differentiate among MS-related diseases, especially for SP-NMOSD/SN-NMOSD.

In conclusion, our applied proteomic pattern analysis facilitated the effective distinction of similar MS-related disorders, and revealed a possibility that these patterns, themselves, can be used as biomarkers for each disorder.

Acknowledgments

We thank Dr. Keiko Tanaka for analysis of the serum anti aquaporin-4 antibody in patient samples and Etsuko Nishiguchi for assistance in treating CSF samples. We also acknowledge Atsuhiro Kanzaki for excellent assistance in our statistical analysis.

References

1. Sospedra M, Martin R. Immunology of multiple sclerosis. Annu Rev Immunol 2005;23:683–747.
2. Wingerchuk DM, Lennon VA, Lucchinetti CF, et al. The spectrum of neuromyelitis optica. Lancet Neurol 2007;6:805–815.
3. Tanaka M, Tanaka K, Komori M. Interferon-beta(1b) treatment in neuromyelitis optica. Eur Neurol 2009;62:167–170.
4. Wingerchuk DM, Lennon VA, Pittock SJ, et al. Revised diagnostic criteria for neuromyelitis optica. Neurology 2006;66:1485–1489.
5. Lennon VA, Wingerchuk DM, Kryzer TJ, et al. A serum autoantibody marker of neuromyelitis optica: distinction from multiple sclerosis. Lancet 2004;364:2106–2112.
6. Lennon VA, Kryzer TJ, Pittock SJ, et al. IgG marker of optic-spinal multiple sclerosis binds to the aquaporin-4 water channel. J Exp Med 2005;202:473–477.
7. Tanaka M, Tanaka K, Komori M, Saida T. Anti-aquaporin 4 antibody in Japanese multiple sclerosis: the presence of optic spinal multiple sclerosis without long spinal cord lesions and anti-aquaporin 4 antibody. J Neurol Neurosurg Psychiatry 2007;78: 990–992.
8. Pittock SJ, Weinshenker BG, Lucchinetti CF, et al. Neuromyelitis optica brain lesions localized at sites of high aquaporin 4 expression. Arch Neurol 2006;63:964–968.
9. McKeon A, Fryer JP, Apiwattanakul M, et al. Diagnosis of neuromyelitis spectrum disorders: comparative sensitivities and specificities of immunohistochemical and immunoprecipitation assays. Arch Neurol 2009;66:1134–1138.
10. Cabrera-Gomez JA, Bonnan M, Gonzalez-Quevedo A, et al. Neuromyelitis optica positive antibodies confer a worse course in relapsing-neuromyelitis optica in Cuba and French West Indies. Mult Scler 2009;15:828–833.
11. Miller DH, Leary SM. Primary-progressive multiple sclerosis. Lancet Neurol 2007;6:903–912.
12. Koch M, Kingwell E, Rieckmann P, Tremlett H. The natural history of primary progressive multiple sclerosis. Neurology 2009;73: 1996–2002.
13. Frischer JM, Bramow S, Dal-Bianco A, et al. The relation between inflammation and neurodegeneration in multiple sclerosis brains. Brain 2009;132:1175–1189.
14. Hammack BN, Fung KY, Hunsucker SW, et al. Proteomic analysis of multiple sclerosis cerebrospinal fluid. Mult Scler 2004;10: 245–260.
15. Noben JP, Dumont D, Kwasnikowska N, et al. Lumbar cerebrospinal fluid proteome in multiple sclerosis: characterization by ultrafiltration, liquid chromatography, and mass spectrometry. J Proteome Res 2006;5:1647–1657.
16. Stoop MP, Dekker LJ, Titulaer MK, et al. Quantitative matrix-assisted laser desorption ionization-Fourier transform ion cyclotron resonance (MALDI-FT-ICR) peptide profiling and identification of multiple-sclerosis-related proteins. J Proteome Res 2009;8: 1404–1414.
17. Dumont D, Noben JP, Raus J, et al. Proteomic analysis of cerebrospinal fluid from multiple sclerosis patients. Proteomics 2004;4: 2117–2124.
18. Lehmensiek V, Sussmuth SD, Tauscher G, et al. Cerebrospinal fluid proteome profile in multiple sclerosis. Mult Scler 2007;13:840–849.
19. D'Aguanno S, Barassi A, Lupisella S, et al. Differential cerebro spinal fluid proteome investigation of Leber hereditary optic neuropathy (LHON) and multiple sclerosis. J Neuroimmunol 2008;193: 156–160.
20. Irani DN, Anderson C, Gundry R, et al. Cleavage of cystatin C in the cerebrospinal fluid of patients with multiple sclerosis. Ann Neurol 2006;59:237–247.
21. Ottervald J, Franzén B, Nilsson K, et al. Multiple sclerosis: Identification and clinical evaluation of novel CSF biomarkers. J Proteomics 2010;73:1117–1132.
22. Comabella M, Fernandez M, Martin R. et al. Cerebrospinal fluid chitinase 3-like 1 levels are associated with conversion to multiple sclerosis. Brain 2010;133:1082–1093.
23. Misu T, Takano R, Fujihara K, et al. Marked increase in cerebrospinal fluid glial fibrillar acidic protein in neuromyelitis optica: an astrocytic damage marker. J Neurol Neurosurg Psychiatry 2009; 80:575–577.
24. Pasinetti GM, Ungar LH, Lange DJ, et al. Identification of potential CSF biomarkers in ALS. Neurology 2006;66:1218–1222.
25. Huang JT, Leweke FM, Oxley D, et al. Disease biomarkers in cerebrospinal fluid of patients with first-onset psychosis. PLoS Med. 2006;3:e428.
26. Sekiyama E, Matsuyama Y, Higo D, et al. Applying magnetic bead separation/MALDI-TOF mass spectrometry to human tear fluid proteome analysis. J Proteomics Bioinform 2008;1:368–373.
27. Polman CH, Reingold SC, Edan G, et al. Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria." Ann Neurol 2005;58:840–846.
28. Tanaka K, Tani T, Tanaka M, et al. Anti-aquaporin 4 antibody in selected Japanese multiple sclerosis patients with long spinal cord lesions. Mult Scler 2007;13:850–855.
29. Ivosev G, Burton L, Bonner R. Dimensionality reduction and visualization in principal component analysis. Anal Chem 2008;80: 4933–4944.
30. Alexandrov T, Decker J, Mertens B, et al. Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. Bioinformatics 2009;25:643–649.
31. Freiwald A, Sauer S. Phylogenetic classification and identification of bacteria by mass spectrometry. Nat Protoc 2009;4:732–742.
32. Nagy E, Maier T, Urban E, et al. Species identification of clinical isolates of Bacteroides by matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry. Clin Microbiol Infect 2009; 15:796–802.
33. Leray E, Yaouanq J, Le Page E, et al. Evidence for a two-stage disability progression in multiple sclerosis. Brain 2010;133: 1900–1913.
34. Tallantyre EC, Bo L, Al-Rawashdeh O, et al. Greater loss of axons in primary progressive multiple sclerosis plaques compared to secondary progressive disease. Brain 2009;132:1190–1199.
35. Appel SH, Beers DR, Henkel JS. T cell-microglial dialogue in Parkinson's disease and amyotrophic lateral sclerosis: are we listening? Trends Immunol 2010;31:7–17.

RESEARCH ARTICLE

Cerebrospinal Fluid Proteomic Patterns Discriminate Parkinson's Disease and Multiple System Atrophy

Noriko Ishigami, MD,[1] Takahiko Tokuda, MD, PhD,[1,2] Masaya Ikegawa, MD, PhD,[3]* Mika Komori, MD, PhD,[4] Takashi Kasai, MD, PhD,[1] Takayuki Kondo, MD, PhD,[5] Yumiko Matsuyama, PhD,[6] Takashi Nirasawa, PhD,[6] Herbert Thiele, PhD,[7] Kei Tashiro, MD, PhD,[3] and Masanori Nakagawa, MD, PhD[1]

[1]*Department of Neurology, Kyoto Prefectural University of Medicine, Kamigyo-ku, Kyoto, Japan*
[2]*Department of Molecular Pathobiology of Brain Diseases, Kyoto Prefectural University of Medicine, Kamigyo-ku, Kyoto, Japan*
[3]*Department of Genomic Medical Sciences, Kyoto Prefectural University of Medicine, Kamigyo-ku, Kyoto, Japan*
[4]*Department of Neurology, Graduate School of Medicine, Kyoto University, Shogoin, Sakyo-ku, Kyoto, Japan*
[5]*Department of Neurology, Tazuke Kofukai Medical Research Institute and Kitano Hospital, Kita-ku, Osaka, Japan*
[6]*Bruker Daltonics K.K., Yokohama, Kanagawa, Japan*
[7]*Bruker Daltonik GmbH., Bremen, Germany*

ABSTRACT: The differential diagnosis of Parkinson's disease and multiple system atrophy can be challenging, especially in the early stages of the diseases. We developed a proteomic profiling strategy for parkinsonian diseases using mass spectrometry analysis for magnetic-bead-based enrichment of cerebrospinal fluid peptides/proteins and subsequent multivariate statistical analysis. Cerebrospinal fluid was obtained from 37 patients diagnosed with Parkinson's disease, 32 patients diagnosed with multiple system atrophy, and 26 patients diagnosed with other neurological diseases as controls. The samples were from the first cohort and the second cohort. Cerebrospinal fluid peptides/proteins were purified with C8 magnetic beads, and spectra were obtained by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Principal component analysis and support vector machine methods are used to reduce dimension of the data and select features to classify diseases. Cerebrospinal fluid proteomic profiles of Parkinson's disease, multiple system atrophy, and control were differentiated from each other by principal component analysis. By building a support vector machine classifier, 3 groups were classified effectively with good cross-validation accuracy. The model accuracy was well preserved for both cases, training by the first cohort and validated by the second cohort and vice versa. Receiver operating characteristics proved that the peak of *m/z* 6250 was the most important to differentiate multiple system atrophy from Parkinson's disease, especially in the early stages of the disease. A proteomic pattern classification method can increase the accuracy of clinical diagnosis of Parkinson's disease and multiple system atrophy, especially in the early stages. © 2012 *Movement* Disorder Society

Key Words: Parkinson's disease; multiple system atrophy; proteomics; cerebrospinal fluid; biomarkers

Parkinson's disease (PD) is the second most common neurodegenerative disorder, increasing in prevalence with age.[1] Multiple system atrophy (MSA) is a rare atypical parkinsonian disorder and has a relatively poor prognosis compared with PD because of much more widespread neurodegeneration.[2] The diagnoses of PD and MSA are still based on clinical features,[1–4] and differential diagnosis may be challenging, especially in the early-disease stages.[1] The development of reliable biochemical markers would have profound implications for clinical management and basic research.

By its direct communication with the extracellular fluid surrounding brain cells, cerebrospinal fluid (CSF) directly reflects the metabolic and pathological status of the central nervous system and is an ideal source for biochemical markers for parkinsonian disorders.

Noriko Ishigami and Takahiko Tokuda contributed equally to this work.

*Correspondence to: Dr. Masaya Ikegawa, Department of Genomic Medical Sciences, Kyoto Perfectural University of Medicine, 465 Kajii-cho, Kamigyo-ku, Kyoto 602-8566, Japan; mikegawa@koto.kpu-m.ac.jp

Relevant conflicts of interest/financial disclosures: Nothing to report. Full financial disclosures and author roles may be found in the online version of this article.

