Contents lists available at ScienceDirect

Regenerative Therapy

journal homepage: http://www.elsevier.com/locate/reth

Original Article

High-precision multiclass cell classification by supervised machine learning on lectin microarray data

Mayu Shibata [a, b], Kohji Okamura [c], Kei Yura [b, d], Akihiro Umezawa [a, *]

[a] Department of Reproductive Biology, National Center for Child Health and Development, Tokyo, 157-8535, Japan
[b] Graduate School of Humanities and Sciences, Ochanomizu University, Tokyo, 112-8610, Japan
[c] Department of Systems BioMedicine, National Center for Child Health and Development, Tokyo, 157-8535, Japan
[d] School of Advanced Science and Engineering, Waseda University, Tokyo, 162-0041, Japan

ARTICLE INFO

Article history:
Received 16 September 2020
Accepted 25 September 2020

Keywords:
Lectin microarray
Linear classification
Neural network
Artificial intelligence
Pluripotent stem cells

ABSTRACT

Introduction: Establishment of a cell classification platform for evaluation and selection of human pluripotent stem cells (hPSCs) is of great importance to assure the efficacy and safety of cell-based therapy. In our previous work, we introduced a discriminant function that evaluates pluripotency from the cells' glycome. However, it is not yet suitable for general use.
Methods: The current study aims to establish a high-precision cell classification platform introducing supervised machine learning and test the platform on glycome analysis as a proof-of-concept study. We employed linear classification and neural network to the lectin microarray data from 1577 human cells and categorized them into five classes including hPSCs.
Results: The linear-classification-based model and the neural-network-based model successfully predicted the sample type with accuracies of 89% and 97%, respectively.
Conclusions: Because of the high recognition accuracies and the small amount of computing resources required for these analyses, our platform can be a high precision conventional cell classification system for hPSCs.

1. Introduction

Human pluripotent stem cells (hPSCs) such as induced pluripotent stem cells [1,2] and embryonic stem cells [3] play a central role in regenerative medicine, due to their pluripotency of differentiation and unlimited self-renewal abilities [4]. A number of researchers have developed technologies using hPSCs that have advanced to clinical trials [5–10]. However, it is still a challenge to evaluate the characteristics of hPSCs used for treatments [11]. Establishment of a high-precision cell classification method is of great importance in order to define criteria for the selection of hPSCs and to assure the efficacy and safety of cell-based treatments.

Lectin microarrays are recognized as powerful analytical platforms for understanding the type and condition of cells [12]. This is a technology that enables rapid quantitative analysis of the glycome, the whole glycan profile of a sample [13,14]. In this method, glycans of glycoproteins are captured with lectins, which are proteins that specifically interact with glycans. Lectin microarray utilizes a series of lectins to enable the simultaneous detection of multiple kinds of glycans. Owing to their sensitivity and high-throughput performance [15], lectin microarrays have been used to uncover biomarkers or typical glycome patterns of the cells or glycoproteins of interest [16–23]. Such research includes our prior studies on endometrial cancer cells [24] and hPSCs [25]. The latter work describes two lectins that act as markers of pluripotency and generated a discriminant tool that can be used to evaluate the pluripotency of cells based on lectin microarray data. Although it achieved 100% accuracy, it is not practical for three reasons: (i) the tool was derived and tested with a small number of samples (less than 100); (ii) the test samples were composed of a few types of cells; and (iii) the tool depended on data from only a few lectins. Despite the uncertainty of usefulness of this preliminary tool, the

[*] Corresponding author. Center for Regenerative Medicine National Center for Child Health and Development Research Institute, 2-10-1 Okura, Setagaya, Tokyo, 157-8535, Japan. Fax: +81 3 5494 7048.
E-mail address: umezawa@1985.jukuin.keio.ac.jp (A. Umezawa).
Peer review under responsibility of the Japanese Society for Regenerative Medicine.

https://doi.org/10.1016/j.reth.2020.09.005

work indicates that the glycome has a potential to be a powerful information source for building a cell classification platform.

Machine learning is effective for data analysis in many aspects of modern medicine, including diagnosis and prediction of outcome of patients [26–31]. In this study, we employed supervised machine learning method, one of the machine learning methods that analyzes labeled data. Supervised machine learning method allows the model to optimize parameters that define decision boundaries through repetitive training of the dataset. Two commonly used methods of supervised machine learning, linear classification and neural network, were adopted to our cell classification models. One of the most significant differences in these algorithms is the flexibility of a decision boundary. The former represents the boundary as a linear combination of features, whereas the latter does so in a non-linear form. This enables neural-network-based model to form a more complex boundary than that of the linear-classification-based model. In this study, we demonstrate that an analytical platform of supervised machine learning on lectin microarray data exhibited a high capability for multiclass cell classification by cell types. Our study provides an answer to one of the fundamental questions for the establishment of an evaluation system of hPSCs: Which analyses can perform cell classification effectively? The results of this study may pave the way to a wide application of hPSCs to cell-based treatments by enriching the foundational knowledge in constructing cell classification system of the hPSCs.

