JCB*i*   JOURNAL OF
CLINICAL BIOINFORMATICS

METHODOLOGY   Open Access

A diagnostic methodology for Alzheimer's disease

Wen-Chin Hsu[1,2*], Christopher Denq[1,3] and Su-Shing Chen[1,4]

Abstract

Background: Like all other neurodegenerative diseases, Alzheimer's disease (AD) remains a very challenging and difficult problem for diagnosis and therapy. For many years, only historical, behavioral and psychiatric measures have been available to AD cases. Recently, a definitive diagnostic framework, using biomarkers and imaging, has been proposed. In this paper, we propose a promising diagnostic methodology for the framework.

Methods: In a previous paper, we developed an efficient SVM (Support Vector Machine) based method, which we have now applied to discover important biomarkers and target networks which provide strategies for AD therapy.

Results: The methodology selects a number of blood-based biomarkers (fewer than 10% of initial numbers on three AD datasets from NCBI), and the results are statistically verified by cross-validation. The resulting SVM is a classifier of AD vs. normal subjects. We construct target networks of AD based on MI (mutual information). In addition, a hierarchical clustering is applied on the initial data and clustered genes are visualized in a heatmap. The proposed method also performs gender analysis by classifying subjects based on gender.

Conclusions: Unlike other traditional statistical analyses, our method uses a machine learning-based algorithm. Our method selects a small set of important biomarkers for AD, differentiates noisy (irrelevant) from relevant biomarkers and also provides the target networks of the selected biomarkers, which will be useful for diagnosis and therapeutic design. Finally, based on the gender analysis, we observe that gender could play a role in AD diagnosis.

Keywords: Feature selection, Biomarkers, Target networks, Alzheimer's disease, Support vector machine

Background
Overview

Analysis of Alzheimer's disease (AD) has been a challenging problem for diagnosis and therapy. Currently, a definitive clinical diagnosis can be obtained only by historical, behavioral and psychiatric measures, and only when the patient's condition has sufficiently deteriorated [1]. In [2], a dynamic model was proposed for AD diagnosis and has led to several studies of biomarker-based analysis. However, in order to validate the model, continuous studies of biomarkers are necessary to identify critical time points when changes or permutations of biomarkers occur [3]. The specificity and sensitivity of AD diagnosis remain in doubt due to the lack of comparisons of AD with other neurodegenerative disease [4]. In addition, the standards and guidelines for blood-sample biomarkers are still in the process of development [4]. The current methods for biomarker collection are also problematic, due to the need for expensive instrumentation and the invasiveness of the procedures [5]. Recently, Zhang et al. integrated three modalities--MRI, FDG-PET, and CSF biomarker--into a Multi-Kernel SVM to classify AD vs. normal samples [6].

NIH guidelines

In the past few years, the technologies for both biomarker analysis and imaging have provided promising contributions to definitive diagnoses of AD. In 2011, the NIH National Institute on Aging and the Alzheimer's Association also established new guidelines to allow use of biomarkers and imaging for diagnoses [1]. Since the announcement from NIH, research to identify and compare biomarkers has been thriving.

SVM

High-dimensional pattern classifiers such as SVMs (Support Vector Machines) are adapted to contribute classifications. In [7-9], biomarker selections were performed by SVM-RFE, a feature (biomarker) extraction

* Correspondence: wenchin@ufl.edu
[1]Systems Biology Lab, University of Florida, Gainesville, FL 32611, USA
[2]Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA
Full list of author information is available at the end of the article

Figure 1 Distributions of original pair-wise MI values and permuted pair-wise MI values.
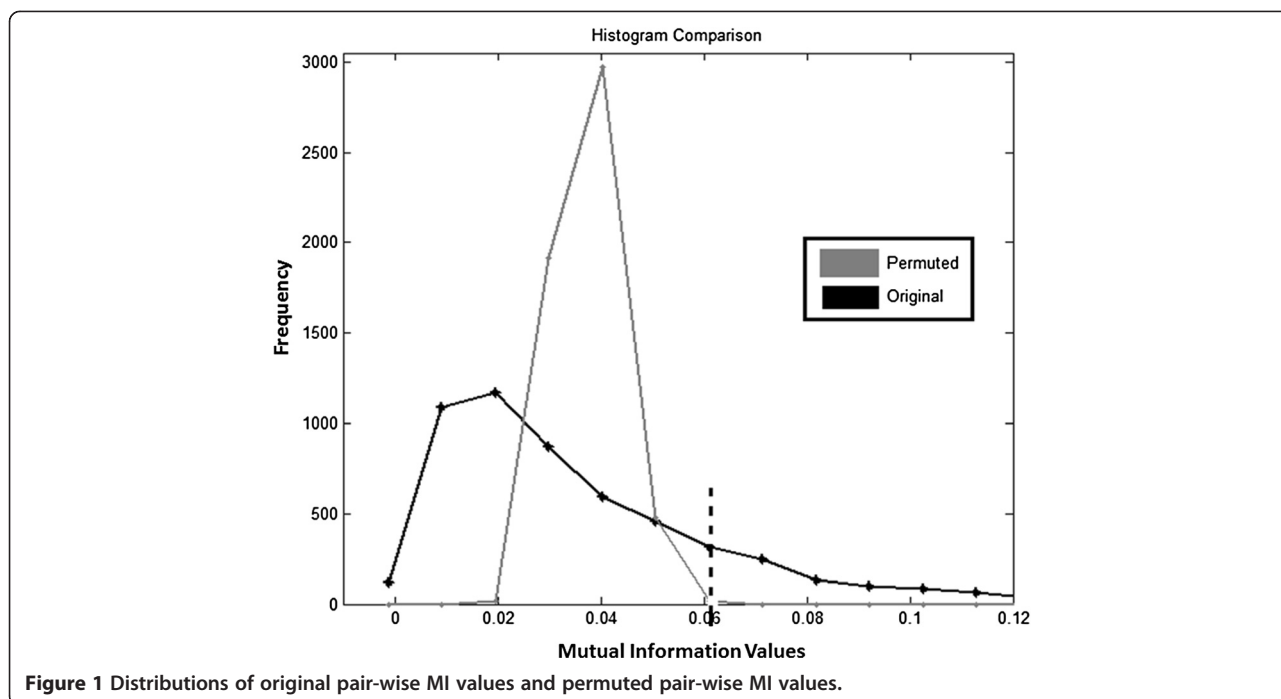
method, by reclusively eliminating features based on the validation accuracy of SVM [10]. During the selection process, the least useful feature is iteratively removed from the original subset. However, a group of weak features can still construct a strong classifier [10]. Once a feature is removed from the original subset, it cannot be evaluated by different combinations of the remaining features. Thus, the SVM-RFE approach usually suffers from selection of a sub-optimal subset since the classification ability of features should be evaluated by subsets instead of by individuals.

AMFES

In [11], we first proposed an efficient algorithm, AMFES (Adaptive Multiple FEatues Selection), to select important biomarkers for cancers. Based on that initial success, this paper reports the extension of previous results on the datasets provided by Maes et al. in an attempt to discover important biomarkers for AD from the blood-based samples [12]. Unlike traditional statistical analyses, AMFES is an SVM-based methodology, which can select a much smaller subset of important biomarkers. In addition, AMFES applies an adaptive method which enables selection of a globally optimal subset of important

biomarkers compared to SVM-RFE. AMFES is particularly useful for differentiating noisy biomarkers from the relevant ones when interferences between biomarkers exist. Our results are supported by a high ROC/AUC (Receiver Operating Characteristic/Area Under Curve) value when we apply a cross-validation verification. Thus, AMFES should play an important role in the classification framework of multi-modalities proposed by Zhang et al. in [6]. In this paper, we shall develop the details of AMFES for blood-based biomarkers.

The target networks of AD with statistical dependencies (mutual information) are constructed by these selected biomarkers. The resulting AD target network is characterized as a signature of the disease, and will enable a more detailed diagnosis. In addition, the MI (Mutual Information) values of AD subjects are found to be lower than those of normal subjects. Based on our method and results, a promising framework for definitive diagnosis is proposed.