Received: 20 June 2011; **Revised:** 27 February 2012; **Accepted:** 9 March 2012
Published online 1 June 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/mds.24994

Although previous studies have shown differences in the levels of certain CSF proteins such as α-synuclein and DJ-1 between patients with PD and those with other forms of parkinsonism, their potential as differentiating biomarkers for these diseases has not yet been validated.[5–8] It may not be realistic to expect to find a single biomarker for complex disease processes involving multiple underlying molecular mechanisms of pathogenic importance. With emerging state-of-the-art technology,[9–13] proteomic pattern analysis, a new method to search for biomarkers, is suitable for this purpose, as it examines a panel of molecules; moreover, this approach can effectively distinguish apparently closely related diseases of a complex nature by computational statistical methods.[14]

In the current study, we asked whether the CSF proteomic profile analyzed by the ClinProt protocol (Bruker Daltonik GmbH) could classify PD, MSA, and controls.[12,13] Various statistical and machine-learning methods have been used to analyze the high-dimensional data generated by mass spectrometry.[14] Principal component analysis (PCA) is an unsupervised dimension reduction method generating orthogonal projections of the data, which is useful to highlight distinctive patterns in multivariate data. In contrast, support vector machine (SVM) is a powerful supervised machine-learning method for classification and pattern recognition. Here we have demonstrated the suitability of the SVM classifier in the CSF proteome when trained in the first cohort to classify the samples in the second cohort and demonstrated the suitability when trained in the second cohort to classify the samples in the first cohort. Furthermore, we evaluated classification performance of the SVM by considering the area under the receiver operating characteristics (ROC) curve, and the most optimal mass was nominated for the classification of PD and MSA.

Our findings suggest that CSF proteomic pattern analysis can increase the accuracy of disease diagnosis of PD-related disorders and may ultimately aid physicians in appropriate therapeutic decision making.

Patients and Methods

Subjects

We enrolled 26 patients with clinically defined PD and 23 patients with probable and possible MSA as the first cohort (Table 1). These subjects were recruited from the Department of Neurology, Kyoto Prefectural University Hospital, Kyoto, Japan, between April 2002 and February 2009. The patients with PD or MSA were diagnosed according to the United Kingdom Parkinson's Disease Society Brain Bank clinical diagnostic criteria[1] and the second consensus criteria for MSA,[3] respectively. Clinical data were retrieved from patient charts and confirmed by 3 board-certified neurologists. We defined the patients with PD and MSA who were examined fewer than 3 years after onset as the patients in the early stage of the disease. The 26 age-matched control subjects (Table 1) in the first cohort were neurologically normal individuals who underwent lumbar puncture as part of the diagnostic process (n = 13), and controls with various neurologic disorders without involvement of the brain (n = 13), including patients with peripheral neuropathy (n = 6), myelopathy (n = 3), epilepsy (n = 3), and myopathy (n = 1).

TABLE 1. Patient demographics and clinical data from the first and second cohorts

Diagnosis	Number of cases	Sex (F/M)	Mean age (± SD) at LP (y)	Duration of disease (n)	
				<3 years	≥3 years
First cohort					
PD (total)	26	11/15	66.3 ± 11.2	13	13
H&Y 1–2	11	5/6	69.5 ± 11.8	7	4
H&Y 3–4	15	6/9	63.9 ± 10.5	6	9
MSA (total)	23	6/17	62.4 ± 7.5	15	8
Probable	7	2/5	60.6 ± 8.5	4	3
Possible	16	4/12	63.2 ± 7.1	11	5
Controls	26	12/14	63.4 ± 12.4	—	—
Second cohort					
PD (total)	11	4/7	64.6 ± 11.7	—	—
MSA (total)	9	5/4	56.1 ± 7.7	—	—

Abbreviations: PD, Parkinson's disease; MSA, multiple system atrophy; H&Y, Hoehn and Yahr stage; LP, lumbar puncture.

The second cohort, which was collected for validation, included 11 clinically defined PD patients whose CSF samples were collected and stored between January 2005 and January 2010 and 9 age-matched MSA patients whose CSF samples were taken and stored between January 1995 and January 2001 (Table 1). Both groups of patients were diagnosed according to the same clinical criteria applied to the first cohort. The diagnosis of each patient was concealed prior to experiments to facilitate blind testing.

All the study subjects provided written informed consent to participate, which was approved by the university ethics committee (Kyoto Prefectural University, Kyoto, Japan). The study procedures were designed and performed in accordance with the Declaration of Helsinki.

Collection and Preparation of CSF Samples

The collected CSF samples were gently mixed to avoid gradient effects, then stored at −80°C. A 20-μL aliquot of the CSF samples was subjected to SDS-PAGE followed by CBB staining to ensure that no samples were contaminated with hemoglobin (data not shown). A 5-μL aliquot of the CSF from each subject was purified using magnetic beads with a functionalized surface (hydrophobic C8-coated magnetic

beads, MB-HIC, Bruker Daltonik GmbH, Bremen, Germany) according to the manufacturer's protocol. For MS analysis, 1 μL of bead eluate was mixed with 10 μL of matrix solution (0.6 g/L α-cyano-4-hydroxycinnamic acid in 2:1 ethanol/acetone), and 1 μL of the mixture was then spotted in quadruplicate on a MALDI target MTP AnchorChip 600/384 (Bruker Daltonik GmbH, Bremen, Germany).

Mass Spectrometry

Samples applied to the AnchorChip were analyzed on an autoflex MALDI-TOF mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany) operating in positive-ion linear mode. To generate a spectrum, 1000 laser shots were acquired from random positions for each matrix spot. Four independent spectra were acquired for each spot. Acquisition was controlled by flexControl 3.0 software (Bruker Daltonik GmbH, Bremen, Germany) using the AutoXecute (Bruker Daltonik GmbH), and fuzzy control of laser intensity. The mass range analyzed was 1000–15,000 *m/z* at a signal-to-noise threshold of 5. Spectra were externally calibrated using a mixture of standardized protein/peptide calibrants (ClinProt Standard, Bruker Daltnik GmbH, Bremen, Germany). The same MS analysis was replicated 3 times on different experimental days.

Analysis of Proteomic Profiles

The resulting spectra were analyzed using ClinProTools 2.2 bioinformatic software (Bruker Daltonik GmbH, Bremen, Germany).[12] Peaks of interest were selected from the total average spectra, using a signal-to-noise threshold of 5. Data normalization was performed as (1) spectra normalization to the total ion current; (2) spectra recalibration using the prominent peaks; (3) baseline subtraction, smoothing, and peak detection; and (4) calculation of peak areas for each spectrum. First, PCA was employed to visualize the distribution of the data.[15] The feature selections of PCA are present in the top principal components (PCs), which separate the samples into homogeneous clusters and can be visualized in 3-D plots in which the calculated values for top PCs serve as *x*, *y*, and *z* axes.[14] The loading plots, which are the variance plots, show how PCs are related to the original peaks. An SVM, another machine-learning approach, was applied to the current mass data for selection of clusters of signals able to discriminate the 2 objective groups.[12] Cross-validation accuracy is the percentage of data correctly classified. To test the classifier accuracy, the training data set and the testing data set were tested vice versa. Finally, ROC analysis for each peak of interest was performed, and the area under the curve (AUC) score was plotted for each selected feature.[14]

Results

Principal Component Analysis for the First and Combined Cohort Data Sets

PCA scores plotted based on MALDI spectra of CSF samples showed a clear difference between MSA and control (Fig. 1C) and a probable difference between PD and control (Fig. 1A) in the first cohort samples. To our surprise, if we replicated this for the first and second cohorts combined, it showed extremely good separation between PD and control (Fig. 1B) and between MSA and control (data not shown). These data indicate that PD and MSA are quite different from the control in terms of CSF proteomic pattern. Between PD versus MSA, compared with the above-described comparisons, the differentiation was a bit decreased. However, in analyzing the second cohort, very good separation was observed. When we compared the early parkinsonian subsets, it also showed good separation. The PCA loadings plot, which provided information about the contribution of single peaks to the variance covered by the respective PC, demonstrated that no single peak significantly contributed to the variance, but many peaks contributed to discrimination between PD and control, between MSA and control, and between PD and MSA (data not shown).

SVM Model for the First and Second Cohorts

For further evaluation of experiment-to-experiment data stability, we performed mass data analysis on 3 different experimental days using the same sample sets (Table 2). Cross-validation analysis provided an estimate of the reliability of the SVM model to classify defined groups of spectra separate from each other (Table 2). After each model was generated, a 20% leave-out cross-validation process was performed with the software. We had also classified the obtained spectra from early-stage PD (ePD) and early-stage MSA (eMSA) by SVM, which resulted in cross-validation accuracy of about 90% (Table 2). Thus, patients with ePD or eMSA were well separated by SVM with better cross-validation accuracy than the patients with PD or MSA in early and more advanced stages combined. It was also unexpected that when the first and second cohorts were combined, the cross-validation score was extremely high between PD and MSA (Table 2). For SVM training, the ClinProtools software had selected several features to enable an efficient model generation by a designated algorithm. The features selected automatically by this software differed from 12 to 24 peaks in each analysis.

Detecting Useful Peaks for Differential Diagnosis between PD and MSA

When spectra were compared between subject groups, the discriminatory peaks were ranked according to the *P* value of a Wilcoxon rank sum test by

