in integer counts while the patient data was given in log-probability. The resulting tree was built on counts, while the classification traversal used a likelihood threshold to decide which branch a patient should be sent down. Method 3.4 required a new data set with an entirely different file structure. The HapMap data gave concrete values for each patient's allele expression; while this worked well with the goals we had for Method 3.4, our algorithm and experimental designs were restricted by the run time and the data sets available.

Our final implementation differed significantly from our original proposal, with the only commonality being the application of SNPs as a classification mechanism. The original goal was to use SNPs for disease prediction; however, we were unable to find relevant data for any disease that was not behind an NIH membership wall. We then changed our project to fit SNP data sets accessible to the public, which resulted in focusing on the use of genetic variation to predict population membership.

Acknowledgements

We would like to thank Professor Soni for his extensive guidance throughout the course of this project. We would also like to thank the 1000 Genomes Project and the HapMap Project for making the SNP data available online,

without which this project would not have been possible.

References

Cao, Qingyi, Zhou, Meng, Wang, Xujun, Meyer, Cliff A, Zhang, Yong, Chen, Zhi, Li, Cheng, and Liu, X Shirley. Casnp: a database for interrogating copy number alterations of cancer genome from snp array data. *Nucleic acids research*, pp. gkq997, 2010.

Cendrowska, Jadzia. Prism: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, 27(4):349–370, 1987.

Cheetham, SW, Gruhl, F, Mattick, JS, and Dinger, ME. Long noncoding rnas and the genetics of cancer. *British journal of cancer*, 108(12):2419–2425, 2013.

Chiang, Derek Y, Getz, Gad, Jaffe, David B, O'Kelly, Michael JT, Zhao, Xiaojun, Carter, Scott L, Russ, Carsten, Nusbaum, Chad, Meyerson, Matthew, and Lander, Eric S. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6(1):99–103, 2008.

Choudhury, Ananyo, Hazelhurst, Scott, Meintjes, Ayton, Achinike-Oduaran, Ovokeraye, Aron, Shaun, Gamieldien, Junaid, Dashti, Mahjoubeh Jalali Sefid, Mulder, Nicola, Tiffin, Nicki, and Ramsay, Michèle. Population-specific common snps reflect demographic histories and highlight regions of genomic plasticity with functional relevance. *BMC Genomics*, 15(1):437, 2014.

Cios, Krzysztof J and Liu, Ning. A machine learning method for generation of a neural network architecture: A continuous id3 algorithm. *Neural Networks, IEEE Transactions on*, 3(2):280–291, 1992.

Consortium, 1000 Genomes Project et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

Edwards, Sophie. Analysis of two competing theories on the origin of homo sapiens sapiens: Multiregional theory vs. the out of africa 2 model. *The Collegiate Journal of Anthropology*, 1, 2012.

Gibbs, Richard A, Belmont, John W, Hardenbol, Paul, Willis, Thomas D, Yu, Fuli, Yang, Huanming, Ch'ang, Lan-Yang, Huang, Wei, Liu, Bin, Shen, Yan, et al. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

Hsieh, Ai-Ru, Chang, Su-Wei, Chen, Pei-Lung, Chu, Chen-Chung, Hsiao, Ching-Lin, Yang, Wei-Shiung, Chang, Chien-Ching, Wu, Jer-Yuarn, Chen, Yuan-Tsong, Chang, Tien-Chun, et al. Predicting hla genotypes using unphased and flanking single-nucleotide polymorphisms in han chinese population. *BMC genomics*, 15(1):81, 2014.

John, George H, Kohavi, Ron, Pfleger, Karl, et al. Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the Eleventh International Conference*, 1994.

Kodaman, Nuri, Aldrich, Melinda C, Smith, Jeffrey R, Signorello, Lisa B, Bradley, Kevin, Breyer, Joan, Cohen, Sarah S, Long, Jirong, Cai, Qiuyin, Giles, Justin, et al. A small number of candidate gene snps reveal continental ancestry in african americans. *Annals of human genetics*, 77(1):56–66, 2013.

Leary, Rebecca J, Kinde, Isaac, Diehl, Frank, Schmidt, Kerstin, Clouser, Chris, Duncan, Cisilya, Antipova, Alena, Lee, Clarence, McKernan, Kevin, Francisco, M, et al. Development of personalized tumor biomarkers using massively parallel sequencing. *Science translational medicine*, 2(20):20ra14–20ra14, 2010.

NCBI. Human variation sets in vcf format. *NCBI*, 2014. URL http://www.ncbi.nlm.nih.gov/variation/docs/human_variation_vcf/.

Network, Genome News. Genome variations. 2003. URL http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp4_1.shtml.

Norrgard, Karen and Schultz, Joanna. Using snp data to examine human phenotypic differences. *Nature Education*, 1(1):85, 2008.

Reiche, Kristin, Kasack, Katharina, Schreiber, Stephan, Lüders, Torben, Due, Eldri U, Naume, Bjørn, Riis, Margit, Kristensen, Vessela N, Horn, Friedemann, Børresen-Dale, Anne-Lise, et al. Long non-coding rnas differentially expressed between normal versus primary breast tumor tissues disclose converse changes to breast cancer-related protein-coding genes. *PloS one*, 9(9):e106076, 2014.

Temam, Stephane, Kawaguchi, Hidetoshi, El-Naggar, Adel K, Jelinek, Jaroslav, Tang, Hongli, Liu, Diane D, Lang, Wenhua, Issa, Jean-Pierre, Lee, J Jack, and Mao, Li. Epidermal growth factor receptor copy number alterations correlate with poor clinical outcome in patients with head and neck squamous cancer. *Journal of Clinical Oncology*, 25(16):2164–2170, 2007.

Wolf, Maija, Mousses, Spyro, Hautaniemi, Sampsa, Karhu, Ritva, Huusko, Pia, Allinen, Minna, Elkahloun, Abdel, Monni, Outi, Chen, Yidong, Kallioniemi, Anne, et al. High-resolution analysis of gene copy number alterations in human prostate cancer using cgh on cdna microarrays: impact of copy number on gene expression. *Neoplasia*, 6(3):240–247, 2004.

Yamaguchi-Kabata, Yumi, Nakazono, Kazuyuki, Takahashi, Atsushi, Saito, Susumu, Hosono, Naoya, Kubo, Michiaki, Nakamura, Yusuke, and Kamatani, Naoyuki. Japanese population structure, based on snp genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *The American Journal of Human Genetics*, 83(4):445–456, 2008.

Zhou, Nina and Wang, Lipo. Effective selection of informative snps and classification on the hapmap genotype data. *Bmc Bioinformatics*, 8(1):484, 2007.