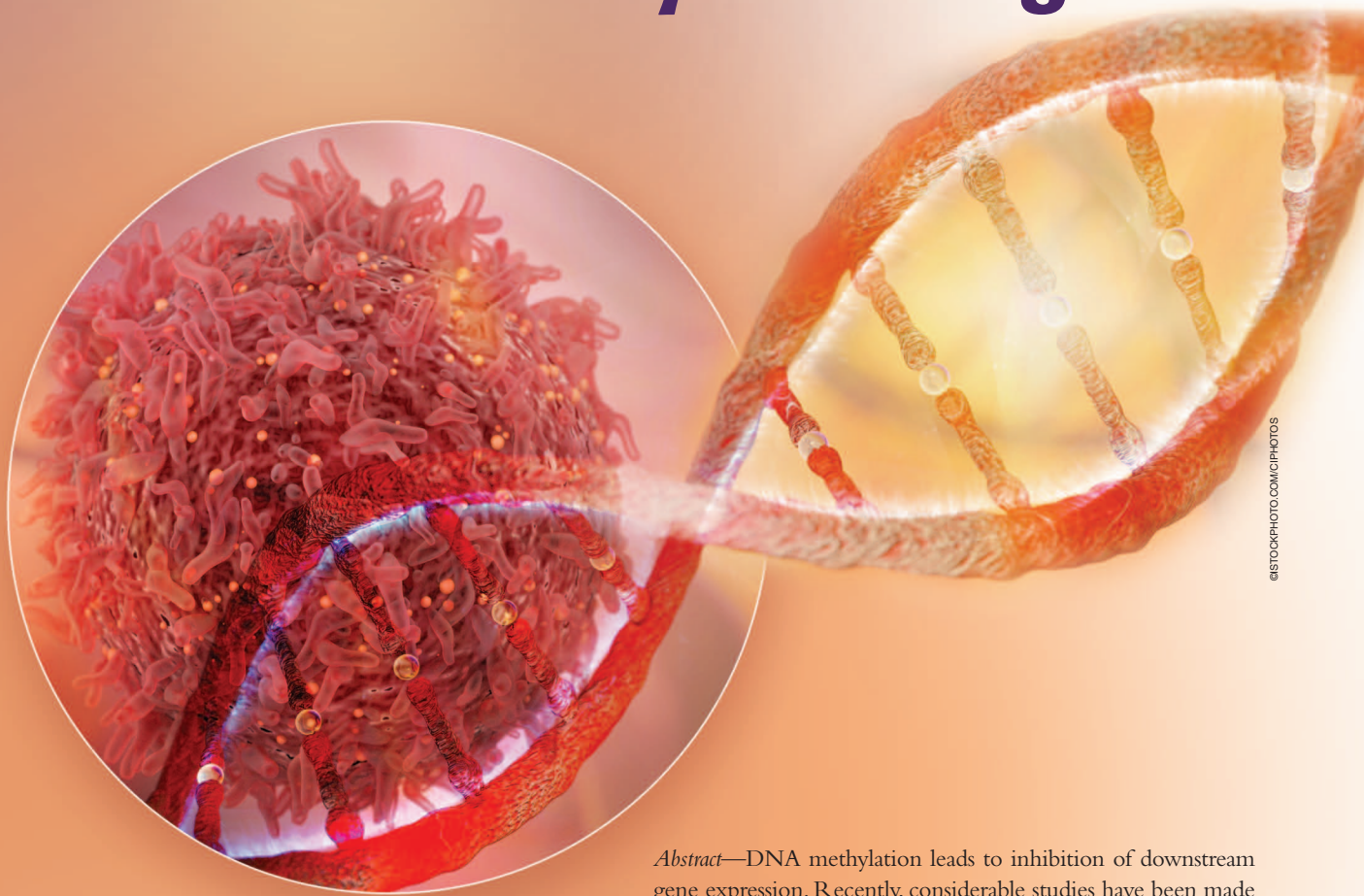
Identifying DNA Methylation Modules Associated with a Cancer by Probabilistic Evolutionary Learning



©ISTOCKPHOTO.COM/CIPHOTOS

Je-Keun Rhee
Cancer Research Institute, College of Medicine, Catholic University of Korea, Seoul, KOREA

Soo-Jin Kim
Research Institute of Agriculture and Life Sciences, College of Agriculture and Life Sciences, Seoul National University, Seoul, KOREA

Byoung-Tak Zhang
School of Computer Science & Engineering, Seoul National University, Seoul, KOREA

Abstract—DNA methylation leads to inhibition of downstream gene expression. Recently, considerable studies have been made to determine the effects of DNA methylation on complex disease. However, further studies are necessary to find the multiple interactions of many DNA methylation sites and their association with cancer. Here, to assess DNA methylation modules potentially relevant to disease, we use an Estimation of Distribution Algorithm (EDA) to identify high-order interaction of DNA methylated sites (or modules) that are potentially relevant to disease. The method builds a probabilistic dependency model to produce a solution that is a set of discriminative methylation sites. The algorithm is applied to array- and sequencing-based high-throughput DNA methylation profiling datasets. The experimental results show that it is able to identify DNA methylation modules for cancer.

1556-603X/18©2018IEEE

I. Introduction

Genomic studies mainly aim to find genetic markers that are associated with a phenotype. Based on DNA sequences, researchers have searched for causal effects on biological processes including gene regulatory mechanisms and diseases. Although several risk factors have been identified by the association studies, the genetic variants do not fully explain the abnormal regulation because the biological regulatory mechanism can be affected by many other factors, as well as DNA sequence modification [1]–[4].

Epigenomics refers to the study of regulation of various genomic functions that are controlled by another partially stable modification, but not DNA sequence variants [5]. Among these, DNA methylation, which typically occurs at CpG dinucleotides catalyzed by DNA methyltransferase, is a crucial epigenetic regulatory mechanism in cellular processes. DNA methylation of CpG sites mostly causes silencing of the downstream gene. The enrichment of the differentially methylated DNA fractions can contribute to specific abnormalities, including complex diseases [6]–[8]. In particular, with the advent of array and next generation sequencing (NGS) technology, many researchers have carried out genome-wide DNA methylation profiling studies [9]–[11], and the genome-wide studies have reported that many genomic regions are differentially methylated in normal and abnormal cells [12]–[14].

However, a complex disease is caused by a combination of dysegulatory effects of multiple genes [15]–[17]. That is, errors of biological processes are not caused by the alteration of an individual methylation level. Recently, Easwaran et al. hypothesized that DNA hypermethylation modules preferentially target important developmental regulators in embryonic stem cells [18]. They found a set of genes whose DNA methylation contributed to the stem-like state of cancer. Horvath et al. studied aging effects of DNA methylation and identified co-methylated modules related to aging in the human brain and blood tissue [19]. Zhang and Huang investigated the DNA co-methylation patterns frequently observed in cancer [20].