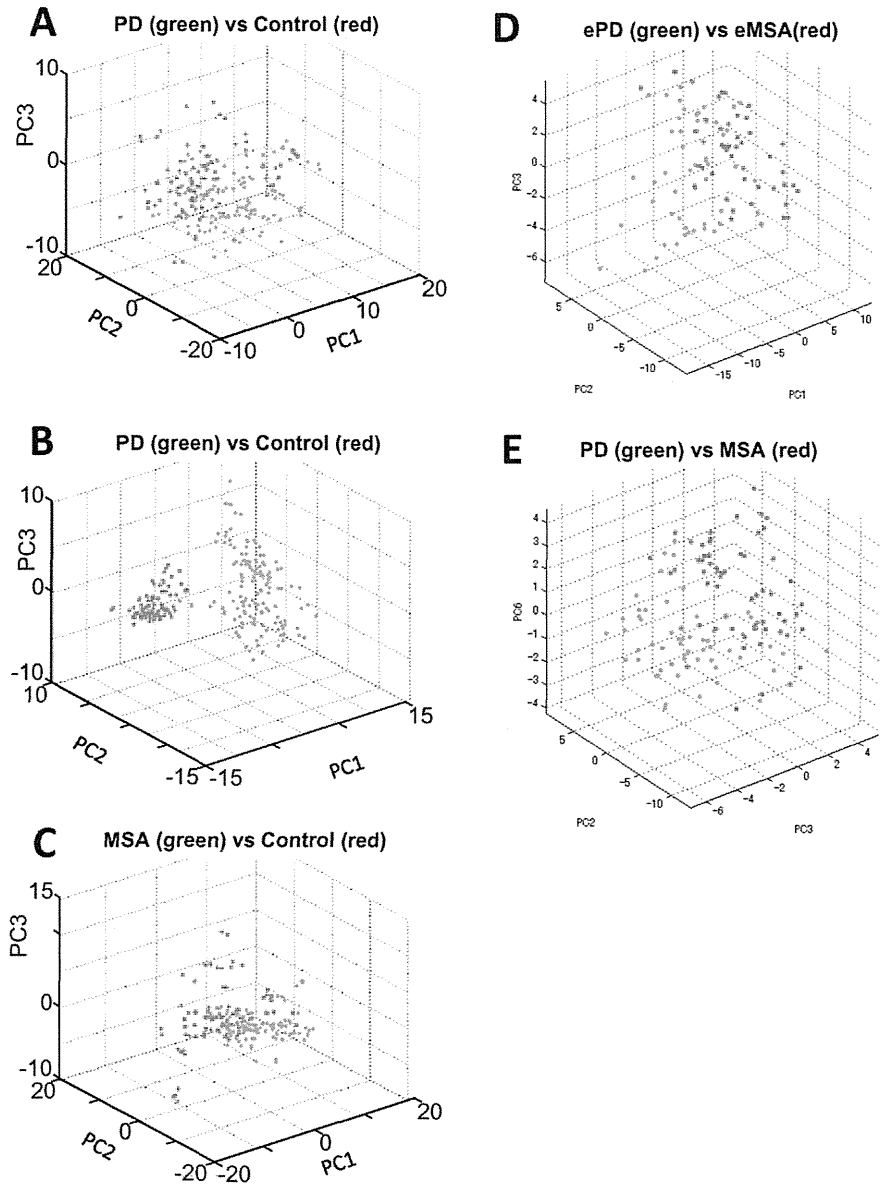

FIG. 1. PCA analyses of CSF proteomic profiles for the discrimination of the 2 indicated groups (**A–E**) displayed as scores on plots (PC, principal component). **A:** PD (green) versus control (red) from the first cohort. **B:** PD (green) versus control (red) from the first and second combined data sets. **C:** MSA (green) versus control (red) from the first cohort. **D:** PD of early-stage samples (ePD) from the first cohort (green) vesus MSA of early-stage samples (eMSA) from the first cohort (red). **E:** PD (green) versus MSA (red) from the second cohort.

ClinProTools. Among the 3 top-ranked peaks in the discrimination of each of the 3 pairs of comparisons, only the peak at *m/z* 6250 was commonly selected in all 3 pairs. The peak at *m/z* 6250 was highly expressed in control patients, but less expressed in PD patients and expressed the least in MSA patients (Fig. 2). In addition, the same respective order of the intensities of that peak was also observed for control, and early-stage patients with PD or MSA (control > ePD > eMSA). ROC curve analysis also proved the diagnostic capability of the peak at *m/z* 6250. The values of the AUC were 0.669 in PD versus control, 0.826 in MSA versus control, and 0.763 in PD versus MSA (Fig. 2B–D, respectively). In the ROC analyses of ePD or eMSA, the AUC values were 0.658 in ePD versus control, 0.886 in eMSA versus control, and 0.956 in ePD versus eMSA (Fig. 2F–H, respectively), suggesting that patients with ePD or eMSA were well discriminated by the peak at *m/z* 6250.

TABLE 2. Cross-validation (%) calculated by the SVM

Differential diagnosis	1st–1	1st–2	1st–3	1st + 2nd (2nd)
PD versus control	85.1	90.3	83.5	98.2 (ND)
MSA versus control	91.9	93.3	90.9	96.7 (ND)
PD versus MSA	85.7	83.8	86.2	90.2 (96.9)
Early PD versus early MSA	89.3	91.5	90.5	ND (ND)

1st–1, 1st–2, and 1st–3, results of 3 independent data analyses for the first cohort samples, taken on 3 different experimental days; early PD, early MSA, patients with disease duration less than 3 years after onset (only nominated in the first cohort); 1st, first cohort; 2nd, second cohort; 1^{st} + 2^{nd}, combining the data sets from the first and second cohorts; ND, not detected (because of a lack of data set information).

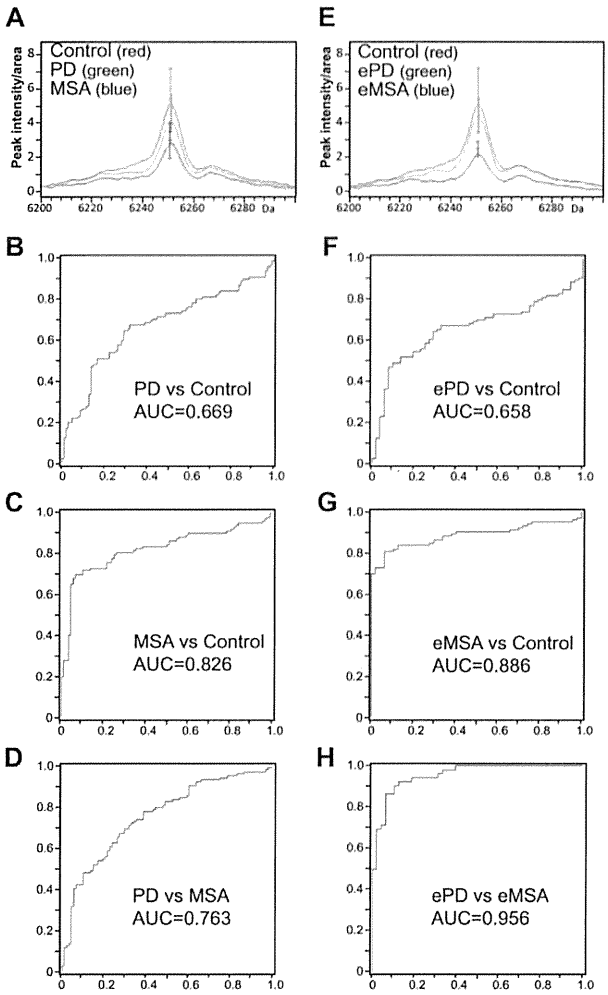

FIG. 2. Mean spectral features at *m/z* 6250 derived from CSF peptide profiling (**A, E**). Bars indicate average values with standard deviation. **B–D, F–H:** ROC curves with an AUC value of peak *m/z* 6250 in the discrimination of the indicated 2 groups (ePD/eMSA, early-stage patients with PD/MSA).

Classification Accuracy by SVM Was Tested for the First and Second Cohorts

The PCA analysis showed that patients with PD or MSA in the first cohort could be efficiently differentiated by their proteomic profiles and also by the peak intensities of the most discriminative peak, at *m/z* 6250. The differentiation ability of proteomic profiling by ClinProTools was put to practical use for the differential diagnosis of PD and MSA. We constructed the classification model to discriminate the groups of PD and MSA patients by using SVM in the training set (Table 3). In this analysis, ClinProTools selected several features identified as useful to classify PD or MSA automatically. Table 3 shows the results of the positive predictive values for MSA and PD (80.0% and 90.0%, respectively), sensitivities (88.9% and 81.8%, respectively), and specificities (81.8% and 88.9%, respectively), when we used the first cohort to build a SVM classifier and test on the second cohort. This process was further validated vice versa, namely, the classification model was generated by the second cohort and was validated on the first cohort. The result was well replicated, but a little lessened for the positive predictive value for MSA, partly because the first cohort contained larger numbers of samples with a variety of clinical situations. However, in both cases, the positive predictive value for PD was more than 80%, indicating that this classification model can efficiently discriminate patients with PD from those with MSA for the other data set.

Discussion

In this study, we have clarified that CSF proteomic profiles could differentiate patients with MSA or PD from each other, even if those with either disease were in the early stage of the illness. ClinProt is a well-established proteomics method that enables proteomic profiling by using the bioinformatics software ClinProTools, which provides algorithms of multivariate statistical analyses.[12,13,14] First by using PCA, we succeeded in differentiating PD, MSA, and the control by the dimension reduction approach. In the next step, we tried to make a classification based on a supervised machine-learning method, SVM. From the first cohort, we have shown that we can classify PD versus MSA with several features selected by SVM. The accuracy was validated by the second cohort, and this step was replicated vice versa. Furthermore, the classification efficiency of SVM and the discrimination power of a data set were proven by ROC analysis. When the number of features is large and the number

TABLE 3. Results of SVM model accuracy tested for the first and second cohorts

Model data set	Validation data set	Clinical diagnosis	Correct rate	Predictive value (%)	Sensitivity (%)	Specificity (%)
1st[a]	2nd	MSA	8/9	80.0	88.9	81.8
		PD	9/11	90.0	81.8	88.9
2nd[b]	1st	MSA	18/22	72.0	81.8	73.1
		PD	19/26	82.6	73.1	81.8

[a]Feature selected for first cohort, 24 peaks, automatically selected by the SVM of the ClinProTools software;
[b]feature selected for second cohort, 12 peaks, automatically selected by the SVM of the ClinProTools software; correct rate, number of truly classified cases for each data set. Sensitivity was calculated as the ratio of true positives against the total number of true-positive and false-negative cases. Specificity was calculated as the ratio of true negatives against the total numbers of true negatives and false positives. Predictive value means the proportion of patients with a positive test who have a disease and was calculated as the ratio of true positives against the total number of true positives and false positives.

of training patterns is comparably small, classification accuracy and the risk of a data-overfitting issue are potential drawbacks. However, this is the first report that has demonstrated the feasibility of the multivariate proteome profiling of CSF obtained from patients with parkinsonian disorders.

Accurate clinical diagnosis of PD and other parkinsonian disorders during life, especially in the early stages of the illness, is surprisingly difficult.[1–4] This unfortunate situation indicates that the development of reliable peptide/protein biomarkers in living subjects would represent a major advance. For example, CSFα-synuclein is so far considered the leading candidate as a single biomarker and has been tested the most extensively. However, published data on the CSF concentration of α-synuclein in patients with PD and controls have been contradictory.[5,6] A recent study clearly showed that CSF α-synuclein decreases in both PD and MSA patients and therefore cannot be used to differentiate these 2 diseases.[17] Similarly, the potential of other CSF proteins previously reported as differentiating biomarkers for parkinsonian disorders has not yet been validated.[5,6,8] A single biomarker may not be sufficient to differentiate PD and MSA by the targeted approach, possibly because of heterogeneity in each disease pathology and pathological overlaps between these 2 disorders.[18] Meanwhile, the combined assessment of multiple biomarkers has been shown to enhance the diagnostic power in neurodegenerative disorders.[19,20] Moreover, a diagnostic panel of multiple CSF proteins, including DJ-1, α-synuclein, Aβ peptides, and tau proteins was demonstrated to aid in Parkinson's disease diagnosis.[21] With these emerging multiple biomarker studies for parkinsonian disorders, we adopted an unbiased approach provided by full MALDI mass spectral profiles based on nontryptic CSF peptides/proteins.

Our results demonstrated a clear separation of PD, MSA, and controls from each other, with good values for cross-validation. In our results, PCA, an unsupervised leaning method to reduce data dimension, demonstrated that PD, MSA, and the control were clearly discriminated from each other by their proteomic profile distributions, and this discrimination was achieved not by a single or a few peaks, but by a combined set of many peaks, namely, the pattern unique to each disease condition. There was no significant association between the clinical variables such as age and sex. In our case, the 3 groups—PD, MSA, and the control—were age-matched as shown in Table 1. The PD group had a larger standard deviation of the ages than those for the MSA group, and we evaluated the PD group for age greater than 65 and age younger than 65 and then compared those 2 groups. However, age had no association with proteomic data, with a cross-validation rate of less than 50%. Sex had no association with proteomic data in each pair of these 3 groups, with a cross-validation rate of less than 50%. For the differential diagnosis of patients with PD and those with MSA, we made an SVM classification model based on a supervised machine-learning method with several features selected by using multiple peaks in the spectra obtained from the blinded test set. When we tested the second cohort of samples on this diagnostic panel, the discriminability of the differential diagnosis groups was clear, with reasonably high sensitivity and specificity, both of which were more than 80% with several features selected. Feature selection can be modulated, and for the current analysis, we adopted about 10–20 peaks for each model generation.

