Network clustering of cancer patients based on DNA methylation variability

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 levels 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 understand how their variability is connected with health and cancerous tissues of colon cancer patients. For the first time in methylation analysis complex structure of network of patients is introduced. Feature selection in this paper is based on values of each methylation site and variability those sites have among patients.

Based on similarity of selected methylation positions patients are connected in network, which is then clustered. Each cluster represents group of patients that have similar variabilities on selected methylation positions. Since there are no studies using network for solving this problem, null model is created to show that network representation and clustering is meaningful.

As expected results present 2-4 clusters for each network we created and usually there is one smaller cluster that represents patients with common high methylation variability. It is easier to detect clusters for cancer tissues data. Feature selection influences patients' clusters the most, while clustering methodology doesn't play significant role.

Introduction

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.

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 [1], cancer [2], [3], [4] and neural system [5] 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 ([11], [6], [11], [7], [8]), analysis ([9], [10], [12]) or the whole process of working with methylation data ([12], [12]). (Figure 1). However, all of these methodologies use only simple statistic for analysis and aim of this paper is to introduce complex structure of network into inspection of methylation variability in patients.

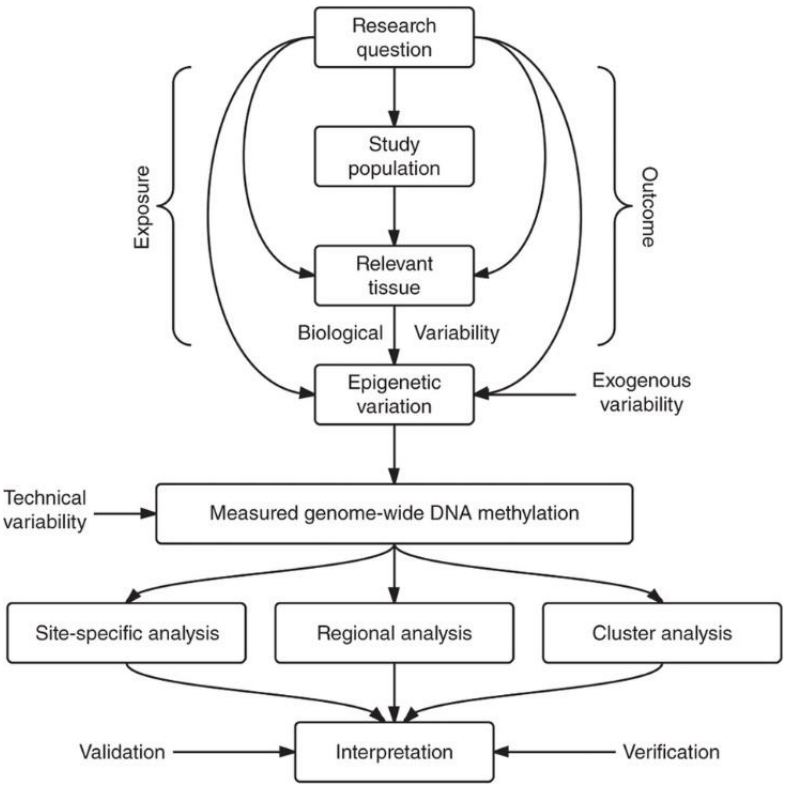

Figure 1: Steps toward a successful epigenome-wide study (EWAS) in cancer

This paper is doing further research of variability of DNA methylation levels in colon cancer patients using advanced data mining techniques and network science. In 2017, it is estimated that 1,688,780 cases will be diagnosed only in United States [21], what is 0.05% of population. Colorectal cancer is second leading cause for cancer-death and third most-diagnosed cancer in US [22]. Expenditures for cancer care in United States is expected to reach $156 billion in 2020 [23].

The research contributions are:

1. Creation of methods for features selection that can extract features with high differentiability in DNA methylation. This problem was tacked in existing literature, but approaches used rather groups of CpG positions than single sites.
2. Designing methodology for network representation and clustering of patients based on their methylation levels with evaluation of different approaches.
3. Showing that patients can be meaningfully partitioned into groups based on their methylation level.

