Genomics & Informatics- Vol. 5(1) 1-5, March 2007

Detection of Neural Fates from Random Differentiation : Application of Support Vector Machine

Min Su Lee[1], Jeong-Hyuck Ahn[2] and Woong-Yang Park[3]*

[1]Department of Computer Science and Engineering, Ewha Womans University, Seoul 120-759, Korea, [2]Laboratory of Molecular and Cellular Neuroscience, Rockefeller University, New York, NY, U.S.A., [3]Human Genome Research Institute and Departmentof Biochemistry, Seoul National University College of Medicine, Seoul 110-799, Korea

Abstract

Embryonic stem cells can be differentiated into various types of cells, requiring a tight regulation of transcription. Biomarkers related to each lineage of cells are used to guide the differentiation into neural or any other fates. In previous experiments, we reported the guided differentiation (GD)-specific genes by comparing profiles of random differentiation (RD). Interestingly 68% of differentially expressed genes in GD overlap with that of RD, which makes it difficult for us to separate the lineages by examining several markers. In this paper, we design a prediction model to identify the differentiation into neural fates from any other lineage. From the profiles of 11,376 genes, 203 differentially expressed genes between neural and random differentiation were selected by random variance T-test with 95% confidence and 5% false discovery rate. Based on support vector machine algorithm, we could select 79 marker genes from the 203 informative genes to construct the optimal prediction model. Here we propose a prediction model for the prediction of neural fates from random differentiation which is constructed with a perfect accuracy.

Keywords: stem cells, microarray, neural differentiation, random differentiation

Introduction

Development of mammalian central nervous system (CNS) is a complex process involving an orchestrated regulation of structural and regulatory genes through differentiation stages of multipotent stem cells into neurons (Sasai, 1998). Mouse embryonic stem (ES) cells retain the characteristics of multipotent stem cells exhibiting the infinite self-renewal activity and the differentiation potential into various kinds of lineages (Czyz and Wobus, 2001). Numerous efforts have been made to induce ES cells into neurons by regulating transcriptions of critical genesin the hope of using these cells for therapy for neurological disorders such as Parkinson's disease.

In previous works, thanks to the microarray technology, we profiled the transcriptions of two differentiation model: guided differentiation into DA neurons (GD) and random differentiation into EB (RD), and reported neural specific genes by comparing the random differentiation (Lee *et al.*, 2006). However, most differentially expressed genes in GD overlapped with those of RD. This makes it difficult for us to separate the differentiation types by examining several markers.

In this paper, we constructed a prediction model to identify neural differentiation from random differentiation using support vector machine which outperforms in binary classification tasks. From the profiles of 11,376 genes, we selected 203 differentially expressed genes between neural and random differentiation using a feature selection method which is random variance T-test with 95% confidence and false discovery rate < 0.5%. Then, we investigated about compact biomarker gene set based on support vector machine algorithm. We could select 79 marker genes among the 203 informative genes to construct optimal prediction model. As a result, a prediction model for discriminate neural differentiation from random differentiation is constructed with perfect accuracy.

Methods

Experimental design and microarray experiments

To detect neural fates from random differentiation, we used microarray dataset which were experimented at previous study (Lee *et al.*, 2006). The microarray dataset consists of neural differentiation and random differentiation of embryonic stem cells. Each differentiation process has 5 stages of cell cultivation.

