Classification of Single Cell Types During Leukemia Therapy using Artificial Neural Networks

Minjie LYU; Milena RADENKOVIC; Derin B. KESKIN; Vladimir BRUSIC

University of Nottingham Ningbo China, 199 Taikang East Road, Ningbo, 315100, Zhejiang, China.

First published 2021

Classification of Single Cell Types During Leukemia Therapy using Artificial Neural Networks

Minjie LYU
School of Computer Science
University of Nottingham
Ningbo China
minjie.lyu@nottingham.edu.cn

Milena RADENKOVIC
School of Computer Science
University of Nottingham
Nottingham UK
milena.radenkovic@nottingham.ac.uk

Derin B. KESKIN
Dana-Farber Cancer Institute
Harvard Medical School
Boston USA
Derin_Keskin@dfci.harvard.edu

Vladimir BRUSIC
School of Computer Science
University of Nottingham
Ningbo China
vladimir.brusic@nottingham.edu.cn

Abstract—We trained artificial neural network (ANN) models to classify peripheral blood mononuclear cells (PBMC) in chronic lymphoid leukemia (CLL) patients. The classification task was to determine differences in gene expression profiles in PBMC pre-treatment (with ibrutinib) and on days 30, 120, 150, and 280 after the start of treatment. Twelve datasets represented clinical samples containing a total 48,016 single cell profiles were used to train and test ANN models to classify the progress of therapy by gene expression changes. The accuracy of ANN classification was >92% in internal cross-validation. External cross-validation, using independent data sets for training and testing, showed the accuracy of classification of post-treatment PBMCs to more than 80%. To the best of our knowledge, this is the first study that has demonstrated the potential of ANNs with 10x single cell gene expression data for detecting the changes during treatment of CLL.

Keywords—ANN, PBMC, CLL, ibrutinib, scRNAseq, Machine Learning

I. INTRODUCTION

Peripheral blood mononuclear cells (PBMC) represent mixed subpopulations of blood cells. PBMC are composed of 5 cell types: B cells, T cells, Natural Killer (NK) cells, monocytes, and dendritic cells (DC) [1]. The proportions of PBMCs subtypes vary between individuals and change over time. In healthy individuals, the frequencies of PBMC cell types fit within broadly defined ranges: B cells are 5-15%, monocytes are 10-30%, DC are 1-2%, NK are 5-10%, and T cells are 40-70% of total PBMC [1,2]. PBMCs are widely used in the study of the immune system, infectious diseases, and vaccine development. PBMC are used in medical diagnosis for many diseases, including cancer, pulmonary fibrosis, viral hepatitis, and many others [3-6]. PBMC cell types and sub-types have well-defined gene expression profiles that are very similar between healthy individuals. These observations were made both by using bulk sequencing [7,8] and by the study of single cell transcriptomes [9].

Traditional bulk RNA sequencing can only measure the expression value of each gene as an average of expression levels across the whole sample [8]. In contrast, single-cell transcriptome technologies (SCT) measure the expression levels of genes from individual cells. SCT provides a higher resolution of transcriptome measurements than bulk RNA sequencing. However, the profiles of individuals are sparse and cover only a fraction of true gene expression. Among the main challenges in single cell gene expression analysis are the lack of standard data sets, manual annotation of results due to the use of unsupervised machine learning tools, very large expression matrices, and sparsity of data in expression matrices [10].

Chronic Lymphocytic Leukemia (CLL) is a common type of leukemia in adults which usually slowly grows and can remain asymptomatic for years. Approximately 95% of CLL is the malignancy of B cells, while the remainder is of T-cell type. The diagnosis of CLL is based on the identification of abnormal populations of B lymphocytes in the blood, bone marrow, spleen, and lymph nodes [11]. Functional and molecular abnormalities are often characterized by changes in gene expression that varies by CLL subtype and the tissue distribution of CLL cells [11].

