Bioinformatics, 31(15), 2015, 2505–2513
doi: 10.1093/bioinformatics/btv173
Advance Access Publication Date: 26 March 2015
Original Paper

OXFORD

Gene expression

Identification of a small set of plasma signalling proteins using neural network for prediction of Alzheimer's disease

Swapna Agarwal[1,*], Pradip Ghanty[2] and Nikhil R. Pal[1]

[1]Electronics and Communication Sciences Unit, Indian Statistical Institute, Calcutta 700108 and [2]Praxis Softek Solutions Pvt. Ltd., Saltlake Electronic Complex, Calcutta 700091, India

*To whom correspondence should be addressed.
Associate Editor: Ziv Bar-Joseph

Received on May 26, 2014; revised on February 25, 2015; accepted on March 22, 2015

Abstract

Motivation: Alzheimer's disease (AD) is a dementia that gets worse with time resulting in loss of memory and cognitive functions. The life expectancy of AD patients following diagnosis is ~7 years. In 2006, researchers estimated that 0.40% of the world population (range 0.17–0.89%) was afflicted by AD, and that the prevalence rate would be tripled by 2050. Usually, examination of brain tissues is required for definite diagnosis of AD. So, it is crucial to diagnose AD at an early stage via some alternative methods. As the brain controls many functions via releasing signalling proteins through blood, we analyse blood plasma proteins for diagnosis of AD.

Results: Here, we use a radial basis function (RBF) network for feature selection called feature selection RBF network for selection of plasma proteins that can help diagnosis of AD. We have identified a set of plasma proteins, smaller in size than previous study, with comparable prediction accuracy. We have also analysed mild cognitive impairment (MCI) samples with our selected proteins. We have used neural networks and support vector machines as classifiers. The principle component analysis, Sammmon projection and heat-map of the selected proteins have been used to demonstrate the proteins' discriminating power for diagnosis of AD. We have also found a set of plasma signalling proteins that can distinguish incipient AD from MCI at an early stage. Literature survey strongly supports the AD diagnosis capability of the selected plasma proteins.

Availability and implementation: The FSRBF code is available at https://sites.google.com/site/agarwalswapna/publications.

Contact: agarwal.swapna@gmail.com or swapna_r@isical.ac.in

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

1 Introduction

Alzheimer's disease (AD) is the most common form of dementia. According to a cohort longitudinal study, ~10–15 persons per 1000 persons per year get dementia of which 5–10 get AD (Bermejo *et al.*, 2008; Carlo *et al.*, 2002). In AD, an unknown process divides Amyloid Precursor Protein into smaller fragments. One of these fragments gives rise to fibrils of beta-amyloid, which gets deposited outside neurons in dense formation known as senile plaques (Hooper *et al.*, 2005). Tau protein becomes hyper-phosphorylated and creates neurofibrillary tangles (Hernandez and Avila, 2007; Tiraboschi *et al.*, 2004).

In recent years researchers have devoted great efforts for the development of AD diagnosis tools (German *et al.*, 2007; Hye *et al.*, 2006; Xiao *et al.*, 2005). Most of the articles have investigated on a small set (two to four) of Cerebro Spinal Fluid (CSF) proteins (Craig-Schapiro *et al.*, 2011; Rentzos *et al.* 2006; Westin *et al.*, 2012). Few attempts have been made using serum proteins (Ray *et al.*, 2007) and using both CSF and serum proteins (Tham *et al.*, 1993). Some have tried to

find biomarker proteins that can discriminate AD and non-AD patients (Hu *et al.*, 2010) and some have tried to predict AD from mild cognitive impairment (MCI) patients (Craig-Schapiro *et al.*, 2011). In the study by Teramoto (2008), a semi-supervised distance metric learning using random forests with label propagation is proposed for prediction of AD. In Hansson *et al.* (2006), concentration of $T-tau$, $P-tau_{181}$ and $A\beta 42$ in CSF are reported to be associated with future development of AD in patients with MCI. Ray *et al.* (2007) have proposed a microarray-based method that has selected 18 blood plasma signalling proteins to classify and predict clinical AD diagnosis. We have used the same dataset and found nine plasma signalling proteins for the same problem with comparable prediction accuracy. We have also found a set of useful plasma signalling proteins that can predict AD from MCI with a better prediction accuracy.

The selection of biomarker plasma proteins from a large set of proteins is a feature selection problem. Feature selection methods can be broadly classified as filter methods and wrapper methods. In filter methods, the features are given importance solely depending on the properties of the features themselves. These methods ignore the tool finally used to recognize the patterns. But the utility of a feature depends on the pattern recognition tool being used and the problem being solved. A set of features good for a particular pattern recognition problem and a tool may not be as good for a different pattern recognition tool. The wrapper-based feature selection methods utilize the classifier itself to find the relevance of the feature. Thus, wrapper methods are generally considered better as compared with filter methods (Kohavi and John, 1997). One of the early filter methods is Relief (Kira and Rendell, 1992). In Relief, given a feature vector, two more instances of feature vectors are considered; one from the same class and the other from the other class. The weight of a feature is decreased if the value of that feature differs more from the value in the instance of the same class than the value in the instance of the different class and vice versa. Relief was modified into ReliefF by Kononenko *et al.* (1997). KernelPLS (Sun *et al.*, 2014) is a kernel-based multivariate feature selection method that selects features taking into account possible non-linear relation between features as well as that between features and target. In the study by Sun *et al.* (2014), a partial least square (PLS) algorithm finds a low-dimensional approximation of the input matrix that can explain as close as possible the target vectors. The support vector machine recursive feature elimination (SVMRFE) method (Guyon *et al.*, 2002) eliminates poor features using an iterative process. It starts with all features and removes one feature at a time based on a feature ranking score that is computed using the coefficients of the weight vector of a linear SVM. At each iteration, the feature with the smallest ranking score is eliminated.

In this study, neural networks (NNs) are used for feature selection as well as for classification. We have used both multilayer perceptron (MLP) and radial basis function (RBF) NNs for classification. RBF NNs are used for feature selection in the studies by Basak and Mitra (1999) and Chakraborty and Pal (2008). The Group Feature Selection RBF (GFSRBF) is proposed by Chakraborty and Pal (2008) for selecting useful groups of features where each feature group corresponds to a sensor. In Pal and Malpani (2012), this network has been adapted for feature selection with controlled redundancy. Here, we have adapted GFSRBF for feature selection without any explicit control on redundancy for selection of plasma proteins that can predict clinical AD. Our approach is described in the following section.