ClinProTools identified the peak at *m/z* 6250 that provided a satisfactory AUC value in the ROC analysis (0.956) only in discrimination of the early-stage patients with PD or MSA, but there were not enough values of this type in the discrimination of other pairs among PD, MSA, and control. From the results of this study, we emphasize that the proteomic pattern, not a single peak, could be a useful diagnostic panel for the differentiation of PD and MSA, even in the early stages. We do not know from which protein the peak at *m/z* 6250 was derived. Previously, Constantinescu et al reported a study of proteomic profiling of CSF in parkinsonian disorders by using SELDI-TOF MS and identified a fragment of chromogranin B, detected as the peak at *m/z* 6250 as a peptide to help differentiate patients with PD and MSA.[11] The peak intensity of the chromogranin B–derived fragment decreased in MSA patients compared with PD patients in their study. The chromogranins are widely distributed in neuroendocrine and nervous system tissues.[22] They suggested that the decrease in chromogranin B–derived fragments in MSA could be related to more aggressive synaptic or neuronal loss in patients with MSA than what is observed in PD.[11] For the moment, we will not explore this possibility any further because single-peak identification was not a top priority compared with testing reproducibility with a larger patient population and fine-tuning the comprehensive diagnostic panel based on these data.

In conclusion, although our results were derived from limited sample numbers, our study is the first to identify a promising application of proteomic pattern analysis to the clinical diagnosis of PD and MSA by profiling their respective CSF proteomes. Further studies are needed to confirm our current findings in larger cohorts of parkinsonian patients, especially to help diagnose disease progression and improve therapeutic efficacy. ■

Acknowledgments: We thank Drs. Yuichi Tokuda and Kengo Yoshii for useful discussion in the statistical analysis.

References

1. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. J Neurol Neurosurg Psychiatry 1992;55: 181–184.

2. Stefanova N, Bucke P, Duerr S, Wenning GK. Multiple system atrophy: an update. Lancet Neurol 2009;8:1172–1178.
3. Gilman S, Wenning GK, Low PA, et al. Second consensus statement on the diagnosis of multiple system atrophy. Neurology 2008;71:670–676.
4. Brooks DJ, Seppi K. Proposed neuroimaging criteria for the diagnosis of multiple system atrophy. Mov Disord 2009;24:949–964.
5. Tokuda T, Salem SA, Allsop D, et al. Decreased alpha-synuclein in cerebrospinal fluid of aged individuals and subjects with Parkinson's disease. Biochem Biophys Res Commun 2006;349:162–166.
6. Hong Z, Shi M, Chung KA, et al. DJ-1 and alpha-synuclein in human cerebrospinal fluid as biomarkers of Parkinson's disease. Brain 2010;133:713–726.
7. Abdo WF, De Jong D, Hendriks JC, et al. Cerebrospinal fluid analysis differentiates multiple system atrophy from Parkinson's disease. Mov Disord 2004;19:571–579.
8. Brettschneider J, Petzold A, Sussmuth SD, et al. Neurofilament heavy-chain NfH (SMI35) in cerebrospinal fluid supports the differential diagnosis of Parkinsonian syndromes. Mov Disord 2006; 21:2224–2227.
9. Abdi F, Quinn JF, Jankovic J, et al. Detection of biomarkers with a multiplex quantitative proteomic platform in cerebrospinal fluid of patients with neurodegenerative disorders,. J Alzheimers Dis 2006;9:293–348.
10. Zhang J, Sokal I, Peskind ER, et al. CSF multianalyte profile distinguishes Alzheimer and Parkinson diseases. Am J Clin Pathol 2008; 129:526–529.
11. Constantinescu R, Andreasson U, Li S, et al. Proteomic profiling of cerebrospinal fluid in parkinsonian disorders. Parkinsonism Relat Disord 2010;16:545–549.
12. Ketterlinus R, Hsieh SY, Teng SH, Lee H, Pusch W. Fishing for biomarkers: analyzing mass spectrometry data with the new ClinProTools software. Biotechniques 2005;Suppl:37–40.
13. Bosso N, Chinello C, Picozzi SC, et al. Human urine biomarkers of renal cell carcinoma evaluated by ClinProt. Proteomics Clin Appl 2008;2:1036–1046.
14. Komori M, Matsuyama Y, Nirasawa T, et al. Proteomic pattern analysis discriminates among multiple sclerosis-related disorders. Ann Neurol 2011;doi:10.1002/ana.22633.
15. Ivosev G, Burton L, Bonner R. Dimensionality reduction and visualization in principal component analysis. Anal Chem 2008;80: 4933–4944.
16. Sekiyama E, Matsuyama Y, Higo D, et al. Applying magnetic bead separation/MALDI-TOF mass spectrometry to human tear fluid proteome analysis. J Proteomics Bioinformatics 2008;1: 368–373.
17. Mollenhauer B, Locascio JJ, Schulz-Schaeffer W, Sixel-Doring F, Trenkwalder C, Schiossmacher MG. α-Synuclein and tau concentrations in cerebrospinal fluid of patients presenting with parkinsonism: a cohort study. Lancet Neurol 2011;10:230–240.
18. van Dijk KD, Teunissen CE, Drukarch B, et al. Diagnostic cerebrospinal fluid biomarkers for Parkinson's disease: a pathogenetically based approach. Neurobiol Dis 2010;39:229–241.
19. Fagan AM, Roe CM, Xiong C, Mintun MA, Morris JC, Holyzman DM. Cerebrospinal fluid tau/beta-amyloid(42) ratio as a prediction of cognitive decline in nondemented older adults. Arch Neurol 2007;64:343–349.
20. Shaw LM, Vanderstichele H, Knapik-Czajka M, et al. Alzheimer's Disease Neuroimaging Initiative. Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. Ann Neurol 2009;65:403–413.
21. Shi M, Bradner J, Hancock AM, et al. Cerebrospinal fluid biomarkers for Parkinson disease diagnosis and progression. Ann Neurol 2011;69:570–580.
22. Helle KB. The granin family of uniquely acidic proteins of the diffuse neuroendocrine system: comparative and functional aspects. Biol Rev Camb Philos Soc 2004;79:769–794.

SpringerPlus
a SpringerOpen Journal

RESEARCH Open Access

An approach to predict the risk of glaucoma development by integrating different attribute data

Yuichi Tokuda[1], Tomohito Yagi[1], Kengo Yoshii[1], Yoko Ikeda[2], Masahiro Fuwa[3], Morio Ueno[2], Masakazu Nakano[1], Natsue Omi[1], Masami Tanaka[1], Kazuhiko Mori[2], Masaaki Kageyama[3], Ikumitsu Nagasaki[4], Katsumi Yagi[5], Shigeru Kinoshita[2] and Kei Tashiro[1*]

Abstract

Primary open-angle glaucoma (POAG) is one of the major causes of blindness worldwide and considered to be influenced by inherited and environmental factors. Recently, we demonstrated a genome-wide association study for the susceptibility to POAG by comparing patients and controls. In addition, the serum cytokine levels, which are affected by environmental and postnatal factors, could be also obtained in patients as well as in controls, simultaneously. Here, in order to predict the effective diagnosis of POAG, we developed an "integration approach" using different attribute data which were integrated simply with several machine learning methods and random sampling. Two data sets were prepared for this study. The one is the "training data set", which consisted of 42 POAG and 42 controls. The other is the "test data set" consisted of 73 POAG and 52 controls. We first examined for genotype and cytokine data using the training data set with general machine learning methods. After the integration approach was applied, we obtained the stable accuracy, using the support vector machine method with the radial basis function. Although our approach was based on well-known machine learning methods and a simple process, we demonstrated that the integration with two kinds of attributes, genotype and cytokines, was effective and helpful in diagnostic prediction of POAG.

Keywords: *Glaucoma*, GWAS, Machine learning, Integration approach

Introduction

Glaucoma is a progressive eye disease that shows characteristic degeneration of the optic nerve and visual field defects (Kwon et al. 2009). Among the subtypes of glaucoma, primary open-angle glaucoma (POAG) is a major cause of blindness worldwide. The results of many studies have suggested that a genetic contribution is one of the risk factors for the development of glaucoma (Ray & Mookherjee 2009). However, it is still unclear if the genetic risk factors contribute to all of the pathogenesis of glaucoma. To investigate the mechanism(s) of common diseases such as glaucoma, genome-wide association studies (GWAS) have been widely performed (Consortium TWTCC 2007; Balding 2006). GWAS is one of the powerful tools to identify genetic association to common diseases with genotype data for single nucleotide polymorphisms (SNPs). Previously, we performed a GWAS to identify the common POAG-associated genetic factors (Nakano et al. 2009) and found a number of SNPs significantly associated with POAG. GWAS for POAG has also been performed by several other research groups (Meguro et al. 2010; Thorleifsson et al. 2010; Burdon et al. 2011), and we also recently published additional GWAS research results on POAG (Nakano et al. 2012). However, compared with the genetic risk for another type of glaucoma, Exfoliation Glaucoma (EG), which was carried out by deCODE using only two SNPs (http://www.decodehealth.com/glaucoma), genetic contribution for POAG seems to be a complex. In EG, SNPs were highly significant on a single gene, LOXL1, by GWAS (Thorleifsson et al. 2007; Williams et al. 2010; Mabuchi et al. 2008; Fan

* Correspondence: tashiro@koto.kpu-m.ac.jp
[1]Department of Genomic Medical Sciences, Kyoto Prefectural University of Medicine, Kajiicho 465, Kawaramachi-Hirokoji, Kamigyo-ku, Kyoto 602-8566, Japan
Full list of author information is available at the end of the article

et al. 2008), while in POAG, several genes are involved as genetic risk factors. In addition, besides the genetic factor, POAG is considered to have other risk factors (Kwon et al. 2009) as well. Thus, precise disease mechanism(s) of POAG remains elusive.

For the purposes of diagnostic prediction or finding out the pathogenesis of diseases, genotype data have been applied in several machine-learning algorithms (Relton et al. 2004; Listgarten et al. 2004; Ritchie et al. 2001; Nelson et al. 2001; Hoh et al. 2001; Wang et al. 2012). Genetic data and the other risk factors (e.g., smoking, body mass index) were combined for these prediction models (Seddon et al. 2009). In such studies, careful extraction of attributes for prediction from large volumes of data and appropriate data selection from several attributes are essential. As the development of common diseases like POAG is influenced by many factors, the contribution of each attribute weighs variously among the patients. Thus, for the diagnostic prediction of POAG, clarification of each attribute obtained for analysis needs to be carefully assessed. In this regard, it is important to develop a new strategy of integrating the data with various attributes for establishing useful diagnostic prediction.

