Bioinformatics, 32(23), 2016, 3611–3618
doi: 10.1093/bioinformatics/btw498
Advance Access Publication Date: 9 August 2016
Original Paper

OXFORD

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/32/23/3611/2525615 by UNIVERSIDAD DE GUANAJUATO user on 21 February 2019

Data and text mining

Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates

S. Cogill and L. Wang*

Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 23, 2016; revised on July 20, 2016; accepted on July 21, 2016

Abstract

Motivation: Autism spectrum disorders (ASD) are a group of neurodevelopmental disorders with clinical heterogeneity and a substantial polygenic component. High-throughput methods for ASD risk gene identification produce numerous candidate genes that are time-consuming and expensive to validate. Prioritization methods can identify high-confidence candidates. Previous ASD gene prioritization methods have focused on *a priori* knowledge, which excludes genes with little functional annotation or no protein product such as long non-coding RNAs (lncRNAs).

Results: We have developed a support vector machine (SVM) model, trained using brain developmental gene expression data, for the classification and prioritization of ASD risk genes. The selected feature model had a mean accuracy of 76.7%, mean specificity of 77.2% and mean sensitivity of 74.4%. Gene lists comprised of an ASD risk gene and adjacent genes were ranked using the model's decision function output. The known ASD risk genes were ranked on average in the 77.4th, 78.4th and 80.7th percentile for sets of 101, 201 and 401 genes respectively. Of 10,840 lncRNA genes, 63 were classified as ASD-associated candidates with a confidence greater than 0.95. Genes previously associated with brain development and neurodevelopmental disorders were prioritized highly within the lncRNA gene list.

Contact: liangjw@clemson.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Autism spectrum disorder (ASD) is the umbrella term for the neurodevelopmental disorders: autistic disorder, Asperger's syndrome, pervasive developmental disorder not otherwise specified, Rett syndrome and childhood disintegrative disorder. It is generally diagnosed at an age greater than four years old based predominantly on the behavioral phenotype described as delayed communication, difficulty acknowledging social cues, and engaging in repetitive behaviors (American Psychiatric Association, 2012). In 2010, the Centers for Disease Control and Prevention (CDC) estimated the prevalence of ASD at 1 in 68 children aged 8, and this was an increase from the 2007 estimate of 1 in 150 (CDC, 2014). The increase in prevalence may be attributable to more public awareness and implementation of prescreening technology (Chlebowski *et al.*, 2013). Twin and sibling studies indicate that ASD etiology is influenced heavily by genetics and to a lesser extent environmental factors (Kim and Leventhal, 2015). The actual physiological cause of ASD is currently unknown, but leading theories include imbalance between excitatory and inhibitory synapses and substandard signaling between brain structures due to poor axonal growth (Fakhoury, 2015; McFadden and Minshew, 2013). It is possible that a myriad of genetic and environmental factors could lead to distinct physiological conditions convergent on the behavioral phenotype.

ASD is complex with hundreds of genes implicated in its etiology. The predominant focus of previous research has been on protein-coding genes, and there is the potential of more genes such

© The Author 2016. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

as non-coding genes being implicated as well. Disease gene identification studies are usually large-scale and high-throughput. Examples include genome-wide association studies (GWAS), copy number variation studies (CNV) and whole exome sequencing (WES). These studies in themselves are time-consuming and expensive especially when considering the sample size required for an effective study. The output can contain numerous potential candidate genes, which are also expensive and time-consuming to validate, with minimal impact on risk of the disease itself. GWAS studies have been shown to be particularly susceptible to weak SNP associations for ASD (Anney *et al.*, 2012). Disease gene prioritization systems seek to determine the confidence for disease association amongst gene lists. While disease gene prioritization methods have become somewhat ubiquitous, they generally do not have methodologies accounting for prioritization of non-coding genes.

Support vector machine (SVM) approaches have previously been applied to ASD research. Bruing *et al.* (2014) provided support for a phenotype-genotype relationship by using symptom profiles as features and genetic disorders as classes for a multi-class extension of SVM. Magnetic resonance image (MRI) offers one of the most potentially fruitful routes for ASD diagnostics, and SVM approaches have been applied to MRI data to classify and further characterize the morphology of the disorder. Retico *et al.* (2016) used SVM to determine differences in the morphologies between young male and female patients with ASD. A similar method was employed by Ecker *et al.* (2010) on whole-brain structural imaging to reveal a correlation for the distance from the decision boundary and the severity of the disorder.

In this study, we have developed a machine learning method to prioritize genes for ASD risk. Based on domain-specific knowledge of ASD, it is hypothesized that expression patterns offer a potential means of prioritization for all gene types for ASD risk. In particular, we design the novel approach of using the normal developmental brain expression patterns found in the BrainSpan dataset to leverage previous research focused primarily on protein-coding genes to classify ASD risk gene candidates. The validity of this approach is supported by weighted gene co-expression network analysis on the BrainSpan dataset, which has shown convergence of ASD risk genes on developmental pathways (Parikshak *et al.*, 2013). It is further supported by evidence of the potential role in ASD of lncRNAs, whose function is closely linked with their expression patterns (Derrien *et al.*, 2012; Necsulea *et al.*, 2014; Ziats and Rennert, 2013). In this study, we first generated a gene list of high-confidence developmental ASD risk genes. We were then able to create a support vector machine (SVM) model for ASD risk gene prediction with a 76.7% accuracy capable of prioritizing ASD candidate genes based solely on expression patterns in the developing brain. Utilizing a wrapper methodology and a best-first search method during feature selection, we were able to drastically reduce the dimensionality and identify biologically relevant and novel temporospatial features within the dataset. The performance of the feature subset showed improvement over the full feature set. To further test our model, we used the ASD risk gene list to generate hypothetical loci similar to what would be expected from an association study. The genes within the loci were prioritized to determine the relative rank within the list of the known risk gene. Finally, the model was applied to the prioritization of long non-coding RNA (lncRNA) genes. Overall, the study demonstrates the effective application of a machine learning approach to ASD risk gene identification using normal tissue expression patterns.