2. Results

Our dataset consists of lectin microarray data from 1577 human cell samples with 45 different lectins (Table S1). Each sample was manually annotated as one of the five classes: pluripotent stem cells, mesenchymal stromal cells, endometrial and ovarian cancer cells, cervical cancer cells, or endometrial cells. After pre-processing of the dataset, we performed principal component analysis (PCA), an unsupervised machine learning method, to visualize the datasets. Then we designed supervised machine learning models with a linear classification and a neural network. Furthermore, we extracted weight coefficients of the lectins in the decision boundaries from linear-classification-based classifiers, i.e. trained models.

2.1. Data visualization by PCA

PCA was performed on 12 different datasets that were generated by the unique combinations of four pre-processing methods (see Experimental Procedures). None of the PCA plots showed clear clusters by class (Fig. S1), including those of the raw dataset (dataset A) and one of the pre-processed datasets (dataset H) (Fig. 1). Dataset H was subjected to the minmax-normalization of the fluorescent values among the samples to remove the experimental bias and among the features to uniform the value range. Dataset H was adopted for further analyses because i) its pre-processing method included data corrections which were important for effective machine learning and ii) its PCA plot was less biased than others.

2.2. Cell classification by supervised machine learning methods

Linear classification and neural network are two commonly used supervised machine learning methods. For both algorithms, we first optimized hyper-parameters, which cannot be tuned by the machine learning model itself, and then generated classifiers (see Experimental Procedures). All of the supervised machine learning analyses for classification employed leave-one-out cross-validation to maximize the number of training samples within the limited data. In this method, one of the samples was assigned as the test sample and the model was trained on the rest of the samples, hence the classifiers were generated as many as the total number of the samples [32]. The recognition accuracy, an index of the recognition ability of the model, was calculated by dividing the number of correct predictions (the sum of the number of true positives and true negatives) by the number of the samples (1,577).

Two hyper-parameters, regularization weight and the number of epochs, were optimized for linear-classification-based model. The former is involved in how much the parameters are updated in each iteration and the latter defines the number of repetitions of learning. The overall recognition accuracy, which is the recognition accuracy for all the samples, showed two peaks (88.7%) at regularization weights 3 and 30 (Fig. 2A). Since a small regularization weight slows the learning compared to a large one, the recognition accuracy of the model with a smaller regularization weight is considered to have more room for improvement when the numbers of epochs are the same. Therefore, regularization weight was set to 3 for further analyses. With the regularization weight at 3, the overall recognition accuracy reached its highest score (89.0%) at 240 epochs (Fig. 2B). With these hyper-parameters (regularization weight: 3, the number of epochs: 240), the overall recognition accuracy was 89.3 ± 0.1% (standard error of the mean). The recognition accuracy of the best predicted class (mesenchymal stromal cell) was 97.7 ± 0.1% and that of the worst predicted class (endometrial and ovarian cancer cell) was 74.8 ± 0.2% (Table 1). In order to test these recognition accuracies were not achieved by chance, the prediction was performed in the same protocol on a shuffled dataset, the dataset of which labels were reassigned randomly. The overall recognition accuracy of the model on the shuffled dataset was 26.8 ± 0.2% (Table S2) and this confirmed that the recognition

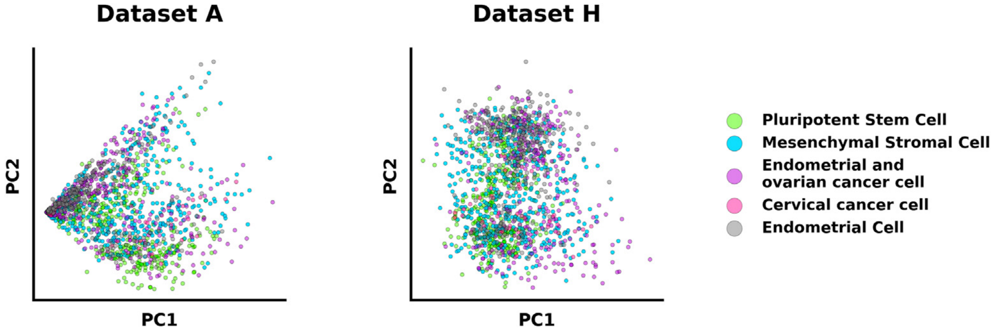

Fig. 1. PCA plots of dataset A (the raw dataset) and dataset H (the dataset subjected to correction of the fluorescent values among samples and probes). Cumulative contribution ratios at PC2 were 0.87 and 0.41, respectively. See also Fig. S1.

