2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)

Compound Analytics using Combinatorics for Feature Selection: A Case Study in Biomarker Detection

Ronald D. Hagan[1], Brett D. Hagan[2], Charles A. Phillips[2], Bradley J. Rhodes[1], Michael A. Langston[2]

[1] BAE Systems FAST Labs^{TM}, Burlington, MA 01803
{ron.hagan,brad.rhodes}@baesystems.com

[2] Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville, TN 37996
{bhagan2, cphill25, langston}@tennessee.edu

Abstract— **Computer and data scientists are increasingly tasked with analyzing data growing at unprecedented rates. These data frequently involve a high level of dimensionality. In this work, we present a novel method for dimension reduction that combines statistical scoring with graph theoretical filtering to distill salient features for machine learning. We apply this method to the timely problem of detecting epigenetic biomarkers in DNA methylation data.**

Keywords—dimension reduction; feature selection; epigenetics; machine learning; graph theoretical algorithms

I. Introduction

Technological advancements over the past decade have resulted in an explosion of data collection techniques. A rapid increase in the number of active sensors and the rise of the internet of things have combined to create exponential growth in data production across the world, with recent estimates putting the amount of data produced daily at 2.5 quintillion bytes [1]. Further complicating the matter, many of these data sets involve a high degree of dimensionality, especially in problem areas such as bioinformatics, image processing, and text translation. The development of effective dimension reduction techniques is therefore key for researchers in today's world of Big Data.

Dimension reduction techniques can be divided into two general classes, feature selection and feature extraction [2]. Feature extraction consists of transforming known features into a lower dimensional space. Examples include principal component analysis, independent component analysis, and machine learning methods such as autoencoders. Unfortunately, mapping to an abstract feature space makes it difficult to interpret data characteristics and the transform itself can be an expensive operation. Feature selection, on the other hand, consists of selecting a subset of known features without transformation. These techniques can be further divided into two broad categories: filters and wrappers. Filters are independent of downstream model selection and include classical statistical methods as well as newer approaches such as Markov Blanket Filtering [3]. Although efficient to implement, filters are generally unable to capture high level dependencies between features. Wrapper methods such as forward feature selection [4] perform selection based on classification accuracy across subsets of features using a preselected learning algorithm. As such, wrappers tend to be computationally expensive.

In previous work, we described a conceptual framework for the analysis of complex data sets that combines machine learning and graph analytics techniques [5]. This modular approach to compound analytics, illustrated in Figure 1, seeks to leverage potential synergies to reveal subtle interactions and patterns that might otherwise remain hidden to a singular modality. As part of our framework, we have developed a novel method for feature selection that combines a statistical scoring function with a filter derived from a classic graph covering problem, minimum dominating set. By using a graph based approach, we maintain computational efficiency while considering high level interactions between features overlooked by traditional filtering approaches.

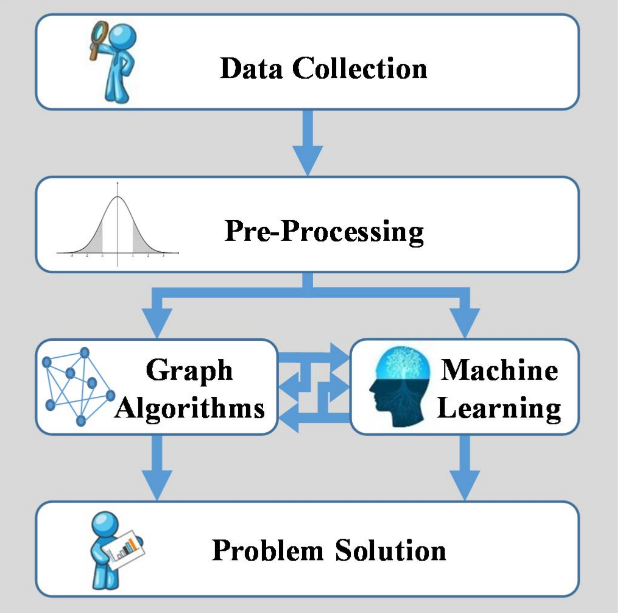

Figure 1. A framework for knowledge discovery. Graph algorithms and machine learning can be tailored to fit processing pipelines for a wide variety of applications.