2 Materials and Methods

The machine learning problem in this study can be defined in the following way: genes serving as instances are to be classified for autism spectrum disorder (ASD) risk using their respective expression profiles which serve as the feature set. A model for this decision would allow the prioritization of gene lists based on the strength of predicted ASD associations. Using known ASD risk genes and non-ASD genes with their expression profiles, we seek to perform supervised training of a model.

2.1 Datasets

The BrainSpan Atlas of the Developing Human Brain is a developmental transcriptome dataset compiled by a consortium consisting of the Allen Institute for Brain Science and five collaborating universities (Hawrylycz *et al.*, 2012). The dataset consists of 524 samples with a developmental time point range from 8 weeks post-conception to 40 years of age from 26 brain structures. While the dataset demonstrates a lack of availability for multiple samples at each temporospatial time point in development, the BrainSpan dataset is currently the most comprehensive transcriptome of the human developing brain. Expression values were RNA-sequencing reads that were assembled and aligned using the GENCODE consortium's annotation release v10 (Harrow *et al.*, 2012). They were in the units of Reads Per Kilobase of transcript per Million mapped reads (RPKM). A $\log_2(RPKM+1)$ transformation was applied to the data. Genes in the dataset were instances, and their expression values for the temporospatial time points acted as features for the training dataset.

To build a model for ASD risk gene classification, negative and positive gene instances were required. While many genes can be considered non-ASD, two considerations were made to enhance potential model performance. Genes associated with diseases unrelated to the disease being studied have previously been used as negative controls in prioritization studies in an effort to reduce potential systematic bias (Erlich *et al.*, 2011; Moreau and Tranchevent, 2012; Thienpont *et al.*, 2010), and here we employ that same methodology by using non-ASD disease-associated genes as our negative instances. Since many individuals afflicted with ASD are also diagnosed with some form of intellectual disability (ID) (Bakken *et al.*, 2010; Hoekstra *et al.*, 2009), and there is considerable overlap between ID and ASD-associated genes (Pinto *et al.*, 2010), ID genes were not among the negative instances. The positive instances were ASD risk genes compiled from the Simons Foundation Autism Research Initiative Gene database (Abrahams *et al.*, 2013), AutismKB (Xu *et al.*, 2012), and De Rubeis *et al.*'s (2014) large exome sequencing study for *de novo* mutations in individuals with ASD. To curate for developmental ASD risk genes, the top 85% of the genes based upon expression variance within the BrainSpan dataset were used (Supplementary Table S1).

2.2 Support vector machines

Support vector machine (SVM) is a supervised machine learning algorithm that is effective for high-dimensional datasets comprised of real numerical as opposed to categorical values. It is commonly used for biological classification problems (Cortes and Vapnik, 1995; Kourou *et al.*, 2014). Genes within our training dataset are vectors defined as $x_i \in R|0 \leq x_i \leq 1$ (after normalization) for $i = 1, \ldots, l$, where i is a temporospatial feature in the BrainSpan dataset and l is the size of the feature vector. The classification of each gene, non-ASD or ASD, is defined here as $y_i \in \{-1, +1\}$. When the model is trained, the algorithm seeks to maximize the distance between

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/32/23/3611/2525615 by UNIVERSIDAD DE GUANAJUATO user on 21 February 2019

margins for a decision boundary separating the positive and negative instances in hyper-dimensional space. The margins are determined from a subset of the total instances nearest in Euclidian distance to the decision boundary referred to as support vectors. The distance between margins is defined as $2/\|\omega\|$ where ω is a vector orthogonal to the decision boundary such that its dot product with a support vector is zero. The sign of the decision function (positive or negative) is used for binary classification. Classification problems generally require richer feature space than what is defined originally within the dataset to separate the variables. Plotting vectors in higher dimensions is computationally expensive, but application of a kernel function allows for model optimization in higher dimensional space without having to plot the points. Popular kernel methods for SVM models include radial basis, linear, polynomial and sigmoidal, and for our initial feature selection and final model after optimization, we used the radial basis function (RBF) kernel:

$$K(x \cdot x') = \exp(-\gamma\|x - x\prime\|^2) \quad (1)$$

The parameter γ determines the 'smoothness' of the decision boundary. For this study, we use the SVM SVC class from the Scikit-learn Python library (Pedregosa *et al.*, 2011) for all SVM implementations with the exception of the feature selection process where the libSVM package from the WEKA data mining software was used (Hall *et al.*, 2009).

The training dataset was imbalanced, consisting of 366 ASD risk genes as positive instances and 1762 non-ASD disease genes as negative instances, a ratio of 1:4.8. For model construction, there are methods of balancing the dataset such as oversampling using synthetic minority over-sampling technique (SMOTE) (Chawla *et al.*, 2002) and randomized under-sampling of the majority class (Kubat and Matwin, 1997). However, adjusting class weights within the parameters of the learning algorithm may offer an optimal solution. The class weights balance the misclassification cost in establishing soft margins (see Section 2.2.2) without altering the underlying data space. In this study, we have empirically demonstrated that there is no performance loss in using the class weight parameter versus randomized under-sampling of the majority class for optimized SVM models with the full feature set (Supplementary Table S2) based on 50 repetitions of tenfold cross-validations.

