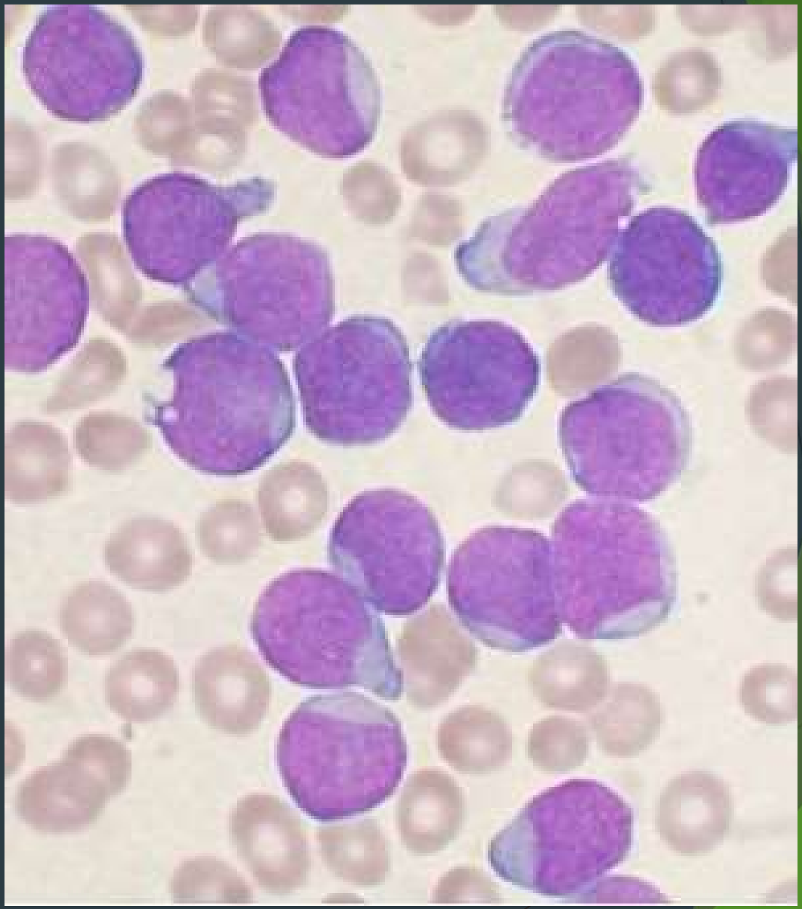
Leukemia Classification

CSE 527 Computational Biology

Andrew Wei, nowei@cs.washington.edu

What is Leukemia?



Wikipedia

The Data - GSE13159

- Name: Microarray Innovations in LEukemia (MILE) study: Stage 1 data
  - Data collected from 11 centers across 3 continents
  - Size: ~616 MB
  - Data released on Sept 30, 2009
- n = 2096 patients
- d = 17,788 genes
- 18 classes

Data breakdown

Name	Count
MDS	207
CLL	448
AML complex aberrant karyotype	52
AML with normal karyotype + other abnormalities	347
c-ALL/Pre-B-ALL without t(9;22)	237
T-ALL	174
CML	76
AML with t(11q23)/MLL	38
ALL with t(12;21)	58

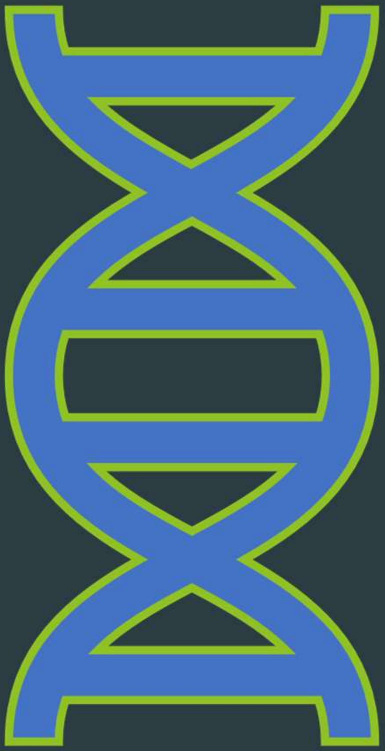
Name	Count
Non-leukemia and healthy bone marrow	73
c-ALL/Pre-B-ALL with t(9;22)	122
AML with t(8;21)	40
ALL with hyperdiploid karyotype	40
ALL with t(1;19)	36
Pro-B-ALL with t(11q23)/MLL	70
AML with t(15;17)	37
AML with inv(16)/t(16;16)	28
mature B-ALL with t(8;14)	13