Ibrutinib is a small molecule drug that binds protein BTK expressed by B cells, alters cellular signaling that promotes CLL survival and spread, resulting in changes of immune microenvironment in CLL [12,13]. Although CLL is a highly heterogeneous disease, ibrutinib has been demonstrated as an effective treatment with an excellent safety profile and does not cause myelosuppression. In a recent study, SCT gene expression was measured in CLL patients before and during ibrutinib therapy [14]. Data analysis in the ibrutinib study was performed using unsupervised machine learning methods, namely principal component analysis, t-SNE analysis, hierarchical clustering, and differential expression analysis [14].

SCT reveals cellular patterns of individual cells, but current analysis options have limitations. Single cell gene expression analysis by 10x SCT technology [15] produces large matrices. A typical study will produce large matrices that have more than 30,000 rows representing features (genes) and tens of thousands of columns representing individual single cell profiles. These individual profiles are sparse – in any one individual cell profile, 95-99% of features are typically zeros. It means that two nearly identical cells may have very different individual profiles. Current analytical tools mainly focus on unsupervised machine learning methods. These methods do not scale up well as they require progressively higher computer power to analyze data combined from multiple studies. They also do not generalize well – they are not accurate when applied to data sets from other studies [16].

We developed a standardized format of the SCT matrices and deployed supervised machine learning to SCT data. The

SCT data sets of the treatment course from multiple patients were standardized and used to trained artificial neural networks (ANN) to classify PBMC in CLL samples treated by ibrutinib. Several research questions were pursued in this study:

- Can we train an ANN on a set of data extracted from PBMC from unrelated CLL patients and accurately classify whether they were treated by Ibrutinib or not?
- Can we train ANN models to predict the treatment periods?
- Is it possible to generate accurate prediction models without feature selection or dimensionality reduction?

II. MATERIALS AND METHODS

A. Data

Data were extracted from the NCBI Gene Expression Omnibus (GEO) database [15]. The sources of our data are 12 sample IDs GSM3020393, GSM3020394, GSM3020395, GSM3020396, GSM3020397, GSM3020398, GSM3020399, GSM3020400, GSM3020401, GSM3020402, GSM3020403, and GSM3020404. These data consist of expression matrices and metadata reported in [14]. The metadata contains descriptions of samples and experimental conditions. PBMC were obtained from 4 CLL patients (patients 1, 5, 6, and 8) before treatment (day 0) and on days 30 and 120. Patient 5 had a sample collected on day 150 instead of day 120, and patient 6 had an additional sample taken on day 280. We cleaned, labeled, and converted these data sets to a standardized format. For binary classification, we labeled the data sets as pre- and post-treatment. The pre-treatment data sets were from day zero samples (before treatment). The post-treatment data sets were from samples taken on days 30, 120, 150, and 280 (during and after the treatment). For multi-class-classification, we labeled data as 0_day, 30_day, 120_day, 150_day, and 280_day. The number of datasets indicating classes and patients is shown in TABLE I. The total number of cells we used in this study is 48,016; the breakdown of cell numbers by sample collection days are shown in Table II.

TABLE I. THE NUMBER OF DATASETS USED IN THIS STUDY

Cell type	Number of Datasets				
	Patient 1	Patient 5	Patient 6	Patient 8	Total
0 day	1	1	1	1	4
30 day	0	1	1	1	3
120 day	1	0	1	1	3
150 day	0	1	0	0	1
280 day	0	0	1	0	1
Total	2	3	4	3	12

TABLE II. TOTAL NUMBER OF CELLS AVAILABLE FOR THIS STUDY

Cell type	Total Number of Cells				
	Patient 1	Patient 5	Patient 6	Patient 8	Total
0 day	2,775	6,655	3,391	2,186	15,007
30 day	0	7,172	5,771	965	13,908
120 day	3,827	0	2,200	2,330	8,357
150 day	0	6,041	0	0	6,041
280 day	0	0	4,703	0	4,703
Total	6,602	19,868	16,065	5,481	48,016

The genes across all data sets for our analysis were mapped to the genomic build GRCh38 patch release 12 (GRCh38.p12) [17]. Each standardized data set in this study contains 31,217 genes. Sparse matrices have 31,217 rows corresponding to each feature, while the number of columns (single cells) ranges from 965 to 7,172 across all data sets. We divided the data sets into training and testing sets. For train-test cycles 1 and 2, each data set was randomly divided into 10 partitions for internal cross-validation (described later in the text). For the train-test cycle 3, the training and testing sets were combined so that the training and test sets are from different patients.