The SVM algorithm was chosen over other machine learning algorithms due to its effectiveness in producing a more generalized model through its maximization of the decision boundary and not the minimization of training errors (Yang, 2004). It has been shown to outperform many other learning algorithms on various biological problems, including prediction of proteolytic cleavage (duVerle and Mamitsuka, 2012), DNA-binding residues in proteins (Si *et al.*, 2015; Wang and Brown, 2006) and linear B-cell epitopes (Wang and Pai, 2014). It is favorable due to its low computational cost for training a model and easily optimized parameters. The SVM algorithm also has the benefit of a function output, which allows for gene prioritization (see Section 2.3). To verify its suitability for ASD risk gene prediction, we compared the performance of the weighted SVM model to other commonly used machine learning algorithms (Supplementary Table S2). The dataset used for this analysis was balanced through randomized under-sampling of the majority class, and the performance evaluation was the same as outlined in Section 2.2.2 for 50 repetitions of tenfold cross-validations.

2.2.1 Feature selection

Gene expression datasets such as the BrainSpan dataset can have high dimensionality. The feature selection process removes redundant and irrelevant features to improve model performance, reduce computational load, and decrease the ratio of features to samples, which reduces the probability of overfitting. Wrapper methods evaluate feature subsets in the context of the learning algorithm. Given that these methods take into account the learning algorithm that is to be optimized, wrapper methods generally perform better than filtering or embedded techniques especially for gene expression data, but they are computationally expensive (Hira and Gillies, 2015) and can potentially lead to overfitting (Saeys *et al.*, 2007). To address potential overfitting in this study, the approach of candidate gene prioritization (see Section 2.3) has been employed as external validation in lieu of an independent test set. In addition to model performance improvement, the methodological approach of the wrapper method allows us further knowledge discovery in evaluating the performance of the model with the addition of each new feature. This is of particular interest in this study given that the features represent temporospatial points in brain development and the molecular etiology of ASD is still unclear.

The number of potential feature subsets for a brute force search is equivalent to a power set or 2^N, where $N = 524$ for our dataset. This is infeasible. Deterministic feature selection methods such as the sequential forward selection (SFS) method, which incrementally adds features in a greedy hill-climbing search, have been used successfully in cancer machine learning studies with expression datasets (Kourou *et al.*, 2014). In a greedy hill-climbing search, for a machine learning algorithm such as SVM, a feature set of size n and a feature subset F,

$$f_{\mathrm{SVM}}(F) \text{ where } F \subseteq \{f_1, f_2, \ldots, f_n\} \quad (2)$$

$$f_i \in \{f_1, f_2, \ldots, f_n\}$$

the performance of the classifier is measured by a scoring metric. For this study, overall accuracy was used (TP = true positive, TN = true negative, FP = false positive and FN = false negative):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

After testing multiple performance measures, overall accuracy was determined to be optimal for searching subsets. It showed steady increases and converged on a subset with minimal processing time. The subset is then built in the following manner where t is representative of iterations:

$$\text{if } f_{\mathrm{SVM}}(F^{t-1} \cup \{f_i\}) > f_{\mathrm{SVM}}(F^{t-1}) \text{ then Set}: \; F^t = F^{t-1} \cup \{f_i\} \quad (4)$$

$$\textit{if } f_{\mathrm{SVM}}(F^{t-1} \cup \{f_i\}) \leq f_{\mathrm{SVM}}(F^{t-1}) \text{ then Feature Set} = F$$

While SFS is an effective wrapper search method, it does present the possibility of local maxima. One way to partially alleviate this while maintaining a heuristic search is to allow for hill traversal. The best-first search algorithm implements a greedy hill climbing algorithm but allows for backtracking and expansion of previously evaluated nodes. It has also been shown to outperform the greedy algorithm (Kohavi and John, 1997). In this study using the WEKA data mining software, we searched the feature subset using the best-first search algorithm with the overall accuracy from a fivefold cross-validation using libSVM at default settings as the performance measure (Hall *et al.*, 2009). The best-first search was forward starting from an empty feature set and allowed for the expansion of five non-improving nodes.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/32/23/3611/2525615 by UNIVERSIDAD DE GUANAJUATO user on 21 February 2019

2.2.2 Model parameter optimization and performance evaluation

To improve SVM classification performance with the full and selected feature sets respectively, we optimized the three parameters, cost (C), γ and kernel using a grid search approach, which evaluates all combinations based on a performance measure. Parameter optimization for high-dimensional datasets can be sensitive to class imbalance if overall accuracy is used as the performance measure. Optimized parameters can favor a model with low sensitivity where positive instances are in the minority such as ASD genes. Therefore, we used G-mean (geometric mean) as the performance metric, which measures the classifier's ability to balance specificity and sensitivity (Lin and Chen, 2012):