Goal

Understand how gene expression patterns are different in different subtypes of leukemia

Challenges:

- Interpreting results
- Determining significance of results
- (I didn’t know we had to do a presentation until last week)



Approach

- 90/10 train/test split
- Finding significant features among different leukemia subtypes
  - Check for significant features (applying Bonferroni correction)
  - Look at most common significant features
- Logistic Regression
  - Check with all features/only with significant features and different data normalization methods
  - Check similarity between learned weights
  - Analyze learned weights
- K-Nearest Neighbors

Finding significant features

- Bonferroni correction (p=0.05, m=17788)
  - Performed on training set
  - "Non-leukemia and healthy bone marrow" vs. others
- 1,408 significant features in total

Leukemia sub-type	Number of significant features
MDS	39
CLL	956
AML complex aberrant karyotype	220
AML with normal karyotype + other abnormalities	522

Leukemia sub-type (cont.)	Number of significant features
c-ALL/Pre-B-ALL without t(9;22)	817
T-ALL	793
CML	195
AML with t(11q23)/MLL	271
ALL with t(12;21)	724
c-ALL/Pre-B-ALL with t(9;22)	731
AML with t(8;21)	373
ALL with hyperdiploid karyotype	581
ALL with t(1;19)	612
Pro-B-ALL with t(11q23)/MLL	689
AML with t(15;17)	428
AML with inv(16)/t(16;16)	351
mature B-ALL with t(8;14)	48

Jaccard similarity of significant features

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

	MDS	CLL	AML complex aberrant karyotype	AML with normal karyotype + other abnormalities	c-ALL/Pre-B-ALL without t(9;22)	T-ALL	CML	AML with t(11q23)/MLL	ALL with t(12;21)	c-ALL/Pre-B-ALL with t(9;22)	AML with t(8;21)	ALL with hyperdiploid karyotype	ALL with t(1;19)	Pro-B-ALL with t(11q23)/MLL	AML with t(15;17)	AML with inv(16)/t(16;16)	mature B-ALL with t(8;14)
MDS	1.00	0.03	0.07	0.04	0.04	0.03	0.03	0.05	0.04	0.04	0.05	0.05	0.04	0.04	0.06	0.04	0.00
CLL	0.03	1.00	0.17	0.37	0.53	0.49	0.14	0.20	0.44	0.47	0.28	0.38	0.39	0.46	0.31	0.26	0.04
AML complex aberrant karyotype	0.07	0.17	1.00	0.28	0.19	0.19	0.11	0.16	0.18	0.19	0.25	0.19	0.18	0.19	0.21	0.22	0.06
AML with normal karyotype + other abnormalities	0.04	0.37	0.28	1.00	0.44	0.40	0.15	0.37	0.34	0.40	0.44	0.35	0.37	0.45	0.41	0.43	0.06
c-ALL/Pre-B-ALL without t(9;22)	0.04	0.53	0.19	0.44	1.00	0.58	0.14	0.22	0.62	0.70	0.33	0.56	0.55	0.60	0.34	0.31	0.05
T-ALL	0.03	0.49	0.19	0.40	0.58	1.00	0.15	0.21	0.50	0.52	0.31	0.43	0.47	0.51	0.33	0.27	0.05
CML	0.03	0.14	0.11	0.15	0.14	0.15	1.00	0.16	0.17	0.15	0.16	0.15	0.17	0.16	0.16	0.16	0.05
AML with t(11q23)/MLL	0.05	0.20	0.16	0.37	0.22	0.21	0.16	1.00	0.19	0.23	0.33	0.21	0.21	0.25	0.28	0.34	0.09
ALL with t(12;21)	0.04	0.44	0.18	0.34	0.62	0.50	0.17	0.19	1.00	0.54	0.28	0.54	0.53	0.48	0.33	0.28	0.04
c-ALL/Pre-B-ALL with t(9;22)	0.04	0.47	0.19	0.40	0.70	0.52	0.15	0.23	0.54	1.00	0.33	0.52	0.51	0.53	0.34	0.31	0.05
AML with t(8;21)	0.05	0.28	0.25	0.44	0.33	0.31	0.16	0.33	0.28	0.33	1.00	0.28	0.30	0.35	0.40	0.41	0.07
ALL with hyperdiploid karyotype	0.05	0.38	0.19	0.35	0.56	0.43	0.15	0.21	0.54	0.52	0.28	1.00	0.49	0.48	0.32	0.30	0.05
ALL with t(1;19)	0.04	0.39	0.18	0.37	0.55	0.47	0.17	0.21	0.53	0.51	0.30	0.49	1.00	0.53	0.31	0.31	0.05
Pro-B-ALL with t(11q23)/MLL	0.04	0.46	0.19	0.45	0.60	0.51	0.16	0.25	0.48	0.53	0.35	0.48	0.53	1.00	0.36	0.32	0.05
AML with t(15;17)	0.06	0.31	0.21	0.41	0.34	0.33	0.16	0.28	0.33	0.34	0.40	0.32	0.31	0.36	1.00	0.32	0.06
AML with inv(16)/t(16;16)	0.04	0.26	0.22	0.43	0.31	0.27	0.16	0.34	0.28	0.31	0.41	0.30	0.31	0.32	0.32	1.00	0.07
mature B-ALL with t(8;14)	0.00	0.04	0.06	0.06	0.05	0.05	0.05	0.09	0.04	0.05	0.07	0.05	0.05	0.05	0.06	0.07	1.00