2 Approach

For diagnosis of AD from plasma samples, we need to identify the plasma proteins that carry AD specific signature. As the initial set of plasma proteins, we consider the set of 120 proteins reported in Ray *et al.* (2007). To select an adequate set of proteins we use a modified RBF NN for feature selection. This is an integrated method where feature selection and system identification are done simultaneously. In this method, a feature attenuator is associated with each feature. For a useful feature, its attenuator allows the feature to get into the network. For an unimportant feature, its attenuator does not allow the feature to affect the network. To verify the AD-specific signature of these selected features (proteins), we subject them to different classifiers. We have also tested if the selected proteins form natural AD and non-AD-specific clusters. We have compared the results with that of Ray *et al.* (2007) as well as those by two filter methods (ReliefF and kernelPLS) and a wrapper method (SVMRFE).

3 Methods

3.1 Feature selection RBF

Let p be the number of features, c be the number of classes and X be the dataset, $X=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}\subset\Re^p$ with associated output in $\Re^c$. In general, an RBF network has three layers. Layer 1, which is called input layer, has p nodes. Layer 3, which is called output layer, has c nodes. The architecture of RBF depends on number (let us denote it by h) of nodes present in layer 2, which is called hidden layer or basis function layer. The required value of h depends on the dataset. In an RBF network, each node in the input layer is connected to each node in the hidden layer, and each node in the hidden layer is connected through some weights to each node in the output layer. There is no connection between the nodes in the same layer. Each node in the hidden layer uses a Gaussian basis function, $\phi_j, j=1,...,h$:

$$\phi_j(\mathbf{x})=\exp\{-\|\mathbf{x}-\mathbf{u}_j\|^2/\sigma_j^2\}, \tag{1}$$

where $\mathbf{u}_j$ is the centre and σ_j is the spread of the jth basis function. Note that $\mathbf{u}$ and $\mathbf{x}$ both are in $\Re^p$. Let O_k be the output of the k^{th} node of the output layer, then

$$O_k=\sum_{j=1}^{h} w_{jk}\phi_j(\mathbf{x}), \tag{2}$$

where w_{jk} is the weight between jth node in hidden layer and kth node in output layer. In (2), O_k is unbounded as w_{jk} can take any value. For classification problems, desired output lies in [0,1]. So, we add a standard sigmoidal function to each output node.

$$O_k'=1/\{1+\exp\Big(-\sum_{j=1}^{h} w_{jk}\phi_j(\mathbf{x})\Big)\}. \tag{3}$$

For classification problem, (3) is used and for regression, (2) is used. We can rewrite (1) as

$$\phi_j(\mathbf{x})=\Pi_{i=1}^{p}\exp\{-(x_i-u_{ij})^2/\sigma_j^2\}=\Pi_{i=1}^{p}C_j^i, \tag{4}$$

where

$$C_j^i=\exp\{-(x_i-u_{ij})^2/\sigma_j^2\}. \tag{5}$$

Value of C_j^i depends only on the ith feature of the input. To eliminate the features which have derogatory effect on the classification/regression problem, as done in Chakraborty and Pal (2008), we design C_j^i as:

$$C_j^i=[\exp\{-(x_i-u_{ij})^2/\sigma_j^2\}]^{1-e^{-\beta_i^2}}. \tag{6}$$

The term, $(1-e^{-\beta_i^2})$, is called the feature attenuator. When β_i approaches 0, C_j^i approaches 1 and therefore C_j^i has almost no effect

on ϕ_j. When β_i is high, the feature attenuator approaches 1 thus there is almost no change to the value of C_j^i. The value of β_i is learnt along with w_{jk} during training through back propagation algorithm. The architecture of the FSRBF is shown in Figure 1.

3.1.1 Learning rules

Let the desired output label associated with a data point be $\mathbf{d} = [d_1, d_2 ..., d_c]^T$. Thus, the instantaneous error for a data point $\mathbf{x}$ is

$$E = \frac{1}{2}\sum_{k=1}^{c} (O_k^{'} - d_k)^2. \quad (7)$$

We use gradient descent technique to learn w_{jk} and β_i.

$$w_{jk}(t+1) = w_{jk} - \eta_w\Big(\delta E/\delta w_{jk}(t)\Big) \quad (8)$$

$$\beta_i(t+1) = \beta_i - \eta_\beta\Big(\delta E/\delta \beta_i(t)\Big), \quad (9)$$

where η_w and η_β are learning rates. For initial assignment of centres to the hidden nodes, we use the k-means clustering of the training dataset. After the k-means clustering, the spread (σ) value for each RBF is chosen as the minimum distance between its cluster centre and all the other cluster centres. The connection weights between the hidden layer and the output layer are initialized with random values in [−0.5, 0.5]. When total error on all training samples goes below a predefined tolerance, we terminate the learning.

3.2 Dataset

We have used the same dataset reported in Ray *et al.* (2007). This dataset consists of the expression values of 120 known blood plasma proteins drawn from individuals with pre-symptomatic to late state AD and some non-demented controls (NDC). The expression values were measured with filter-based arrayed sandwich enzyme-linked immunosorbent assay (ELISA). We have also used data from plasma samples drawn from individuals who had MCI. We have used a training set consisting of 43 AD and 40 NDC samples while three test sets consisting of 81 (42 AD and 39 NDC), 11 (other dementia) and 47 (MCI) samples, respectively. In the rest of the article by 'first test', 'OD' and 'MCI' dataset we mean these three test datasets, respectively. Out of 47 subjects diagnosed with MCI at blood draw, 22 converted to AD within 2–5 years (MCI → AD), eight converted to OD (MCI → OD), whereas 17 were still diagnosed as MCI, 4–6 years later (MCI → MCI) (Ray *et al.*, 2007). The datasets are available at http://www.nature.com/nm/journal/v13/n11/suppinfo/nm1653_S1.html.

3.3 Experiment design

We do a 5-fold cross-validation (CV) on the training data for architecture selection for FSRBF. For each fold we run FSRBF 10 times and in each run we initialize the w_{jk} and initial centres of k-means clustering randomly. We take the result averaged over these 10 runs as the final result of that fold, thus decreasing the effect of these random initializations. The number of hidden nodes is varied between 2 and 30. FSRBF is run 10 times on the best architecture selected from the 5-fold CV to get 10 sets of β values. The β values are then averaged over different runs. FSRBF assigns an importance (weight) to each feature during the training in terms of β values. The features are sorted in descending order of the values of β. We need to put a threshold either on the number of features to be selected or on the value of β to select the features. Features for which β value is <0.1135 produce component basis function value > 0.95 even at 2σ distance, therefore, do not have much effect on ϕ_j (Chakraborty and Pal, 2008). Here also we choose m features (proteins) which have average β value > 0.1135. To further condense the feature set, we take the feature with the highest average β value and do a 5-fold CV for RBF with different architectures. We increase the number of features by one, i.e. take the two features with highest average β value and again do a 5-fold CV for RBF with different architectures. This process is repeated until 5-fold CV is done on all m features. We select the number of features f' for which validation result is the best. This approach is briefly depicted in Figure 2. Note that this may not be the only or the best strategy. This is one of the strategies to select a small set of proteins for early detection of AD. Before applying FSRBF on AD data, we test FSRBF on some synthetic datasets with known characteristics.