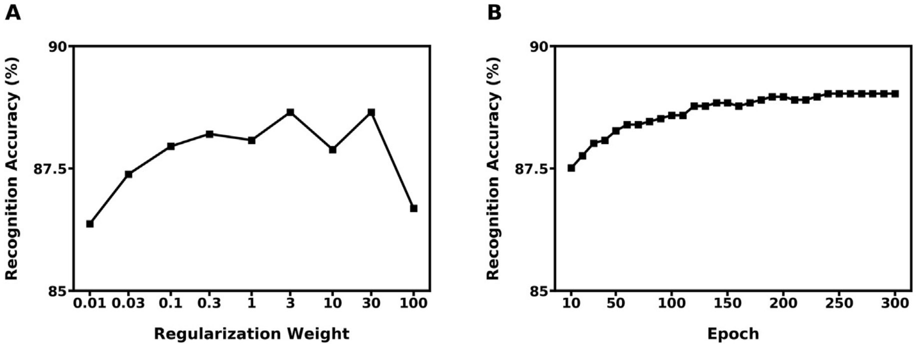

Fig. 2. Hyper-parameter optimization of the linear-classification-based model (A) Regularization weight (B) Number of epochs.

Table 1
Recognition accuracy of linear-classification-based classifiers.

Class	Number of samples	Recognition accuracy [%]
Pluripotent stem cell	391	92.7 ± 0.1
Mesenchymal stromal cell	511	97.7 ± 0.1
Endometrial and ovarian cancer cell	313	74.8 ± 0.2
Cervical cancer cell	48	84.0 ± 1.1
Endometrial cell	314	86.7 ± 0.2
Total	1577	89.3 ± 0.1

accuracy of the model on the original dataset was not achieved randomly.

Three hyper-parameters, i.e. the number of hidden layers of a model, the number of nodes in each hidden layer and the number of epochs, were tuned for a neural-network-based model. The first two are the parameters that regulate the complexity of the model. The highest recognition accuracy was achieved by the three-hidden-layer model, however, the two-hidden-layer model was adopted for the further analysis, considering the trade-off between the performance and the complexity of the model. The recognition accuracies of the models with two to five hidden layers were roughly within the same range (95.4%–95.8%) and were higher than that of the single-hidden-layer model (94.7%), therefore, the model was expected to have at least two hidden layers for the best performance. However, the number of hidden layers were encouraged to be kept minimal in order to simplify the model itself and the optimization of the number of nodes in each hidden layer. Hence, the model with two hidden layers were adopted to compromise these requirements. The overall recognition accuracy of this model marked the highest score of 97.1%, when both hidden layers had 300 nodes (Fig. 3B–E). The increase in the complexity of the model accompanied both rise and drop of the overall recognition accuracies within this search area. This implies that complexity contributed to better prediction within the range where the recognition accuracies rose, however, it induced overfitting in the area where the accuracies dropped (Fig. S2). Overfitting is a situation where generalization ability of a model drops due to exceeding the adaptation of the decision boundary to the limited training dataset. It becomes an issue when the flexibility of a model exceeds the need [33]. Taking these observations into consideration, the model with 300 nodes in each hidden layer was selected. The best recognition accuracy of this model was 97.8% at 120 epochs (Fig. 3F). With the optimized hyper-parameters (the number of hidden layers: 2, the number of nodes in each hidden layer: 300, the number of epochs: 120), the model was run three times to test its performance. The overall recognition accuracy of the model reached 97.4 ± 0.2% (Table 2). The best class prediction was for mesenchymal stromal cells, with a recognition accuracy of 98.6 ± 0.2%, and the worst class prediction was for cervical cancer cells, with an accuracy of 95.6 ± 0.5%. The recognition ability of the neural-network-based model was also tested on the shuffled dataset. The overall recognition accuracy on the shuffled dataset was 24.8 ± 0.5% (Table S3). This confirmed that the recognition accuracy of the model on the original dataset was not achieved by chance.

2.3. *Glycome pattern extraction via linear classification*

Weight coefficients of the features were extracted from the decision boundaries to better understand the contribution of each lectin to the coefficients. The regularization weight and the number of epochs were set to 3 and 240, respectively, as previously optimized. All 1577 samples were used as training data in this analysis in order to maximize the number of training samples. Each class demonstrated distinctive patterns of weight coefficients (Fig. 4). The lectins that strongly influenced the prediction of the cells were detected as those with large absolute values of weight coefficients in the decision boundaries.

