NIH Public Access
Author Manuscript
ACM Conf Bioinform Comput Biol Biomed Inform (2013). Author manuscript; available in PMC 2015 January 15.

Published in final edited form as:
ACM Conf Bioinform Comput Biol Biomed Inform (2013). 2013 ; 2013: 569. doi: 10.1145/2506583.2506637.

Classification of Alzheimer Diagnosis from ADNI Plasma Biomarker Data

Jue Mo[1],
Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, MD 20892, USA

Stuart Maudsley,
Receptor Pharmacology Unit, National Institute on Aging, National Institutes of Health, Baltimore, MD 21224, USA

Bronwen Martin,
Metabolism Unit, National Institute on Aging, National Institutes of Health, Baltimore, MD 21224, USA

Sana Siddiqui,
Receptor Pharmacology Unit, National Institute on Aging, National Institutes of Health, Baltimore, MD 21224, USA

Huey Cheung, and
Div. of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, MD 20892, USA

Calvin A. Johnson
Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, MD 20892, USA

for the Alzheimer's Disease Neuroimaging Initiative*

Jue Mo: jue.mo@nih.gov; Stuart Maudsley: maudsleyst@mail.nih.gov; Bronwen Martin: martinbro@mail.nih.gov; Sana Siddiqui: sana.siddiqui@nih.gov; Huey Cheung: cheungh@mail.nih.gov; Calvin A. Johnson: johnson@mail.nih.gov

Abstract

Research into modeling the progression of Alzheimer's disease (AD) has made recent progress in identifying plasma proteomic biomarkers to identify the disease at the pre-clinical stage. In contrast with cerebral spinal fluid (CSF) biomarkers and PET imaging, plasma biomarker diagnoses have the advantage of being cost-effective and minimally invasive, thereby improving our understanding of AD and hopefully leading to early interventions as research into this subject advances. The Alzheimer's Disease Neuroimaging Initiative* (ADNI) has collected data on 190

[1]Jue Mo works under contract to NIH in the Health Group, SRA International Inc., Fairfax, VA, 22033, USA

*Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

plasma analytes from individuals diagnosed with AD as well subjects with mild cognitive impairment and cognitively normal (CN) controls. We propose an approach to classify subjects as AD or CN via an ensemble of classifiers trained and validated on ADNI data. Classifier performance is enhanced by an augmentation of a selective biomarker feature space with principal components obtained from the entire set of biomarkers. This procedure yields accuracy of 89% and area under the ROC curve of 94%.

Keywords

Alzheimer's Disease Neuroimaging Initiative (ADNI); feature clustering; feature augmentation

1. INTRODUCTION

In the working model for Alzheimer's Disease (AD) progression, a cascade of events starts with the buildup of amyloid plaque, followed by tau-mediated neuronal injury, and then by memory loss and finally clinical diagnosis of AD [1]. Recently, Prestia *et al*. [2] have provided clinical evidence that the core biomarker patterns are c[1] onsistent with this model. Specifically, the model predicts that tracer retention on amyloid PET imaging and low Aβ-42 concentration in the cerebral spinal fluid (CSF) should become abnormal earlier in the disease course, followed by cortical hypometabolism on F18-FDG-PET, and finally brain atrophy in structural MRI. Although biomarkers obtained through invasive collection of CSF and expensive PET imaging are the most consistent and reliable, predictive biomarkers that can be collected cost-effectively and in a minimally invasive manner would be preferred [3].

A number of investigators have reported progress in identifying plasma-based proteomic biomarkers and their effectiveness in predicting AD and mild cognitive impairment (MCI). In 2007, Ray *et al*. [4] identified 18 signaling proteins in blood plasma that can be used to classify blinded samples from MCI subjects who progressed to AD two to six years later. This study incorporated both unsupervised and supervised machine learning methodology. Ravetti and Moscato [5] re-analyzed the dataset of Ray *et al*. and obtained equivalent results with smaller 6-protein and 5-protein signatures using standard classification algorithms. Multivariate linear regressions correlating plasma and CSF biomarkers were investigated by Hu *et al*. [6] using ADNI data. Among these, changes in APOE, BNP, CRP, and pancreatic polypeptide levels were also associated with AD diagnosis and CSF AD biomarkers. APOE has been identified as the most predictive biomarker by Johnstone *et al*. [7], who also identified a limited set of paired biomarkers via univariate entropy filtering and the α-β-k feature selection process, achieving accuracy in excess of 85%.