Here, we identify combinatorial modules of DNA methylation sites associated with human diseases using an evolutionary learning approach (Figure 1). Evolutionary algorithms can approximate solutions well for a variety of problems [21]–[25]. They generate a new population through iterative updates and selection using a guided search process in a feature space. We utilized an Estimation of Distribution Algorithm (EDA)-based learning approach to identify combinations of cancer-related DNA methylation sites. In the EDA, the population is evolved according to the probabilistic distribution in selected individuals without conventional genetic operators such as crossover and mutation. As a result, the EDA can provide answers in combinatorial optimization problems [26]–[29]. The EDA-based methods have been previously applied in several biological studies, and it has offered promising results for complex problems where other methods failed to find a good solution [30]–[32].

We investigated DNA methylation modules relevant to cancer, using the DNA methylation profiling datasets produced by array- and sequencing-based approaches. The experimental results showed that our method could identify DNA methylation modules related to cancer.

II. Methods

A. Evolutionary Learning Procedure to Identify a Set of DNA Methylation Sites Associated with a Disease

EDAs evolve a population to find the optimal solution probabilistically. The initial population is constructed by randomly selecting individuals. The individuals represent higher order interactions of the methylated sites. The population size m is decided empirically and the initial weight w_j of the individual $j(0 < j < m)$ is randomly assigned with a small value $(-1 < w_j < 1)$.

In the evolutionary process, each individual is evaluated for how discriminative the interaction is for the datasets. Better individuals are then selected and a dependency tree is built by fitting to the selected individuals. New individuals of the next generation are generated using the probability distribution within the tree structure, and replace the previous individuals. The overall procedure is as follows:

Step 1) Set $g \leftarrow 0$
Step 2) Initialize population $X(g)$ by random generation
Step 3) Evaluate individuals in $X(g)$
Step 4) Select a set of individuals by tournament selection from $X(g)$
Step 5) Construct a dependency tree $G(g)$ by measuring Kullback-Leibler divergence between variables
Step 6) Learn parameters using a probability distribution of the set of selected at step 4
Step 7) Generate new individuals by sampling with joint distribution from the $G(g)$, and create a new population $X(g+1)$
Step 8) Set $g \leftarrow g+1$
Step 9) If the termination criterion is not met, go to Step 3

Further details for steps 3 and 5 are explained in following sections.

B. Learning Dependency Tree

The dependency tree is built from the selected individuals by searching conditional dependencies between random variables. The model is then optimized by a series of incremental updates [33], [34], as follows:

Suppose that X is a population and $X = \{X_1, X_2, ..., X_n\}$ represents a vector of variables with n features, i.e., DNA methylation sites. The probability distribution is denoted by a joint probability $P(X_1, X_2, ..., X_n)$ as to:

Digital Object Identifier 10.1109/MCI.2018.2840659
Date of publication: 18 July 2018

Corresponding Author: Byoung-Tak Zhang (Email: btzhang@bi.snu.ac.kr)

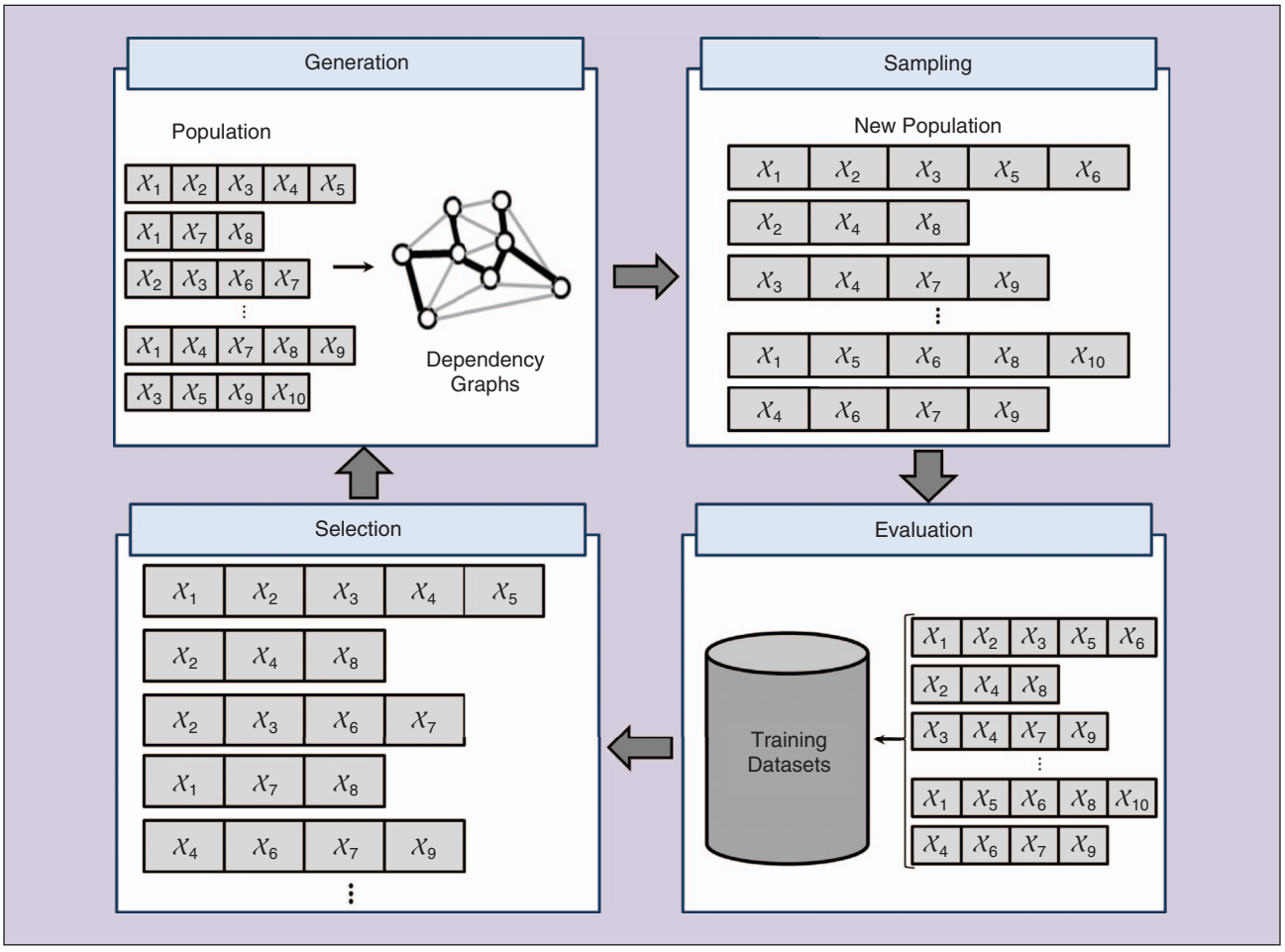


FIGURE 1 Schematic overview for probabilistic evolutionary learning to identify DNA methylation modules.