Related work

It is shown in [16] that methylation level influences gene expression in animals. As mentioned in [17] hypervariable DNA methylation is related with a lack of order in gene expression in humans. Results of

[18] explain genetic apparatus for variable methylation and it shows that high variability of DNA methylation is related to evolution.

As written in [19] 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. Results of [20] have shown that cancer risk markers can be identified better if differential variability of DNA methylation is used instead of DNA methylation mean.

Many studies ([24], [25], [26]) discussed influence of genetics, behavior and environment (age, BMI) on cancer development through analysis of DNA methylation level. Research in progress is trying to understand relation between risk for getting tumor and age using DNA methylation. That association is especially visible in colon cancer which is usually happening in older people, since average age of diagnosis is 72.

Studies showed that difference of mean level of DNA methylation can help to identify those features that influence cancer strongly [27]. Models developed based on this discovery were successful in finding important features and understanding their influence on cancer development.

Recent papers ([18], [28], [17], [29], [20]) discussed that variance of DNA methylation is also important for understanding disease. Study [30] 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.

Methodology

To create meaningful network of patients and get clusters that would be made based on DNA methylation variability, method of five steps is created. All steps are described in subsections below. Networks are not used in explaining methylation values at patients till now, so there is no any methodology that can be used as reference. Because of this, in each step more different techniques are described, used and compared.

Features selection

Existing research is operating with mean and deviations of clusters of methylation values. Most recent methodologies for understanding DNA methylation and its relation to diseases, age or BMI was determined by analysis of differentially methylated regions (DMR) as mentioned in [31]. However, according to [31] there are numerous problems with grouping features, especially related to the fact that groups are not of equal size, so it is advised to use individual CpG site analysis which is case in this paper.

Purpose of this research is to group patients based on the variability in methylation level in order to understand if there is group of patients for which certain CpG positions are highly deviant and to investigate further similarities between those patients and difference between them and other patients.

However, most of the methylation features don't have high variability and all patients methylation values are close to the mean value. Therefore, it is important to select only those CpG sites that have high variability and/or elevated number of patients with very deviant methylation values for that position.

Methylation features were selected if they satisfied condition using next methodologies:

- Standard deviation higher than threshold (**ST**): $stddev > threshold$
- Standard deviation over mean higher than threshold (**SMT**): $\frac{stddev}{mean} > threshold$
- Number of people whose absolute variation is higher than 2 standard deviations is above threshold (**D**): $count(abs(deviation) > 2 * stddev) > threshold$

Using each of those techniques two different sets were created, one that contained around 10% of methylation features and another that contained around 0.5% of existing features. Percentages were chosen from previous research that initiated hypothesis for this research. That work experimented with total variance of patients if different number of features are selected. According to expected number of selected features closest threshold value was chosen with three decimal places.

Correlation

Two highest objectives for network are to keep as much information as possible from given methylation values and to understand if same patients are having high methylation variability for all methylation position. That is why correlation between patients' methylation arrays (that contain selected features) is chosen to represent strength of link between two patients.

Three different correlation methodologies are chosen, since as described in [32] their properties are similar but they measure different relations between variables.

Pearson correlation $\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$, where μ_X, μ_Y are mean and σ_X, σ_Y are standard deviation values for X and Y variables, which in this case are methylation arrays for patient. Pearson correlation assumes normally distributed variables and correlation measures linear relation between two variables. However, in literature [33] methylation features are rather presented with beta-binomial model and values are in [0, 1] range and not in [-∞, ∞] as expected for normal distribution. Therefore, even though Pearson correlation is the most used one, it might not be the best solution for this problem.

Spearman rank correlation $\rho_{X,Y} = 1 - \frac{\sigma \sum d_i^2}{n(n^2 - 1)}$, where d_i is difference between ranks of X_i and Y_i values (in this case methylation values between two people on the same site) and n is number of all methylation positions. There is no assumption about the distribution and this correlation measures monotonic relation between two variables.

