molecules

MDPI

Article

An Efficient Classifier for Alzheimer's Disease Genes Identification

Lei Xu [1], Guangmin Liang [1], Changrui Liao [2], Gin-Den Chen [3,*] and Chi-Chang Chang [4,5,*]

1 School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen 518055, China; csleixu@szpt.edu.cn (L.X.); gmliang@szpt.edu.cn (G.L.)
2 Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Optoelectronic Engineering, Shenzhen University, Shenzhen 518060, China; cliao@szu.edu.cn
3 Department of Obstetrics and Gynecology, Chung Shan Medical University Hospital, Taichung 40201, Taiwan
4 School of Medical Informatics, Chung Shan Medical University, Taichung 40201, Taiwan
5 IT Office, Chung Shan Medical University Hospital, Taichung 40201, Taiwan
* Correspondences: gdchentw@hotmail.com (G.-D.C.); threec@csmu.edu.tw (C.-C.C.); Tel.: +86-64-2473-0022 (ext.12218) (C.-C.C.)

Academic Editor: Xiangxiang Zeng
Received: 24 October 2018; Accepted: 19 November 2018; Published: 29 November 2018

Abstract: Alzheimer's disease (AD) is considered to one of 10 key diseases leading to death in humans. AD is considered the main cause of brain degeneration, and will lead to dementia. It is beneficial for affected patients to be diagnosed with the disease at an early stage so that efforts to manage the patient can begin as soon as possible. Most existing protocols diagnose AD by way of magnetic resonance imaging (MRI). However, because the size of the images produced is large, existing techniques that employ MRI technology are expensive and time-consuming to perform. With this in mind, in the current study, AD is predicted instead by the use of a support vector machine (SVM) method based on gene-coding protein sequence information. In our proposed method, the frequency of two consecutive amino acids is used to describe the sequence information. The accuracy of the proposed method for identifying AD is 85.7%, which is demonstrated by the obtained experimental results. The experimental results also show that the sequence information of gene-coding proteins can be used to predict AD.

Keywords: Alzheimer's disease; gene coding protein; sequence information; support vector machine; classification

1. Introduction

Prior research has shown that there were more than 26.6 million people with AD worldwide in 2010 [1]. It has been predicted that there will soon be a further significant increase in prevalence: specifically, it is expected that there will be 70 million people with AD in 2030 and more than 115 million people with AD in 2050, respectively. In other words, in 2050, one in 85 people are expected to have AD. Unfortunately, to date, there is no treatment in existence that can cure AD. During disease progression, the neurons of AD patients are destroyed gradually, resulting in the loss of cognitive ability and ultimately death. Thus, it is important to identify AD, an age-related disease [2], as early as possible so as to manage the advancement of the condition.

Most existing diagnosis methods focus on identifying AD by way of magnetic resonance imaging (MRI). The MRI method is based on neuroimaging data, for the reason that the imaging data can reflect the structure of brain. Using this technique, the results of classification accuracy are encouraging. However, MRI scans are expensive and the time required for scanning is significant because of the

large size of the images. A diffusion map is extended to identify AD in Mattsson [3] and principal component analysis (PCA) is used to reduce features before classification.

Many biomarkers have been discovered for AD identification, such as structural MRI for brain atrophy measurement [4–6] functional imaging for hypometabolism quantification [7–9], and cerebrospinal fluid for the quantification of specific proteins [6,10,11]. Multimodel data have been employed by multiple biomarkers for identifying AD. Zu et al. [12] predicted AD by using multimodality data to mine the hidden information between features. In Zu [12], the subjects with the same label on a different modal are closer in the selected feature space; as such, a multikernel support vector machine (SVM) can be used to classify the multimodal data, which are represented by the selected features.

As is known, machine learning methods can learn a model from a training sample and then subsequently predict the label of the testing samples. Some machine learning methods have been used to predict AD and mild cognitive impairment (MCI) [13–20]. The information obtained via structural MRI—for example, hippocampal volumes [21,22], cortical thickness [23,24], voxel-wise tissue [23,25,26], and so on—is extracted to classify AD and MCI. Functional imaging, such as fluorodeoxyglucose positron-emission tomography [14,27,28] can also be used for AD and MCI prediction.

Although most existing research has focused on classifying AD based on MRI methods, the cost is expensive. Furthermore, patients often have to have their brain scanned several times in order to inspect the changes in its structure during whole process, increasing the cost even more. Thus, it would be beneficial to find other options for AD identification. Several researches proved that coding genes/noncoding RNAs/proteins were related to diseases, including AD [29–36]. Other investigations [12] have shown that protein structure is related to AD. The gene coding is related to Alzheimer's disease [37–39]. Different from previous work, in the present study, AD is predicted based on protein information. The information of every sequence is represented by a 400-dimension vector, and each dimension represents the frequency of two consecutive amino acids.

The flow chart of AD identification is shown in Figure 1. First, the data are selected by using the CD-HIT method to remove the most similar sequences. In this step, the input are the proteins related with AD, and the output are selected proteins. Second, the features are extracted from the selected sequences. Each sequence is represented by a 400-dimension (400D) vector. In the third step, the data are classified by a support vector machine method. The input are the feature vectors, and the output are peptides with labels. To the best of our knowledge, this study represents the first effort to identify AD by protein sequence information without the use of MRI. Moreover, a dataset including AD and non-AD samples was created in this work. The experimental results show that the classification accuracy for AD prediction is 85.7%. The contributions of our work include:

(1) A method for predicting AD is proposed in this work. The experimental results demonstrate that the classification accuracy of the proposed method is 85.7%.
(2) Our method is based on protein sequence information. The frequencies of two consecutive amino acids are extracted from the sequence with a 400-dimension vector.
(3) A dataset with AD and non-AD samples is created. This dataset could also be used for additional AD prediction studies.

The rest of the paper is organized as follows: Section 2 introduces the experimental results of the proposed method. The dataset and the proposed method are introduced in Section 3. Finally, the conclusion is made in Section 4.

