Heuristic principal component analysis-based unsupervised feature extraction applied to gene expression analysis of amyotrophic lateral sclerosis data sets

Y-h. Taguchi[1,a)] Mitsuo Iwadate[2,b)] Hideaki Umeyama[2,c)]

Abstract: We applied principal component analysis (PCA)-based unsupervised feature extraction (FE) to amyotrophic lateral sclerosis (ALS) gene expression profiles. ALS is a debilitating neurodegenerative disorder with no effective therapy. The relevant gene expression profiles contained a small number of samples (from a few to tens) with a large number of features (several tens of thousands). Although it is important to recognize critical genes from gene expression profiles, a small-sample-large-feature situation makes FE difficult. In PCA-based unsupervised FE, features rather than samples are embedded into a low dimensional space, and critical genes are identified as outliers that are supposed to obey group-oriented behavior. The 29 candidate genes identified as critical for ALS by this methodology turned out to be biologically feasible based on comparisons with numerous previous studies. Together, they formed a collected gene regulation/protein binding network within which the known, but not explicitly identified in this study, three ALS-causing genes, *SOD1*, *TDP-43*, and *SETX*, could be naturally placed/embedded. Among the 29 genes, the translated chemokine receptor CCR6 protein was considered to be a potential therapy target and its antagonists/agonists were identified using the *in silico* drug discovery software ChooseLD. The ten top-ranked antagonists/agonists shared structures with many compounds that were previously known to bind to various proteins [This paper is the digest version of a conference paper [1]. For more details, see the conference paper version].

1. Introduction

The small-sample-large-feature situation is very common in bioinformatics. For example, in gene expression analysis it is relatively easy to measure the gene expression of tens of thousands of genes at once; however, getting many samples is much more costly, and thus more difficult. Although it is important to specify which genes contribute to the outcomes derived from given experimental treatments, it is well known that feature extraction (FE) is a difficult task under the small-sample-large-feature situation. To resolve this difficulty, principal component analysis (PCA)-based unsupervised FE was proposed and successfully applied to various bioinformatics problems [2–12]. Features, rather than samples, are embedded into a low dimensional space by PCA-based unsupervised FE, and critical genes are identified as outliers that are supposed to obey group-oriented behavior. We can identify features that obey group behavior because the gene expressions attributed to each sample are combined to generate the principal components (PCs) when features are embedded into low dimensional spaces. If many features share the same gene expression profiles attributed to each sample, then the combination of these features has more of a tendency to be employed as primary PCs, while gene expression profiles attributed to each sample that are not associated with many genes have less opportunity to be employed as primary PCs. To demonstrate the small-sample-large-feature situation we first ran the analysis using simulated data.

2. PCA-based unsupervised FE applied to a simulated data set

N features were embedded into low dimensional space. Because there were no orders other than those generated by a given group behavior, the first PC (PC1) is expected to reflect the group behavior. To confirm this exception, we selected 10 top outliers along PC1 (i.e., the features with large absolute PC1 scores, PCS^1_i, were extracted). Fig. 1(A) shows a typical example of PCA-based unsupervised FE applied to this data set. Using 100 ensembles with $N = 100, M = 20$, and $N_0 = 10$, when $\mu = 1(2)$ we found that 52.6% (89.5%) of N_0 extracted features were outliers correctly identified from features associated with group behavior ($1 \le i \le N_0$). Although this result suggests that PCA-based unsupervised FE can extract features that obey group behavior, because performance is clearly parameter dependent (in this example, it depends on μ) it is important to apply the methodology to a real data set to see whether PCA-based unsupervised FE works well with actual data.

[1] Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, 112-8551 Tokyo,Japan
[2] Department of Biological Science, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, 112-8551 Tokyo, Japan
[a)] tag@graular.com
[b)] iwadate@bio.chuo-u.ac.jp
[c)] umeyama@bio.chuo-u.ac.jp

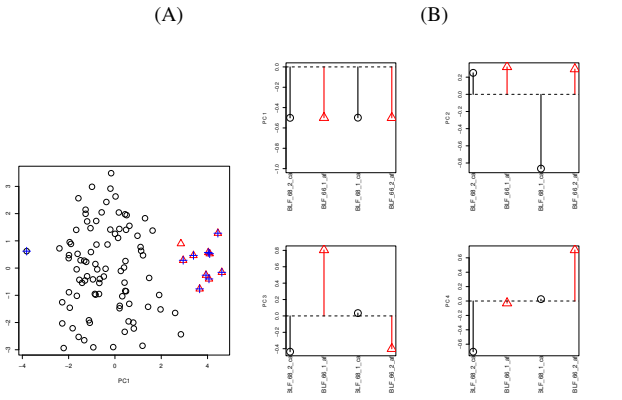

Fig. 1 (A) Typical example of PCA-based unsupervised FE applied to simulated data ($\mu = 2, N = 100, M = 20, N_0 = 10$). Two-dimensional embeddings of 100 features by PCA are represented. The 10 red triangles correspond to features associated with group behavior, and the 90 black circles are the others. Blue crosses indicate the 10 features identified by PCA-based unsupervised FE as outliers. In this particular example nine features associated with group behaviors were recognized correctly by choosing the outliers. (B) Contribution of ALS patient and control samples to the first, second, third, and fourth PCs (PC1, PC2, PC3, and PC4, respectively). Black bars with circles correspond to the patients; red bars with triangles correspond to the healthy controls. Horizontal broken lines indicate the baseline (zero).