The organization of this paper is as follows: The Methods section describes AMFES [11], as well as the computations of mutual information between two biomarkers and the construction of the target networks. In the Results section,

Table 1 Descriptions of 3 datasets: GSE4226, GSE4227, and GSE4229

	GSE4226	GSE4227	GSE4229
Number of Biomarkers	9600	9600	9600
Type of Biomarkers	RNAs	RNAs	RNAs
Number of Samples	28 (14 AD vs. 14 Normal)	34(14 AD vs. 18 normal)	40(18 AD vs. 22 normal)

Table 2 Results of selected subsets of genes

	GES4226	GSE4227	GSE4229
Number of Biomarkers Selected	74	52	395

we describe the PBMC (Peripheral Blood Mononuclear Cells) datasets of sporadic AD: GSE4226, GSE4227, and GSE4229 [12-14]. In addition, we describe the biomarkers and target networks of AD selected according to our approach on these 3 datasets.

Methods
AMFES

Selecting a small subset out of hundreds and thousands of features is always a challenging task due to the COD (Curse of Dimensionality) problem for microarray datasets. To tackle this problem, we use a gene selection methodology, AMFES, to select an optimal subset of genes by training an SVM with subsets of genes generated adaptively [11]. AMFES is developed based on two fundamental processes, ranking and selection.

The gene ranking process contains several stages. In the first stage, all genes are ranked by their ranking scores in a descending order. Then, in the next stage, only the top half of the ranked genes are ranked again, while the bottom half holds the current order in the subsequent stage. The same iteration repeats recursively until only three genes remain to be ranked again to complete one ranking process.

Assume at a given ranking stage, there are k genes indexed from 1 to k. To rank these k genes, we follow 4 steps below. (I) We first generate m independent subsets $S_1 \ldots S_m$. Each subset S_i, $i = 1, 2 \ldots m$, has j genes which are selected randomly and independently from the k genes (II) Let C_i be the SVM classifier that is trained on each subset of genes, $i = 1, 2 \ldots m$. For each gene of k genes, we compute the ranking score $\theta_m(g)$ of the gene g, as equation (1) below [11]. (III) We use the average weight of gene g, given by the summation of weights of g in m subsets divided by the number of subsets for which g is randomly selected. The $weight_i(g)$ is then defined as

the change in the objective function due to g as equation (2) [11] and the m value is obtained when θ_m satisfies the equation (3) in [11]. This increases the robustness to represent the true classifying ability of gene g. (IV) The k genes are then ranked in descending order by their ranking scores.

$$\theta_m(g) = \frac{\sum_{i=1}^{m} I_{\{g\ \in S_i\}} weight_i(g)}{\sum_{i-1}^{m} I_{\{g\ \in S_i\}}} \quad (1)$$

where I is an indicator function, such that $I_{proposition} = 1$ if the proposition is true; otherwise, $I_{proposition} = 0$. In other words, if gene g is randomly selected for the subset S_i, it is denoted as $g \epsilon S_i$ and $I_{proposition} = 1$.

We denote the objective function of C_i as $obj_i(v_1, v_2, \ldots, v_5)$ where $v_1, v_2 \ldots v_s$ are support vectors of C_i. The $weight_i(g)$ is then defined as the change in the objective function due to g, i.e., [6-8].

$$weight_i(g) = \left| obj_i(v_1, v_2, \ldots v_s) - obj_i\left(v_1^{(g)}, v_2^{(g)}, \ldots, v_3^{(g)}\right) \right| \quad (2)$$

Note that if $\mathbf{v}$ is a vector, $\mathbf{v}^{(g)}$ is the vector obtained by dropping gene g from $\mathbf{v}$. Let θ_m be a vector comprising the ranking scores derived from the m gene subsets generated thus far and θ_{m-1} be the vector at the previous stage. The m value is determined when θ_m satisfies equation (3) by adding a gene to an empty subset once a time.

$$\frac{\|\boldsymbol{\theta}_{m-1} - \boldsymbol{\theta}_m\|^2}{\|\boldsymbol{\theta}_{m-1}\|^2} < 0.01 \quad (3)$$

where $\|\theta\|$ is understood as the Euclidean norm of vector θ.

The ranking process is performed by ranking both artificial and original features together. The use of artificial features has been demonstrated as a useful tool to distinguish the relevant features from the irrelevant ones, as in [15-17]. When a set of genes is given, we generate artificial
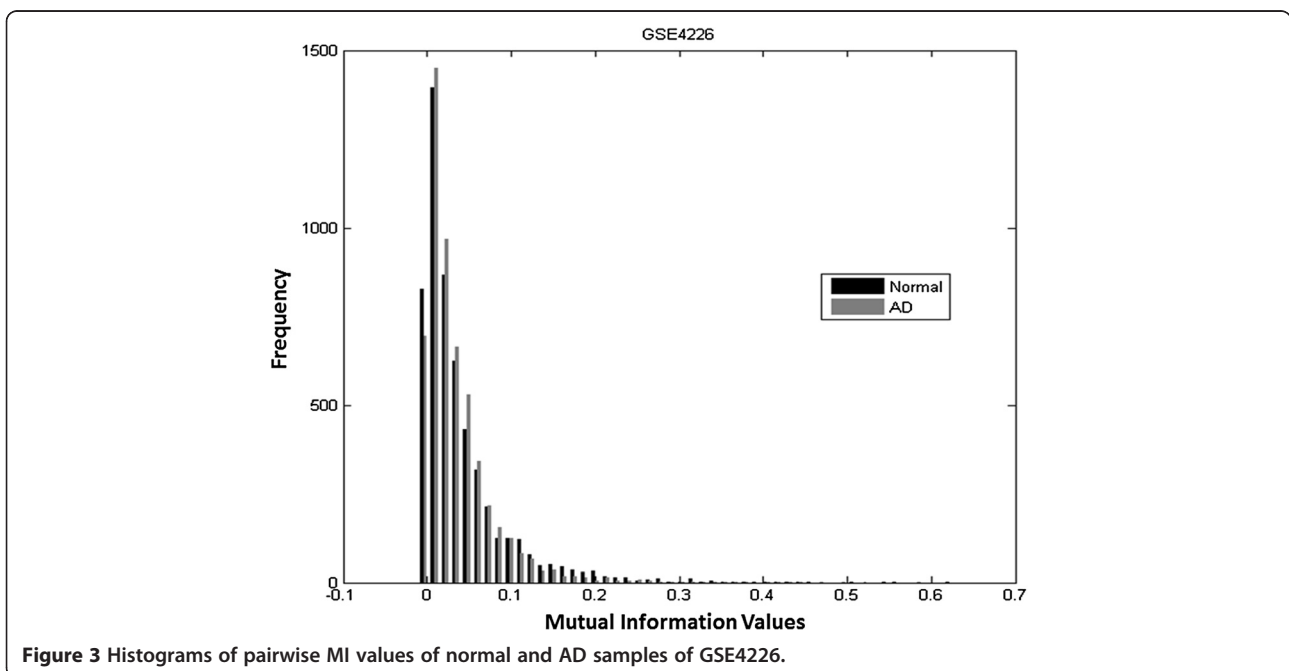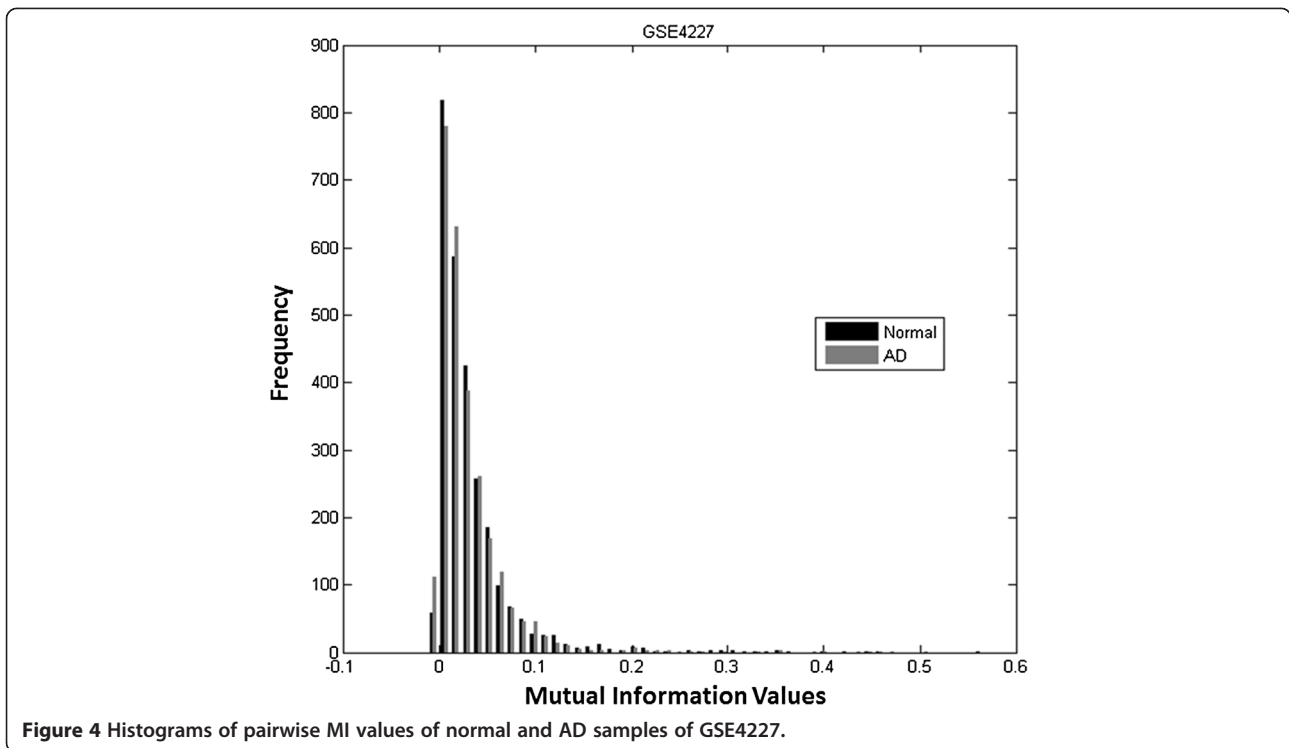
Table 3 Results of analysis of MI matrices