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

$$G_{\text{mean}} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{7}$$

C is the penalty assigned for misclassifications. In the maximization of $2/\|\omega\|$, which alternatively is the minimization of $\|\omega\|^2/2$ during training of the SVM model, when misclassifications are allowed, the problem takes on the form,

$$\min_{\omega,\varepsilon}\{\frac{1}{2}\|\omega\|^2 + C\sum\nolimits_i \varepsilon_i\} \tag{8}$$

where ε is the quantification of the misclassification. The γ parameter can be found in Equation (1). Previous work has shown that the most efficient method of optimization for C and γ is to exponentially increase the values across a range (Hsu *et al.*, 2003). In this study, we used $C \in \{2^{-5}, 2^{-3}, \ldots, 2^{13}, 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, \ldots, 2^{1}, 2^{3}\}$. Using the radial basis function (RBF) kernel, we measured the performance for each parameter pairing of C and γ. Using the linear kernel, we measured the performance for each C. The optimal kernel, C and γ parameter combinations were selected based on the highest G-mean returned from tenfold cross-validations. Model performance was evaluated using sensitivity (Equation 5), specificity (Equation 6), overall accuracy (Equation 3) and the Matthews correlation coefficient (MCC), which measures the correlation between the predicted and actual classifications on a scale of $MCC = R| - 1 \leq MCC \leq 1$ (Matthews, 1975):

$$\text{MCC} = \frac{(TP + TN) - (FP + FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{9}$$

2.3 Candidate gene prioritization

To rank candidate genes, we used the output from the SVM model with greater output corresponding to higher rank. To test the ability of the model to prioritize ASD risk genes, we generated gene lists containing at least one ASD risk gene. This methodology was adapted from the work by Piro *et al.* (2010). For each ASD risk gene in the training dataset, we identified N flanking genes on the same chromosome using the GENCODE release v10 annotation. We then constructed a hypothetical locus with $2N + 1$ genes centered on the ASD risk gene. If an ASD risk gene was close to a chromosome terminal, the number of genes in the opposing flank were extended to ensure that each gene list was of equal length. Three gene list lengths were tested: 101, 201 and 401. Genes present in the training dataset were removed from each gene list. The SVM model was trained for each hypothetical locus, and the candidate gene list was prioritized. Model performance was evaluated by the percentile rank assigned to the known ASD risk gene in its respective candidate gene list:

$$\text{Percentile Rank} = \frac{L}{N} \times 100\% \tag{10}$$

In the above equation, L is the number of SVM scores less than the target, and N is the total number of candidate genes.

2.4 Long non-coding RNA gene candidate prioritization

The 10 840 lncRNA genes in the GENCODE release v10 were prioritized using the SVM model to further demonstrate its capabilities and performance. The model was built with all instances from the training dataset, the feature subset from feature selection, and the optimized parameters. Confidence measures of the classification (ASD or non-ASD associated) were assigned for each gene. If the instance was classified as positive, the confidence was (1 – false positive rate), and if the instance was classified as negative, the confidence was (1 – false negative rate) (Wang and Brown, 2006).

3 Results

3.1 Support vector machine classification of ASD risk genes

Table 1 shows the performance of the support vector machine (SVM) classifier for 50 repetitions of tenfold cross-validations using all 524 features available in the dataset. The model was optimized on the G-means performance measure and used the radial basis function (RBF) kernel with a cost (C) = 8 and $\gamma = 0.0078125$. The SVM model with the full feature set achieved a mean accuracy of 0.739 with 0.748 sensitivity, 0.737 specificity and 0.385 Matthews Correlation Coefficient (MCC). The receiver operator characteristic (ROC) curve for ASD risk gene prediction using the full feature set is shown in Figure 1. The ROC curve is a plot of the true positive rate (sensitivity) versus the false positive rate (1 – specificity) for varying output thresholds of the SVM classifier (Hajian-Tilaki, 2013), and in this study it was generated using the ROC class from the Scikit-learn Python library (Pedregosa *et al.*, 2011). The ROC curve and the area under the curve (ROC-AUC) are considered to be the most robust measures of model performance. The ROC-AUC for the full feature set is 0.8045. This value is significantly greater than the random guess value of 0.5. Given the novelty of the study, there are no real means of comparison to other model performance, and therefore the SVM model performance can serve as a benchmark for future models. The heterogeneous nature of ASD warrants against over-optimization as the probability of overfitting may increase, and the current performance appears to be indicative of an effective generalized model.

Table 1. The mean sensitivity, specificity, overall accuracy and Matthews Correlation Coefficient (MCC) of each model for 50 repetitions of tenfold cross-validations

	Full feature set	Selected feature set
Sensitivity	0.748 ± 0.006	0.744 ± 0.005
Specificity	0.737 ± 0.003	0.772 ± 0.002
Accuracy	0.739 ± 0.003	0.767 ± 0.002
MCC	0.385 ± 0.003	0.419 ± 0.005

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/32/23/3611/2525615 by UNIVERSIDAD DE GUANAJUATO user on 21 February 2019

3.2 Wrapper method with best-first heuristic search for feature selection

Reduction of the dimensionality for a dataset decreases computational load and the probability of overfitting. Here we demonstrate SVM model performance gain with a selected feature subset over the full feature set. Our methodology applied a forward heuristic best-first feature selection search using a wrapper method. Forward searches begin with empty sets and build them incrementally with the addition of one feature at a time, and this approach allows for evaluation of each feature added to what will become the final subset. The best-first search is a modification of the greedy stepwise method and allows for backtracking if no improvement is seen with further feature additions. Table 2 shows the incremental building of the feature subset. The features are in the order of their additions to the subset, and the overall accuracy is listed for each resulting subset. For example, the feature subset of the primary somatosensory cortex (area S1, areas 3,1,2) (S1C) at 13 post-conception weeks (pcw) had an accuracy of 0.662, and the subset of S1C at 13 pcw and the dorsolateral prefrontal cortex (DFC) at 8 pcw had an accuracy of 0.694. The selected feature model was evaluated with the same parameters as the full feature set model. The specificity of 0.772, accuracy of 0.767 and MCC of 0.419 are all significantly improved over the model with the full feature set, and there was no significant difference in the specificity (Table 1). The selected feature model used the RBF kernel with $C = 32$ and $\gamma = 0.03125$. Its ROC-AUC of 0.8194 shows a performance improvement over the full feature model (Fig. 1).

