Generalized procedure for selecting methylation CpGs associated with cancer

Marija Stanojevic
Center for Data Analytics and Biomedical Informatics, Temple University, Pennsylvania, PA
marija.stanojevic@temple.edu

Zoran Obradovic
Center for Data Analytics and Biomedical Informatics, Temple University, Pennsylvania, PA
zoran.obradovic@temple.edu

Abstract

Recent studies work on connecting methylation values with cancer and identifying methylation sites or corresponding DNA parts as tumor markers. This effort is based on statistical methods. Since the size of methylation arrays is huge, current research rather works with groups of methylation features, which has numerous disadvantages, including fact that it can't understand role of single position.

This paper is analyzing each methylation site individually trying to select positions which can separate in best possible way normal and cancerous tissues in different cancers. Generalized procedure for positions selection is proposed that is robust to imbalance in classes and small datasets which are two biggest problems of methylation datasets. This method performs better than state-of-the art method having around 1% better results for large datasets and 5-30% better results for very small datasets. This method doesn't require both normal and cancerous samples from patient. Model doesn't require any assistance from experts.

Regularization value for least absolute shrinkage and selection operator (LASSO) is shown to be dependent of sample size and not much related to the cancer type. By comparing LASSO and Elastic net regression models, we show that using single methylation positions is better than using spatial correlation between features for 27,000 array.

Selected features are consistent despite changes in dataset and Cox analysis of time till death gives better results than with features selected with state-of-the-art model. Some positions selected for one cancer are good or spatially close to good features to distinguish other cancers as well. Future work will focus more on understanding this experimental findings from medical point of view.

Introduction

Understanding causes and influences that increase chance of cancer development is crucial problem in medicine, drug development, biology and genetics. Assessment is that 39.6% of men and women will be diagnosed with cancer at some point during their lifetimes and only 66.5% of patients with cancer have survived for more than 5 years. In 2017, it is estimated that 1,688,780 cases were diagnosed only in United States [1], what is 0.05% of population. Most common cancers are breast cancer, lung and bronchus cancer, prostate cancer, colon and rectum cancer, bladder and melanoma of the skin cancer. Cancer mortality is higher among men than among women. Every year 0.02% population dies from cancer in US only. Therefore, expenditures for cancer care in United States is expected to reach $156 billion in 2020. [2]. This paper is focuses on finding of DNA methylation points which are strong differentiators between cancerous and normal tissues.

DNA part in which cytosine is followed by guanine nucleotide in sequence along 5’ -> 3’ direction is called CpG site. CpG is short for cytosine -> phosphate -> guanine, since in this case those two nucleotides are separated with phosphate. DNA regions with high frequency of CpG positions are called CpG islands.

Methylation is process of adding 5-methyl group to cytosine in CpG site and it can change gene expression. In mammals, 70-80% of cytosine of CpG sites are methylated. In humans, methylation at the 5’ position of the cytosine pyrimidine ring at CpG position creates 5-methylcytosines. Methylation beta value, also known as methylation level, is estimated as ratio of intensities of the methylated and total of methylated and unmethylated alleles. It can be in range [0, 1], where 0 corresponds to unmethylated, and 1 means fully methylated.

As the most stable and experimentally accessible epigenetic mark, DNA methylation is of great interest to the research community. It can be used to understand DNA level changes that are associated to cancer and also to predict risk of cancer or survival time for persons with cancer. Methylation is influenced by both genetics and environment, but it doesn’t change as fast as gene expression which makes it very good and stable indicator of health change.

Objective of this paper is to create generalized procedure for selecting methylation CpG positions that are associated with cancer and predicting survival time based on selected features. To the best of authors knowledge there is no generalized procedure that is applicable to all cancers.

Hypothesis 1: It is possible to create generalized procedure for classifying samples as normal and cancerous. This procedure gives better classification than state-of-the-art model [3].

Hypothesis 2: Created generalized procedure is able to extract CpG points consistently, regardless the size or structure (number of cancerous/normal samples) of samples. Selected CpG points don’t depend on samples.

Additionally, this paper wants to answer questions:

1. Is there an overlap in selected features between different common cancers?
2. Is it better to use methylation difference or methylation deviation difference or a mixture of those to understand cancers?
3. Do CpG islands play role in selecting important CpG features?

Contributions of this paper are:

1. Generalized procedure is created that can be applied to any given cancer and it outperforms state-of-the-art procedure which is not generalizable, i.e. requires human judgment and attention for each cancer type.
2. Created model is robust to data size. It gives very good accuracy even for 20 samples (5-30% better than state-of-the-art [3]), which is important because methylation extraction is expensive process even for more common cancers. This can especially help with rare cancers.
3. This procedure is robust to imbalance of classes and performs well as long as there are samples from both types of tissues and doesn’t require normal and cancerous tissue sample to be from the same patient as in [3].