$$P(X) = P(X_1, X_2, ..., X_n) = P(X_1|X_2, ..., X_n)\,P(X_2|X_3, ..., X_n)\,....P(X_{n-1}|X_n)\,P(X_n). \quad (1)$$

However, it is hard to measure all the joint probabilities exactly when n, the number of variables, is large. Thus it is necessary to approximate the probability distribution. In this study, we used a dependency tree, and the distribution is approximated as follows:

$$P(X_1, X_2, ..., X_n) = P(X_r)\prod_{i \neq r} P(X_i|X_{pa(i)}), \quad (2)$$

where $X_1, X_2, ..., X_n$ are random variables, r is an index of a root node, and *pa(i)* denotes the index of the parent node of X_i. The tree structure is built by searching based on Kullback-Leibler divergence between two random variables. The dependency graph is constructed optimally in a direction to maximize total mutual information as follows:

$$\text{argmax}_{r,pa} \prod_{i \neq r} I(X_i; X_{pa(i)}), \quad (3)$$

$$I(X_i; X_{pa(i)}) = \sum_x \sum_y P(X_i = x, X_{pa(i)} = y) \log \frac{P(X_i = x, X_{pa(i)} = y)}{P(X_i = x)\,P(X_{pa(i)} = y)}. \quad (4)$$

The complete graph G searches the maximum spanning tree, and then the best dependency tree is constructed.

For parameter learning, the most likely values are calculated from the frequencies in the selected individuals. That is, the model parameters are represented as marginal probabilities in a root node and conditional probabilities in the other nodes. The marginal probabilities in the root nodes and the conditional probabilities in the child nodes are calculated by Eqs. (5) and (6), respectively, as follows:

$$P(X_r = x) = \frac{c(X_r = x)}{N}, \quad (5)$$

$$P(X_i = x|X_{pa(i)} = y) = \frac{c(X_i = x, X_{pa(i)} = y)}{c(X_{pa(i)} = y)}, \quad (6)$$

where c is the count of a variable X with a specific value and N is the total number of individuals.

C. Fitness Evaluation in a Population

The fitness function represents how informative the chromosome is to classify the samples. That is, the fitness for an individual is evaluated by measuring the classification accuracy for interaction of the features. To determine and update the fitness for each individual, we introduce a gradient descendant rule for training data $\mathbf{D}$ as follows:

$$w_i = w_i + \eta(t_j - f(\mathbf{D}_j))\, v_{ji}, \tag{7}$$

where w_i is the weight value for i-th feature and t_j is the target class in the j-th training instance $\mathbf{D}_j$. η is the learning rate and v_{ji} is the value of the i-th attribute in the j-th instance. $f(\mathbf{D}_j)$ is the predicted output value of the j-th training instance by our model and determined as follows:

$$f(\mathbf{D}_j) = \begin{cases} 1, & \text{if } \sum_{i=0}^{n} w_i \cdot v_{ji} > 0, \\ -1, & \text{otherwise.} \end{cases} \tag{8}$$

The difference between the predictions and the target values specified in the training sequence is used to represent the error of the current weight vector. The target function is optimized to minimize the classification error. The weight values are evaluated against a sequence of training samples and are updated to improve the classification accuracy. The weight update processes are repeated until they converge after a number of epochs.

Using the learning scheme, we identify the most informative individuals for classification, where the absolute values of their weights are large. In addition, it is better to find the DNA methylation module, whose number of features is small. Finally, the fitness function for the k-th individual X^k, $Fitness(X^k)$ is defined as follows:

$$Fitness(X^k) = Acc(X^k) - Order(X^k), \tag{9}$$

where $Acc(X^k)$ is the classification accuracy for training datasets and $Order(X^k)$ denotes the number of methylation sites which are selected in the individual X^k.

D. Dataset

The high-throughput DNA methylation profiles of large genomic regions can be produced by both array and NGS technologies. We applied our approach to these two types of datasets. The array data were generated by the Illumina Infinium 27 k Human DNA methylation BeadChip, for surveying genome-wide DNA methylation profiles in breast cancer and normal samples [35]. We downloaded the dataset from Gene Expression Omnibus accession number GSE32393, and removed the samples with missing values. Sequence-based datasets were produced by MethylCap-seq in matched normal and colorectal cancer samples and collected at GSE39068 [36]. Normalization and preprocessing were carried out using the approaches detailed by Simmer et al. [36].

The DNA methylation levels of the two datasets were represented as beta-values, which were bounded between 0 (unmethylated) and 1 (totally methlyated).

III. Results

A. DNA Methylation Module Associated with Breast Cancer

This analysis was carried out based on DNA methylation profiling datasets that experimentally measured the methylation statuses using DNA Methylation BeadChip [35]. We extracted data for DNA methylation profiles on chromosome 17 from breast cancer and normal samples. Then, the data used at our experiment consist of total 99 samples with 82 cancer and 17 normal samples with 1,587 features. Figure 2 shows the learning curves in the evolutionary process. The fitness value was improved when the number of generations increased. We introduced a term, in the fitness function, for the number of the methylation sites to find an individual with a shorter length;