Kendall tau rank correlation $\rho_{X,Y} = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$, where n_c is number of concordant (ordered in the same way), n_d is number of discordant (ordered differently) values X_i and Y_i on the same site and n is number of all methylation positions. It doesn't have assumptions about the distribution.

Network creation and normalization

Network node is a patient and link is correlation value of methylation arrays (containing only selected features) of the two patients that are connected with that link. For later processing, it is required for weights of links to be in range [0, 1], however all correlation metrics give results in range [-1, 1], therefore normalization of correlation is done using two approaches:

- $abs(corr)$, which gives high weights (≈1) to both direct and reverse strong correlations
- $\frac{1+corr}{2}$, which gives high weight (≈1) only to directly strongly correlated variables

Even though correlation of patient with itself is 1, elements on diagonal of network adjacency matrix is set to 0, since patient is not connected to itself.

Clustering

Two different graph clustering methods are used on each combination of previously mentioned methodologies for feature selection, correlation and normalization of weights.

Louvain clustering, as described in [35] is modularity based greedy algorithm with time complexity O (n log n). It starts by assigning each node to a separate cluster and then it evaluates gain in modularity if we move node i to cluster of its neighbor j. This is repeated for each i and j. Modularity gain is calculated using formula $\Delta Q = \left[\frac{\sum_{in}+2k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$ that changes little with change of i and j what makes this algorithm so fast. Result of algorithm are clusters that maximize formula for modularity $Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$.

This algorithm is used to determine optimal number of clusters for each network.

Spectral clustering algorithm defined in [36] is used to determine network clusters. Provided input is affinity matrix (A), which is given as adjacency matrix of network and number of clusters (k) which is learned from Louvain clustering. Another option was to run spectral clustering on adjacency matrix for different number of clusters and optimize modularity function for each clusters number and then chose best optimization. However, that approach would give different number of clusters than Louvain algorithm, so it is not considered.

Algorithm finds Laplacian matrix which is defined as $L = D^{-1/2}AD^{-1/2}$, where D is diagonal matrix whose elements represent degrees of nodes, then it finds k largest eigenvectors of L, orthogonal to each other

which create matrix X, which is then normalized to matrix Y. Finally, k-means algorithm is run on points, where each point is a row in Y matrix. Nodes of network are assigned to clusters based on results of k-means algorithm.

Both algorithms results are non-overlapping communities.

Evaluation

Evaluating graph clusters in unsupervised way is hard problem and there is no single best evaluation metric, so in this paper three often used and complementary evaluation techniques are utilized. All of them are described in [36] and short overview is given below.

Modularity maximizes function $Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$ and gives output in range [-1, 1] where higher value is better. This is most complex optimization and often used for communities' evaluation. However, it is also optimization function for Louvain algorithm, so Louvain algorithm will always perform better than spectral algorithm for this evaluation methodology.

Conductance minimizes function $\varphi(G) = avg(\varphi(Ci))$, where C_i are clusters, $\varphi(C_i) = \frac{\sum_{u\in C_i}\sum_{v\in C_i} w(u,v)}{\min(a(C_i), a(\bar{C}_i))}$, $a(C_i) = \sum_{u\in C_i}\sum_{v\in C_i} w(u,v)$ and $\bar{C}_i$ is complement of C_i. Conductance measure ratio of sum of weights of links between clusters and sum of weights of links in cluster. It can be in range [0, ∞] and smaller results refer to better clustering.

Coverage maximizes function $\mathrm{cov}erage(C) = \frac{w(C)}{w(G)}$, where $w(C) = \sum_{i=1}^{k} w(E(v_x, v_y))$ for each v_x and v_y from cluster C_i, where G represents whole graph. Coverage measure ratio between weights in certain cluster and weights in the whole graph and final result is average of coverage values for each cluster. Range of results is [0, 1] and higher value means that clustering is better.

Results

Methodology is applied on methylation data from colon cancer, on both health and cancer tissues. All data preprocessing and methodologies implementation is done in python, using PyCharm Community Edition 2016 IDE on machine with 64GB RAM and Intel® Core™ i7-6700 CPU with 3.41 GHz speed. Whole procedure lasts few hours if it is run sequentially, but most of the work can be parallelized. Speed is not optimized because it was not objective of this project.