In order to evaluate the risk factor of POAG, we integrated cytokine data together with genetic data as a new strategy. We focused on the serum cytokines because the relation between glaucomatous neurodegeneration and immune response was previously suggested (Tezel 2011), and several cytokines were reported to be linked with glaucoma (Huang et al. 2010; Yang et al. 2001). Cytokines, which include both chemokines and lymphokines, are small soluble proteins that play a pivotal role in immune system. The concentration of serum cytokines may reflect the physiological condition of the hosts affected by environmental and postnatal factors as one of the important indices useful for the diagnostic prediction of certain diseases. Obviously, cytokine data as an attribute weigh differently from those of the genotype data. In addition, the equipments that many cytokines can measure simultaneously under the same condition could have been developed and applied to diagnostic analysis (Ray et al. 2007; Lambeck et al. 2007). Therefore, we especially tried to measure and handle many cytokines simultaneously.

Here, for predicting the risk of POAG development, we attempted to establish a new integration approach with a good potential as a useful and simple tool. This procedure performs the integration of data with various kinds of attributes by using several machine learning methods with random sampling. In particular, because both genotyping and cytokines attributes were obtained from blood sample, our approach is considered to be useful for assessment of the risk of POAG and predicting the onset possibility before consulting ophthalmologists. This strategy may give us with new prototype for a clinical approach in understanding the underlying mechanism(s) of various diseases, not limited to POAG.

Methods

Sample Information

To obtain the peripheral blood samples, 115 POAG patients and 94 healthy control volunteers were recruited at the University Hospital of Kyoto Prefectural University of Medicine (Kyoto, Japan). This study was approved by the institutional review board of Kyoto Prefectural University of Medicine and conducted in accordance with the principles set forth in the Helsinki Declaration. All participants were interviewed about their familial history of glaucoma and other diseases and diagnosed either POAG or control by three ophthalmologists (YI, MU, and KM). The 115 POAG patients had peak intraocular pressure ≥ 22 mmHg without treatment. Peripheral blood samples were collected simultaneously from each participant for obtaining genomic DNA for genotyping and serum for cytokine measurement. DNA and sera were stored at −80°C until examined.

These samples were divided into two groups, since the cytokine data was obtained with two conditions. The first was defined as the "training data set" and the other as the "test data set" (Table 1). The former consisted of 42 POAG and 42 healthy control samples and was utilized in the training process of the machine learning. The latter consisted of 73 POAG and 52 healthy control samples, which were applied for the diagnostic prediction of POAG.

Genotype data

All genotype data were obtained by GeneChip® Human Mapping 500K Array platform (Affymetrix) according to the manufacturer's instructions. Although this array system carries the probes for more than five hundred thousand SNPs, we needed a number of SNPs significantly associated with POAG for the tests. Our previous study (Nakano et al. 2009) suggested that 40 SNPs were significantly POAG-associated which had both Mantel-Haenszel p-value of less than 0.01 and a p-value of Cochran's Q test (Ioannidis et al. 2007) equal to or more than 0.05 in the two stage GWAS. Because the pairs of SNPs showing high linkage disequilibrium (LD) could cause a multicollinearity problem, the Haploview program (Barrett et al. 2005) was applied to calculate LD. As a result, 11 of the 40 SNPs were excluded because of their high LD and remaining 29 SNPs were employed in this study (Table 2). All of the genotype data except for the missing by genotyping failure, which were

Table 1 Clinical characteristic of samples

	Training data set		Test data set	
	POAG	Control	POAG	Control
Number of sample	42	42	73	52
Famale / male ratio	1.00	0.83	0.62	1.74
Age at blood sampling	56.4±5.5	55.3±3.4	70.9±10.7	61.8 ± 11.3
Storage period of blood (days)	880.1±112.0	865.7±106.0	1044.0±114.4	892.2 ± 129.9

represented by a pair of letters (e.g., AA, AT and TT), were converted into discrete numerical values according to the number of allele with higher frequency in the POAG (i.e., risk allele) as followed: risk allele homozygote, 2; risk allele heterozygote, 1; and other allele homozygote, 0. Then, all the genotype data were normalized using the equations in EIGENSTRAT (Price et al. 2006), so that the missing data were set to 0.0. According to the allele frequency and the average of numeric genotypes calculated from the training data set, this normalization was carried out and the normalized data represented discrete values.

Table 2 Summary of 29 SNPs used in this study

dbSNP ID	Chr.	SNP type	Nearest gene	Genotype frequency
rs547984	1	intergenic	ZP4	AA(0.263) AC(0.488) CC(0.249)
rs1892116	1	intronic	AHCTF1	AA(0.507) AG(0.445) GG(0.048)
rs4666488	2	intergenic	OSR1	AA(0.100) AG(0.397) GG(0.503)
rs2268794	2	intronic	SRD5A2	AA(0.005) AT(0.319) TT(0.676)
rs7574012	2	intergenic	QPCT	AA(0.373) AG(0.459) GG(0.168)
rs1990702	2	intergenic	LRP2	GG(0.120) GA(0.433) AA(0.447)
rs10930437	2	intergenic	SP5	AA(0.429) AG(0.454) GG(0.117)
rs779701	3	intronic	GRM7	AA(0.490) AG(0.413) GG(0.097)
rs6550783	3	intergenic	UBE2E1	AA(0.412) AG(0.442) GG(0.146)
rs6550308	3	intergenic	ARPP21	GG(0.215) GA(0.488) AA(0.297)
rs3922704	3	intronic	PLCXD2	CC(0.034) CG(0.254) GG(0.712)
rs17279573	4	intergenic	KIAA0922	GG(0.120) GA(0.483) AA(0.397)
rs818725	5	intronic	ADAMTS12	CC(0.019) CG(0.226) GG(0.755)
rs11750584	5	intergenic	HEATR7B2	CC(0.029) CG(0.292) GG(0.679)
rs9640055	7	intronic	GLCCI1	GG(0.038) GA(0.344) AA(0.618)
rs2966712	7	intergenic	LOC285965	AA(0.005) AG(0.211) GG(0.784)
rs411102	9	intergenic	KRT8P11	GG(0.749) GA(0.242) AA(0.009)
rs7850541	9	intergenic	GBGT1	GG(0.514) GA(0.361) AA(0.125)
rs7081455	10	intergenic	PLXDC2	AA(0.644) AC(0.293) CC(0.063)
rs493622	11	intergenic	CHORDC1	AA(0.565) AC(0.383) CC(0.052)
rs610160	11	intronic	GRIA4	AA(0.693) AG(0.262) GG(0.045)
rs7961953	12	intronic	TMTC2	GG(0.522) GA(0.397) AA(0.081)
rs10492680	13	intergenic	FLJ42392	GG(0.005) GA(0.187) AA(0.808)
rs1571379	14	intergenic	SEL1L	AA(0.440) AG(0.454) GG(0.106)
rs9788983	17	intronic	RPH3AL	AA(0.770) AG(0.215) GG(0.015)
rs16940484	18	intronic	TTC39C	GG(0.469) GA(0.450) AA(0.081)
rs2864107	19	intergenic	ZNF175	GG(0.684) GA(0.301) AA(0.015)
rs6115865	20	intergenic	C20orf194	AA(0.125) AG(0.428) GG(0.447)
rs5765558	22	intergenic	ATXN10	AA(0.287) AG(0.478) GG(0.235)

The dbSNP ID represents with build 130. Chr. denotes the number of chromosome. The Nearest genes are positioned nearest by each SNP and referred to NCBI Build 36. Genotype frequencies are calculated by total samples used in this study, which are 115 POAG patients and 94 healthy control volunteers.

Cytokine data

Serum cytokines were measured by the bead flow-cytometry analysis by the Becton Dickinson (BD, San Diego, CA) Cytometric Bead Array (CBA™) Flex Set System according to the manufacturer's protocol. The data was examined by a BD FACSArray™ (BD) flow cytometer with FCAP Array™ software and the BD FACSArray™ Bioanalyzer (BD).

In this study, we first assayed 29 cytokines in the sera from "the training data set", and each cytokine concentration was calculated from each raw data by the Four Parameter Logistic Model (FPLM), which was recommended by the manufacturer (http://www.bdbiosciences.com/documents/Analysis_of_data_from_CBA_using_FCAPArray.pdf). Before we performed the statistical analysis, the quality of the cytokine data was evaluated. Of 29 cytokines, 21 cytokines were excluded; 7 were for measurement failures (over 5% of the 84 samples) and 14 for concentration of zero (over 5% of the 84 samples). The remaining 8 cytokines were tested by the Student's *t*-test between the POAG and control samples, of which 5 cytokines were excluded with a p-value over 5%. Eventually, only 3 cytokines, i.e., Fas Ligand, Eotaxin, and MIG, were picked up to be significantly associated with POAG from the training data set samples (Table 3).

Subsequently, these 3 cytokines were determined with the same assay procedure on 126 samples (73 POAG and 53 controls) from the "test data set" samples. Data were obtained from 125 samples, excluding one control sample of failed assay (Table 3). For statistical analysis, the cytokine concentration data were standardized in order to minimize the biases among the assay conditions as followed. Let c_{ij} be the cytokine concentration measured for cytokine i and sample j, where i = 1 to 3 and j = 1 to M (M is 84 in the training data set; 125 in the test data set). Let m_i and s_i be the mean and standard deviation of cytokine i, respectively. At each data set, m_i and s_i were calculated only for the control samples because it was considered that the cytokine concentration of healthy control samples might act fairly consistently under each experimental condition. The standardized value n_{ij} was calculated using the following equation: $n_{ij} = (c_{ij} - m_i)/s_i$. Notably, the cytokine concentration data was obtained as continuous values when they were calculated by FPLM.

Finally, results of a total of 32 attributes, which consisted of 29 SNPs (Table 2) and 3 cytokines (Table 3), were applied for "integration approach" in this study.

Base classifiers

In this study, well-known machine learning methods, i.e., Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Naive Bayes Classifier (NBC), and Decision Tree (DT) were applied. We defined these methods as "base classifiers".

LDA is a method used in statistics and machine learning to find a discriminant function by which two or more groups can be separated. LDA seeks a linear function of the variables (e.g., genotype and cytokine) in the training data set that maximizes the distance among means in each group as it minimizes the within-group variance. Hence, a discriminant function can be computed explicitly and used as a linear classifier.

SVM is a supervised machine learning method based on the idea of classifying two groups by a hyperplane with a large margin. SVM maps the data in the training data set into a possibly higher dimension of space by using a kernel function. In the space, SVM learns the classifier by seeking a hyperplane that may separate the two groups by a certain distance. If the training data set is not separated linearly, SVM optimizes the separation between the two groups. The kernel function in SVM is decided according to the attribute of the data. In this study, we used SVM for learning with three kernel functions: linear, polynomial, and radial basis function (RBF).

NBC is a simple and efficient probabilistic classifier based on Bayes' theorem. Assuming there is independence between each set of attribute data (e.g., genotype or cytokine); NBC calculates the probabilities used for the prediction from the training data set. As each sample in the test data set is given to the NBC, it predicts to which

Table 3 Summary of the three cytokines used in the integration approach

Cytokine		Training data set		Test data set	
		Concentration	P-value*	Concentration	P-value*
Fas Ligand	POAG	63.5 (52.2-87.3)	0.002	37.5 (31.8-46.6)	0.877
	Control	53.3 (34.9-63.4)		36.2 (28.0-45.4)	
Eotaxin	POAG	309.1 (273.6-342.9)	0.038	70.6 (54.9-90.8)	0.013
	Control	268.5 (236.7-311.6)		63.5 (54.4-73.9)	
MIG	POAG	410.9 (306.8-524.9)	0.021	318.1 (182.9-511.7)	0.109
	Control	340.4 (198.9-470.1)		148.4 (117.7-241.9)	

"Concentration" represents the median concentration and interquartile range. * P-value of the comparison between POAG and control calculated by Student's *t*-test.

group (e.g., POAG or control) the sample belongs by the highest conditional probability.

