Predicting Pathogenicity using Aberrant Splicing Predictions of Single Nucleotide Variants in Autism Spectrum Disorder

Bioinformatics Laboratory, BIMM 185

Kevin Chau

University of California, San Diego
16 June 2017

Abstract

Distributions of splicing likelihood scores for single nucleotide variants (SNVs) were analyzed with the hopes that these scores could be used to predict their pathogenicity, defined as a pathogenicity score multiplied by a risk score for that gene. SNVs implicated in autism as well as those found in control phenotypes were gathered and scored for their likelihood to cause cassette exon skipping and their pathogenicity. Gene coexpression networks were constructed for affected genes per eight developmental periods using publicly available RNA-seq data. Genes were scored for risk, per period, based on relative numbers of coexpression partners. A poserior probability function was calculated as the predictive model relating splice score to disease phenotype likelihood and was benchmarked for predicitve power.

1 Introduction

Prediction of pathogenicity of genetic variants is paramount to current understanding of diseases with heavy genetic influence and, by extension, development of treatments. In addition, it has been shown that aberrant splicing can have major contributions to expression of disease phenotypes, likely due to a propensity to disrupt biological pathways through alteration of gene products. Therefore, one might deduce that variants predicted to cause alternative splicing in genes central to biological processes are likely to result in the expression of the pathogenic phenotype. With numerous tools currently available to the public for academic use, a predictive model may be constructed with the hopes that the likelihood that a genetic variant causes aberrant splicing can be used to predict the contribution of that variant to development of a given disease. This knowledge could provide new insight into therepeutic targets at the gene or even isoform level.

2 Data and Processing

A list of variants classified by expressed phenotype was downloaded from DenovoDB, a database of nucleotide variants. Only variants tagged with the autism phenotype and control phenotype were downloaded. Overall, the taken from DenovoDB included the phenotype, the nucleotide variant, chromosome, affected gene, and starting position of the variant. This table was uploaded to a locally hosted MySQL database for simplified querying. Since only SNVs were analyzed, the dataset was filtered for only mutations that substitute one nucleotide for another. The SPIDEX splicing scoring tool was downloaded in order to score the nucleotide variants for their likelihood to cause cassette exon skipping.

2.1 Splicing Scores with SPIDEX

The SPIDEX tool was used to score each SNV for the likelihood that it causes cassette exon skipping. The scoring metric used was delta PSI (dpsi), the change in percent inclusion rate; negative values correspond to decreased inclusion (exon skipping) and positive values correspond to increased inclusion (exon retention). The criteria for scoring was that the SNV must have occurred within 300 nucleotides from a splice junction; that is, within 300 nu-

cleotides from an exon-to-intron or intron-to-exon transition position. Since not all variants fell within that distance, only scorable variants were retained for further analysis.

The SPIDEX tool is downloaded as a tab-indexed gzip-compressed file. The contents of the file was queried using the *tabix* utility provided by the htslib C library (formerly packaged with samtools). Example query and output follows:

```
tabix spidex_public_noncommercial_v1.0/
    spidex_public_noncommercial_v1_0.tab.gz chr1:0 | head -10 |
    cut -f1,2,3,4,5,7
chr1    861181  G       A       0.5983  SAMD11
chr1    861181  G       C       0.3903  SAMD11
chr1    861181  G       T       0.6564  SAMD11
chr1    861182  T       A       0.4921  SAMD11
chr1    861182  T       C       0.4723  SAMD11
chr1    861182  T       G       0.7517  SAMD11
chr1    861183  G       A       -0.0870 SAMD11
chr1    861183  G       C       -0.0874 SAMD11
chr1    861183  G       T       0.1979  SAMD11
chr1    861184  G       A       -0.1115 SAMD11
```

with fields as Chromosome, Position, Reference Allele, Mutant Allele, dPSI, and Gene.

2.2 Pathogenicity Scores with UMD Predictor

Pathogenicity of the given variants were scored with the web tool UMD Predictor. This online tool takes in a list of formatted variants and positions and returns a score for pathogenicity of the variant for each transcript affected. These scores ranged from 0 to 100, with 0 being non-pathogenic and 100 being confirmed pathogenicity. These scores were converted to percentages in order to more closely conform with the splicing scores.