Data preprocessing

Data is taken from TCGA project for colon cancer patients using GDC Data Portal API. Beside methylation values, biospecimen and clinical supplement files were downloaded since they contain information about samples and patients from which methylation is measured, respectively. For the purpose of study it was important to work with both methylation values of health and cancer tissues for same patients.

Only 75 patients had data from health tissue, so for them data from health and cancer tissues are taken, making 150 tissues in total.

Analysis is done separately on health and cancer tissue methylation data. Some of those patients contained multiple samples from cancer tissues, but only the first sampling is taken. Files contained 485578 or 27579 methylation positions, so intersection is taken. In further preprocessing, methylation sites from sex chromosomes X and Y are removed as well as positions that had more than 20% of missing values. In most cases, all values at those position were missing. The remaining missing values are imputed with mean for that feature.

After preprocessing, methylation arrays contained 22385 methylation sites per each patient sample. Biospecimen supplement data is used to understand which methylations belong to which samples and to which patients. Clinical data will be used to understand better who are the patients from the same community and how they differ from patients from other clusters.

Null model

In existing papers networks are never used for matching patients and understanding methylations, therefore, it is important to prove that network representation is meaningful. In that light, null model is created. Methylation values are randomly permuted for each site among patients and this data was used through the whole process of data selection, network creation and clustering. Networks created with random permutation are not meaningful as shown in Figure 2. Such networks usually give best modularity if they contain only one cluster (best for normalization $\frac{1+corr}{2}$) or if they contain high number of clusters (best for normalization $abs(corr)$). When we compare clustering evaluation results between null model and health/cancer networks, null model always performs much worse and even the best results have inferior performance.

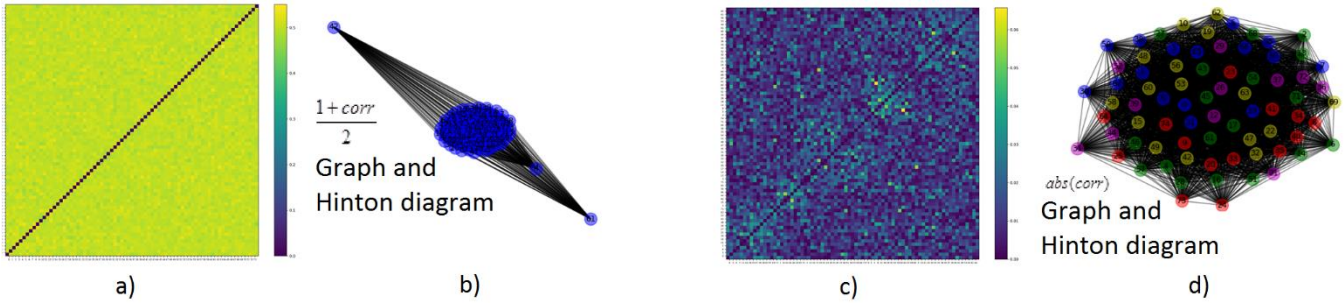

Figure 2: Graphs and Hinton diagrams for null model with different normalization values

General results

Results are shown using Hinton diagram, which is plot of weights in adjacency matrix, while patients order is in respect with their cluster. First are shown patients from first cluster, then from second and so on. Inside the clusters patients are ordered by their initial id. Patients order is shown on x and y axis. Weights are displayed with different colors where color is dark blue for weights 0, green for weight 0.5 and bright yellow for weight 1. In this representation squares around minor diagonal that are more

yellow than their surrounding represent clusters. Additionally, results are drawn as networks, where different node colors represent different clusters. Most of the networks pictures have “outlier” nodes which are consequence of drawing implementation in python and don’t have any meaning.