B. Artificial Neural Networks

In the first part of the study we trained a fully connected feed-forward ANN with 31,217 input units, one 10-nodes hidden layer, and 2 output units for binary classification (pre- and post-treatment) – architecture 31217-10-2. For the second part of the study, we trained a network with architecture 31217-10-5 (multi-class temporal classification). For neural network implementation, we used the multi-layer perceptron classifier MLPClassifier from python library scikit-learn. The activation function for hidden nodes was logistic sigmoid function (Logistic), $f(x) = 1/(1 + exp^{-x})$. The model was trained with a maximum of 300 iterations. A first-order gradient-based optimization Adam algorithm [18] was used for training the network. The initial learning rate was 10^{-3}. Other parameters included early stopping, and 10% of the training data was used. The training stopped when the accuracy of the model assessed by validation data did not improve for > 10 iterations.

C. Study Design

The study design involves three cycles of training and testing designed to answer different research questions and assess the generalization properties of the ANN models. The train-test cycles were:

Cycle 1: Train ANN using data before treatment and data after treatment labeled as 'Pre' and 'Post' and test using 10-fold cross-validation (internal cross-validation).

Cycle 2: Train ANN using data before treatment and data after treatment labeled as '0_day', '30_day', '120_day', '150_day' and '280_day', test using 10-fold cross-validation (internal cross-validation)

Cycle 3: Train ANN using data before treatment, after 30 days of the start of treatment, after 120 days, after 150 days, and after 280 days. Because there is only one data set for days 150, 80% of this data was used for training and 20% for testing. The same procedure was used with 280 days data set.

D. Assessment of Performance

We made confusion matrices and calculated the Precision, Recall, and overall Accuracy using expressions:

$$Accuracy = \frac{Number\ of\ correctly\ classified\ cells}{Total\ number\ of\ cells}$$

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{Precision * Recall}{Precision + Recall}, where$$

TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

III. RESULTS

A. Incremental Training Results

The ANN with the training set was trained using more than 36,000 training instances in each of the three cycles. The training usually took between 70 and 80 epochs (iterations) before stopping. A typical learning curve displaying the changes in loss with respect to the number of epochs is shown in Fig. 1. ANN training cycles showed convergence, usually after 60 to 70 epochs.

B. Internal Cross-Validation

10-fold cross-validation was performed in cycle 1 and cycle 2. The overall internal cross-validation results showed very high accuracy. Cycle 1, the binary classification, had 96.67% of correctly classified instances; 6.2% of post-treatment cells were misclassified as "Day 0" and 2.1% pre-treatment cells were misclassified as not "Day 0", (Fig. 2). Cycle 2 had 95.42% correctly classified instances. In cycle 2, 4.43% of 0_day cells, 3.87% of 30_day cells, 7.16% of 120_day cells, 4.6% of 150_day cells, and 3.44% of 280_day cells were misclassified. The highest misclassified rate was for 0_day (1.8% of experimental 0_day cells classified as 30_day cells), 30_day (2.1% of 30_day cells classified as 120_day cells), 120_day (3.3% of 120_day cells classified as 30_day cells), 150_day (1.8% of 150_day cells classified as 30_day), and 280_day (2.9% of 280_day cells classified as 120_day) (Fig. 3). These results were confirmed by additional classification performance metrics shown in Fig. 4. The best performance determined by internal cross-validation was for days 0 and 280 (F1>0.96 in both cases). The performance for days 30 and 150 was good (F1>0.95). The lowest performance was observed for days 120 (F1=0.92).