2.3 Gene Networks and Risk Analysis

Gene coexpression networks were constructed in order to take into account gene risk factors, with the reasoning that genes with many coexpression partners are more likely to play an important role in their respective biological pathways; disruption in their splicing would thus lead to pathogenicity. These

coexpression networks were created using RNA-seq expression data from the BrainSpan database. A single network was developed for each of eight developmental periods, since expression values are likely to vary depending on the given stage of life. These developmental periods are as follows: 8 weeks postconception (PCW) to 12PCW, 13PCW to 18PCW, 19PCW to 23PCW, 24PCW to 37PCW, 0 months after birth (M) to 11M, 1 year (Y) to 11Y, 12Y to 19Y, and $21Y^{+}$. Pairwise comparisons were performed using the Pearson correlation metric and all gene-gene coexpression pairs with Pearson correlation coeffections of ≥ 0.7 were retained. The number of coexpression partners for each gene was calculated and divided by the maximum of the set, yielding the risk score for that gene. These scores were appended to all variants that affect that gene; the relevant variant information along with the calculated scores were all uploaded to a local MySQL database.

3 Data Analysis

The final dataset to analyze was compiled into a MySQL database with the following layout:

```
mysql> DESCRIBE scored_denovo_db;
```

Field	Type	Null	Key	Default
PrimaryPhenotype	**varchar**(125)	**NO**	MUL	**NULL**
Gene	**varchar**(125)	**NO**	MUL	**NULL**
Transcript	**varchar**(125)	**NO**	MUL	**NULL**
Chromosome	**varchar**(125)	**NO**		**NULL**
Position	bigint(15)	**NO**		**NULL**
Variant	**varchar**(500)	**NO**		**NULL**
SpliceScore	double	**NO**		**NULL**
PathogenScore	double	**NO**		**NULL**
P1Risk	double	**NO**		**NULL**
P2Risk	double	**NO**		**NULL**
P3Risk	double	**NO**		**NULL**
P4Risk	double	**NO**		**NULL**
P5Risk	double	**NO**		**NULL**
P6Risk	double	**NO**		**NULL**
P7Risk	double	**NO**		**NULL**
P8Risk	double	**NO**		**NULL**

From this data, distributions could be plotted. First, a kernel density estimate was performed for the frequencies of splice scores for both the autism phenotype and control phenotype.

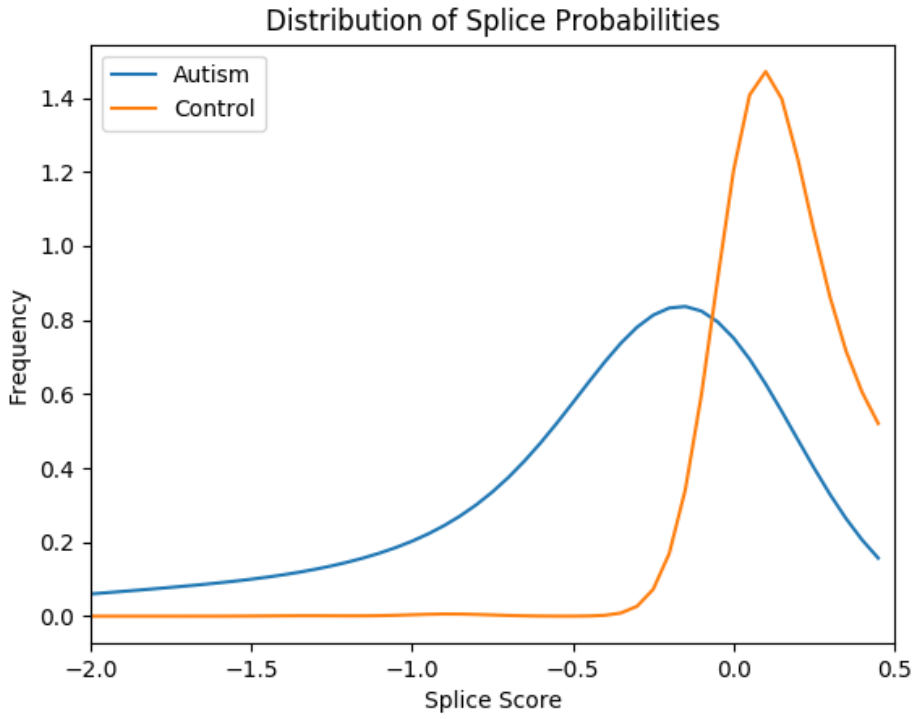

Figure 1: Kernel density estimates of autism SNV splice scores and control SNV splice scores

As shown, there seems to be a clear distinction between the distributions of splice scores between the autism phenotype and control phenotype.

Scatter plots between splicing scores and pathogenicity scores may also be drawn in order to illustrate any correlation between the two over each developmental period and in each condition.