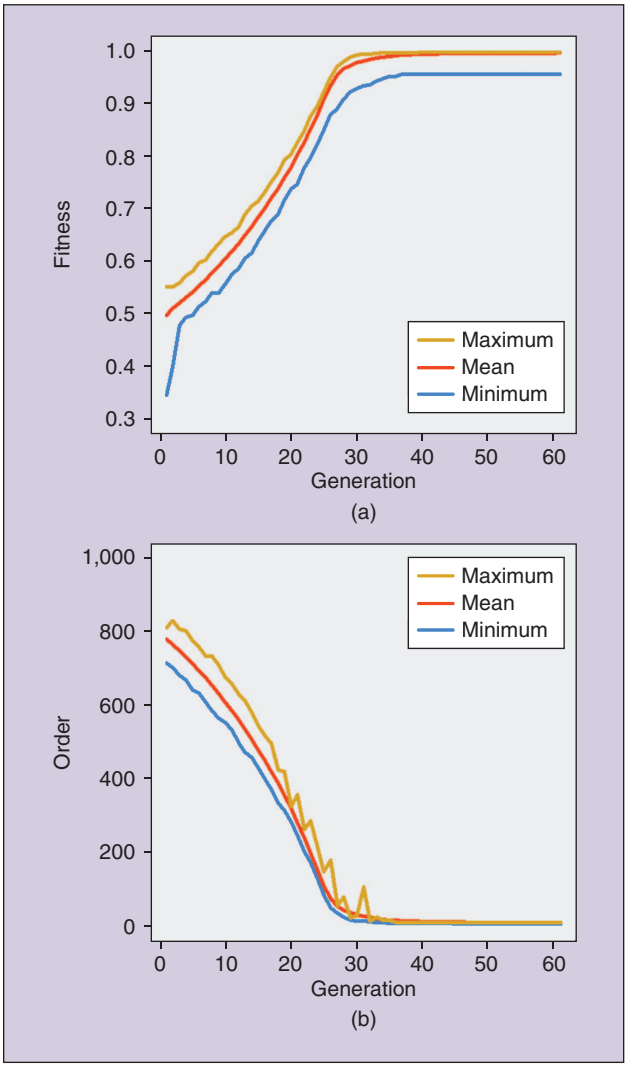


FIGURE 2 Learning curve using breast cancer datasets. The *x*-axis is the number of generations and the *y*-axis shows (a) fitness values and (b) the number of methylation sites.

therefore, the order decreased with the learning process (Figure 2(b)). After convergence, six sites were selected for the discrimination. These six sites were related to genes, KIAA1267, CD79B, ALOX12, TMEM98, KRT19 and FOXJ1 (Table 1).

ALOX12 has a role in the growth of breast cancer and its inhibition may be a strategy for inhibiting tumor growth [37]. The gene can be used as a serum marker for breast cancer [38].

TABLE 1 Finally Selected Methylation Sites.

ID	POSITION	GENE	CGI LOCATION
CG02301815	41605268	KIAA1267	41605074-41605445
CG07973967	59363339	CD79B	25467633-25468370
CG08946332	6840612	ALOX12	6839463-6841283
CG11833861	28279748	TMEM98	28278827-28279833
CG16585619	36938776	KRT19	NO CGI*
CG24164563	71647990	FOXJ1	71647419-71649480

*This site is not located within a CGI.

TABLE 2 Classification Performance by Splitting Training and Test Data.

ALGORITHM	ACCURACY	SENSITIVITY	SPECIFICITY
LOGISTIC REGRESSION	0.947	0.919	0.768
SVM	0.908	0.975	0.476
DECISION TREE	0.928	0.894	0.768
NAIVE BAYES	0.935	0.928	0.772

In addition, it has been reported that hypermethylation of ALOX12 is associated with cancer [39]–[42]. Indeed, the ALOX12 gene is closely related to apoptosis, and alterations in its expression caused by DNA methylation can cause a malfunction in cell death [43]–[45]. Therefore, it is reasonable to hypothesize that a change of methylation in the gene is linked to cancer, including breast tumors. KRT19 is a well-known marker for breast cancer [46], [47], and the KRT19 promoter can be aberrantly methylated in cancer cell lines [48]. The CD79B gene has also been shown to be related to breast cancer in several studies [49], [50]. FOXJ1, a member of the forkhead box (FOX) family, may function as a tumor suppressor gene in breast cancer [51]. FOXJ1 is hypermethylated and silenced in breast cancer cell lines [52]. TMEM98 is a transmembrane protein. Recently, Grimm et al. investigated transmembrane proteins specific for cancer cells, and showed that the transmembrane proteins can be targets for antibodies and may form biomarkers for tumor diagnosis, prognosis, and treatment [53]. The function of KIAA1267 is unclear yet, but this gene encodes KAT8 regulatory NSL complex subunit 1, and KAT8 regulates p53, a tumor suppressor gene [54], [55]. Our results suggest that KIAA1267 also can have a role in breast cancer.

To verify that our method produced good classification performance generally, we calculated the classification performance by randomly separating the original dataset into training and test datasets. Table 2 shows the average accuracy, sensitivity and specificity for 20 times repetition of random

TABLE 3 Classification Performance Using the Selected Sites and Randomly Selected Sites.

ALGORITHM	FEATURE*	ACCURACY	SENSITIVITY	SPECIFICITY
LOGISTIC REGRESSION	SELECTED	0.939	0.987	0.762
	$f = 5$	0.834	0.968	0.191
	$f = 6$	0.839	0.967	0.224
	$f = 10$	0.855	0.949	0.405
	$f = 20$	0.893	0.950	0.621
SVM	SELECTED	0.929	0.941	0.857
	$f = 5$	0.829	0.999	0.008
	$f = 6$	0.830	0.998	0.018
	$f = 10$	0.833	0.995	0.054
	$f = 20$	0.867	0.986	0.304
DECISION TREE	SELECTED	0.939	0.952	0.867
	$f = 5$	0.822	0.936	0.269
	$f = 6$	0.822	0.93	0.302
	$f = 10$	0.826	0.908	0.431
	$f = 20$	0.849	0.910	0.555
NAIVE BAYES	SELECTED	0.919	0.951	0.765
	$f = 5$	0.774	0.817	0.568
	$f = 6$	0.769	0.802	0.613
	$f = 10$	0.795	0.804	0.753
	$f = 20$	0.837	0.843	0.810

*At the column Feature, f is the number of randomly selected sites, and selected means the selected sites by our method.

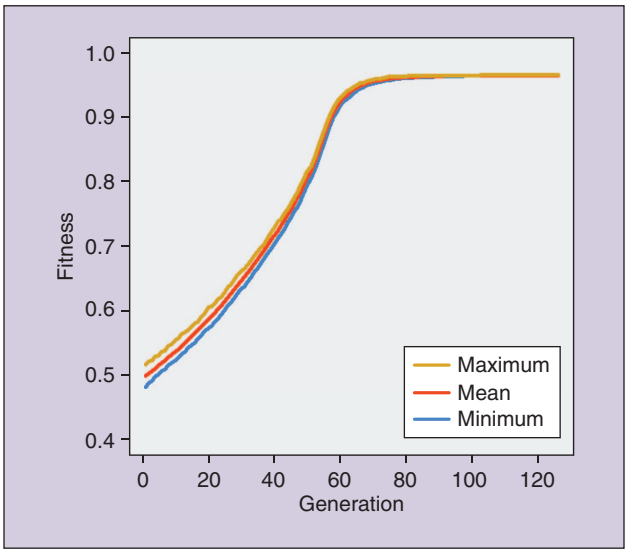


FIGURE 3 Learning curve using colorectal cancer datasets. The *x*-axis is the number of generations and the *y*-axis is fitness values.

splitting, measured by conventional classification algorithms. Our algorithm showed good classification results even at the independent test set. For further verification, we randomly extracted the methylation sites with 100 times repetition, then measured the classification performance in each dataset by 10-fold cross-validation. Table 3 shows that our method produced better results than the others, regardless of the number of the randomly selected sites. In particular, it was noted that the specificity using the selected sites by our method was much better than the others, even though the original data was highly imbalanced.

