Supervised machine learning for high-precision cell classification on glycome

Mayu Shibata[1,2], Kohji Okamura[3], Kei Yura[2,4], Akihiro Umezawa[1]
[1]Dept. of Reproductive Biology, NCCHD; [2]Grad. Sch. of Humanities and Sciences, Ochanomizu Univ.; [3]Dept. of Systems BioMedicine, NCCHD;
[4]Sch. of Advance Science and Engineering, Waseda Univ.

グライコームに基づく高精度な細胞分類のための教師あり機械学習

柴田眞侑[1,2]、岡村浩司[3]、由良敬[2,4]、梅澤明弘[1]
[1]成育医療セ・再生医療、[2]お茶大・人間文化創成科学、[3]成育医療セ・システム医学、[4]早大・先進理工

❖ Introduction

Cell classification is an important technology to address the cell evaluation problem of cell-based products used in regenerative therapy. **This study aims to demonstrate that the combination of supervised machine learning and glycome is effective for high-precision cell classification by cell types** as a starting point to address this challenge for the wide-application of regenerative medicine.

❖ Data Preparation

Glycan expressions were quantified as fluorescent values using lectin microarray. The data consisted of 1,577 human cell samples by 45 lectins. Each sample was annotated as one of the 5 classes. Then, the data were subjected to fluorescent value corrections.

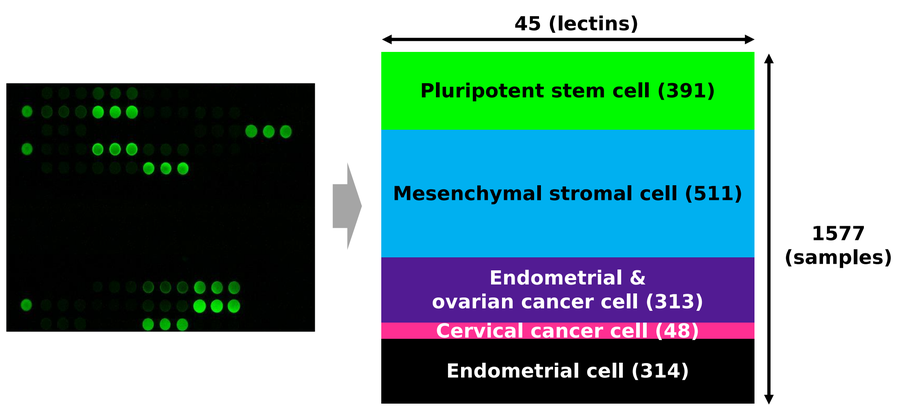
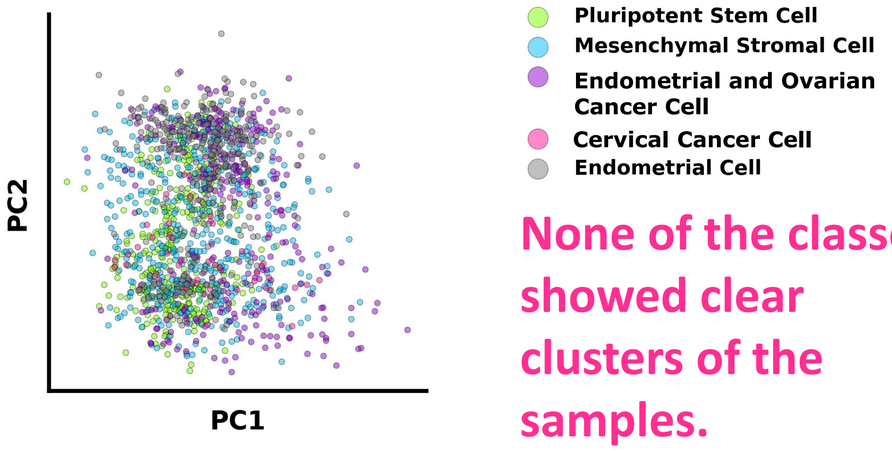


Fig. 1 Data extraction

❖ Data Visualization

The data distribution of the pre-processed dataset was visualized by principal component analysis.



None of the classes showed clear clusters of the samples.

Fig. 1 Principal component analysis

❖ Cell Classification by Supervised Machine Learning

Two widely used supervised machine learning methods, linear classification and neural network was used.