Guided differentiation to the DA neuron of mouse embryonic stem cells was induced as described previously (Lee *et al.*, 2000). Briefly undifferentiated ES cells (stage 1) were grown on gelatin-coated tissue culture plates in KO-DMEM media. To induce EB formation (stage 2), the cells were dissociated into a single-cell suspension and plated onto non-adherent bacterial culture dishes at a density of 2.5×104 cells/cm^2 in the KO medium. After four days, the cells were transferred to the original tissue culture dish in a serum-free Insulin/Transferrin/Selenium/Fibronectin (ITSF) medium to

*Corresponding author: E-mail wypark@snu.ac.kr
Tel +82-2-740-8241, Fax +82-2-744-4534
Accepted 4 Dec 2006

select the nestin-positive cells (stage 3). After 6 days of selection, the cells were expanded (stage 4) by transferring to the plate coated with polyornithine and laminin in N^2 medium supplemented with laminin/bFGF/ SHH/FGF8. After 6 days, bFGF was removed to induce the differentiation (stage 5) in N^2 medium supplemented with laminin and ascorbic acid for 6 days. Three independent biological replicates were taken at four stages of dopaminergic differentiation.

For random differentiation, EBs were dissociated and plated onto a tissue culture dish in DMEM with fetal bovine serum and antibiotics for indicated periods.For random differentiation model, three biological replicates were made at day 4 (stage 2), 8 (stage 3), 15 (stage 4), and 21 (stage 5) to extract total RNA.

Total RNAs from undifferentiated mouse ES cells were used as a reference group in all the experiments. Total RNA was prepared by using TriZol reagent (Invitrogen, Calsbard, CA). The array used in this experiment was the Macrogen Mouse Oligo 11K Chip (Macrogen, Seoul, Korea) as described previously (Park *et al.*, 2002; Kim *et al.*, 2004). Cy3 and Cy5 fluorescent intensities were determined using the GenePix scanner (Axon Instruments, Union City, CA), and images were analyzed using the GenePix Pro to calculate relative ratios and to determine confidence intervals.

Data preprocessing

Fluorescent intensity data was imported to an in-house microarray database. Variance stabilizing normalization by Huber *et al.* was applied with the 'vsn'package in Bioconductor using the R statistical package (Huber *et al.*, 2002). After performing intensity-dependent global LOWESS regression, spatial and intensity dependent effects were managed by pin-group LOWESS normalization following by the approach of Yang *et al.* (Yang *et al.*, 2000).

The gene expression dataset was registered to the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/ geo/) under the accession numbers, GSE3527 for random differentiation and GSE3528 for guided differentiation. In this study, 18 slides which consist of stage 3, 4, and 5 expression profiles of GD and RD with biological triplicates are used, because stage 2 was common state between them.

Feature Selection

When analyzing expression profiles using machine learning methods, one of the most important processes is feature (gene) selection for the target classes. Although there are huge amount of genes in a microarray, the number of genes that show strong correlation with a certain class is