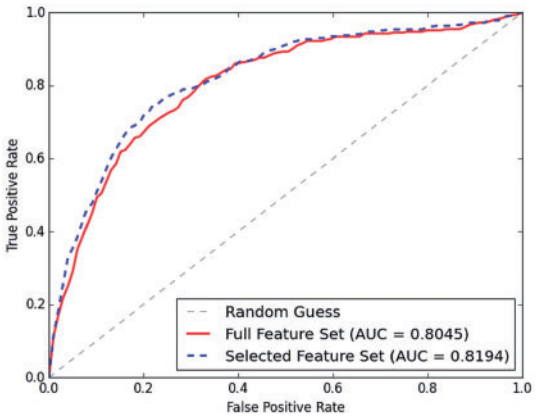

Fig. 1. ROC curves of the selected and full feature set SVM models. The AUCs for the ROC curves are given in the legend

3.3 Prioritization of ASD risk gene candidates using SVM model output

Given the relatively low number of high-confidence ASD risk genes and the heterogeneity of the disorder, all of the positive instances were used in the training of the model to maximize the available data space. This however precludes the use of an independent test dataset as a means of external validation. As an alternative to the use of an independent test dataset, both models were evaluated on their ability to prioritize hypothetical loci. We generated gene lists for each ASD risk gene and its surrounding genes, and model performance was measured by the ability to highly prioritize the known ASD risk genes. While the lists may contain previously unknown ASD risk genes, a liberal estimate of ASD risk gene frequency in the genome at 2% (De Rubeis *et al.*, 2014) and varied location of CNV's (Liu and Takumi, 2014) would allow us to assume that a high-performing model would prioritize a known ASD risk gene within the 90th percentile. Since our interests are in the relative ranks of the genes for the hypothetical loci, we perform a form of ordinal regression in ordering by the decision function output. Here we compare the prioritization capability for the selected and full feature set SVM models.

Table 3 shows the mean percentile rank for known ASD risk genes in their respective hypothetical loci of varying sizes. Again, the selected feature model outperformed the full feature model. The selected feature model showed 2–3% greater mean percentile rank of ASD risk genes for each locus length than the full feature model.

The mean percentile ranks increase with the size of the loci, which is to be expected given the assumption that ASD risk genes highly prioritized would remain so in an expanding list. Figure 2 shows the distributions of ASD risk genes grouped by percentile rank for the two models. There were little to no genes ranked in or near the 50th percentile for the two models. ASD risk genes were either ranked in the low or high percentiles, and the amounts in each are consistent with the overall accuracy of the models. ASD risk genes classified as positive instances were predominantly in the 95th percentile or above. This is consistent with expectations and

Table 2. The selected features from the best-first search algorithm

Developmental time point	Structure	Accuracy	Sensitivity	Specificity
13 pcw	Primary somatosensory cortex (area S1, areas 3,1,2)	0.661071	0.73918	0.644847
8 pcw	Dorsolateral prefrontal cortex	0.694323	0.703607	0.692395
9 pcw	Parietal neocortex	0.709699	0.721585	0.70723
37 pcw	Mediodorsal nucleus of thalamus	0.719803	0.70153	0.723598
1 yr	Dorsolateral prefrontal cortex	0.728092	0.745902	0.724393
4 yrs	Dorsolateral prefrontal cortex	0.746203	0.722787	0.751067
1 yr	Primary visual cortex (striate cortex, area V1/17)	0.747989	0.740929	0.749455
16 pcw	Orbital frontal cortex	0.750761	0.755628	0.74975
8 pcw	Orbital frontal cortex	0.754389	0.751694	0.754949
30yrs	Primary auditory cortex (core)	0.754474	0.754754	0.754415
8 pcw	Occipital neocortex	0.756053	0.752787	0.756731
40 yrs	Primary motor cortex (area M1, area 4)	0.756147	0.75153	0.757106
21 yrs	Inferolateral temporal cortex (area TEv, area 20)	0.758778	0.748798	0.760851
8 pcw	Hippocampus (hippocampal formation)	0.759596	0.751093	0.761362
8 yrs	Primary somatosensory cortex (area S1, areas 3,1,2)	0.759746	0.747978	0.762191

The features are described by the time point when the sample was collected and the brain structure where the sample was collected. They are listed in order of their addition to the cumulative set. The overall accuracy, sensitivity and specificity of each subset is listed. pcw = post conception weeks, yr(s) = years of age.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/32/23/3611/2525615 by UNIVERSIDAD DE GUANAJUATO user on 21 February 2019

indicative of strong performance for both models. Genes in the lower percentiles were principally classified as negative instances. Given the heterogeneity of the disorders, it is also possible that the misclassified genes are false positives or that their etiology for ASD is independent of brain development.