The internal cross-validation results indicate that ANN learning is effective when combining multiple treatment periods and cell data sets, even if they were collected from different patients. Those datasets are randomly split and a period cell from the same patient is represented in both training and testing sets; the misclassification rate for any treatment periods is lower than 7%.
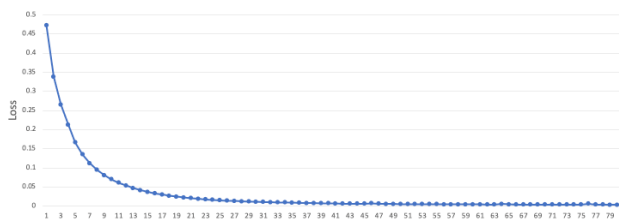

Fig. 1. A representative ANN training cycle in our study. Training stopped after 10 cycles of no improvement of the Loss function.
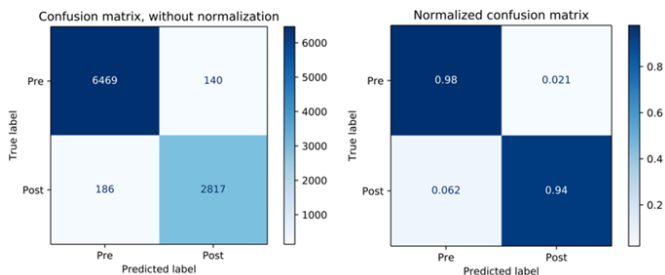

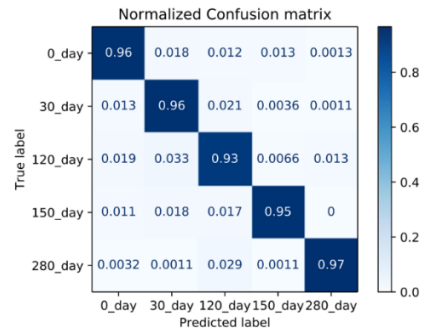Fig. 2. Confusion matrix for Day 0 cycle 1, raw and normalized.

Fig. 3. Normalized confusion matrix for Days 0, 30, 120, 150, and 280, cycle 2.

C. Prospective Validation

After demonstrating that ANN can accurately classify different treatment periods in the training set (not identical to the cell instances in the test set), we explored the generalization ability of training models. The process included using samples from one patient for testing, and samples from all other patients for training.

In cycle 3, we trained ANN using patient 1 (0 day, 120 day), patient 5 (0 day, 30 day, 150 day), patient 6 (30 day, 120 day 280 day) and used patient 6 (0 day), patient 8 (0 day, 30 day, 120 day) for testing.

The same model that could make highly accurate predictions using internal cross-validation (Cycle 2) could not predict previously unseen data sets with similar accuracy as in prospective validation. The accuracy of predictions in Cycle 3 was 67.14% for the ANN with 10 hidden layer nodes, 72.56% for ANN with 20 hidden layer nodes, and 70.9% for ANN with two hidden layers; none of the periods' cells showed useful prediction (150_day and 280_day using internal cross-validation because of the shortage of data sets) (Fig. 4, and Fig. 5). These results are fascinating – the classification results of day 30 shows that 80% of cells are classified as post-treatment cells. Day 120 shows that more than 89% of PBMC are classified as post-treatment.
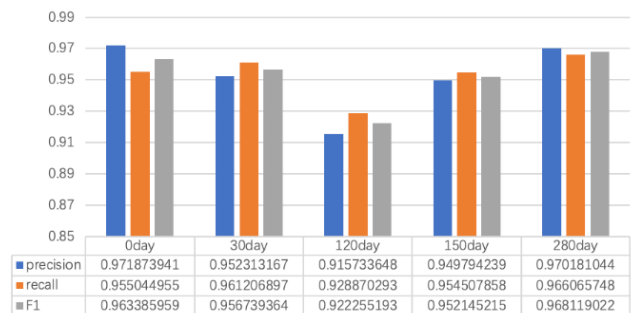