3. PCA-based unsupervised FE applied to amyotrophic lateral sclerosis

Amyotrophic lateral sclerosis (ALS) is a debilitating neurodegenerative disorder without any effective therapy. A critical reason that no effective therapies exist is that the genetic mechanisms that cause ALS are unknown. Identifying critical genes relevant to ALS will accelerate our understanding of the disease and expedite the development of effective therapies. A potential difficulty of this task is the sparseness of available samples. ALS itself is a rare disease (one in 10,000 in a population), and obtaining the neurons required for experimental study is difficult without injuring the patients' nerves.

3.1 Data set from two ALS patients and two controls

Recently, Fogel et al. [13] generated fibroblast (i.e., skin) cell lines from two healthy controls and two patients (thus, $M = 4$), and measured gene expression using microarray technology. Because 24,525 genes were mapped onto each microarray plate (thus, $N = 24,525$), this is a typical small-sample-large-feature situation. To demonstrate the difficulty of the problem, we applied FE based on categorical regression analysis to this data set. We found no significant regression for any is, almost certainly because of the sparseness of samples. Next, we applied PCA-based unsupervised FE to this data set. Fig. 1(B) shows the contributions of each of the samples to the PCs, PCk_j. In contrast to the previous example with simulated data, because this is a real biological example PC1 may no longer reflect what we are interested in. Indeed, PC1 contributed more than 99% to the PCs but did not exhibit any sample dependence (Fig. 1(B)); thus, gene expression showed very little sample dependence. (This could be another reason why the previous regression analysis failed.) Although PC2, PC3, and PC4 contributed as little as 0.4%, 0.2%, and 0.1%, respectively, and thus exhibited clearer sample dependence, the sample dependence was not coincident with the distinction between ALS patients and healthy controls. However, if the sample dependence does indeed reflect some hidden group behavior attributed to a limited number of features, then the extraction of outliers along these PCs may have biological meaning. FE was performed as follows. First, the Z score was summed from the PC score for each i as, $Z_i^2 = \sum_{k=2}^{4} Z_{ik}^2$, $Z_{ik} = \frac{PCS_i^k - \langle PCS_i^k \rangle_k}{\delta PCS^k}$, $\langle PCS_i^k \rangle_k = \frac{1}{N}\sum_i PCS_i^k$, $\delta PCS^k = \sqrt{\frac{1}{N}\sum_i \left(PCS_i^k - \langle PCS_i^k \rangle_k\right)^2}$ where Z_{ik} is the Z score derived from the PCk score of the ith feature. Then assuming Z_i^2 obeys the χ^2 distribution, P-values are attributed to each i. After adjusting the P-values with the BH criterion, features associated with adjusted P-values less than 0.01 were extracted. In this way, a total of 708 features were selected.

To validate this selection biologically, we uploaded the list of 708 genes to the DAVID server [14, 15] and performed enrichment analyses to obtain enriched biological terms and pathways for these genes. DAVID compares a specified set of genes against a background that consists of all genes in a particular organism. The relevant metabolic pathways in the KEGG pathway enrichment analysis [16] are listed in Table 1.

Although the adjusted P-values attributed to some KEGG pathways were not always significant, it is remarkable that the most significant pathways included other neurodegenerative diseases, including Alzheimer's, Parkinson's, and Huntington's diseases (AD, PD, and HD). This result suggests that PCA-based unsupervised FE successfully selected biologically meaningful genes. Furthermore, using the alternative pathway analysis tool, TargetMine [17], we confirmed that AD and PD were significantly enriched KEGG pathways for the selected genes. DAVID also reported several other plausible upregulated features and/or tissues associated with the selected genes, including fetal brain cortex (despite the fact that the cell line was made from fibroblasts), Bethlem myopathy (a congenital, autosomal dominant form of muscular dystrophy), Ullrich congenital muscular dystrophy and other neuron-related features among the biological process gene ontology terms assigned by DAVID. We also uploaded the selected gene list to the g:Profiler web server [18], which is another toolset that reports enriched biological terms and/or concepts in a set of genes. g:Profiler reported the enrichment of genes targeted by several transcription factors, notably *Sp1*, *LRF*, *VDR*, and *E2F*, all of which have been reported to be related to neuron development and neural diseases. Transcription factors trigger gene expression by binding to the promoter regions (genomic regions that control expression) of their associated genes. The finding that the target genes of neuron-related transcription factors were enriched in our gene set demonstrates the ability of PCA-based unsupervised FE to detect genes with group behavior.

We also checked whether our selected genes overlapped with the ALS genes reported by the Gendoo server [19], which attributes P-values to the simultaneous appearances in published literature of genes and diseases. We downloaded 211 genes associated with ALS with P-values less than 0.05. We found 13 common genes that were common between this ALS-related gene list and the 708 genes that we identified with PCA-based unsupervised FE. This overlap was associated with P-values of 4×10^{-3};