978-1-5386-5555-9/19/$31.00 ©2019 IEEE
DOI 10.1109/IPDPSW.2019.00050

In this paper we present our method, together with the results of its application to the problem of detecting putative biomarkers in epigenetic data. As we shall see, testing shows that our method is capable of identifying a small set of core biomarkers (features) that allow for classification of tissue samples as case vs. control with high fidelity in a variety of diseases/conditions. The paper is organized as follows. In the next section, we provide some background on DNA methylation and its role in human disease. We then discuss our tools and methods for feature selection. In a fourth section, we present the results of our testing on publicly available data for osteoarthritis, breast cancer, liver cancer, and schizophrenia. Finally, we close with a brief summary.

II. DNA Methylation and Human Disease

When the Human Genome Project undertook its mission to map the entire DNA sequence of the human genome in 1990, it carried the hope of transforming our understanding of biology. In 2001 it was even chronicled in an episode of NOVA on PBS entitled "Cracking the Code of Life [6]." While certainly representing a great leap forward in our fundamental knowledge of genetics, it has become clear since its completion in 2003 that there are mechanisms at play in the actual expression of genes that go far beyond the physical arrangement of the underlying genetic code. While researchers have identified a variety of these epigenetic mechanisms, perhaps the most studied and well understood is DNA methylation.

DNA methylation generally occurs when a methyl group is added at the 5′ position of the cytosine ring, transforming the cytosine to 5-methylcytosine. Usually this occurs at CpG dinucleotides, although non-CpG methylation has been seen to occur more frequently in specific contexts such as neural development and in embryonic stem cells [7]. The process is believed to be regulated by DNA methyltransferases including DNMT1, DNMT3a, and DNMT3b. DNMT1 works to maintain methylation patterns by recognizing and copying them to the unmethylated daughter strands during DNA replication. DNMT3a and DNMT3b are thought to be responsible for *de novo* methylation events. Mutations in the DNMT3b gene have been found to be responsible for ICF (Immunodeficiency, centromeric instability, facial anomalies) syndrome [8], while mutations to any of DNMT1, DNMT3a, or DNMT3b have been found to be embryonically lethal in mice [9, 10].

In humans, some 70% of CpG dinucleotides throughout the genome are methylated [11]. At the same time, there are genomic regions with a heavy concentration of CpG content that can be found in the promoter regions of many genes. The cytosines in these CpG rich regions, termed CpG islands, tend normally to be unmethylated with exceptions in the context of the inactive X chromosome [12] and imprinted genes [13, 14]. Aberrant methylation patterns have been found to play a role in many diseases. In particular, it has been shown to play a dual role in many forms of cancer through both a pattern of global hypomethylation, allowing aberrant overexpression and ensuing oncogenesis, together with hypermethylation of CpG islands in the promoter regions of tumor suppressor genes, leading to their silencing [15-17]. These discoveries provide a compelling impetus for the development of methods for the discovery of novel methylation biomarkers capable of differentiating between healthy and diseased states. Such markers could then potentially be used in screening and diagnosis, and as guides for the selection of therapeutic targets for DNA-demethylating agents.

III. Methods and Tools

In this section, we describe our toolchain for the analysis of methylation data (as illustrated in Figure 2).

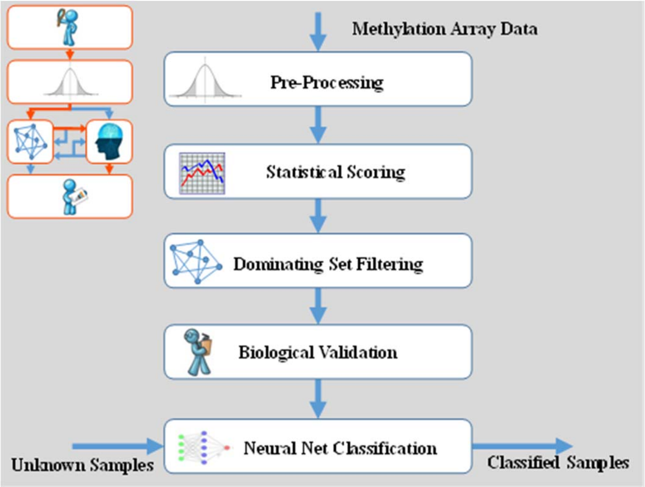

Figure 2. An illustration of our workflow for methylation analysis.