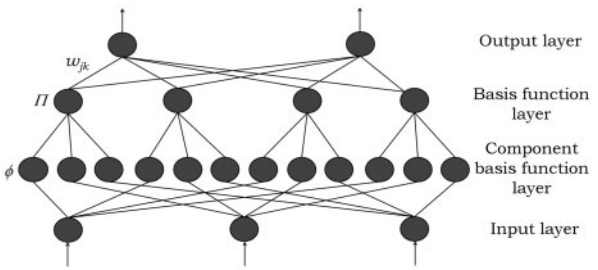

Fig. 1. Architecture of FSRBF

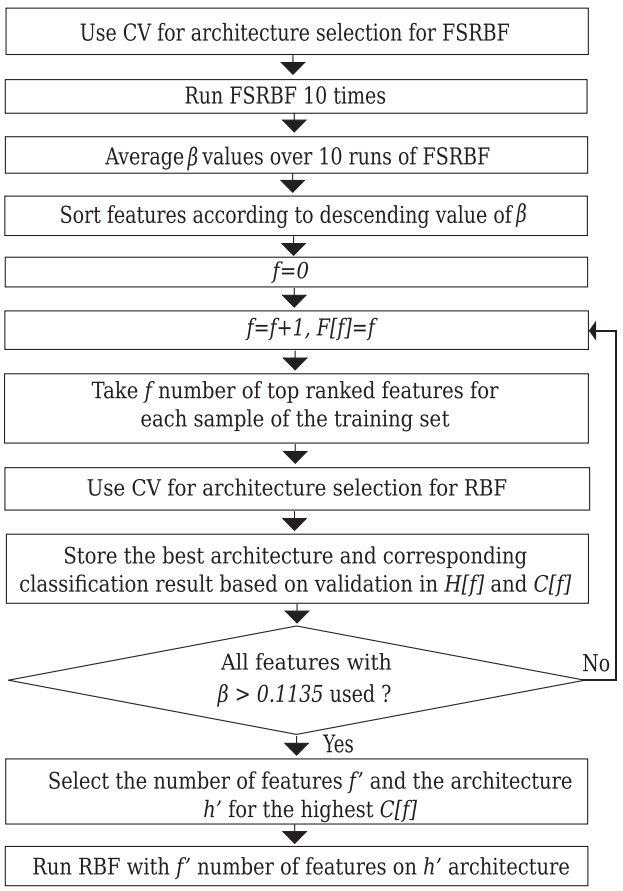

Fig. 2. Overview of the approach: $F[f]$ and $C[f]$ represent feature count and the highest validation accuracy, $H[f]$ represents number of hidden nodes for $C[f]$

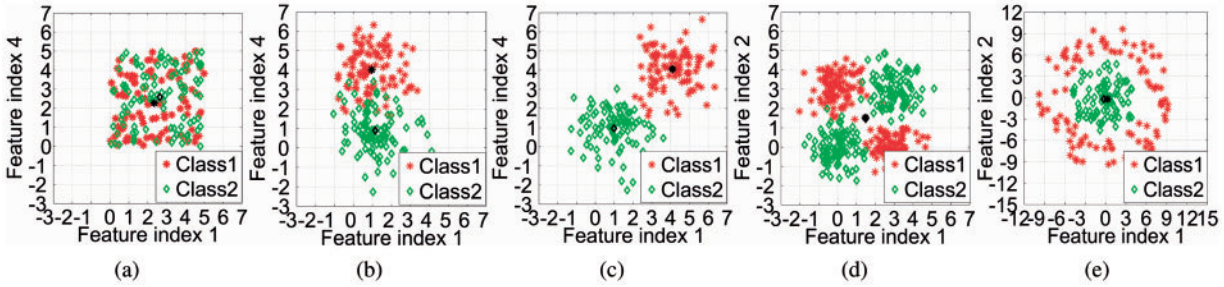

Fig. 3. Scatter plots of the synthetic datasets. Datasets A, B and C are four variate and the fourth feature is important. Datasets D and E are two variate and both the features are important for classification. The panels (**a**), (**b**) and (**c**) plot the first versus the fourth feature of the datasets A, B and C, respectively. Panels (**d**) and (**e**) plot the first versus the second feature of the datasets D and E, respectively

4 Results

4.1 Performance on synthetic data

We generate five types of synthetic data. We name them 'A', 'B', 'C', 'D' and 'E'. To test whether the feature selection algorithm can recognize the situation when none of the features carry any class information, we create the dataset A. All the four features of dataset A are assigned random values. The class labels are also assigned randomly to the samples. Therefore, in dataset A, none of the features is important. To test the feature selection algorithm on cases where the classes are linearly separable, we create datasets B and C. We also test the algorithm's efficiency on identifying correlation between features. The first and the second features of dataset C are highly correlated, correlation coefficient being 0.97. We also need to test the cases where the classes are not linearly separable. Therefore, datasets D and E are created. Whereas datasets B and C are linearly separable, dataset D cannot be separated by a single straight line. Dataset D resembles XOR data in appearance. In dataset E, samples belonging to one class surround the other class. Therefore, the classes are not linearly separable. The scatter plots of the datasets are displayed in Figure 3.

We also test the behaviour of FSRBF with increasing number of noisy features. We compare our results with two state-of-the-art filter methods: kernelPLS (Sun *et al.*, 2014) and ReliefF (Kononenko *et al.*, 1997) and a wrapper method SVMRFE (Guyon *et al.*, 2002) in terms of type I (false positive) and type II (false negative) errors. For dataset A where none of the four features carry class information, FSRBF selects two features incurring type I error, whereas KernelPLS and SVMRFE select all the four features as useful incurring a high type I error. Only ReliefF assigns low weights to all the features resulting in zero type I and type II error. For linearly separable dataset B all feature selection methods select the right feature. For dataset C that contains correlated features, the three methods FSRBF, KernelPLS and ReliefF select all important (including correlated) features. None of the unimportant features get selected even in the presence of 48 noisy features (zero type I and type II errors). Though SVMRFE is able to recognize correlated features, as the number of noisy features increases, SVMRFE tends to select some unimportant features and discards some important features. For dataset D, FSRBF and for dataset E, FSRBF and ReliefF select only the important features (zero error). KernelPLS and ReliefF for dataset D and KernelPLS for dataset E select only some unimportant features discarding both the important features. Therefore, type I and type II errors both become high. As the number of noisy features increases, SVMRFE, for datasets D and E, usually discards one of the important features and selects one of the unimportant features.