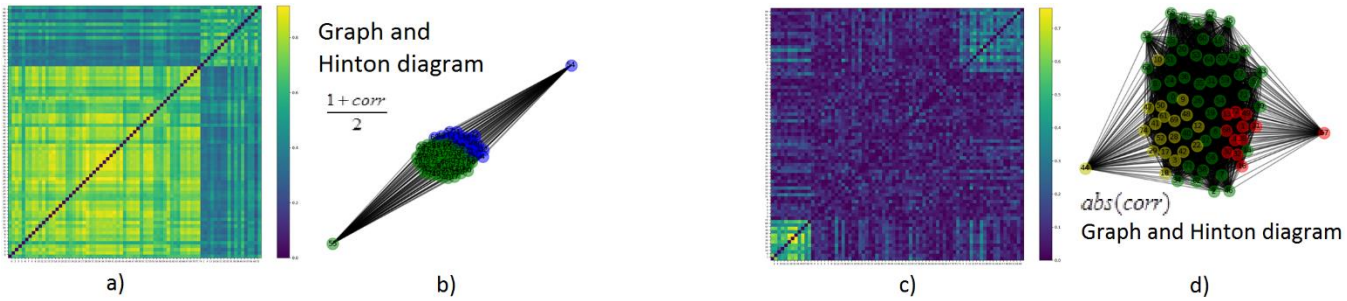

Figure 3: Graphs and Hinton diagrams for health tissues with different normalization values

Hinton graphs show that weights are bigger if we normalize correlations with $\frac{1+corr}{2}$ than if we normalize them with $abs(corr)$ in any case, including null models. As shown in Figure 3 and Figure 4 clusters are meaningful and there are always one or two small clusters that present people with high deviations. Those are sometimes part of bigger clusters, as in example of Figure 4 a) right top corner.

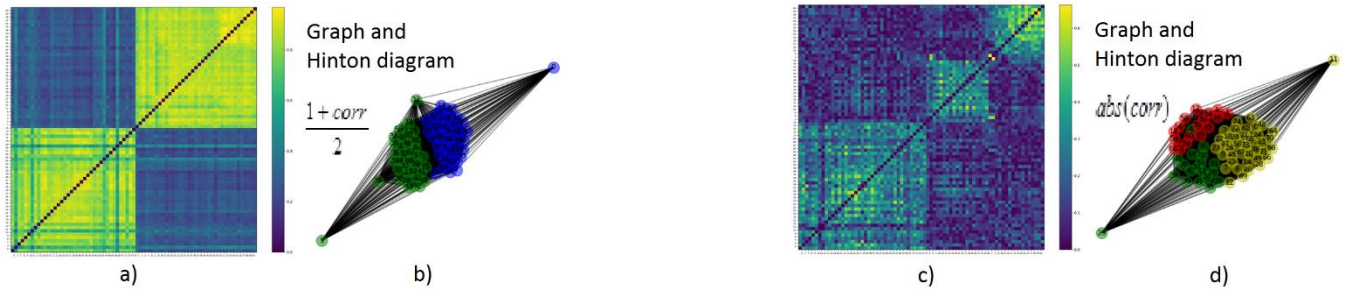

Figure 4: Graphs and Hinton diagrams for health tissues with different normalization values

From figures and cluster evaluation results, it can be concluded that cancer methylation values are forming better clusters, weights are more substantial inside cancer tissue data clusters than inside the health tissue data clusters.

In

Table 1, results are represented for random, health and cancer data for each normalization type, because research showed that normalization type influences highly best clustering (columns). For each of abovementioned tissue and normalization types, data is shown for both clustering and all three evaluation methods (rows).

A cell contain best value for certain clustering method, evaluation method, tissue type and normalization technique. The best value is chosen from results of three correlation techniques (Pearson, Spearman, Kendall), three different feature selection methodologies (ST, SMT, D) and two sizes of feature selection

sets (10% and 0.5%). For random data sometimes whole network is one cluster, in which case coverage is 1, which is the best possible value, but this clustering is not useful.

