Constructing prototypes for classification using epigenetic and genetic analysis

Christopher L. Bartlett

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

1 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 project seeks to apply the innovative and adaptive machine learning methodology in case-based reasoning (CBR) to examine DNA methylation levels in breast cancer. Instead of relying on a generalized knowledge-base, CBR uses highly specific information extracted from similar cases which can also greatly expedite the process of finding a solution. Further, this can locate targeted biomarkers by reusing homogenous factors, or revising to locate novel biomarkers in highly heterogeneous samples. While locating these biomarkers, this project proposes to use CBR to classify samples, predict prognoses and determine survival factors.

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. 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.

Within the domain of case-based reasoning (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.

A bit unorthodox, Yao and Li, [4], considered microarray samples in each class as one case-base. Then, given a sample, they retrieve several similar cases from each of the case-bases. Testing on leukemia, colon, and cancer data, Yao and Li retrieved results that outperformed several classic algorithms, including a few which used case-based reasoning.

Ramos-Gonzalez et al., [3] 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 [2] 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.

Advantages of CBR are its ability to generalize, and explainability. These factors will lend to an informative view of the epigenetic state of a cancer sample, and will hopefully assist in determining the heterogeneity of specific subgroups of samples.

2 Research Plan

The proposed research project seeks to employ CBR in an investigation of the epigenetic factors of breast cancer. Feature selection methods will be tested and evaluated to hone in on highly specific areas of the epigenome that have been impacted. A CBR framework to classify cancer samples, predict cancer prognoses and calculate survival is planned, with the underlying pathophysiological impacts of the cancer being investigated along the way. Prototypical representations of the the cancer and the clinical subgroups will also be researched.

2.1 Research Aims

1. To construct a case-based reasoning framework for classification of epigenetic data in breast cancer which takes covariate factors into account. Primary work here will focus on retrieving similar cases based on clinical and epigenetic similarity and using previously located labels to classify novel cases. In areas of dissimilarity, prior cases will be adapted to conform to the novel case. Integrating clinical factors has been shown to increase prediction ability (van Vliet et al., 2012) and prognostic performance (Zhu et al., 2017). It is hypothesized that the inclusion of these factors will lead to greater heterogeneity of found biomarkers as well as greater biological relevance.
2. To extend the established framework to predicting cancer prognoses. After the construction of a CBR framework for classification, prediction becomes a natural and swift process. Here, sample similarities will be retrieved and used to determine patient outcomes with modifications occurring where its necessary.
3. To further extend the established framework for survival analyses. Similar to Aim 1 and 2, similar samples will be retrieved though the goal at this phase is to locate the epigenetic signatures relevant to prolonged patient survival.
4. To locate deep pathophysiological pathways that have been impacted by cancer.
5. To establish a prototypical representation of cancer and clinical subgroups.
6. Extend the model for the reuse of prototypes for classification, prediction and survival analysis.

3 Progress-To-Date

Work was just completed using DNA methylation to classify breast cancer samples from normal tissue samples. The first stage was to investigate the most diverse of these cases, stage 4 cancer versus normal tissue. Classification was performed using naive bayes (NB), random forest (RF), and k-nearest neighbor with 3 iterations of k at a stage after surrogate variable analyses, after differentially-methylated position analyses, and after differentially-methylated region analyses. Finally, methylation probes at each genomic region within a particular gene were averaged and features were selected to find the highest performing genomic regions. The genes with the highest performing genomic regions

were then mapped to KEGG functional pathways and for the top 4 functional pathways, the associated genes were used to classify a larger set of cancer samples from a variety of stages to normal tissue. The four pathways were olfaction transduction, neuroactive ligand-receptor interaction, nicotine addiction, and GABAergic synapse. Results of this classification process are in Table 1.

Functional Pathway	NB	RF	K1	K2	K3
1. Olfaction Transduction	95.4	95	94.975	95.4	96.45
2. Neuroactive-Ligand	95.92	93.45	96.8	96.27	98.05
3. Nicotine Addiction	94	87.25	97.05	95.95	97.8
4. GABAergic	93.5	88.9	97.05	96.225	97.95

Table 1: Classification results from Naive Bayes, Random Forest and K-Nearest Neighbor with 3 instances of K for genes located in the most associated functional pathways

While this methodology held strong results, all iterations of the dataset suffered from a class-imbalance and whether or not overfitting occurred cannot yet be deduced. With these issues in mind, it is hopeful that the generation of a strong prototype through which to compare samples will allow a one-to-one correspondence that eliminates class-imbalance and strengthens classification results. If the prototype is able to be visualized, it would expand its strength and allow for downstream views into which biological mechanisms lend to the prototype's accuracy. Further, stage 4 samples were selected to represent a heterogeneous group in regards to the epigenetic state, but the small sample size removed the possibility of separating by clinical factors and still locating meaningful information. It is believed that a case-based reasoning approach would mitigate these issues and produce stronger results.

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. 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
3. 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
4. Yao, B., Li, S.: Anmm4cbr: a case-based reasoning method for gene expression data classification. Algorithms for Molecular Biology **5**(1) (2010). https://doi.org/10.1186/1748-7188-5-14