Case-based Reasoning for the Analysis of Methylation Data in Oncology

Christopher L. Bartlett[1] and Isabelle Bichindaritz[1]

[1] Intelligent Bio Systems Laboratory, Biomedical and Health Informatics
State University of New York at Oswego, 7060 NY-104, Oswego, NY 13126
cbartle3@oswego.edu

Abstract. Researchers seek to identify biological markers which accurately differentiate cancer subtypes and their severity from normal controls. One such biomarker, DNA methylation, has recently become more prevalent in genetic research studies in oncology. This paper proposes to apply these findings in a study of the diagnostic accuracy of DNA methylation signatures for classifying metastasis samples. Very high classification performance measures were obtained from differentially methylated positions and regions, as well as from selected gene signatures. Perfect accuracy was achieved with the top 5 feature-selected genes using three similar cases and the K-nearest neighbor classifier. This work contributes to the path toward the identification of biological signatures for oncology samples using case-based reasoning.

Keywords: machine learning, case-based reasoning, bioinformatics, breast cancer

1 Introduction

The term epigenetics was first introduced into modern biology by Conrad Waddington as a means of defining interactions between genes and their products that result in phenotypic variations. Waddington's landscape presents a cell becoming more differentiated as time goes on. One of the events that can cause this differentiation is methylation. Methylation is a covalent attachment of a methyl group to cytosine. Figure 1 shows the addition of this methyl group to cytosine. Cytosine (C) is one of the four bases that construct DNA and one of only two bases that can be methylated. While adenine can be methylated as well, cytosine is typically the only base that's methylated in mammals. Once this methyl group is added, it forms 5-methylcytosine where the 5 references the position on the 6-atom ring where the methyl group is added. Under the majority of circumstances, a methyl group is added to a cytosine followed by a guanine (G) which is known as CpG. While the methyl group is added onto the DNA, it doesn't alter the underlying sequence but it still has profound effects on the expression of genes and the functionality of cellular and bodily functions. Methylation at these CpG sites has been known to be a fairly stable epigenetic biomarker that usually results in silencing the gene. Further, the amount of methylation can be

increased (known as hypermethylation) or decreased (known as hypomethylation) and improper maintenance of epigenetic information can lead to a variety of human diseases.

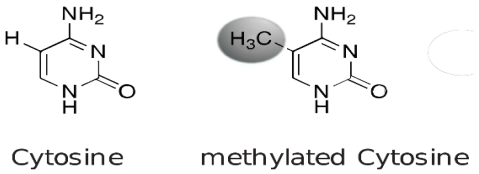

Fig. 1: Attachment of a methyl group to the 5 position of cytosine.

DNA methylation, has recently become more prevalent in genetic research studies in oncology. This paper proposes to apply these findings in a study of the diagnostic accuracy of DNA methylation signatures for classifying metastatic samples in breast cancer. This paper outlines the methods used to be able to apply case-based reasoning (CBR) and instance-based learning to methylation data, most often analyzed through statistical methods. Methylation data require a preprocessing pipeline leading to improved analysis, as this article shows. First, potential confounding factors such as batch effect and potential covariates are eliminated. Following, varied methods for the selection of subsets of methylation probes from the 485,577 highly dimensional dataset are applied. Feature selection methods further refine and select appropriate probes, eventually grouping them in genomic regions. These stages amount to case elaboration and constitute the bulk of the work for classification or prediction. This paper shows that the case elaboration mechanisms greatly improves the classification capability of case-based reasoning. Following sophisticated case elaboration processes, very high classification performance measures were obtained from differentially methylated positions and regions, as well as from selected gene signatures.

Specifically, we offer the following significant contributions:

1. **One of the first applications of CBR using methylation data.** While studies using gene expression data in a CBR context have been performed previously, very few (if any), applications using methylation data have been produced.
2. **Multi-level case elaboration and refinement which examine biological and statistical differences.** Significantly different methylation levels in the DNA, both at the microarray probe level and with a higher-order cluster of probes that serve similar functions were utilized and compared. Lastly, these probes are mapped to genes and ranked through a feature selection stage that attempted to locate the smallest possible signature of differential methylation.

2 Related Work

The utility of DNA methylation for the purposes of classification has been recently studied to differentiate blood samples in mental disorder subtypes [2] and cancer tumor tissue from normal tissue. This section will discuss a few such examples before concluding with the inspiration for the project outlined in this paper. The first such example is a prognostic classifier developed by Dos Reis et al., [10] for well-differentiated thyroid carcinoma (WDTC) based on 21 DNA methylation probes that predicted a poor outcome in patients with 63% sensitivity and 92% specificity for their internal data and 64% sensitivity and 88% specificity for data from The Cancer Genome Atlas. Similarly, Mundbjerg et al., [8] constructed an aggressiveness classifier from 25 methylation probes that could determine aggressive versus non-aggressive subtypes of prostate cancer. Testing on 496 prostate samples from tumors and adjacent-normal (AN) tissue, they found 97.4% specific and a 96.2% sensitivity.