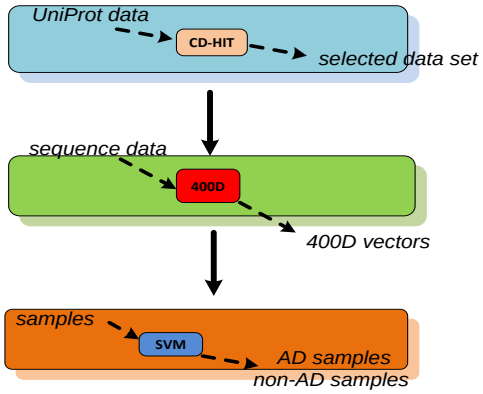

Figure 1. The flow chart of AD identification.

2. Results and Discussion

2.1. Results

Identifying AD by way of using protein sequence information has not been widely done yet. Moreover, most existing works use AD Neuroimaging Initiative (ADNI) database [40], which is based on MRI. Existing methods also use MRI information for classification, which is different from our method. Thus, it is difficult to compare the performance evaluation of our proposed method with the performance of existing methods. The performance of our method is shown in Table 1.

Table 1. The performance of our proposed method.

Performance Evaluation	Accuracy
ACC	0.8565
Precision	0.857
Recall	0.857
F-measure	0.856
MCC	0.714
AUC	0.857

As noted in the table, the method was evaluated according to accuracy, precision, recall, F-measure, Mathew coefficient (MCC), and receiver operating characteristic (ROC). The accuracy of the proposed method was 85.7%, which means that the more than 85% of AD and non-AD samples were able to be classified correctly using the method in question. F-measure is based on precision and recall. The recall of our method was 0.857, and the result shows that 85.7% of AD samples in the dataset could be identified in the experiment. Area under the curve (AUC) is related to the metrics of receiver operating characteristic (ROC). ROC is used to measure sensitivity and specificity, while AUC describes the area under the ROC curve. When the AUC is larger, the performance of the algorithm is better. The value of AUC for our method was 0.857 according to the UniProt dataset [41]. The experimental results show that the performance quality of our method in terms of accuracy, precision, and four other metrics as well as the results obtained are acceptable and encouraging.

2.2. The Comparison of Performance Evaluation on Feature Selection Methods

To demonstrate the efficiency of the feature extraction method we used, we compared the 400D features with information theory, which is another feature extraction method. Information theory is proposed in Wei [42], for exploring sequential information from multiple perspectives. Figure 2 shows that 400D performs better than information theory method on accuracy, precision, F-measure, AUC and MCC. The value of recall is higher by using information theory method than using 400D.

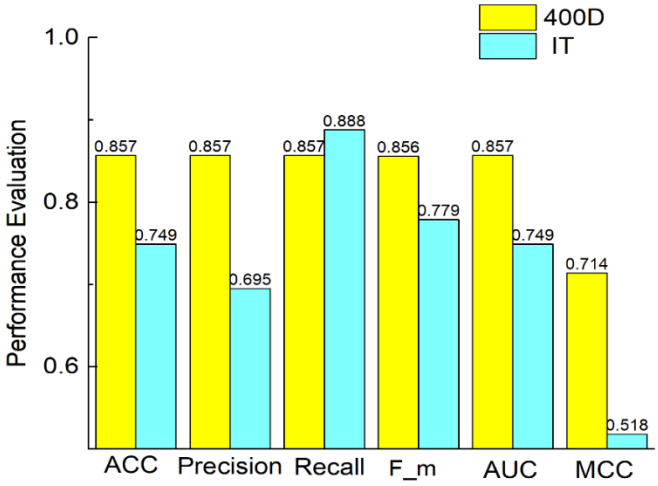

Figure 2. Comparison of 400D with information theory on SVM.

2.3. The Comparison of Performance Evaluation on Existing Classification Methods

Our method's performance is evaluated according to other classifiers, such as random forest, naïve Bayes, LibD3C, Adaptive Boosting (AdaBoost), and Bayes network. The classifiers are introduced briefly as follows:

- Random forest is an ensemble classifier, which learns more than one decision tree together. The decision will be made by voting process.
- Naïve Bayes assumes the features are independent of one other. The samples will be assigned to a class with the maximum posterior probability.
- LibD3C [43] is a hybrid ensemble model, which is based on k-means clustering and the framework of dynamic selection and circulating in combination with a sequential search method.
- AdaBoost can assemble classifiers together and, during the training process, the weights of the samples which are classified incorrectly will be increased. The weights of the samples classified correctly will be decreased.
- Bayes network is a probabilistic graph model. The variables and their relationships are represented by a directed acyclic graph.

Figure 3 shows the comparison of accuracy according to the six classifiers. The comparisons of precision, recall, F-measure, MCC, and AUC are shown in Figures 3–7. In Figure 3, we can see that accuracy performs better than the other classifiers. The value of accuracy of AdaBoost, Bayes network, and naïve Bayes is about 0.8, while the accuracy of SVM is 0.857. The accuracy of LibD3C is 0.84. The accuracy of random forest is 0.85, which is comparative with that of SVM. Thus, SVM improves the accuracy of other classifiers by nearly 1% to 7%.

Figures 4–7 show the comparisons of the classifiers on precision, recall, and F-measure. The results are similar to those of Figure 3. SVM performs better than the other methods. The performance is improved by SVM by approximately 1% to 7.5% as compared with in the case of the other methods. F-measure is calculated based on precision and recall, so the result here is consistent with that of precision and recall. AUC reflects the area under the ROC curve. AUC refers to the ratio of the specificity and sensitivity. The value of AUC on random forest is 0.93, which is better than the values achieved via other methods. The values of AUC for AdaBoost, Bayes network, SVM, and naïve Bayes are similar to one another. Figure 8 shows that the MCC of SVM is 0.714, which is better than the MCCs of the other mentioned methods. The values of MCC for random forest and SVM reach a level of 0.7. Moreover, the value of MCC is improved by 0.8% to 20% by using SVM. As a result, SVM performs better than other classifiers evaluated by the metrics.