Checking the common significant features

# shared labels	# genes
16	1
15	17
14	31
13	30
12	51
11	64
10	75
9	110
8	101
7	128
6	111
5	103
4	116
3	125
2	130
1	215

# shared labels	Name of gene
16	10487_at
15	1116_at
15	116362_at
15	10123_at
15	1118_at
15	10562_at
…	…

* We have information on which labels the gene is shared by, but it wouldn’t fit on the slide.

Example: Looking at the most common significant gene

Shared by:
CLL, AML complex aberrant karyotype, AML with normal karyotype + other abnormalities, c-ALL/Pre-B-ALL without t(9;22), T-ALL, CML, AML with t(11q23)/MLL, ALL with t(12;21), c-ALL/Pre-B-ALL with t(9;22), AML with t(8;21), ALL with hyperdiploid karyotype, ALL with t(1;19), Pro-B-ALL with t(11q23)/MLL, AML with t(15;17), AML with inv(16)/t(16;16), mature B-ALL with t(8;14)

Not shared by: MDS

Gene: 10487_at

Shared by: 16 leukemia subtypes

Description: CAP1 - CAP, adenylate cyclase-associated protein 1 (yeast)

Checking significance

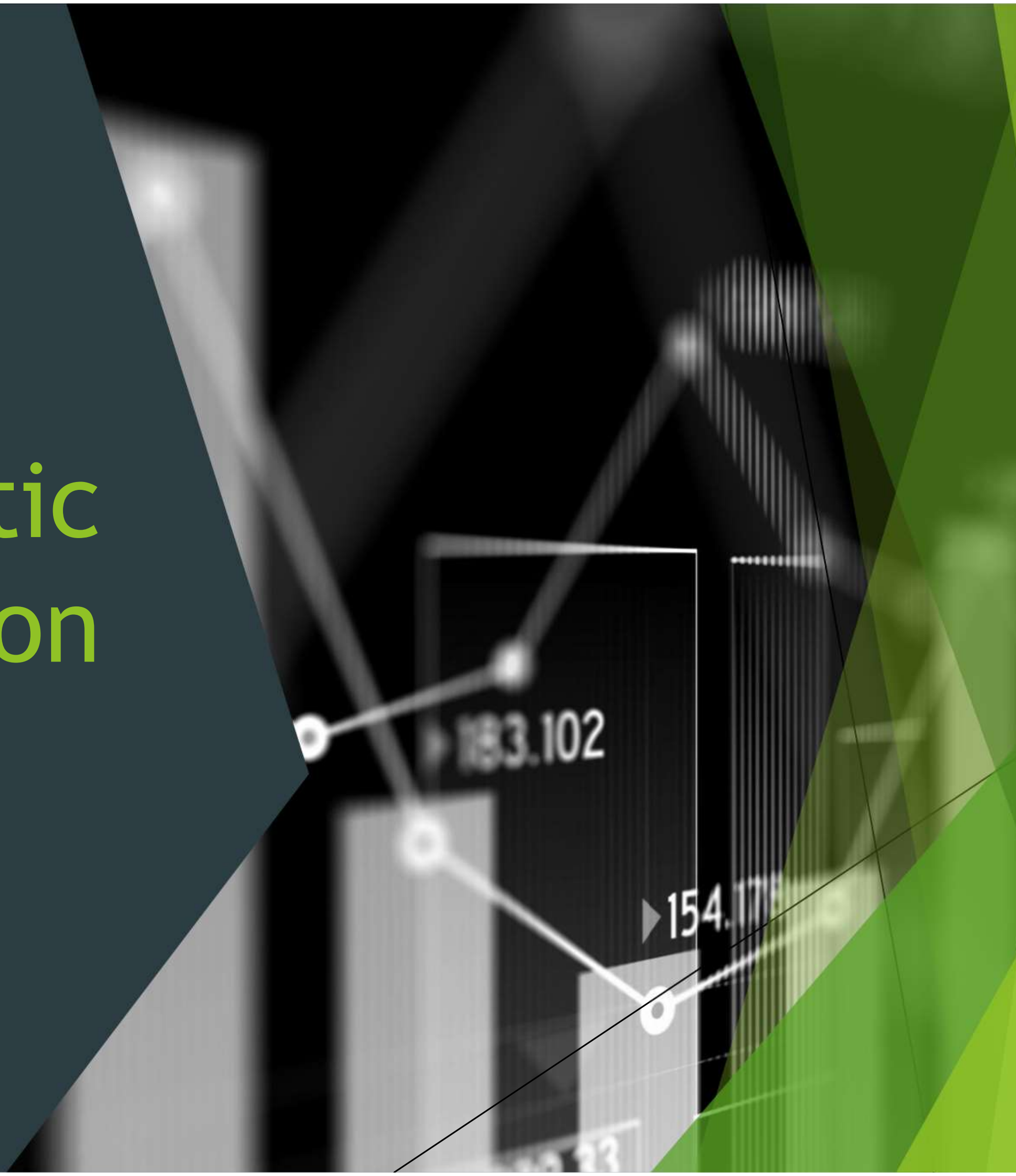
Gene: 10487_at

Shared by: 16 leukemia types

Description: CAP1 - CAP, adenylate cyclase-associated protein 1 (yeast)

From [Xie, Shen, Tan, Li, Song, Wang 2017]: “CAP1 […] was under-expressed in breast and leukemia cancers as compared to that in normal tissue.”

Logistic Regression



Learning settings

- Used 5-fold cross validation
- L1 regularization
- With all features vs. with significant features
- Different normalization schemes
- 1 vs. all classification scheme

Normalization schemes

- Don’t normalize
- Normalize across entire training dataset
- Normalize by healthy patient data in training dataset

Results of 5-fold cross-validation

Normalization Scheme	w/ all features		w/ significant features	
	Top 1 acc.	Top 5 acc.	Top 1 acc.	Top 5 acc.
Don’t normalize	0.895	**0.995**	0.870	**0.987**
Normalize across entire training dataset	0.898	0.988	0.881	0.986
Normalize by healthy patient data in training dataset	**0.907**	0.991	**0.884**	0.984