A. Statistical Scoring

To begin, each methylation site is assigned a merit score by means of the following function:

$$score(site\ i) = |\mu_i(case) - \mu_i(control)| - \alpha|\sigma_i(case) + \sigma_i(control)|$$

Where μ_i and σ_i are the mean and standard deviations, respectively, of the indicated sample group and α is a constant used as a tuning factor with $0 < \alpha \leq 1$. We start with $\alpha = 1$ and adjust it downward as necessary until we are able to identify sites with positive merit scores.

Ultimately the goal is to be able to identify sites that are capable of providing a clean separation between case and control. To that end, we next calculate inter-sample scores. The score comparing sample i and sample j is assigned via:

$$\sum score(site_k) \cdot \left(1 - |methylation_{value_{ik}} - methylation_{value_{jk}}|\right)$$

This metric is designed in such a way so as to favor homogeneous over heterogeneous sample pairs. Thus, case-case pairs and control-control matched pairs will tend to receive higher scores than mismatched case-control pairs.

B. Dominating Set Filter

Graph theoretical algorithms have found a wealth of applications across domains. The maximum clique problem for example has been used to mine putative gene networks associated with disease [18] and to detect fraudulent trading patterns in financial markets [19]. Covering problems such as vertex cover and dominating set have been used in applications ranging from image processing [20] to wireless ad-hoc routing [21]. These minimum covering problems in particular focus on finding a reduced set of nodes or edges that in some way capture the global structure of the network.

Formally, let $G=(V,E)$ be a simple, undirected graph with vertex set V and edge set E. A dominating set for G is a set of vertices $S \subseteq V$ such that for all $u \in V$ either $u \in S$ or there exists some $v \in S$ with $(u,v) \in E$. The minimum dominating set problem (MDS) then seeks to find, for a given input graph, a dominating set of minimum cardinality. Although MDS is *NP*-hard we have found in practice our implementation can reliably solve instances for graphs with thousands of nodes derived from real-world data in a few seconds.

Our scoring function has the potential to return a large number of sites with positive merit scores. We would like to be able to winnow these sites down to those with the best potential as discriminatory markers. In such situations, we apply a filter based on a variant of MDS, namely red-blue dominating set, in which the vertices are first colored red or blue, and then we seek the smallest set of red vertices that dominate the blue.

We first construct a bipartite graph in which one partite set contains red vertices representing sites and the other partite set contains blue vertices representing samples. For each site, we calculate the p-value of its observed methylation level for each sample. This p-value is calculated using the distribution of the levels at that site across all the group samples of the same type, be it case or control. A site is said to cover a sample and an edge is added between them in the graph if the p-value is greater than .05. This culls from the tails of the distributions and leaves us with observed methylation values that are in some sense "normal" for the sample within its group at each site.

Unfortunately, a straight application of minimum dominating set might sacrifice sites with high merit scores for those with lower discriminatory power based solely on the size of the returned set. To guard against this, we begin with the top scoring site and iteratively add the next highest scoring site until we obtain a dominating set. We then take a minimum set that dominates the graph from among this collection. In order to visualize the effectiveness of our reduced set in discriminating between case and control samples, we examine the distribution of the inter-sample scores.

C. *Machine Learning for Classification*

For the task of classification of samples as case or control (healthy vs. diseased), we employed boosted decision trees trained on the reduced feature set. Our models were implemented in Python using the popular sklearn package's built-in AdaBoost classifier. In all experiments described in the next section, we used 30 estimators and employed five-fold cross validation.