Table 1 KEGG pathway enrichment analysis using DAVID

	(A)			(B)		
KEGG pathway	Count	%	*P*-value	Count	%	*P*-value
hsa03010: Ribosome	42	7.25	7.68 E-29	34	5.67	7.68 E-20
hsa04510: Focal adhesion	30	5.18	2.65 E-06	30	5.00	2.47 E-06
hsa04512: ECM-receptor interaction	15	2.59	2.47 E-04	13	2.16	2.63 E-03
hsa05130: Pathogenic *Escherichia coli* infection	12	2.07	3.09 E-04	14	2.33	1.30 E-05
hsa00190: Oxidative phosphorylation	19	3.28	3.78 E-04	20	3.33	1.20 E-04
hsa05110: Vibrio cholerae infection	—	—	—	11	1.83	1.07 E-03
hsa05010: Alzheimer's disease	20	3.45	2.21 E-03	18	3.00	1.09 E-02
hsa05012: Parkinson's disease	17	2.93	2.38 E-03	15	2.50	1.36 E-02
hsa03050: Proteasome	—	—	—	8	1.33	1.61 E-02
hsa00010: Glycolysis / Gluconeogenesis	10	1.72	6.51 E-03	8	1.33	5.29 E-02
hsa05016: Huntington's disease	18	3.10	2.75 E-02	21	3.50	2.91 E-03
hsa00030: Pentose phosphate pathway	—	—	—	6	1.00	1.21 E-02
hsa00620: Pyruvate metabolism	—	—	—	7	1.16	2.43 E-02
hsa04350: TGF-beta signaling pathway	—	—	—	11	1.86	2.53 E-02
hsa04670: Leukocyte transendothelial migration	—	—	—	13	2.17	3.55 E-02
hsa04722: Neurotrophin signaling pathway	—	—	—	13	2.17	4.91 E-02
hsa00630: Glyoxylate and dicarboxylate metabolism	—	—	—	4	o.67	5.01 E-02
hsa04540: Gap junction	—	—	—	10	1.67	6.56 E-02
hsa04142: Lysosome	13	2.24	3.43 E-02	—	—	—
hsa00480: Glutathione metabolism	7	1.20	6.38 E-02	7	1.17	6.30 E-02
hsa04810: Regulation of actin cytoskeleton	19	3.28	6.51 E-02	19	3.17	6.35 E-02
hsa05412: Arrhythmogenic right ventricular cardiomyopathy (ARVC)	9	1.55	6.73 E-02	—	—	– -
hsa04260: Cardiac muscle contraction	9	1.55	7.60 E-02	9	1.50	7.48 E-02
hsa04530: Tight junction	—	—	—	13	2.17	7.89 E-02

(A) 708 genes identified in the data set from two ALS patients and two controls. *P*-values are not adjusted.

(B) 715 genes identified in the data set from transfected cell lines by PCA-based unsupervised FE. *P*-values are not adjusted.

thus, we concluded that the two lists were significantly overlapped. These enrichment and association analyses all supported the ability of PCA-based unsupervised FE to select biologically feasible genes.

3.2 Data set from transfected cell lines

Fogel et al. [13] performed another experiment with the fibroblast (i.e. skin) cell lines from two healthy controls and two ALS patients by transfecting of artificially mutated genes into the cell lines. They generated three mutations, reported to be related to ALS [13], and transfected them, as well as control non-mutated versions of the same genes, into the cell lines. Because four cell lines were used, four times the three mutated genes, plus the non-mutated control, resulted in sixteen samples ($M = 16$). This is another small-sample-large-feature situation because there are $N = 24,525$ features and 16 samples. To demonstrate the difficulty of the task, we again performed regression analysis, and showed that, as with the previous data set, no significant associated adjusted *P*-values were found. Next, we applied PCA-based unsupervised FE and found that, once more, PC1 contributed 97.6 %, and did not exhibit any sample dependency (data not shown). Fig. 2 shows the contributions of PC2 and PC3, which correspond to the differences between the mutated and non-mutated gene transfections (PC2 and PC3 contributed only 0.7 % and 0.4 %, respectively). This result indicates that the contributions of samples transfected with the non-mutated genes to the PCs were extracted from the contributions of samples transfected with mutated genes, as in $\Delta PCk_j = PCk_j - PCk_{c0}$ where PCk_{c0} is the contribution of a sample when a non-mutated gene instead of a mutated gene is transfected to the cth cell line, and the jth sample belongs to the cth cell line. No contributions consistent with the difference between the mutated and non-mutated transfections were apparent because both up- and down-regulation were observed for PC2 and PC3 independent of the transfected mutated gene. However, when we tried the following alternative regression analysis, we found significant regression for $k = 2$ and 3, $\Delta PC_{kj} = \sum_{c=1}^{4} a_{kc}\psi_{jc} + g_k$, where ΔPC_{kj} represents the differential contributions of the jth sample ($j = 1, \cdots 12$) to the kth PC (or loading) defined above, and ψ_{jc} is only assigned 1 when the jth sample is taken from the cth cell line ($c = 1, \cdots, 4$), as shown in Fig. 1(B), and is otherwise assigned 0. This situation is equivalent to ignorance of transfected gene dependence, which meant that no matter which mutated gene was transfected, the cell lines exhibited the same gene expression profile as the cell lines transfected with the non-mutated gene. Thus, there is still a possibility

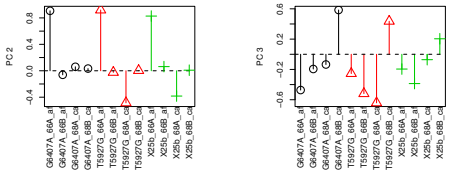

Fig. 2 Contributions of samples to the differential second and third PCs (PC2 and PC3) corresponding to mutated and non-mutated gene transfection. Black, red, and green bars with a circle, triangle, and cross respectively, correspond to the transfection of three distinct mutated genes. Horizontal broken lines are the baseline (zero).

that PCA-based unsupervised FE can extract biologically feasible features. Next we performed the following analysis by assuming that the *PC* scores, PCS_i^k, obey a normal distribution for $k = 2$, and 3. *P*-values were computed and attributed to the ith feature as P_{ik}, then, geometric mean values were computed with $k = 2$ and 3, $P_i = \sqrt{P_{i2} \cdot P_{i3}}$. The P_i obtained was then adjusted by the BH criterion, and features with adjusted *P*-values less than 0.01 were extracted. We obtained 715 features as a result of this procedure.