4. Connection between regularization value and sample size is shown. Additionally, it’s shown that there is no strong connection between regularization value and cancer type.

Related work

Methylation extraction and analysis is very complex process, but it is often used in understanding diseases and cell level processes which are currently not explained, such as obesity [4], cancer [5], [6], [7] and neural system [8] functioning. Methylation analysis for purpose of understanding and explaining cancer is very often used and there are many papers that explain protocols for methylation values extraction ([9], [10], [11]), analysis ([12], [13], [14]) or the whole process of working with methylation data ([14]) as shown in **Figure 1**.

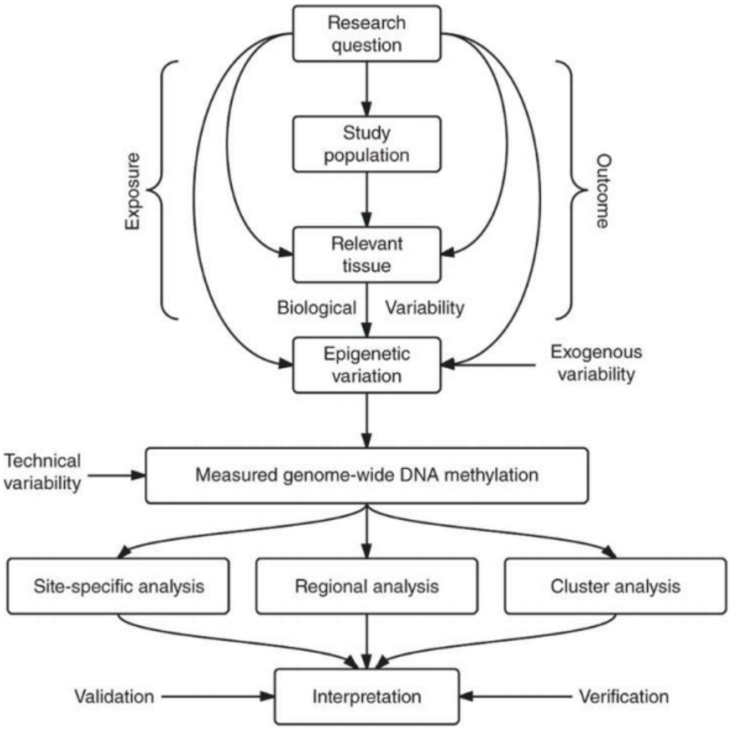

Figure 1: Steps toward a successful epigenome-wide study (EWAS) in cancer

It is shown in [15] that methylation level influences gene expression in animals. As mentioned in [16] hypervariable DNA methylation is related with a lack of order in gene expression in humans. Results of [17] explain genetic apparatus for variable methylation and it shows that high variability of DNA methylation is related to evolution.

As written in [18] DNA methylation has possible influence on many diseases and biological processes, therefore it is important to understand it better. Recent research is using DNA methylation as markers for certain cancers or in diagnosis of other complex diseases and influence of drugs in treatment of patients. Methylation is suitable for such analysis since it is influenced by all factors: genetics, behavior and environment.

Many studies ([20], [21], [22]) discussed influence of genetics, behavior and environment (age, BMI) on cancer development through analysis of DNA methylation level.

Studies showed that difference of mean level of DNA methylation can help to identify those features that influence cancer strongly [23]. Models developed based on this discovery were successful in finding important features and understanding their influence on cancer development. Recent papers ([16], [17], [19], [24], [25]) discussed that variance of DNA methylation is also important for understanding disease. Results of [19] have shown that cancer risk markers can be identified better if differential variability of DNA methylation is used instead of DNA methylation mean. Study [26] showed that features that have extraordinary high variance in normal tissue also have extraordinary high variance in cancer tissue when comparing different DNA methylation features.

In previous experimental work on colon cancer samples (unpublished) it is found that features selected using mean difference and deviation difference, as described in papers referenced above,

are very different. Also, features highly depended on pre-processing of data and bounds that are determined by experts and have to be determined separately for each cancer.

Methodology of [3] uses better models comparing to above described methods, however, it still requires experts to determine bounds in pre-processing step. This method consists of volcano plot which shows $\log_{10}$(p-value) of t-test on y-axis and $\log_2$(fold change) on x-axis between cancer and normal tissues of same patient. Experts select bounds based on volcano plot and features that are more extreme than given bounds are pre-selected. Then, method applies LASSO regression using pre-selected features to determine the most important features and it applies Cox analysis for time till death using the LASSO selected features. Method [3] has to be separately developed for each cancer (since each cancer would have different bounds). Additionally, model is not very robust to changes in data set and can’t work with imbalanced datasets at all.

Methodology

Data are taken from GDC Data Portal (https://portal.gdc.cancer.gov) for four types of cancer: kidney (1307 samples), breast (1234 samples), lung (1230 samples) and colorectal (735 samples) since those cancers are represented with highest number of samples in the portal. All those cancers have subtypes except breast cancer. Study [3] considers pairs of samples of methylation between normal and cancerous tissue for the same patient, however there are only around 90, 70, 250 and 90 of such patients in those data for each cancer respectively.