All the four feature selection methods under consideration assign a numerical importance (β for FSRBF) to each feature. For an ideal feature selection method, it is expected that the important features

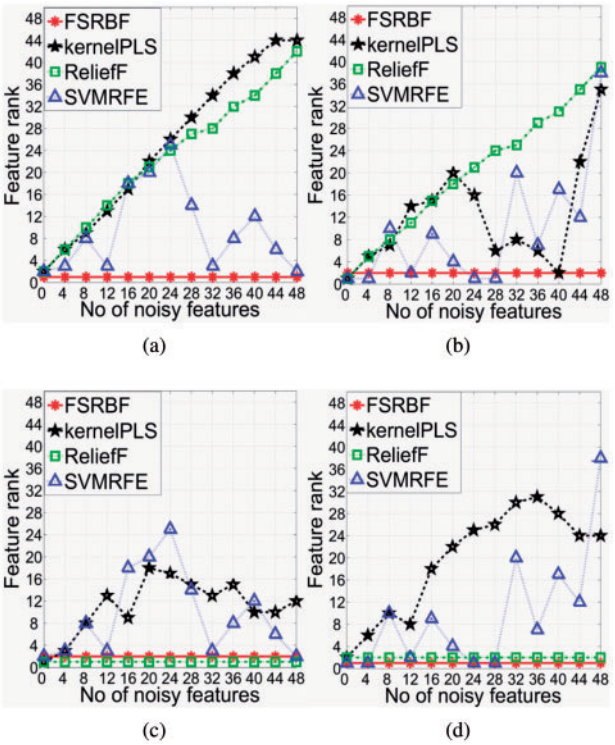

Fig. 4. Change of feature rank assigned by FSRFB, kernelPLS, ReliefF and SVMRFE with increasing number of noisy features. Rank 1 stands for the most important feature and rank 50 for the least important feature. The panels (**a**), (**b**), (**c**) and (**d**) show the change of rank of the first (important) feature of the dataset D, the second (important) feature of the dataset D, the first (important) feature of the dataset E and the second (important) feature of the dataset E, respectively

be assigned the highest weights and therefore the highest ranks even in presence of noisy features. We analyse in Figure 4, the change of rank of the two important features of datasets D and E (with non-linearly separable classes) with increasing number of noisy features for the four competing feature selection methods. We are more interested in datasets D and E as all the four competing methods produce zero error only for linearly separable datasets. The three methods KernelPLS, ReliefF and SVMRFE produce high or moderate type I and type II errors for non-linearly separable datasets D and E whereas FSRBF produces zero type I and type II errors in non-linear cases also. From Figure 4 it can be seen that ReliefF maintains the highest ranks for both the important features, feature 1 and feature 2, for dataset E, but fails to maintain the same for dataset D. For both the datasets D and E, only FSRBF consistently maintains the

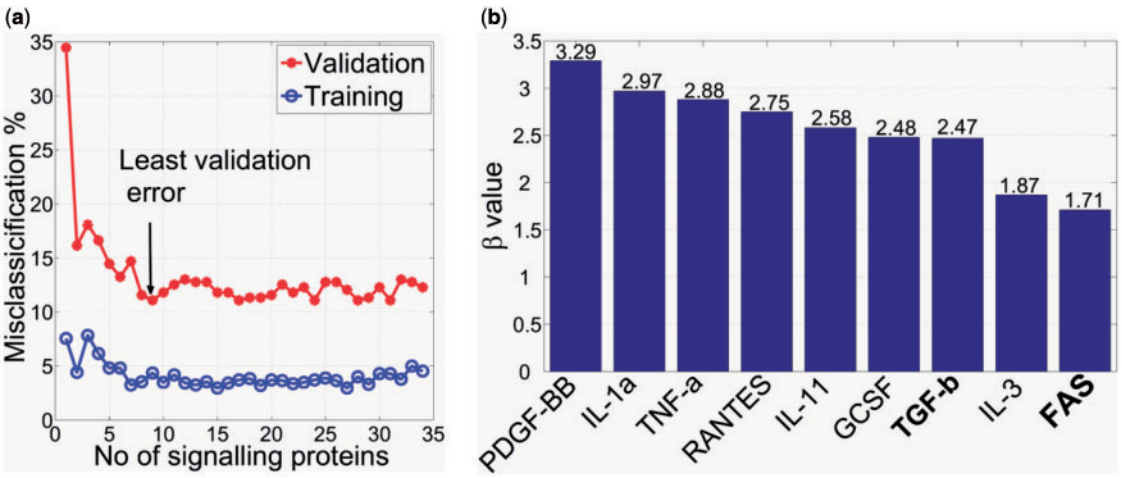

Fig. 5. (**a**) Change of training and validation error with increasing number of signalling proteins. (**b**) The feature importance values (β) of the nine plasma proteins suggested by FSRBF

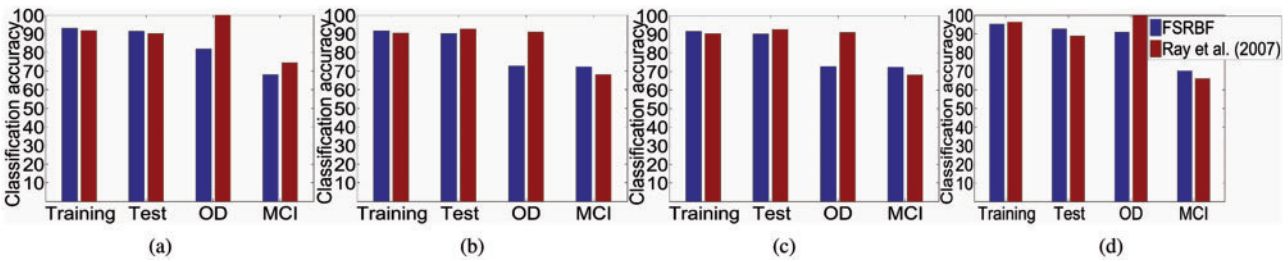

Fig. 6. Performance comparison in terms of classification accuracy % of the nine signalling proteins selected by FSRBF with 18 signalling proteins reported by Ray *et al.* (2007) using (**a**) RBF, (**b**) SVM with linear kernel, (**c**) SVM with RBF kernel and (**d**) MLP classifiers

highest ranks (1 and 2) for the two important features, even in presence of noisy features. Note that these results are average over 10 runs of FSRBF, where the network architecture is selected using 5-Fold CV, but in some individual runs FSRBF may not maintain the highest ranks for features 1 and 2.