	Mean value of MI	Standard deviation of MI	Num of positive values	Num of negative values	Min value	Max value
GSE4226_normal	0.0245	0.0793	4636	840	−0.0037	1.4451
GSE4226_AD	0.0225	0.0804	4382	1094	−0.0032	1.7946
GSE4227_normal	0.0086	0.0487	2086	618	−0.0029	1.5557
GSE4227_AD	0.0076	0.0466	1972	732	−0.0035	1.4860
GSE4229_normal	0.000095	0.0045	100275	55750	−0.0041	1.1717
GSE4229_AD	0.000081	0.0045	97987	58038	−0.0036	1.5503

Figure 2 ROC curve of selected genes of GSE4226, 74 genes.
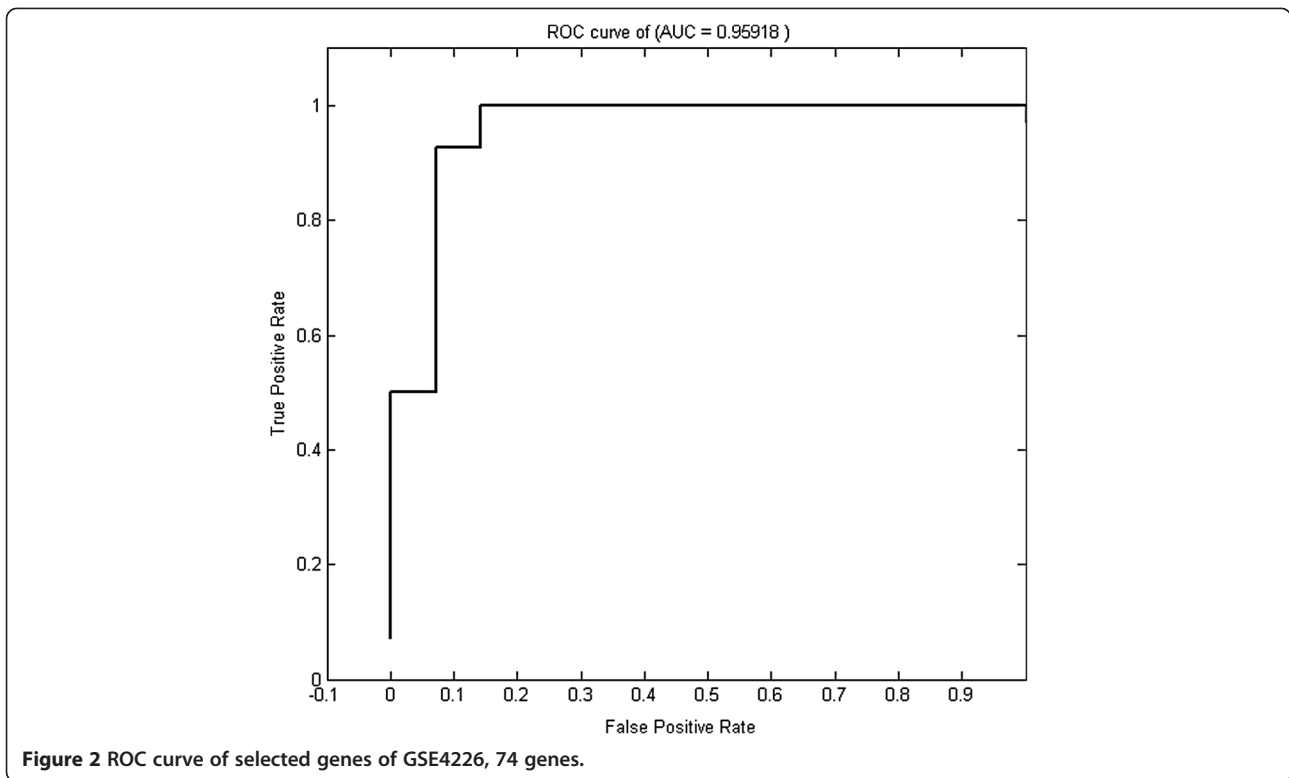
genes and rank them together with original ones. After finishing the ranking of the set, we assign a gene-index to each original gene by the proportion of artificial ones that are ranked above it, where the gene-index is a real numerical value between 0 and 1. Then, we generate a few subset candidates from which the optimal subset is chosen.

Each subset has a subset value, p_i, and it contains original genes whose indices are smaller than or equal to p_i [11]. We train an SVM on every subset, and compute its validation accuracy $v(p_i)$. We stop at the first p_k at which its validation accuracy is better than baseline (i.e., the case in which all features are involved in training [11]).

Figure 3 Histograms of pairwise MI values of normal and AD samples of GSE4226.

Figure 4 Histograms of pairwise MI values of normal and AD samples of GSE4227.

When starting to apply AMFES, we first divide all samples into either learning samples or testing samples. Then, we randomly extract r training-validation pairs from the learning samples according to the heuristic rule $r = \max\left(5, (\mathrm{int})\frac{500}{n+0.5}\right)$, where n is the number of learning samples in the dataset. The heuristic ratio and rule are chosen based on experience of the balance of time consumption and performance. The ranking and selection processes from previous sections correspond to one training-validation pair. To increase the reliability of validation, we generate r pairs to find the optimal subset.