Other investigators have modeled the longitudinal progression of clinical AD assessments. Doody *et al.* [8] performed mixed effects regression modeling to predict longitudinal performance on standard clinical measures of AD. A sigmoidal model of the longitudinal changes in AD assessment cognitive sub-scale (ADAScog) was developed by Samtani *et a*l. [9]. Yet, the main contributors in their predictive model were demographic factors and

clinical assessment. To our knowledge, there are no studies that incorporate the full set of AD biomarker data in a comprehensive model.

Complex processes associated with AD are mediated by interactions of functionally related proteins [10]. Since these interactions between the plasma biomarkers are not fully understood, a model that incorporates as many of the biomarker data as practical could be valuable. In this paper, our goal is to build a predictor of clinical assessments from plasma protein biomarker data, which takes advantage of the full set of available ADNI biomarker data, and improves prediction accuracy, compared to previous investigations into plasma biomarkers prediction.

2. DATASET AND EVALUATION†

Data used in our study were obtained from the ADNI database (adni.loni.ucla.edu). The set of biomarkers and the experimental procedures used to obtain them are described in [11].

Participants received a diagnosis at their first or baseline visit to one of the consortium clinics of cognitively normal (CN), mild cognitive impairment (MCI) or AD based on clinical and neuropsychological testing in accordance with the guidelines in [12]. Twelve months after the baseline assessment, plasma samples were collected again but only from a subset of these participants. In our study, we analyzed data from CN and AD participants, whose plasma samples were collected at both baseline and 12 months. Hence, we used a selected subset of the plasma samples, including 39 CN and 65 AD patients.

A rigorous quality-control procedure was observed on each of the set of 190 analytes as described in [11]. Analytes with more than 10% of samples below the assay detection limits were excluded by the ADNI Analysis Team. As such, 44 of the 190 analytes were excluded from our consideration, leaving 146 analytes for the feature selection and augmentation process.

As is customary, a log transformation was applied prior to feature selection because concentrations for these analytes are generally not normally distributed.

†Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.
The Principal Investigator of ADNI is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many co- investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

2.1 Evaluations

Previous studies on the AD plasma biomarkers investigated the predictive power of individual biomarkers as well as group of biomarkers [4–7]. In selecting biomarkers for a feature set, entropy heuristic was first applied to filter out non-informative analytes. A common method for feature selection is the (*α-β)-k feature selection process* [14]. Our work is an extension of these explorations, attempting to increase the predictive power by enhancing the feature space. We first re-evaluated three sets of biomarkers identified in [7], as summarized in Table 1. We then evaluated augmentations of these biomarker sets informed by clustering and dimensionality-reduction methods.

Classifiers were trained on various enhancements to the biomarker feature space using the provided AD and CN labels. Due to the limited number of available samples, we used a 13-fold cross validation procedure for each classifier to evaluate performance. Based on the predicted output, we calculated various measures of the classifier's performance, including accuracy, specificity, sensitivity and area under the ROC curve. Note that as the dataset is imbalanced with more AD than CN subjects, a single measure cannot represent overall performance.

3. METHODS

A naïve selection of all 146 analytes without feature selection would lead to an overfit to the noisy and uninformative features, leading to high variance in the classification. The feature selection process treats this problem of variance but may result in too few features and hence an underfit model that exhibits classification bias. The bias problem is confounded by the relatively small (65 positive, 39 negative) training and cross-validation set. Previous studies have concentrated on reduction of variance via feature selection but have ignored the unintended effect on bias that the smaller feature sets can yield. Our goal is to improve the trade-off between over-fit and under-fit by augmenting the feature space comprising selected features identified by previous literature [7] with additional features obtained from clustering and dimensionality reduction. Although the sample size of our data is not large, it is large enough to allow feature augmenting without incurring over-fitting problem.

In this work, we assume that the unselected features contain useful information yet are too noisy in their raw form to present individually. The literature is full of examples of improved classifier and regression performance from enhancement of feature sets with clusters of the original data. In [14] for example, an analysis of DNA microarray data was enhanced with Gene Set Enrichment Analysis, which improved the statistical significance of diabetic versus normal predictions. In our study, we have implemented and evaluated a number of clustering and dimensionality reduction methods to enhance the feature spaces of select plasma analytes. We have also investigated methods for improving classification performance via an ensemble of different classification algorithms. We evaluated the efficacy of various feature augmentations on the various classifier topology schemes.