Table 1 Recognition accuracies of linear classification classifiers

Class	Recognition accuracy [%]
Pluripotent stem cell	92.7 ± 0.1
Mesenchymal stromal cell	97.7 ± 0.1
Endometrial and ovarian cancer cell	74.8 ± 0.2
Cervical cancer cell	84.0 ± 1.1
Endometrial cell	86.7 ± 0.2
Total	89.3 ± 0.1

Overall recognition accuracy was about 89%

Table 2 Recognition accuracies of neural network classifiers

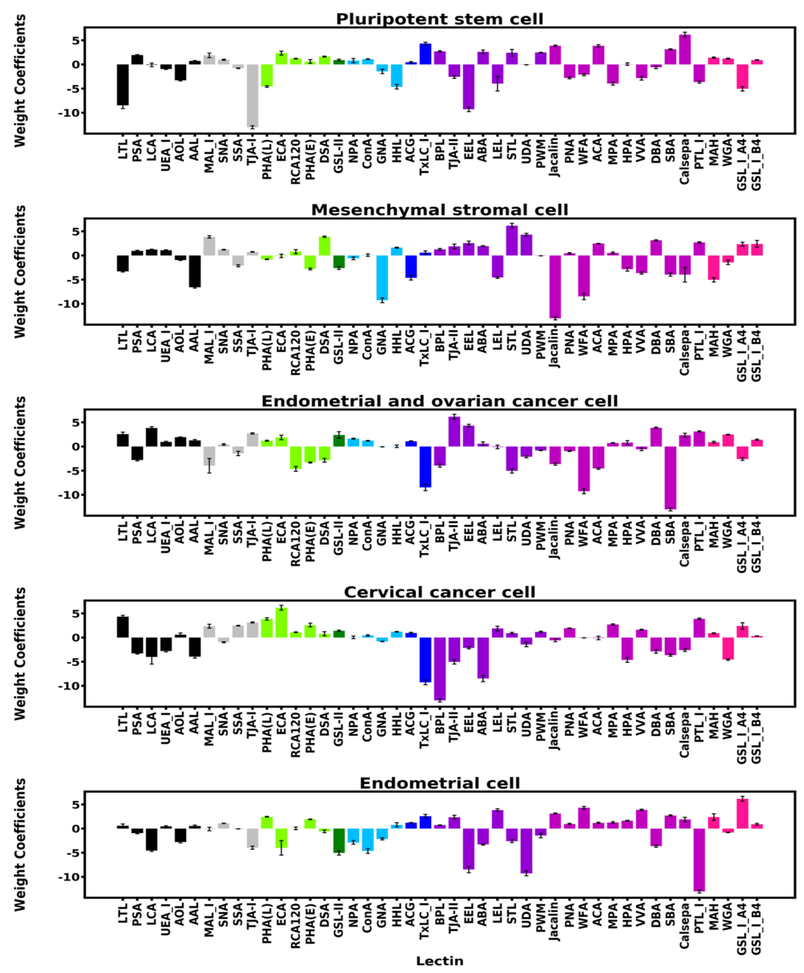
Class	Recognition accuracy [%]
Pluripotent stem cell	97.8 ± 0.2
Mesenchymal stromal cell	98.6 ± 0.2
Endometrial and ovarian cancer cell	95.6 ± 0.5
Cervical cancer cell	96.5 ± 1.5
Endometrial cell	96.7 ± 0.2
Total	97.4 ± 0.2

Overall recognition accuracy was about 97%

The overall recognition accuracy of the neural-network-based classifiers was about 7% higher than the linear classification classifiers.

❖ Decision Boundaries

Weight coefficients of the lectins were extracted from the decision boundaries drawn by linear classification classifiers.



Influential lectins do not necessarily correspond to marker lectins of the class, suggesting that the decision boundaries were defined based on relative differences of the fluorescent values among the classes.

Fig. 2 Weight coefficients of the decision boundaries

❖ Summary

Supervised machine learning and lectin microarray enabled high-precision multiclass cell classification by cell types. Neural-network-based model achieved higher recognition accuracy. Linear-classification-based model was useful in understanding contributions of the lectins to the decision boundaries.