Figure 5 Histograms of pairwise MI values of normal and AD samples of GSE4229.
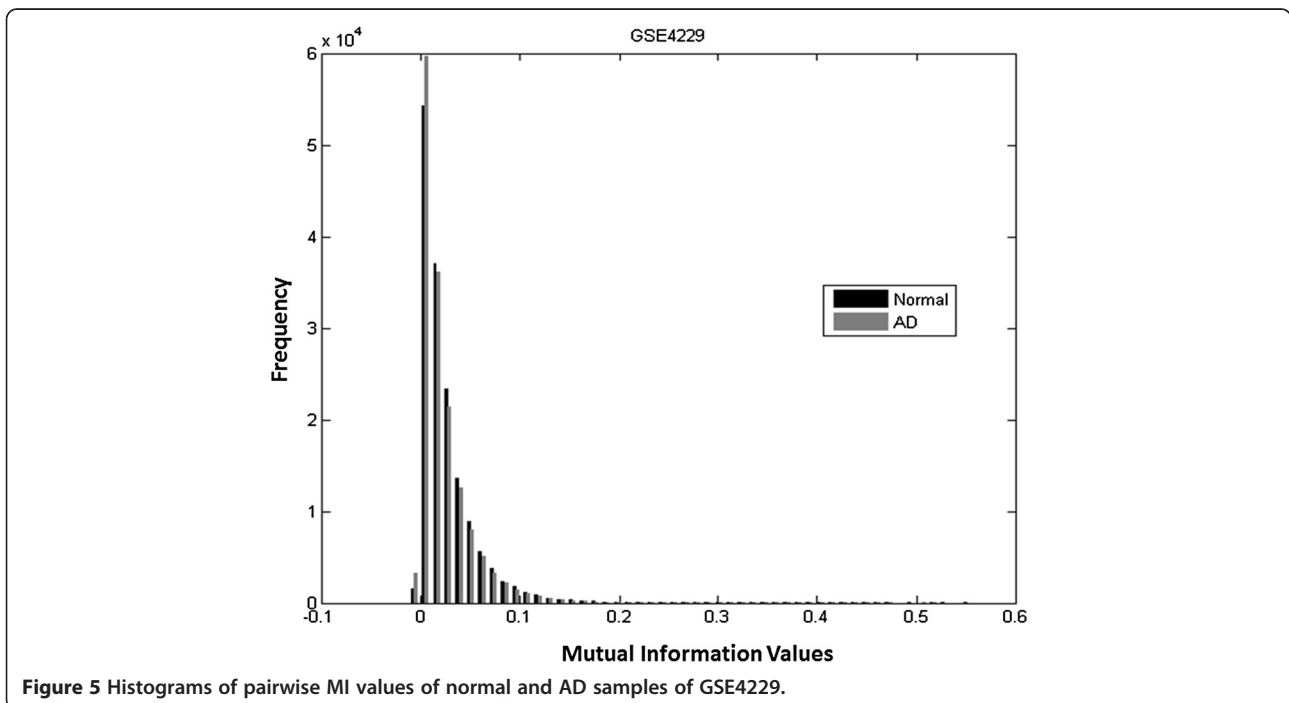
We calculate the validation accuracy of all pairs and the average accuracy, $av(p_i)$. Then, we perform the subset search as explained in the previous section to find the optimal p_i value, denoted as p^*. However, p^* is a derived value and does not belong to a unique subset. Thus, we have to adapt all training samples and repeat the entire process in order to find a unique subset.

We generate artificial genes and rank them together with the original genes. Finally, we select the original genes whose indices are smaller than or equal to the p^* derived previously as the subset of genes we select for the dataset [11].

Mutual information

To treat a complex disease or injury such as AD, an optimal approach is to discover important biomarkers for which we can find certain treatments. These biomarkers form a certain dependency network as a framework for diagnosis and therapy [18]. We call such a network a target network of these biomarkers [11].

Mutual information has been used to measure the dependency between two random variables. Assume the two random variables X and Y are continuous numbers. The mutual information is defined as [19]:

$$I(X, Y) = \iint f(x, y) \log\left(\frac{f(x, y)}{f(x)f(y)}\right) dx\, dy \tag{7}$$

where $f(x,y)$ denotes the joint probability distribution, and $f(x)$ and $f(y)$ denote the marginal probability distributions

of X and Y. By using the Gaussian kernel estimation, the $f(x, y), f(x)$ and $f(y)$ can be further represented as [20]:

$$f(x,y) = \frac{1}{M} \frac{1}{2\pi h^2} \sum e^{-\frac{1}{2h^2}\left((x-x_u)^2 + (y-y_u^2)\right)} \tag{8}$$

$$f(x) = \frac{1}{M} \sum \frac{1}{\sqrt{2\pi h^2}} e^{\frac{1}{2h^2}(x-x_u)^2} \tag{9}$$

$$f(y) = \frac{1}{M} \sum \frac{1}{\sqrt{2\pi h^2}} e^{\frac{1}{2h^2}(y-y_u)^2} \tag{10}$$

where M represents the number of samples for both X and Y, u is index of samples $u = 1, 2, \ldots M$, and h is a parameter controlling the width of the kernels. Thus, the mutual information $I(X,Y)$ can then be represented as:

$$I(X, Y) = \frac{1}{M} \sum_i \log \frac{M \sum_i e^{-\frac{1}{2h^2}\left((x_w-x_u)^2 + (y_w-y_u)^2\right)}}{\sum_j e^{-\frac{1}{2h^2}(x_w-x_u)^2} \sum_j e^{-\frac{1}{2h^2}(y_w-y_u)^2}} \tag{11}$$

where both w, u are indices of samples $w, u = 1, 2, \ldots M$.

Computation of pairwise genes of a microarray dataset usually involves a nested loops calculation which takes a dramatic amount of time. Assume a dataset has N genes and each gene has M samples. To calculate the pairwise mutual information values, the computation usually first finds the kernel distance between any two samples for a given gene. Then, the same process goes through every
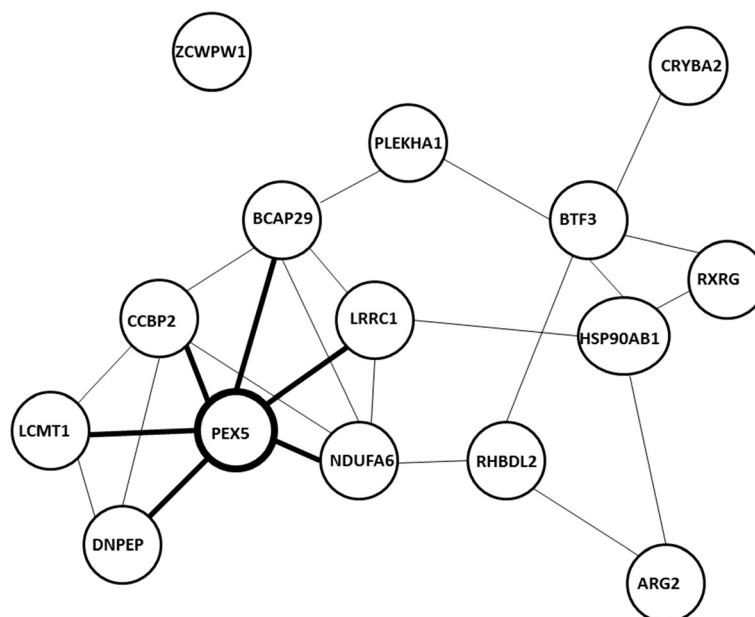

Figure 6 A target network of top ranked 15 genes selected by AMFES for GSE4226.

pair of genes in the dataset. In order to be computationally efficient, two improvements are applied [21]. The first one is to calculate the marginal probability of each gene in advance and use it repeatedly during the process [21,22]. The second improvement is to move the summation of each sample pair for a given gene to the most outer for-loop rather than inside a nested for-loop for every pairwise gene. As a result, the kernel distance between two samples is only calculated twice instead N times, thereby saving considerable computational time. LNO (Loops Nest Optimization) which changes the order of nested loops is a common time-saving technique in computer science field [23].

Target network

In our approach, a constructed target network is represented by an undirected graph in which nodes represent genes in the system and edges represent the dependency between gene pairs [18]. For each gene pair, we use MI (Mutual Information) to measure the dependency between them and represent the weight of linkages. Assuming that the graph contains N nodes (genes), there should be $\frac{N*(N-1)}{2}$ pairwise MI values for all genetic pairs. An adjacency matrix of N × N elements is used to hold MI values of all the linkages in the graph. The adjacency matrix can be visualized as a heatmap. In addition, hierarchical clustering is often used to help verify the dependency between genes. In this paper, we adapt the Matlab clustergram() function, which uses Euclidean distance as the default method to calculate pairwise distance to visualize the heatmap after clustering.

