IOP Conf. Series: Materials Science and Engineering **719** (2020) 012024 doi:10.1088/1757-899X/719/1/012024

Improved Disease Gene Predication Method

Gerui He[1,b], Zhiming Liu[1,a,*], Lingyun Luo[1] and Yaping Wan[1]

[1]School of Computer, University of South China, Hunan, Hengyang , 421001, China
Email: [a]nhdxlzm@usc.edu.cn, [b]616281384@qq.com

Abstract. The prediction of disease genes has always been a hot topic in the field of bioinformatics. Machine learning methods can effectively dig out the hidden relationship between disease-causing genes and predict disease genes. At present, the prediction algorithm of Gene Ontology (GO) combined with GO annotation has limitations. It is believed that disease genes will only accumulate on the biological process branches of GO, ignoring the cellular components and molecular function branches. Disease gene prediction is performed by considering data from three branches of biological processes, cell components, and molecular functions. Multiple sets of experiments were performed. The data showed that the use of three branches to predict disease genes increased the accuracy from 78% to 91%, indicating that the disease genes not only aggregate on the branches of biological processes but also aggregate on molecular functions and cellular components.

1. Introduction

Some diseases in the traditional concept belong to mental illness, such as schizophrenia, autistic spectrum disorder(ASD), major depression. But the progress of genetic technology has sent a voice of opposition. These diseases are no longer considered to be simple mental illnesses but are also considered to be genetic diseases, and in some cases identify potentially relevant genes [1].

Autism Spectrum Disorder is a complex mental disorder that has a strong genetic influence, multiple causes, and hundreds of different associated genes. A mass of genetic studies related to ASD has identified hundreds of disease-causing genes [2,3]. However, the massive data generated by these large-scale research institutes bring more invalid information while bringing effective information. It has become a challenge and a necessary task to dig out effective information from the vast amount of data and identify the true pathogenic genes. At present, 20%-25% of genetic factors of ASD disease are found, and there is still much room for improvement, which has research significance.

In recent years, machine learning has been applied to various research fields, and the analysis of genomic data sets is no exception [4]. A supervised machine learning approach can identify hidden relationships between disease genes in a data set and then use this information to distinguish disease genes from non-disease genes [5,6]. At present, how to use existing genomic data to reliably predict disease genes and improve the accuracy of prediction accuracy is one of the problems that need to be solved to transform genetic technology into biomedical applications.

2. Related Work

Krishnan used the weighted support vector machine (SVM) to predict the association probability between brain genes and ASD [7]. They use weighted SVM to train on specialized gene networks that integrate gene expression, protein-protein interactions, and regulatory sequences of brain genes. The classifier evaluates on high confidence disease gene of ASD. Besides, there are limitations in the network which is difficult to express weak interactions. Therefore, protein-protein interactions and co-

expression network data cannot express weak interactions that limit the prediction range of disease genes.

To overcome the shortcomings of network data not expressing weak interactions, Muhammad used gene ontology to predict gene functional similarity as training data [8]. Genes with similar functions have a similar expression trend. For example, in ASD, a disease-related gene that is disrupted by genetic variation tends to accumulate on specific biological processes [9]. That indicates that disease genes may belong to the same hierarchical path of GO and have a higher functional similarity. Based on this hypothesis, Muhammad used GO predictor gene functional similarity to predict potential ASD candidate genes.

Muhammad believes that disease genes only accumulate in biological processes, ignoring molecular function and cellular components. The effect of the calculation is that some genes have a functional similarity to any gene that is zero, including itself. The reason for this illogical phenomenon is that the term annotating the gene is not on the branch of the biological process. For example, the gene “AAAS”, whose annotation terminology is shown in Figure 1, is all located in the branch of the cell component and has no annotations located in the branch of the biological process. The genetic similarity between all genes and “AAAS “is 0, include “AAAS”.

