Identification of cancer-relevant Variations in a Novel Human Genome Sequence

Robert Bruggner, Amir Ghazvinian[1], & Lekan Wang[1]

CS229 Final Report, Fall 2009

1. Introduction

Cancer affects people of all ages worldwide and can afflict many different parts of the body. It is one of the leading causes of death in the world, accounting for 12% of all deaths and has a mortality rate of approximately one hundred deaths per 100,000 population. An individual's specific configuration of DNA (genotype) influences susceptibility to cancer by modulating the effects of environmental factors, affecting an individual's behavior, and also by altering normal biological functions of the body. Naturally, many extensive genome-wide association studies (GWAS) of variability have been performed that attempt to associate specific genomic variations with cancer. However, due to technical limitations, those studies have focused on the evaluation of variations that occur commonly within a population while ignoring equally potent individual-specific variations (private variations). More specifically the dominant technology for measuring single nucleotide polymorphisms (SNPs) in the genome, microarray SNP Chips, have a limited number (between 200,000 and 1,000,000) of static, pre-selected SNPs that it can detect in a genome. Thus, the size of SNP microarray chips limit us to the evaluation of a subset of the estimated 2-3 million SNPs that occur in any individual's genome. Furthermore, as the set of target SNPs that a chip is able to detect is fixed, they are unable to detect variations unique to an individual. Many experts have attributed the marginal success of GWAS studies to the limited scope of variation analyzed with these chips [1].

Recent advances in sequencing technologies, specifically whole genome sequencing, have enabled the study of disease association with an individual's complete genotype. Using an individual's complete genetic sequence to correlate mutations with disease risk would allow us to find rarer, private variations that affect risk and likely contribute much more to susceptibility than previously identified common variants. Thus, as full-genome sequencing becomes cheaper and easier, methods for variation analysis should shift their focus from analysis of common SNPs to and leverage the entirety of genomic variation. The method presented here takes such an approach and our aims are as follows:

1) Use supervised learning to build a model of cancer-relevant SNPs using a range of variation topographical, chemical, and functional features.
2) Evaluate method accuracy by validating model against known cancer-relevant SNPs.
3) Identify cancer-relevant private SNPs in a newly sequenced genome using our model.

2. Materials and Methods

2.1. *Training Data*

As with any supervised learning problem, building a model of a cancer-relevant SNP and its validation requires substantial training data. To assemble such data, we extracted 5,902 cancer-associated SNPs from two manually curated variations databases, the Human Gene Mutation Database (HGMD) [2] and the Swiss-Var component of the UniProt knowledgebase [3]. Both of these repositories are populated with variations extracted from literature where the variant has been show to be statistically correlated with disease. As such, these variations were considered our gold standard for both training and validation of our model. All variations in our aggregated training data encoded different amino acids than the reference base (non-synonymous SNPs) and we were therefore limited to the classification of non-synonymous SNPs in our target genome.

In addition to our positive training set, we extracted 29,185 neutral polymorphisms from the Swiss-Var database. These non-synonymous mutations had not been associated with any diseases and were used as our negative training set.

[1] These individuals are not in CS229 but worked on the project in conjunction with BMI212.

2.2. *Features*

For each SNP, we extracted a variety of features including cancer-relevant functional associations, involvement in known cancer pathways, and variation of amino acid physiochemical properties.

SNP functional associations are binary features indicating whether a SNP is present in transcripts that have a functional role in cancer-relevant processes. To calculate these features, we mapped each SNP to corresponding coding transcripts, extracted transcript gene ontology (GO) [4] annotations from the EnsEMBL genome database [5], and determined if the associated transcript was involved in any of the following biological processes or subprocesses: "response to tumor" (GO:0002347), "apoptosis" (GO:0006915), "regulation of cell cycle" (GO:0007049), or "DNA repair" (GO:0006281). It naturally follows that the successful determination of these features relies heavily on the availability of correct, complete annotation of genes by GO terms.

We included another binary feature indicating whether each SNP is present in a gene known to be part of an established cancer pathway. To calculate this feature, we used the EnsEMBL Perl API to map each SNP to corresponding coding transcripts and map transcripts to genes. We then cross-referenced these genes against all genes in all known cancer pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [6]. The utility of this feature is dependent on the availability of previously annotated and correct descriptions of cancer pathways.