In order to remove irrelevant linkages in a graph, it is necessary to choose a suitable MI threshold which determines the topology of networks formed. The value of 0 or 1 is assigned to a matrix element based on the chosen MI threshold. References in [24] and [25] describe a method to determine a suitable threshold using permutations of MI. The procedure involves permuting MI values of gene pairs and then choosing the largest MI to be the threshold. Using this procedure for 30 repetitions of the permutation on the MI matrix, we choose 0.06 as the threshold. The distributions of the original and permuted MI values are shown in Figure 1.

Results
Microarray datasets descriptions
The gene expressions used for this paper are based on PBMC (Peripheral Blood Mononuclear Cells) blood-based biomarkers [12-14]. Subject AD and normal elderly patients all took the MMSE (Mini-Mental State Examination). Those with chronic metabolic conditions such as diabetes, rheumatoid arthritis and other chronic illnesses or familial AD problems are not included in the analysis [12-14]. Fields such as immunology, transplant immunology, vaccine development often use PBMCs.

GSE4226
AMFES is used to analyze the gene expressions from the BMC (Blood Mononuclear Cell) of AD patients [12]. The dataset contains 9600 features from 14 normal elderly control samples (7 females and 7 male) and 14 AD patient
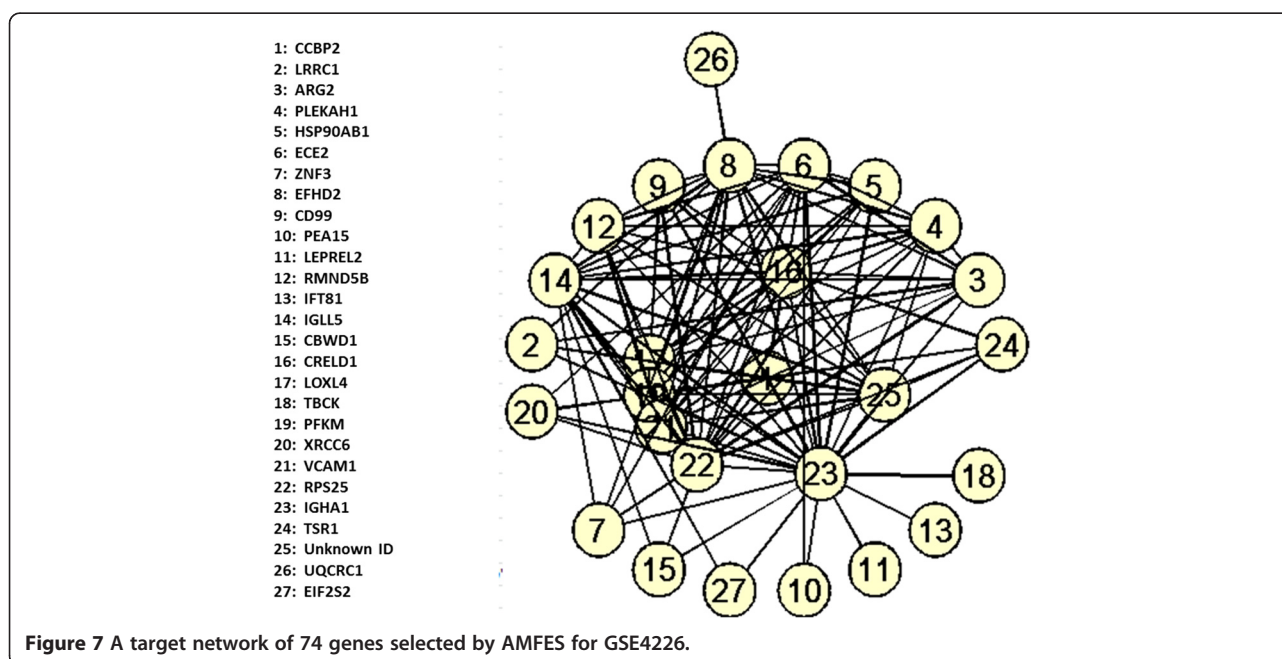

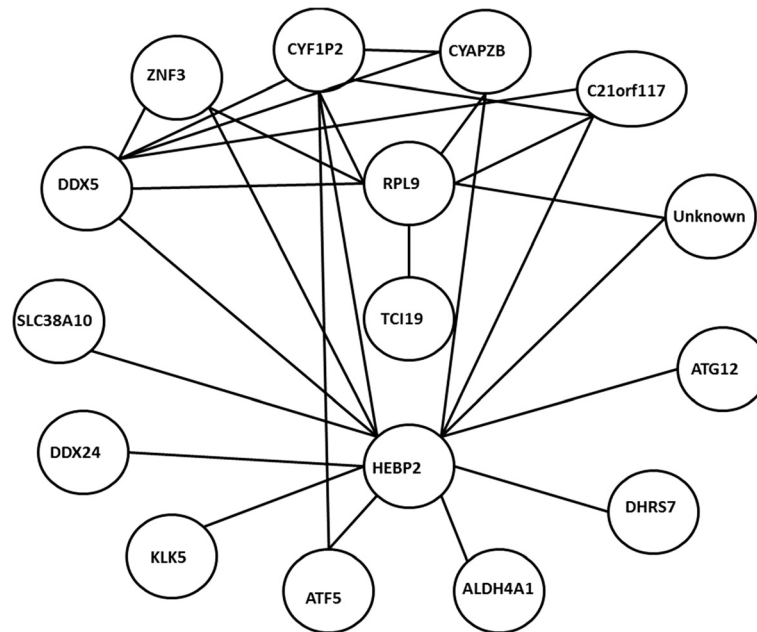**Figure 7 A target network of 74 genes selected by AMFES for GSE4226.**

1: CCBP2
2: LRRC1
3: ARG2
4: PLEKAH1
5: HSP90AB1
6: ECE2
7: ZNF3
8: EFHD2
9: CD99
10: PEA15
11: LEPREL2
12: RMND5B
13: IFT81
14: IGLL5
15: CBWD1
16: CRELD1
17: LOXL4
18: TBCK
19: PFKM
20: XRCC6
21: VCAM1
22: RPS25
23: IGHA1
24: TSR1
25: Unknown ID
26: UQCRC1
27: EIF2S2

Figure 8 A target network of 52 genes selected by AMFES for GSE4227.

samples (7 females and 7 males). The average age of the patients is 79 ± 5 years with 11 ± 4 years of formal educational background. The platform of the dataset is GPL1211 and gene expressions are extracted by using the technology of NIA (National Institution on Aging) Human MGC (Mammalian Genome Collection) cDNA microarray. The raw normalized dataset can be found in Additional file 1.

GSE4227

The dataset GSE4227 was extracted from BMC and under the same GPL1211 platform as GSE4226. It was used to identify the genes with expressions associated with GSTM3 (Glutathione S-Transferase Mu 3) [14]. The dataset contains 9600 features and 34 samples (16 sporadic AD samples and 18 normal elderly control samples). The raw normalized dataset can be found in Additional file 2.
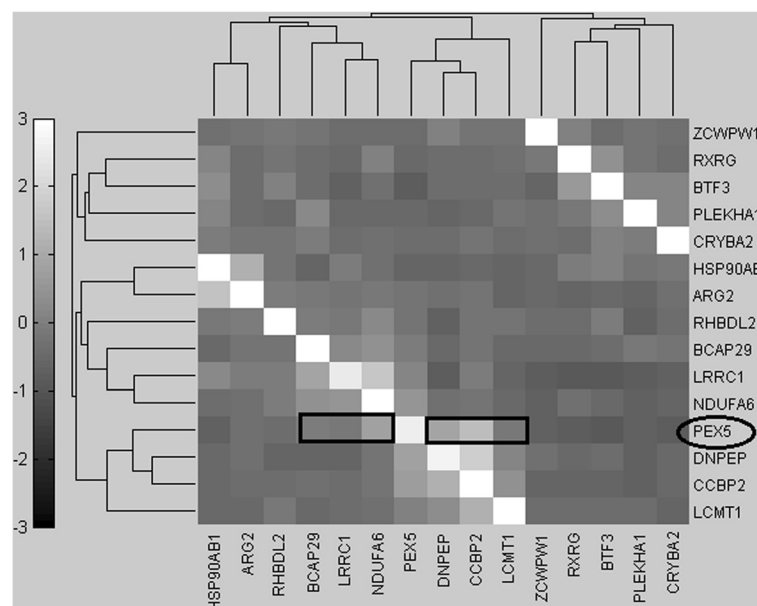

Figure 9 Clustergram of first 15 genes selected by AMFES for GSE4226.