Hao et al., [5] determined that DNA methylation could predict cancer versus normal tissue with accuracies above 95% in a three-cohort study of four common cancers. Testing in breast, colon, liver and lung cancer, differentially methylated CpG sites were used to classify tumor versus normal tissue. Hao et al., [5] used whole-genome methylation data from The Cancer Genome Atlas to construct a training cohort of 1,619 tumor samples and 173 matched adjacent normal tissue samples, and a validation cohort of 791 tumor samples and 93 matched adjacent normal tissue samples. The correct diagnosis rate for their training set was 98.4%, which was then replicated in the validation cohort for a statistically similar rate of 97.1%. A third, independent cohort of Chinese cancer samples (394 tumor samples and 324 matched adjacent normal tissue samples) resulted in a correct diagnosis rate of 95.0%. Methylation patterns were also able to correctly identify 29 of 30 colorectal cancer metastases in the liver, 32 of 34 colorectal cancer metastases in the lung and 19 of 20 breast cancer metastases [5]. This particular study promoted a positive outlook on the utility of DNA methylation for the classification and characterization of cancer.

Within the domain of CBR, there exist several applications using microarray data. Anaissi, Goyal, Catchpoole, Braytee, and Kennedy [1], for example, attempted to navigate the complexity of the highly-dimensional and imbalanced datasets often found in microarray analysis by focusing on case retrieval. Their framework uses a k-nearest neighbor (kNN) classifier with a weighted feature-based similarity measure to retrieve similar patients from a case base of acute lymphblastic leukemia. Gene expression data is employed to determine this similarity, and the treatment and outcome is used to propose solutions. Feature selection, dimensionality reduction, and feature weighting is used to handle the high-dimensionality of the data and removal of irrelevant features. They utilize oversampling to deal with the imbalanced classes. More specifically, they use the synthetic minority oversampling technique (SMOTE) methodology which artificially creates minority samples based on interpolation between members of the original minority class. After these pre-processing stages, a new sample is given to the kNN classifier to retrieve similar cases.

Ramos-Gonzalez et al., [9] used a two-level feature selection process for gene expression data in squamous cell carcinoma and adenocarcinoma. Their methodology has a preliminary feature selection which uses a non-parametric Mann-Whitney test to locate genes whose expression levels variation are statistically differentiated between subtypes. Following is a feature selection stage with Gradient Boosted Regression Trees that further refines the feature list into a greatly reduced subset that still maintains a high classification accuracy. A distance-based approach is used to retrieve similar cases, while additional diagnostic information may be requested that assists in correcting the prediction.

More recently, Lamy, Sekar, Guezennec, Bouaud and Seroussi [7] proposed a CBR method that visualizes results. The CBR system was rather straightforward, retrieving cases through a distance measure, though their specialization was in the explainability. Qualitative attributes between cases were shown using *rainbow boxes*, where labeled and colored rectangles extend through columns that represent the cases, clearly showing what was similar or dissimilar between cases. Quantitative attributes are provided in scatter plots that center on the query case and accurately displays the similar cases.

3 Material

Methylation data for breast cancer (BRCA, [1]) was downloaded from The Cancer Genome Atlas (TCGA, [2]) using the R package TCGAbiolinksGUI [3]. Molecular data was filtered for only the Illumina Human Methylation 450 platform and prepared as an RStudio object. This data pertained to 892 samples and the 485,577 probes that exist on the Illumina Human Methylation 450 beadchip. The methylation β values were then extracted. β values are an estimation of the methylation levels between 0 and 1 with 0 being completely non-methylated and 1 being completely methylated. Similarly, the BRCA clinical data was downloaded and subset for variables of relevance. These variables were the sample definition (describes whether the sample is a primary solid tumor, normal tissue, or the metastatic site), tumor stage, year of birth, tissue of origin, gender and race. This study focuses on classifying sample as either normal or a sample from a metastatic, stage 4 tumor.