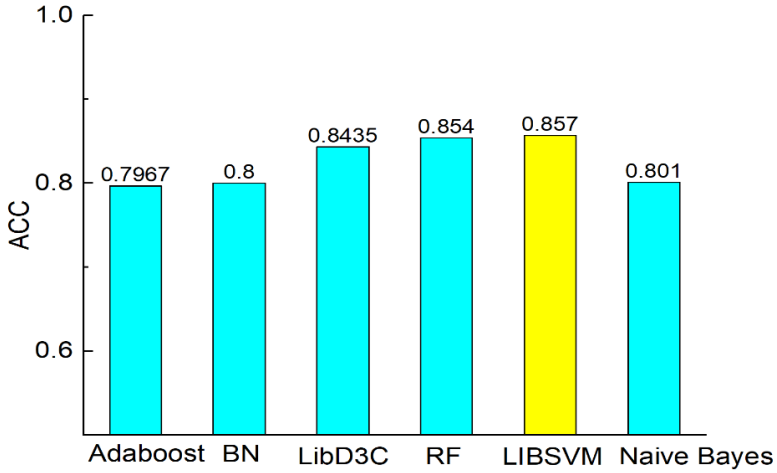

Figure 3. Comparison of ACC on different classifiers.

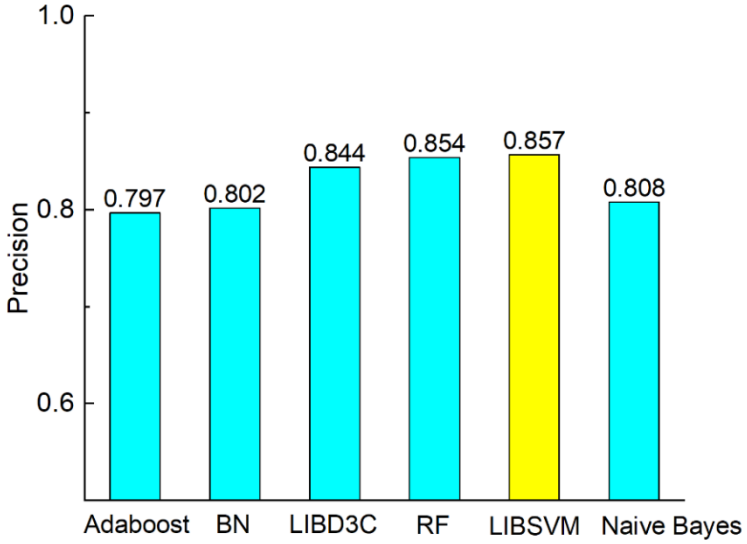

Figure 4. Comparison of precision on different classifiers.

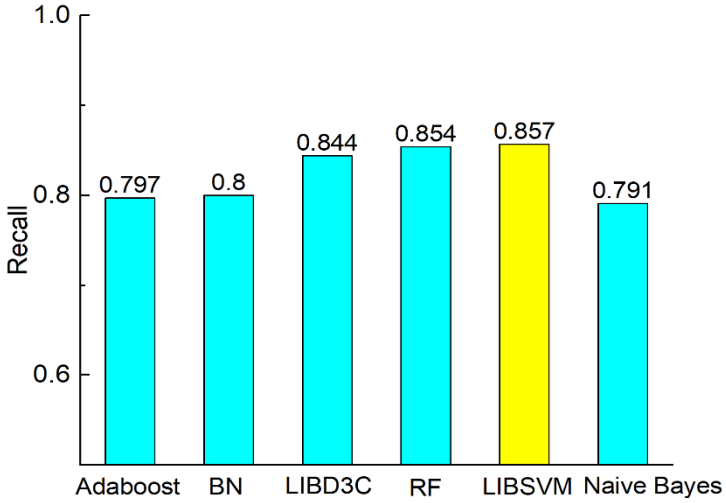

Figure 5. Comparison of recall on different classifiers.

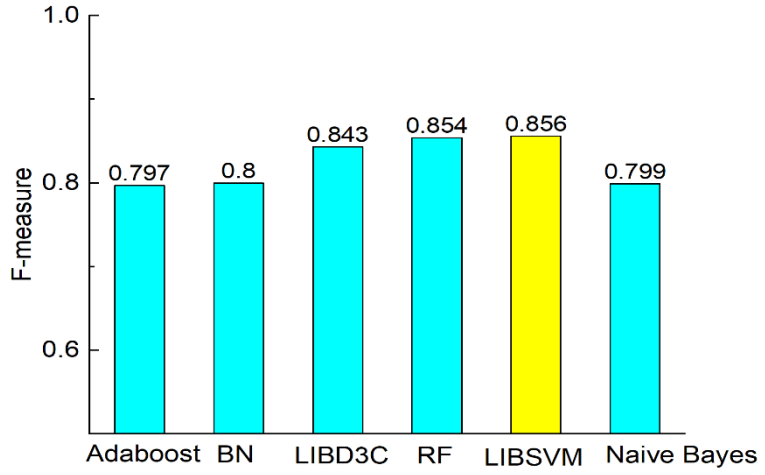

Figure 6. Comparison of F-measure on different classifiers.