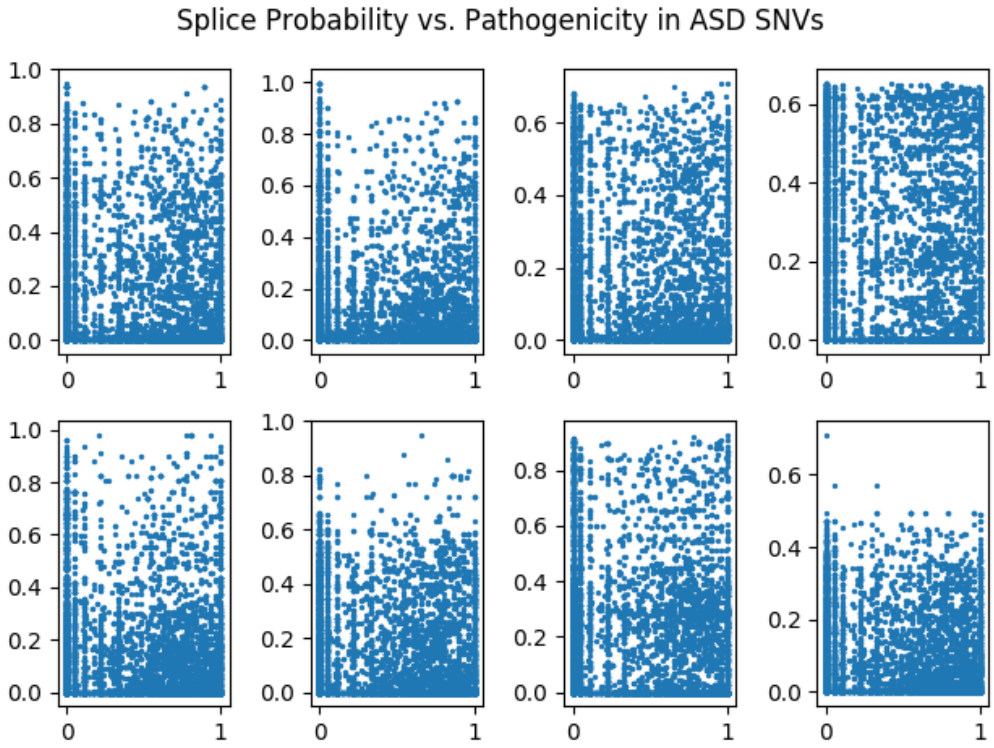

Figure 2: Splicing probabilities (x-axis) plotted against pathogenicity probabilities (y-axis) for each developmental period in SNVs implicated in autism

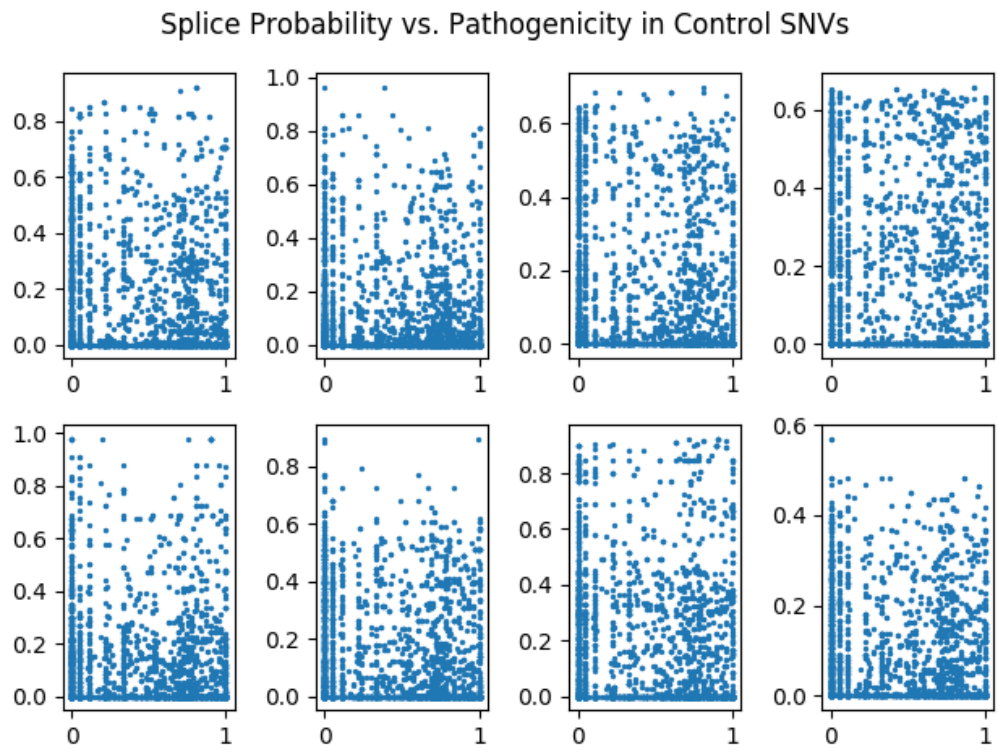

Figure 3: Splicing probabilities (x-axis) plotted against pathogenicity probabilities (y-axis) for each developmental period in control SNVs