Fig. 4. Results of 5-Class classification for cycle 2, assessed by internal cross-validation.
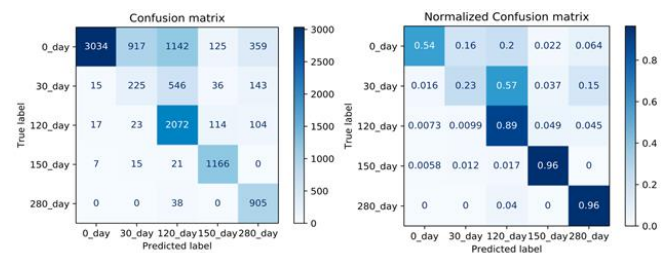

Fig. 5. Confusion matrix in cycle 3. The results for days 150 and 280 are only indicative because the assessment for these days was done using internal cross-validation.

IV. Conclusion and Discussion

We performed a cyclical refinement of ANN models by combining data collected from multiple patients for prediction of treatment periods. We achieved an overall accuracy of classification Acc>92% in internal cross-validation. These results indicate that ANN classification models are useful and show potential for clinical applications. Specifically, these results indicate that it is possible to follow-up an individual patient in a longitudinal study and observe the effects on therapy as early as 30 days after the start of therapy. These results indicate that ANN applied to single cell data can identify changes in PBMC due to the treatment in leukemia patients.

The analysis of predictive models using external cross-validation indicates that ANN models developed on data sets from other patients are also useful as they identify changes due to treatment in >80% of the cells. These observations are significant because they indicate that biological processes resulting from ibrutinib therapy are similar in different patients and can classify PBMC with accuracy that has practical significance, even on previously unseen data. We speculate that it is possible to observe the effect of the therapy even earlier, perhaps after 15 days. The analysis of differences between time points before, during, and after therapy, therefore, may be used to assess both the efficiency and the effectiveness of the therapy. Further analyses using data from more patients are needed to corroborate this hypothesis.

We observed that ANN with 20 hidden layer nodes was more sensitive in detecting post-treatment cells than other architectures. This indicates that the optimization of ANN architecture may be useful in improving the classification of PBMC cells after treatment. This can be useful for the early assessment of the effects of therapy.

In this study, we applied supervised machine learning, using ANN models, to classify data sets from SCT data multiple unrelated patients to predict the changes in PBMC of CLL patients treated by ibrutinib. ANNs were successfully used before in the diagnosis of CLL and other leukemias by using gene expression markers [19,20]. To the best of our knowledge, this is the first study that has demonstrated the usefulness of ANNs for detecting the changes during treatment of CLL with 10x single cell gene expression data.

References

[1] C. R. Kleiveland, "Peripheral blood mononuclear cells". In the Impact of Food Bioactives on Health, pp. 161-167, Springer, Cham, 2015.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] R.A. Shaikh, J. Zhong, M. Lyu, S. Lin, D. Keskin, G.L. Zhang, L. Chitkushev and V. Brusic, "Classification of five cell types from PBMC samples using single cell transcriptomics and artificial neural network", In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2207-2213, IEEE, 2019.

[3] M.J. Baine, S. Chakraborty, L.M. Smith, K. Mallya, A.R. Sasson, R.E. Brand and S.K. Batra, "Transcriptional profiling of peripheral blood mononuclear cells in pancreatic cancer patients identifies novel genes with potential diagnostic utility", PloS One, vol. 6, no. 2, p.e17014, 2011.

[4] W.S. Wang, L.X. Liu, G.P. Li, Y. Chen, C.Y Li, D.Y. Jin and X.L. Wang, "Combined serum CA19-9 and miR-27a-3p in peripheral blood mononuclear cells to diagnose pancreatic cancer", Cancer Prev. Res., vol. 6, no. 4, pp. 331-338, 2013.