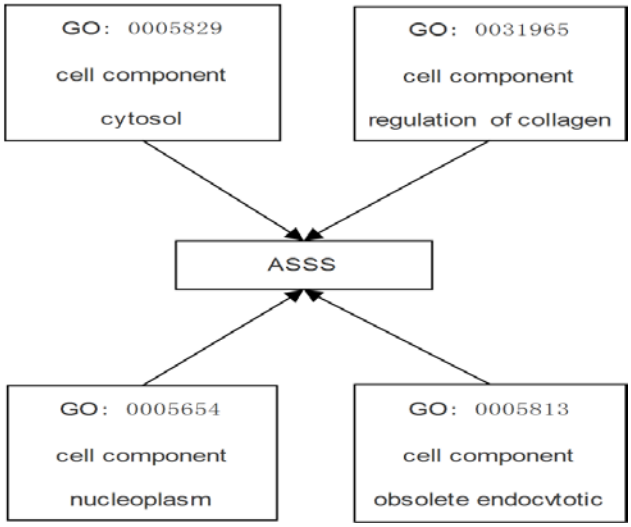

Figure 1. Annotation relationship of Gene Ontology term and gene of AAAS

In response to the above illogical shortcomings, it is believed that disease genes not only accumulate on the biological processes but also aggregate on cellular components and molecular functions. When calculating the functional similarity of genes, considering the three branches of GO. And then the final prediction accuracy is improved.

3. Improved Method

3.1. Overview

The method of Muhammad consider that disease genes will only accumulate on branches of biological processes, Named it biological process aggregation (BPA) methods. The full-branch aggregation (FBA) method in this paper believes that disease genes not only aggregate on biological process branches but also aggregate on molecular functions and cell component branches.

Figure 1 shows the steps of the FBA method. A gene functional similarity matrix is generated for a given data set. Gene ontology annotations for molecular functions and cellular components are no longer ignored when calculating gene functional similarities. Classifier training and testing on the gene functional similarity matrix. Classifiers from machine learning methods used: random forest, naive Bayes, support vector machine. Also, during training and prediction, the undersampling method is

used to strictly keep the number of positive class (disease gene) and negative class (non-disease gene) consistent to avoid partial estimation. Each classifier is trained and tested repeatedly for 20 averaging.

3.2. Gene Ontology Semantic Similarity

Gene functional similarity calculated using gene ontology semantic similarity when constructing similarity matrix. This article uses four different semantic similarity method: resnik [10], rel [11], wang [12], Netsim [13]. The resnik and rel algorithms are based on information content. The information content of gene ontology terms is defined as follows:

$$\mathrm{IC(c)} = -\log\big(\mathrm{p(c)}\big) \tag{1}$$

The resnik method is defined as follows

$$\mathrm{Sim_{resnik}(t_1, t_2)} = \max_{\mathrm{t}} \mathrm{IC(t)(t \in T)} \tag{2}$$

Where T is a set of common ancestor terms for t1and t2.

The rel method is defined as follows:

$$Sim_{rel}(t_1, t_2) = \max_t \frac{2 \times IC(t)}{IC(t_1) + IC(t_2)} (t \in T) \tag{3}$$

Where T is a set of common ancestor terms for t1and t2.

Wang's method is based on semantic contribution. For a given term, calculate the semantic contribution of the ancestor node of the term to the node. The similarity of two terms is expressed by the semantic contribution of the common ancestor to the two terms. The formula is as follows:

$$Sim_{wang}(t_1, t_2) = \sum_t \frac{S_A(t) + S_B(t)}{SV(t_1) + SV(t_2)} \ (t \in T) \tag{4}$$

Where T is a set of common ancestor terms for t1, t2.S_A(t) is the semantic contribution of the term t to the term t1.S_B(t) is the semantic contribution of the term t to the term t2.SV(t1) is the sum of the semantic contributions of all ancestor terms of the term t1 to t1. SV(t2) is the sum of the semantic contributions of all ancestral terms to t2 in the term t2.

Netsim is a newer method that combines GO, GO annotation and co-function data. The method performs better on metabolic pathways.