Table 1. Top 30 genes among 79 biomarkers

Rank	Genbank id	Title	RD	GD
1	NM_016887	claudin 7; cldn7	0.67±0.33	-0.39±0.16
2	NM_009692	apolipoprotein a-i; apoa1	2.40±0.83	0.15±0.11
3	NM_018816	apolipoprotein m; apom	1.87±0.65	0.01±0.22
4	NM_019714	b lymphocyte gene 1; bce1-pending	-0.35±0.20	0.48±0.25
5	NM_010016	decay accelerating factor 1; daf1	1.24±0.47	0.00±0.22
6	NM_023065	interferon gamma inducible protein 30; ifi30	0.42±0.22	-0.31±0.20
7	NM_009695	apolipoprotein cii; apoc2	2.39±1.06	-0.05±0.21
8	NM_008655	growth arrest and dna-damage-inducible 45 beta; gadd45b	0.30±0.29	-0.57±0.23
9	AF426411	zinc ring finger-containing protein grail	0.91±0.44	-0.24±0.24
10	NM_010421	hexosaminidase a; hexa	0.46±0.18	-0.09±0.10
11	NM_007423	alpha fetoprotein; afp	2.96±1.20	0.22±0.31
12	AK016614	homolog to human melanoma-associated antigen 10 (mage-10 antigen)	1.66±0.73	-0.20±0.41
13	NM_010664	keratin complex 1, acidic, gene 18; krt1-18	1.63±0.52	-0.41±0.76
14	BC019790	similar to hypothetical protein flj14779	-0.42±0.12	0.12±0.17
15	AK013489	homolog to alanine:glyoxylate aminotransferase 2 homolog 1, splice form 1	0.52±0.16	-0.23±0.28
16	AK018713	cyba	0.56±0.30	-0.29±0.22
17	NM_016697	glypican 3; gpc3	1.30±0.50	0.15±0.21
18	NM_021291	glycoprotein-associated amino acid transporter b0,+at1; slc7a9	1.55±0.81	-0.29±0.38
19	AK011614	cdna clone transforming growth factor beta regulated gene1	0.51±0.16	-0.20±0.28
20	AF440692	transferrin	2.14±1.24	-0.51±0.38
21	NM_008663	myosin viia; myo7a	0.33±0.24	-0.40±0.23
22	NM_133255	similar to hook2 protein; mgc28586	-0.18±0.11	-0.63±0.14
23	NM_010360	glutathione-s-transferase, mu 5; gstm5	0.05±0.19	0.59±0.12
24	NM_018777	claudin 6; cldn6	0.90±0.40	-0.23±0.38
25	NM_008642	microsomal triglyceride transfer protein; mttp	1.13±0.53	-0.13±0.31
26	NM_007469	apolipoprotein ci; apoc1	0.49±0.21	-0.76±0.58
27	NM_010902	nuclear, factor, erythroid derived 2, like 2; nfe2l2	0.26±0.13	-0.35±0.25
28	NM_008906	protective protein for beta-galactosidase; ppgb	0.70±0.34	-0.05±0.13
29	X12905	properdin (aa 5 - 441)	0.74±0.29	0.07±0.15
30	NM_015775	transmembrane protease, serine 2; tmprss2	1.20±0.61	-0.11±0.30

relatively small. Hence the selection of the most relevant and informative genes for target classes is important. Good feature set consists of those highly correlated with a class but are uncorrelated with other classes.

Since support vector machine algorithm doesn't embedfeature selection mechanism, informative features should be selected before the data mining algorithm is run, using some approach that is independent of the data mining task. To select feature set with discriminative power, various feature selection methods have been used, such as information gain, Gini index, and statistical tests. In this study, multivariate permutation test with t-statistics was used to identify discriminative genes between RD and GD.

There are a variety of benefits to feature selection (Tan *et al*., 2005). Many data mining algorithms work better if the number of attributes in the data is lower, because feature selection can eliminate irrelevant features and reduce noise. Moreover, feature selection can lead to a more understandable model because the model may involve fewer attributes. Finally, the amount of time and memory required by the data mining algorithm is reduced with feature selection.

Support vector machine

Support vector machines (SVMs) which are supervised machine learning techniques, have exhibited superb performance in binary classification tasks (Tan *et al*., 2005; Vapnik, 1998; Christianini *et al*., 2002). Due to robust performance, SMVs have been adopted in a number of applications from text categorization to bioinformatics problems. Intuitively, SVMs aim at searching for a hyperplane that separates the two classes of data with the largest margin between the hyperplane and the point closest to it.

We used a sequential minimal optimization (SMO) algorithm with a logistic regression model and RBF kernel for training a support vector classifier (Christianini *et al*., 2002 ; Platt, 1998; Keerthi *et al*., 2001 ; Witten *et al*., 2005). The SMO algorithm was derived by taking the idea of the decompositionmethod to its extreme and optimizing a minimal subset of just two points at each iteration. The power of this technique resided in the fact that the optimization problem for two data points admitted an analytical solution, eliminating the need to use an iterative quadratic optimizer as part of the algorithm.