IV. Experimental Results

We applied our method to five sets of publicly available data obtained from the Gene Expression Omnibus (GEO). Datasets were chosen to span a variety of diseases. We selected only those containing a relatively large number of case and control samples. All the sets are from the Illumina Infinium HumanMethylation450 BeadChip array, often referred to as the Illumina 450k methylation array. The set of probes on the HM450 BeadChip targets over 450 thousand CpG sites across the human genome, covering not only promoters, but also gene bodies and untranslated regions. The data sets and some details of the results are summarized in Table 1.

A. *Osteoarthritis*

According to the Arthritis Foundation, osteoarthritis is the most common chronic condition of the joints. Sometimes called degenerative joint disease, it has no specific cause, but is influenced by several factors including age, occupation, obesity, injury and overuse. As well as having a known genetic component, several studies have been conducted that point to epigenetic mechanisms such as DNA methylation [22, 23] and histone modifications [24, 25]. The GEO series GSE63695 consists of methylation data from chondrocyte DNA samples drawn from the hip cartilage of 23 patients with osteoarthritis, knee cartilage of 73 osteoarthritis patients, and 21 hip samples from healthy controls. For the purpose of this study, we discarded the data from the knee cartilage in order to avoid possible confounding issues due to a mixture of tissue types.

With $\alpha = 1$, we identified 777 sites with positive merit scores for separation. The top scoring sites mapped to the genes ALX4, ANK1, and ARNT2. Differential expression of, or differential methylation in the promoter regions for, each of these genes has been indicated in the literature as playing a role in the development of osteoarthritis [26-28]. Our dominating set filter identified a set of three methylation sites that covered all samples. As seen in Figure 3, the homogeneous sample pairs exhibit a strong tendency to cluster toward the high end of the inter-sample scores, yielding a readily apparent separation. Training our classifier on this dataset achieves a mean accuracy of 0.933.

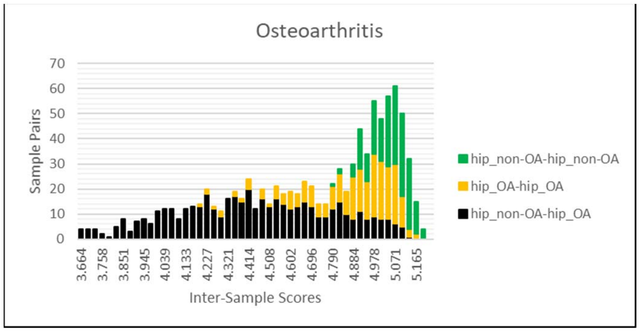

Figure 3. Stacked histogram of Osteoarthritis inter-sample scores. In this and following figures, green and gold indicate scores for homogeneous control-control and case-case pairs respectively. Black is scores for mismatched case-control pairs. Note that homogeneous scores cluster significantly to higher values.

Table 1. A summary of the data sets and results of our experimental testing. The mean accuracy reported is from classification testing using five-fold cross validation.

GEO Series Number	Disease	Case Samples	Control Samples	Number of Original Features	Features After Scoring and Dominating Set Filtering	Classification Mean Accuracy
GSE63695	Osteoarthritis	23	21	485,512	3	0.933
GSE66695	Breast Cancer	80	40	485,577	5	0.95
GSE54503	Liver Cancer	66	66	485,577	4	0.932
GSE61107	Schizophrenia	24	24	485,577	4	0.876
GSE40360	Multiple Sclerosis	28	19	481,917	7	0.871

B. Breast Cancer

Breast cancer in women accounts for one in ten of all new cancers diagnosed worldwide annually [29]. As with all cancers, DNA methylation is known to play a large role in its progression. A host of studies have been undertaken in efforts to improve our understanding of that relationship. For example, see [30-32]. Series GSE66695 consists of methylation data drawn from 40 normal and 80 breast cancer tissue samples.

In the breast cancer data, our scoring metric produced 12,107 sites with positive merit scores for $\alpha = 1$. The top scoring sites mapped to ZFP106, MXRA7, and the tumor suppressor gene ST7. MXRA7 has been found to be differentially expressed in a number of cancers [33]. We were able to uncover a dominating set consisting of five sites separating the data. The distribution of inter-sample scores shows a near total separation of the homogeneous and heterogeneous sample pairs, lending strong evidence to support the utility of our five sites as biomarkers for breast cancer. See Figure 4. This is further supported by the mean accuracy of 0.95 achieved by our classifier.