B. Modules Associated with Colorectal Cancer using High-Throughput Sequencing Data

Recently, high-throughput sequencing technologies have been used to determine DNA methylation profiles. We applied our method to the sequencing-based methylation profile datasets produced by Simmer et al. [36].

The experiments were carried out using 25 cancer and 25 normal samples with 10,393 genomic regions on chromosome 17. Figure 3 depicts the improvement of the fitness in iterative learning procedures using these datasets, and finally 348 regions were selected to discriminate colorectal cancer and normal samples after convergence. Table 4 shows the average classification performance by 10-fold cross-validation using the selected sites.

We annotated the 348 selected regions using GPAT [56] and investigated which genes were located close to the selected regions. We determined which genes were enriched within the KEGG pathway using the genes whose transcription start sites are located within 5,000 bp from the selected genomic regions [57], [58]. Table 5 summarizes the significantly enriched pathways with low *p*-values and shows that most of these are closely associated with cancer-related networks. Note that the enriched signaling pathways were related to colorectal cancer. In colorectal cancer, the roles of the wnt signaling pathway and MAPK signaling pathway have been studied intensively [59]–[62]. Genetic mutations affecting the pathway components and the alteration of their expression can enhance tumorigenicity in cancer cells. In addition, the neurotrophin signaling pathway could be related to growth of colorectal cancer cells [63] and the chemokine signaling pathway suppresses colorectal cancer metastasis [64], [65].

TABLE 4 Classification Performance Using Only the Selected Sites in Colorectal Cancer.

ALGORITHM	ACCURACY	SENSITIVITY	SPECIFICITY
LOGISTIC REGRESSION	0.900	0.920	0.880
SVM	0.940	0.960	0.920
DECISION TREE	0.640	0.680	0.600
NAIVE BAYES	0.900	0.920	0.880

TABLE 5 Enriched Geneset in Colorectal Cancer Data.

GENE SET	*p*-VALUE	FDR *q*-VALUE
NON-SMALL CELL LUNG CANCER	2.61E-05	4.25E-03
GLIOMA	4.56E-05	4.25E-03
NEUROTROPHIN SIGNALING PATHWAY	3.25E-04	1.85E-02
PATHWAYS IN CANCER	3.99E-04	1.85E-02
WNT SIGNALING PATHWAY	5.52E-04	2.05E-02
ALDOSTERONE-REGULATED SODIUM REABSORPTION	9.09E-04	2.22E-02
ENDOCYTOSIS	9.62E-04	2.22E-02
VASOPRESSIN-REGULATED WATER REABSORPTION	9.97E-04	2.22E-02
CHEMOKINE SIGNALING PATHWAY	1.07E-03	2.22E-02
FOCAL ADHESION	1.26E-03	2.34E-02
ENDOMETRIAL CANCER	1.39E-03	2.35E-02
BASAL CELL CARCINOMA	1.55E-03	2.41E-02
COLORECTAL CANCER	1.97E-03	2.73E-02
PANCREATIC CANCER	2.50E-03	2.73E-02
MELANOMA	2.57E-03	2.73E-02
CHRONIC MYELOID LEUKEMIA	2.72E-03	2.73E-02
CYTOKINE-CYTOKINE RECEPTOR INTERACTION	2.82E-03	2.73E-02
MAPK SIGNALING PATHWAY	2.82E-03	2.73E-02
PHOSPHATIDYLINOSITOL SIGNALING SYSTEM	2.94E-03	2.73E-02
VEGF SIGNALING PATHWAY	2.94E-03	2.73E-02
FC EPSILON RI SIGNALING PATHWAY	3.17E-03	2.81E-02
SMALL CELL LUNG CANCER	3.58E-03	2.98E-02
ERBB SIGNALING PATHWAY	3.83E-03	2.98E-02
APOPTOSIS	3.92E-03	2.98E-02
PROSTATE CANCER	4.01E-03	2.98E-02

The phosphatidylinositol signaling pathway plays an important role in the growth, survival and metabolism of cancer cells, and targeting this pathway has the potential to lead to treatments for colorectal cancer [66], [67]. VEGF and ErbB may be valid therapeutic targets for patients with colorectal cancer [68]–[71].

IV. Discussion and Conclusion

DNA methylation may be associated significantly with complex diseases and many genomic regions are differentially methylated in various cancers, comparing to normal samples. In this study, we presented a method to identify combinatorial effects of DNA methylation at multiple sites. From a systematic perspective, the relationship between DNA methylation regions and a specific disease is learned by the presented probabilistic evolutionary learning method. The fitness value of a DNA methylation module measures the level of its responses to the cancer. In a computational view, our method can solve a large number of feature problems by identifying modules with both compactness and high coverage of cancer-related genes. Applying our method to breast cancer and colorectal cancer data produced by high-throughput technologies, we detected cancer-related modules that were confirmed by the literature and functional enrichment analysis. Interestingly, we observed that the selected regions were located around genes that are significantly enriched in cancer-related gene set categories, which provided evidence that the identified modules in our study are biologically meaningful.

Moreover, from the result for the array-based dataset, we could obtain a good accuracy with a very small number of random features. However, the specificity was very low in the experiments with random features. The result suggested that our method could generate well-balanced classification performance even with a highly imbalanced dataset, although conventional classifiers would not work well with imbalanced circumstances. Also in the second experiment using the NGS-based dataset with large number of features and small sample size, our method could find the informative DNA methylation sites with good classification performances, even though the decision tree, necessary to be discretized in each value, showed relatively lower results.

Studies on DNA methylation could reveal the process of tumorigenesis as well as identify biomarkers. Our approach, which identifies multiple DNA methylation sites that might be epigenetically regulated, could provide a useful strategy to detect the epigenetic association related to cancer. By applying our method to array- and NGS-based data, we showed that it is applicable to a variety of data types and various disease contexts. Moreover, recent studies suggest a complex relationship between genetic variation and DNA methylation. Systems genetics and epigenetics approaches are required to examine these relationships. Although our framework is based on DNA methylation profile datasets, it could be used to identify the combinatorial association of various factors, including gene expression levels, microRNAs, copy number variations, genetic variations, and environmental factors. The integration of a variety of data would provide the basis for new hypotheses and experimental approaches in the model of a complex disease. Moreover, the systematic identification of causal factors and modules would provide insights into mechanisms underlying complex diseases and help to develop efficient therapies or effective drugs.