GSE4229
This dataset contains new subjects and some subjects from GSE4226 and GSE4227 [13]. The blood samples were extracted by phlebotomy in an EDTA vacutainer. The dataset also contains 9600 features and 40 samples (18 AD samples and 22 normal elderly control samples). The platform is the same as GSE4226 and GES4227. The raw normalized dataset can be found in Additional file 3.

Results of AMFES
Table 1 contains the descriptions of the three datasets, GSE 4226, 4227 and 4229. AMFES selected 74 genes for GSE4226, 52 for GSE4227, and 395 for GSE4229 and the selected results are shown in Table 2. The complete lists of the 74, 52 and 395 selected genes can be found in Additional file 4: GSE4226_74_Biomakrers.xlsx, Additional file 5: GSE4227_52_Biomakrers.xlsx and Additional file 6: GSE4229_395_Biomakrers.xlsx. The statistical results of MI values are shown in Table 3.

ROC/AUC analysis
To show the classification ability of selected genes, we calculate the AUC (Area Under the Curve) and draw the ROC (Receiver Operating Characteristic) curves for the expressions of 74 genes by using the LIBSVM Matlab ROC tool, as shown in Figure 2 [26]. The ROC/AUC value is verified based on cross-validation [26]. The AUC metric represents the probability that the classifier constructed by the selected genes has higher performance than the classifier constructed using randomly chosen genes. Thus, the closeness of the AUC value to the value of 1 indicates the importance of the selected

genes. The ROC/AUC value (0.95918) therefore supports the significance of verification for the selected genes.

Mutual information analysis
The pair-wise MI values of selected genes of AD or normal samples are calculated separately. The histograms of MI values of GSE4226 are shown in Figure 3, where the black bars represent MI values of normal samples, and the grey bars are for AD samples. The histograms for GSE4227 and GSE4229 are displayed in Figures 4 and 5, respectively. The pair-wise MI files of AD and normal samples are shown in Additional file 7: GSE4226 AD MI.xlsx, Additional file 8: GSE4226-Normal MI.xlsx, Additional file 9: GSE4227-AD MI.xlsx, Additional file 10: GSE4227-Normal MI.xlsx, Additional file 11: GSE4229-AD MI.xlsx and Additional file 12: GSE4229-Normal MI.xlsx. The analysis results are shown in Table 3. Interestingly, the mean MI value of the GSE4226 AD samples is larger than those of the other datasets. In addition, the ratio of the number of positive MI values to the negative MI values for GSE4226 AD samples is also larger than the ratio for other datasets.

Clustergram/Target network examples
The clustergram function of the genes selected from the GSE4226 dataset is described as an example to support the target networks constructed. Only the top ranked 15 genes are used for the analysis shown in Figure 6. In fact, any number of genes can be selected to follow the same procedure. The complete target network constructed by 74 biomarkers for GSE4226 and 52 biomarkers for GSE4227 are shown in Figures 7 and 8, respectively.
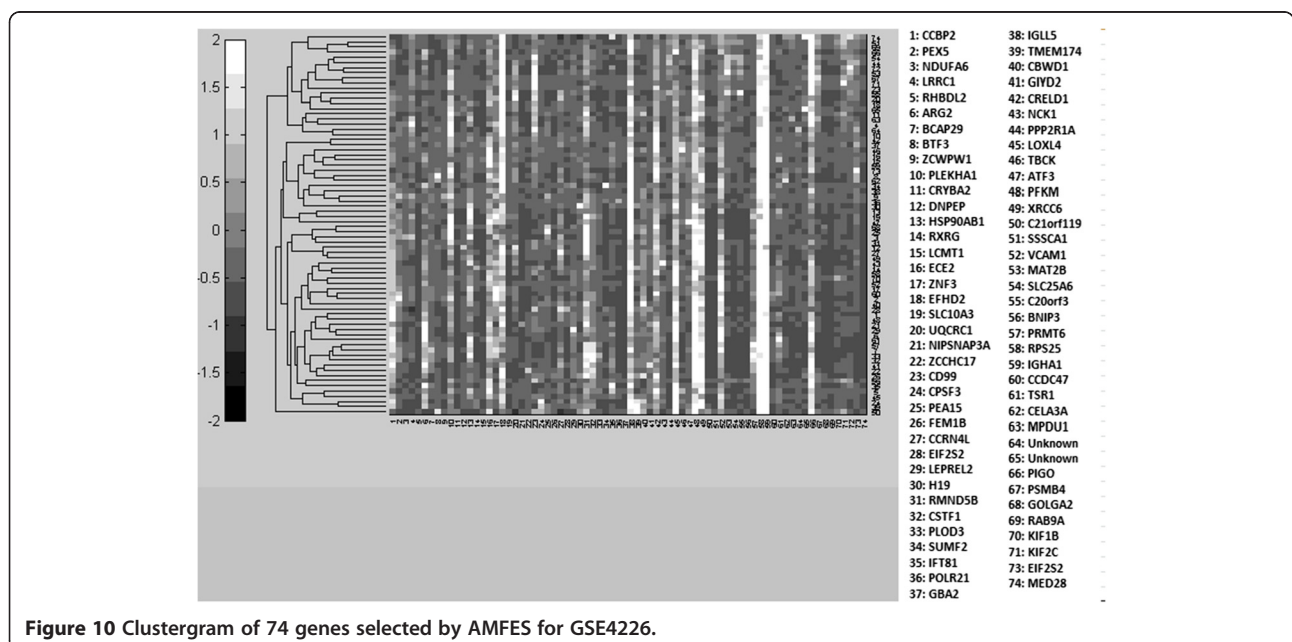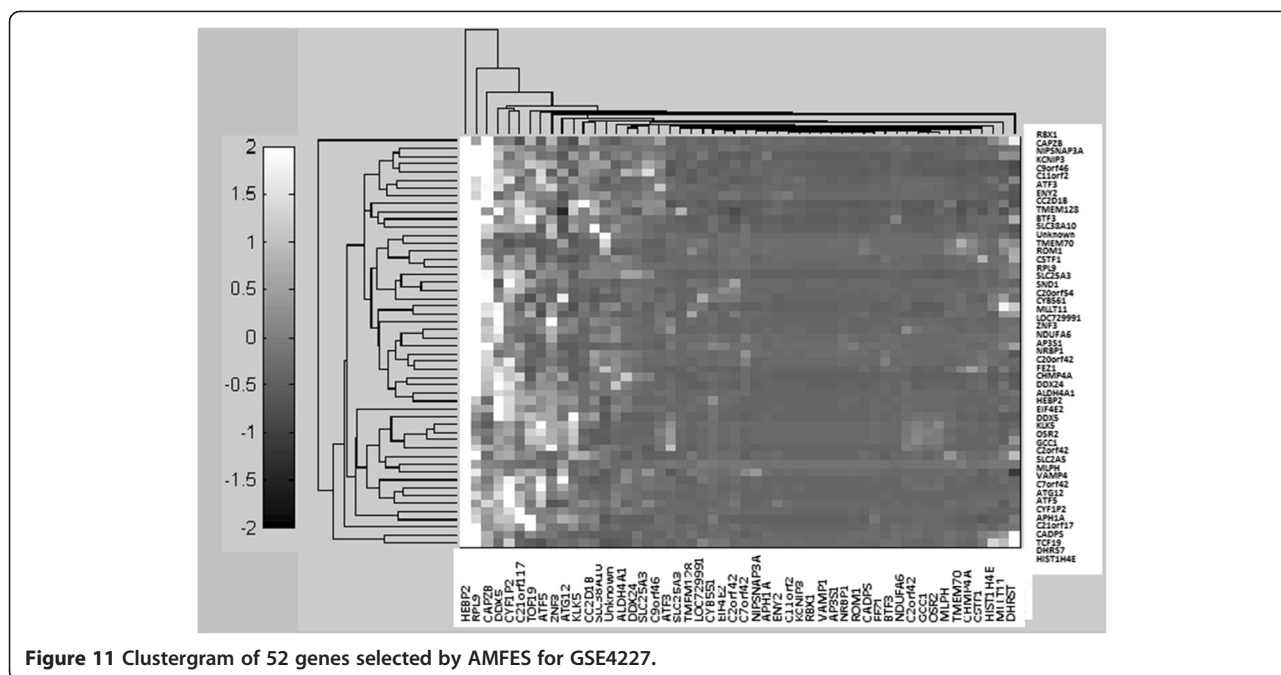

Figure 10 Clustergram of 74 genes selected by AMFES for GSE4226.