Clearly, since there is no correlation between the UMD Predictor predicted pathogenicity scores, transformed by gene risk factors, and splice scores, this data cannot be used as a prior probability. Therefore, a posterior probability function is calculated using prior probability of 0.5; that is, an unweighted posterior probability function is created. Additionally, only a random sample of half of the scores was used to create the posterior probability function so as to better gauge the predictive power of the model through benchmarking tests utilizing the various attributes of confusion matrix values over a range of thresholding powers.

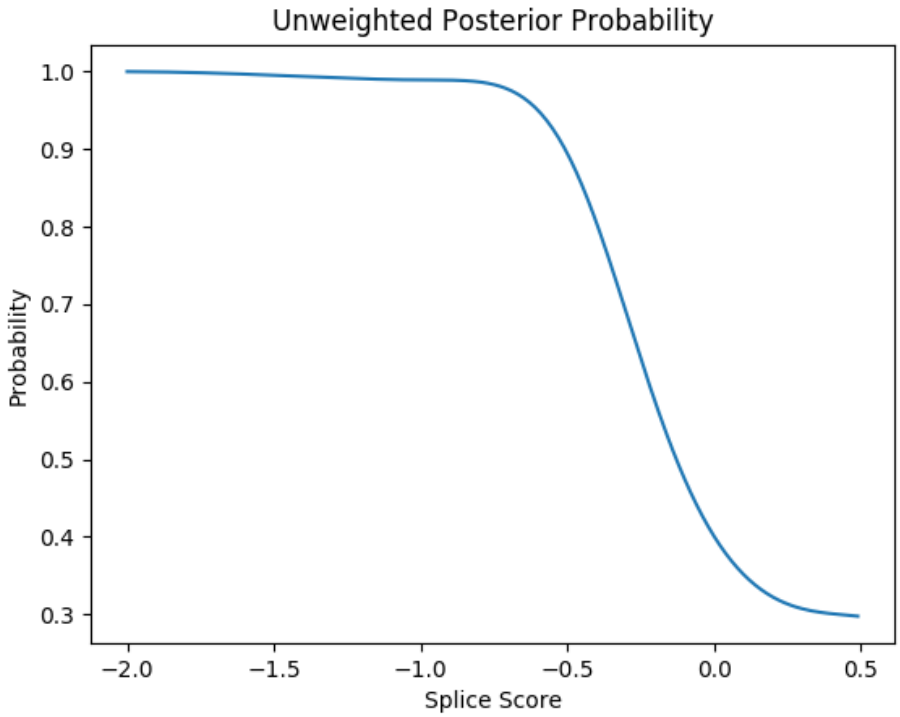

Figure 4: Posterior probability function

Benchmarking of the model was performed in order to gauge its predictive power. Sensitivity and specificity plots were drawn to illustrate the performance of the inference model.