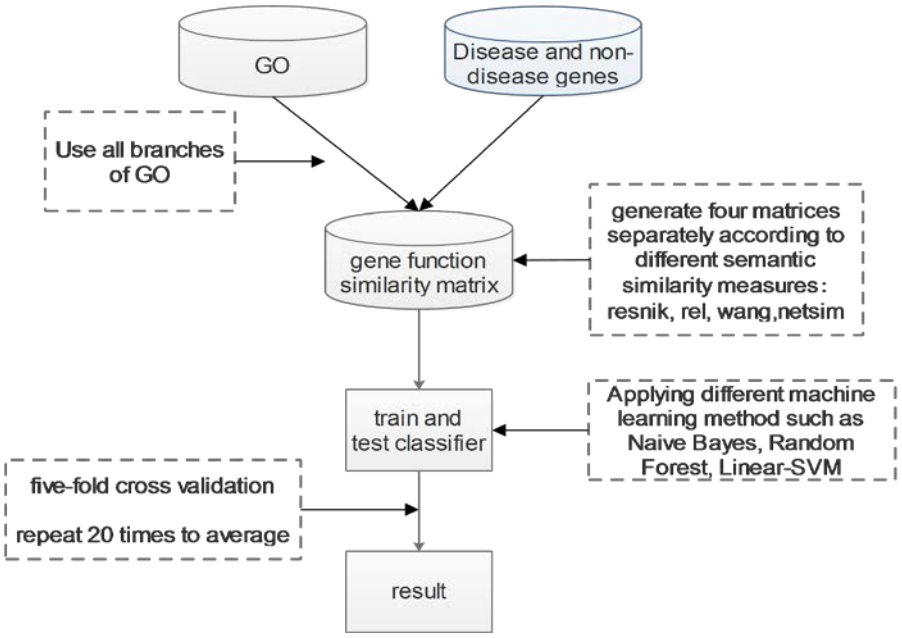

Figure 2. Step of proposed method to predict disease genes

3.3. Gene Function Similarity

The calculation methods of gene functional similarity are divided into two categories: the pairwise comparison method and overall comparison method. Since the overall comparison method does not pass the semantic similarity between gene ontology, the overall comparison method is not considered. In the pairwise comparison method, each gene corresponds to a set of gene ontology terms, and comparing the similarities of the two genes is equivalent to comparing the corresponding term sets. Common methods for integrating the similarity of two-term sets include average, maximum, and sum. This paper chooses the most widely used maximum method.

For a given two genes g1, g2. Through the GO annotation data, the corresponding term set T1, T2 is obtained. For example, in Figure 1, the set of terms corresponding to the gene ASSS is {"GO: 0005829", "Go: 0031965", "GO: 0005654", "GO: 0005813"}

The gene functional similarity of g1, g2 is defined as follows:

$$(g_1, g_2) = max\left(\left(t_i, t_j\right)\right)\left(t_i \in T_1, t_j \in T_2\right) \tag{5}$$

3.4. Data Set

Disease gene of ASD was obtained from the Simons Foundation Autism Research Initiative (SFARI) gene database(https://gene.sfari.org/)

Genes are classified into 7 categories according to the degree of correlation with the disease in the SFARI data set. The most relevant, reproducible evidence is classified as a class 1 gene. Reproducible under certain constraints is classified as a class 2 gene. Class 3, 4 genes have small research evidence, 5 genes are indirectly related, and 6 genes are completely unrelated. There is another type of gene that does not indicate a score. In this study, only 746 genes of 1, 2, 3, and 4 genes were used, of which 1, 2 were used as High confidence Disease gene HD, a total of 87, 3, 4 As a low confidence disease gene (LD), a total of 649. A total of 1111 non-ASD genes were from Muhammad's paper [13].

In the experiment, by dividing the positive class (ie disease-related genes) into HD and LD, two sub-datasets were generated, one being HD as a positive class and NoASD as a negative class, which is called an HD data set. The other is HD+LD as a positive class and NoASD as a negative class called the HD+LD dataset.

Furthermore, due to the update of SFARI data, 8 of the LD genes intersected with non-ASD. These 8 genes were used as expired invalid data, and these 8 genes were deleted in both non-ASD and LD. The classification of the final gene is shown in Figure 2.