3.1. Classifier Ensemble

Ensembles of classifiers reduce the potential for over-fitting that exists with high dimensional data and limited number of samples [15, 16]. As a result, such ensembles have been successfully applied to many bioinformatics applications, e.g. classification in microarray and proteomic data. An ensemble was constructed consisting of five conventional classification algorithms: libSVM [17] with linear kernel, binary decision tree, naïve Bayes, logistic regression and perceptron. The latter four methods were provided by Matlab (The Math Works, Natick, MA). All classifiers were trained on and performed prediction on the same sets of data. The topology of the ensemble includes an aggregating libSVM classifier, as depicted in Figure 1. The feature space of the aggregating classifier consists of the votes of the five first-layer classifiers. The aggregating classifier was trained on the same labels as the first-later classifier. We performed testing on the individual classifiers as well the ensemble result.

3.2. Feature Clustering Methods

As described in Section 2.1, a number of clustering and dimensionality reduction methods were implemented and tested against the task of cross-validating the individual classifiers in Section 3.1 as well as the ensemble. These clustering methods are described below.

Latent Process Decomposition (LPD)—LPD is an adaptive version of Latent Dirichlet Allocation (LDA) [18]. In LDA, the dependency between features is explained by the unobserved topics [19]. Instead of imposing a multinomial distribution on each feature, the LPD assumes the observations of each feature follow a Gaussian distribution, which is more suitable for proteomic data. First, the hyper-parameters for latent processes were estimated with the mean and variance of each feature with an iterative method with training data. Second, the probability of an observation conditioned on each latent processes was estimated. The vector of conditional probabilities was used as the feature vector in the classifiers.

Mixture of Gaussian Model Clustering (GMM)—GMM clustering is similar to that “soft K-means” method, in that it considers clusters as Gaussian distributions centered on their means [20]. The algorithm maximizes the conditional probability of data given the center of clusters. The vector of conditional probability was used as the feature vector in the classifiers, similar to the manner in which LPD uses conditional probabilities.

Self-Organizing Feature Map (SOFM)—SOFM involves training a neural network by using unsupervised learning to produce a low-dimensional representation of the input space of the training samples [21].

Principal Component Analysis (PCA)—PCA decomposes of the covariance matrix of features into principal components, or eigenvectors ordered by decreasing eigenvalue load [22]. Only the most significant components are retained, thereby reducing the dimensionality of the representation.

3.3. Feature Augmentation

In constructing an augmented feature space, we evaluated two different methods to combine feature sets, as described in [23]. The first method, known as 'early fusion,' uses a single classifier trained on all the feature sets, i.e., the feature vector itself is augmented prior to training the model. This method has the advantage of simplicity and can potentially capture interactions among different features. However, features from different sources might require different preprocessing or scaling or scaling procedures, and may be suitable for different kernels (in the case of SVM). The second method, known as 'late fusion,' combines the outputs of autonomous classifiers trained on each feature type separately. When we tested late fusion with the ensemble, we implemented ten first-layer classifiers, two for each algorithm corresponding to the two feature spaces, namely the select biomarkers, and the clustering representations, respectively.

4. RESULTS

The first experiment illustrates the effect of the ensemble on the select biomarkers, without augmentation. We used three sets of biomarkers identified by Johnstone *et al.* [7], as described in Section 2.1. We found that combining the three feature sets described in Table 1 into a single feature space yields better results than the individual feature spaces. We refer to this combined feature space comprising 11 single-features, 8 meta-features and 8 longitudinal-features identified, as the *selective feature set.* Table 2 provides cross-validation results by libSVM against the three sub-spaces in the selective feature set. Table 3 provides the cross-validation results of the various classifiers in the ensemble, on the combined selective feature set. The effect of combining the three sub-spaces of the selective feature set can be appreciated by comparing Table 2 to the first row in Table 3. The last row of Table 3 provides the cross-validation performance of the ensemble, which is clearly preferable to the individual classifiers on these data.

The ensemble's improvement in accuracy and the area under the ROC curve (AUC), when compared with libSVM along, is largely due to improved specificity. The accuracy that we have obtained, 86% for the ensemble, is similar to the 85% reported in [7], despite the fact that we have used less training data than Johnstone *et al.*, who used 112 AD and 58 CN subjects. By subsampling our labeled data, we discovered that performance measures decrease by approximately 5% as the training data as cut in half. It appears that effect of combining the individual select feature sub-spaces coupled with the effect of the ensemble, compensates for the effect of a smaller training set.