Finally, for each SNP we calculated basic physiochemical differences between the encoded wild type amino acid and variation-encoded amino acid. These features include change in amino acid polarity, change in amino acid hydrophobicity, Grantham Matrix score (a score quantifying differences between any two amino acids) [7], and transformation of an amino-acid encoding codon to a stop codon. These features, although not specific to cancer, were included because they are easily computable for all amino acids and are not dependent on prior manually curated data.

2.3. *Model Building*

2.3.1. Feature Selection

A common problem when building a classifier is the use of correlated features that make the model overly sensitive to specific factors. To avoid this problem, we applied correlated feature selection (CFS) to our 9 features in order to discover any redundant information within our feature set. Next, we evaluated each feature independently to determine their respective predictive powers. Finally, although issues of dimensionality and overfitting can potentially occur when building models from many features, our low feature set to training set size ratio minimizes the risks of these effects.

2.3.2. Classifier Selection

Using our training data, calculated features, and the Weka machine learning software framework [8], we constructed four models of a cancer-relevant SNP. We first used Weka's J48 decision tree learner with pruning and boosting. Next, we trained a support Vector Machine (SVM) using the SMO algorithm. For our SVM, we used a standard linear kernel due to the linear nature of our features. Finally, we tested two types of Bayesian classifiers: the simple to train naïve Bayes classifier and the more powerful Bayes net classifier. Of note, all our features besides the Grantham matrix score do not have any immediately known relationships and accordingly the Bayes net classifier performed similarly to naïve Bayes classifier.

For each of these classifiers, we trained a model using our cancer-associated SNP training set described in section 2.1, performed 10-fold cross-validation on each model, and use ROC curves to evaluate the performance of each model. We evaluated each model's performance and ultimately chose the naïve Bayes classifier to identify cancer-relevant SNPs in the recently sequenced genome of a male of European decent, hereafter referred to as P0 [9].

2.4. *Application to P0 SNPs*

Having trained a model for cancer-relevant SNPs, we next sought to identify potentially deleterious private non-synonymous coding SNPs in the P0 genome. To establish our subset of SNPs for classification, we first started with the set of all 2.7 million P0 SNPs as provided by the team that assembled the P0 genome. In order to remove population-common variants, we leveraged knowledge of common patterns of human genetic variation as established by the International HapMap project. We removed all P0 SNPs corresponding to SNPs established by the International HapMap project as having a population prevalence greater than 0.5%. Removal of population common variants left 1.9 million private SNPs. Next, non-coding SNPs were eliminated by mapping all private SNPs to the NCBI v36 human reference genome using the EnsEMBL core Perl API. SNPs that failed to map to coding regions of transcripts were removed, leaving 13,533 private coding SNPs. Finally, we computed variation effects of coding private SNPs using the EnsEMBL variation API. Synonymous SNPs were removed, leaving us with a final total of 6,718 private non-synonymous coding variations for classification. Post extraction, each SNP was classified using our trained naïve Bayes classifier and cancer-relevance was assessed.

3. Results

3.1. *Feature Evaluation and Classifier Performance*

3.1.1. Feature Selection

Using CFS, we found no correlations among our features and thus chose to use all of them when building our model. Results of our evaluation of feature predictive power are shown in Table 1. This evaluation revealed that Cell Cycle, Member of Cancer Pathway, and DNA repair were the most informative of our features. Other functional properties, such as tumor response and apoptosis, effectively had no predictive power because none of our training SNPs were functionally mapped to these categories via GO annotations.

Table 1. Results of classification using single features.