Interestingly, although the simple comparison between healthy controls and ALS patients described in the previous section is a distinct measurement, the results of the biological investigations using the 708 gene set described previously were nearly identical to the results we obtained with this new set of 715 selected genes.

As before, we uploaded the list of 715 genes to DAVID. The result of the KEGG pathway enrichment analysis is shown in Table 1. Although some of the adjusted P-values attributed to particular KEGG pathways were not always significant, again the most significant pathways included the AD, PD, and HD neurodegenerative diseases, suggesting that PCA-based unsupervised FE successfully selected biologically meaningful genes. Furthermore, using TargetMine, we confirmed that AD ($P = 2.13 \times 10^{-2}$), PD ($P = 1.98 \times 10^{-2}$), and HD ($P = 1.09 \times 10^{-2}$) were significantly enriched KEGG pathways for our selected genes (BH adjusted P-values). Additionally, DAVID reported that fetal brain cortex (despite the fact that the cell lines were from fibroblasts, $P = 3.22 \times 10^{-23}$) was an upregulated tissue (BH adjusted P-value). g:Profiler also reported the enrichment of genes targeted by several transcription factors that were identified previously, e.g., *SP1*, *LRF*, *VDR*, *E2F-1*, and *E2F* (P-values ranged from 6.29×10^{-11} to 3.19×10^{-2}; only significant P-values are presented because of the implementation of g:Profiler).

We identified 14 genes that were common between the Gendoo ALS-related gene list and the 715 genes we identified with the PCA-based unsupervised FE method. This overlap was associated with P-values of 2×10^{-3}; thus, we considered that these two lists were significantly overlapped.

Despite the coincidence of biological terms and concepts between the two sets of selected genes, the two lists of are not identical; indeed, the overlap between the two sets was only 393 genes, which is only slightly more than half of all of the selected genes (708 versus 715 genes). This result further suggests the robustness of PCA-based unsupervised FE in selecting biologically feasible sets of genes.

3.3 Identification of genes especially critical for ALS

We identified a large number of genes that were possibly critical for ALS in the two *in vitro* data sets. To better identify these genes, we selected the 100 top-most significant genes with low P-values from among the 708 genes (identified by the comparison between controls and ALS patients) and the 715 genes (identified by the mutated genes transfection). Then, we selected the common genes between the 100 top-ranked genes in the two sets and found 29 genes that were included in both data sets. We considered that this amount of overlap was highly significant; therefore, we supposed that the 29 selected genes (Table 2) were especially critical genes for ALS.

Table 2 Common genes between the 100 top-most significant genes in two *in vitro* data sets

ACTA2 ADM ANXA1 CCR6 CFL1 COL8A1 CRYAB CTGF FBLN1 HIST1H4C **ID1 ID3 IGFBP7 M6PRBP1** MGC16703 MRCL3 PSG4 PODXL RPS4Y1 **S100A10** SNHG5 **TAGLN TGFBI THBS1 TMEM119** TNFRSF11B **TPM1 UBC**

The two data sets were the two ALS patients and two controls set and the mutated genes transfection set. The genes in bold have been associated with ALS in at least one previous study (data not shown).

Using STRING [20], we found that PPIs were enriched among the genes in Table 2 (31 PPIs were detected, while the expected PPIs were 14; $P = 5.45 \times 10^{-5}$). However, none of three neurodegenerative diseases (AD, HD, or PD) reported in the KEGG pathway enrichment analysis (Table 1) were associated with the 29 selected genes (data not shown). This result apparently indicated that our identification of critical ALS genes was unsuccessful. However, after performing an extensive literature search, summarized in the pathway map in Fig. 3, we found that the initial assumption was incorrect. For example, although no reports directly related ACTA2 to ALS, ACTA2 was reported to bind to LRRTM2 (Fig. 3(a)), which was targeted by TDP43 (Fig. 3(b)), one of the major ALS-causing proteins [21]. Most of the relations between the 29 selected genes and ALS that we determined based on the literature search were indirect, and this may explain why the relations between neurodegenerative diseases and these genes were not identified in the KEGG pathway enrichment analysis. In addition, many genes in Table 2 were associated with at least one previous report that suggested a relation with ALS (data not shown). Constructing a biologically informative pathway of this kind is difficult without successful identifications; therefore, we believe that we have shown that our methodology can successfully screen for ALS critical genes.