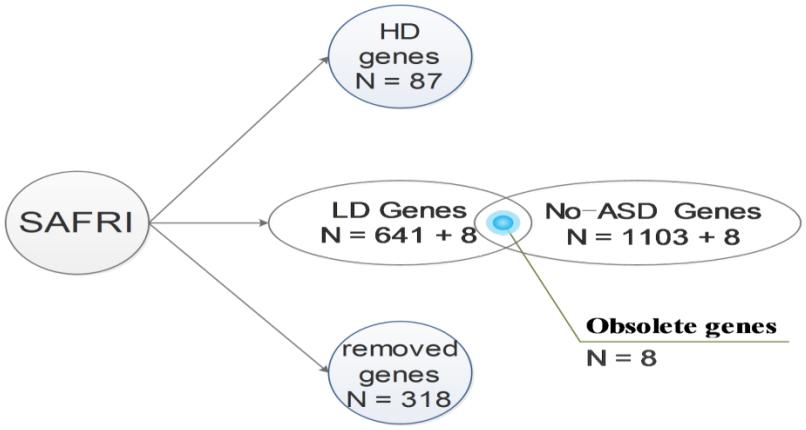

Figure 3. ASD data set gene classification

3.5. *Class Imbalance Problem*

There is a class imbalance in the dataset. The non-ASD gene is more than the ASD gene, especially in the HD data set, reaching a ratio of 87:1103. Therefore, this paper adopts a random undersampling method to solve the problem of class imbalance. At the time of training and testing, random undersampling of non-ASD genes, random acquisition of non-ASD gene construction training sets and

test sets consistent with the number of ASD genes. For example, when performing HD gene training and testing, there are 87 genes in the normal HD gene, and the sample size of the negative class is also 87.

However, under-sampling introduces the risk of data specialization, it is easy for the model to learn only a part of the features. Therefore, the experiment was repeated several times and then averaged to avoid the problem of data specialization.

4. Experimental Result

Four different ontology similarity algorithms generate four matrices, which are classified according to HD and HD+LD. The results are shown in Table 1.

Table 1. Experiment result

		HD		HD+LD	
		BPA	FBA	BPA	FBA
Resnik	RF	0.793	0.908	0.759	0.897
	NB	0.701	0.707	0.667	0.718
	SVM	0.793	0.897	0.756	0.885
Rel	RF	0.753	0.891	0.787	0.897
	NB	0.660	0.741	0.655	0.747
	SVM	0.736	0.799	0.736	0.810
Wang	RF	0.787	0.914	0.793	0.897
	NB	0.690	0.724	0.701	0.747
	SVM	0.770	0.862	0.782	0.839
Netsim	RF	0.770	0.793	0.770	0.805
	NB	0.598	0.517	0.506	0.529
	SVM	0.672	0.615	0.684	0.615

In Muhammad's paper, the performance of HD datasets is significantly better than HD+LD, and the experimental data in this paper has the same performance. The BPA method based on resnik similarity has better performance of HD dataset than HD+LD on all classifiers. In rel, wang and netsim methods, the performance on the HD dataset is not much different from the performance of HD+LD.

There is no statistically significant difference in the performance of the classifier using the FBA method on the two data sets. This indicates that the full branch aggregation method has lower quality requirements than the biological process aggregation method.

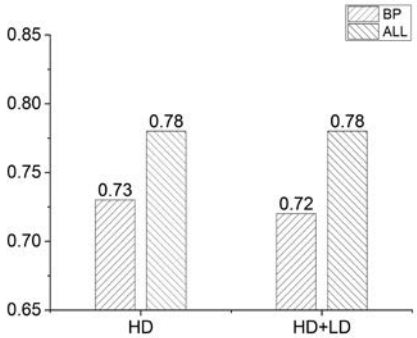

Figure 4. Average performance comparison of data sets

The average classification accuracy is shown in Figure 3 also shows that the HD performance of the algorithm for biological process aggregation is better than HD+LD, while the performance of full-branch aggregation is consistent.

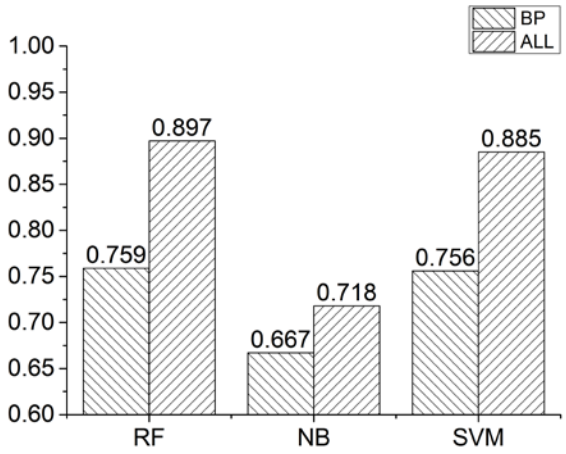

Figure 5. Classification result based on HD+LD in resnik method

The results of the four gene ontology similarity methods, resnik, rel, and wang all showed the same trend: whether in the HD dataset or the HD+LD dataset, the full branch aggregation is better than the biological process aggregation algorithm. The classification accuracy has a 10% improvement. Figure 4 is a comparison of the resnik algorithm on the HD+LD data set.