3. Discussion

Establishment of a high-precision cell classification method is a challenge that calls for solutions to assure efficacy and safety of cell-based treatments through evaluation and selection of hPSCs. As a starter for approaching this challenge, our research proposed that a supervised machine learning platform on lectin microarray data was one of the prospective methods to realize a practical cell classification platform. We demonstrated that an analytical platform that employed supervised machine learning on lectin microarray data exerted high capability in multiclass cell classification. Our linear-classification-based and neural-network-based models predicted the samples into one of the five classes with high recognition accuracies of 89% and 97%, respectively. These results show that both of the supervised learning models succeeded in capturing the difference of the data distribution of the classes in multi-dimensional space, whereas PCA, one of the widely-used

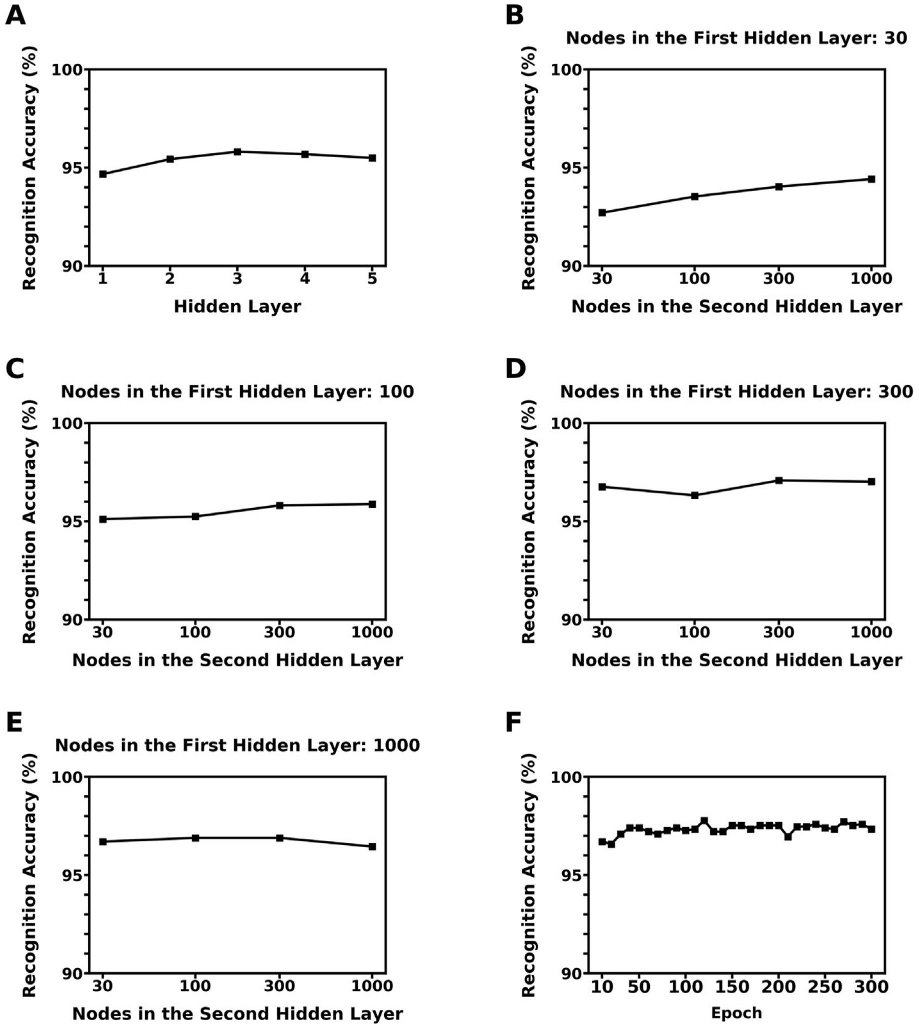

Fig. 3. Hyper-parameter optimization of the neural-network-based model (A) Number of hidden layer(s) (B–E) Number of nodes in each hidden layer (F) Number of epochs.

Table 2
Recognition accuracy of neural-network-based classifiers.

Class	Number of samples	Recognition accuracy [%]
Pluripotent stem cell	391	97.8 ± 0.2
Mesenchymal stromal cell	511	98.6 ± 0.2
Endometrial and ovarian cancer cell	313	95.6 ± 0.5
Cervical cancer cell	48	96.5 ± 1.5
Endometrial cell	314	96.7 ± 0.2
Total	1577	97.4 ± 0.2

unsupervised machine learning methods to analyze lectin microarray data, failed. Moreover, they suggest that the majority of the data points can be separated by linear hyperplane in a high-dimensional space. Comparing our two supervised machine learning models, the one designed with the neural network performed better (8% higher score in overall recognition accuracy). This suggests that the neural-network-based model succeeded in drawing decision boundaries among the data points including those that were inseparable by linear hyperplanes. This flexibility of the decision boundaries is considered to stem from the nature of the neural network to draw nonlinear decision boundaries.