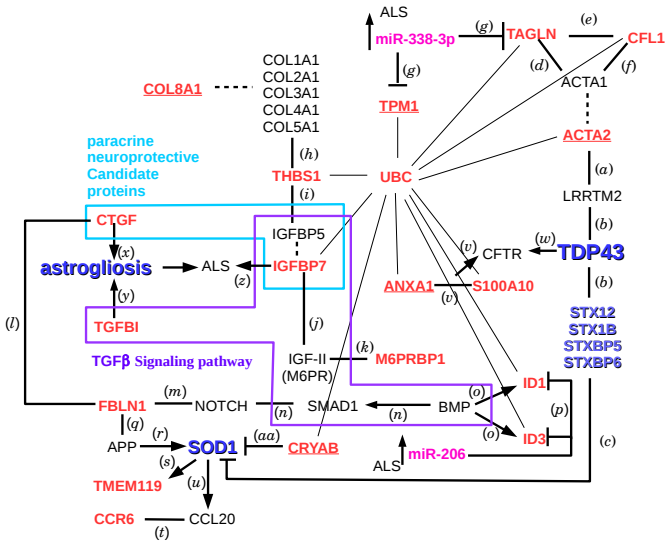

Fig. 3 Pathway map summarizing the results of an extensive literature search. Genes in Table 2 are shown in red. Arrows and T-shaped segments represent up/downregulation, respectively. Solid black bold lines represent protein binding interactions. The segment that included UBC (ubiquitin C) is represented by thin black lines because the connections are too ubiquitous. Dotted black lines represent connections between proteins in the same superfamily/isoforms that may be expected to bind to common protein partners. The underlined genes also were detected in a proteomic analysis (data not shown). *TDP43*, *SOD1*, and *SETX* are major ALS drivers genes [22]. Astrogliosis was reported to play critical roles in ALS [23]. Some of the proteins have been identified as paracrine neuroprotective candidate proteins [24] that may play critical roles in neurogenesis. Finally, the map includes genes related to the TGFβ pathway that was recognized recently as an important contributor to the causes of ALS [25]. References used to generate the pathway map are: (a) [26], (b) [27], (c) [28], (d) [29], (e) [30], (f) [31], (g) [32], (h) [33], (i) [34], (j) [35], (k) [36], (l) [37], (m) [38], (n) [39], (o) [40], (p) [41], (q) [42], (r) [43], (s) [44], (t) [45], (u) [46], (v) [47], (w) [48], (x) [49], (y) [50], (z) [51], (aa) [52].

3.4 CCR6 as a therapy target of ALS and *in silico* drug discovery

Although we successfully screened ALS critical genes, it is important to identify the gene that would make the most promising therapy target. We identify *CCR6* as the most promising gene for the following reasons. Although pathological details of ALS are still unknown, the infiltration of T lymphocytes and dendritic cells into the spinal cord has been observed in the initial phases of ALS [53] and was regarded as a potential causing factor of ALS through inflammation. However, Saresella et al. [54] confirmed that T helper type 17 (Th17) cells, which are thought to induce inflammation, were increased in ALS peripheral blood mononuclear cells (PBMC) while regulatory T cells, which are thought to suppress inflammation, were reduced in ALS PBMC. This finding suggested that populations of these two cell types largely affected the progression of ALS. Yamasaki et al. [55] recently confirmed that experimental autoimmune encephalomyelitis (EAE) was induced by CCL20-associated CCR6-mediated Th17 cell migration, while EAE was suppressed by CCL20-unassociated CCR6-mediated regulatory T cell migration. CCL20-associated CCR6-mediated Th17 cell migration to the affected diseased part (in this case, inflamed joints) also has been observed in other diseases [56] where an anti-CCR6 monoclonal antibody substantially inhibited the symptoms of ALS.

And as discussed above, CCR6 binds to CCL20 (Fig. 3(*t*)), which is enhanced by SOD1 in ALS (Fig. 3(*u*)). Thus, CCR6, which was identified in this study, was considered to be a potential ALS therapy target.

3.4.1 Inference of tertiary structure of CCR6

The tertiary structure of CCR6 is required for *in silico* drug discovery. Because the tertiary structure of CCR6 was not available in the PDB, we predicted its structure by homology modeling. We identified the homology modeling candidate structures of β_2 adrenergic receptors (ADRB2s) with an agonist as well as three antagonists bound to the ligand binding sites in the PDB.

3.4.2 *in silico* drug discovery to identify agonist candidates

The ten top-ranked compounds that were selected as candidate ligands for CCR6 are shown in Table 3. In the list of "Related Compounds with Annotation" in PubChem, six of the ten top-ranked compounds were associated with at least one known active ligand for some proteins. This suggested that our list of candidate agonist compounds was promising, because the lack of associated known active ligands may simply have reflected the lack of trials/experiments.

3.4.3 *in silico* drug discovery to identify antagonist candidates

The 10 top-ranked compounds that were selected as candidate ligands are shown in Table 3. In the list of "Related Compounds with Annotation" in PubChem, five of the 10 top-ranked compounds (for CXCR4 and OPRM1) and eight of 10 top-ranked compounds (for OPRK1) were associated with at least one known active ligand for some proteins. Although the top three compounds were not associated with known protein binding ligands, they were associated with inhibitors of microbial proliferation. Because compounds are likely to affect microbial proliferation by binding to proteins, we supposed that these three compounds also probably bind to some proteins. In particular, all three of the first ranked compounds were associated with at least one known active ligand of some proteins when microbial proliferation inhibitors were included. This finding suggested that our list of candidate antagonist compounds was promising, because the lack of associated known active ligands may simply reflect the lack of trials/experiments.

Table 3 Ten top ranked agonists/antagonists for CCR6 inferred by chooseLD.