Table 1: Best clustering results for random, health and cancer tissues clustered and measured with different techniques

Clustering + Evaluation Method	**Null model** $abs(corr)$	**Null model** $\frac{1+corr}{2}$	**Health tissue** $abs(corr)$	**Health tissue** $\frac{1+corr}{2}$	**Cancer tissue** $abs(corr)$	**Cancer tissue** $\frac{1+corr}{2}$
Modularity (Louvain)	0.099249, Pearson, 0.5%, SMT	0.005466 Pearson, 0.5%, D	0.162885 Kendall, 0.5%, ST	0.065379 Pearson, 10%, SMT	0.204630 Kendall, 0.5%, D	0.243234 Spearman, 0.5%, SMT
Conductance (Louvain)	4.745966 Spearman, 10%, D	1.982202 Pearson, 0.5%, SMT	1.393586 Kendall, 0.5%, SMT	1.622324 Pearson, 10%, ST	1.091775 Kendall, 10%, ST	0.690853 Spearman, 0.5%, SMT
Coverage (Louvain)	0.348093 Kendall, 10%, D	1 – single cluster Multiple	0.642946 Kendall, 10%, D	0.661576 Spearman, 10%, SMT	0.694147 Kendall, 10%, ST	0.743313 Spearman, 0.5%, SMT
Modularity (Spectral)	0.075320 Pearson, 0.5%, SMT	0.004795 Pearson, 0.5%, D	0.149705, Kendall, 0.5%, ST	0.064327 Pearson, 10%, SMT	0.202401 Kendall, 0.5%, D	0.243234 Spearman, 0.5%, SMT
Conductance (Spectral)	4.485229 Kendall, 10%, D	1.978286 Pearson, 0.5%, ST	1.393586 Kendall, 0.5%, SMT	1.657455 Pearson, 10%, SMT	1.057027 Kendall, 10%, ST	0.690853 Spearman, 0.5%, SMT
Coverage (Spectral)	0.436908 Pearson, 0.5%, SMT	1 – single cluster Multiple	0.593260 Kendall, 0.5%, SMT	0.744199 Spearman, 10%, SMT	0.66919 Spearman, 10%, ST	0.743313 Spearman, 0.5%, SMT

Discussion

All results for health or cancer data divide network in 2-4 clusters as expected. Normalization $\frac{1+corr}{2}$ gives always better clustering results for cancer tissues, while $abs(corr)$ gives better results for health tissue, except if evaluated by coverage metric. This can be explained by the fact that cancer tissue methylation data is more deviant and it is therefore easier to preselect important features and correlation coefficients are higher and $\frac{1+corr}{2}$ normalization metric gives enough information for making good clusters. It also mean that for cancer sample methylation data, patients that are negatively correlated shouldn't be grouped, but rather present different classes of people and might have different biological processes. On the other side, it looks like health tissues data is not enough deviant, so $abs(corr)$ normalization metric helps patients that are strongly or reversely correlated to separate out.

Those groups should be carefully examined from biological point of view in order to understand if all those people methylation values are deviant in same way.

Feature selection method SMT (standard deviation over mean is higher than threshold) gives best results in 19 out of 36 cases which means that feature selection method is important for clustering. When it comes to number of features, choosing 0.5% or 10% of features seems to be best option in similar number of cases.

Best values are most often created by Kendall correlation (11/24), than by Spearman correlation (9/24) and least often by Pearson correlation (4/24) if we exclude null model results. Inferior results of Pearson correlation can be explained by its requirement for normal distribution which is not distribution of methylation data.

Both clustering methods give similar best results under all evaluation metrics, so we can conclude that network is robust and clustering method is the least important factor for getting good communities. It is important to say here that evaluation metrics are complemental, chosen to show different characteristics of resulting communities.

Future work

This is first research in which network is used to explain methylation values between people and to understand variability of methylation sites. Also, there is no good features selection method, so more research can be done in that area.

There is place for more data analysis from perspective of graphs as well as for improvement of methodology. Concretely, it would be interesting to see clustering of patients based on their methylation values and compare results with network clustering results in order to understand if and how much network structure improves communities detection.

Additionally, it would be interesting to get results using more different clustering and evaluation (such as NMI, RI, F and Silhouette) methodologies.

Since features selection appears to have the highest influence on results, it is important to understand them in more details and to find even better techniques for their extraction. One option would be to apply PCA that would transfer n-dimensional methylation arrays into 2-dimensional arrays or t-SNE. This can help us understand better how many features should be selected and what features that are correlated.