To evaluate performance of the prediction model, we trained support vector classifier based on Leave-One-Out Cross-Validation (LOOCV). LOOCV is simple n-fold cross-validation, where *n* is the number of samples in the dataset (Tan *et al*., 2005 ; Witten *et al*., 2005). Each sample in turn was left out, and the SVM algorithm trained all the remaining samples. It was judged by its correctness on the remaining instance – one or zero for success or failure, respectively. The results of all *n* judgments, one for each member of the dataset, were averaged, and that average represented the final error estimate. LOOCV had the advantage of utilizing as much data as possible for training. In addition, the testsets were mutually exclusive and they effectively covered the entire dataset. The drawback of this approach was that it was computationally expensive to repeat the procedure n times. Nevertheless, LOOCV seemed to offer a chance of squeezing the maximum out of a small dataset and obtaining as accurate an estimate as possible (Tan *et al*., 2005).

Results

Differentially expressed genes

Multivariate permutation test was applied to evaluate statistical significance of changes in the gene expression between GD and RD. Qualite-quantile plot analysis showed that the residual quantiles were deviated from the theoretic quantiles. Hence we regarded the distribution of the data as abnormal (Wilk *et al*., 1968). Typical T-test assumes that the input dataholds normal distribution. Therefore we used the multivariate permutation test to collect the genes at a 95% confidence with the false discovery rate of less than 5% (Korn *et al*., 2004). The test statistics used were random variance version of T-test for each gene (Wright *et al*., 2003). Although T-test was used, the multivariate permutation test was non-parametric and did not require the assumption of normal distributions. By random variance T-test analysis, we selected 203 differentially expressed genes (DEGs) between GD and RD.

Construction of the optimal prediction model

When we constructed prediction model using support vector machine based on weka environment with aforementioned 203 discriminative genes, the model could predict the target class perfectly by LOOCV (Wright *et al*., 2003). Since the number of discriminative genes was relatively numerous, we found the optimal set of marker genes with a perfect accuracy. A set of experiments were conducted using support vector machine by decreasing the number of the genes which were sorted by their p-value. When the number of selected genes was less than 79, one RD example was predicted incorrectly as GD (94.4% accuracy). Therefore, we selected 79 genes as a marker gene set to construct the optimal prediction model (Table 1).

Gene ontology analyses

To characterize the 79 marker genes, we analyzed gene

ontology (GO) categories of them in GD and RD (The Gene Ontology Consortium, 2000). This procedure finds gene ontology categories that have higher than expected number of genes differentially expressed among the different classes of the samples based on random variance T-tests (Wright and Simon, 2003). Gene ontology analyses were performed based on BRB-arraytools version 3.2.3. Then we arranged significant GO categories into molecular function, biological process, and cellular component which are presented to Table 2, Table 3, and Table 4 respectively.

Interestingly in the aspect of molecular function, the genes related with lipid transporter activity, structural molecular activity, peptidase activity, DNA binding, hydrolase activity, and protein binding showed a significant variance (Table 2).

Discussion

Though we constructed the optimal prediction model with perfect accuracy, the number of training samples was relatively small. Dataset was made up of triplicates biological replicates for each of the 3 differentiation stages of GD and RD. To avoid the problem of overfitting and bias, we evaluated our prediction model based on LOOCV. LOOCV makes prediction model more robust because the

Table 2. Gene ontology analysis – molecular function

GO category	GO description	# of genes	P-value	Genbank id
5319	lipid transporter activity	5	0.004	NM_007469, NM_009692, NM_009695, NM_009696, NM_018816
5198	structural molecule activity	6	0.178	NM_008471, NM_009449, NM_010664, NM_016887, NM_018777, NM_031170
8233	peptidase activity	6	0.278	NM_008198, NM_008906, AF426411, NM_011670, NM_015775, NM_016907
3677	DNA binding	5	0.352	AK011614, BC019790, NM_010193, NM_010902, U49507
16787	hydrolase activity	12	0.364	NM_007423, NM_008198, NM_008548, NM_008906, NM_009449, AF426411, NM_010421, NM_011670, NM_015775, NM_016907, NM_023587, AK008492
16740	transferase activity	5	0.409	AK013489, NM_009177, NM_010360, NM_011479, NM_031170
5515	protein binding	13	0.437	BC006588, NM_008655, NM_008663, NM_008906, NM_009031, NM_009692, NM_010217, NM_011503, NM_011670, NM_011804, NM_021344, NM_023653, NM_133255