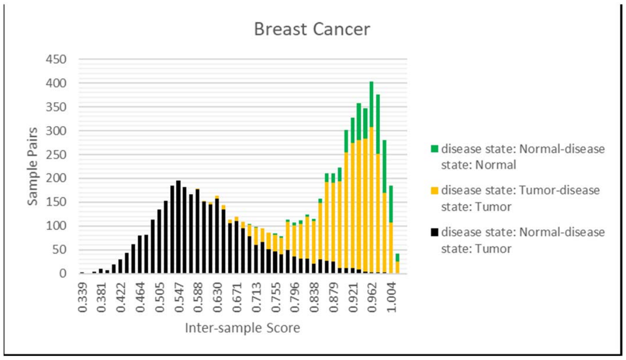

Figure 4. Distribution of Breast Cancer inter-sample scores.

C. Liver Cancer

The most common type of primary liver cancer is hepatocellular carcinoma (HCC). It ranks as the fifth most common type of cancer globally and is responsible for the third most deaths due to cancers. Despite its high global rankings, the distribution of cases is strongly centred in sub-Saharan Africa and Eastern Asia with China accounting for more than 50% of all cases worldwide [34]. GSE54503 is made up of methylation data drawn from 66 pairs of hepatocellular carcinoma (HCC) liver tumors and adjacent non-tumor tissues.

With $\alpha = 1$, our scoring produced a set of 30,576 methylation sites having positive merit scores. Dominating set filtering produced a set of four probes covering all samples. These sites map to the genes KCNQ2, C1orf70, GRASP, and PTPRN2. All four can be found in the literature as being involved in HCC, see [35-37]. As can be seen in Figure 5, the distribution of inter-sample scores again provides a nearly ideal separation between like and mixed sample pairs. Training on this set of four features achieved a mean classification accuracy of 0.932.

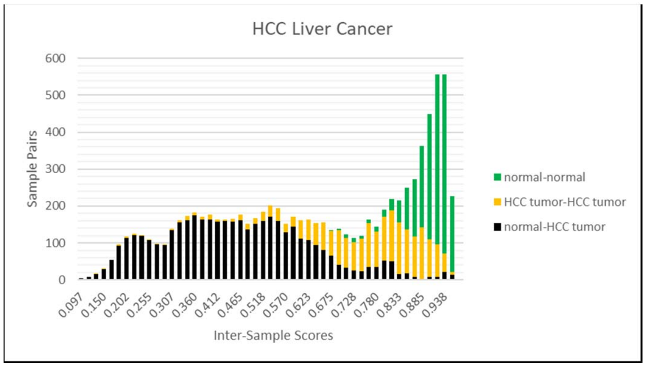

Figure 5. Distribution of HCC inter-sample scores.

D. Schizophrenia

GSE61107 comes from a genome-wide methylation analysis of brain tissue in schizophrenia patients [38]. It contains data drawn from frontal cortex post-mortem tissue from 24 individuals diagnosed with schizophrenia and 24 controls. The tissue samples themselves were provided by the Human Brain and Spinal Fluid Resource Centre.

With $\alpha = 1$, only four sites were identified with positive merit scores. Three of these sites mapped to TNRC6C, ZNF787, and HOXA13, while the fourth mapped to an intragenic region on chromosome 6. HOXA13 appears repeatedly as a potential biomarker for schizophrenia in the literature. See for example [39-41]. While the separation we obtain in this case is not to the level observed with the cancer datasets, we still observe a marked upshift in the distribution of homogeneous inter-sample scores as can be seen in Figure 6. As to be expected, our classifier does not perform quite as well on this set, but still achieves a mean accuracy of 0.876.