Except for dataset A, FSRBF produces zero error for all the cases. Given the application of feature selection in identifying biomarkers of critical biological phenomena (e.g. existence of diseases like Alzheimer, cancer), it is important that none of the features that carry important class information gets discarded. Our limited experiments with simple synthetic datasets reveal the success of FSRBF in this regard, but this does not mean that this will always be the case with real life datasets. Further results on synthetic datasets can be found in Section 1 of the Supplementary Material. Next we apply FSRBF on AD dataset.

4.2 Performance on AD data

Our architecture selection scheme suggests an FSRBF with 13 hidden nodes (validation accuracy 90.25%). Therefore, we run FSRBF NN with 13 hidden nodes on AD training data. We have executed FSRBF on an 80 core computing system involving @2.13 GHz Xeon(R) E7-L8867 processors with 512 GB RAM. Note that though the above-mentioned system is a multi-core multiprocessor system we did not explicitly exploit the parallel processing capability of the system. The CPU time (using only single core of a single processor) for one run of FSRBF with 500 iterations for the parameter updates (Equations 8 and 9) is 20.14 s. It is worth noting that there is a natural parallelism in an RBF network. The computations done at the hidden nodes can be done in parallel which can drastically reduce the computation time. Using the average β values obtained after applying the first four steps in Figure 2, we find 34 signalling proteins with β value > 0.1135. To condense the feature set further, 5-fold CV on n signalling proteins (sorted according to β values) where $n = 1, 2, ..., 34$ is performed. Figure 5a shows the change of training and validation error with increasing number of signalling proteins. It is seen in Figure 5a that validation error is least for nine signalling proteins. We select these nine signalling proteins as our final set of predictors. The selected nine proteins along with their importance in terms of β values averaged over 10 runs of FSRBF are shown in Figure 5b. Seven predictors out of these nine are common with 18 predictors reported by Ray *et al.* (2007). The remaining two (marked in bold font in Fig. 5b) are new findings with FSRBF. Correlation analysis reveals that the maximum correlation between any two of the nine proteins is 0.39 and five of the nine proteins have high correlation (>0.61) with one or more proteins from the remaining set of 111 proteins. That is FSRBF has selected only one of the several highly correlated and important proteins. The details are given in Section 2 of the Supplementary Material.

4.3 Prediction result with the selected proteins

To test the usefulness of the features selected by FSRBF, we subject the AD dataset with selected features to different classifiers, e.g. RBF, SVM with linear kernel, SVM with RBF kernel and MLP. For selection of an appropriate architecture/hyperparameters for these classifiers, we do a 5-fold CV on the training dataset. Each classifier is trained with the 'train' dataset using both sets of features (nine features found by FSRBF and 18 features found by Ray *et al.* 2007) and tested on the four datasets, 'train', 'first test', 'OD' and 'MCI'.

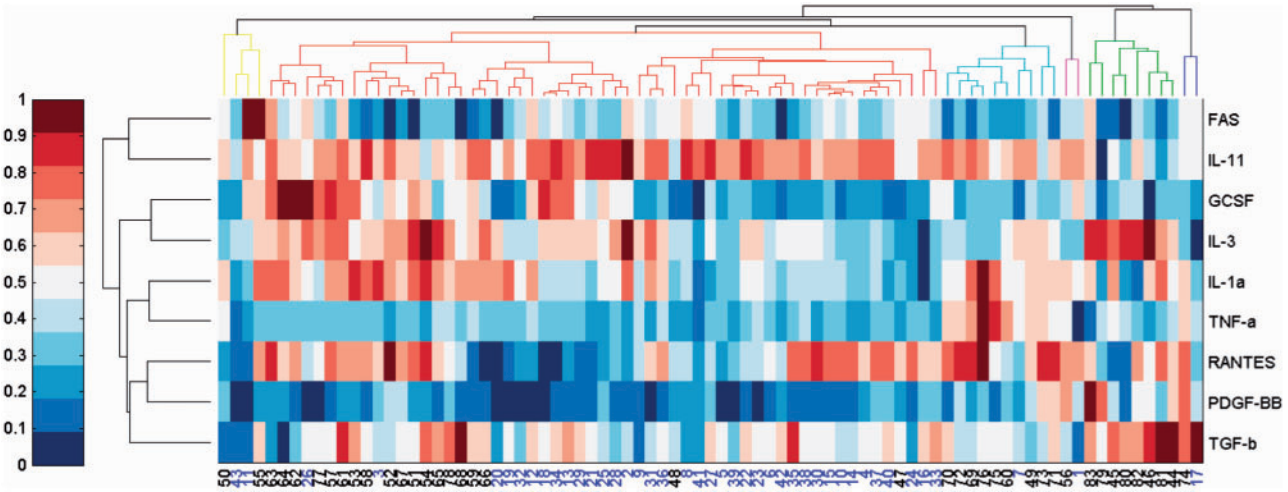

Fig. 7. Heat-map of AD training set with our nine selected plasma proteins. The numbers at the bottom of the heat-map represent indices of data samples in the train set. Data sample numbers 1–43 are AD data and from 44 to 83 are non-AD data samples. For ease of understanding, blue numbers represents AD samples and black non-AD samples. The dendrograms shown are constructed using 'correlation' distance and 'average' linkage. The arrangement of the coloured data sample indices, as ordered by the bottom up clustering, reveals that the samples are efficiently clustered into AD and non-AD categories. The names of the proteins are shown on the right of the heat-map (Color version of this figure is available at *Bioinformatics* online.)

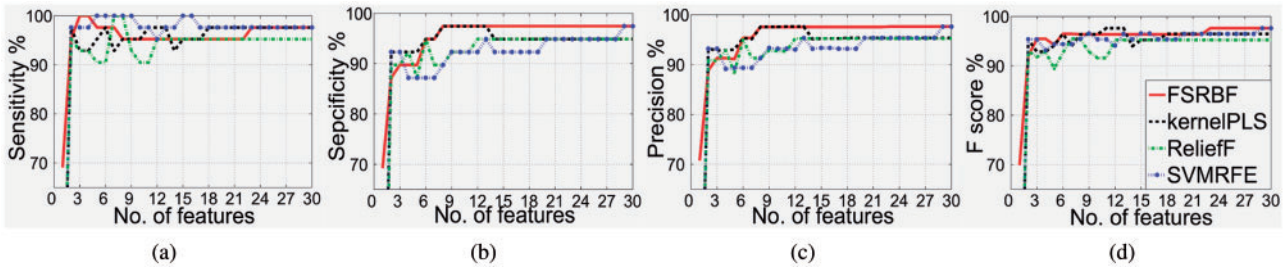

Fig. 8. Performance comparison of FSRBF, kernelPLS, ReliefF and SVMRFE. (**a**) Sensitivity, (**b**) specificity, (**c**) precision and (**d**) *F*-score