Table 3. Gene ontology analysis – biological process

GO category	GO description	# of genes	P-value	Genbank id
50874	organismal physiological process	10	0.014	BC006588, NM_007423, NM_008198, NM_008663, NM_009692, AF426411, NM_010016, NM_010217, NM_023065, X12905
6955	immune response	5	0.037	NM_008198, AF426411, NM_010016, NM_023065, X12905
51234	establishment of localization	17	0.182	AK018713, BC006588, NM_007423, NM_007469, NM_007819, NM_008625, NM_009449, NM_009692, NM_009695, NM_009696, NM_010217, NM_010240, AF440692, NM_011503, NM_018816, NM_021291, NM_026536
50789	regulation of biological process	12	0.188	BC019790, NM_007709, NM_008655, NM_009031, NM_009692, AF426411, NM_010193, NM_010217, NM_010902, NM_011276, NM_011804, NM_016697
50791	regulation of physiological process	11	0.236	BC019790, NM_007709, NM_008655, NM_009031, NM_009692, AF426411, NM_010193, NM_010217, NM_010902, NM_011276, NM_011804
6508	proteolysis	5	0.268	NM_008198, NM_008906, AF426411, NM_011670, NM_015775
6950	response to stress	5	0.290	NM_008198, NM_010016, NM_010497, NM_031170, X12905
7010	cytoskeleton organization, biogenesis	6	0.455	NM_008471, NM_008663, NM_009449, NM_010664, NM_031170, NM_133255

Table 4. Gene ontology analysis – cellular component

GO category	GO description	# of genes	P-value	Genbank id
5886	plasma membrane	6	0.195	NM_008471, NM_016697, NM_016887, NM_018777, NM_025278, NM_031170
16021	integral to membrane	18	0.198	AK018713, BC027328, NM_008548, NM_008625, NM_009177, AF426411, NM_010016, NM_010330, NM_010941, NM_011479, NM_015775, NM_016697, NM_016887, NM_016907, NM_018777, NM_021291, NM_023587, NM_033603
43231	intracellular membrane-bound organelle	23	0.234	AK011614, AK013489, AK018713, BC007174, BC019790, NM_007709, NM_007819, NM_008131, NM_008548, NM_008655, NM_008906, NM_009031, NM_009177, AF426411, NM_010421, AF440692, NM_010664, NM_010902, NM_011479, NM_011804, NM_013602, NM_023065, NM_026536
5739	mitochondrion	6	0.276	AK013489, AK018713, NM_008131, NM_008906, NM_010664, NM_026536
5634	nucleus	8	0.364	AK011614, BC007174, BC019790, NM_007709, NM_008655, NM_009031, NM_010902, NM_011804
44444	cytoplasmic part	18	0.37414	AK013489, AK018713, NM_007819, NM_008131, NM_008471, NM_008548, NM_008906, NM_009177, AF426411, NM_010421, NM_010497, AF440692, NM_010664, NM_011479, NM_013602, NM_023065, NM_026536, NM_031170
43232	intracellular non-membrane-bound organelle	6	0.45506	NM_008471, NM_008663, NM_009449, NM_010664, NM_031170, NM_133255

greatest possible amount of data is used for training in each case, which presumably increases the chance that the classifier is an accurate one.

Since each GD and RD class had various samples which were taken at different stages, there were some variations among samples in each class according to the degree of differentiation. Owing to the principle of feature selection, uninformative genes which showed variation of expression in a classwere filtered out. Through the feature selection process, most of genes whose expression profiles had dynamical changes according to the degree of differentiation were excluded from training prediction model. Hence, the dynamics of gene expression according to differentiation stages was not considered in our prediction model. On the other hand, the characteristic of feature selection made it easy to detect neural fates from random differentiation regardless of differentiation stages.