3.4 Application of the SVM model to lncRNAs

Long non-coding RNA (lncRNA) genes code for transcripts greater than 200 nucleotides in length, which are not translated to peptide sequences. They are ideal genes for testing the performance of an expression-based prioritization model since they lack protein product, generally lack functional annotation, may be more numerous than protein-coding genes in the genome, and are highly expressed in the brain (Derrien *et al.*, 2012). They also have high temporospatial expression specificity, and conservation studies have identified them as potentially key developmental regulators (Necsulea *et al.*, 2014). LncRNAs have been found to be differentially expressed in individuals with ASD (Ziats and Rennert, 2013), but the role of lncRNAs in ASD is still an emerging field of research.

To further evaluate the model performance, we prioritized a gene list comprised of the available lncRNA genes within the dataset using the selected feature model. For each gene, confidence values for the prediction were assigned based upon the SVM output for the gene (Supplementary Table S3). Of the 10 840 lncRNA genes, 962 (8.87%) were classified as potential ASD risk genes, but only 63 had a confidence measure greater than 0.95.

While an exhaustive investigation of the high-priority candidate lncRNAs and their potential ASD association is outside the scope of this study, we highlight the most interesting genes based on existing annotations as a means of further demonstrating the validity of the approach. Table 4 shows four lncRNA genes that are highly ranked for ASD association. CHL1-AS1 is antisense to the ASD risk gene CHL1 (Salyakina *et al.*, 2011). MALAT1 has been shown to be a regulator of synapse formation (Bernard *et al.*, 2010). MIAT has been shown to influence cell fate during neurogenesis (Aprea *et al.*, 2013). TUG1 has been linked previously to neurodegenerative disorders (Wüu *et al.*, 2013). Given the limited knowledge of the role of lncRNAs in ASD, the high prioritization of genes with roles in brain development or adjacency to known ASD risk genes demonstrates the high performance of the SVM model.

Table 3. The mean percentile rank of known ASD risk genes for the selected and full feature set SVM models

Number of genes	Selected features model mean percentile rank	Full feature set model mean percentile rank
101	77.4	75.6
201	78.4	75.6
401	80.7	77.2

Three different gene set sizes were tested for the hypothetical loci.

4 Discussion

Autism spectrum disorder (ASD) has a complex physiologic etiology. Combinations of environmental and genetic factors causing aberrant development of brain regions have been linked to the disorder (Fakhoury, 2015). Studies utilizing non-invasive brain imaging procedures such as magnetic resonance imaging (MRI) and positronic emission tomography (PET) have also implicated brain structures and biological processes contributing to ASD (Ecker *et al.*, 2015; Zürcher *et al.*, 2015). While the underlying mechanisms remain poorly understood, there is a large body of knowledge in the field to be utilized for further study. Although treatment of ASD has been shown to be effective, it is dependent on early detection. Imaging can offer diagnostic input, but the process is expensive and not always practical. Biomarkers may offer the best means for early detection of the disorder, but the complexity of the disorder and the increase in sequencing capacity have led to a multitude of potential targets too numerous to test. The need for an ASD risk gene prioritization system encompassing all gene types is evident. In this study, we employ a novel approach to bridging this gap.

By reframing the task of ASD risk gene identification as a supervised machine learning problem, we were able to construct an accurate classification model. Leveraging the existing extensive research on protein-coding genes in ASD and the comprehensive view of the transcriptome for the developing human brain offered by the BrainSpan dataset, we were able to discern a pattern from the expression profile, which distinguishes ASD risk genes. Appling a heuristic search through the possible feature space, we were also

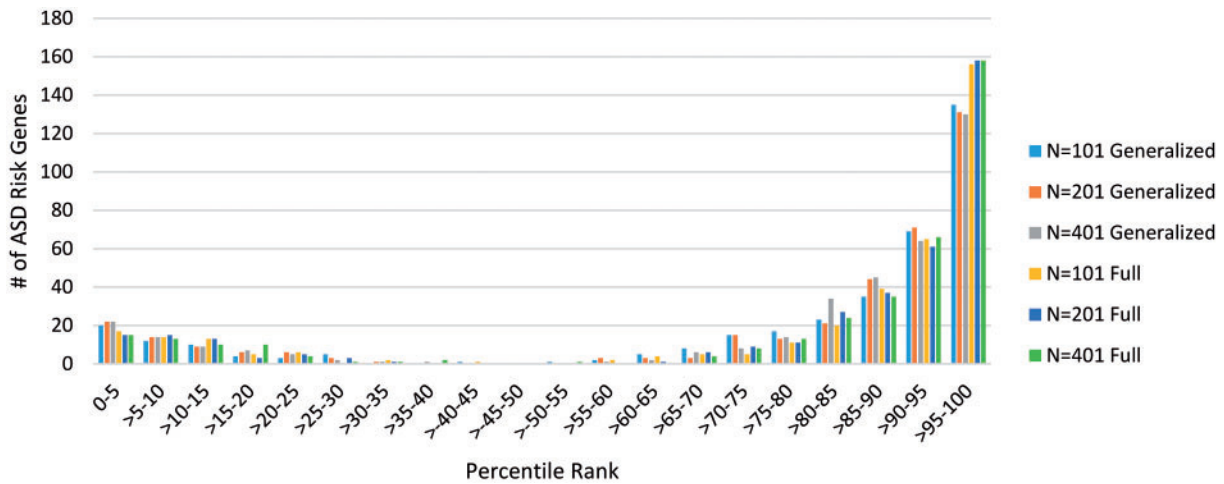

Fig. 2. Histogram of ASD risk gene count grouped by percentile rank for the selected and full feature set SVM models for the three gene set sizes. For each hypothetical locus, the percentile rank of the known ASD risk gene within the gene list was calculated, and here those ranks are grouped in 5 percentile point increments