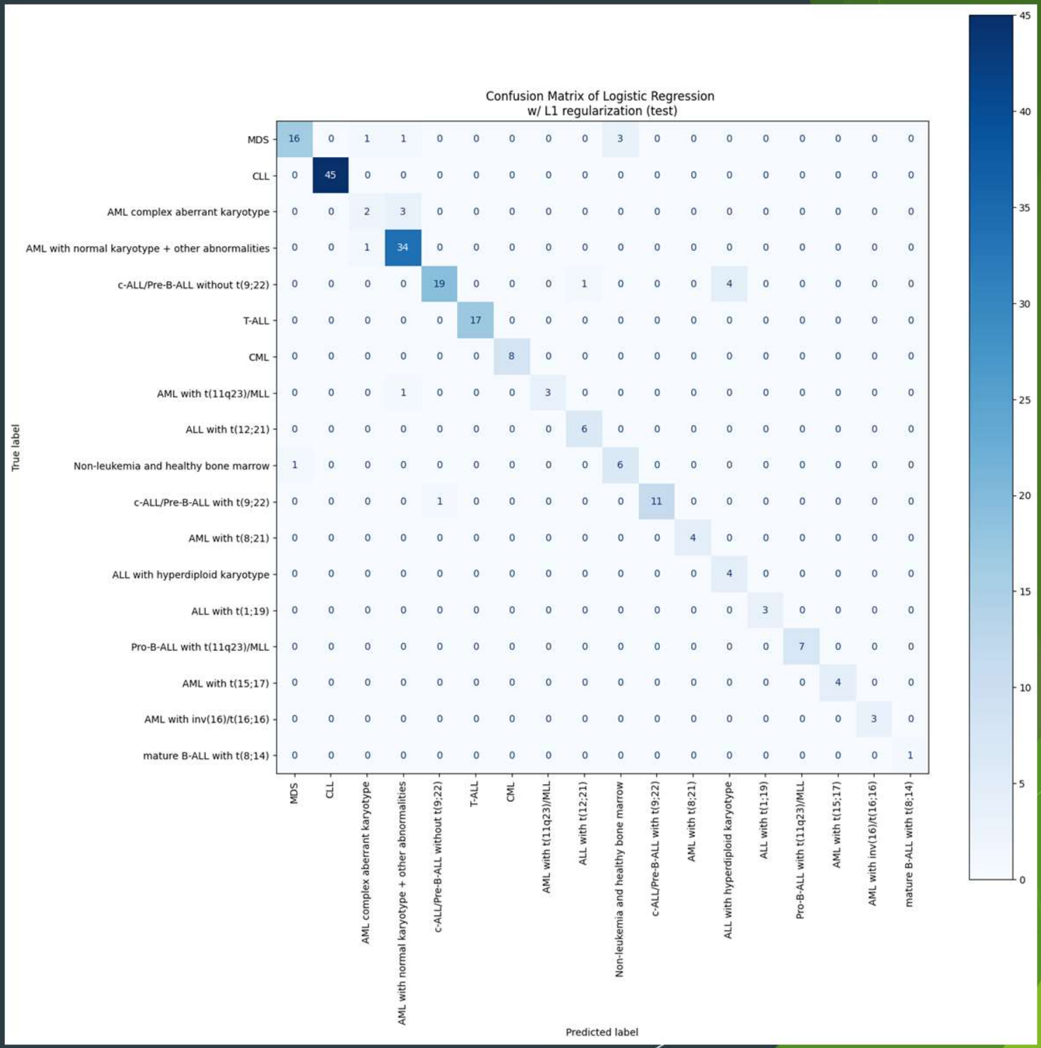
Results

Train

Normalization Scheme	w/ all features		w/ significant features	
	Top 1 acc.	Top 5 acc.	Top 1 acc.	Top 5 acc.
Don’t normalize	1.0	1.0	1.0	1.0
Normalize across entire training dataset	1.0	1.0	1.0	1.0
Normalize by healthy patient data in training dataset	1.0	1.0	1.0	1.0

Test

Normalization Scheme	w/ all features		w/ significant features	
	Top 1 acc.	Top 5 acc.	Top 1 acc.	Top 5 acc.
Don’t normalize	**0.919**	**1.0**	**0.881**	**0.990**
Normalize across entire training dataset	0.919	0.995	0.857	0.981
Normalize by healthy patient data in training dataset	0.900	0.995	0.843	0.971


Confusion Matrix of Logistic Regression
w/ L1 regularization (test)

True label \ Predicted label	MDS	CLL	AML complex aberrant karyotype	AML with normal karyotype + other abnormalities	c-ALL/Pre-B-ALL without t(9;22)	T-ALL	CML	AML with t(11q23)/MLL	ALL with t(12;21)	Non-leukemia and healthy bone marrow	c-ALL/Pre-B-ALL with t(9;22)	AML with t(8;21)	ALL with hyperdiploid karyotype	ALL with t(1;19)	Pro-B-ALL with t(11q23)/MLL	AML with t(15;17)	AML with inv(16)/t(16;16)	mature B-ALL with t(8;14)
MDS	16	0	1	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0
CLL	0	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AML complex aberrant karyotype	0	0	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AML with normal karyotype + other abnormalities	0	0	1	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c-ALL/Pre-B-ALL without t(9;22)	0	0	0	0	19	0	0	0	1	0	0	0	4	0	0	0	0	0
T-ALL	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0
CML	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0
AML with t(11q23)/MLL	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0
ALL with t(12;21)	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0
Non-leukemia and healthy bone marrow	1	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0
c-ALL/Pre-B-ALL with t(9;22)	0	0	0	0	1	0	0	0	0	0	11	0	0	0	0	0	0	0
AML with t(8;21)	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
ALL with hyperdiploid karyotype	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
ALL with t(1;19)	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
Pro-B-ALL with t(11q23)/MLL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0
AML with t(15;17)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0
AML with inv(16)/t(16;16)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
mature B-ALL with t(8;14)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Reflecting on results

* results from using all features and no normalization

Features used by leukemia sub-type model

Leukemia Sub-type	# zeros	# non-zero
MDS	15400	2388
CLL	17004	784
AML complex aberrant karyotype	17258	530
AML with normal karyotype + other abnormalities	15227	2561
c-ALL/Pre-B-ALL without t(9;22)	15381	2407
T-ALL	16507	1281
CML	17040	748
AML with t(11q23)/MLL	17080	708
ALL with t(12;21)	17072	716