4 Methods

4.1 Data Preprocessing

Metastatic tissue samples (those pertaining to the metastasized site, not the primary cancer site) were discarded, as well as samples from males. Year of birth was subtracted from the current year as a measure of the subject's age, regardless of whether the subject was alive or deceased. These subjects were then assigned

[1] https://portal.gdc.cancer.gov/projects/TCGA-BRCA
[2] https://www.cancer.gov/tcga

an age group with those less than 50 being in group 1, between 50 and 60 being in group 2, 60 to 70 in group 3, 70 to 80 in group 4, 80 to 90 in group 5, and those over 90 in group 6. The 10 stage 4 primary solid tumor samples were used to define the Metastatic group (M), while 95 solid tissue normal samples defined the Normal group (N). Removal of probes associated with covariate variables were then performed using the R package SVA and ComBat. The resulting dataset after pre-processing was 120,681 sites for 105 samples.

4.2 Case-based Classification

Classification was performed in several stages that further elaborated and refined the cases and carried out using the Waikato Environment for Knowledge Analysis (WEKA) [4] and the K-nearest neighbor algorithm. In each classification step, training was first performed through iterative removal of one sample for testing while the other samples were retained for training. For the testing sample, the nearest one, two or three cases were retrieved by calculating the Euclidean distance based on the similarity of features. The classification label for these cases were retrieved, with the label that was in the majority being reused for the testing sample. Despite the class imbalance, we elected not to use oversampling or undersampling. Oversampling the minority class can swiftly lead to overfitting, while undersampling the majority class can potentially lead to leaving out an important instance with crucial differences that could aid in the identification of the minority class. Instead, we utilized performance measures that adjusted for the class imbalance by calculating a balanced accuracy (BACC, computed using the average of per-class accuracy) and the weighted average area under the ROC curve (AUC).

4.3 Case Elaboration and Refinement

First results on all features after pre-processing proved to be unsatisfactory and attempts were made to refine the cases by focusing on the differentially methylated probes between the two groups (normal and metastatic), then on the differentially methylated regions. Finally feature selection methods were attempted to define a methylation signature on the metastatic samples.

Differentially Methylated Positions Differentially methylated positions (DMP) were identified using the Chip Analysis Methylation Pipeline (ChAMP) for R. This package uses limma to identify statistically different probes between the groups, using a Benjamin-Hochberg adjusted p-value of 0.05 for significance. 107,497 probes were found to be differentially methylated, and these sites were tested to observe differences in classification performance.

Differentially Methylated Regions The DMRcate method within ChAMP was used to extract the differentially methylated regions (DMR). Regions are

clusters of probes that serve a similar function in gene transcriptional regulation. Cross-hybridizing probes and sex-chromosome probes were removed prior to operation to further account for potential confounding factors such as gender. A false-discovery rate of 0.05 and a minimum probe number of 15 were provided as primary thresholding parameters with an adjusted p-value of 0.01 as the significance threshold. Probes within the located regions were then used to build the dataset for this stage. 788 probes were located within these regions.

4.4 Feature Selection

Feature selection was carried out on the dataset after initial pre-processing measures were performed, as well as on the data after differentially methylated position analyses. Prior to feature selection, each probe was mapped to its associated gene. Four algorithms in WEKA consisting of the Information Gain Attribute Evaluation, Correlation Attribute Evaluation, SMO Classifier Attribute Evaluation and Naive Bayes Classifier Attribute Evaluation were performed. An ensemble was then created using all of the results by tallying the rankings for each gene in the results of each algorithm. In each list, the best gene would be ranked first and the second best would be ranked second and so forth. The first stage was to take the top 5 percent of genes. The top 5 percent after pre-processing equated to 6,036 genes, while the top 5 percent after DMP equated to 5,377 genes. Balanced accuracy and the AUC were again used as performance measures.

Finally, features were ranked and a search by trial-and-error was performed to determine the smallest possible methylation signature.

5 Results

5.1 Classification after pre-processing

The resulting dataset after pre-processing was classified using leave-one-out-cross-validation (LOOC) using one, two or three cases to serve as a baseline for the comparison of case elaboration and refinement strategies. Table 1 displays the classification results as well as the number of metastatic samples (M) identified out of 10 total M samples. The results show that only 75% of the samples were correctly classified. This is a difficult problem due to the very large number of features (120,681).

5.2 Differentially Methylated Positions

The first case elaboration strategy consisted in selecting differentially methylated probes between normal and metastatic cases. 107,497 probes were found to be differentially methylated, and these sites were tested to observe differences in classification performance. The resulting balanced accuracies, AUC, and M samples identified is in Table 2. This table shows that results improved only slightly. The problem remains hard to the still large number of features (107,497).