Agonist candidates based on 3SN6_A (3P0G_A) and ADRB2 (4LDE_A)
1. 008072156, 2. **008206651**, 3. 007915138, 4. **006405195**, 5. **006898371**, 6. **007416814**, 7. **006898272**, 8. 007360992, 9. 007318671, 10. **007920051**
Antagonist candidates based on CXCR4 (3ODU_A)
1. **005222305**, 2. **002168831**, 3. **002168773**, 4. 002168537, 5. 005227329, 6. 002168546, 7. **005227330**, 8. 007649892, 9. 016880807, 10. **005227219**
Antagonist candidates based on OPRM1 (4DKL_B)
1. **015910967**, 2. 006616964, 3. **015911601**, 4. 006409130, 5. 006913742, 6. **015994619**, 7. 006889306, 8. **007804964**, 9. 007333860, 10. **007429734**
Antagonist candidates based on OPRK1 (4DJH_B)
1. **007816975**, 2. **007802415**, 3. **007802145**, 4. 005427196, 5. **006913516**, 6. **007150068**, 7. **007818089**, 8. **007801751**, 9. 007045063, 10. **007044891**

Numbers indicate rank order. Bold compound IDs are listed in "Related Compounds with Annotation" and suggested to be protein binding ligands. All compound IDs should be prefixed by "AKOS", for example "AKOS012345678".

Acknowledgments This study was supported by KAKENHI 23300357 and 26120528 and a Chuo University Joint Research Grant. We thank Dr. Katsuichiro Komatsu who helped with the *in silico* drug screening using ChooseLD.

References

[1] Taguchi, Y.-h., Iwadate, M. and Umeyama, H.: Heuristic Principal Component Analysis Based unsupervised Feature Extraction and its Application to Gene Expression Analysis of Amyotrophic Lateral Sclerosis, *Computational Intelligence and Bioinformatics and Computational Biology (CIBCB), IEEE Symposium on*, New York, NY, USA, IEEE (2015).

[2] Taguchi, Y. H., Iwadate, M. and Umeyama, H.: Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for posttraumatic stress disorder-mediated heart disease, *BMC Bioinformatics*, Vol. 16, No. 1, p. 139 (2015).

[3] Taguchi, Y.-h., Iwadate, M., Umeyama, H., Murakami, Y. and Okamoto, A.: Heuristic principal component analysis-aased unsupervised feature extraction and its application to bioinformatics, *Big Data Analytics in Bioinformatics and Healthcare* (Wang, B., Li, R. and Perrizo, W., eds.), IGI global, pp. 138–162 (2015).

[4] Taguchi, Y.-h. and Okamoto, A.: Principal Component Analysis for Bacterial Proteomic Analysis, *Pattern Recognition in Bioinformatics* (Shibuya, T., Kashima, H., Sese, J. and Ahmad, S., eds.), LNCS, Vol. 7632, Springer Berlin Heidelberg, pp. 141–152 (2012).

[5] Murakami, Y., Toyoda, H., Tanahashi, T. et al.: Comprehensive miRNA expression analysis in peripheral blood can diagnose liver disease, *PLoS ONE*, Vol. 7, No. 10, p. e48366 (2012).

[6] Ishida, S., Umeyama, H., Iwadate, M. and Taguchi, Y. H.: Bioinformatic Screening of Autoimmune Disease Genes and Protein Structure Prediction with FAMS for Drug Discovery, *Protein Pept. Lett.*, Vol. 21, No. 8, pp. 828–39 (2014).

[7] Taguchi, Y. H. and Murakami, Y.: Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers, *PLoS ONE*, Vol. 8, No. 6, p. e66714 (2013).

[8] Kinoshita, R., Iwadate, M., Umeyama, H. and Taguchi, Y. H.: Genes associated with genotype-specific DNA methylation in squamous cell carcinoma as candidate drug targets, *BMC Syst Biol*, Vol. 8 Suppl 1, p. S4 (2014).

[9] Taguchi, Y. H. and Murakami, Y.: Universal disease biomarker: can a fixed set of blood microRNAs diagnose multiple diseases?, *BMC Res Notes*, Vol. 7, p. 581 (2014).

[10] Murakami, Y., Tanahashi, T., Okada, R. et al.: Comparison of Hepatocellular Carcinoma miRNA Expression Profiling as Evaluated by Next Generation Sequencing and Microarray, *PLoS ONE*, Vol. 9, No. 9, p.