In summary, we presented a method for searching the higher-order interaction of DNA methylation sites using a probabilistic evolutionary learning method. We also examined the potential for the combined effects of various sites on the genome. The results suggested that the alteration of DNA methylations at multiple sites affects cancer. Similar to genome-wide association studies, our approach provided an opportunity to capture the complex and multifactorial relationships among DNA methylation sites and to find new factors for future study. Therefore, our approach would facilitate a comprehensive analysis of genome-wide DNA methylation datasets and help the interpretation for the effects of DNA methylation on multiple sites.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT, Republic of Korea (grant no. NRF-2015R1C1A1A01053824, NRF-2018R1C1B6005304, NRF-2016R1D1A1B03935676, and NRF-2018R1D1A1B07050393).

References

[1] P. A. Jones and S. B. Baylin, "The epigenomics of cancer," *Cell*, vol. 128, no. 4, pp. 683–962, 2007.
[2] B. Sadikovic, K. Al-Romaih, J. Squire, and M. Zielenska, "Cause and consequences of genetic and epigenetic alterations in human cancer," *Current Genomics*, vol. 9, no. 6, pp. 394–408, 2008.
[3] A. E. Handel, G. C. Ebers, and S. V. Ramagopalan, "Epigenetics: Molecular mechanisms and implications for disease," *Trends Mol. Med.*, vol. 16, no. 1, pp. 7–16, 2010.
[4] J. Sandoval and M. Esteller, "Cancer epigenomics: Beyond genomics," *Current Opinion Genetics Develop.*, vol. 22, no. 1, pp. 50–55, 2012.
[5] L. Bonetta, "Epigenomics: Detailed analysis," *Nature*, vol. 454, pp. 795–798, 2008.
[6] K. Robertson, "DNA methylation and human disease," *Nature Rev. Genetics*, vol. 6, pp. 597–610, 2005.
[7] A. Portela and M. Esteller, "Epigenetic modifications and human disease," *Nature Biotechnol.*, vol. 28, no. 10, pp. 1057–1068, 2010.
[8] P. Jones, "Functions of DNA methylation: Islands, start sites, gene bodies and beyond," *Nature Rev. Genetics*, vol. 13, no. 7, pp. 484–492, 2012.
[9] P. Laird, "Principles and challenges of genomewide DNA methylation analysis," *Nature Rev. Genetics*, vol. 11, no. 3, pp. 191–203, 2010.
[10] V. K. Hill, C. Ricketts, I. Bieche, S. Vacher, D. Gentle, C. Lewis, E. R. Maher, and F. Latif, "Genome-wide DNA methylation profiling of CpG islands in breast cancer identifies novel genes associated with tumorigenicity," *Cancer Res.*, vol. 71, no. 8, pp. 2988–2999, 2011.
[11] J.-K. Rhee, K. Kim, H. Chae, J. Evans, P. Yan, B.-T. Zhang, J. Gray, P. Spellman, T. H.-M. Huang, K. P. Nephew, and S. Kim, "Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer," *Nucl. Acids Res.*, vol. 41, no. 18, pp. 8464–8474, 2013.
[12] H. Cheung, T. Lee, A. Davis, D. Taft, O. Rennert, and W. Chan, "Genome-wide DNA methylation profiling reveals novel epigenetically regulated genes and non-coding RNAs in human testicular cancer," *Br. J. Cancer*, vol. 102, no. 2, pp. 419–427, 2010.
[13] G. Toperoff, D. Aran, J. D. Kark, M. Rosenberg, T. Dubnikov, B. Nissan, J. Wainstein, Y. Friedlander, E. Levy-Lahad, B. Glaser, and A. Hellman, "Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood," *Hum. Mol. Genetics*, vol. 21, no. 2, pp. 371–383, 2012.
[14] B. A. Walker, C. P. Wardell, L. Chiecchio, E. M. Smith, K. D. Boyd, A. Neri, F. E. Davies, F. M. Ross, and G. J. Morgan, "Aberrant global methylation patterns affect the molecular pathogenesis and prognosis of multiple myeloma," *Blood*, vol. 117, no. 2, pp. 553–562, 2011.
[15] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Rev. Genetics*, vol. 6, no. 2, pp. 95–108, 2005.
[16] A. Janssens and C. van Duijn, "Genome-based prediction of common diseases: Advances and prospects," *Hum. Mol. Genetics*, vol. 17, pp. R166–R173, 2008.
[17] A. Kiezun, K. Garimella, R. Do, N. O. Stitziel, B. M. Neale, P. J. McLaren, N. Gupta, P. Sklar, P. F. Sullivan, J. L. Moran, C. M. Hultman, P. Lichtenstein, P. Magnusson, T. Lehner, Y. Y. Shugart, A. L. Price, P. I. de Bakker, S. M. Purcell, and S. R.

