Predicting Type 1 Diabetes Status From Genetics

Sara Hillenmeyer

December 15, 2011

1 Introduction

Genome wide association studies (GWAS) have identified many single nucleotide polymorphisms (SNPs) that are associated with disease. GWAS typically compare allele frequencies of $\sim$ 500K individual SNPs in case patients to those of control patients using single-feature statistics and hundreds or thousands of patients to gain enough power to overcome the serious multiple testing hypothesis correction burden. The original Wellcome Trust Case Control Consortium (WTCCC) paper published 24 high-confidence SNPs that are associated with 7 major diseases in a study with 14,000 cases and 3,000 shared controls ([1])

Currently, identifying SNPs associated with disease is motivated by two goals: first, to find a parsimonious panel of testable biomarkers for disease, and second, to learn about the biology of disease by examining which DNA markers are correlated with incidence. Given that the cost of genome sequencing is dropping exceptionally fast, and the cost of genotyping tag SNPs on a SNP-chip is even cheaper, the motivation for this first goal is waning—in the future, physicians will not have to prioritize which SNPs to genotype, but will have access to whole-chip or sequencing data for each patient. Additionally, it is becoming increasingly clear that multiple genetic loci, and the interactions between them, are associated with disease states, and the individual-SNP approach does not capture the biological intricacies of many disorders. In this report, I suggest that we can build whole-chip classifiers for disease, with selected features and interaction terms that will help elucidate the biology of disease. I also hope to show that using all of the data results in a better clinical predictor than the state-of-the-art predictors that include only the top individually associated SNPs.

2 The Data

I downloaded the GWAS data from the Wellcome Trust. In my initial work, I have analyzed the 58C controls versus Type 1 Diabetes (T1D) cases.

Abbreviation:	Description:	Participants:
T1D	Type 1 Diabetes	1963
58C	1958 Birth Cohort (control)	1480

2.1 Featurization

For each patient, I have genotype data from 447,221 SNPs. To convert this genotype data into a feature vector, I first defined the *minor allele* to be the allele at each SNP locus that was less common in the whole population (cases and controls). Then, for each person, I counted the number of minor alleles that they had at each position. The final feature vector for each patient i is $\{x_1^{(i)}, x_2^{(i)}, , x_{447221}^{(i)}\}$ where $x_j^{(i)} \in \{0, 1, 2\}$.

Genotype:	Minor Allele:	Score:
AA	G	0
AG	G	1
GG	G	2

2.2 Feature Ranking

Since using all 447,221 SNPs from the original GWAS study would both result in over fitting, and be exceptionally computationally intensive, I performed a preliminary feature ranking. For each training set in my 5-fold cross validation, I examined the SNPs individually, and calculated the chi-squared test statistic for independence. To do this, I calculated the expected frequencies of each genotype (0,1,2) in cases and controls given the marginal totals. Then, I compared the observed frequencies to these expected values. I ranked each the SNPS from largest to smallest C, where $C = \sum_{cells}((observed - expexted)^2/expected)$. For further analyses, I used the top r SNPs, where $1 \leq r \leq 102,400$, though I only report up to $r = 3,200$ in this document. These top SNPs represent the strongest individual features that discriminate between cases and controls in the training set.

3 Naive Bayes

Using the feature vectors described above, I implemented a multinomial-event model Naive Bayes classifier using Laplace Smoothing. I performed 5-fold cross validation (the SNPs were ranked and selected on the training data). See Table 1 for results from the cross validation. The % error is lowest for the top-10-feature model (6.3%) and levels off at around 19% when more features are included.

4 SVM with Linear Kernel

After Naive Bayes, I implemented an SVM with no kernel ($u'v$). I used 5-fold cross validation again, and varied the cost parameter from 10^{-4} to 10^4 by factors of 10 to create ROC curves for each classifier. In Figure 1, the x-axis is 1-Specificity ($1-(TrueNegatives/(TrueNegatives+FalsePositives))$) and the y-axis is Sensitivity ($TruePositives/(TruePositives+FalseNegatives)$). The top feature alone achieved an area under the ROC curve (AUC) of 0.65. Adding more features increased the AUC. The top 5, 10, and 25 features all had AUC $\approx$ .92. This is a great improvement over the single-feature model implemented here and other single-feature models reported in the literature.

5 SVM with Polynomial Kernel

To get interaction terms in the SVM, I added a polynomial kernel of the form: $(\gamma * u'v)^d$. I used γ =1/(num features) and did not vary γ during my cross validation and ROC building trials. Even without optimizing γ, the polynomial model achieved very high area under the curve, as is shown in Figure 2. Given the number of terms in these models, and the relatively small number of training examples, I suspect that these models are over fitting the data, and would not perform well on an external validation sample.