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/32/23/3611/2525615 by UNIVERSIDAD DE GUANAJUATO user on 21 February 2019

Table 4. Genes of interest from the prioritization of lncRNA genes with their biotype, SVM output and confidence score

Gene	Type	SVM output	Confidence
CHL1-AS1	lncRNA-antisense	3.150	0.996
MIAT	lincRNA	2.933	0.974
MALAT1	lincRNA	2.266	0.919
TUG1	lincRNA-antisense	1.915	0.846

lincRNA, long intervening non-coding RNA.

able to refine the model through feature selection, which improved performance and implicated potentially critical temporospatial features in the onset of ASD. To test the performance of our model in the absence of an independent test set, we used both standard cross-validations and the prioritization of candidate genes within hypothetical loci. The performance measures confirmed that the model achieved high accuracy and had the capability to highly prioritize ASD risk genes within gene lists of varying lengths. In light of the gathering evidence that lncRNAs are associated with ASD, we further demonstrated the utility of our model through the prioritization of lncRNA candidates. Biologically significant candidates were highly prioritized, providing further validation to our approach.

Collectively, the feature subset has interesting aspects (Table 2). It is intriguing that while ASD as early developmental disorders can be diagnosed by the age of two years old with standard methods (American Psychiatric Association, 2012), the developmental time points of the selected features span the entirety of the transcriptome studied (from 8 pcw to 40 years of age). Most notably, the time points are enriched for early development, particularly from 8 pcw to one year of age, and these points were predominantly added to the subset before later developmental time points such as those greater than four years of age. Therefore, the early developmental period appears to have a larger influence on ASD risk gene identification. Co-expression modules enriched with ASD risk genes have been shown to have either dramatic upward or downward trends in expression patterns between 8 pcw to one year of age, which indicates a critical timespan in relation to ASD etiology (Parikshak *et al.*, 2013). Not surprisingly, the structures that were selected are mainly cortical regions, which are associated with sensory input processing and behavior. Generally, cortical regions have been found to be enlarged in children with ASD around three years of age (Schumann *et al.*, 2010). Two non-cortical regions selected are the hippocampus and the mediodorsal nucleus of thalamus (MD). Both the hippocampus and the thalamus have been found to be proportionately smaller for individuals between the ages of four and eighteen years old with ASD (Sussman, *et al.*, 2015). However, there is little to no evidence of a role for the MD in ASD. The selected feature set also contained later development cortical structures. These unanticipated features may present new avenues of research for ASD.

5 Conclusion

In this study, a novel approach is proposed for knowledge transfer from known ASD risk protein-coding genes to all gene types. We have demonstrated that a model built using only expression patterns within normal brain development as features can accurately classify and prioritize ASD risk genes. This provides a distinct advantage over previous models. It does not require *a priori* knowledge and allows for the prioritization of non-protein coding genes. It is our hope that this will lay the groundwork for an accurate prioritization tool utilizing this model.

Acknowledgements

We would like to thank Dr. Anand Srivastava for his discussion, and Jose Guevara for his review and comments on the manuscript.

Funding

This work was supported by a grant from the Self Regional Healthcare Foundation.

Conflict of Interest: none declared.

References

Abrahams,B. *et al.* (2013) SFARI Gene 2.0: A Community-Driven Knowledgebase for the Autism Spectrum Disorders (ASDs). *Mol. Autism*, **4**, 36.

American Psychiatric Association. (2012) Diagnostic and statistical manual of mental disorders (5th ed., text rev.).

Anney,R. *et al.* (2012) Individual common variants exert weak effects on the risk for Autism Spectrum Disorders. *Hum. Mol. Genet.*, **21**, 4781–4792.

Aprea,J. *et al.* (2013) Transcriptome sequencing during mouse brain development identifies long non-coding RNAs functionally involved in neurogenic commitment. *EMBO J*, **32**, 3145–3160.

Bakken,T. *et al.* (2010) Psychiatric disorders in adolescents and adults with Autism and intellectual disability: a representative study in one county in Norway. *Res. Dev. Disabil.*, **31**, 1669–1677.

Bernard,D. *et al.* (2010) A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO J*, **29**, 3082–3093.

Bruing,H. *et al.* (2014) Behavioral signatures related to genetic disorders in autism. *Mol. Autism*, **5**, 11.

Chawla,N. *et al.* (2002) SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.*, **16**, 321–357.

Chlebowski,C. *et al.* (2013) Large-scale use of the modified checklist for autism in low-risk toddlers. *Pediatrics*, **131**, e1121–e1127.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

De Rubeis,S. *et al.* (2014) Synaptic, transcriptional and chromatin genes disrupted in Autism. *Nature*, **515**, 209–215.

Derrien,T. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, **22**, 1775–1789.

Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators and Centers for Disease Control and Prevention (CDC). (2014) Prevalence of Autism Spectrum Disorder among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2010. *MMWR Surveil. Summ.*, **63**, 1–21.

duVerle,D. and Mamitsuka,H. (2012) A review of statistical methods for prediction of proteolytic cleavage. *Brief. Bioinf.*, **13**, 337–349.

Ecker,C. *et al.* (2010) Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. *Neuroimage*, **49**, 44–56.