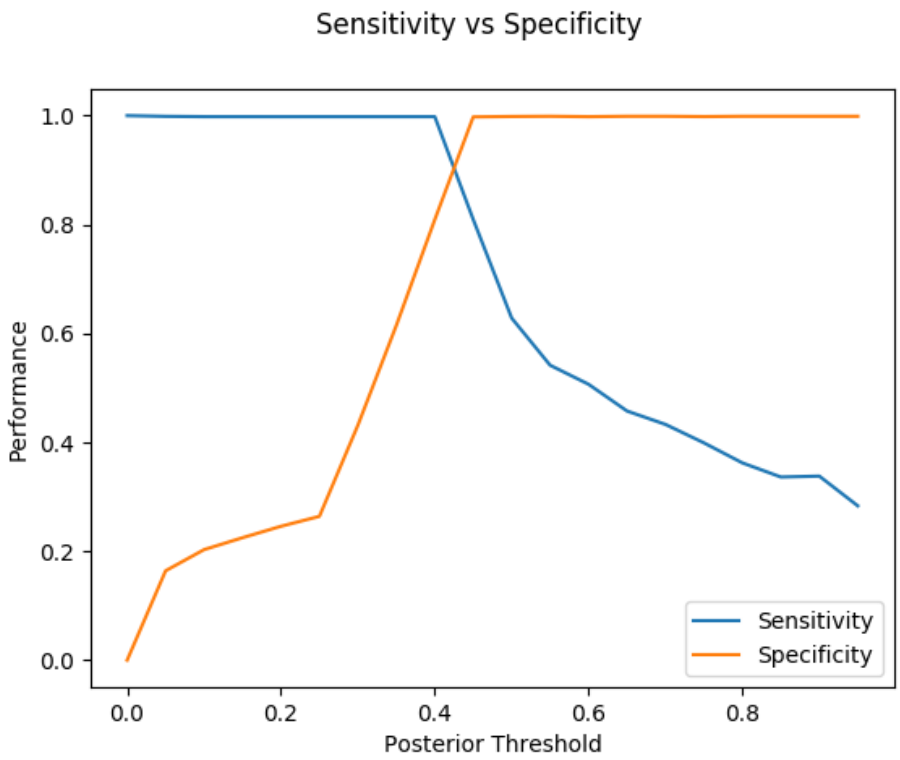

Figure 5: Sensitivity and specificity curves.

In addition, an accuracy curve was plotted to illustrate the relationship between the true predictions and false predictions

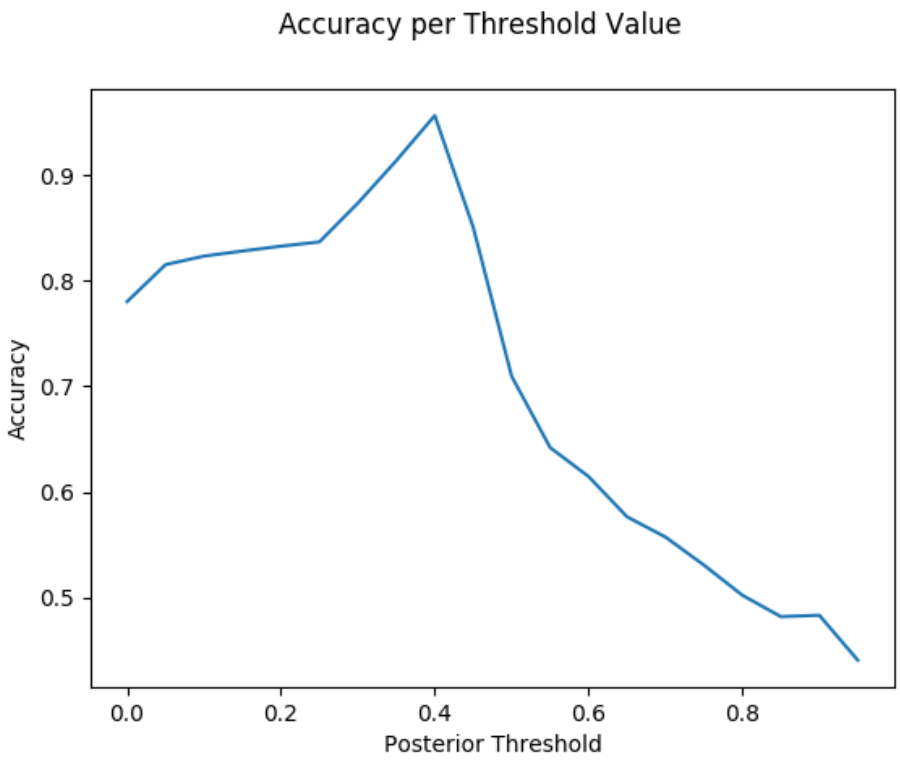

Figure 6: Accuracy curve

4 Discussion

Given the data as well as the benchmarking curves, the inference model proves to be remarkably accurate. However, one should note that the model proposed in this study is based on very specific criteria; only single nucleotide variants within three hundred nucleotides of a splice junction in brain-expressed genes that were able to be scored by the SPIDEX tool and were either implicated in autism spectrum disorder or were not associated with any other disease phenotype expression were considered out of the entire dataset. Additionally, the posterior function was calculated without any prior probabilities; the likelihood that a single nucleotide variant could be implicated in ASD versus the control phenotype were assumed to be the same. Appropriation of this probabilistic model to other datasets should take these notes into account.

5 Supplementary Materials

Relevant code:
https://github.com/kkchau/BIMM-185-FINAL-PROJECT/