Figure 11 Clustergram of 52 genes selected by AMFES for GSE4227.

Because of the limited visibility of the figure constructed by a large number of biomarkers for GSE4229, we did not include it in the manuscript. For the target networks constructed, we observed that only a few of the interactions between biomarkers are reliable after trimming by a permutation test. This observation can help focus on a smaller set of the more important interactions.

As the heatmap generated for the 15 genes (Figure 9) shows, if a few genes share high pairwise MI values with a specific gene, they tend to cluster together as indicated by rectangles and have fewer "hops" (number of connections between a pair of gene) than other genes (Figure 6). For example, PEX5 shares similar MI values with DNPEP, CCBP2, LCMT1, BCAP29, LRRC1 and NDUFA6, which are clustered together in Figure 9. From the graphical view of the target network, these genes have direct connections to PEX5, as shown in Figure 6. On the other hand, a gene such as PLEKHA1 which is two hops away from PEX5 displays an obvious color difference in the MI clustergram. The clustered heatmaps of 74 genes for GSE4226 and of

52 genes for GSE4227 are shown in Figures 10 and 11, respectively.

Gender analysis

AMFES is used to analyze the data by gender for all three datasets and the results are shown in Table 4. The complete

Table 4 The comparisons of female genes and male gene selected by AMFES

Datasets	Number of features selected for female	Number of features selected for male
GSE4226	19	9
GSE4227	12	13
GSE4229	36	13

Table 5 The selected female genes of GSE4226

Number	Gene symbols
1	Unknown ID
2	SKIP
3	Unknown ID
4	MGC15504
5	DKFZP434J046
6	FLJ20591
7	ELA3B
8	CADPS
9	FLJ10707
10	FLJ14639
11	ZDHHC1
12	Unknown ID
13	LOXL4
14	PTPN4
15	STK25
16	CNP
17	TERF21P
18	TAF12
19	BAP29

Table 6 The selected male genes of GSE4226

Number	Gene symbols
1	RNASE1
2	FLJ22729
3	FLJ12571
4	IDH2
5	HCS
6	E1B-AP5
7	PIGO
8	ZDHHC3
9	H4FJ

lists of selected female and male genes for GSE4226 are shown in Tables 5 and 6 as examples, respectively. The complete lists of selected female and male genes for GSE4227 and GSE4229 are shown in Additional file 13, Additional file 14, Additional file 15, and Additional file 16. Interestingly, from all the results there are no overlapped biomarkers between female and male genes.

Discussion

In this paper, the GSE 4226 dataset is studied in more detail because the number of female and male subjects is equal, thereby avoiding the biased sampling problem of the datasets (i.e., when the number of samples is unbalanced for two classes). Traditionally, statistical software such as SAM (Significance Analysis of Microarrays)

[27], PAM (Prediction Analysis for Microarrays) [28] or ANOVA (Analysis of Variance) are used for analyses of biomarkers [12-14]. Compared to the results in [12-14], AMFES selects a much smaller, yet important, set of biomarkers which are supported by the cross-validation. In [7,8,29], the researches were performed based on SVM-RFE for AD biomarker analyses. Here, AMFES can appreciably improve the performance of biomarker analysis. In our current research, we are extending the framework of Zhang et al. [6] by AMFES, and this work will be reported shortly. Finally, interestingly for the gender analysis, when we compare results for female AD subjects with those for male AD subjects, there are no overlapping genes, indicating that the important biomarkers may differ according to gender.

Conclusions

Based on above results, we have proposed a methodology for improving the diagnosis of AD, which is summarized in the Figure 12. As shown in Figure 12, after labeling the AD vs. normal samples, AMFES can select a small subset of important biomarkers by evaluating them on an SVM. For a new patient, the proposed method can select the biomarkers accordingly and construct the corresponding target networks to provide a definitive diagnosis. As in [11], the target networks can be used for further development of a synergistic strategy to improve the therapy of AD in the future.
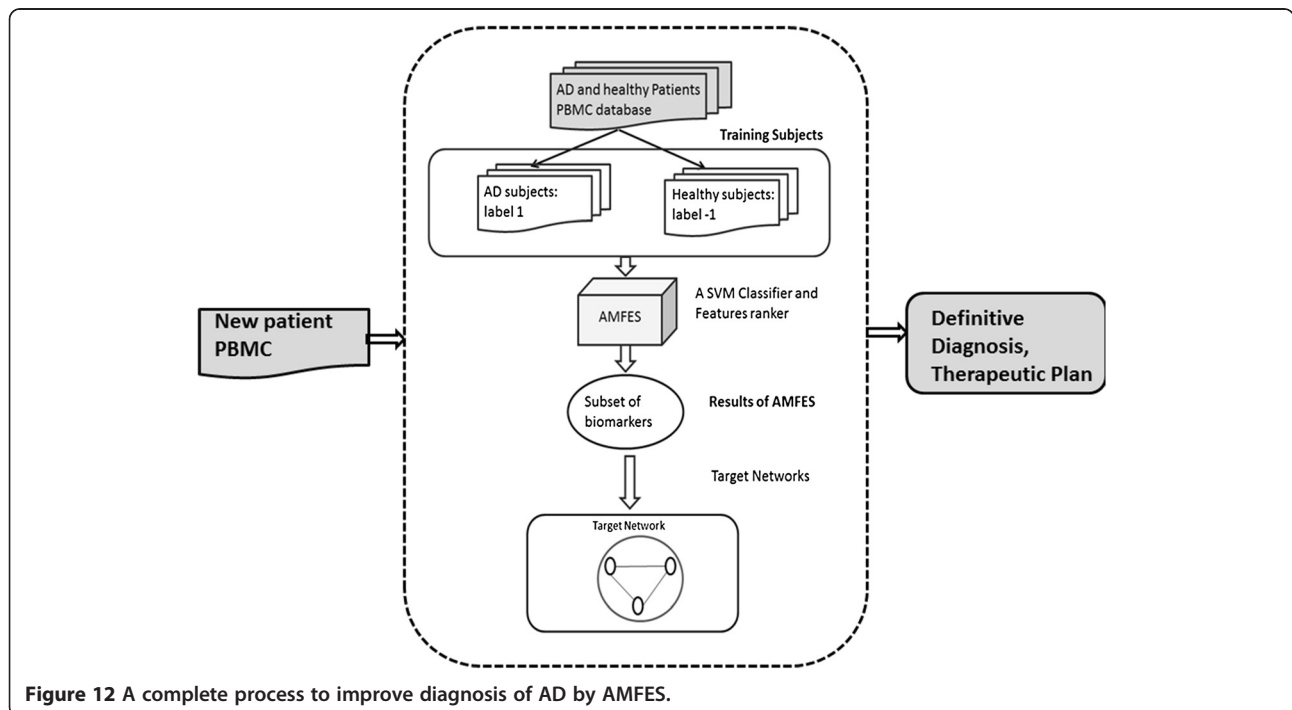

Figure 12 A complete process to improve diagnosis of AD by AMFES.