Ecker,C. *et al.* (2015) Neuroimaging in Autism spectrum disorder: brain structure and function across the lifespan. *Lancet. NEURO*, **14**, 1121–1134.

Erlich,Y. *et al.* (2011) Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. *Genome Res.*, **21**, 658–664.

Fakhoury,M. (2015) Autistic spectrum disorders: a review of clinical features, theories and diagnosis. *Int. J. Dev. Neurosci.*, **43**, 70–77.

Hajian-Tilaki,K. (2013) Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J. Intern. Med.*, **4**, 627–635.

Hall,M. *et al.* (2009) The WEKA data mining software: an update. *ACM SIGKDD Explorations Newslett.*, **11**, 10–18.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/32/23/3611/2525615 by UNIVERSIDAD DE GUANAJUATO user on 21 February 2019

Harrow,J. *et al*. (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res*., **22**, 1760–1774.

Hawrylycz,M. *et al*. (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, **489**, 391–399.

Hira,Z. and Gillies,D. (2015) A review of feature selection and feature extraction methods applied on microarray data. *Adv. Bioinf*., **2015**, 198363.

Hoekstra,R. *et al*. (2009) Association between extreme autistic traits and intellectual disability: insights from a general population twin study. *Br. J. Psychiatry*, **195**, 531–536.

Hsu,C. *et al*. (2003) A practical guide to support vector classification. *Tech. Rep*., Department of Computer Science, National Taiwan University.

Kim,Y. and Leventhal,B. (2015) Genetic epidemiology and insights into interactive genetic and environmental effects in autism spectrum disorders. *Biol. Psychiatry*, **77**, 66–74.

Kohavi,R. and John,G. (1997) Wrappers for feature subset selection. *Artif. Intell*., **97**, 273–324.

Kourou,K. *et al*. (2014) Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J*., **13**, 8–17.

Kubat,M. and Matwin,S. (1997) Addressing the Curse of Imbalanced Training Sets: One Sided Selection. In: *Proceedings of the Fourteenth International Conference on Machine Learning*, Nashville, Tennesse. Morgan Kaufmann, pp. 179–186.

Lin,W. and Chen,J. (2012) Class-imbalanced classifiers for high-dimensional data. *Brief. Bioinform*., **12**, 13–26.

Liu,X. and Takumi,T. (2014) Genomic and genetic aspects of autism spectrum disorder. *Biochem. Biophys. Res. Commun*., **452**, 244–253.

Matthews,B. (1975) 'Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme.' *Biochim. Biophys. Acta*, **405**, 442–451.

McFadden,K. and Minshew,N. (2013) Evidence for dysregulation of axonal growth and guidance in the etiology of ASD. *Front. Hum. Neurosci*., **7**, 671.

Moreau,Y. and Tranchevent,L. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet*., **13**, 523–536.

Necsulea,A. *et al*. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, **505**, 635–640.

Parikshak,N. *et al*. (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in Autism. *Cell*, **155**, 1008–1021.

Pedregosa,F. *et al*. (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*., **12**, 2825–2830.

Pinto,D. *et al*. (2010) Functional impact of global rare copy number variation in Autism spectrum disorders. *Nature*, **466**, 368–372.

Piro,R. *et al*. (2010) Candidate gene prioritization based on spatially mapped gene expression: an application to XLMR. *Bioinformatics*, **26**, i618–i624.

Retico,A. *et al*. (2016) The effect of gender on the neuroanatomy of children with autism spectrum disorders: a support vector machine case-control study. *Mol. Autism*, **7**, 5.

Saeys,Y. *et al*. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Salyakina,D. *et al*. (2011) Copy number variants in extended Autism spectrum disorder families reveal candidates potentially involved in Autism risk. *PloS One*, **6**, e26049.

Schumann,C. *et al*. (2010) Longitudinal magnetic resonance imaging study of cortical development through early childhood in Autism. *J. Neurosci*., **30**, 4419–4427.

Si,J. *et al*. (2015) An overview of the prediction of protein DNA-binding sites. *Int. J. Mol. Sci*., **16**, 5194–5215.

Sussman,D. *et al*. (2015) The Autism puzzle: diffuse but not pervasive neuroanatomical abnormalities in children with ASD. *NeuroImage Clin*., **8**, 170–179.

Thienpont,B. *et al*. (2010) Haploinsufficiency of TAB2 causes congenital heart defects in humans. *Am. J. Hum. Genet*., **86**, 839–849.

Wang,H. and Pai,T. (2014) Machine learning-based methods for prediction of linear B-cell epitopes. *Methods Mol. Biol*., **1184**, 217–236.

Wang,L. and Brown,S. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res*., **34**, W243–W248.

Wüu,P. *et al*. (2013) Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. *Brain Res. Bull*., **97**, 69–80.

Xu,L. *et al*. (2012) AutismKB: an evidence-based knowledgebase of Autism genetics. *Nucleic Acids Res*., **40**, D1016–D1022.

Yang,Z. (2004) Biological applications of support vector machines. *Brief. Bioinf*., **5**, 328–338.

Ziats,M. and Rennert,O. (2013) Aberrant expression of long noncoding RNAs in Autistic brain. *J. Mol. Neurosci*., **49**, 589–593.

Zürcher,N. *et al*. (2015) A systematic review of molecular imaging (PET and SPECT) in Autism spectrum disorder: current state and future research opportunities. *NeBiobehav. Rev*., **52**, 56–73.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/32/23/3611/2525615 by UNIVERSIDAD DE GUANAJUATO user on 21 February 2019