BRIGHAM YOUNG UNIVERSITY
DEPARTMENT OF MICROBIOLOGY AND MOLECULAR BIOLOGY

Distinguishing Leukemic Cells Using Fractal Chromatin Patterns and Machine Learning

[1]Abigail Gordhamer, [1]Paul Young, [1]Ryan Cordner
[1]Department of Microbiology and Molecular Biology, Brigham Young University.

PURPOSE

One of the most important tests in the clinical laboratory is the Complete Blood Count, which involves identifying the white blood cells in a patient's blood. The respective counts of the different white blood cell types correlate with various states of health and disease and are critical to diagnosing diseases such as leukemia. Leukemic cells (blasts) are considered especially difficult to distinguish, and it is of the upmost importance that these cells are identified correctly. To aid in the process of leukemic cell identification, we quantified fractal patterns in the chromatin of white blood cells and used the data to identify cells with a random forest algorithm. By distinguishing between cells with the help of a machine learning algorithm, we hope to improve accuracy and efficiency in the clinical laboratory and more easily identify leukemic cells.

METHODS

We compiled image banks of 300-500 images for fifteen types of white blood cells by taking pictures of patient blood samples. We then isolated the nucleus in each image and used a program called TWOMBLI to generate a mask image and high-definition matrix image for each nucleus. From these images, TWOMBLI calculated parameters that indicate fractal patterns in the nucleus such as lacunarity, curvature, branchpoints, endpoints, etc. Using these parameters, we calculated the average values for each cell type and compared those values to one another. Additionally, we ran our data through a random forest algorithm and calculated the accuracy, precision, specificity, and sensitivity from the confusion matrix.

RESULTS

The random forest algorithm was able to identify five different types of leukemic cells with up to 92% accuracy and 90% precision. We also found that the algorithm could distinguish leukemic cells from non-leukemic cells with 97% accuracy and 95% precision. The most important parameters used in the algorithm were endpoints and branch points. A t-test revealed that certain parameters, such as lacunarity and percent high density matrix, have a p-value of 2.4e-25 or lower when compared amongst cell types.

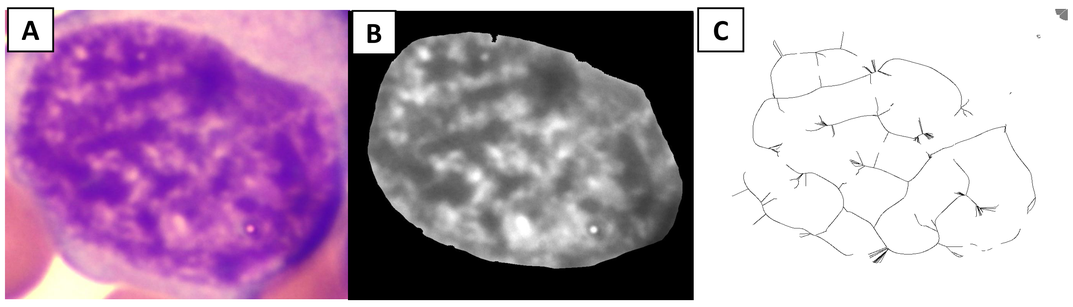

Figure 1. The progression of cell images as they are processed. **A.** A cropped image of a monoblast taken at 100x magnification. **B.** The high density matrix calculated by TWOMBLI and used to calculate parameters such as percent high density matrix. **C.** The image mask generated by TWOMBLI and used to calculate parameters such as branchpoints, endpoints, and lacunarity.

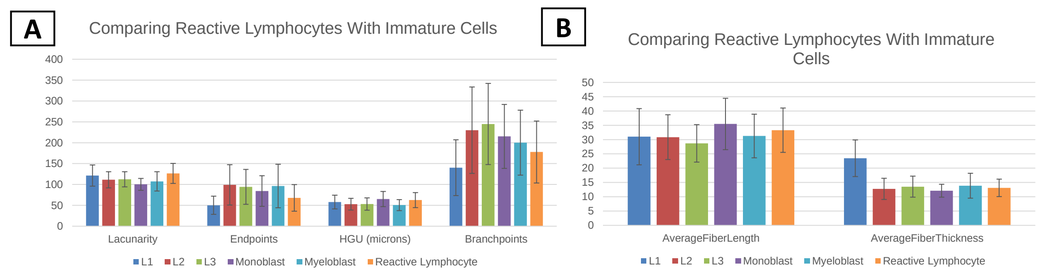

Figure 2. Differentiation of reactive lymphocytes from immature cells. **A.** Reactive lymphocytes are most distinct from L1 lymphoblasts in branchpoints and lacunarity. **B.** Average fiber length was calculated by dividing total fiber length by branchpoints and endpoints. Fiber thickness was calculated by dividing percent high density matrix by total length. Error bars show the standard deviation in both A and B.