This study aim is to work with imbalanced data and without requirement of having two types of tissues from patient. Beside methylation arrays, clinical and biospecimen data are collected, even though they are not present for all the samples. Clinical data contain demographic information (ethnicity, gender, race, year of birth and year of death), diagnoses (age of diagnosis, days from birth, days to death, days from last follow up, last known disease status, classification of tumor, alcohol intensity, ICD code, tumor grade, tumor stage and vital status) and exposure (alcohol history, BMI, cigarets per day, height, weigh, years smoked). Biospecimen data contain information about sample type, tumor code and days to collection. Clinical and biospecimen data have high percentage of missing data, so selected featuers will be used mainly for interpretation of results in this work.

Unfortunately, data taken from GDC program is not ready for analysis and it requires a lot of preprocessing since methylation measures come from different laboratories and projects. Old methylation measures were able to extract 27,000 positions methylation, while newer methods extract 480,000 positions and they are not compatible, so first step in data preprocessing is to **find overlap of features** which is around 25,000 positions. In such dataset some **features are missing almost all the data and have to be removed**. Final dataset contains around 23,500 features. Other features that have missing data usually have around 1% of missing data, if any. Those **missing data are imputed with mean for that feature**. Boundary for removal of the feature is 20% or more of missing values.

Methylations are measured using small chips and then scaled to [0,1] interval. Study [14] and many other suggest to account for **chip bias** when using data measured by multiple chips. Since chip effect was experimentally confirmed in unpublished work by author, in this work **mean of the chip for feature is subtracted from each methylation point**, so new interval for methylation values is [-1, 1].

While methylation data is in .csv file and easy to extract, clinical and speciment data are encoded in xml page. For the purpose of this project, most important features from those files are extracted. If variable is continuous, it is recorded as such (those variables are in R^+). If variable is binary, given values (usually strings) are transformed into numbers 0 and 1. **Missing data is encoded with -1 for binary variables and with 10^{-8} in case of continues variables**. Because of high percentage of missing data, those features are mainly used for interpretation. Only variable used in all the analysis in this work is type of sample (normal – 0 and cancerous – 1).

LASSO is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. It is used as part of the [3] procedure, however, idea of this project is to use LASSO only for selection of important features in order to avoid need for experts created boundaries and to be able to generalize the approach to different cancers. Study [3] selects features based on mean difference, but ([16], [17], [19], [24], [25]) have shown that variance difference is also important differentiator between cancerous and normal tissues. By removing pre-selection based on mean we allow for features to be selected even if they have just different variance and they don't have different mean.

LASSO regression is linear regression with l1-norm regularization which Lagrangian form is: $\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 \right\}$ subject to $\|\beta\|_1 \leq t.$, which can be rewritten as $\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$

In this paper, different values of lambda are tested with LASSO model and lambda is chosen to minimize R-squared score of LASSO fitted model, where R-squared is defined as $1 - \frac{u}{v}$ where $u = \sum_i (y_i - predicted_i)^2$ and $u = \sum_i (y_i - mean(y))^2$. Additionally, classification accuracy is measured and reported and in most cases model that minimizes R-squared error was also maximizing classification accuracy. Classification accuracy is $\frac{\text{\# well classified samples}}{\text{\# all samples}}$.

Elastic net is regression method that uses both l1-norm and l2-norm regularizations. It is tested in here as alternative methodology since LASSO regression is very strict in selecting features. **If there is group of highly correlated variables, LASSO will selected only one from them and ignore others.** Elastic net regression, on the other side, have higher tolerance for correlated variables. Therefore, LASSO and Elastic net are compared here to understand if and how result will be influenced by spatial dependency between methylation positions. Formula of Elastic net can be written as:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1)$$

Once model is fitted, lambda values are chosen to minimize R-squared score described above. Also, classification accuracy is measured.

In order to make full comparison, Ridge regression (regression with l2-norm regularization) is tested using few runs to confirm hypothesis that its performance will be much worse than performance of proposed model because it's feature selection is very relaxed.

Lastly, **LASSO regularization is tested on methylation data extended with the standard deviation** for each of the features. This methodology is created to see how additional methylation variance features will influence classification and R-squared results.

To ensure robustness of the methodology despite data imbalance and data size, methodology decides on final **lambda value as average of lambdas in 10 runs**. Each run is created by randomly splitting original dataset into training (80%) and test (20%) data. Features that have coefficient larger than 0.05 are selected as important from one run, but final list of features consist of features that were selected in at least 50% of runs. This procedure selects around 10 features per cancer. Without last step, each cancer has around 20-30 selected features using LASSO regression. For smaller datasets (20 or 40 samples only) features coefficients are smaller since model is less sure about them, so all features with coefficient bigger than 0.02 are selected and rest of the procedure is the same.