Additional files

Additional file 1: GSE4226_RawData_Normalized.xlsx. An MS Office Excel file which contains normalized raw data of GSE4226 samples.

Additional file 2: GSE4227_RawData_Normalized.xlsx. An MS Office Excel file which contains normalized raw data of GSE4227 samples.

Additional file 3: GSE4229_RawData_Normalized.xlsx. An MS Office Excel file which contains normalized raw data of GSE4229 samples.

Additional file 4: GSE4226_74_Biomarkers.xlsx. An MS Office Excel file which contains a list of gene symbols of 74 biomarkers of GSE4226 samples.

Additional file 5: GSE4227_52_Biomarkers.xlsx. An MS Office Excel file which contains a list of gene symbols of 52 biomarkers of GSE4227 samples.

Additional file 6: GSE4229_395_Biomarkers.xlsx. An MS Office Excel file which contains a list of gene symbols of 395 biomarkers of GSE4229 samples.

Additional file 7: GSE4226 AD MI.xlsx. An MS Office Excel file which contains a matrix of the pairwise MI values of 74 biomarkers of AD samples.

Additional file 8: GSE4226 Normal MI.xlsx. An MS Office Excel file which contains a matrix of the pairwise MI values of 74 biomarkers of normal samples.

Additional file 9: GSE4227 AD MI.xlsx. An MS Office Excel file which contains a matrix of the pairwise MI values of 52 biomarkers of AD samples.

Additional file 10: GSE4227 Normal MI.xlsx. An MS Office Excel file which contains a matrix of the pairwise MI values of 52 biomarkers of normal samples.

Additional file 11: GSE4229 AD MI.xlsx. An MS Office Excel file which contains a matrix of the pairwise MI values of 395 biomarkers of AD samples.

Additional file 12: GSE4229 Normal MI.xlsx. An MS Office Excel file which contains a matrix of the pairwise MI values of 395 biomarkers of normal samples.

Additional file 13: GSE4227 Female Gene List.xlsx. An MS Office Excel file which contains a complete list of gene symbols of female genes selected for GSE4227.

Additional file 14: GSE4227 Male Gene List.xlsx. An MS Office Excel file which contains a complete list of gene symbols of male genes selected for GSE4227.

Additional file 15: GSE4229 Female Gene List.xlsx. An MS Office Excel file which contains a complete list of gene symbols of female genes selected for GSE4229.

Additional file 16: GSE4229 Male Gene List.xlsx. An MS Office Excel file which contains a complete list of gene symbols of female genes selected for GSE4229.

Authors' contributions
Please see sample text in the instructions for authors. WH, CD: Implementation of project. SC: Design the project. All authors read and approved the final manuscript.

Author details
[1]Systems Biology Lab, University of Florida, Gainesville, FL 32611, USA. [2]Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA. [3]Van Nuys Senior High School, Van Nuys, CA 91411, USA. [4]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611, USA.

References
1. Fraller DB: **State of the science: use of biomarkers and imaging in diagnosis and management of Alzheimer disease.** *J Neurosci Nurs* 2013, **45**:63–70.
2. Jack CR Jr, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, Petersen RC, Trojanowski JQ: **Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade.** *Lancet Neurol* 2010, **9**:119–128.
3. Frisoni GB, Blennow K: **Biomarkers for Alzheimer's: the sequel of an original model.** *Lancet Neurol* 2013, **12**:126–128.
4. Laws Sm DJDMRN: **Blood-based biomarkers in alzheimer disease: where are we now and where have we to go?—reply.** *JAMA Neurol* 2013, **70**:133–134.
5. Guo LH, Alexopoulos P, Wagenpfeil S, Kurz A, Perneczky R: **Plasma proteomics for the identification of Alzheimer Disease.** *Alzheimer Dis Assoc Disord* 2013, **00**:000–000.
6. Zhang D, Wang Y, Zhou L, Yuan H, Shen D: **Multimodal classification of Alzheimer's disease and mild cognitive impairment.** *Neuroimage* 2011, **55**:856–867.
7. Fan Y, Batmanghelich N, Clark CM, Davatzikos C: **Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline.** *Neuroimage* 2008, **39**:1731–1743.
8. Fan Y, Shen D, Gur RC, Gur RE, Davatzikos C: **COMPARE: classification of morphological patterns using adaptive regional elements.** *IEEE T Med Imaging* 2007, **26**:93–105.
9. Davatzikos C, Fan Y, Wu X, Shen D, Resnick SM: **O3-06-01: Complex patterns of structural brain abnormality identified in MCI via high-dimensional pattern classification of MR images may be early markers of Alzheimer's disease.** *Alzheimer's Dement* 2006, **2**:S64.
10. Guyon I, Weston J, Barnhill S, Vapnik V: **Gene selection for cancer classification using support vector machines.** *Mach Learn* 2002, **46**:389–422.
11. Hsu WC, Liu CC, Chang F, Chen SS: **Cancer classification: mutual information, target network and strategies of therapy.** *J Clin Bioinforma* 2012, **2**:16.
12. Maes OC, Xu S, Yu B, Chertkow HM, Wang E, Schipper HM: **Transcriptional profiling of Alzheimer blood mononuclear cells by microarray.** *Neurobiol Aging* 2007, **28**:1795–1809.
13. Maes OC, Schipper HM, Chertkow HM, Wang E: **Methodology for discovery of Alzheimer's disease blood-based biomarkers.** *J Gerontol A Biol Sci Med Sci* 2009, **64**:636–645.
14. Maes OC, Schipper HM, Chong G, Chertkow HM, Wang E: **A GSTM3 polymorphism associated with an etiopathogenetic mechanism in Alzheimer disease.** *Neurobiol Aging* 2010, **31**:34–45.
15. Bi J, Bennett K, Embrechts M, Breneman C, Song M: **Dimensionality reduction via sparse support vector machines.** *J Mach Learn Res* 2003, **3**:1229–1243.
16. Stoppiglia H, Dreyfus G, Dubois R, Oussar Y: **Ranking a random feature for variable and feature selection.** *J Mach Learn Res* 2008, **3**:1399–1414.
17. Tuv E, Borisov A, Torkkola K: **Feature Selection Using Ensemble Based Ranking Against Artificial Contrasts.** In *Neural Networks, 2006 IJCNN '06 International Joint Conference on; 0–0 0.* Vancover, BC,Canada: IEEE; 2006:2181–2186.
18. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5**:101–113.
19. Shannon CE: **A mathematical theory of communication.** *SIGMOBILE Mob Comput Commun Rev* 2001, **5**:3–55.
20. Beirlant J, Dudewicz EJ, Ouml LG, rfi, Meulen ECVD: **Nonparametric entropy estimation: an overview.** *Int J Math Stat Sci* 1997, **6**:17–39.
21. Qiu P, Gentles AJ, Plevritis SK: **Fast calculation of pairwise mutual information for gene regulatory network reconstruction.** *Comp Methods Programs Biomed* 2009, **94**:177–180.
22. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, Califano A: **ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context.** *BMC Bioinformatics* 2006, **7**:S7.
23. Wolf ME, Lam MS: **A data locality optimizing algorithm.** *SIGPLAN Not* 1991, **26**:30–44.
24. Butte AJ, Kohane IS: **Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.** *Pac Symp Biocomput* 2000:418–429.

25. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A:
    Reverse engineering of regulatory networks in human B cells. *Nat Genet*
    2005, **37**:382–390.
26. Chang C-C, Lin C-J: **LIBSVM: a library for support vector machines.**
    ACM Trans Intell Syst Technol 2011, **2**:1–27.
27. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays
    applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001,
    98:5116–5121.
28. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer
    types by shrunken centroids of gene expression.** *Proc Natl Acad Sci USA*
    2002, **99**:6567–6572.
29. Fan Y, Shen D, Davatzikos C: **Classification of structural images via high-
    dimensional image warping, robust feature extraction, and SVM. MICCAI
    2005.** In *Lecture Notes in Computer Science*, Volume 3749. Edited by Duncan J,
    Gerig G. Palm Springs, California, USA: Springer Berlin Heidelberg; 2005:1–8.