Sunyaev, "Exome sequencing and the genetic basis of complex traits," *Nature Genetics*, vol. 44, no. 6, pp. 623–630, 2012.
[18] H. Easwaran, S. Johnstone, L. Van Neste, J. Ohm, T. Mosbruger, Q. Wang, M. Aryee, P. Joyce, N. Ahuja, D. Weisenberger, E. Collisson, J. Zhu, S. Yegnasubramanian, W. Matsui, and S. Baylin, "A DNA hypermethylation module for the stem/progenitor cell signature of cancer," *Genome Res.*, vol. 22, pp. 837–849, 2012.
[19] S. Horvath, Y. Zhang, P. Langfelder, R. Kahn, M. Boks, K. van Eijk, L. van den Berg, and R. Ophoff, "Aging effects on DNA methylation modules in human brain and blood tissue," *Genome Biol.*, vol. 13, no. 10, p. R97, 2012.
[20] J. Zhang and K. Huang, "Pan-cancer analysis of frequent DNA co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers," *BMC Genomics*, vol. 18, no. 1, p. 1045, 2017.
[21] M. Kumar, M. Husian, N. Upreti, and D. Gupta, "Genetic algorithm: Review and application," *Int. J. Inf. Technol. Knowl. Manage.*, vol. 2, no. 2, pp. 451–454, 2010.
[22] K. Deb and R. Datta, "A fast and accurate solution of constrained optimization problems using a hybrid bi-objective and penalty function approach," in *Proc. IEEE Congr. Evolutionary Computation*, 2010, pp. 1–8.
[23] J.-G. Joung, S.-J. Kim, S.-Y. Shin, and B.-T. Zhang, "A probabilistic coevolutionary biclustering algorithm for discovering coherent patterns in gene expression dataset," *BMC Bioinform.*, vol. 13, no. Suppl 17, p. S12, 2012.
[24] R. Wang, R. C. Purshouse, and P. J. Fleming, "On finding well-spread pareto optimal solutions by preference-inspired co-evolutionary algorithm," in *Proc. 15th Annu. Conf. Genetic and Evolutionary Computation Conf.*, New York, NY, 2013, pp. 695–702.
[25] S.-J. Kim, J.-W. Ha, and B.-T. Zhang, "Bayesian evolutionary hypergraph learning for predicting cancer clinical outcomes," *J. Biomed. Inform.*, vol. 49, pp. 101–111, 2014.
[26] T. Chen, P. Lehre, K. Tang, and X. Yao, "When is an estimation of distribution algorithm better than an evolutionary algorithm?" in *Proc. IEEE Congr. Evolutionary Computation*, 2009, pp. 1470–1477.
[27] A. Zhou, Q. Zhang, and Y. Jin, "Approximating the set of pareto-optimal solutions in both the decision and objective spaces by an estimation of distribution algorithm," *IEEE Trans. Evol. Comput.*, vol. 13, no. 5, pp. 1167–1189, 2009.
[28] V. Shim, K. Tan, J. Chia, and A. Al Mamun, "Multi-objective optimization with estimation of distribution algorithm in a noisy environment," *Evol. Comput.*, vol. 21, no. 1, pp. 149–177, 2013.
[29] J. Ceberio, E. Irurozqui, A. Mendiburu, and J. Lozano, "A distance-based ranking model estimation of distribution algorithm for the flowshop scheduling problem," *IEEE Trans. Evol. Comput.*, vol. 18, no. 2, pp. 286–300, 2014.
[30] S. Pal, S. Bandyopadhyay, and S. Ray, "Evolutionary computation in bioinformatics: A review," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 36, no. 5, pp. 601–615, 2006.
[31] R. Santana, A. Mendiburu, N. Zaitlen, E. Eskin, and J. Lozano, "Multi-marker tagging single nucleotide polymorphism selection using estimation of distribution algorithms," *Artif. Intell. Med.*, vol. 50, no. 3, pp. 193–201, 2010.
[32] K. Shelke, S. Jayaraman, S. Ghosh, and J. Valadi, "Hybrid feature selection and peptide binding affinity prediction using an EDA based algorithm," in *Proc. IEEE Congr. Evolutionary Computation*, 2013, pp. 2384–2389.
[33] M. Pelikan, "Implementation of the dependency-tree estimation of distribution algorithm in C++," 2006.
[34] M. Pelikan, S. Tsutsui, and R. Kalapala, "Dependency trees, permutations, and quadratic assignment problem," in *Proc. 9th Annu.Conf. Genetic and Evolutionary Computation*, New York, NY, 2007, pp. 629–629.
[35] J. Zhuang, A. Jones, S.-H. Lee, E. Ng, H. Fiegl, M. Zikan, D. Cibula, A. Sargent, H. B. Salvesen, I. J. Jacobs, H. C. Kitchener, A. E. Teschendorff, and M. Widschwendter, "The dynamics and prognostic potential of DNA methylation changes at stem cell gene loci in women's cancer," *PLoS Genetics*, vol. 8, no. 2, p. e1002517, 2012.
[36] F. Simmer, A. Brinkman, Y. Assenov, F. Matarese, A. Kaan, L. Sabatino, A. Villanueva, D. Huertas, M. Esteller, T. Lengauer, C. Bock, V. Colantuoni, L. Altucci, and H. Stunnenberg, "Comparative genome-wide DNA methylation analysis of colorectal tumor and matched normal tissues," *Epigenetics*, vol. 7, no. 12, pp. 1355–1367, 2012.
[37] A. kumar Singh, R. Singh, F. Naz, S. S. Chauhan, A. Dinda, A. A. Shukla, K. Gill, V. Kapoor, and S. Dey, "Structure based design and synthesis of peptide inhibitor of human lox-12: In vitro and in vivo analysis of a novel therapeutic agent for breast cancer," *PLoS One*, vol. 7, no. 2, p. e32521, 2012.
[38] A. Singh, S. Kant, R. Parshad, N. Banerjee, and S. Dey, "Evaluation of human LOX-12 as a serum marker for breast cancer," *Biochem. Biophys. Res. Commun.*, vol. 414, no. 2, pp. 304–308, 2011.
[39] A. C. Tan, A. Jimeno, S. H. Lin, J. Wheelhouse, F. Chan, A. Solomon, N. Rajeshkumar, B. Rubio-Viqueira, and M. Hidalgo, "Characterizing DNA methylation patterns in pancreatic cancer genome," *Mol. Oncol.*, vol. 3, no. 5, pp. 425–438, 2009.
[40] S. Alvarez, J. Suela, A. Valencia, A. Fernández, M. Wunderlich, X. Agirre, F. Prósper, J. I. Martín-Subero, A. Maiques, F. Acquadro, S. Rodriguez Perales, M. J. Calasanz, J. Roman-Gómez, R. Siebert, J. C. Mulloy, J. Cervera, M. A. Sanz, M. Esteller, and J. C. Cigudosa, "DNA methylation profiles and their relationship with cytogenetic status in adult acute myeloid leukemia," *PLoS One*, vol. 5, no. 8, p. e12197, 2010.
[41] O. Ammerpohl, J. Pratschke, C. Schafmayer, A. Haake, W. Faber, O. von Kampen, M. Brosch, B. Sipos, W. von Schönfels, K. Balschun, C. Röcken, A. Arlt, B. Schniewind, J. Grauholm, H. Kalthoff, P. Neuhaus, F. Stickel, S. Schreiber, T. Becker, R. Siebert, and J. Hampe, "Distinct DNA methylation patterns in cirrhotic liver and hepatocellular carcinoma," *Int. J. Cancer*, vol. 130, no. 6, pp. 1319–1328, 2012.