Once features are selected, **Cox regression** is used to model time to death for patients using their time from last visit in case they are alive or time to death if they are dead. Both times are measured from the point when cancer was detected up to event (last visit/death). People alive are given value 1 and people dead are given value 0. Survival models have two purposes: 1) to understand how the risk of event changes over time and 2) to understand how the event varies in response to explanatory variables. We can model Cox proportional hazard as function at time *t* for subject *i* with explanatory variables X_i.

$$\lambda(t|X_i) = \lambda_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}) = \lambda_0(t)\exp(X_i \cdot \beta).$$

Likelihood and log-likelihood of event occurring at time Y_i with subject *i* can be modeled with:

$$L_i(\beta) = \frac{\lambda(Y_i|X_i)}{\sum_{j:Y_j \geq Y_i} \lambda(Y_i|X_j)} = \frac{\lambda_0(Y_i)\theta_i}{\sum_{j:Y_j \geq Y_i} \lambda_0(Y_i)\theta_j} = \frac{\theta_i}{\sum_{j:Y_j \geq Y_i} \theta_j}, \qquad \ell(\beta) = \sum_{i:C_i=1} \left(X_i \cdot \beta - \log \sum_{j:Y_j \geq Y_i} \theta_j \right).$$

The log-likelihood function can be maximized over parameter β by taking first derivative over β and then calculating Hessian matrix of the partial log likelihood.

Results and Discussion

To prove that accuracy of our model is better than state-of-the art methodology [3], first experiments compare methodology from [3] with our methodology using LASSO and Elastic Net. State-of-the-art method doesn’t mention averaging multiple runs or selected features that appear the best in multiple runs. In order to give equal opportunities to the models, wrapper developed around LASSO for the purpose of this paper is added as wrapper of state-of-the-art method. From the differences between min and max values for 10 runs we can see that running experiment only once is bad practice (since results can vary up to 7%) and also features selected in different runs can differ for around 10-20%.

When comparing LASSO and Elastic net models (**Table 1**), we can see that LASSO always outperforms Elastic net for up to 1% of classification accuracy and for up to 0.04 of R-squared value. This shows that allowing for more flexibility in selecting features that would allow influence of spatial correlation between features worsens the results. Also, When comparing lists of features selected by those two models, we can see that there is around 80% match.

Figures 2-5 show different methylation patterns for Kidney cancer, where y**-axis is methylation value** and **x-axis is sample number**. **Figure 2** shows example of feature selected by Elastic net and not selected by LASSO. We can see that **cancerous (red) and normal (green) samples** have similar variations and means, explained by flexibility of Elastic net. **Figure 3** shows example of feature selected by LASSO and not selected by Elastic net model. We can see that both mean and variance are very different between the cancerous and normal samples due to rigorous feature selection process of

LASSO. Also, list of features selected by Elastic net in at least one run is broader 30-40 features (comparing to 20-30 selected by LASSO), but amount of features that appeared in more than 50% of runs is similar to LASSO. Data is also tested with **Ridge regression** with few runs, but p**erformance was in range 10-80% for classification task,** so rest of the experiments are canceled and it is confirmed that Ridge regression is not a solution for this problem.

	l1			l12			state-of-the-art[3]		
	avg	max	min	avg	max	min	avg	max	min
breast (classification)	**94.797**	97.561	91.870	94.756	97.967	93.089	94.186	95.934	91.463
breast (R-sq)	**0.341**	0.446	0.268	0.340	0.462	0.133	0.305	0.434	0.202
breast (lambda)	0.002			0.004			0.001		
colorectal (classification)	**95.009**	97.945	93.835	94.073	96.599	89.116	94.049	97.945	91.095
colorectal (R-sq)	**0.356**	0.455	0.289	0.345	0.464	0.147	0.311	0.406	0.167
colorectal (lambda)	0.003			0.007			0.003		
kidney (classification)	96.123	97.701	94.615	96.092	97.701	95.402	**96.127**	98.084	94.252
kidney (R-sq)	0.681	0.696	0.653	0.680	0.718	0.631	**0.685**	0.723	0.634
kidney (lambda)	0.003			0.006			0.002		
lung (classification)	**96.551**	98.312	94.222	95.650	97.414	94.396	95.928	97.436	92.703
lung (R-sq)	**0.492**	0.563	0.325	0.452	0.502	0.389	0.460	0.524	0.386
lung (lambda)	0.002			0.003			0.002		

Table 1: Comparison of different methods average, maximum and minimum classification accuracy and R-squares for four cancers: breast, colorectal, kidney and lung