[5] M.K. Scott, K. Quinn, Q. Li, R. Carroll, H. Warsinske, F. Vallania, S. Chen, M.A. Carns, K. Aren, et al. "Increased monocyte count as a cellular biomarker for poor outcomes in fibrotic diseases: a retrospective, multicentre cohort study", Lancet Resp. Med., vol. 7, no. 6, pp. 497-508, 2019.

[6] M.K. El-Awady, S.M. Ismail, M. El-Sagheer, Y.A. Sabour, K.S. Amr and E.A. Zaki, "Assay for hepatitis C virus in peripheral blood mononuclear cells enhances sensitivity of diagnosis and monitoring of HCV-associated hepatitis", Clin. Chim. Acta, vol. 283, no. 1-2, pp. 1-14, 1999.

[7] G. Monaco, B. Lee, W. Xu, S. Mustafah, Y.Y. Hwang, C. Carre, N. Burdin, L. Visan, M. Ceccarelli, et al., 2019. "RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types". Cell Rep., vol. 26, no. 6, pp. 1627-1640, 2019.

[8] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, J.L. Rinn and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks", Nat. Protoc., vol. 7, no. 3, pp. 562-578, 2012.

[9] L. Yang, Y. Zhang, Lou Chitkushev, N. Mitic, D. Keskin, G. Zhang and V. Brusic, "Single cell mRNA profiles in PBMC" , In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), this issue, IEEE, 2020.

[10] D. Lähnemann, D. Köster, J. Szczurek, E. McCarthy, D.J. Hicks, S.C. Robinson, M.D. Vallejos, C.A. Campbell, K.R. Beerenwinkel et al., "Eleven grand challenges in single-cell data science", Genome Biol., vol. 21, no, 1, pp.1-35, 2020.

[11] S. Zhang and T.J. Kipps, "The pathogenesis of chronic lymphocytic leukemia", Annu. Rev. Pathol., vol. 9, pp. 103-118, 2014.

[12] E.D. Deeks, "Ibrutinib: a review in chronic lymphocytic leukaemia", Drugs, vol. 77, no. 2, pp. 225-236, 2017.

[13] P. Bachireddy, C. Ennis, V. N. Nguyen, S. H. Gohil, K. Clement, S. A. Shukla, et al., "Distinct evolutionary paths in chronic lymphocytic leukemia during resistance to the graft-versus-leukemia effect", Sci Transl. Med.;12(561):eabb7661, 2020.

[14] A.F. Rendeiro, T. Krausgruber, N. Fortelny, F. Zhao, T. Penz, M. Farlik, Schuster LC, Nemc A, Tasnády S, et al., "Chromatin mapping and single-cell immune profiling define the temporal dynamics of ibrutinib response in CLL. Nat. Commun., 2020 Jan 29;11(1):1-4.

[15] G.X.Y. Zheng, J.M. Terry, P. Belgrader, P. Ryvkin, Z.W. Bent, R. Wilson, S.B. Ziraldo, T.D. Wheeler, G.P. McDermott, et al., "Massively parallel digital transcriptional profiling of single cells". Nat. Commun., vol. 8, no. 1, pp. 1-2, 2017.

[16] W. Hou, Z. Ji, H. Ji and S.C. Hicks, "A systematic evaluation of single-cell RNA-sequencing imputation methods", Genome Biol., vol. 21, 218, 2020.

[17] GRCh38, www.ncbi.nlm.nih.gov/assembly/GCF_000001405.12

[18] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization", arXiv:1412.6980, 2014.

[19] F.S Aghamaleki, B. Mollashahi, M. Nosrati, A. Moradi, M. Sheikhpour and A. Movafagh, "Application of an artificial neural network in the diagnosis of chronic lymphocytic leukemia", Cureus, vol. 11, no 2, 2019.

[20] H.T. Salah, I.N. Muhsen, M.E. Salama, T. Owaidah and S.K. Hashmi, "Machine learning applications in the diagnosis of leukemia: Current trends and future directions", Int. J. Lab. Hematol., vol. 41, no. 6, pp. 717-25, 2019.