The high recognition accuracies of the models, especially that of the model based on neural network, satisfied our expectation for this study. This success is partially due to the characteristics of lectin microarray data. The data in this study were derived from 45 lectins, which means the number of the input features for our models was 45. This number is quite small compared with other applications of supervised machine learning where input data with more than thousand features are not uncommon. This small number of features contributed to the high recognition accuracies of our models, though enabling a detailed hyper-parameter optimization by shortening the computational time the models required. This indicates that unlike gene expression microarray data which often have 10,000 probes, lectin microarray yields intriguing biological data to be used with supervised machine learning.

Furthermore, we showed the weight coefficients of lectins to the decision boundary of each class by the linear-classification-based

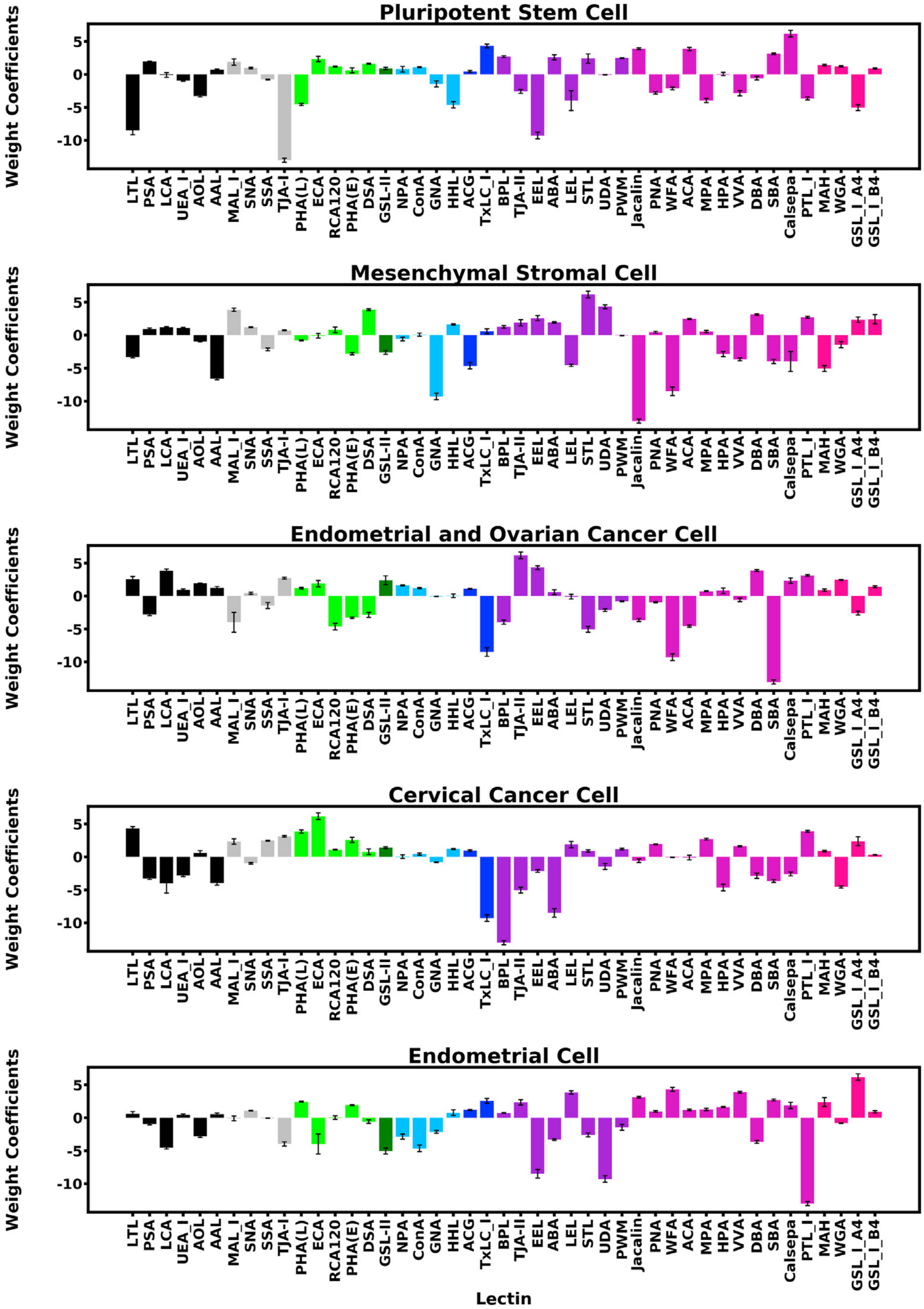

Fig. 4. Weight coefficients of the lectins in each decision boundary drawn by the linear-classification-based classifiers. See also Table S1.