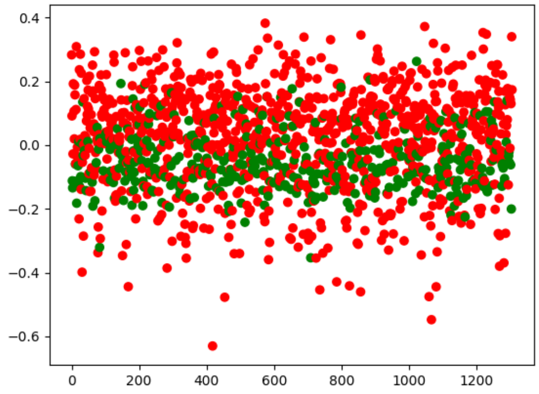

Figure 3: Kidney cancer: position cg06290096 selected by Elastic net and not by LASSO

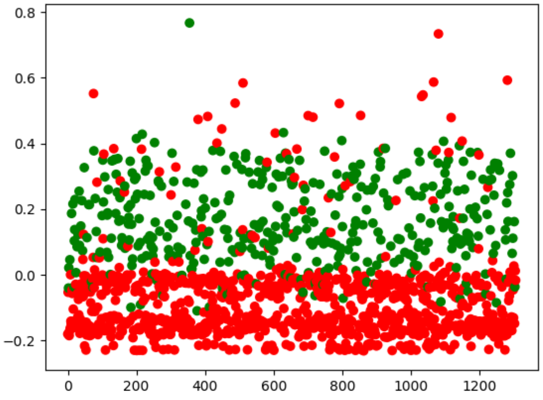

Figure 4: Kidney cancer: position cg00363813: example of very different means

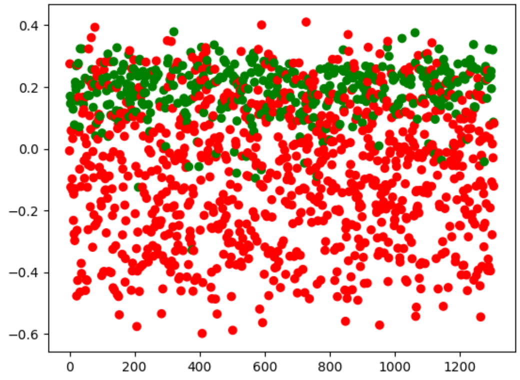

Figure 2: Kidney cancer: position cg04312209: selected by LASSO and not by Elastic net

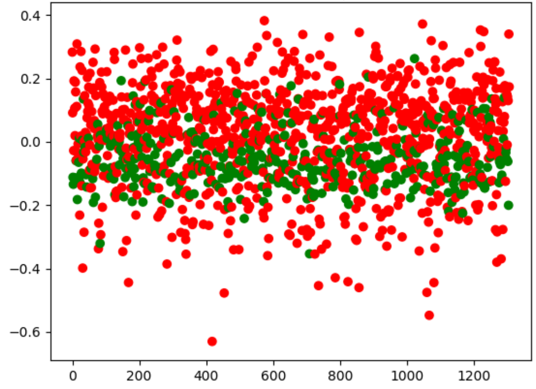

Figure 5: Kidney cancer: position cg04563996: example of very different variabilities and similar means.

Figures 4 shows example of methylation position in which difference between means of cancerous (red) and normal (green) samples is large, while deviation seam to be similar. On the other side, **Figure 5** shows example in which means are similar, but the deviation of cancerous points is much bigger. As visible in those four figures, **if there is difference in variation of samples, cancerous tissues are usually more various** which is already discussed in literature. However, **in literature, usually cancerous tissues are hypo-methylated, while in features selected by LASSO, we can see that chance of them being hypo-methylated comparing to normal tissue is 50%**.

Let's compare LASSO as better model to the state-of-the-art model [3]. LASSO performs up to 1% better than state-of-the-art model wrapped in robust procedure developed for this project, except for kidney cancer where results are similar. Also, R-squared value is up to 0.04 better than state-of-the-art, except for kidney where it's similar. **It is important to mention that LASSO and wrapped state-of-the-art model select similar features, but LASSO is more rigorous and selects less features.**

Additionally, wrapped state-of-the-art model isn't generalizable because log$_{10}$ of p-value and log$_{2}$ of fold change (FC) can differ a lot. So, it is required to plot volcano plot and ask expert to estimate bounds for each cancer separately. Finding transferable bounds is attempted. Examples of best such bounds are shown in **Figures 6 and 7**, however while they work pretty well for breast cancer, they work poorly for colorectal cancers, selecting too many features as extreme. Also, while log of p-value for breast cancer is maximum 20, log of p-value of colorectal cancer is more than 200. Finally, state-of-the art model requires pair of samples (normal-cancerous) from the same patient.