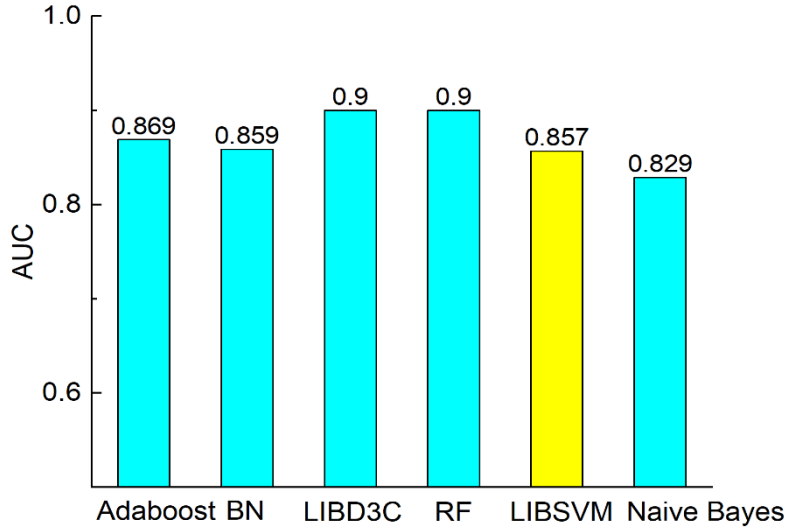

Figure 7. Comparison of AUC on different classifiers.

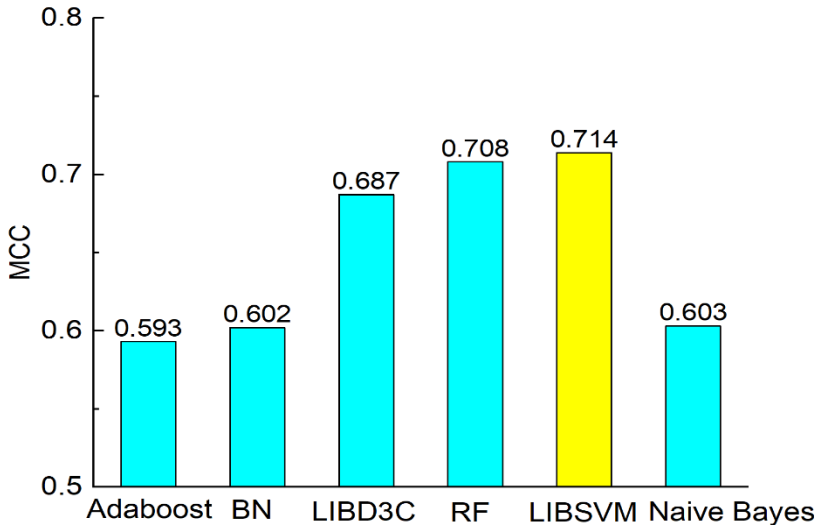

Figure 8. Comparison of MCC on different classifiers.

3. Materials and Methods

3.1. Benchmark Dataset

The data were selected from the UniProt database [41,44]. To guarantee the validity of the dataset, the proteins with ambiguous meanings (such as "B", "X", and so on) is removed, and only the proteins related to "Alzheimer's disease" are kept.

The benchmark dataset (D) is represented by a positive subset (D^+) and a negative subset (D^-), formulated as seen in Equation (1):

$$D = D^+ \cup D^- \tag{1}$$

where the symbol "$\cup$" represents the union of the sets in the set theory. After the selection process, there are 310 proteins related to AD and 312 non-AD proteins left in the benchmark dataset. Because some sequences are significantly similar, the redundancy of the sequences is considered. To avoid the overestimation of the performance of the methods, the homologous sequences with more than 60% similarity were removed from the dataset by using CD-HIT program [45]. As a result, a benchmark dataset with 279 proteins related to AD and 1,463 proteins not related to AD was used for the prediction model. In other words, the benchmark dataset contains 279 positive samples in the positive subset (D^+) and 1,463 negative samples in the negative subset (D^-), respectively.

3.2. Support Vector Machine

SVM is a supervised machine learning model. The labeled samples are trained based on the goal of maximizing the margin between the classes. Since SVM performs better than some the-state-of-art supervised learning methods, SVM is widely used in classification problems. Most works in bioinformatics [41,46–59] also use SVM for classification. SVM was used to identify AD in our work.

The principles of SVM were introduced in Chou and Cai [60,61], and more details are provided in Cristianini [62]. Above all, the key idea of SVM is that two groups are separated with a maximum margin by building a hyperplane. The objective function of SVM is described in Equation (2), as follows:

$$\underset{w,b}{argmax}\{\frac{1}{\|w\|}\min_{i=1,2,\ldots,n}[y_i(w^T\varphi(x^{(i)})+b)]\} \tag{2}$$

In Equation (2), the input variable $x^{(i)}$ is mapped into a high dimensional feature space by the kernel function $\varphi(\cdot)$. Radial kernel function (RBF) is used in the experiment. RBF is used widely because of its effectiveness and efficiency. Equation (2) can be transferred to optimize Equation (3), as follows:

$$\max\frac{1}{\|w\|},\ s.t.\ y_i(w^T\varphi(x^i)+b)\geq 1, i=1,\ldots,n \tag{3}$$

where n is the number of training samples. The condition ($y_i(w^T\varphi(x^i)+b)\geq 1$) should be satisfied in Equation (3), which means that the samples must be classified correctly by the optimized hyperplane. However, the problem of overfitting will be caused. Soft SVM is proposed to tackle the problem. The objective function is refined into Equation (4), as follows:

$$\begin{gathered}\min_{w,b}(\tfrac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\delta_i)\\ s.t.\ y_i(w^T\varphi(x^i)+b)\geq 1-\delta_i, i=1,\ldots,n\\ \delta_i\geq 0\end{gathered} \tag{4}$$

where δ_i is the slack variable and C is the penalty parameter. The SVM used in our work is the package named LIBSVM written by Chang and Lin [63].

3.3. Sequence Representation

AD is classified based on protein sequence information, so, in this paper, we used the features extracted from the peptides. The sequence is represented by a 400-dimension vector, and each dimension describes the frequency of two consecutive amino acids. The feature extraction will be introduced later. To describe the information more clearly, the symbols used in the paper are summarized in Table 2.

Table 2. The symbols used in the present paper.