classifiers. Unlike the neural network, a decision boundary from linear classification is expressed as the linear combination of the features. Hence, it is easier to interpret the contributions of each probe to the decision boundaries drawn by the model designed with linear classification. In this study, the model was built for multiclass classification, where decision boundaries are formed to distinguish the class of interest against all the others. The pattern of weight coefficients of each decision boundary was unique, supporting that the five classes in this work were indeed composed of types of cells that had different glycome patterns. The absolute values of the weight coefficients demonstrate how influential the values of the features are for the samples to be predicted as the class of interest. The negative coefficients in the hyperplane mean that the lectins were expressed less in the class than in all the other classes and positive coefficients mean that the lectins were expressed more in the class than the others.

For instance, to the decision boundary between pluripotent stem cells and other four classes of cells, Calsepa, a lectin that binds specifically to mannose and maltose, showed the largest positive coefficient, while TJA-I, whose binding specificity is Siaα2-6Gal/GalNAc, had the largest value in negative coefficients. TJA-I is highly expressed not only in pluripotent stem cells, but also in other types of cells [24,25]. The preprocessing of the dataset used for the classification included normalization of the value range of the fluorescent intensity among lectins. This means that this dataset focuses on the relative differences of the fluorescent values among the cell classes rather than the absolute intensity values of the signals among lectins. In the case of TJA-I, the subtle change in signal contributes significantly to the cell classification. Hence, our interpretation here does not contradict the observation that TJA-I signal is highly expressed in hPSCs as well as other classes of cells.

The high recognition accuracies of the supervised machine learning classifiers supported our expectation that the combination of supervised machine learning and lectin microarray was an effective approach for multiclass cell classification by cell types. However, there are still a large gap between our models in this study and a model for quality evaluation of hPSC-based products in regenerative medicine. A predominant contributor for this gap is lack of information on efficacy and safety. Priority of the future work is to prepare lectin microarray dataset of hPSC-based products including annotation of safety and efficacy and build a supervised machine leaning model to test whether our combinatorial analysis platform can predict safety/efficacy with accuracy as high as that for cell type prediction. We conclude that supervised machine learning analysis on lectin microarray data is a powerful candidate for a high-precision multiclass cell classification system. This study will serve as the first step to assure efficacy and safety of cell-based products by providing the knowledge for one of the effective analysis methods to perform cell classification.

4. Experimental Procedures

4.1. Lectin microarray data and data visualization

The lectin microarray data was derived from 1577 samples and 45 lectins and were retrieved from our previous studies [24,25]. The fluorescent values of each sample were measured on a TIFF file of the microarray chip by bundled software provided by GlycoTechnica. Each sample was manually annotated as one of the five classes: pluripotent stem cell, mesenchymal stromal cell, endometrial and ovarian cancer cell, cervical cancer cell, and endometrial cell (Table S4). The annotated data were then subjected to different pre-processing methods to generate 12 unique datasets (dataset A to L). The pre-processing methods consisted of none or combinations of the following four methods: (i) min–max normalization of a sample (rescale fluorescent values of a sample so that their minimum becomes 0 and maximum becomes 1); (ii) logarithm conversion (take the logarithm of all the fluorescent values to base 10 after substituting 0 by 0.00001); (iii) min–max normalization of a feature (rescale fluorescent values of a feature, a set of the fluorescent values that correspond to one feature for all the samples, to make their minimum 0 and maximum 1); (iv) standardization of a feature (convert fluorescent values of a feature to make their mean 0 and standard deviation 1) (Table S5). The aims of these methods were to correct the differences in fluorescent value distributions that come from using different microarray chips (method (i)), to magnify the difference among small fluorescent values (method (ii)), and to make the data distribution suitable for machine learning (method (iii) and (iv)). The 12 datasets were visualized by scatter plots of PCA (Fig. S1). Scikit-learn (version 0.19.1) [34] was used to perform PCA.

4.2. Cell classification by supervised machine learning

Two supervised machine learning algorithms, linear classification and neural network, were applied to the dataset which underwent the correction of the fluorescent values among samples and features. Hyper-parameters were manually optimized based on the recognition accuracy calculated from the result of a single run. The number of epochs was set to 20 in order to try many different hyper-parameter values with small computational cost. The recognition accuracy of prediction was calculated as the mean recognition accuracies of the triple runs. Each run was performed in a manner of leave-one-out cross-validation and the order of the training samples was randomized in order to prevent biased learning.