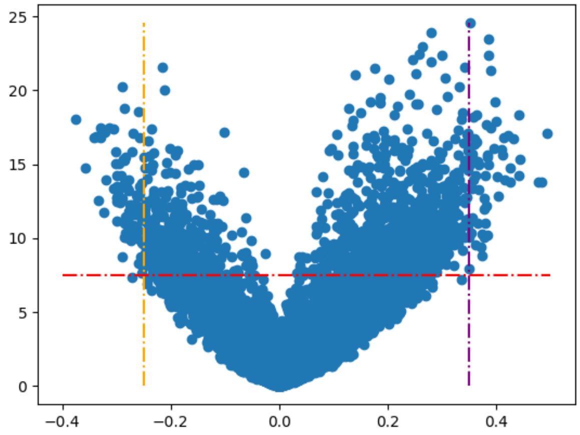

Figure 6: Breast cancer volcano plot

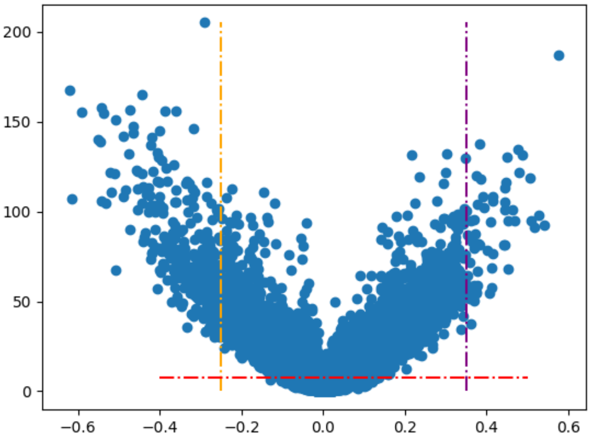

Figure 7: Colorectal cancer volcano plot

On the other side, from Table 1 we can see that average lambda values for different cancers are very close to each other. Ranging from 0.0017 for lung cancer to 0.0032 for colorectal cancer. Experimental results show that if the same lambda regularization (for example average of all cancers, 0.00267) is applied for each of the cancers, both classification accuracy and R-squared score don't change much.

This can be confirmed from R-squared curve given in **Figure 8** for LASSO model for colorectal cancer. Scale for x-axis is 1:2000, so actual values are 0.0005 till 0.0095. Whichever of those values we take as long as it's between 0.001 and 0.0095 we will get similar R-squared score (range of this score is (-inf, 1]). **Based on this evidence we can confirm that lamda is transferable between the cancers as long as sample size is the similar.**

Next problem with methylation data is that it's hard to obtain a lot of samples, so in this project we compared results of proposed methodology with state-of-the-art [3] wrapped in robust procedure developed in this project on very small datasets. Since state-of-the-art requires patients with both samples, those are extracted and from them 10, 20, 50 patients are randomly selected making 20, 40 and 100 samples in total. For proposed methodology 20, 40 and 100 samples are randomly selected from training data. This means that data selected for proposed methodology has same balance of classes as original data and doesn't require balanced datasets.

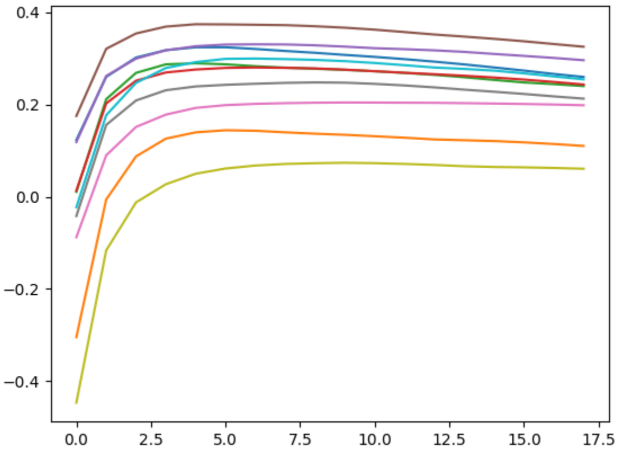

Figure 8: Colorectal cancer R-square curve (LASSO)

In **Table 2** we can see that **proposed methodology performs 5-30% better in classification and has R-square score up to 1.5 higher**. Most importantly difference between 20 sample and 1000 sample classification accuracy for proposed methodology is 5-10% only comparing to state-of-the-art for which difference is 15-35%. Selected features have smaller coefficients, so it is important to lower boundary for selecting features for smaller dataset to 0.02 or even 0.01. Most importantly, **lambda score is again similar between different cancers for the same size of the data** which means that same lambda can be used for different cancers, even for those cancers for which we don't have huge datasets to test on.