Feature	Accuracy	Specificity	Sensitivity	AUC
Stop Mutation	83.9483	1.0	0.093	.5421
Hydrophobicity Change	82.3105	1.0	0.0	.4998
Grantham Score	83.9891	1.0	.095	.5434
Polarity Change	82.3105	1.0	0.0	.4998
Member of Cancer Pathway	88.8617	.976	.482	.7239
Apoptosis	82.3105	1.0	0.0	.4998
Cell Cycle	90.1104	.981	.53	.7502
Tumor Response	82.3105	1.0	0.0	.4998
DNA Repair	89.1504	.983	.468	.7222

3.1.2. Model Evaluation

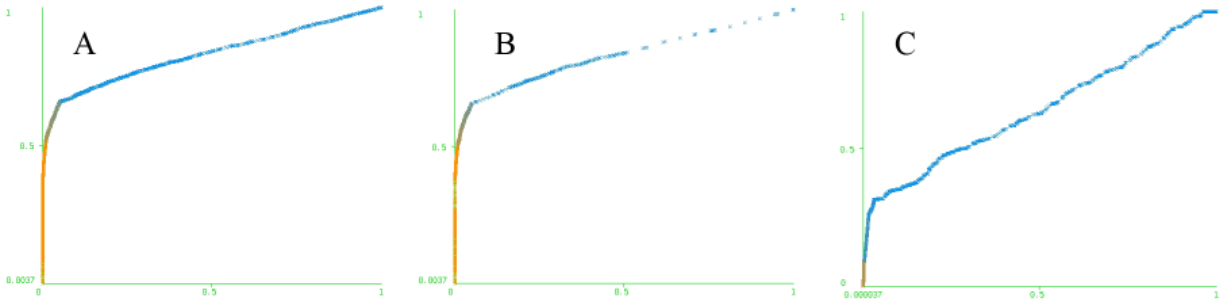
We performed 10-fold cross-validation on each one of our classifiers in training. The results are summarized in Table 2. The two Bayes methods performed best in AUC and sensitivity, while the decision tree and SVM were both more specific. We chose to use naïve Bayes as our final classifier, as it had the highest sensitivity among the classifiers we used, and the second highest AUC. The specificity was above 0.95 so we can be reasonably confident of our positive predictions. We determined the learning curves for several of our classifiers as a final test for bias and variance problems. Neither bias nor variance was significant issues for any of the classifiers.

Finally, as a sanity check, we classified HGMD diabetes mutations with a naïve Bayes classifier trained on our gold standard cancer SNPs. As expected, we have less than 10% sensitivity for predicting diabetes with our classifier, which demonstrates its specificity to cancer.

Table 2. Results of classifier performance using 10-fold cross validation.

Type	AUC	Accuracy	Specificity	Sensitivity
J48	.8026	.90813	.982	.564
SVM	.7622	.90346	.981	.544
Naïve Bayes	.8286	.90136	.963	.614
Bayes Net	.8299	.90496	.969	.607
Naïve Bayes - Diabetes	*.6424*	*.98115*	*.997*	*.087*

Figure 2. **A** and **B** show ROC curves for naive Bayes, Bayes net classifiers. **C** shows performance of naive Bayes classifier on diabetes data



3.2. *Identifying Novel SNPs*

Using our naïve Bayes classifier, we predicted with confidence greater than 95% that 210 private, non-synonymous SNPs in the P0 genome have potential cancer relevance. Of these 210 variations, 65% changed an amino acid-encoding codon to a stop codon, 5% occurred within known cancer pathway genes, 4% occurred within genes involved in cell cycle regulation, and 2% occurred within DNA repair genes. Additionally, of the 1,794 training SNPs extracted from HGMD, 32 were found to occur within the P0 genome, 10 of which we successfully re-predicted. Factors contributing to low recall of our gold standard SNPs are examined in the discussion section.

4. Discussion

Using our naïve Bayes classifier, we were only able to re-predict correctly 10 of the 32 SNPs that were included in both P0's genome and our gold standard for cancer-associated mutations; however, these results are actually quite promising given that our model has 96.3% specificity. The high specificity of our model gives great confidence that any mutation our classifier predicts as being cancer-associated is not a false positive, which is important in the context of identifying candidates for biological exploration. Pursuing any of these candidates further would require significant effort both in terms of time and funds, so ensuring the quality of these results is essential.

Looking in further detail at putative cancer-relevant private SNPs, it is straightforward to hypothesize about the detrimental effects of these mutations and how they might contribute to cancer susceptibility. A mutation on Chromosome 14 at base 53,487,272 changes an encoded amino acid from valine to alanine in the Bone Morphegenetic Protein 4 gene (BMP4) protein product. This mutation occurs within a transformation growth factor domain and could potentially alter the protein's ability to control cell proliferation and differentiation. Furthermore, other mutations in BMP4 have been shown to cause defects during eye, brain, and digit development [10].