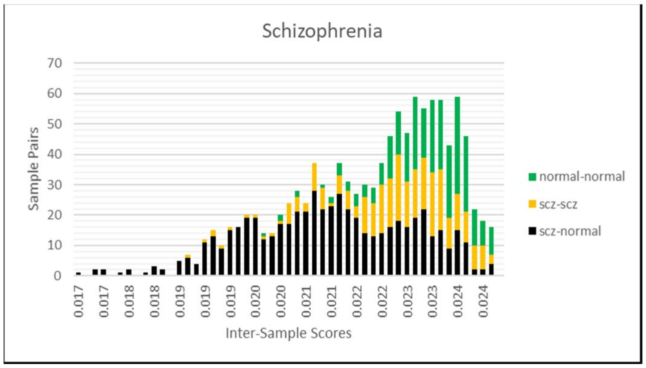

Figure 6. Distribution of Schizophrenia inter-sample scores.

E. *Multiple Sclerosis*

GSE40360 originated from a study seeking to identify differences in methylation patterns in pathology-free regions of brain tissue in persons affected by multiple sclerosis [42]. Drawn from brain bank samples of normal appearing white matter dissected from the frontal lobe, it consists of post-mortem samples from 28 multiple sclerosis patients as well as 19 healthy controls. This particular set turned out to be quite dirty, with numerous missing values. As such, an initial preprocessing step was required. We chose to discard records for all probes missing entries for a sample, leaving us with data for 460,421 probes.

This is the only one of the five datasets that returned no positive scores for a tuning factor of 1. Reducing to $\alpha = 0.9$, we obtained a set of seven sites with positive merit scores. As can be seen in figure 7, we start to see decreased separation in the distribution commiserate with the need to lower α. Notice, however, that the homogeneous sample pairs still produce scores that fall primarily in the top third of the distribution; our classifier still performs well with a mean accuracy of 0.871.

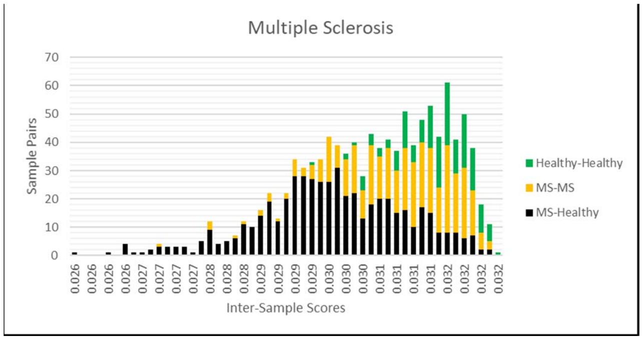

Figure 7. Distribution of Multiple Sclerosis inter-sample scores.

V. Conclusion

In this paper we described a novel method for dimension reduction that employs a statistical scoring function together with a red-blue dominating set filter. This method again reinforces the advantages of our conceptual framework encouraging a compound analytics approach combining both graph algorithms and traditional machine learning. We demonstrated the efficacy of our method by using it for biomarker detection in DNA methylation data.

Extracting low-dimensional feature sets for the training of boosted decision tree models for the classification of tissue samples from five different diseases demonstrated that the method is capable of identifying a small set of training features allowing for classification at a high fidelity with average mean accuracies ranging from .871 to .95 when testing with five-fold cross validation.

VI. References

[1] Bernard Marr. 2018. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. (July 2018). Retrieved October 13, 2018 from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#6240eb9560ba

[2] Pavel Pudil and Jana Novovičová. 1998. Novel Methods for Feature Subset Selection with Respect to Problem Knowledge. Feature Extraction, Construction and Selection (1998), 101–116. DOI:http://dx.doi.org/10.1007/978-1-4615-5725-8_7

[3] Zena M. Hira and Duncan F. Gillies. 2015. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics* 2015 (2015), 1–13. DOI:http://dx.doi.org/10.1155/2015/198363

[4] Francisco Macedo, M.Rosário Oliveira, António Pacheco, and Rui Valadas. 2018. Theoretical Foundations of Forward Feature Selection Methods based on Mutual Information. Neurocomputing (2018). DOI:http://dx.doi.org/10.1016/j.neucom.2018.09.077

[5] Hagan, R.D., Phillips, C.A., Rhodes, B.J ., and Langston, M.A. Compound Analytics: Templates for Integrating Graph Algorithms and Machine Learning. *IPDPS Workshops 2017*, 1550-1556.

[6] Krulwich, R. and LANDER, E. *Cracking the Code of Life*. Public Broadcasting Service, City, 2001.