DT is a tree-like data structure used for learning a method to classify data hierarchically by sequential decision process. Basically, DT is a binary tree and each node splits the data by each feature (i.e., large/small, male/female). In this study, DT was performed by CART (Classification and Regression Trees), and used to classify SNPs (each data consisted four discrete; three genotypes and missing data) and cytokines (each data was continuous).

All the data analysis and drawing figures were performed with R software (version 2.14.0) (R Development Core Team 2011); the LDA was implemented by the MASS (version 7.3-16) R package; the SVM and NBC functions were implemented by the e1071 (version 1.6) R package (Dimitriadou et al. 2011); and the DT functions were implemented by the mvpart (version 1.4-0) R package. In addition, each classifier was performed with default parameter settings.

Accuracy, sensitivity and specificity of the data (genotype and cytokine) for the POAG prediction were calculated by these analytical procedures.

Integration approach

In this study, the data consists of two kinds of attributes in that the genotype data are discrete and the cytokine data are continuous. In most cases, it is easy and no problem to apply these data for each method simply and simultaneously. However, one must be careful to integrate them while considering each attribute, especially to note how each attribute contributes. The prediction may be made possible from analytical results for each type of attribute data instead of applying the data directly, because of the difference in the attributes. In addition, if the analytical results show differences between each attribute, the prediction for each sample has interesting information how each attribute contributes. For these reasons, we performed the integration approach so that after the genotype and cytokine data are separately applied in the processes, their results are integrated after the last process. To enable an effective analysis by integrating these two kinds of data, this approach is based on the idea of ensemble learning (e.g., Bootstrap aggregating (Bagging) (Breiman 1996)). Bagging is one of the powerful prediction tools for improving other basic classifier. For example, bagging is used for the purpose of improving the diagnosis of Valvular Heart Disease by SVM (Sengur 2012), or assessing the interactions of SNPs (Schwender et al. 2011).

For the training data set L consisted of cases $(l_1^p,\ldots,l_p^p)$ and controls $(l_1^c,\ldots,l_q^c)$ and the test data set $T = \{t_1,\ldots,t_r\}$, the integration approach consists of the following steps:

1) Obtain Sg, which is the subset of the training data set, by random sampling without replacement from L so that the same number of samplings is taken from the cases as from the controls.
2) Apply the base classifiers to the genotype data of Sg to obtain a predictor Pg as a training result.
3) Repeat above steps (1) and (2) K times; this process produces genotype data predictors $\{Pg_1,\ldots,Pg_K\}$ from $\{Sg_1,\ldots,Sg_K\}$.
4) In addition, repeat the same process as in (1) and (2) above N times for cytokine data; cytokine data predictors $\{Pc_1,\ldots,Pc_N\}$ are produced from the subset of the training data set $\{Sc_1,\ldots,Sc_N\}$.
5) For each t_j in the test data T, the predictor gives a result which predicts whether t_j belongs to the cases (positive) or the controls (negative). Thus for each t_j in the test data T, the genotype data predictors $\{Pg_1,\ldots,Pg_K\}$ produce K prediction results $\{Rg_1,\ldots,Rg_K\}$ and the cytokine data predictors $\{Pc_1,\ldots,Pc_N\}$ produce N prediction results $\{Rc_1,\ldots,Rc_N\}$.
6) For each t_j in the test data T, the majority vote of the $N + K$ prediction results is the final prediction for t_j.

This procedure adopted the same number of samplings, for example, 20 POAG and 20 healthy controls were sampled from 42 POAG and 42 healthy controls in the training data set, respectively. This reason is that the contribution of the characteristics of POAG and control should be as close to equal possible. Besides, it is preferable for the genotype and cytokine data to be evaluated as equally as possible (e.g., $K = N$.) However, it may be impossible to predict one group by dividing it in half if the total number of sampling repeats is an even number. In this study, since the size of the genotype data set was greater than that of the cytokines, K is taken as $N + 1$ to avoid the situation of a tie vote. In addition, note that use of the base classifier should be limited to one kind of classifier from the beginning of this procedure to the end.

Results

Single classifier analysis

Single classifier analysis was performed for each base classifier on 29 SNPs and 3 cytokines each and both integrated (Table 4). All of these tests were first done by the training data set and evaluated to predict the test data set. Except for DT, the accuracy of genotype data prediction was higher than that of cytokines for each base classifier. The integrated accuracy was better than each base classifier, when tested with use of the polynomial SVM, RBF SVM, and NBC. However, the integrated sensitivity (0.521) was lower than the genotype (0.589) or cytokine (0.658) prediction alone, when tested by polynomial SVM, in spite of increasing the integrated specificity (0.846) from the genotype (0.731) or cytokine (0.308) prediction alone. By contrast, RBF SVM test

Table 4 Summary of the three cytokines used in the integration approach

Base classifier			Single analysis			Analysis with sampling*		
			Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
LDA		Genotype	0.688	0.712	0.654	0.671 ± 0.011	0.693 ± 0.015	0.639 ± 0.014
		Cytokine	0.592	0.466	0.769	0.584 ± 0.010	0.457 ± 0.012	0.763 ± 0.010
		Integrated	0.632	0.616	0.654	0.655 ± 0.022	0.611 ± 0.034	0.717 ± 0.015
SVM	linear	Genotype	0.664	0.699	0.615	0.683 ± 0.013	0.754 ± 0.023	0.584 ± 0.016
		Cytokine	0.568	0.452	0.731	0.577 ± 0.008	0.458 ± 0.012	0.745 ± 0.013
		Integrated	0.659	0.648	0.673	0.668 ± 0.014	0.640 ± 0.024	0.706 ± 0.012
	polynomial	Genotype	0.648	0.589	0.731	0.633 ± 0.010	0.539 ± 0.026	0.764 ± 0.018
		Cytokine	0.512	0.658	0.308	0.457 ± 0.012	0.275 ± 0.077	0.713 ± 0.086
		Integrated	0.656	0.521	0.846	0.624 ± 0.010	0.480 ± 0.065	0.827 ± 0.078
	RBF	Genotype	0.688	0.712	0.654	0.676 ± 0.010	0.685 ± 0.016	0.664 ± 0.013
		Cytokine	0.648	0.712	0.558	0.662 ± 0.006	0.701 ± 0.011	0.607 ± 0.020
		Integrated	0.744	0.767	0.712	0.740 ± 0.013	0.805 ± 0.020	0.650 ± 0.014
NBC		Genotype	0.640	0.671	0.596	0.630 ± 0.006	0.651 ± 0.013	0.601 ± 0.014
		Cytokine	0.624	0.479	0.827	0.621 ± 0.006	0.489 ± 0.013	0.807 ± 0.019
		Integrated	0.744	0.767	0.712	0.698 ± 0.013	0.644 ± 0.027	0.775 ± 0.051
DT		Genotype	0.536	0.342	0.808	0.562 ± 0.025	0.411 ± 0.070	0.774 ± 0.043
		Cytokine	0.624	0.904	0.231	0.605 ± 0.018	0.874 ± 0.099	0.226 ± 0.126
		Integrated	0.600	0.959	0.096	0.617 ± 0.013	0.668 ± 0.032	0.545 ± 0.040

*These values are represented as the mean and SD of each statistics. The mean of each statistics included extremely good or bad result, especially small sampling size and few sampling repeat time.

increased all of the accuracy (0.744), sensitivity (0.767) and specificity (0.712) on the integrated data from either genotype or cytokine prediction. These results suggested that both genotype and cytokine attributes contributed, especially when integrated, to improve the diagnostic prediction based on the base classifier.

Integration approach analysis

The results of single use with base classifier demonstrated fluctuations on each or both applying attribute (Table 4; Single analysis). Therefore, the further integrated approach was performed using each base classifier by changing the size and time of parameters (Table 4; Analysis with sampling). One of the changed parameters was the size of the subset sampling from the training data set (defined as "sampling size"), and the other was the sampling repeat times (defined as "sampling time"). The sampling size was increased from 40 (consisted of 20 POAG and 20 healthy controls) to 80 (consisted of 40 POAG and 40 healthy controls) with an equal number of samples from POAG and controls. (i.e., 21 steps were tested) On the other hand, the sampling time for each genotype and cytokine was also increased from 25 to 1,500 by 60 steps. (i.e., 25, 50, 75, ···, 1,450, 1,475 and 1,500 repeat times were tested) Moreover, because the sampling time for the genotype data was increased by one, the total sampling repeat times increased from 51 to 3,001. As a result, the integration approach was performed on 1,260 tests (21 steps of sampling sizes × 60 steps of sampling times) per each base classifier.

These results are summarized in "Analysis with sampling" in Table 4. The LDA, Linear SVM, and DT methods improved the mean of integrated accuracy from single analysis (from 0.632 to 0.655, from 0.659 to 0.668, and from 0.600 to 0.617, respectively), although those values included fluctuations due to parameter settings. The mean of the integrated accuracy (0.740 ± 0.013; mean ± SD) assessed by the RBF SVM method was the best results in analysis with sampling, however, it was slightly lower than that in single analysis in association with the higher integrated sensitivity (0.805 ± 0.020) than that in single analysis (0.767). Moreover, the specificities of genotype (0.664 ± 0.013) and cytokine (0.607 ± 0.020) by SVM RBF method in analysis with sampling were better than those in single analysis (0.654 and 0.558, respectively). In addition, some accuracy in the 1,260 tests was achieved over the single analysis.

In order to understand how the test results improved by changing the sampling size and time of parameters and each attribute contributed to the prediction, the integration results were demonstrated graphically (Figure 1). The schematic presentations of the genotype and cytokine data were plotted on horizontal and vertical axes, respectively, as

shown in Figure 1a. One example of the unstable results was shown in Figure 1b. Because those parameters were comparatively smaller, the positive ratios of each attribute were generally unsatisfactory with several samples being plotted in the vicinity of the diagonal threshold. By contrast, when the sampling size was 70 (consisted of 35 POAG and 35 healthy controls) and sampling times was 2,001 (1,001 times at genotype data and 1,000 times at cytokine data), most of the samples were plotted in the vicinity of the axes (Figure 1c). Using these parameters, the accuracy was improved for 0.768. This result was also obtained by many other conditions when the sampling size and time were comparatively larger; therefore it was considered as the best stable results of the integration approach. Thus, the predictions were improved by changing the size and time of parameters in either the genotype or cytokine test.