Leukemia Sub-type	# zeros	# non-zero
Non-leukemia and healthy bone marrow	17235	553
c-ALL/Pre-B-ALL with t(9;22)	16414	1374
AML with t(8;21)	17336	452
ALL with hyperdiploid karyotype	16999	789
ALL with t(1;19)	17132	656
Pro-B-ALL with t(11q23)/MLL	16959	829
AML with t(15;17)	17202	586
AML with inv(16)/t(16;16)	17481	307
mature B-ALL with t(8;14)	17544	244

Total number of genes: 17788

Cosine Similarity to compare learned weights

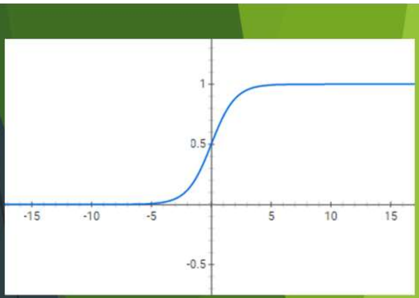
$$C(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

* results from using all features and no normalization

	MDS	CLL	AML complex aberrant karyotype	AML with normal karyotype + other abnormalities	c-ALL/Pre-B-ALL without t(9;22)	T-ALL	CML	AML with t(11q23)/MLL	ALL with t(12;21)	Non-leukemia and healthy bone marrow	c-ALL/Pre-B-ALL with t(9;22)	AML with t(8;21)	ALL with hyperdiploid karyotype	ALL with t(1;19)	Pro-B-ALL with t(11q23)/MLL	AML with t(15;17)	AML with inv(16)/t(16;16)	mature B-ALL with t(8;14)
MDS	1.000	-0.002	-0.034	-0.145	0.006	-0.010	-0.065	-0.013	-0.037	-0.206	-0.017	-0.024	0.016	0.012	-0.011	-0.015	-0.018	0.022
CLL	-0.002	1.000	-0.024	-0.029	-0.028	-0.016	0.007	-0.017	-0.002	-0.007	0.009	0.023	-0.015	0.001	0.004	0.015	-0.013	-0.021
AML complex aberrant karyotype	-0.034	-0.024	1.000	-0.122	-0.012	-0.001	-0.011	0.001	-0.006	-0.037	-0.021	0.022	0.016	0.006	0.010	-0.004	0.020	-0.004
AML with normal karyotype + other abnormalities	-0.145	-0.029	-0.122	1.000	-0.028	-0.058	-0.070	-0.104	-0.009	-0.049	0.004	-0.062	-0.016	0.000	-0.016	-0.021	-0.064	-0.009
c-ALL/Pre-B-ALL without t(9;22)	0.006	-0.028	-0.012	-0.028	1.000	-0.012	-0.019	-0.001	-0.126	0.005	-0.207	-0.008	-0.109	-0.050	-0.026	0.021	-0.007	-0.006
T-ALL	-0.010	-0.016	-0.001	-0.058	-0.012	1.000	0.010	-0.004	-0.014	-0.003	-0.018	-0.009	-0.006	-0.001	-0.015	-0.012	0.018	0.012
CML	-0.065	0.007	-0.011	-0.070	-0.019	0.010	1.000	-0.007	-0.004	-0.014	0.007	-0.008	-0.026	0.007	0.011	-0.049	-0.029	-0.010
AML with t(11q23)/MLL	-0.013	-0.017	0.001	-0.104	-0.001	-0.004	-0.007	1.000	0.011	-0.008	0.000	0.001	-0.011	0.007	0.018	-0.011	0.024	-0.003
ALL with t(12;21)	-0.037	-0.002	-0.006	-0.009	-0.126	-0.014	-0.004	0.011	1.000	0.021	-0.025	0.008	-0.026	-0.018	-0.026	0.000	-0.001	0.003
Non-leukemia and healthy bone marrow	-0.206	-0.007	-0.037	-0.049	0.005	-0.003	-0.014	-0.008	0.021	1.000	-0.008	-0.004	0.017	-0.012	0.012	-0.006	0.012	0.007
c-ALL/Pre-B-ALL with t(9;22)	-0.017	0.009	-0.021	0.004	-0.207	-0.018	0.007	0.000	-0.025	-0.008	1.000	0.018	-0.019	-0.022	-0.030	-0.021	0.005	-0.008
AML with t(8;21)	-0.024	0.023	0.022	-0.062	-0.008	-0.009	-0.008	0.001	0.008	-0.004	0.018	1.000	-0.005	-0.002	-0.012	0.016	0.009	-0.014
ALL with hyperdiploid karyotype	0.016	-0.015	0.016	-0.016	-0.109	-0.006	-0.026	-0.011	-0.026	0.017	-0.019	-0.005	1.000	-0.013	-0.004	0.022	0.002	0.011
ALL with t(1;19)	0.012	0.001	0.006	0.000	-0.050	-0.001	0.007	0.007	-0.018	-0.012	-0.022	-0.002	-0.013	1.000	-0.019	0.005	0.008	0.006
Pro-B-ALL with t(11q23)/MLL	-0.011	0.004	0.010	-0.016	-0.026	-0.015	0.011	0.018	-0.026	0.012	-0.030	-0.012	-0.004	-0.019	1.000	-0.002	0.005	0.010
AML with t(15;17)	-0.015	0.015	-0.004	-0.021	0.021	-0.012	-0.049	-0.011	0.000	-0.006	-0.021	0.016	0.022	0.005	-0.002	1.000	0.011	-0.001
AML with inv(16)/t(16;16)	-0.018	-0.013	0.020	-0.064	-0.007	0.018	-0.029	0.024	-0.001	0.012	0.005	0.009	0.002	0.008	0.005	0.011	1.000	0.020
mature B-ALL with t(8;14)	0.022	-0.021	-0.004	-0.009	-0.006	0.012	-0.010	-0.003	0.003	0.007	-0.008	-0.014	0.011	0.006	0.010	-0.001	0.020	1.000