4.1. Feature Clustering

We applied the 5 different clustering methods described in Section 3.2 on the 146 analytes available in ADNI dataset. Selection of the reduced dimension size s (in LDA- the number of topics; GMM Clustering- the number of clusters; SOFM – number of nodes in the network; PCA- number of principal components) is critical to the desired effect of improving the classification result. As an example, Figure 2 shows the effect on classifier performance on the number of principal components from PCA selected for the augmented

feature space used in classification. An inspection of Figure 2 suggests that, for PCA, the value s=20 yields the best result. Although 21 or more features can be selected without compromising performance, these additional features would require unnecessary computational resources.

Using similar procedures as those represented in Figure 2, we determined the best value of *s* for each clustering method on these data. Table 4 lists these method-specific *s* values, as well as the corresponding classification results on the clustered features alone. Note that the first row of Table 4 presents the classification result when all 146 analytes are included in the feature space by "brute force." As we have discovered that the 20 PCA components yield the best classification results, we have concluded that PCA is the more effective of the clustering methods. Subsequent tests are performed on only the 20 PCA features.

We tested both single SVM (linear kernel) and the ensemble on the 20 PCA features alone as shown in Table 5. In this result, we find that linear SVM actually outperforms the ensemble.

4.3. Feature Augmentation

We compared the early fusion and late fusion methods of combining the selective feature set with the PCA features as described in Section 3.3. We assert that since the PCA features were computed from the entire set of 146 analytes, that the PCA feature space contains information not present in the selective feature set. For each fusion method, both single linear SVM and the ensemble classifier are implemented. For early fusion, we constructed a feature vector by concatenating the respective feature vectors from the selective feature-set and the 20 PCA feature-set. For late fusion, selective feature-set and PCA feature-set were used parallel as inputs for two classifiers, e.g. two independent SVMs. Then, the outputs of two SVMs are aggregated for final decision using an aggregating SVM. The result of combining the feature sets with early fusion and late fusion is shown in Table 6. Early fusion on a linear SVM appears to yield the best result, an accuracy of 89% and an area under the ROC curve of 94%.

5. CONCLUSION AND FUTURE WORK

Feature augmentation by PCA improved classification performance by 8% for accuracy, 9% for sensitivity, 13% for specificity and 3% for AUC by 3%, compared to using the full selected feature set without augmentation. Our overall best result of 89% accuracy compares favorably with the 85% accuracy reported by Johnstone *et al.* on a larger sample of the same ADNI dataset. As such, we believe that the PCA augmentation approach proposed here represents a clear improvement in predicting AD assessment from plasma biomarker data.

While we are encouraged by the effect of augmenting a biomarker feature space with features such as PCA features that were derived directly from the original data, there may be alternative clusters of the data based on other sources of evidence such as the literature. As such, we intend to investigate additional feature-space augmentations using latent semantic indexing [24] along with the augmentations proposed here. We also intend to explore

dynamic state models to exploit the additional longitudinal data available in ADNI. As we have discovered that a linear SVM yields the best result in an augmented feature space, we are encouraged that a linear model may be appropriate to describe the observations.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