[42] R. S. Ohgami, L. Ma, L. Ren, O. K. Weinberg, M. Seetharam, J. R. Gotlib, and D. A. Arber, "DNA methylation analysis of ALOX12 and GSTM1 in acute myeloid leukaemia identifies prognostically significant groups," *Br. J. Haematol.*, vol. 159, no. 2, pp. 182–190, 2012.
[43] X.-Z. Ding, C.A. Kuszynski, T. H. El-Metwally, and T. E. Adrian, "Lipoxygenase inhibition induced apoptosis, morphological changes, and carbonic anhydrase expression in human pancreatic cancer cells," *Biochem. Biophys. Res. Commun.*, vol. 266, no. 2, pp. 392–399, 1999.
[44] G. P. Pidgeon, M. Kandouz, A. Meram, and K. V. Honn, "Mechanisms controlling cell cycle arrest and induction of apoptosis after 12-lipoxygenase inhibition in prostate cancer cells," *Cancer Res.*, vol. 62, no. 9, pp. 2721–2727, 2002.
[45] G. P. Pidgeon, K. Tang, Y. L. Cai, E. Piasentin, and K. V. Honn, "Overexpression of platelet-type 12-lipoxygenase promotes tumor cell survival by enhancing $\alpha_v\beta_3$ and $\alpha_v\beta_5$ integrin expression," *Cancer Res.*, vol. 63, no. 14, pp. 4258–4267, 2003.
[46] A. Ring, I. E. Smith, and M. Dowsett, "Circulating tumour cells in breast cancer," *Lancet Oncol.*, vol. 5, no. 2, pp. 79–88, 2004.
[47] M. Lacroix, "Significance, detection and markers of disseminated breast cancer cells," *Endocrine-Relat. Cancer*, vol. 13, no. 4, pp. 1033–1067, 2006.
[48] M. Morris, D. Gentle, M. Abdulrahman, N. Clarke, M. Brown, T. Kishida, M. Yao, B. Teh, F. Latif, and E. R. Maher, "Functional epigenomics approach to identify methylated candidate tumour suppressor genes in renal cell carcinoma," *Br. J. Cancer*, vol. 98, no. 2, pp. 496–501, 2008.
[49] R. Ellsworth, C. Heckman, J. Seebach, L. Field, B. Love, J. Hooke, and C. Shriver, "Identification of a gene expression breast cancer nodal metastasis profile," *J. Clin. Oncol.*, vol. 26, no. 15 Suppl, p. 1022, 2008.
[50] A. Prat, J. S. Parker, O. Karginova, C. Fan, C. Livasy, J. I. Herschkowitz, X. He, and C. M. Perou, "Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer," *Breast Cancer Res.*, vol. 12, no. 5, p. R68, 2010.
[51] B. C. Jackson, C. Carpenter, D. W. Nebert, and V. Vasiliou, "Update of human and mouse forkhead box (FOX) gene families," *Hum. Genomics*, vol. 4, pp. 345–352, 2010.
[52] B. Demircan, L. M. Dyer, M. Gerace, E. K. Lobenhofer, K. D. Robertson, and K. D. Brown, "Comparative epigenomics of human and mouse mammary tumors," *Genes Chromosomes Cancer*, vol. 48, no. 1, pp. 83–97, 2009.
[53] D. Grimm, J. Bauer, J. Pietsch, M. Infanger, J. Eucker, C. Eilles, and J. Schoenberger, "Diagnostic and therapeutic use of membrane proteins in cancer cells," *Current Med. Chem.*, vol. 18, no. 2, pp. 176–190, 2011.
[54] X. Li, L. Wu, C. A. S. Corsa, S. Kunkel, and Y. Dou, "Two mammalian MOF complexes regulate transcription activation by distinct mechanisms," *Mol. Cell*, vol. 36, no. 2, pp. 290–301, 2009.
[55] S. Zhang, X. Liu, Y. Zhang, Y. Cheng, and Y. Li, "RNAi screening identifies KAT8 as a key molecule important for cancer cell survival," *Int. J. Clin. Exp. Pathol.*, vol. 6, no. 5, pp. 870–877, 2013.
[56] A. Krebs, M. Frontini, and L. Tora, "GPAT: Retrieval of genomic annotation from large genomic position datasets," *BMC Bioinform.*, vol. 9, no. 1, p. 533, 2008.
[57] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 43, pp. 15 545–15 550, 2005.
[58] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.
[59] E. Å. Jansson, A. Are, G. Greicius, I.-C. Kuo, D. Kelly, V. Arulampalam, and S. Pettersson, "The WNT/β-catenin signaling pathway targets PPARY activity in colon cancer cells," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 5, pp. 1460–1465, 2005.
[60] S. Segditsas and I. Tomlinson, "Colorectal cancer and genetic alterations in the WNT pathway," *Oncogene*, vol. 25, no. 57, pp. 7531–7537, 2006.
[61] J. Y. Fang and B. C. Richardson, "The MAPK signalling pathways and colorectal cancer," *Lancet Oncol.*, vol. 6, no. 5, pp. 322–327, 2005.
[62] M. L. Slattery, A. Lundgreen, and R. K. Wolff, "Map kinase genes and colon and rectal cancer," *Carcinogenesis*, vol. 33, no. 12, pp. 2398–2408, 2012.
[63] H. Akil, A. Perraud, C. Mélin, M.-O. Jauberteau, and M. Mathonnet, "Fine-tuning roles of endogenous brain-derived neurotrophic factor, TrkB and sortilin in colorectal cancer cell survival," *PLoS One*, vol. 6, no. 9, p. e25097, 2011.
[64] T. Kitamura, T. Fujishita, P. Loetscher, L. Revesz, H. Hashida, S. Kizaka-Kondoh, M. Aoki, and M. M. Taketo, "Inactivation of chemokine (C-C motif) receptor 1 (CCR1) suppresses colon cancer liver metastasis by blocking accumulation of immature myeloid cells in a mouse model," *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 29, pp. 13 063–13 068, 2010.
[65] H. J. Chen, R. Edwards, S. Tucci, P. Bu, J. Milsom, S. Lee, W. Edelmann, Z. H. Gümüs, X. Shen, and S. Lipkin, "Chemokine 25-induced signaling suppresses colon cancer invasion and metastasis," *J. Clin. Invest.*, vol. 122, no. 9, pp. 3184–3196, 2012.
[66] D. W. Parsons, T.-L. Wang, Y. Samuels, A. Bardelli, J. M. Cummins, L. DeLong, N. Silliman, J. Ptak, S. Szabo, J. K. Willson, S. Markowitz, K. W. Kinzler, B. Vogelstein, C. Lengauer, and V. E. Velculescu, "Colorectal cancer: Mutations in a signalling pathway," *Nature*, vol. 436, no. 7052, pp. 792–792, 2005.
[67] T. Yuan and L. Cantley, "PI3K pathway alterations in cancer: Variations on a theme," *Oncogene*, vol. 27, no. 41, pp. 5497–5510, 2008.
[68] L. M. Ellis and D. J. Hicklin, "VEGF-targeted therapy: Mechanisms of anti-tumour activity," *Nature Rev. Cancer*, vol. 8, no. 8, pp. 579–591, 2008.
[69] T. Winder and H.-J. Lenz, "Vascular endothelial growth factor and epidermal growth factor signaling pathways as therapeutic targets for colorectal cancer," *Gastroenterology*, vol. 138, no. 6, pp. 2163–2176, 2010.
[70] R. Roskoski, Jr., "The ERBB/HER receptor protein-tyrosine kinases and cancer," *Biochem. Biophys. Res. Commun.*, vol. 319, no. 1, pp. 1–11, 2004.
[71] J. Spano, R. Fagard, J.-C. Soria, O. Rixe, D. Khayat, and G. Milano, "Epidermal growth factor receptor signaling in colorectal cancer: Preclinical data and therapeutic perspectives," *Ann. Oncol.*, vol. 16, no. 2, pp. 189–194, 2005.