Linear classification was performed using Jubatus (version 1.1.1, http://jubat.us/en/). The model adopted the normal herd [35] for learning algorithm. Two hyper-parameters, regularization weight and the number of epochs, were optimized. Regularization weight was tested from 0.01 to 100 (0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100) with the number of epochs set at 20. Then the number of epochs was tested from 10 to 300 by 10. After the hyper-parameters were optimized, the model was trained and its recognition ability was tested. In addition to the cell classification, the weight coefficients of the lectins in each decision boundary were extracted.

Classification of neural-network-based model was performed by Keras (version 2.2.4) [36] using Tensorflow (version 1.14.0) [37] backend. Three hyper-parameters, i.e. the number of hidden layers, the number of nodes in each hidden layer, and the number of epochs, were optimized. The number of hidden layers was tested from 1 to 5 with the number of nodes in each hidden layer(s) 100 and the number of epochs 20. Then, the numbers of nodes were set to be one of these 4 values, 30, 100, 300 and 1000 for each hidden layer (therefore, $4^2 = 16$ patterns of the numbers of nodes in total for the model with 2 hidden layers, the adopted model) with epoch set at 20. Finally, the number of epochs was tested from 10 to 300 by 10. Throughout the analyses, the model consisted of 45 nodes in the input layer and 5 nodes in the output layer, corresponding to the number of features (different type of lectins) and classes (different type of cells), respectively. ReLU was adopted as an activation function for all the layers except the output layer, where the softmax function was used. We employed Adam [38,39] for the optimizer with default parameters (lr = 0.001, beta 1 = 0.9, beta 2 = 0.999, ε = None, decay = 0.0, amsgrad = False) and He normal initializer [40] as kernel weights initializer. In addition, dropout technique [41] was adopted to all hidden layers to prevent the model from overfitting (rate = 0.1). Out of all the training data, validation data accounted for 20% and the batch size was set to be 32. Batch normalization [42] was used for all the layers except the output layer.

After the recognition accuracies of the final model were calculated, those on the shuffled dataset were also tested. In this dataset, the labels of the dataset were shuffled randomly, resulting in only about 23% of all the samples to retain the original annotation. Prediction of the labels was performed in the same protocol: three trials by the optimized model on the shuffled dataset in leave-one-out cross-validation method.

Author contributions

AU conceived the experiments. MS and KO conducted the experiments. MS, KO, KY, and AU discussed the data and manuscript. MS, KO, and AU wrote the text and prepared the figures.

Funding information

This research was supported by grants from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan; by Ministry of Health, Labor and Welfare (MHLW) Sciences research grants. KY was supported by Platform Project for Supporting Drug Discovery and Life Science Research (Basis for

Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number JP20am0101065.

Declaration of competing interest

The authors declare that there is no conflict of interest regarding the work described herein.

Acknowledgments

We express our sincere appreciation to Catherine Ketcham for reviewing the manuscript, Masao Yamada and Koichiro Nishino for fruitful discussion, Yoshihiro Nishijima, Masashi Toyoda and Mayu Yamazaki-Inoue for providing the lectin microarray data. Computational time was provided by the HA8000/RS210 cluster and cluster of HPE ProLiant DL360 Gen 10 at the Center for Regenerative Medicine, National Research Institute for Child Health and Development. This work was supported by National Institute for Child Health and Development Research Institute, Japan.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.reth.2020.09.005.