References

1. Jack CR, 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 Neurology. 2010; 9(2010):119–128.
2. Prestia A, van der Flier WM, Ossenkoppele R, Van Berckel BV, Barkhov F, Teunissen CE, Wall AE, Carter SF, Scholl M, Choo IH, Nordberg A, Scheltens P, Frisconi GB. Prediction of dementia in MCI patients based on core diagnostic markers for Alzheimer disease. Neurology. 2013; 80(2013):1–9.
3. Williams R. Warning signs. Nature. 2011; 475(2011):S5–S7. [PubMed: 21760581]
4. Ray S, Britschgi M, Herbert C, Takeda-Uchimura Y, Boxer A, Bleddow K, Friedman LF, Galasko DR, Jutel M, Karydas A, Kaye JA, Leszek J, Miller BL, Minthon L, Quinn JF, Rabinocici GD, Robinson W, Sabbagh MN, So YT, Sparks DL, Tabaton M, Tinklenberg J, Yesavage JA, Tibshirani R, Wyss-Coray T. Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. Nature Medicine. 2007; 13(11):1359–1362.
5. Ravetti MG, Moscato P. Identification of a 5-Protein Biomarker Molecular Signature for Predicting Alzheimer's Disease. PLOS One. 2008; 3(9):e3111. [PubMed: 18769539]
6. Hu W, Holtzman DM, Fagan AM, Shaw LM, Perrin R, Arnold SE, Grossman M, Xiong C, Craig-Schapiro, Clark CM, Pickering E, Kuhn M, Chen Y, Van Deerlin VM, McCluskey L, Elman L, Karlawish J, Chen-Plotkin A, Hurtig HI, Siderowf A, Swenson F, Lee VM-Y, Morris JC, Trojanowski JQ, Soares H. Plasma multivariate profiling in mild cognitive impairment and Alzheimer disease. Neurology. 2012; 79:897–905. [PubMed: 22855860]
7. Johnstone D, Milward EA, Berretta R, Moscato P. Multivariate Protein Signatures of Pre-Clinical Alzheimer's Disease in the Alzheimer's Disease Neuroimaging Initiative (ADNI) Plasma Proteome Dataset. PLoS One. 2012; 7(4):e34341. [PubMed: 22485168]
8. Doody RS, Pavlik V, Massman P, Rountree S, Darby E, Chan W. Predicting progression of Alzheimer's disease. Alzheimer's Research & Therapy. 2010; 2:2.
9. Samtani MN, Farnum M, Lobanov V, Yang E, Raghavan N, Dibernardo A, Narayan V. Alzheimer's Disease Neuroimaging Initiative. An improved model for disease progression in patients from the Alzheimer's disease neuroimaging initiative. The Journal of Clinical Pharmacology. 2012; 52(5): 629–44.

10. Chadwick W, Brenneman R, Martin B, Maudsley S. Complex and Multidimensional Lipid Raft Alterations in a Murine Model of Alzheimer's Disease. Int J Alzheimers Dis. 2012; 604792:21151659.
11. Biomarkers Consortium. Use of Targeted Multiplex Proteomic Strategies to Identify Plasma-Based Biomarkers in Alzheimer's Disease – Data Primer. 2010. http://adni.loni.ucla.edu/wp-content/uploads/2010/11/BC_Plasma_Proteomics_Data_Primer.pdf
12. Alzheimer's Disease Neuroimaging Initiative. ADNI Procedures Manual. 2006. http://www.adni-info.org/Scientists/Pdfs/adniproceduresmanual12.pdf
13. Berretta R, Costa W, Moscato P. Combinatorial optimization models for finding genetic signatures from gene expression datasets. Methods in Molecular Biology. 2008; 453:363–377. [PubMed: 18712314]
14. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics. 2003; 34(3):267–73. [PubMed: 12808457]
15. Yang P, Hwa Yang Y, Zhou B, Zomaya YA. A Review of Ensemble Methods in Bioinformatics. Current Bioinformatics. 2010; (13):296–308.
16. Dietterich TG. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization. Machine Learning. 2000; 40:139–158.
17. Chang CC, Lin C-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2(3):1–27.
18. Rogers S, Girolami M, Campbell C, Breitling R. The Latent Process Decomposition of cDNA Microarray Data Sets. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2005; 2(2):143–56. [PubMed: 17044179]
19. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003; 3(4–5):993–1022.
20. Bishop, C. Pattern recognition and machine learning. New York: Springer; 2006.
21. Kohonen T. Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics. 1982; 43(1):59–69.
22. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010; 2:433–459.
23. Madani O, Georg M, Ross DA. On Using Nearly-Independent Feature Families for High Precision and Confidence. JMLR Workshop and Conference Proceeding. 2012; 25:269–284.
24. Chen H, Martin B, Daimon CM, Siddiqui S, Luttrell LM, Maudsley S. Textrous!: Extracting Semantic Textual Meaning from Gene Sets. PLoS One. 2013; 8(4):e62665. [PubMed: 23646135]