In these test plot presentations, we focused on the contribution of the genotype and cytokine data to the

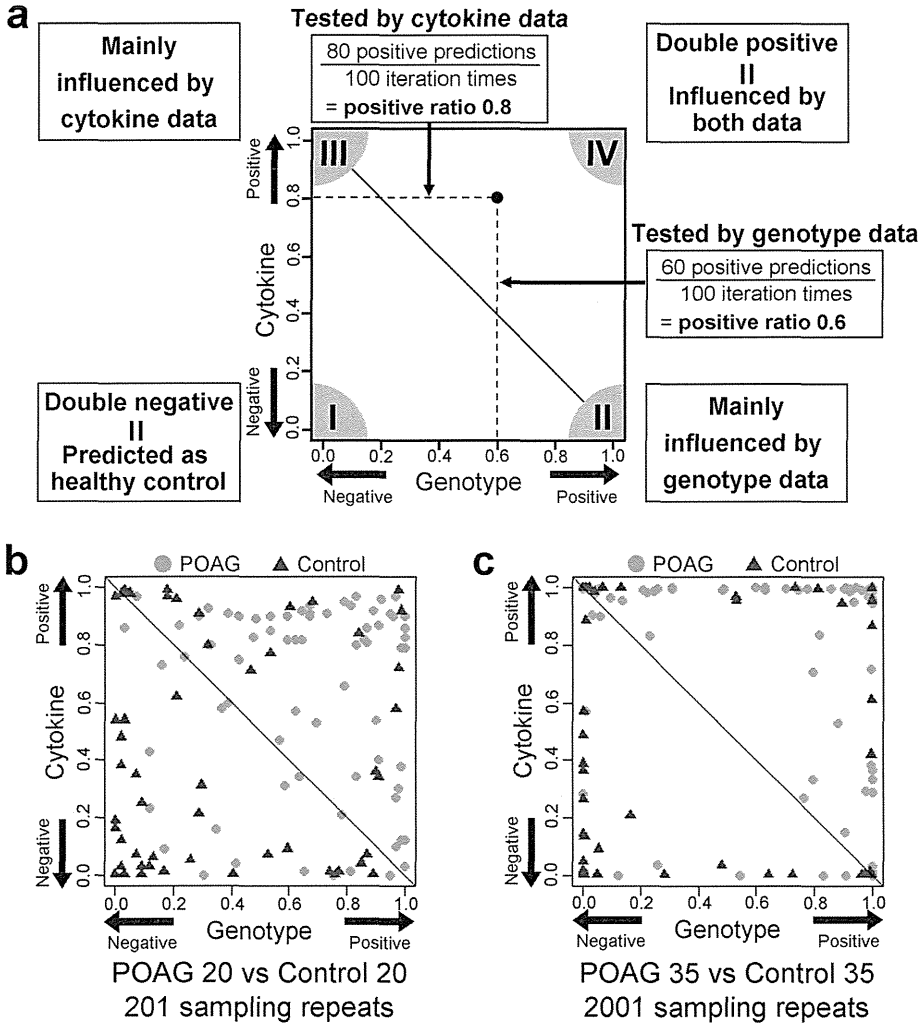

Figure 1 Scatter plot showing the ratio of POAG prediction for each sample. Figure 1 (**a**) The example figure for the scatter plot. The horizontal axis represents the ratio of positive prediction using genotype data. The positive prediction indicated the sample with POAG feature, and the negative prediction indicated the sample with control feature. The ratio was obtained by dividing the number of positive predictions by the total test number. Thus, "1" and "0" indicate 100% prediction as positive and negative, respectively. The vertical axis similarly represents the ratio using the cytokine data. Dots and triangles represent POAG and control samples, respectively. The figure can be read as, if one POAG sample was predicted as positive 60 times using the genotype data and 80 times using the cytokine data each with 100 sampling repeat times, the sample is plotted at (0.6, 0.8) by dot. If the approach has a good performance (means; highly negative or positive prediction) for samples with interaction between those two attributes, more samples will be plotted in the corner I or corner IV. If either the genotype or cytokine data is at risk for POAG, such samples will be plotted in the corner II or corner III, respectively. The diagonal line shows the threshold of the prediction by the integration approach. If a sample is plotted above or below the threshold, the final prediction result is positive or negative, respectively. Figure 1 (**b**) shows one of the examples as the comparatively smaller and unstable, which is the result with 40 sampling size and 201sampling times by RBF SVM method. Figure 1 (**c**), one of the examples as the best stable result, which is the result with 70 sampling size and 2,001sampling times by RBF SVM method.

stable results among the POAG samples, 23 (31.5%) showed more than 90% accuracy for both positive ratios (i.e., plotted in the corner IV in Figure 1c). On the other hand, 14 (26.9%) of the control samples showed more than 90% accuracy (i.e., plotted in the corner I in Figure 1c).

Discussion

Bootstrap methods, such as Bagging (Breiman 1996), are generally applied in approaches using random sampling techniques. In a typical procedure, bootstrap can provide us with an estimated distribution for statistical analysis by random sampling with replacement from all samples in the data set. In this study, the method of random sampling was independent for each group, and an equal number of samples were adopted in order to avoid bias by the difference in sample numbers among each group. Additionally, our approach adopted random sampling without replacement due to the potential for multicollinearity. Because genotype data show discrete values consisted of three genotypes and one missing data, the combinations of values were easy to be limited as much as causing multicollinearity. Especially, this phenomenon was apparent when LDA method was applied with the small sampling size. For this reason, the changing parameters of the sampling size were started with 40 samples by random sampling without replacement. Besides, the accuracy did not improve without any relation to the iteration times even when the sampling size was increased enough as showed in Figure 1c. This tendency was considered to be caused by highly correlated samples. To solve this problem, it might be better to adopt the data for random sampling with replacement than without replacement according to the size of the training data set.

Using genotype data, the diagnostic prediction of POAG by RBF SVM method generally performed well also in our study (Ban et al. 2010; Rojas et al. 2009). The applied 29 SNPs were selected by the statistical result of GWAS from enormous genotype data. Employment of the SNPs selected by some large size of population was useful for this type of diagnostic prediction study without complex procedures. Thus, simple strategy might be suitable for the post GWAS analysis. The bagging is generally considered to reduce variance of classifier such as DT method; therefore, the classifier with less variant such as SVM method was considered to be improved a little by bagging. However the result of our study was effective even when SVM, DT methods with bagging was not improved.

Using cytokine data, the diagnostic prediction of POAG by RBF SVM method also performed well, regardless of some fluctuation between two data sets. Thus, RBF SVM method was thought to be successfully suitable for each attribute data, genotype as well as cytokine, in our study. In other words, the base classifier is necessary to select suitably according to each attribute. However, the effectiveness of cytokine data analysis using SVM has been reported for selecting the significant cytokines to elucidate the pathway of inflammatory response (McKinney et al. 2006).

In this study, we found 3 cytokines that are associated with POAG in 29 cytokines. In our approach, some samples was certainly predicted by only cytokine attributes as shown in Figure 1b or c. These results demonstrated that POAG patients with low genetic risk were predicted by cytokine attributes effectively.

In terms of the integration approach, one of our goals is to predict the diagnosis and/or prognosis by the patterning of different types of experimental data. In the process, an interaction between genotype and cytokine might indicate a risk of disease development, because approximately 30% of the samples in the test data set were performed with a high prediction from both types of data. Our approach also elicited a good classification of same sample when one of the two data sets was used individually before integrating them. The classification was made successful by using one data set because either genotype or cytokine behaved as a risk of disease development in these samples. For such reasons, our approach is considered to be one of the good tools to analyze the mixed data, irrespective of their interaction.

In conclusion, we demonstrated that our integration approach improved the diagnostic prediction of POAG with use of two attributes, SNPs as genotype and serum cytokines. Although two attribute data are applied independently, this approach is not affected by the differences of attribute, because the base classifier was first set according to each type of attribute data. It was confirmed that when the setting of the base classifier for one data set is successfully optimized, the integration approach might be applied using additional data with other attributes. In view of the versatility and simplicity, our approach was thought to be effective and useful for various clinical applications in future.

Competing interests

The authors declare that they have no conflict of interest.

Authors' contributions

KM, MK, IN, SK, and KT designed the research. YI, MU, KM, SK and KT recruited POAG patients and healthy volunteers. YI, MU and KM performed their clinical diagnosis, and collected and managed blood samples with NO. NO and MT processed blood samples and prepared DNA samples. MN, NO and MT analyzed and processed the genotyping data. TY, MF and MT measured and analyzed the cytokine data. YT, TY and KYoshii preprocessed and evaluated the genotype and cytokine data. YT and KYoshii developed and improved the integration approach and base classifiers used in it. IN and KYagi helped evaluate the integration approach. YT and TY drafted the manuscript. All the authors read and approved the final manuscript.

Acknowledgements

We appreciate all the patients and volunteers enrolled in our study. We also thank Ms. Sayaka Ohashi, Naoko Saito, Hiroko Adachi, Yumi Yamashita, and Yuko Konoshima for processing blood samples and performing genotyping;

Mrs. Hiromi Yamada, Ms. Aiko Hashimoto, Ms. Keiko Nirasawa, and Mrs. Akemi Tanaka for assisting with the clinical information analysis; Mr. Ryuichi Sato and Ms. Fumiko Sato (SASA Plus Co., Ltd., Fukuoka, Japan) for the management of genotyping data; and Ms. Tomoko Ichikawa for excellent secretarial assistance. This work was supported by grants from Collaborative Development of Innovative Seeds of Japan Science and Technology Agency (JST) to MK and KT, and Researches on Sensory and Communicative Disorders from the Ministry of Health, Labour and Welfare in Japan to KM, IN, SK, and KT.

Author details
[1]Department of Genomic Medical Sciences, Kyoto Prefectural University of Medicine, Kajiicho 465, Kawaramachi-Hirokoji, Kamigyo-ku, Kyoto 602-8566, Japan. [2]Department of Ophthalmology, Kyoto Prefectural University of Medicine, Kajiicho 465, Kawaramachi-Hirokoji, Kamigyo-ku, Kyoto 602-8566, Japan. [3]Research and Development Center, Santen Pharmaceutical Co. Ltd, 8916-16 Takayama-cho, Ikoma, Nara 630-0101, Japan. [4]Department of Mathematics, Kyoto Prefectural University of Medicine, Kajiicho 465, Kawaramachi-Hirokoji, Kamigyo-ku, Kyoto 602-8566, Japan. [5]Louis Pasteur Center for Medical Research, 103-5, Tanakamonzen-cho, Sakyo-ku, Kyoto City, Kyoto 606-8225, Japan.