References

[1] Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 2007;131:861–72.
[2] Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, et al. Induced pluripotent stem cell lines derived from human somatic cells. Science 2007;318:1917–20.
[3] Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, et al. Embryonic stem cell lines derived from human blastocysts. Science 1998;282:1145–7.
[4] Cahan P, Daley GQ. Origins and implications of pluripotent stem cell variability and heterogeneity. Nat Rev Mol Cell Biol 2013;14:357–68.
[5] Schwartz SD, Hubschman J-P, Heilwell G, Franco-Cardenas V, Pan CK, Ostrick RM, et al. Embryonic stem cell trials for macular degeneration: a preliminary report. Lancet 2012;379:713–20.
[6] Ilic D, Devito L, Miere C, Codognotto S. Human embryonic and induced pluripotent stem cells in clinical trials. Br Med Bull 2015;116:19–27.
[7] Trounson A, Thakar RG, Lomax G, Gibbons D. Clinical trials for stem cell therapies. BMC Med 2011;9:52.
[8] Ilic D, Ogilvie C. Concise review: human embryonic stem cells-what have we done? What are we doing? Where are we going? Stem Cell 2017;35:17–25.
[9] Angelos MG, Kaufman DS. Pluripotent stem cell applications for regenerative medicine. Curr Opin Organ Transplant 2015;20:663–70.
[10] Menasché P, Vanneaux V, Hagège A, Bel A, Cholley B, Cacciapuoti I, et al. Human embryonic stem cell-derived cardiac progenitors for severe heart failure treatment: first clinical case report. Eur Heart J 2015;36:2011–7.
[11] Heslop JA, Hammond TG, Santeramo I, Tort Piella A, Hopp I, Zhou J, et al. Concise review: workshop review: understanding and assessing the risks of stem cell-based therapies. Stem Cells Transl Med 2015;4:389–400.
[12] Ribeiro JP, Mahal LK. Dot by dot: analyzing the glycome using lectin microarrays. Curr Opin Chem Biol 2013;17:827–31.
[13] Angeloni S, Ridet JL, Kusy N, Gao H, Crevoisier F, Guinchard S, et al. Glycoprofiling with micro-arrays of glycoconjugates and lectins. Glycobiology 2005;15:31–41.
[14] Pilobello KT, Krishnamoorthy L, Slawek D, Mahal LK. Development of a lectin microarray for the rapid analysis of protein glycopatterns. Chembiochem 2005;6:985–9.
[15] Kuno A, Uchiyama N, Koseki-Kuno S, Ebe Y, Takashima S, Yamada M, et al. Evanescent-field fluorescence-assisted lectin microarray: a new strategy for glycan profiling. Nat Methods 2005;2:851–6.
[16] Sun Y, Cheng L, Gu Y, Xin A, Wu B, Zhou S, et al. A human lectin microarray for sperm surface glycosylation analysis. Mol Cell Proteomics 2016;15:2839–51.
[17] Tateno H, Toyota M, Saito S, Onuma Y, Ito Y, Hiemori K, et al. Glycome diagnosis of human induced pluripotent stem cells using lectin microarray. J Biol Chem 2011;286:20345–53.
[18] Huang W-L, Li Y-G, Lv Y-C, Guan X-H, Ji H-F, Chi B-R. Use of lectin microarray to differentiate gastric cancer from gastric ulcer. World J Gastroenterol 2014;20:5474–82.
[19] Ebe Y, Kuno A, Uchiyama N, Koseki-Kuno S, Yamada M, Sato T, et al. Application of lectin microarray to crude samples: differential glycan profiling of lec mutants. J Biochem 2006;139:323–7.
[20] Kuno A, Itakura Y, Toyoda M, Takahashi Y, Yamada M, Umezawa A, et al. Development of a data-mining system for differential profiling of cell glycoproteins based on lectin microarray. J Proteonomics Bioinf 2008;1:68–72. https://doi.org/10.4172/jpb.1000011.
[21] Krishnamoorthy L, Bess Jr JW, Preston AB, Nagashima K, Mahal LK. HIV-1 and microvesicles from T cells share a common glycome, arguing for a common origin. Nat Chem Biol 2009;5:244–50.
[22] Tao S-C, Li Y, Zhou J, Qian J, Schnaar RL, Zhang Y, et al. Lectin microarrays identify cell-specific and functionally significant cell surface glycan markers. Glycobiology 2008;18:761–9.
[23] Tateno H, Uchiyama N, Kuno A, Togayachi A, Sato T, Narimatsu H, et al. A novel strategy for mammalian cell surface glycome profiling using lectin microarray. Glycobiology 2007;17:1138–46.
[24] Nishijima Y, Toyoda M, Yamazaki-Inoue M, Sugiyama T, Miyazawa M, Muramatsu T, et al. Glycan profiling of endometrial cancers using lectin microarray. Gene Cell 2012;17:826–36.
[25] Toyoda M, Yamazaki-Inoue M, Itakura Y, Kuno A, Ogawa T, Yamada M, et al. Lectin microarray analysis of pluripotent and multipotent stem cells. Gene Cell 2011;16:1–11.
[26] Holder LB, Haque MM, Skinner MK. Machine learning for epigenetics and future medical applications. Epigenetics 2017;12:505–14.
[27] Lee S, Mohr NM, Street WN, Nadkarni P. Machine learning in relation to emergency medicine clinical and operational scenarios: an overview. West J Emerg Med 2019;20:219–27.
[28] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017;12: e0174944.
[29] Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics 2017;37:505–15.
[30] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.
[31] Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep 2018;8:3395.
[32] Bishop CM. Pattern recognition and machine learning. Springer Verlag; 2006.
[33] Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci 2004;44: 1–12.
[34] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[35] Crammer K, Lee DD. Learning via Gaussian herding. Adv Neural Inf Process Syst 2010;23.
[36] Chollet F. keras. Keras; 2015. https://keras.io/.
[37] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv.org 2016.
[38] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv.org 2014.
[39] Reddi SJ, Kale S, Kumar S. On the convergence of Adam and beyond. arXiv.org 2019.
[40] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. arXiv.org 2015.
[41] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.
[42] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv.org 2015.