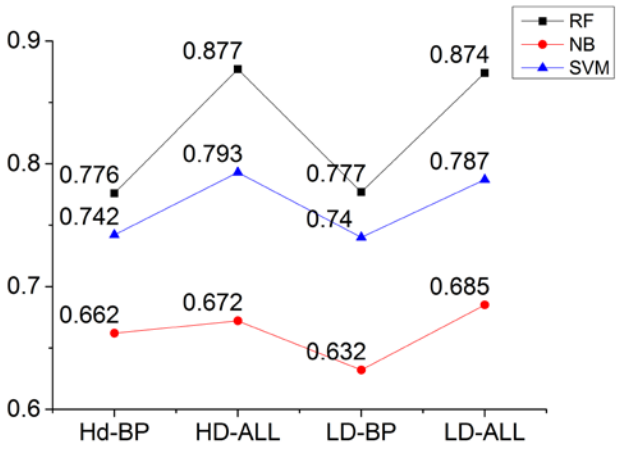

Figure 6. Average accuracy of classification of three classifiers

But the results based on netsim semantic similarity have a different performance. FBA on random forest classifiers has some improvement over BPA. However, there is no improvement in Naive Bayes and SVM classifiers, but it has declined. The main reason for this is that NETSIM uses other data sources. NETSIM not only considers GO and annotation, but also considers gene functional distance, and considers the similarity of gene function when calculating semantic similarity.

In this paper, three classifiers are used. The results are shown in Figure 5. In different cases, the RF performance is higher than SVM, and SVM is higher than NB. The best performing classifier is the RF classifier running on the HD dataset. The matrix is generated using wang's semantic similarity method and FBA with an accuracy of 91.4%.

5. Summary

The improved disease gene prediction method effectively improves the prediction accuracy of identifying ASD disease genes. Besides, disease gene prediction based on full-branch aggregation is superior to biological process branch aggregation. This indicates that disease genes not only accumulate in biological process but also aggregate on molecular functions and cellular components, which is also helpful for predicting other diseases in the future.

6. Acknowledgement

This work is supported by the Hunan Natural Science Foundation Youth Project(2019JJ50520)

7. References

[1] Geschwind D H , Flint J . Genetics and genomics of psychiatric disease[J]. Science, 2015, 349(6255):1489-1494.
[2] Sanders S J . First glimpses of the neurobiology of autism spectrum disorder[J]. Current Opinion in Genetics & Development, 2015, 33:80-92.
[3] Stephan Ripke etc al. Biological insights from 108 schizophrenia-associated genetic loci[J]. Nature, 2014, 511(7510):421-427.
[4] Libbrecht M W , Noble W S . Machine learning applications in genetics and genomics[J]. Nature Reviews Genetics, 2015, 16(6):321-332.
[5] Luo P, et al. Identifying disease genes from PPI networks weighted by gene expression under different conditions[C] //2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2016: 1259-1264.
[6] Radivojac P , et al. An integrated approach to inferring gene-disease associations in humans.[J]. Proteins-structure Function & Bioinformatics, 2010, 72(3):1030-1037.
[7] Krishnan A , Zhang R , Yao V , et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder[J]. Nature Neuroscience, 2016. (2016): 1454-1462.
[8] Asif M, et al. Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology[J]. PloS one, 2018, 13(12): e0208626.
[9] Voineagu I, Eapen V. Converging pathways in autism spectrum disorders: interplay between synaptic dysfunction and immune responses[J]. Frontiers in human neuroscience, 2013, 7: 738.
[10] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language[J]. Journal of artificial intelligence research, 1999, 11: 95-130.
[11] Schlicker A, et al. A new measure for functional similarity of gene products based on Gene Ontology[J]. BMC bioinformatics, 2006, 7(1): 302.
[12] Wang J Z, et al. A new method to measure the semantic similarity of GO terms[J]. Bioinformatics, 2007, 23(10):1274-1281.
[13] Peng J, Uygun S, Kim T, et al. Measuring semantic similarities by combining gene ontology annotations and gene co-function networks.[J]. Bmc Bioinformatics, 2015, 16(1):1-14.