These results are depicted in Figure 6. Figure 6 indicates that for most of the classifiers, the set of nine proteins improves the classification accuracies of the MCI dataset (47 samples) into AD and non-AD classes by ∼4%. On the other hand, for most classifiers, the set of 18 proteins better classifies the OD dataset (11 samples) by ∼18%. Therefore, when the set of 18 proteins performs better, the improvement in performance appears higher than the cases when the set of nine proteins perform better. But, the OD dataset has only 11 samples and hence ∼18% better classification amounts to just two extra correct classifications. Moreover, we have identified a much smaller set of uncorrelated proteins which leads to low computational complexity and may result in decreased diagnostic expenses. We draw a heat-map using our nine proteins in Figure 7. Observe that most AD samples, shown in blue, form a cluster in the middle and the non-AD samples form two clusters on the two sides. This pattern shows that our nine proteins form AD-specific signature. A heat-map for the 18 proteins identified in Ray *et al.* (2007) is shown in Supplementary Figure S1a and is compared in Supplementary Section 2 with the heat-map presented in Figure 7.

To compare the efficiency of FSRBF with that of kernelPLS, ReliefF and SVMRFE, we run them on the same training set (Section 3.2) for AD-specific feature selection. Each feature selection method assigns a numerical value to each feature according to the importance of the feature. For each method we have considered the top 20 features and then identified the features that are common to all four sets. We have found eight such common biomarkers. Of these eight proteins, seven are present in the set of 18 reported in Ray *et al.* (2007). These seven are shown in Figure 5b in non-bold font. The biomarker FAS which is selected by FSRBF as the ninth important feature, but not reported in Ray *et al.* (2007), is also selected by ReliefF, kernelPLS and SVMRFE. For the next set of experiments, we train the classifiers with the train set of 83 samples and test the performance on the first test set of 81 samples (Section 3.2). Figure 8 compares the performances of the features selected by different methods. Different metrics: recall/sensitivity, specificity, precision and *F*-score have been used for performance comparison. These metrics are plotted against increasing number of features as selected by different feature selection methods. The classifier used is SVM with linear kernel. We choose LIBSVM implementation of SVM with default value (0.5) of the

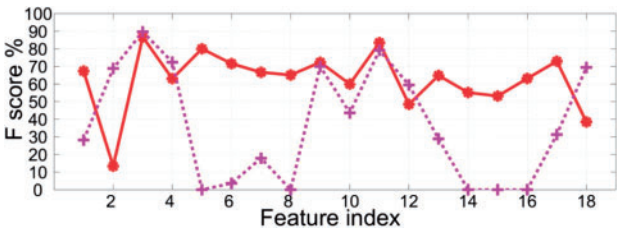

Fig. 9. Performance increase of each of the 18 signalling proteins suggested by Ray *et al.* (2007) by addition of the two newly selected plasma proteins, TGF-b and FAS. The numbers along the horizontal axis show the index of each individual feature in the set of 18 signalling proteins. The dashed line shows the *F*-score for each of the 18 signalling proteins individually. The solid line indicates the *F*-score when TGF-b and FAS are added to each of the 18 signalling proteins

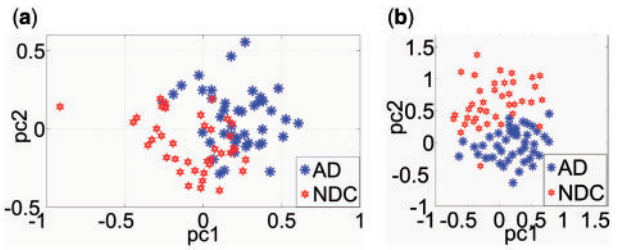

Fig. 10. Scatter plot of 'first test' data using nine selected proteins in (**a**) 2D principal space and after (**b**) Sammon mapping to 2D space

parameter ν. Observe that FSRBF shows the highest sensitivity (100%), specificity (97.4%) and precision (97.6%) with the minimum number of features (three, eight and eight, respectively).

We further analyse the contribution of the two newly selected proteins (TGF-b and FAS) which were not mentioned by Ray *et al.* (2007). Experiments show that addition of TGF-b and FAS to the set of 18 proteins reported by Ray *et al.* (2007) does not make any change to the performance of the set of 18 proteins. This shows that TGF-b and FAS are not derogatory features. We further analyse how the two newly identified proteins affect the performance of each of the proteins reported by Ray *et al.* (2007). Figure 9 shows the experimental results in terms of *F*-score. Note that high *F*-score indicates both high sensitivity and high precision. For 14 proteins identified by Ray *et al.* (2007), addition of TGF-b and FAS increases the *F*-score. It is observable that each of the proteins MCP-3, PARC, ICAM, IGFBP-6 and IL-11 (feature indices 5, 8, 14, 15 and 16, respectively, in Fig. 9) identifies all the samples as non-AD, resulting in zero sensitivity and thus no *F*-score. TGF-b and FAS, when added to each of these five proteins, results in a significant *F*-score.

4.4 PCA and Sammon mapping of selected proteins

To further analyse the characteristics of the selected nine proteins, we perform principal component analysis (PCA). Experiments show that for our nine proteins, 92.69% variance is concentrated in seven principal components (PCs) whereas, for 18 proteins, 91.12% variance is concentrated in nine PCs. Figure 10a shows the scatter plot of the first test dataset in 2D principal domain of our nine proteins and Supplementary Figure S2 shows similar scatter plots for the train set and the rest two test sets: OD and MCI. For a fair comparison, Supplementary Figure S2 also shows similar scatter plots in 2D principal space for the 18 proteins reported by Ray *et al.* (2007). The two figures reveal that clustering capability of both the protein sets (nine and 18) is comparable.

Since we have used non-linear architecture for selection of plasma signalling proteins, linear PCA may not reflect the appropriate discriminating power of the selected predictors. So, we have also explored non-linear Sammon mapping. The 2D Sammon projection of the first test set using nine signalling proteins and that of the 18 signalling proteins (reported by Ray *et al.*, 2007) are shown in Figure 10b and Supplementary Figure S3, respectively. For the sake of completeness, the 2D Sammon projections of the training set and the remaining two test sets, OD and MCI, using our nine signalling proteins as well as using the set of 18 signalling proteins (reported in Ray *et al.*, 2007) are also shown in Supplementary Figure S3. In the two scatter plots of Figure 10, especially in Figure 10b, two distinct clusters corresponding to AD and non-AD samples can be observed.