Table 1: Classification results of 120,681 sites after pre-processing. **BACC:** Balanced Accuracy. **AUC:** Area Under the Curve. **M:** Metastatic

Classifier	BACC	AUC	Correct M Samples
1 case	70%	0.700	4
2 cases	65%	0.750	3
3 cases	75%	0.800	5

Table 2: Classification results of 107,497 sites after differentially methylated position analyses. **BACC:** Balanced Accuracy. **AUC:** Area Under the Curve. **M:** Metastatic

Classifier	BACC	AUC	Correct M Samples
1 case	75%	0.750	5
2 cases	70%	0.800	4
3 cases	75%	0.850	5

5.3 Differentially Methylated Regions

The second case elaboration and refinement strategy consisted in selecting differentially methylated regions. 788 probes were located within these regions, which greatly reduced the number of features. The balanced accuracies, AUC and M samples identified after classification at this stage is in Table 3. It is interesting to notice that the classification results significantly improve as measures by AUC. In addition the classification efficiency is greatly improved due to the significant reduction in number of features.

Table 3: Classification results of 788 sites after differentially methylated region analyses. **BACC:** Balanced Accuracy. **AUC:** Area Under the Curve. **M:** Metastatic

Classifier	BACC	AUC	Correct M Samples
1 case	75%	0.750	5
2 cases	70%	0.850	4
3 cases	75%	0.897	5

5.4 Feature Selection

Finally feature selection was applied to determine a methylation signature of the metastatic cases. The top 5 percent after pre-processing equated to 6,036 genes, while the top 5 percent after DMP equated to 5,377 genes. Balanced accuracy and the AUC were again used as performance measures. The results for the top 5 percent after pre-processing is available in Table 4 and after DMP in Table 5. This method generates significantly improved balanced accuracy and AUC over the previous methods.

Table 4: Classification results of 6,036 feature selected genes after pre-processing was conducted. **BACC:** Balanced Accuracy. **AUC:** Area Under the Curve. **M:** Metastatic

Classifier	BACC	AUC	Correct M Samples
1 case	90%	0.900	8
2 cases	90%	0.900	8
3 cases	85%	0.900	7

Table 5: Classification results of 5,377 feature selected genes after DMP was conducted. **BACC:** Balanced Accuracy. **AUC:** Area Under the Curve. **M:** Metastatic

Classifier	BACC	AUC	Correct M Samples
1 case	90%	0.900	8
2 cases	85%	0.900	7
3 cases	85%	0.900	7

5.5 Incremental Testing of the Highest Ranked Features

To determine a methylation signature, the top 1 feature-selected gene, top 2 feature-selected genes and so forth were selected, until reaching the top 15 feature-selected genes. The balanced accuracies and AUC for the top 1, top 5, top 10 and top 15 genes after pre-processing are available in Table 6. The balanced accuracies and AUC for the top 1, top 5, top 10 and top 15 genes after DMP are available in Table 7. Comparisons between case-based classification and alternate methods such as Naive Bayes and Random Forest, which showed highest classification performance, were performed. These tables show that the case-based classifiers performed at least as well as the best classifiers in this domain. Therefore, the case elaboration and refinement strategies proved very effective at reducing the search space and once this task accomplished the case-based approach is just as effective, if not more, with the advantage of being more explainable through the possibility of showing the cases used for the classification process.

Table 6: Sequential classification of the top 1, top 5, top 10, and top 15 genes from feature selection after pre-processing. **BACC:** Balanced Accuracy. **AUC:** Area Under the Curve

Number of Genes	1	1	5	5	10	10	15	15
Classifier	*BACC*	*AUC*	*BACC*	*AUC*	*BACC*	*AUC*	*BACC*	*AUC*
NB	100%	1.0	99%	0.999	99%	0.999	98%	0.999
RF	50 %	0.040	85%	1.0	95%	1.0	74%	0.853
1 case	95%	0.950	95%	0.950	100%	1.0	100%	1.0
2 cases	95%	0.950	99%	0.999	100%	1.0	100%	1.0
3 cases	95%	0.949	100%	1.0	100%	1.0	100%	1.0

Table 7: Sequential classification of the top 1, top 5, top 10, and top 15 genes from feature selection after DMP. **BACC:** Balanced Accuracy. **AUC:** Area Under the Curve

Number of Genes	1	1	5	5	10	10	15	15
Classifier	*BACC*	*AUC*	*BACC*	*AUC*	*BACC*	*AUC*	*BACC*	*AUC*
NB	100%	1.0	99%	0.999	99%	0.999	99%	0.999
RF	50%	0.040	90%	1.0	95%	1.0	95%	1.0
1 case	95%	0.950	100%	1.0	100%	1.0	100%	1.0
2 cases	95%	0.950	100%	1.0	100%	1.0	100%	1.0
3 cases	95%	0.949	100%	1.0	100%	1.0	100%	1.0

6 Discussion

These experiments show the usefulness of feature selection to both improve the efficiency and effectiveness of classification on highly dimensional data. Whatever the feature selection method selected, classifying on 1 to 15 features yielded improved results in most cases. In comparison with Anaissi et al., [1], we retrieve similar cases and perform classification using multiple levels which further refine the case information. We also opted out of using a synthetic oversampling technique which we believed may have reduced variance and impacted feature selection.