Symbol	Meaning
P_L	Peptide with L residual
R_i	The i-th residual
f_i	The frequency of the i-th amino acid
F_P	The feature vector of peptide P

P_L is a peptide with L residue, so P_L can be written into a sequence as $\{R_1R_2R_3 \ldots R_i \ldots R_L$. R_i represents the i-th residual of P_L in the sequence. The symbol f_i represents the normalized occurrence frequency of the i-th type of native amino acid in the peptide. There are, in total, 20 types of native amino acids. The peptide P can be represented by $F_p = [f_1, \ldots, f_i, \ldots, f_{20}]$, reflecting the occurrence frequency of every amino acid of P. It is obvious that the sequence information is lost in F_p. To overcome this limitation, we extracted the occurrence frequency of the combination of two consecutive amino acids, such as AR (A and R representing the amino acids). Since there are 20 native amino acids, the number of features of the combination of two consecutive amino acids is 400 (20^2). Thus, we call it a 400D sequence-based feature. The peptide P is straightly represented by $(f_{AA}, f_{AR}, \ldots, f_{VV})$.

3.4. Performance Evaluation

The classification quality is evaluated by accuracy, recall, precision, F-measure, MCC, and AUC. The metrics are used in evaluating the performance frequently [64–72]. In the experiments, n is the number of samples, so n^+ is the number of positive samples and n^- is the number of negative samples. TP (true positive) represents the number of samples that are labeled positive by the method correctly. FP (false positive) is the number of samples that are labeled positive but which are in fact negative. TN (true negative) means the number of sample which are classified correctly as negative sample. FN (false negative) is the number of samples that are positive but which are labeled as negative. The accuracy (ACC_G) represents the correct classification rate of a method G, which is shown in Equation (5). $Precision_G$, $recall_G$ and $F\text{-}measure_G$ are calculated in Equations (5) through (8). AUC is the area size of the ROC curve. The X-axis of ROC curve is the false positive rate, while the Y-axis is true positive rate. The MCC describes the rate of specificity and sensitivity, which is calculated by Equation (9). Specificity and sensitivity are used in evaluating the performance of protein prediction, such as in the case of Feng [47,48] and so on. Specificity (Sp, calculated by Equation (10)) is the rate of misclassification of AD proteins. Sensitivity (Sn, calculated by Equation (11)) is the rate of correctly classified AD proteins:

$$ACC_G = \frac{TP + TN}{n^+ + n^-} \tag{5}$$

$$Precision_G = \frac{TP}{TP + FP} \tag{6}$$

$$Recall_G = \frac{TP}{TP + FN} \tag{7}$$

$$F - measure_G = \frac{(1 + b^2) \times P \times R}{b^2 \times P + R} \tag{8}$$

$$MCC = \frac{Sp}{Sn} \tag{9}$$

$$Sp = \frac{TN}{TN + FP} \tag{10}$$

$$Sn = \frac{TP}{TP + FN} \tag{11}$$

4. Conclusions

In this paper, a computational method based on protein sequence information was introduced to predict the onset of AD. In our proposed method, the sequences are represented by the frequency of two consecutive amino acids, and then the data are classified by SVM. Our work is different from previous work that was completed using MRI, which is time-consuming and expensive. As demonstrated by the presented experimental results, the classification accuracy of our proposed method is 85.7%. Moreover, a dataset used for AD classification was created in our work. In future work, we will try to mine the relationships between the features to improve the classification performance of the predictions method. Furthermore, due to the wide use of webservers in bioinformatics, such as the work of RNA secondary structure comparison [73], we will also develop the a webserver for AD prediction.

Author Contributions: L.X. initially drafted the manuscript and did most of the codes work and the experiments. C.L. collected the features and analyzed the experiments. G.L., G.-D.C. and C.-C.C. revised to draft the manuscript. All authors read and approved the final manuscript.

Funding: This research was funded by the Natural Science Foundation of Guandong Province (grant no. 2018A0303130084), the Science and Technology Innovation Commission of Shenzhen (grant nos. JCYJ20160523113602609, JCYJ20170818100431895), the Grant of Shenzhen Polytechnic (grant no. 601822K19011) and National Nature Science Foundation of China (grant no. 61575128), Chung Shan Medical University Hospital (grant no. CSH-2018-D-002), and Research projects of Shenzhen Institute of Information Technology (No. ZY201714).

Acknowledgments: We thank the reviewers for their great comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Brookmeyer, R.; Johnson, E.; Zieglergraham, K.; Arrighi, H.M. Forecasting the global burden of alzheimer's disease. *Alzheimers Dement.* **2007**, *3*, 186–191. [CrossRef] [PubMed]
2. Yang, J.; Huang, T.; Petralia, F.; Long, Q.; Zhang, B.; Argmann, C.; Zhao, Y.; Mobbs, C.V.; Schadt, E.E.; Zhu, J.; et al. Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Sci. Rep.* **2015**, *5*, 15145. [CrossRef] [PubMed]
3. Mattsson, N. Csf biomarkers and incipient alzheimer disease in patients with mild cognitive impairment. *JAMA* **2009**, *302*, 385. [CrossRef] [PubMed]
4. McEvoy, L.K.; Fennema-Notestine, C.; Roddey, J.C.; Hagler, D.J.; Holland, D.; Karow, D.S.; Pung, C.J.; Brewer, J.B.; Dale, A.M. Alzheimer disease: Quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment. *Radiology* **2009**, *251*, 195–205. [CrossRef] [PubMed]
5. Du, A.T.; Schuff, N.; Kramer, J.H.; Rosen, H.J.; Gorno-Tempini, M.L.; Rankin, K.; Miller, B.L.; Weiner, M.W. Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. *Brain* **2007**, *130*, 1159–1166. [CrossRef] [PubMed]
6. Fjell, A.M.; Walhovd, K.B.; Fennema-Notestine, C.; McEvoy, L.K.; Hagler, D.J.; Holland, D.; Brewer, J.B.; Dale, A.M. Csf biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and alzheimer's disease. *J. Neurosci.* **2010**, *30*, 2088–2101. [CrossRef] [PubMed]
7. de Leon, M.J.; Mosconi, L.; Li, J.; De Santi, S.; Yao, Y.; Tsui, W.H.; Pirraglia, E.; Rich, K.; Javier, E.; Brys, M.; et al. Longitudinal CSF isoprostane and MRI atrophy in the progression to AD. *J. Neurol.* **2007**, *254*, 1666–1675. [CrossRef] [PubMed]
8. Morris, J.C.; Storandt, M.; Miller, J.P.; McKeel, D.W.; Price, J.L.; Rubin, E.H.; Berg, L. Mild cognitive impairment represents early-stage Alzheimer disease. *Arch. Neurol.* **2001**, *58*, 397–405. [CrossRef] [PubMed]