[7] Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P. and Jaenisch, R. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences*, 97, 10 (2000), 5237-5242.

[8] Hansen, R. S., Wijmenga, C., Luo, P., Stanek, A. M., Canfield, T. K., Weemaes, C. M. and Gartler, S. M. The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. *Proceedings of the National Academy of Sciences*, 96, 25 (1999), 14412-14417.

[9] Okano, M., Bell, D. W., Haber, D. A. and Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99, 3 (1999), 247-257.

[10] Li, E., Bestor, T. H. and Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69, 6 (1992), 915-926.

[11] Strichman-Almashanu, L. Z., Lee, R. S., Onyango, P. O., Perlman, E., Flam, F., Frieman, M. B. and Feinberg, A. P. A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome research*, 12, 4 (2002), 543-554.

[12] Yen, P. H., Patel, P., Chinault, A. C., Mohandas, T. and Shapiro, L. J. Differential methylation of hypoxanthine phosphoribosyltransferase

genes on active and inactive human X chromosomes. *Proceedings of the National Academy of Sciences*, 81, 6 (1984), 1759-1763.

[13] Razin, A. and Cedar, H. DNA methylation and genomic imprinting. *Cell*, 77, 4 (1994), 473-476.

[14] Barlow, D. P. Gametic imprinting in mammals. *Science*, 270, 5242 (1995), 1610.

[15] Jones, P. A. and Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nature reviews genetics*, 3, 6 (2002), 415-428.

[16] Herman, J. G. and Baylin, S. B. Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine*, 349, 21 (2003), 2042-2054.

[17] Ehrlich, M. DNA methylation in cancer: too much, but also too little. *Oncogene*, 21, 35 (2002), 5400.

[18] Voy, B. H., Scharff, J. A., Perkins, A. D., Saxton, A. M., Borate, B., Chesler, E. J., Branstetter, L. K. and Langston, M. A. Extracting gene networks for low-dose radiation using graph theoretical algorithms. *PLoS Comput Biol*, 2 (2006).

[19] Wang, J., Zhou, S. and Guan, J. Detecting potential collusive cliques in futures markets based on trading behaviors from real data. *Neurocomputing*, 92 (2012), 44-53.

[20] Uyttendaele, M., Eden, A. and Skeliski, R. *Eliminating ghosting and exposure artifacts in image mosaics*. IEEE, City, 2001.

[21] Wan, P.-J., Alzoubi, K. M. and Frieder, O. Distributed construction of connected dominating set in wireless ad hoc networks. *Mobile Networks and Applications*, 9, 2 (2004), 141-149.

[22] Reynard, L. N., Bui, C., Canty-Laird, E. G., Young, D. A. and Loughlin, J. Expression of the osteoarthritis-associated gene GDF5 is modulated epigenetically by DNA methylation. *Human molecular genetics*, 20, 17 (2011), 3450-3460.

[23] Iliopoulos, D., Malizos, K. N. and Tsezou, A. Epigenetic regulation of leptin affects MMP-13 expression in osteoarthritic chondrocytes: possible molecular target for osteoarthritis therapeutic intervention. *Annals of the rheumatic diseases*, 66, 12 (2007), 1616-1621.

[24] Barter, M., Bui, C. and Young, D. Epigenetic mechanisms in cartilage and osteoarthritis: DNA methylation, histone modifications and microRNAs. *Osteoarthritis and cartilage*, 20, 5 (2012), 339-349.

[25] El Mansouri, F. E., Chabane, N., Zayed, N., Kapoor, M., Benderdour, M., Martel□Pelletier, J., Pelletier, J. P., Duval, N. and Fahmi, H. Contribution of H3K4 methylation by SET□1A to interleukin□1–induced cyclooxygenase 2 and inducible nitric oxide synthase expression in human osteoarthritis chondrocytes. *Arthritis & Rheumatology*, 63, 1 (2011), 168-179.

[26] Aref-Eshghi, E., Zhang, Y., Liu, M., Harper, P. E., Martin, G., Furey, A., Green, R., Sun, G., Rahman, P. and Zhai, G. Genome-wide DNA methylation study of hip and knee cartilage reveals embryonic organ and skeletal system morphogenesis as major pathways involved in osteoarthritis. *BMC musculoskeletal disorders*, 16, 1 (2015), 287.