e106314 (2014).
[11] Umeyama, H., Iwadate, M. and Taguchi, Y. H.: TINAGL1 and B3GALNT1 are potential therapy target genes to suppress metastasis in non-small cell lung cancer, *BMC Genomics*, Vol. 15 S9, p. S2 (2014).
[12] Taguchi, Y.-h.: Integrative Analysis of Gene Expression and Promoter Methylation during Reprogramming of a Non-Small-Cell Lung Cancer Cell Line Using Principal Component Analysis-Based Unsupervised Feature Extraction, *Intelligent Computing in Bioinformatics*, LNCS, Vol. 8590, Springer, Heidelberg, pp. 445–455 (2014).
[13] Fogel, B. L., Cho, E., Wahnich, A. et al.: Mutation of senataxin alters disease-specific transcriptional networks in patients with ataxia with oculomotor apraxia type 2, *Hum. Mol. Genet.*, Vol. 23, No. 18, pp. 4758–4769 (2014).
[14] Huang, d. a. W., Sherman, B. T. and Lempicki, R. A.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc*, Vol. 4, No. 1, pp. 44–57 (2009).
[15] Huang, d. a. W., Sherman, B. T. and Lempicki, R. A.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Res.*, Vol. 37, No. 1, pp. 1–13 (2009).
[16] Kanehisa, M., Goto, S., Sato, Y. et al.: Data, information, knowledge and principle: back to metabolism in KEGG, *Nucleic Acids Res.*, Vol. 42, No. Database issue, pp. 199–205 (2014).
[17] Chen, Y. A., Tripathi, L. P. and Mizuguchi, K.: TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery, *PLoS ONE*, Vol. 6, No. 3, p. e17844 (2011).
[18] Reimand, J., Arak, T. and Vilo, J.: g:Profiler–a web server for functional interpretation of gene lists (2011 update), *Nucleic Acids Res.*, Vol. 39, No. Web Server issue, pp. W307–315 (2011).
[19] Nakazato, T., Bono, H., Matsuda, H. and Takagi, T.: Gendoo: functional profiling of gene and disease features using MeSH vocabulary, *Nucleic Acids Res.*, Vol. 37, No. Web Server issue, pp. W166–169 (2009).
[20] Szklarczyk, D., Franceschini, A., Wyder, S. et al.: STRING v10: protein-protein interaction networks, integrated over the tree of life, *Nucleic Acids Res.*, Vol. 43, No. Database issue, pp. D447–452 (2015).
[21] Tsuiji, H., Iguchi, Y., Furuya, A. et al.: Spliceosome integrity is defective in the motor neuron diseases ALS and SMA, *EMBO Mol Med*, Vol. 5, No. 2, pp. 221–234 (2013).
[22] Chen, S., Sayana, P., Zhang, X. and Le, W.: Genetics of amyotrophic lateral sclerosis: an update, *Mol Neurodegener*, Vol. 8, p. 28 (2013).
[23] Vargas, M. R. and Johnson, J. A.: Astrogliosis in amyotrophic lateral sclerosis: role and therapeutic potential of astrocytes, *Neurotherapeutics*, Vol. 7, No. 4, pp. 471–481 (2010).
[24] Hauck, S. M., Gloeckner, C. J., Harley, M. E. et al.: Identification of paracrine neuroprotective candidate proteins by a functional assay-driven proteomics approach, *Mol. Cell Proteomics*, Vol. 7, No. 7, pp. 1349–1361 (2008).
[25] Katsuno, M., Adachi, H., Banno, H., Suzuki, K., Tanaka, F. and Sobue, G.: Transforming growth factor- signaling in motor neuron diseases, *Curr. Mol. Med.*, Vol. 11, No. 1, pp. 48–56 (2011).
[26] : High-Throughput Proteomic Mapping of Human Interaction Networks via Affinity-Purification Mass Spectrometry:Pre-Publication Dataset in BioGRID, http://thebiogrid.org/166968/publication/high-throughput-proteomic-mapping-of-human-interaction-networks-via-affinity-purification-mass-spectrometry.html. Accessed: 2015-04-06.
[27] Narayanan, R. K., Mangelsdorf, M., Panwar, A. et al.: Identification of RNA bound to the TDP-43 ribonucleoprotein complex in the adult mouse brain, *Amyotroph Lateral Scler Frontotemporal Degener*, Vol. 14, No. 4, pp. 252–260 (2013).
[28] Suraweera, A., Lim, Y., Woods, R. et al.: Functional role for senataxin, defective in ataxia oculomotor apraxia type 2, in transcriptional regulation, *Hum. Mol. Genet.*, Vol. 18, No. 18, pp. 3384–3396 (2009).
[29] Fu, Y., Liu, H. W., Forsythe, S. M. et al.: Mutagenesis analysis of human SM22: characterization of actin binding, *J. Appl. Physiol.*, Vol. 89, No. 5, pp. 1985–1990 (2000).
[30] Ewing, R. M., Chu, P., Elisma, F. et al.: Large-scale mapping of human protein-protein interactions by mass spectrometry, *Mol. Syst. Biol.*, Vol. 3, p. 89 (2007).
[31] Moriyama, K. and Yahara, I.: Human CAP1 is a key factor in the recycling of cofilin and actin for rapid actin turnover, *J. Cell. Sci.*, Vol. 115, No. Pt 8, pp. 1591–1601 (2002).
[32] De Felice, B., Annunziata, A., Fiorentino, G. et al.: miR-338-3p is over-expressed in blood, CFS, serum and spinal cord from sporadic amyotrophic lateral sclerosis patients, *Neurogenetics*, Vol. 15, No. 4, pp. 243–253 (2014).
[33] Galvin, N. J., Vance, P. M., Dixit, V. M. et al.: Interaction of human thrombospondin with types I-V collagen: direct binding and electron microscopy, *J. Cell Biol.*, Vol. 104, No. 5, pp. 1413–1422 (1987).
[34] Son, H. N., Nam, J. O., Kim, S. and Kim, I. S.: Multiple FAS1 domains and the RGD motif of TGFBI act cooperatively to bind $\alpha v\beta 3$ integrin, leading to anti-angiogenic and anti-tumor effects, *Biochim. Biophys. Acta*, Vol. 1833, No. 10, pp. 2378–2388 (2013).