Lastly, once features are selected it is possible to do Cox analysis to predict time to live for patients. **Figures 9 and 10** show Cox prediction for the sample of 5 randomly chosen samples. Figure 9 shows results of Cox analysis using features selected by LASSO and Figure 10 results of Cox analysis using features selected by Elastic net. It is visible that better modeling or additional features are required to get excellent results, but if LASSO performs better. It is able to understand differences between people, while Cox-Elastic net has similar prediction for very different people. Also, LASSO accurately gives higher probability of survival to the alive patient.

	small state-of-the-art [3]						small l1					
	20 samples		40 samples		100 samples		20 samples		40 samples		100 samples	
	avg	max	avg	max	avg	max	avg	max	avg	max	avg	max
breast (classification)	59.187	78.048	67.520	76.016	75.406	79.675	**91.260**	93.089	**89.919**	93.089	**91.219**	93.496
breast (R-sq)	-1.360	-0.499	-1.310	-0.729	-0.988	-0.660	**0.018**	0.216	**0.063**	0.267	**0.190**	0.335
breast (lambda)	0.054		0.037		0.009		0.012		0.010		0.008	
colorectal (classification)	60.775	74.829	70.655	79.452	71.101	78.231	**89.154**	95.238	**88.140**	91.837	**90.709**	95.238
colorectal (R-sq)	-1.412	-0.764	-0.790	0.065	-0.886	-0.597	**0.065**	0.308	**0.125**	0.291	**0.222**	0.358
colorectal (lambda)	0.081		-0.092		0.017		0.020		0.018		0.008	
kidney (classification)	80.381	89.655	83.045	88.077	83.662	88.462	**85.278**	92.720	**88.719**	92.720	**93.829**	96.935
kidney (R-sq)	0.358	0.522	0.446	0.573	0.452	0.539	**0.428**	0.609	**0.500**	0.558	**0.620**	0.673
kidney (lambda)	0.015		0.010		0.008		0.012		0.011		0.010	
lung (classification)	73.933	90.717	78.689	84.322	81.306	85.897	**90.335**	94.390	**91.247**	94.421	**92.451**	96.507
lung (R-sq)	-1.105	0.684	-0.688	-0.242	-0.768	-0.096	**0.188**	0.343	**0.261**	0.475	**0.357**	0.446

Table 2: Classification accuracy and R-square score results for small datasets for state-of-the-art and proposed methods

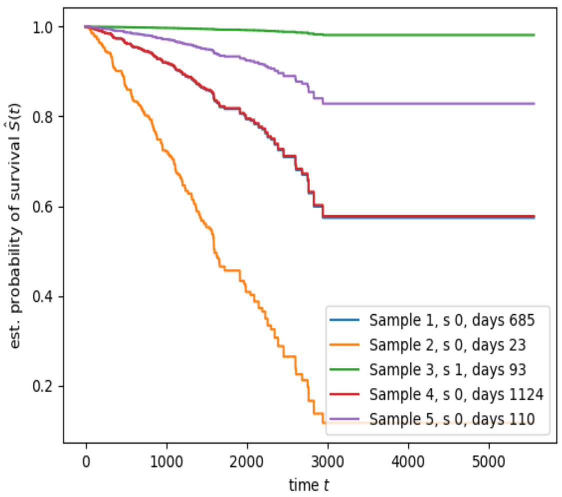

Figure 9: Kidney cancer Cox-LASSO time to survival

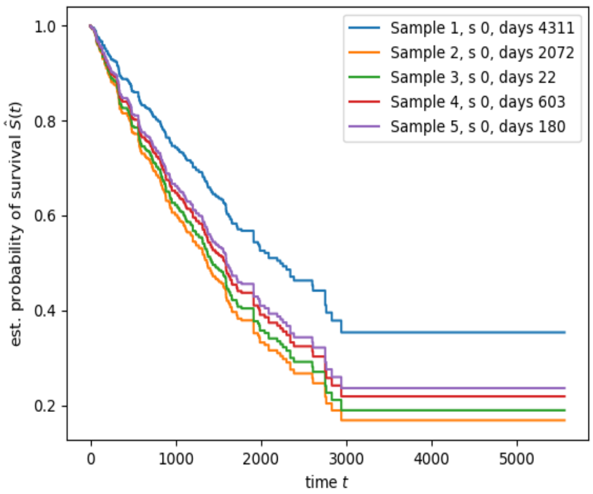

Figure 10: Kidney cancer Cox-Elastic net time to survival

Conclusion and future work

Proposed method has highest accuracy comparing to the state-of-the-art or to models in which LASSO regression is substituted with Elastic net or Ridge regressions. Also, when variance of each feature is added to the dataset, performance drops for few percent, meaning that enforcing variance into data is not beneficial for this model.

Regularizaton value depends mostly on sample size, but it can be used on different cancers as long as sample size is same. This is beneficial for detecting rare cancers for which there is no way to execute whole process and determine best regularization. There is indication that some selected features are appearing in different cancers. Understanding and interpretation of this indication is focus of future work.

It is shown that proposed method performs much better than state-of-the-art on classifying smaller datasets, which is beneficial for rare cancers or cancers subtypes. Also, it can help lowering costs of cancer research since data that has to be collected is much smaller. Additionally, proposed procedure doesn’t require to have normal and cancerous tissue from same patient and it can work even with imbalanced data.