Most heavily weighted weights

* results from using all features and no normalization



MDS

Highest

Name	Coefficient
66000_at	0.100164
2706_at	0.091647
158809_at	0.088469
100130703_at	0.083912
9518_at	0.083138

…

Lowest

Name	Coefficient
285299_at	-0.073488877
388951_at	-0.074094855
51673_at	-0.079185137
7266_at	-0.144620158
56884_at	-0.205755898

CCL

Name	Coefficient
27033_at	0.079577
130367_at	0.067018
5923_at	0.052974
81537_at	0.043758
2823_at	0.041918

…

Name	Coefficient
79872_at	-0.04386
100131644_at	-0.04524
85358_at	-0.04816
390058_at	-0.11016
3892_at	-0.16053

K-Nearest Neighbors

Instance-based learning

Parameters to tune:

- K, i.e. # of neighbors
  - n = 2096 patients

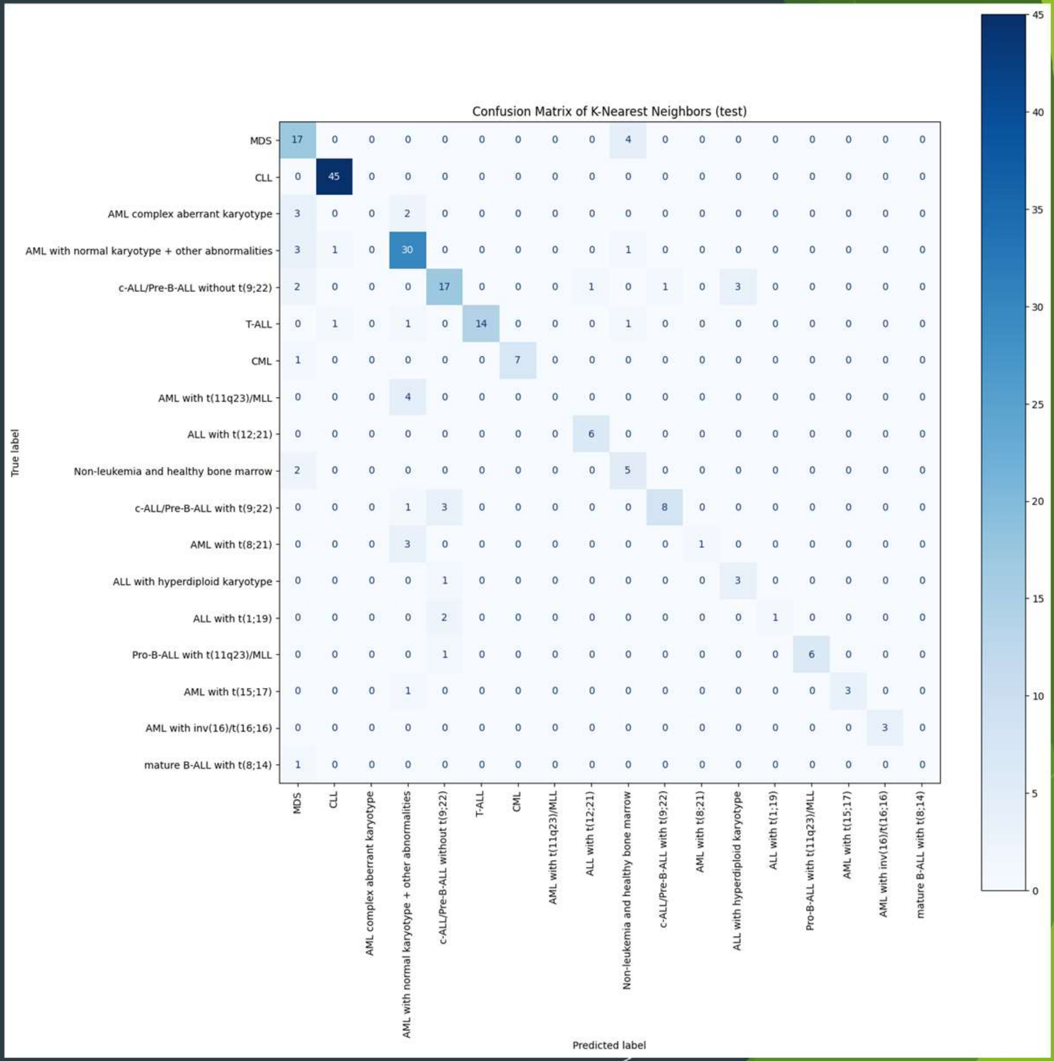
- Whether to use significant features or all the features