6 Conclusion

The combination of smart pre-ranking of the features and a support vector machine produced an excellent classifier for Type 1 Diabetes. The pre-ranking by chi-square test statistic provided an O(n log n) way to reduce the feature space such that it included only the most informative SNPs. This reduction allowed for much faster computation of the support vector machine. For the most part, the features selected in each round of the cross validation were the same, regardless of which random 20% of the data was left out.

The next step in this project is to try the best-fitting model on genotypes from a different cohort of patients. Additionally, since many of the top ranked features are SNPs that are implicated in other autoimmune disorders, I am curious about whether the model trained on Type 1 Diabetes has any predictive value on Rheumatoid Arthritis, Crohn's Disease, or other autoimmune disorders, though this application is less scientifically useful.

References

[1] Wellcome Trust Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007.

#Features	True Positives:	False Negatives:	False Positives:	True Negatives	% Error:
1	1963	0	1480	0	43.0%
5	1869	94	222	1258	9.2%
10	1895	68	150	1330	6.3%
25	1818	145	279	1201	12.3%
50	1826	137	331	1149	13.6%
100	1745	218	318	1162	15.6%
200	1700	263	295	1185	16.2%
400	1627	336	285	1195	18.0%
800	1589	374	278	1202	18.9%
1600	1581	382	278	1202	19.02%
3200	1576	387	281	1199	19.4%

Table 1: Naive Bayes Results

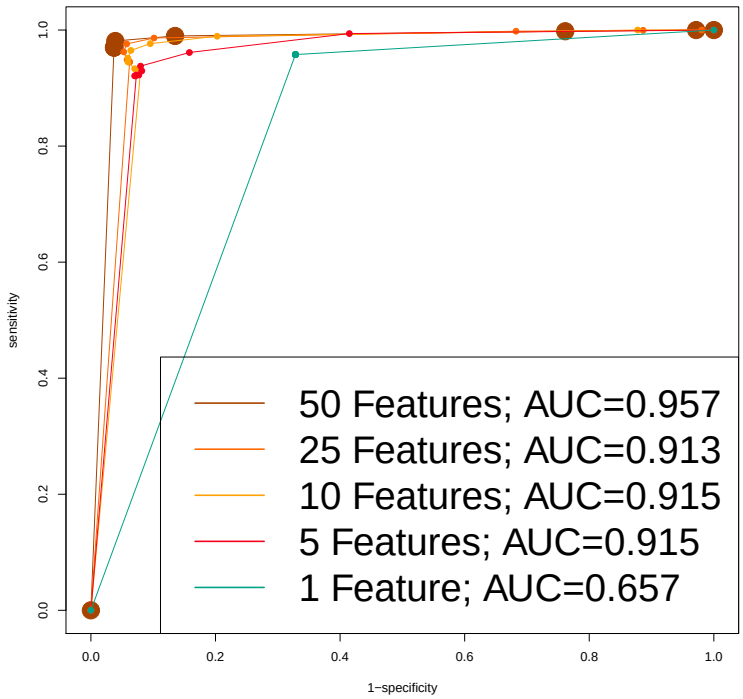

Figure 1: SVM with a linear kernal

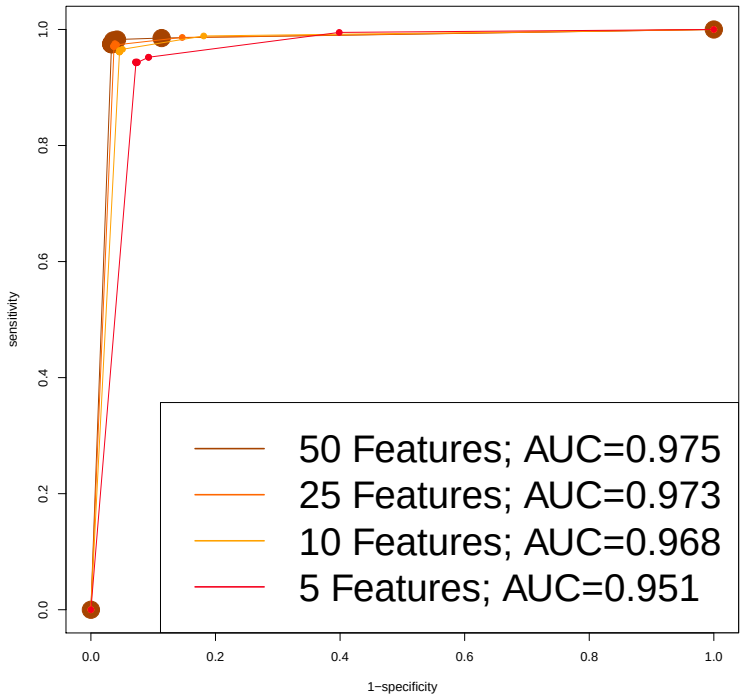

Figure 2: SVM with a 3-d polynomial kernel