4.5 Identifying incipient AD from MCI samples

The last columns of the bar-graphs in Figure 6, and Supplementary Figure S2d and h, show that both the set of nine and the set of 18

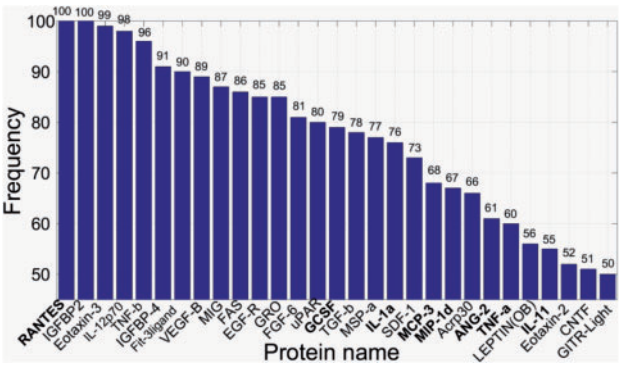

Fig. 11. Frequency count of the features selected for prediction of AD from MCI

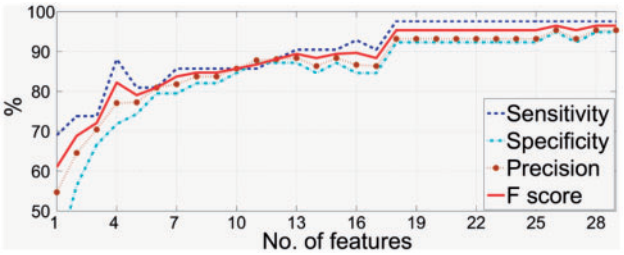

Fig. 12. Performance of the plasma proteins that carry AD specific signature in MCI patients and are selected by FSRBF. The vertical axis represents performance percentage in terms of sensitivity, specificity, precision and *F*-score

signalling proteins are not good enough to identify those pre-symptomatic individuals with MCI who will eventually suffer from AD. To identify the bio-markers that can single out the MCI patients who will gradually convert to AD, we do the feature selection experiments afresh on the MCI data. Since limited numbers (47) of samples are available, we do the experiments in 10 folds. First, the set of 47 MCI data samples is divided into 10 parts say, $D_1, D_2, ..., D_{10}$. We use each D_i as the test data and the remaining data $T_i = \cup_{j \neq i} D_j$ as the training data. We perform 10-fold CV on T_i to choose the architecture with the highest validation accuracy. Once the architecture for T_i is selected, FSRBF is run on T_i 10 times with different initializations and tested on D_i as test data. So we get 10 sets of β values for each D_i. So for 10 iterations, we get a total of 100 sets of β values. The features (signalling proteins) for which β value is >0.1135, are chosen from each set. Then we select 29 signalling proteins for which frequency count is >50. The frequency of the selected signalling proteins in these 100 sets of β values is shown in Figure 11. Note that this set of 29 proteins is not necessarily the minimal set of plasma proteins that carry AD specific signature in MCI patients. The frequency of occurrence of these proteins being more that 50 in 100 runs of FSRBF indicates that these proteins may carry important AD specific signature in MCI patients. A heat-map along with a dendrogram constructed using 'correlation' distance and 'average' linkage for the 29 proteins is shown in Supplementary Figure S1b. To visualize how the MCI data are distributed in this 29-dimensional feature space, we have performed PCA and have plotted the 47 MCI data in the domain spanned by top two PCs in Supplementary Figure S4a. The corresponding Sammon plot is shown in Supplementary Figure S4b. Supplementary Figure S4a and b clearly show that the selected 29 proteins naturally form clusters for AD and non-AD specific signatures in MCI samples.

Figure 12 depicts the performance versus number of proteins graph for the 29 selected proteins. Observe that the *F*-score becomes

~95% for a set of 18 features and attains its maximum for a set of 26 features. It should also be noted that neither the set of nine proteins suggested by FSRBF nor the set of 18 proteins suggested by Ray *et al.* (2007) could attain more that 73% accuracy with SVM classifier with linear kernel in predicting incipient AD from MCI samples. These findings indicate that the phenotype signature of AD hidden in MCI samples may be different from that of confirmed AD samples.

5 Discussion

5.1 Biological relevance of the selected proteins

In Section 4, we notice that seven of our nine proteins selected from AD 'train' set are the same as those reported in Ray *et al.* (2007). To further investigate the biological significance of the two newly selected proteins, TGF-b and FAS, we manually search PubMed and other research articles. Craig-Schapiro *et al.* (2011) indicate FAS, and Lee *et al.* (2010) and Town *et al.* (2008) indicate TGF-β 3 as important for discriminating AD from non-AD samples. Peress and Perillo (1995) have noticed strikingly selective staining of Hirano bodies produced by TGF-β 3. Note that patients with AD are found to have more Hirano bodies than normal persons in the same age group. We also search through integrative multi-species prediction (http://imp.princeton.edu/) interactive web server and find that FAS is related to the protein TNF for AD. These findings strengthen the idea that FAS and TGF-b are important biomarkers for AD.

We have reported a set of 29 plasma proteins in Section 4.5 that can single out the MCI patients who will gradually convert to AD. Of these 29 proteins, eight are reported as carrying AD specific signature in Ray *et al.* (2007). These eight proteins are marked in bold in Figure 11. Of the set of 29 proteins, seven (IL-11, IL-1a, GCSF, TNF-a, RANTES, TGB-b and FAS) are also present in the set of nine proteins.

Literature survey shows that different subsets of this set of 29 proteins are found to play significant roles in different biological processes that are related to AD and/or other neurological brain disorders. This is discussed in detail in Section 3 of the Supplementary Material. Some regulatory protein interaction networks for different biological processes created using gene mania online (http://www.genemania.org/) and defined by the 29 plasma proteins are shown in Supplementary Figure S5. The result of DAVID query (http://david.abcc.ncifcrf.gov/conversion.jsp) for specific biological processes controlled by the indicated 29 proteins is shown in Supplementary Figure S6. Literature further reveals that many of the 29 proteins as listed in Figure 11 are mentioned to play important roles either in distinguishing AD and non-AD disorders or brain injury or related diseases. For example, in connection with dementia/AD we find mention of IFGBP-2 in Tham *et al.* (1993); Eotaxin-3 in Westin *et al.* (2012), Craig-Schapiro *et al.* (2011), Hu *et al.* (2010); IL-12 in Tobinick and Gross (2008), Rentzos *et al.* (2006); IGFBP-4 in Beilharz *et al.* (1993); VEGF-B in Eriksson (2013) and TNF-a in Tobinick and Gross (2008). These facts show that the set of 29 plasma proteins suggested by FSRBF may carry AD-specific signature in MCI patients.