	L1	L2	L3	Lymphocyte	Monoblast	Monocyte	Myeloblast	Myelocyte	Reactive Lymphocyte
L1	85	9	1	0	1	0	16	0	2
L2	13	56	3	0	16	0	58	0	2
L3	8	25	3	0	3	0	11	0	3
Lymphocyte	0	0	0	43	0	4	1	1	1
Monoblast	0	19	2	0	61	0	27	1	1
Monocyte	3	0	0	10	0	46	0	0	0
Myeloblast	21	23	4	0	10	0	100	0	12
Myelocyte	0	2	0	0	5	0	9	27	5
Reactive Lymphocyte	4	10	3	0	5	0	12	1	18

Table 1. Confusion Matrix. Results from the random forest classifier algorithm's performance on test data. Columns represent the actual cell identity while rows represent the random forest classification algorithm's determination of the cell identity.

	L1	L2	L3	Lymphocyte	Monoblast	Monocyte	Myeloblast	Myelocyte	Reactive Lymphocyte
Specificity	0.93	0.87	0.98	0.98	0.94	0.99	0.79	0.996	0.97
Sensitivity	0.75	0.38	0.06	0.86	0.55	0.78	0.59	0.56	0.34
Accuracy	0.90	0.78	0.92	0.98	0.89	0.98	0.75	0.97	0.92
Precision	0.63	0.39	0.19	0.81	0.60	0.92	0.43	0.90	0.41
Misidentification Rate	0.25	0.62	0.94	0.14	0.45	0.22	0.41	0.43	0.66

Table 2. Cell Prediction Metrics. The model's performance for accuracy, precision, sensitivity, and specificity for each cell type from the test data are reported.

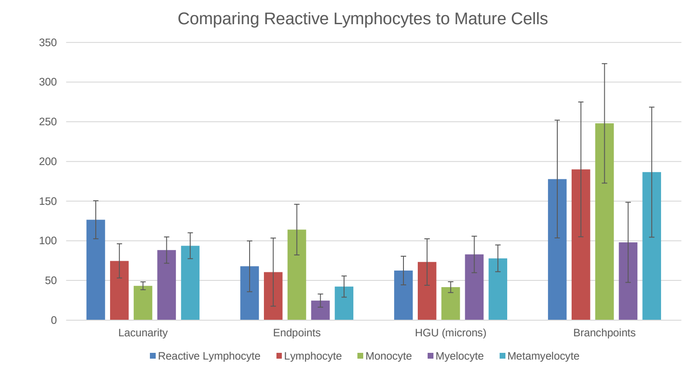

Figure 3. Reactive lymphocytes are most readily distinguished from mature cells using lacunarity. Error bars show the standard deviation.

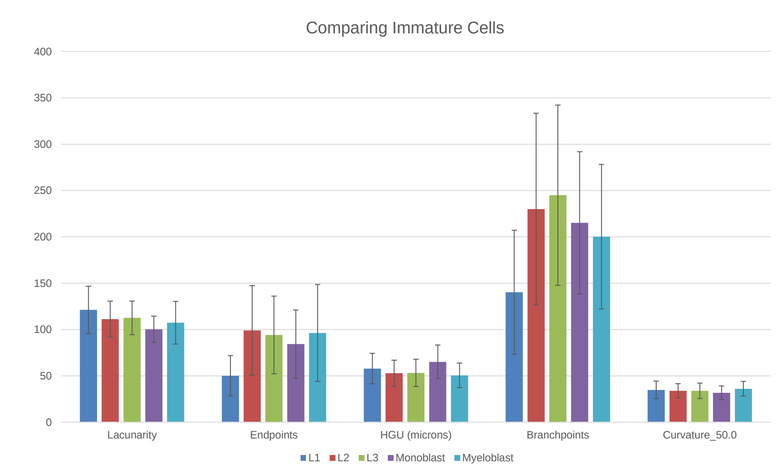

Figure 4. L1 lymphoblasts appear to be the most distinct from other types of blasts. Error bars show the standard deviation.

CONCLUSIONS

Our results suggest that a random forest algorithm can effectively distinguish between leukemic cells based on fractal chromatin patterns. It is possible that a similar algorithm could be used in the clinical laboratory to assist medical laboratory scientists and pathologists in distinguishing between reactive lymphocytes and blasts for both routine and abnormal blood counts and diagnoses.