The results of gene ontology analysis helped interpretation of the characteristics of 79 biomarkers. This analysis was different than annotating a gene list using GO categories. The GO analysis computed statistically significant GO categories for given gene list.

We have presented a method to detect neural fates from random differentiation of mouse stem cells based on gene expression profiles. Proposed prediction model which were constructed based on effective statistical feature selection method and support vector machine algorithm performed. We investigated the compact set of biomarkers with perfect prediction accuracy. The 79 informative genes were selected as novel candidates of biomarkers for neural fates.

The dataset currently available contains relatively few training examples and a more robust prediction model needs to be constructed using plenty of microarray experiments to put our study to practical use.

Acknowledgement

This work was supported by grant to W.-Y. Park from Korea Food and Drug Administration (06940-034, KFDA2006-7100, 2006) and from the Stem Cell Research Center, KOREA (SC11021, 2003).

References

Sasai, Y. (1998). Identifying the missing links: genes that connect neural induction and primary neurogenesis in vertebrate embryos. *Neuron*. 21, 455-458.

Czyz, J. and Wobus A. (2001). Embryonic stem cell di ffe-rentiation: the role of extracellular factors. *Differen- tiation*. 68, 167-174.

Lee, M.S., Jun, D.H., Hwang, C.I., Park, S.S., Kang, J.J., Park, H.S., Kim, J., Kim, J.H., Seo, J.S., and Park, W.Y. (2006). Selection of neural differentiation-specific genes by comparing profiles of random differentiation. *Stem Cells* 24, 1946-1955.

Lee, S.H., Lumelsky, N., Studer, L., Auerbach, J.M., and McKay, R.D. (2000). Efficient generation of midbrain and hindbrain neurons from mouse embryonic stem cells. *Na.t Biotechnol*. 18, 675-679.

Park, W.Y., Hwang, C.I., Im, C.N., Kang, M.J., Woo, J.H., Kim, J.H., Kim, Y.S., Kim, J.H., Kim, H., Kim, K.A., Yu, H.J., Lee, S.J., Lee, Y.S., and Seo, J.S. (2002). Identification of radiation-specific responses from gene expression profile. *Oncogene* 21, 8521-8528.

Kim, J.H., Ha, I.S., Hwang, C.I., Lee, Y.J., Kim, J., Yang, S.H., Kim, Y.S., Cao, Y.A., Choi, S., and Park, W.Y. (2004). Gene expression profiling of anti-GBM glomerulonephritis model: the role of NF-kappaB in immune complex kidney disease. *Kidney Int*. 66, 1826-1837.

Huber, W., von Heydebreck, A., Sultmann, H., Poustkam, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18(Suppl 1), S96-S104.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*. 30, e15.

Tan, P.N., Stenbach, M., and Kumar, V. (2005). *Introduction to Data Mining*, Addison Wesley.

Vapnik, V.N. (1998). *Statistical Learning Theory*, Wiley, New York, NY.

Christianini, N. and Shawe-Taylor, J. (2002). *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization, *Advances in kernel methods-support vector learning*. MIT Press, Boston.

Keerthi, S.S., Shevade, S.K., Bahattacharyya, C., and Murthy, K.R.K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*. 13, 637-649.

Witten, I.H. and Frank E. (2005). *Data mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann.

Wilk, M.B. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika* 55:1-17.

Korn, E.L., Troendle, J.F., McShane, L.M., and Simon, R. (2004). Controlling the number of false discoveries: applications to high-dimensional genomic data. *J. Statist. Planng Inf*. 379-398.

Wright, G.W. and Simon R. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*. 19, 2448-2455.

The Gene Ontology Consortium. (2000). Gene Ontology: tool for the unification of biology. *Nat Genet*. 25,