9. De, S.S.; de Leon, M.J.; Rusinek, H.; Convit, A.; Tarshish, C.Y.; Roche, A.; Tsui, W.H.; Kandil, E.; Boppana, M.; Daisley, K.; Wang, G.J.; et al. Hippocampal formation glucose metabolism and volume losses in MCI and AD. *Neurobiol. Aging* **2001**, *22*, 529–539.
10. Bouwman, F.H.; van der Flier, W.M.; Schoonenboom, N.S.; van Elk, E.J.; Kok, A.; Rijmen, F.; Blankenstein, M.A.; Scheltens, P. Longitudinal changes of CSF biomarkers in memory clinic patients. *Neurology* **2007**, *69*, 1006–1011. [CrossRef] [PubMed]
11. Shaw, L.M.; Vanderstichele, H.; Knapik-Czajka, M.; Clark, C.M.; Aisen, P.S.; Petersen, R.C.; Blennow, K.; Soares, H.; Simon, A.; Lewczuk, P.; Dean, R.; Siemers, E.; Potter, W.; Lee, V.M.-Y.; Trojanowski, J.Q. Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects. *Ann. Neurol.* **2009**, *65*, 403–413. [CrossRef] [PubMed]
12. Zu, C.; Jie, B.; Liu, M.; Chen, S.; Shen, D.; Zhang, D. Label-aligned multi-task feature learning for multimodal classification of alzheimer's disease and mild cognitive impairment. *Brain Imaging Behav.* **2015**, *10*, 1148–1159. [CrossRef] [PubMed]
13. Xu, H.J.; Hu, S.X.; Cagle, P.T.; Moore, G.E.; Benedict, W.F. Absence of retinoblastoma protein expression in primary non-small cell lung carcinomas. *Cancer Res.* **1991**, *8*, 2735–2739.
14. Foster, N.L.; Heidebrink, J.L.; Clark, C.M.; Jagust, W.J.; Arnold, S.E.; Barbas, N.R.; DeCarli, C.S.; Turner, R.S.; Koeppe, R.A.; Higdon, R.; Minoshima, S. FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. *Brain* **2007**, *130*, 2616–2635. [CrossRef] [PubMed]
15. Dai, Z.; Yan, C.; Wang, Z.; Wang, J.; Xia, M.; Li, K.; He, Y. Discriminative analysis of early alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (m3). *NeuroImage* **2012**, *59*, 2187–2195. [CrossRef] [PubMed]
16. Huang, S.; Li, J.; Ye, J.; Wu, T.; Chen, K.; Fleisher, A.; Reiman, E. Identifying Alzheimer's Disease-Related Brain Regions from Multi-Modality Neuroimaging Data using Sparse Composite Linear Discrimination Analysis. *Adv. Neural Inf. Process. Syst.* **2011**, 1431–1439.
17. Westman, E.; Muehlboeck, J.-S.; Simmons, A. Combining mri and csf measures for classification of alzheimer's disease and prediction of mild cognitive impairment conversion. *NeuroImage* **2012**, *62*, 229–238. [CrossRef] [PubMed]
18. Liu, F.; Shen, C. Learning Deep Convolutional Features for MRI Based Alzheimer's Disease Classification. *arXiv*, 2014; arXiv:1404.3366.
19. Herrera, L.J.; Rojas, I.; Pomares, H.; Guillén, A.; Valenzuela, O.; Baños, O. Classification of MRI Images for Alzheimer's Disease Detection. In Proceedings of the 2013 International Conference on Social Computing, Alexandria, VA, USA, 8–14 September 2013.
20. Liu, X.; Tosun, D.; Weiner, M.W.; Schuff, N. Locally linear embedding (LLE) for MRI based Alzheimer's disease classification. *Neuroimage* **2013**, *83*, 148–157. [CrossRef] [PubMed]
21. Gerardin, E.; Chételat, G.; Chupin, M.; Cuingnet, R.; Desgranges, B.; Kim, H.-S.; Niethammer, M.; Dubois, B.; Lehéricy, S.; Garnero, L.; Eustache, F.; Colliot, O. Multidimensional classification of hippocampal shape features discriminates alzheimer's disease and mild cognitive impairment from normal aging. *NeuroImage* **2009**, *47*, 1476–1486. [CrossRef] [PubMed]
22. West, M.J.; Kawas, C.H.; Stewart, W.F.; Rudow, G.L.; Troncoso, J.C. Hippocampal neurons in pre-clinical Alzheimer's disease. *Neurobiol. Aging* **2004**, *25*, 1205–1212. [CrossRef] [PubMed]
23. Desikan, R.S.; Cabral, H.J.; Hess, C.P.; Dillon, W.P.; Glastonbury, C.M.; Weiner, M.W.; Schmansky, N.J.; Greve, D.N.; Salat, D.H.; Buckner, R.L.; Fischl, B. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain* **2009**, *132*, 2048–2057. [CrossRef] [PubMed]
24. Oliveira, P.P., Jr.; Nitrini, R.; Busatto, G.; Buchpiguel, C.; Sato, J.R.; Amaro, E., Jr. Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease. *J. Alzheimers Dis.* **2010**, *19*, 1263–1272. [CrossRef] [PubMed]
25. Fan, Y.; Shen, D.; Gur, R.C.; Gur, R.E.; Davatzikos, C. COMPARE: Classification of morphological patterns using adaptive regional elements. *IEEE Trans. Med. Imaging* **2007**, *26*, 93–105. [CrossRef] [PubMed]
26. Magnin, B.; Mesrob, L.; Kinkingnéhun, S.; Pélégrini-Issac, M.; Colliot, O.; Sarazin, M.; Dubois, B.; Lehéricy, S.; Benali, H. Support vector machine-based classification of alzheimer's disease from whole-brain anatomical mri. *Neuroradiology* **2008**, *51*, 73–83. [CrossRef] [PubMed]