Another mutation on chromosome 17 at base 42,589,359 changes a leucine-encoding codon to a stop codon in the Cell Division Cycle Protein 27 (CDC27) gene product. This mutation reduces the encoded peptide to 254 amino acids from 824 and eliminates the presence of a tetratrico peptide repeat domain at the C-terminus of the peptide, thus altering the protein product to properly bind with other proteins and accordingly, mediate cell division. Such results give us confidence that variations identified as deleterious by our method are biologically relevant.

Through analysis of our selected features and their predictive power for cancer susceptibility, we found that there appears to be a tradeoff between the information content of a feature and its ease of obtainment. We can make the best predictions when we know whether or not the mutation was part of a transcript involved in cell cycle regulation or DNA repair or was a member of a known cancer pathway because those three features had the highest accuracy and sensitivity when they were used individually to predict cancer risk. We attribute the high information content of these features to the central nature of these processes to all cancers. Unfortunately, because these features are dependent on human annotation, they were inconsistent in availability due to incomplete genomic annotation.

Surprisingly, some of the features that we hoped would be informative did not contribute to the accuracy of our classifier. Those features which have an AUC <= 0.5 when evaluated individually essentially provide the same amount of information as random chance, so mutations involved in tumor response and apoptosis were not any more likely to be predicted as cancer-associated than mutations without these functional properties. On closer examination however, we found that this lack of predictive power occurs because there are almost no SNPs associated with either of these two properties in our training data sets; only one of our training cancer SNPs is associated with apoptosis and none of our training cancer SNPs were associated with tumor response. We do not know for sure whether these features could improve our results given more complete gene annotation, but based on results for our other functionally defined features, we believe that they are still promising candidates for prediction.

5. Conclusions and Future Work

Despite inherent data biases, our method successfully leverages the entire genome in order to provide personalized, highly specific analysis of an individual's private mutations. Our method allowed us to identify 210 putative cancer-relevant SNPs in the P0 genome that are similar to our training set in terms of the defined feature set. These high confidence predictions serve as an excellent starting point for further biological exploration.

Functional associations and other cancer-specific features had among the highest information content, but their predictive powers were hampered by incomplete annotations. Future iterations of our method will could avoid such biases of manual curation by incorporating more complex yet universally computable features such as SNP presence within conserved structural domains.

Finally, given the 210 putative cancer-relevant SNPs, there is currently no way for us to quantify risk conferred by each variation given our available training data. To enable such analyses in the future, models of cancer-relevant SNPs should be partially derived from variations present in the genomes of well-documented cancer patients. Given the technology used to sequence the genome of P0, such data should be available within a few years. Much in the same way that full genome sequencing has enabled more specific, personal analyses of variation versus disease, so too will it enable better models of deleterious variation through the availability of full genomic sequences of diseased patients.

References

1. Goldstein DB. Common genetic variation and human traits. *N Engl J Med* 2009; 360:1696-8.
2. Stenson PD, et al. The Human Gene Mutation Database: 2008 update. *Genome Med* 1(1):13, 2009
3. Yip, Y.L, et al. Retrieving mutation-specific information for human proteins in uniprot/swiss-prot knowledgebase. *Journal of Bioinformatics and Computational Biology*, 2007, 5(6): 1215-1231.
4. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat. Genet.*. May2000;
5. T. J. P. Hubbard, et al. Ensembl 2009. *Nucleic Acids Research* 2009 37 Database issue:D690-D697
6. Kanehisa, M. et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28, 27-30
7. Grantham R. Amino acid difference formula to help explain protein evolution. *Science* 1974;185:862–4
8. Mark Hall, et al. The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, Volume 11, Issue 1.
9. Dmitry Pushkarev et al. Single-molecule sequencing of an individual human genome. *Nature Biotechnology* 27, 847 - 850 (2009)
10. Bakrania et al. Mutations in BMP4 cause eye, brain, and digit developmental anomalies: overlap between the BMP4 and hedgehog signaling pathways. *Am J Hum Genet*. 2008 Feb;82(2):304-19