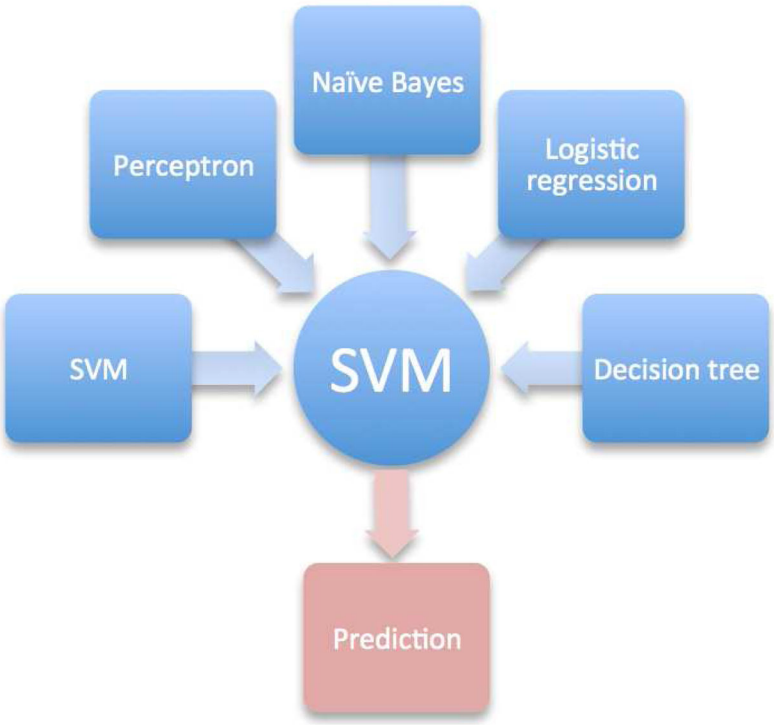

Figure 1.
Schematic of the ensemble of classifier methods.