[27] Chen, H.-C. *Genetics and Biomarkers of Osteoarthritis and Joint Hypermobility*. Duke University, 2009.

[28] Saito, T., Fukai, A., Mabuchi, A., Ikeda, T., Yano, F., Ohba, S., Nishida, N., Akune, T., Yoshimura, N. and Nakagawa, T. Transcriptional regulation of endochondral ossification by HIF-2 [alpha] during skeletal growth and osteoarthritis development. *Nature medicine*, 16, 6 (2010), 678-686.

[29] Ferlay, J., Héry, C., Autier, P. and Sankaranarayanan, R. Global burden of breast cancer. *Breast cancer epidemiology*. Springer New York, 2010.

[30] Huang, T. H.-M., Perry, M. R. and Laux, D. E. Methylation profiling of CpG islands in human breast cancer cells. *Human molecular genetics*, 8, 3 (1999), 459-470.

[31] Dobrovic, A. and Simpfendorfer, D. Methylation of the BRCA1 gene in sporadic breast cancer. *Cancer Research*, 57, 16 (1997), 3347-3350.

[32] Ottaviano, Y. L., Issa, J.-P., Parl, F. F., Smith, H. S., Baylin, S. B. and Davidson, N. E. Methylation of the estrogen receptor gene CpG island marks loss of estrogen receptor expression in human breast cancer cells. *Cancer Research*, 54, 10 (1994), 2552-2555.

[33] Pihur, V., Datta, S. and Datta, S. Finding common genes in multiple cancer types through meta–analysis of microarray experiments: A rank aggregation approach. *Genomics*, 92, 6 (2008), 400-403.

[34] El–Serag, H. B. and Rudolph, K. L. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology*, 132, 7 (2007), 2557-2576.

[35] Shen, J., LeFave, C., Sirosh, I., Siegel, A. B., Tycko, B. and Santella, R. M. Integrative epigenomic and genomic filtering for methylation markers in hepatocellular carcinomas. *BMC medical genomics*, 8, 1 (2015), 28.

[36] Shen, J., Wang, S., Zhang, Y.-J., Wu, H.-C., Kibriya, M. G., Jasmine, F., Ahsan, H., Wu, D. P., Siegel, A. B. and Remotti, H. Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. *Epigenetics*, 8, 1 (2013), 34-43.

[37] Yamada, N., Yasui, K., Dohi, O., Gen, Y., Tomie, A., Kitaichi, T., Iwai, N., Mitsuyoshi, H., Sumida, Y. and Moriguchi, M. Genome-wide DNA methylation analysis in hepatocellular carcinoma. *Oncology reports*, 35, 4 (2016), 2228-2236.

[38] Wockner, L., Noble, E., Lawford, B., Young, R. M., Morris, C., Whitehall, V. and Voisey, J. Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients. *Translational psychiatry*, 4, 1 (2014), e339.

[39] Vawter, M. P., Ferran, E., Galke, B., Cooper, K., Bunney, W. E. and Byerley, W. Microarray screening of lymphocyte gene expression differences in a multiplex schizophrenia pedigree. *Schizophrenia research*, 67, 1 (2004), 41-52.

[40] Pickard, B. S. Schizophrenia biomarkers: translating the descriptive into the diagnostic. *Journal of Psychopharmacology*, 29, 2 (2015), 138-143.

[41] Mamdani, F., Martin, M. V., Lencz, T., Rollins, B., Robinson, D. G., Moon, E. A., Malhotra, A. K. and Vawter, M. P. Coding and noncoding gene expression biomarkers in mood disorders and schizophrenia. *Disease markers*, 35, 1 (2013), 11-21.

[42] Huynh, J. L., Garg, P., Thin, T. H., Yoo, S., Dutta, R., Trapp, B. D., Haroutunian, V., Zhu, J., Donovan, M. J. and Sharp, A. J. Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains. *Nature neuroscience*, 17, 1 (2014), 121-130.