5.2 Pros and cons of FSRBF

Although we have used FSRBF for discovering genetic markers for AD, the FSRBF algorithm described in Section 3.1 is a general algorithm for feature selection, which judiciously combines the feature selection process and the universal function approximation capability of RBF NN. Therefore, FSRBF can be used for any feature selection task including biological data as well as data where classes are not linearly separable. FSRBF can also be used for function approximation type problems. While selecting features, FSRBF exploits the non-linear interactions among features and also that between the features and the target outputs. This is a unique advantage of FSRBF and that is why it can yield better results compared with several filter-based methods. However, the results of FSRBF depend on the initialization and this dependence may become more prominent when we have many correlated features because FSRBF cannot control the level of redundancy among the set of selected features. Moreover, since the present version of FSRBF uses Euclidean distance in the activation function of the hidden layer nodes, if the data are in really very high dimension, FSRBF may not yield the most desired results. For such datasets a hierarchical version of FSRBF may be developed, which we plan to investigate in future.

6 Conclusions

In this study, we have used FSRBF, an adapted version of the GFSRBF neural network for selection of plasma signalling proteins that can predict clinical AD. Compared with a state-of-the-art method, FSRBF finds a smaller set of plasma proteins exhibiting comparable power in discriminating Alzheimer's patients from NDC. FSRBF is also found to be very effective in finding a set of plasma signalling proteins that can distinguish incipient AD from MCI. The biological relevance of the selected proteins in AD and MCI is discussed. The utility of the selected proteins is further established using PCA, heat-map and Sammon projections. FSRBF is a general purpose tool and can be used for finding markers for other diseases as well as for non-biological numeric data.

Conflict of Interest: none declared.

References

Basak,J. and Mitra,S. (1999) Feature selection using radial basis function networks. *Neural Comput. Appl.*, **8**, 297–302.

Beilharz,E.J. *et al.* (1993) Differential expression of insulin-like growth factor binding proteins (IGFBP) 4 and 5 mRNA in the rat brain after transient hypoxic-ischemic injury. *Mol. Brain Res.*, **18**, 209–215.

Bermejo,P.F. *et al.* (2008) Incidence and subtypes of dementia in three elderly populations of central Spain. *J. Neurol. Sci.*, **264**, 63–72.

Carlo,D. *et al.* (2002) Incidence of dementia, Alzheimer's disease, and vascular dementia in Italy. The ILSA Study. *J. Am. Geriatr. Soc.*, **50**, 41–48.

Craig-Schapiro,R. *et al.* (2011) Multiplexed immunoassay panel identifies novel CSF biomarkers for Alzheimer's disease diagnosis and drognosis, *PLoS ONE*, **6**, e18850.

Chakraborty,D. and Pal,N.R. (2008) Selecting useful groups of features in a connectionist framework. *IEEE Trans. Neural Networks*, **19**, 381–396.

Eriksson,U. (2013) Targeting VEGF-b regulation of fatty acid transporters to modulate human diseases. *US Patent 8, 383,112*.

German,D.C. *et al.* (2007) Serum biomarkers for Alzheimer's disease: prot discovery. *Biomed. Pharmacother.*, **61**, 383–389.

Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Hansson,O. *et al.* (2006) Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. *Lancet Neurol.*, **53**, 228–234.

Hernandez,F. and Avila,J. (2007) Taopathies. *Cell. Mol. Life Sci.*, **64**, 2219–2233.

Hooper,N.M. *et al.* (2005) Roles of proteolysis and lipid rafts in the processing of the amyloid precursor protein and prion protein. *Biochem. Soc. Trans.*, **33**, 335–338.

Hu,W.T. *et al.* (2010) Novel CSF biomarkers for Alzheimers disease and mild cognitive impairment. *Acta Neuropatholica*, **119**, 669–678.

Hye,A. *et al*. (2006) Proteome-based plasma biomarkers for Alzheimer's disease. *Brain*, **129**, 3042–3050.

Kira,K., Rendell,L.A. (1992) The feature selection problem: traditional methods and a new algorithm. *AAAI*, **2**, 129–134.

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

Kononenko,I. *et al*. (1997) Overcoming the myopia of inductive learning algorithms with RELIEFF. *Appl. Intell.*, **7**, 39–55.

Lee,M.H. *et al*. (2010) TGF-b induces TIAF1 self-aggregation via type II receptor-independent signaling that leads to generation of amyloid b plaques in Alzheimer's disease. *Cell Death Dis.*, **1**, e110.

Pal,N.R. and Malpani,M. (2012) Redundancy-constrained feature selection with radial basis function networks. In: Proceedings of the WCCI 2012 IEEE World Congress on Computational Intelligence, June 10–15, 2012. Brisbane, Australia, doi:10.1109/IJCNN.2012.6252638.

Peress,N.S. and Perillo,E. (1995) Differential expression of TGF-beta 1, 2 and 3 isotypes in Alzheimer's disease: a comparative immunohistochemical study with cerebral infarction, aged human and mouse control brains. *J. Neuropathol. Exp. Neurol.*, **54**, 802–811.

Ray,S. *et al*. (2007) Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. *Nat. Med.*, **13**, 1359–1362.

Rentzos,M. *et al*. (2006) Interleukin-12 is reduced in cerebrospinal fluid of patients with Alzheimer's disease and frontotemporal dementia. *J. Neurol. Sci.*, **249**, 110–114.

Sun,S. *et al*. (2014) A kernel-based multivariate feature selection method for microarray data classification. *PLoS ONE*, **9**, e102541.

Teramoto, R. (2008) Prediction of Alzheimer's diagnosis using semi-supervised distance metric learning with label propagation. *Comput. Biol. Chem.*, **32**, 438–441.

Tham,A. *et al*. (1993) Insulin-like growth factors and insulin-like growth factor binding proteins in cerebrospinal fluid and serum of patients with dementia of the Alzheimer type. *J. Neural Transm. Park. Dis. Dement. Sect.*, **5**, 165–176.

Tiraboschi,P. *et al*. (2004) The importance of neuritic plaques and tangles to the development and evolution of AD. *Neurology*, **62**, 1984–1989.

Tobinick,E.L. and Gross,H. (2008) Rapid cognitive improvement in Alzheimer's disease following perispinal etanercept administration. *J. Neuroinflammation*, **5**, 2, doi:10.1186/1742-2094-5-2.

Town,T. *et al*. (2008) Blocking TGF-Smad2/3 innate immune signaling mitigates Alzheimer-like pathology. *Nat. Med.*, **14**, 681–687.

Westin,K. *et al*. (2012) CCL2 is associated with a faster rate of cognitive decline during early stages of Alzheimer's disease. *PLoS ONE*, **7**, e30525.

Xiao,Z. *et al*. (2005) Proteomic patterns: their potential for disease diagnosis. *Mol. Cell. Endocrinol.*, **230**, 96–106.