It is important to understand which features are commonly selected by different selection methodologies and what is the overlap of features and to further examine those methylation positions from biological point of view. Additionally, it is important to understand similarities between patients in same clusters with especial focus on smaller clusters of people and commonly highly deviant methylation values for this group.

Different combinations of feature selections, number of selected features and correlation methodologies give best results in distinct cases. Combining all those options and finding best clustering using multiple graphs can be done using multi-graph clustering techniques.

References

[1] 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

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

[3] 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.

[4] 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

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

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

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

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

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

[10] 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

[11] Cruickshanks H. A., McBryan T., Nelson D. M., Vanderkraats N. D., Shah P. P., van Tuyn J., Singh Rai T., Brock C., Donahue G., Dunican D. S. et al. (2013), *Senescent cells harbor features of the cancer epigenome*, Nat Cell Biol., 15, 1495–1506

[12] 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

[13] Mensaert K., Denil S., Trooskens G., Van Criekinge W., Thas O., De Meyer T. (2014), *Next-generation technologies and data analytical approaches for epigenomics*, Environ Mol. Mutagen., 55:155–70.

[14] Assenov, Y. et al. (2014), *Comprehensive analysis of DNA methylation data with RnBeads*, Nat. Methods 11, 1138–1140

[15] Michels K.B., Binder A. M., Dedeurwaerder S., Epstein C. B., Greally, J. M., Gut I., Houseman E. A., Izzi B., Kelsey K.T., Meissner A., et al. (2013), *Recommendations for the design and analysis of epigenome-wide association studies*, Nat. Methods, 10, 949–955

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

[17] Issa J. P. (2011), *Epigenetic variation and cellular Darwinism*, Nat. Genet., 43, 724-726
[18] 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
[19] Bock C. (2012), *Analysing and interpreting DNA methylation data*, Nature Rev. Genet. 13, 705-719
[20] 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
[21] American Cancer Society (2017), *Cancer Facts & 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)
[22] *Colon Cancer Alliance, Statistics*, https://www.ccalliance.org/get-information/what-is-colon-cancer/statistics/ (accessed on 03/30/2017)
[23] National Cancer Institute (NIH) (2017), *Cancer Statistics*, https://www.cancer.gov/about-cancer/understanding/statistics (accessed on 03/30/2017)
[24] 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
[25] Videtic-Paska A. and Hudler P. (2015), *Aberrant methylation patterns in cancer: a clinical view,* Biochemia Medica, 25(2), 161-176
[26] 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
[27] Bengtsoon H. et al. (2001), *Identifying differentially expressed genes in cDNA microarray experiments authors*, Sci. Aging Knowl. Environm., 2001, vp8
[28] 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
[29] Jaffe A. E. et al. (2012), *Significance analysis and statistical dissection of variably methylated regions*, Biostatistics, 13, 166-178
[30] Hansen K. D. et al. (2011), *Increased methylation variation in epigenetic domains across cancer types*, Nat. Genet., vol. 43 (pg. 768-775)
[31] Robinson M. D., Kahraman A., Law C.W., Lindsay H., Nowicka M., Weber L.M., et al. (2014), *Statistical methods for detecting differentially methylated loci and regions*, Front Genet 2014, 5(324): eCollection, doi:10.3389/fgene.2014.00324
[32] Chok N. S (2010), *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data*, Master thesis, Graduate School of Public Health, University of Pittsburgh
[33] Lea A. J., Tung J., and Zhou X. (2015), *A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data*, PLoS Genet. 11, e1005650
[34] Blondel V. D., Guillaume J. L., Lambiotte R., Lefebvre E. (2008), *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008
[35] Ng A., Jordan M., and Weiss Y. (2001), On spectral clustering: analysis and an algorithm. In Advances in Neural Information Processing Systems 14 (NIPS-01)
[36] Almeida H., Guedes D., Meira W. Jr, Zaki M. J. (2011), *Is there a best quality metric for graph clusters?*, 15th European conference on principles and practice of knowledge discovery in databases