[35] Kihira, T., Suzuki, A., Kubo, T., Miwa, H. and Kondo, T.: Expression of insulin-like growth factor-II and leukemia inhibitory factor antibody immunostaining on the ionized calcium-binding adaptor molecule 1-positive microglias in the spinal cord of amyotrophic lateral sclerosis patients, *Neuropathology*, Vol. 27, No. 3, pp. 257–268 (2007).
[36] Diaz, E. and Pfeffer, S. R.: TIP47: a cargo selection device for mannose 6-phosphate receptor trafficking, *Cell*, Vol. 93, No. 3, pp. 433–43 (1998).
[37] Perbal, B., Martinerie, C., Sainson, R. et al.: The C-terminal domain of the regulatory protein NOVH is sufficient to promote interaction with fibulin 1C: a clue for a role of NOVH in cell-adhesion signaling, *Proc. Natl. Acad. Sci. U.S.A.*, Vol. 96, No. 3, pp. 869–874 (1999).
[38] Wang, J., Huo, K., Ma, L. et al.: Toward an understanding of the protein interaction network of the human liver, *Mol. Syst. Biol.*, Vol. 7, p. 536 (2011).
[39] Cook, B. D. and Evans, T.: BMP signaling balances murine myeloid potential through SMAD-independent p38MAPK and NOTCH pathways, *Blood*, Vol. 124, No. 3, pp. 393–402 (2014).
[40] Kersten, C., Dosen, G., Myklebust, J. H. et al.: BMP-6 inhibits human bone marrow B lymphopoiesis–upregulation of Id1 and Id3, *Exp. Hematol.*, Vol. 34, No. 1, pp. 72–81 (2006).
[41] Toivonen, J. M., Manzano, R., Olivan, S. et al.: MicroRNA-206: a potential circulating biomarker candidate for amyotrophic lateral sclerosis, *PLoS ONE*, Vol. 9, No. 2, p. e89065 (2014).
[42] Ohsawa, I., Takamura, C. and Kohsaka, S.: Fibulin-1 binds the amino-terminal head of β-amyloid precursor protein and modulates its physiological function, *J. Neurochem.*, Vol. 76, No. 5, pp. 1411–20 (2001).
[43] Bryson, J. B., Hobbs, C., Parsons, M. J. et al.: Amyloid precursor protein (APP) contributes to pathology in the SOD1(G93A) mouse model of amyotrophic lateral sclerosis, *Hum. Mol. Genet.*, Vol. 21, No. 17, pp. 3871–3882 (2012).
[44] Chiu, I. M., Morimoto, E. T., Goodarzi, H. et al.: A neurodegeneration-specific gene-expression signature of acutely isolated microglia from an amyotrophic lateral sclerosis mouse model, *Cell Rep*, Vol. 4, No. 2, pp. 385–401 (2013).
[45] Schutyser, E., Struyf, S. and Van Damme, J.: The CC chemokine CCL20 and its receptor CCR6, *Cytokine Growth Factor Rev.*, Vol. 14, No. 5, pp. 409–426 (2003).
[46] Fiala, M., Chattopadhay, M., La Cava, A. et al.: IL-17A is increased in the serum and in spinal cord CD8 and mast cells of ALS patients, *J Neuroinflammation*, Vol. 7, p. 76 (2010).
[47] Borot, F., Vieu, D. L., Faure, G. et al.: Eicosanoid release is increased by membrane destabilization and CFTR inhibition in Calu-3 cells, *PLoS ONE*, Vol. 4, No. 10, p. e7116 (2009).
[48] Subramanian, S., Ayala, P., Wadsworth, T. L. et al.: CCR6: a biomarker for Alzheimer's-like disease in a triple transgenic mouse model, *J. Alzheimers Dis.*, Vol. 22, No. 2, pp. 619–629 (2010).
[49] Spliet, W. G., Aronica, E., Ramkema, M. et al.: Increased expression of connective tissue growth factor in amyotrophic lateral sclerosis human spinal cord, *Acta Neuropathol.*, Vol. 106, No. 5, pp. 449–457 (2003).
[50] Yun, S. J., Kim, M. O., Kim, S. O. et al.: Induction of TGF-beta-inducible gene-h3 (βig-h3) by TGF-beta1 in astrocytes: implications for astrocyte response to brain injury, *Brain Res. Mol. Brain Res.*, Vol. 107, No. 1, pp. 57–64 (2002).
[51] Wilczak, N., de Vos, R. A. and De Keyser, J.: Free insulin-like growth factor (IGF)-I and IGF binding proteins 2, 5, and 6 in spinal motor neurons in amyotrophic lateral sclerosis, *Lancet*, Vol. 361, No. 9362, pp. 1007–1011 (2003).
[52] Marino, M., Papa, S., Crippa, V. et al.: Differences in protein quality control correlate with phenotype variability in 2 mouse models of familial amyotrophic lateral sclerosis, *Neurobiol. Aging* (2014).
[53] Holmøy, T.: T cells in amyotrophic lateral sclerosis., *European journal of neurology : the official journal of the European Federation of Neurological Societies*, Vol. 15, No. 4, pp. 360–366 (online), DOI: 10.1111/j.1468-1331.2008.02065.x (2008).
[54] Saresella, M., Piancone, F., Tortorella, P. et al.: T helper-17 activation dominates the immunologic milieu of both amyotrophic lateral sclerosis and progressive multiple sclerosis, *Clinical Immunology*, Vol. 148, No. 1, pp. 79–88 (online), DOI: 10.1016/j.clim.2013.04.010 (2013).
[55] Yamazaki, T., Yang, X. O., Chung, Y. et al.: CCR6 regulates the migration of inflammatory and regulatory T cells., *Journal of immunology (Baltimore, Md. : 1950)*, Vol. 181, No. 12, pp. 8391–8401 (online), DOI: 10.4049/jimmunol.181.12.8391 (2008).
[56] Hirota, K., Yoshitomi, H., Hashimoto, M. et al.: Preferential recruitment of CCR6-expressing Th17 cells to inflamed joints via CCL20 in rheumatoid arthritis and its animal model, *J. Exp. Med.*, Vol. 204, No. 12, pp. 2803–2812 (2007).