Bioinformatics is particularly interested in finding gene signatures for diseases, therefore appreciates feature selection over other methods [6]. It is therefore not surprising that this paper confirms the importance of this method in bioformatics and its usefulness to deal with high dimensional data.

7 Conclusion

In this paper, we have proposed to apply case-based classification to the task of classifying samples between normal and primary tumor with metastasis. Different strategies for case elaboration and refinement were attempted to reduce the high dimensionality of the methylation data. Our results show that case-based classification performs at least as well as the best classifiers in this domain, after selecting a pertinent methylation signature. This methylation signature will be invaluable for interpreting the deeper pathophysiological processes involved in the disease process. Some limitations of this work is that we have analyzed only one type of cancer - breast - which yielded a small dataset with only 105 cases, including 10 primary tumors from metastatic cancer. More work on independent data remains to perform to confirm 1) the reproducibilty of the results on these independent datasets, and 2) the validity of the selected genetic signature.

8 Acknowledgements

We thank the State University of New York EIPF grant #172 for their support of this work.

References

1. Anaissi, A., Goyal, M., Catchpoole, D.R., Braytee, A., Kennedy, P.J.: Case-based retrieval framework for gene expression data. Cancer Informatics **14** (2015). https://doi.org/10.4137/cin.s22371
2. Bartlett, C.L., Glatt, S.J., Bichindaritz, I.: Machine Learning and Feature Selection for the Classification of Mental Disorders from Methylation Data. Artificial Intelligence in Medicine Lecture Notes in Computer Science p. 311321 (2019). https://doi.org/10.1007/978-3-030-21642-9$_4$0
3. Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T., Malta, T.M., Pagnotta, S.M., Castiglioni, I., Ceccarelli, M., Bontempi, G., Noushmehr, H.: Tcgabiolinks: An r/bioconductor package for integrative analysis of tcga data. Nucleic Acids Research (2015). https://doi.org/10.1093/nar/gkv1507, http://doi.org/10.1093/nar/gkv1507
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1), 10–18 (2009)
5. Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., Flagg, K., Hou, J., Zhang, H., Yi, S., et al.: Dna methylation markers for diagnosis and prognosis of common cancers. Proceedings of the National Academy of Sciences **114**(28), 74147419 (2017). https://doi.org/10.1073/pnas.1703577114
6. Jurisica, I., Glasgow, J.: Applications of case-based reasoning in molecular biology. Ai Magazine **25**(1), 85–85 (2004)
7. Lamy, J.B., Sekar, B., Guezennec, G., Bouaud, J., Sroussi, B.: Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artificial Intelligence in Medicine **94**, 4253 (2019). https://doi.org/10.1016/j.artmed.2019.01.001
8. Mundbjerg, K., Chopra, S., Alemozaffar, M., Duymich, C., Lakshminarasimhan, R., Nichols, P.W., Aron, M., Siegmund, K.D., Ukimura, O., Aron, M., Stern, ., Gill, P., Carpten, J.D., Ørntoft, T.F., Sørensen, K.D., Weisenberger, D.J., Jones, P.A., Duddalwar, V., Gill, I., Liang, G.: Identifying aggressive prostate cancer foci using a DNA methylation classifier. Genome Biology **18**(1), 3 (2017). https://doi.org/10.1186/s13059-016-1129-3, http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1129-3
9. Ramos-Gonzlez, J., Lpez-Snchez, D., Castellanos-Garzn, J.A., Paz, J.F.D., Corchado, J.M.: A CBR framework with gradient boosting based feature selection for lung cancer subtype classification. Computers in Biology and Medicine **86**, 98106 (2017). https://doi.org/10.1016/j.compbiomed.2017.05.010
10. dos Reis, M.B., Barros-Filho, M.C., Marchi, F.A., Beltrami, C.M., Kuasne, H., Pinto, C.A.L., Ambatipudi, S., Herceg, Z., Kowalski, L.P., Rogatto, S.R.: Prognostic classifier based on genome-wide DNA methylation profiling in well-differentiated thyroid tumors. The Journal of Clinical Endocrinology & Metabolism **102**(November), 4089–4099 (2017). https://doi.org/10.1210/jc.2017-00881