Results

# of neighbors	w/ all features		w/ significant features	
	Top 1 acc.	Top 5 acc.	Top 1 acc.	Top 5 acc.
3	0.767	0.881	0.790	0.895
5	0.771	0.914	0.790	0.929
10	**0.795**	0.962	**0.819**	**0.962**
15	0.786	0.967	**0.819**	0.952
20	0.790	**0.971**	0.786	0.952

Reflecting on results

- Similar results, but more different errors


Confusion Matrix of K-Nearest Neighbors (test)

	MDS	CLL	AML complex aberrant karyotype	AML with normal karyotype + other abnormalities	c-ALL/Pre-B-ALL without t(9;22)	T-ALL	CML	AML with t(11q23)/MLL	ALL with t(12;21)	Non-leukemia and healthy bone marrow	c-ALL/Pre-B-ALL with t(9;22)	AML with t(8;21)	ALL with hyperdiploid karyotype	ALL with t(1;19)	Pro-B-ALL with t(11q23)/MLL	AML with t(15;17)	AML with inv(16)/t(16;16)	mature B-ALL with t(8;14)
MDS	17	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
CLL	0	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AML complex aberrant karyotype	3	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AML with normal karyotype + other abnormalities	3	1	0	30	0	0	0	0	0	1	0	0	0	0	0	0	0	0
c-ALL/Pre-B-ALL without t(9;22)	2	0	0	0	17	0	0	0	1	0	1	0	3	0	0	0	0	0
T-ALL	0	1	0	1	0	14	0	0	0	1	0	0	0	0	0	0	0	0
CML	1	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0
AML with t(11q23)/MLL	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ALL with t(12;21)	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0
Non-leukemia and healthy bone marrow	2	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0
c-ALL/Pre-B-ALL with t(9;22)	0	0	0	1	3	0	0	0	0	0	8	0	0	0	0	0	0	0
AML with t(8;21)	0	0	0	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ALL with hyperdiploid karyotype	0	0	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0	0
ALL with t(1;19)	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0
Pro-B-ALL with t(11q23)/MLL	0	0	0	0	1	0	0	0	0	0	0	0	0	0	6	0	0	0
AML with t(15;17)	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	0	0
AML with inv(16)/t(16;16)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
mature B-ALL with t(8;14)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Future Work

- Further examine and interpret the results and errors obtained
- Alternative 1 vs. all classification
  - Make a separate model for each leukemia subtype made by selecting out and training on the significant features
- Try different regularizers and machine learning methods
- Address data imbalance

Conclusion

- Analyzing shared significant features seems to be useful
- Logistic regression models can get close to 100% accuracy with MILE dataset
- K-Nearest Neighbors is okay, but not that great

Questions?

Thanks!