Received: 29 June 2012 Accepted: 15 October 2012
Published: 24 October 2012

References

Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7(10):781–791. doi:nrg1916 [pii] 10.1038/nrg1916

Ban HJ, Heo JY, Oh KS, Park KJ (2010) Identification of type 2 diabetes-associated combination of SNPs using support vector machine. BMC Genet 11:26. doi:1471-2156-11-26 [pii] 10.1186/1471-2156-11-26

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263–265. doi:10.1093/bioinformatics/bth457 bth457 [pii]

Breiman L (1996) Bagging Predictors. Mach Learn 24(2):123–140. doi:10.1023/a:1018054314350

Burdon KP, Macgregor S, Hewitt AW, Sharma S, Chidlow G, Mills RA, Danoy P, Casson R, Viswanathan AC, Liu JZ, Landers J, Henders AK, Wood J, Souzeau E, Crawford A, Leo P, Wang JJ, Rochtchina E, Nyholt DR, Martin NG, Montgomery GW, Mitchell P, Brown MA, Mackey DA, Craig JE (2011) Genome-wide association study identifies susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. Nat Genet 43(6):574–578. doi:ng.824 [pii] 10.1038/ng.824

Consortium TWTCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661–678. doi:nature05911 [pii] 10.1038/nature05911

Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A (2011) e1071: Misc Functions of the Department of Statistics (e1071). TU Wien

Fan BJ, Pasquale L, Grosskreutz CL, Rhee D, Chen T, DeAngelis MM, Kim I, del Bono E, Miller JW, Li T, Haines JL, Wiggs JL (2008) DNA sequence variants in the LOXL1 gene are associated with pseudoexfoliation glaucoma in a U.S. clinic-based population with broad ethnic diversity. BMC Med Genet 9:5. doi:1471-2350-9-5 [pii] 10.1186/1471-2350-9-5

Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case–control association studies. Genome Res 11(12):2115–2119. doi:10.1101/gr.204001

Huang P, Qi Y, Xu YS, Liu J, Liao D, Zhang SS, Zhang C (2010) Serum cytokine alteration is associated with optic neuropathy in human primary open angle glaucoma. J Glaucoma 19(5):324–330. doi:10.1097/IJG.0b013e3181b4cac7

Ioannidis JP, Patsopoulos NA, Evangelou E (2007) Heterogeneity in meta-analyses of genome-wide association investigations. PLoS One 2(9):e841. doi:10.1371/journal.pone.0000841

Kwon YH, Fingert JH, Kuehn MH, Alward WL (2009) Primary open-angle glaucoma. N Engl J Med 360(11):1113–1124. doi:360/11/1113 [pii] 10.1056/NEJMra0804630

Lambeck AJ, Crijns AP, Leffers N, Sluiter WJ, ten Hoor KA, Braid M, van der Zee AG, Daemen T, Nijman HW, Kast WM (2007) Serum cytokine profiling as a diagnostic and prognostic tool in ovarian cancer: a potential role for interleukin 7. Clin Cancer Res 13(8):2385–2391. doi:13/8/2385 [pii] 10.1158/1078-0432.CCR-06-1828

Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A, Mackey J, Wishart D, Greiner R, Zanke B (2004) Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clin Cancer Res 10(8):2725–2737

Mabuchi F, Sakurada Y, Kashiwagi K, Yamagata Z, Iijima H, Tsukahara S (2008) Lysyl oxidase-like 1 gene polymorphisms in Japanese patients with primary open angle glaucoma and exfoliation syndrome. Mol Vis 14:1303–1308

McKinney BA, Reif DM, Rock MT, Edwards KM, Kingsmore SF, Moore JH, Crowe JE Jr (2006) Cytokine expression patterns associated with systemic adverse events following smallpox immunization. J Infect Dis 194(4):444–453. doi:JID36092 [pii] 10.1086/505503

Meguro A, Inoko H, Ota M, Mizuki N, Bahram S (2010) Genome-wide association study of normal tension glaucoma: common variants in SRBD1 and ELOVL5 contribute to disease susceptibility. Ophthalmology 117(7):1331–1338. doi: S0161-6420(09)01404-3 [pii] 10.1016/j.ophtha.2009.12.001

Nakano M, Ikeda Y, Taniguchi T, Yagi T, Fuwa M, Omi N, Tokuda Y, Tanaka M, Yoshii K, Kageyama M, Naruse S, Matsuda A, Mori K, Kinoshita S, Tashiro K (2009) Three susceptible loci associated with primary open-angle glaucoma identified by genome-wide association study in a Japanese population. Proc Natl Acad Sci U S A 106(31):12838–12842. doi:0906397106 [pii] 10.1073/pnas.0906397106

Nakano M, Ikeda Y, Tokuda Y, Fuwa M, Omi N, Ueno M, Imai K, Adachi H, Kageyama M, Mori K, Kinoshita S, Tashiro K (2012) Common Variants in CDKN2B-AS1 Associated with Optic-Nerve Vulnerability of Glaucoma Identified by Genome-Wide Association Studies in Japanese. PLoS One 7(3): e33389. doi:10.1371/journal.pone.0033389 PONE-D-11-17292 [pii]

Nelson MR, Kardia SL, Ferrell RE, Sing CF (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 11(3):458–470. doi:10.1101/gr.172901

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909. doi:ng1847 [pii] 10.1038/ng1847

R Development Core Team (2011) R: A Language and Environment for Statistical Computing

Ray K, Mookherjee S (2009) Molecular complexity of primary open angle glaucoma: current concepts. J Genet 88(4):451–467

Ray S, Britschgi M, Herbert C, Takeda-Uchimura Y, Boxer A, Blennow K, Friedman LF, Galasko DR, Jutel M, Karydas A, Kaye JA, Leszek J, Miller BL, Minthon L, Quinn JF, Rabinovici GD, Robinson WH, Sabbagh MN, So YT, Sparks DL, Tabaton M, Tinklenberg J, Yesavage JA, Tibshirani R, Wyss-Coray T (2007) Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. Nat Med 13(11):1359–1362. doi:nm1653 [pii] 10.1038/nm1653

Relton CL, Wilding CS, Pearce MS, Laffling AJ, Jonas PA, Lynch SA, Tawn EJ, Burn J (2004) Gene-gene interaction in folate-related genes and risk of neural tube defects in a UK population. J Med Genet 41(4):256–260

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69(1):138–147. doi:S0002-9297(07)61453-0 [pii] 10.1086/321276

Rojas J, Fernandez I, Pastor JC, Garcia-Gutierrez MT, Sanabria RM, Brion M, Sobrino B, Manzanas L, Giraldo A, Rodriguez-de la Rua E, Carracedo A (2009) Development of predictive models of proliferative vitreoretinopathy based on genetic variables: the Retina 4 project. Invest Ophthalmol Vis Sci 50(5):2384–2390. doi: iovs.08-2670 [pii] 10.1167/iovs.08-2670

Seddon JM, Reynolds R, Maller J, Fagerness JA, Daly MJ, Rosner B (2009) Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. Invest Ophthalmol Vis Sci 50(5):2044–2053. doi:iovs.08-3064 [pii] 10.1167/iovs.08-3064

Sengur A (2012) Support vector machine ensembles for intelligent diagnosis of valvular heart disease. J Med Syst 36(4):2649–2655. doi:10.1007/s10916-011-9740-z

Schwender H, Bowers K, Fallin MD, Ruczinski I (2011) Importance measures for epistatic interactions in case-parent trios. Ann Hum Genet 75(1):122–132. doi:10.1111/j.1469-1809.2010.00623.x

Tezel G (2011) The immune response in glaucoma: a perspective on the roles of oxidative stress. Exp Eye Res 93(2):178–186. doi:S0014-4835(10)00220-4 [pii] 10.1016/j.exer.2010.07.009

Thorleifsson G, Magnusson KP, Sulem P, Walters GB, Gudbjartsson DF, Stefansson H, Jonsson T, Jonasdottir A, Stefansdottir G, Masson G, Hardarson GA, Petursson H, Arnarsson A, Motallebipour M, Wallerman O, Wadelius C, Gulcher JR, Thorsteinsdottir U, Kong A, Jonasson F, Stefansson K (2007) Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. Science 317(5843):1397–1400. doi:1146554 [pii] 10.1126/science.1146554

Thorleifsson G, Walters GB, Hewitt AW, Masson G, Helgason A, DeWan A, Sigurdsson A, Jonasdottir A, Gudjonsson SA, Magnusson KP, Stefansson H, Lam DS, Tam PO, Gudmundsdottir GJ, Southgate L, Burdon KP, Gottfredsdottir MS, Aldred MA, Mitchell P, St Clair D, Collier DA, Tang N, Sveinsson O, Macgregor S, Martin NG, Cree AJ, Gibson J, Macleod A, Jacob A, Ennis S, Young TL, Chan JC, Karwatowski WS, Hammond CJ, Thordarson K, Zhang M, Wadelius C, Lotery AJ, Trembath RC, Pang CP, Hoh J, Craig JE, Kong A, Mackey DA, Jonasson F, Thorsteinsdottir U, Stefansson K (2010) Common variants near CAV1 and CAV2 are associated with primary open-angle glaucoma. Nat Genet 42(10):906–909. doi:ng.661 [pii] 10.1038/ng.661

Wang CH, Liu BJ, Wu LS (2012) The Association Forecasting of 13 Variants Within Seven Asthma Susceptibility Genes on 3 Serum IgE Groups in Taiwanese Population by Integrating of Adaptive Neuro-fuzzy Inference System (ANFIS) and Classification Analysis Methods. J Med Syst 36(1):175–185. doi:10.1007/s10916-010-9457-4

Williams SE, Whigham BT, Liu Y, Carmichael TR, Qin X, Schmidt S, Ramsay M, Hauser MA, Allingham RR (2010) Major LOXL1 risk allele is reversed in exfoliation glaucoma in a black South African population. Mol Vis 16:705–712

Yang J, Yang P, Tezel G, Patil RV, Hernandez MR, Wax MB (2001) Induction of HLA-DR expression in human lamina cribrosa astrocytes by cytokines and simulated ischemia. Invest Ophthalmol Vis Sci 42(2):365–371

doi:10.1186/2193-1801-1-41
Cite this article as: Tokuda *et al.*: An approach to predict the risk of glaucoma development by integrating different attribute data. *SpringerPlus* 2012 **1**:41.

RIMS Kôkyûroku 1816

Developments in Geometry of Transformation Groups

May 28～June 1, 2012

edited by Toshio Sumi

October, 2012

Research Institute for Mathematical Sciences

Kyoto University, Kyoto, Japan

ON BORSUK-ULAM GROUPS

Ikumitsu NAGASAKI (京都府立医科大学 医学部・長崎　生光)
Department of Mathematics
Faculty of Medicine
Kyoto Prefectural University of Medicine

Fumihiro USHITAKI (京都産業大学 理学部・牛瀧　文宏)
Department of Mathematics
Faculty of Science
Kyoto Sangyo University

ABSTRACT. A Borsuk-Ulam group is a group for which the isovariant Borsuk-Ulam theorem holds. A fundamental question is: which groups are Borsuk-Ulam groups? In this article, we shall recall some properties and previous results on a Borsuk-Ulam group. After that, we provide a new family of Borsuk-Ulam groups. We also pose some open questions.

1. NOTATION AND TERMINOLOGY

Let G be a compact Lie group and V an (orthogonal or unitary) representation space of G. We denote by SV the unit sphere of V, called a G-representation sphere. A G-equivariant map (or G-map for short) $f : X \to Y$ is a continuous map between G-spaces satisfying

$$f(gx) = gf(x),\ \forall x \in X, g \in G.$$

It is easy to see that if f is G-equivariant, then

(1) $f(X^H) \subset Y^H$, so we have the restriction map

$$f^H : X^H \to Y^H.$$

(2) $G_x \leq G_{f(x)}$ $(\forall x \in X)$.

Definition. A continuous map $f : X \to Y$ is called a G-isovariant map if f is a G-equivariant map satisfying $G_x = G_{f(x)}$ $(\forall x \in X)$.

It is easy to see that $f : X \to Y$ is G-isovariant if and only if f is a G-equivariant map such that $f_{|G(x)} : G(x) \to Y$ is injective for any $x \in X$, where $G(x)$ is the orbit of x. Similarly we define an isovariant homotopy as follows.

2000 *Mathematics Subject Classification.* 57S17, 55M20,
The first author was partially supported by JSPS KAKENHI Grant Number 23540101.