27. Chetelat, G.; Desgranges, B.; de la Sayette, V.; Viader, F.; Eustache, F.; Baron, J.-C. Mild cognitive impairment: Can FDG-PET predict who is to rapidly convert to Alzheimer's disease? *Neurology* **2003**, *60*, 1374–1377. [CrossRef] [PubMed]
28. Higdon, R.; Foster, N.L.; Koeppe, R.A.; DeCarli, C.S.; Jagust, W.J.; Clark, C.M.; Barbas, N.R.; Arnold, S.E.; Turner, R.S.; Heidebrink, J.L.; Minoshima, S. A comparison of classification methods for differentiating fronto-temporal dementia from alzheimer's disease using fdg-pet imaging. *Stat. Med.* **2004**, *23*, 315–326. [CrossRef] [PubMed]
29. Zeng, X.; Liao, Y.; Liu, Y.; Zou, Q. Prediction and Validation of Disease Genes Using HeteSim Scores. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *14*, 687–695. [CrossRef] [PubMed]
30. Liu, Y.; Zeng, X.; He, Z.; Zou, Q. Inferring microrna-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *14*, 905–915. [CrossRef] [PubMed]
31. Zeng, X.; Liu, L.; Lü, L.; Zou, Q. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* **2018**, *34*, 2425–2432. [CrossRef] [PubMed]
32. Wang, L.; Ping, P.; Kuang, L.; Ye, S.; Pei, T. A Novel Approach Based on Bipartite Network to Predict Human Microbe-Disease Associations. *Curr. Bioinform.* **2018**, *13*, 141–148. [CrossRef]
33. Liao, Z.J.; Li, D.; Wang, X.; Li, L.; Zou, Q. Cancer Diagnosis Through IsomiR Expression with Machine Learning Method. *Curr. Bioinform.* **2018**, *13*, 57–63. [CrossRef]
34. Yang, J.; Huang, T.; Song, W.M.; Petralia, F.; Mobbs, C.V.; Zhang, B.; Zhao, Y.; Schadt, E.E.; Zhu, Y.; Tu, Z. Discover the network mechanisms underlying the connections between aging and age-related diseases. *Sci. Rep.* **2016**, 6. [CrossRef] [PubMed]
35. Xiao, X.; Zhu, W.; Liao, B.; Xu, J.; Gu, C.; Ji, B.; Yao, Y.; Peng, L.; Yang, J. BPLLDA: Predicting lncRNA-Disease Associations Based on Simple Paths with Limited Lengths in a Heterogeneous Network. *Front. Genet.* **2018**, *9*. [CrossRef] [PubMed]
36. Lu, M.; Xu, X.; Xi, B.; Dai, Q.; Li, C.; Su, L.; Zhou, X.; Tang, M.; Yao, Y.; Yang, J. Molecular Network-Based Identification of Competing Endogenous RNAs in Thyroid Carcinoma. *Genes* **2018**, *9*, 44. [CrossRef] [PubMed]
37. Liu, G.; Wang, T.; Tian, R.; Hu, Y.; Han, Z.; Wang, P.; Zhou, W.; Ren, P.; Zong, J.; Jin, S.; Jiang, Q. Alzheimer's Disease Risk Variant rs2373115 Regulates GAB2 and NARS2 Expression in Human Brain Tissues. *J. Mol. Neurosci.* **2018**, *66*, 37–43. [CrossRef] [PubMed]
38. Jiang, Q.; Jin, S.; Jiang, Y.; Liao, M.; Feng, R.; Zhang, L.; Liu, G.; Hao, J. Alzheimer's disease variants with the genome-wide significance are significantly enriched in immune pathways and active in immune cells. *Mol. Neurobiol.* **2016**, *54*, 594–600. [CrossRef] [PubMed]
39. Liu, G.; Xu, Y.; Jiang, Y.; Zhang, L.; Feng, R.; Jiang, Q. Picalm rs3851179 variant confers susceptibility to alzheimer's disease in chinese population. *Mol. Neurobiol.* **2017**, *54*, 3131–3136. [CrossRef] [PubMed]
40. Wei, L.; Luan, S.; Nagai, L.A.E.; Su, R.; Zou, Q. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* **2018**. [CrossRef] [PubMed]
41. Guo, S.-H.; Deng, E.-Z.; Xu, L.-Q.; Ding, H.; Lin, H.; Chen, W.; Chou, K.-C. Inuc-pseknc: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* **2014**, *30*, 1522–1529. [CrossRef] [PubMed]
42. Wei, L.; Xing, P.; Tang, J.; Zou, Q. Phospred-rf: A novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE Trans. NanoBiosci.* **2017**, *16*, 240–247. [CrossRef] [PubMed]
43. Lin, C.; Chen, W.; Qiu, C.; Wu, Y.; Krishnan, S.; Zou, Q. Libd3c: Ensemble classifiers with a clustering and dynamic selection strategy. *Neurocomputing* **2014**, *123*, 424–435. [CrossRef]
44. Available online: https://www.uniprot.org (accessed on 1 January 2007).
45. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. Cd-hit: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]
46. Liu, B.; Fang, L.; Liu, F.; Wang, X.; Chen, J.; Chou, K.C. Identification of Real MicroRNA Precursors with a Pseudo Structure Status Composition Approach. *PLoS ONE* **2015**, *10*. [CrossRef] [PubMed]
47. Feng, P.; Chen, W.; Lin, H. Identifying antioxidant proteins by using optimal dipeptide compositions. *Interdiscip. Sci. Comput. Life Sci.* **2015**, *8*, 186–191. [CrossRef] [PubMed]