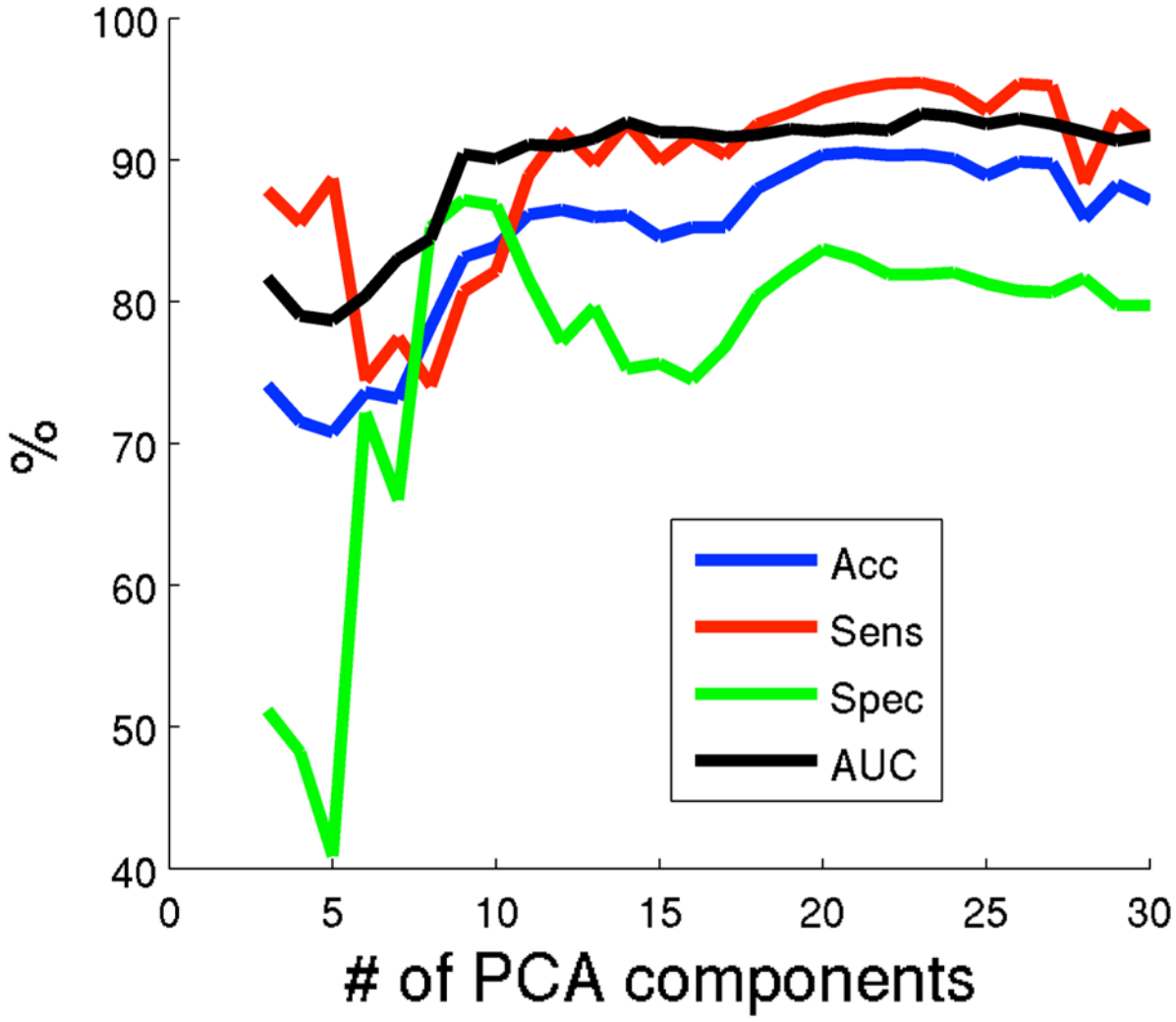

Figure 2.
Classification performance as function of principal components

Table 1

List of analyte signature sets identified in [7].

Single feature	a2-Macroglobulin, Angiotensinogen, Apolipoprotein A-II, Apolipoprotein E, Betacellulin, Fas Ligand, Heparin-Binding EGF-Like Growth Factor, Macrophage Inflammatory Protein-1a, Peptide YY, Serum Glutamic Oxaloacetic Transaminase, Transthyretin
Meta-feature	Angiopoietin-2 & Interleukin-16 Apolipoprotein A-II & Betacellulin Apolipoprotein E & Brain Natriuretic Peptide Apolipoprotein E & Serotransferrin Apolipoprotein E & Thrombopoietin Chromogranin-A & Heparin-Binding EGF-Like Growth Factor Interleukin-6 receptor & Macrophage Inflammatory Protein-1a Macrophage Inflammatory Protein-1a & Pulmonary and Activation-Regulated Chemokine
Longitudinal feature	Chemokine CC-4, Complement Factor H, Cystatin C, Interleukin-16, Kidney Injury Molecule 1, Macrophage Inflammatory Protein- 1a, Resistin, Sortilin

Table 2

Accuracy of selective feature set by linear SVM

Feature set	Accuracy	Sens.	Spec.	AUC
11 single- feature	0.82	0.86	0.71	0.85
8 meta-feature	0.81	0.90	0.63	0.83
8 longitudinal feature	0.64	1	0.05	0.67

Table 3

Accuracy of concatenated selective feature set (11 single-features + 8 meta-features + 8 longitudinal-features) by linear SVM and ensemble method

Classifier	Accuracy	Sens.	Spec.	AUC
Linear SVM	0.83	0.86	0.72	0.85
Naïve Bayes	0.28	0.08	0.62	0.35
Logistics regression	0.63	1	0.04	0.52
Perceptron	0.38	0.02	1	0.51
Decision Tree	0.75	0.78	0.69	0.73
Ensemble	**0.86**	**0.87**	**0.78**	**0.89**

Table 4

Accuracy of features extracted with clustering method, with SVM

Clustering	*s*	Accuracy	Sens.	Spec.	AUC
No clustering	146	0.81	0.86	0.72	0.89
LPD	5	0.61	0.80	0.28	0.56
GMM	10	0.87	0.89	0.74	0.90
SOFM	4	0.74	0.86	0.54	0.7
PCA	20	0.89	0.95	0.85	0.92

ACM Conf Bioinform Comput Biol Biomed Inform (2013). Author manuscript; available in PMC 2015 January 15.

Table 5

Accuracy of PCA feature by single SVM and ensemble method

Classifier	Accuracy	Sens.	Spec.	AUC
Linear SVM	0.89	0.95	0.85	0.92
Ensemble	0.86	0.92	0.74	0.87

Table 6

Accuracy of combining PCA and selective feature-set with early fusion (with single SVM classifier on concatenated features) and late fusion (5 independent classifiers on each feature family)

Classifier	Accuracy	Sens.	Spec.	AUC
Early Fusion (SVM)	0.89	0.97	0.85	0.94
Early Fusion (Ensemble)	0.84	0.89	0.82	0.91
Late Fusion (SVM)	0.86	0.92	0.74	0.87
Late Fusion (Ensemble)	0.89	0.92	0.85	0.92