Beside interpretation of selected features and possible similarity between selected features between the cancers, future work will focus on:

1. Adding clinical and biospecimen features to improve cox analysis.
2. Numerical evaluation of Cox analysis performance.
3. Development of more complex model to predict time to live with missing data. If possible model should optimize for Cox and Lasso regressions in the same time.
4. Understanding which features contribute to Cox stratification.

References

[1] American Cancer Society (2017) Cancer Facts &amp; Figures, https://www.cancer.org/content/dam/cancer-org/research/cancer- facts-and- statistics/annual-cancer-facts- and-figures/2017/cancer- facts-and- figures-2017.pdf (accessed on 03/30/2017)

[2] National Cancer Institute (NIH) (2017) Cancer Statistics, https://www.cancer.gov/about-cancer/understanding/statistics (accessed on 03/30/2017)

[3] Wei, J. H., Haddad, A., Wu, K. J., Zhao, H. W., Kapur, P., Zhang, Z. L., ... &amp; Wang, B. (2015). A CpG-methylation- based assay to predict survival in clear cell renal cell carcinoma. Nature communications, 6, 8699.

[4] Xu X., Su S., Barnes V. A., De Miguel C., Pollock J., Ownby D., Shi H., Zhu H., Snieder H., Wang X. (2013), A genome-wide methylation study on obesity: differential variability and differential methylation, Epigenetics 8(5):522–533

[5] Stirzaker C., Taberlay P. C., Statham A. L., Clark S. J. (2013), Mining cancer methylomes: prospects and challenges, Trends Genet, 30(2):75–84.

[6] Wittenberger T., Sleigh S., Reisel D., Zikan M., Wahl B., Alunni-Fabbroni M., et al. (2014), DNA methylation markers for early detection of women's cancer: promise and challenges, Epigenomics, 6:311–27.

[7] Caviglia G. P., Cabianca L., Fagoonee S. et al (2016), Colorectal cancer detection in an asymptomatic population: faecal immunochemical test for haemoglobin vs. faecal M2-type pyruvate kinase, Biochem Med (Zagreb) 26:114–120

[8] Maze I. et al. (2014), Analytical tools and current challenges in the modern era of neuroepigenomics, Nat. Neurosci. 17, 1476–1490

[9] Hebestreit K., Dugas M., Klein H. U. (2013), Detection of significantly differentially methylated regions in targeted bisulfite sequencing data, Bioinformatics, 29:1647–1653

[10] Wilhelm-Benartzi C. S. et al. (2013), Review of processing and analysis methods for DNA methylation array data, Br. J. Cancer 109, 1394–1402

[11] Kuan P. F., Song J., He S. (2017), methyIDMV: Simultaneous detection of differential DNA methylation and variability with confounder adjustment, Pacific Symposium on Biocomputing

[12] Schubeler D. (2015), Function and information content of DNA methylation, Nature, 517 (7534):321– 326

[13] Plongthongkum N., Diep D. H. & Zhang, K. (2014), Advances in the profiling of DNA modifications: cytosine methylation and beyond, Nat. Rev. Genet. 15, 647–661

[14] Vanderkraats N. D., Hiken J. F., Decker K. F. & Edwards, J. R. (2013), Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes, Nucleic Acids Res. 41, 6816–6827

[15] Razin, A. & Cedar, H. (1991), DNA methylation and gene expression, Microbiol. Rev. 55, 451–458 12

[16] Issa J. P. (2011), Epigenetic variation and cellular Darwinism, Nat. Genet., 43, 724-726

[17] Feinberg A. P. and Irizarry R. A. (2010), Stochastic epigenetic variation as a driving force of development, evolutionary adaptation and disease, Proc. Natl Acad. Sci. USA, 107, 1757-1765

[18] Bock C. (2012), Analysing and interpreting DNA methylation data, Nature Rev. Genet. 13, 705-719

[19] Teschendorff A. E. and Widschwendter M. (2012), Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions, Bioinformatics, 28, 1487-1494

[20] Kevin C. J. et al. (2014), Age-related DNA methylation in normal breast tissue and its relationship with invasive breast tumor methylation, Epigenetics, 9:2, 268-275

[21] Videtic-Paska A. and Hudler P. (2015), Aberrant methylation patterns in cancer: a clinical view, Biochemia Medica, 25(2), 161-176

[22] Teschendorff A. E. et al. (2016), DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer, Nat. Communication, 7:10478

[23] Bengtsoon H. et al. (2001), Identifying differentially expressed genes in cDNA microarray experiments authors, Sci. Aging Knowl. Environm., 2001, vp8

[24] Feinberg A. P. et al. (2010), Personalized epigenomic signatures that are stable over time and covary with body mass index, Sci. Transl. Med., 2, 49ra67

[25] Jaffe A. E. et al. (2012), Significance analysis and statistical dissection of variably methylated regions, Biostatistics, 13, 166-178

[26] Hansen K. D. et al. (2011), Increased methylation variation in epigenetic domains across cancer types, Nat. Genet., vol. 43 (pg. 768-775)