48. Feng, P.M.; Lin, H.; Chen, W. Identification of Antioxidants from Sequence Information Using Naïve Bayes. *Comput. Math. Methods Med.* **2013**, *2013*. [CrossRef] [PubMed]
49. Wei, L.; Xing, P.; Shi, G.; Ji, Z.L.; Zou, Q. Fast prediction of methylation sites using sequence-based feature selection technique. *IEEE/ACM Tran. Comput. Biol. Bioinform.* **2017**. [CrossRef] [PubMed]
50. Zhang, N.; Sa, Y.; Guo, Y.; Wang, L.; Wang, P.; Feng, Y. Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine. *Curr. Bioinform.* **2018**, *13*, 50–56. [CrossRef]
51. Chen, W.; Tang, H.; Ye, J.; Lin, H.; Chou, K.C. iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids* **2016**, *5*. [CrossRef]
52. Lin, H.; Deng, E.-Z.; Ding, H.; Chen, W.; Chou, K.-C. Ipro54-pseknc: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* **2014**, *42*, 12961–12972. [CrossRef] [PubMed]
53. Tseng, C.J.; Lu, C.J.; Chang, C.C.; Chen, G.D. Application of machine learning to predict the recurrence-proneness for cervical cancer. *Neural Comput. Appl.* **2014**, *24*, 1311–1316. [CrossRef]
54. Tseng, C.J.; Lu, C.J.; Chang, C.C.; Chen, G.D.; Cheewakriangkrai, C. Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. *Artif. Intell. Med.* **2017**, *78*, 47–54. [CrossRef] [PubMed]
55. Cheng, C.S.; Shueng, P.W.; Chang, C.C.; Kuo, C.W. Adapting an Evidence-based Diagnostic Model for Predicting Recurrence Risk Factors of Oral Cancer. *J. Univers. Comput. Sci.* **2018**, *24*, 742–752.
56. Zou, Q.C.; Chen, C.W.; Chang, H.C.; Chu, Y.W. Identifying Cleavage Sites of Gelatinases A and B by Integrating Feature Computing Models. *J. Univers. Comput. Sci.* **2018**, *24*, 711–724.
57. Ye, L.L.; Lee, T.S.; Chi, R. Hybrid Machine Learning Scheme to Analyze the Risk Factors of Breast Cancer Outcome in Patients with Diabetes Mellitus. *J. Univers. Comput. Sci.* **2018**, *24*, 665–681.
58. Das, A.K.; Pati, S.K.; Huang, H.H.; Chen, C.K. Cancer Classification by Gene Subset Selection from Microarray Dataset. *J. Univers. Comput. Sci.* **2018**, *24*, 682–710.
59. Xu, L.; Liang, G.; Wang, L.; Liao, C. A Novel Hybrid Sequence-Based Model for Identifying Anticancer Peptides. *Genes* **2018**, *9*, 158. [CrossRef] [PubMed]
60. Chou, K.-C. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* **2002**, *277*, 45765–45769. [CrossRef] [PubMed]
61. Cai, Y.-D.; Zhou, G.-P.; Chou, K.-C. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.* **2003**, *84*, 3257–3263. [CrossRef]
62. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000; pp. 1–28.
63. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM TIST* **2011**, *2*, 1–27. [CrossRef]
64. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef] [PubMed]
65. Chou, K.C. Using subsite coupling to predict signal peptides. *Protein Eng.* **2001**, *14*, 75–79. [CrossRef] [PubMed]
66. Lai, H.Y.; Chen, X.X.; Chen, W.; Tang, H.; Lin, H. Sequence-based predictive modeling to identify cancerlectins. *Oncotarget* **2017**, *8*, 28169–28175. [CrossRef] [PubMed]
67. Su, R.; Wu, H.; Xu, B.; Liu, X.; Wei, L. Developing a Multi-Dose Computational Model for Drug-induced Hepatotoxicity Prediction based on Toxicogenomics Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**. [CrossRef] [PubMed]
68. Wei, L.; Xing, P.; Zeng, J.; Chen, J.; Su, R.; Guo, F. Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* **2017**, *83*, 67–74. [CrossRef] [PubMed]
69. Wei, L.; Ding, Y.; Su, R.; Tang, J.; Zou, Q. Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.* **2018**, *117*, 212–217. [CrossRef]
70. Wei, L.; Chen, H.; Su, R. M6APred-EL: A Sequence-Based Predictor for Identifying N6-methyladenosine Sites Using Ensemble Learning. *Mol. Ther. Nucleic Acids* **2018**, *12*, 635–644. [CrossRef] [PubMed]
71. Liu, X.; Yang, J.; Zhang, Y.; Fang, Y.; Wang, F.; Wang, J.; Zheng, X.; Yang, J. A systematic study on drug-response associated genes using baseline gene expressions of the cancer cell line encyclopedia. *Sci. Rep.* **2016**, *6*, 22811. [CrossRef] [PubMed]

72. Xu, L.; Liang, G.; Shi, S.; Liao, C. SeqSVM: A Sequence-Based Support Vector Machine Method for Identifying Antioxidant Proteins. *Int. J. Mol. Sci.* **2018**, *19*, 1773. [CrossRef] [PubMed]
73. Li, Y.; Shi, X.; Liang, Y.; Xie, J.; Zhang, Y.; Ma, Q. RNA-TVcurve: A Web server for RNA secondary structure comparison based on a multi-scale similarity of its triple vector curve representation. *BMC Bioinform.* **2017**, *18*, 51. [CrossRef] [PubMed]

Sample Availability: Samples of the compounds are not available.