Epigenetic modifications are closely related to a number of human diseases. Clinical application of these modifications represents a promising opportunity for the development of new disease biomarkers or devices to monitor the impact of therapies. Here, we show strategies to apply data mining including machine learning technologies to DNA methylation signals in blood. The data originates from cohort studies and is used to build classifiers to identify smokers vs. non-smokers and elucidate the effect of smoking on DNA methylation, while applying computational biology to find associated genetic mechanisms.

Figure 1: How epigenetics mechanism can affect health (adapted from https://commonfund.nih.gov/epigenomics/figure)
THE DATA

Methylation data are available at the Gene Expression Omnibus from the NCBI (https://www.ncbi.nlm.nih.gov/geo). The study “Distinct Epigenetic Effects of Tobacco Smoking in Whole Blood and among Leukocyte Subtypes” was originally published by Su et al. 2016 in PLoS ONE. The data are available under the accession number GSE85210. This study contains methylation data from 253 individuals, specifically 172 from current smokers and 81 from never smokers. The data were analyzed by the authors with the Illumina Infinium HumanMethylation450 BeadChip array which comprises over 450,000 methylation sites and covers CpGs for about 99% of human RefSeq genes.

THE APPROACH – DATA MINING WITH INCORPORATION OF MACHINE LEARNING TECHNOLOGIES

Modern molecular biology consists of technologies which measure data on a global scale. In genomics, genetic modifications of an individual can be measured concurrently. In transcriptomics, gene expression on the transcript level can be assessed with single cell resolution. For epigenetic analyses epigenetic modifications like DNA methylation or histone modifications can be measured simultaneously. DNA methylation is an epigenetic process which is characterized by the addition of methyl groups to the DNA molecule. Methylation can change the activity of a DNA segment without changing the sequence. In this example, we wanted to find specific methylation changes in a cohort of smokers and non-smokers (referred to as “never-smokers”). Finding a specific informative element in hundreds of thousands simultaneously measured methylation features is a gargantuan task. Therefore we decided to apply machine learning-based data mining.

Machine learning is the sub-field of artificial intelligence which focuses on methods to construct computer programs that learn from experience with respect to some class of tasks and a performance measure (Mitchell 1997). Machine learning methods are suitable for molecular biology data due to the learning algorithm’s ability to construct classifiers/hypotheses that can explain complex relationships in the data.

Generally, there are two types of learning schemes: supervised and unsupervised learning. In supervised approaches, the algorithm learns a function from given pairs of inputs and outputs. During learning, a “teacher” provides the correct function value for input. The goal of supervised learning is that the algorithm creates classifiers for correct associations. In unsupervised learning no prior information is given to the learner regarding the data or the output and the algorithm generates a model for a given set of inputs that describes the inputs and allows for predictions automatically. As in this case a clearly defined training set of data was available, we decided to apply a supervised approach.

In practice the data were uploaded into the Waikato Environment for Knowledge Analysis (Weka) (https://www.cs.waikato.ac.nz/ml/weka/) and different supervised algorithms were applied. The system was trained for classifiers which allowed separating smokers from non-smokers.

Figure 2: The data uploaded into the Weka system. The data is derived from 253 individuals. Each dataset had 496, 697 data points (CpGs probed for their methylation status).
In addition to classical regression approaches (e.g., simple logistics), we used several tree-based algorithms like J48 and random forest. The results are shown for the J48 algorithm which is a Java implementation of the C4.5 (Quinlan, 1993) decision tree algorithm. This algorithm is known to be highly efficient.

According to the authors of the Weka machine learning software C4.5 algorithm is “a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date”. J48 builds decision trees from a set of training data using the concept of information entropy. At each node of the tree, J48 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other (smokers or never-smokers). The results are then checked against randomly sampled subsets of the data. The elements of the tree can be directly interpreted and compared with existing biological knowledge.

Applying the J48 algorithm to the methylation data (more than 450,000 methylation sites) of 253 individuals (172 from smokers / 81 never-smokers) the algorithm builds a classifier which can correctly classify 244 (96.44%) individuals in the referring class. Only 9 (3.56%) individuals were incorrectly classified.

Figure 3 shows the decision tree:

Figure 3: Decision Tree: The most informative feature is cg05575921. Beta values above 0.7984 are mainly never-smokers, and below this value are smokers. Smoking leads to de-methylation of this specific CpG
RESULTS INTERPRETATION

The data mining approach clearly allows the building of a clear classifier that distinguishes smokers from never-smokers. Given the complex nature of the DNA methylation status in more than 250 individuals, the accuracy of the classification at 96% is impressive. In addition, an interesting and unexpected finding was that primarily one CpG (cg05575921) is the key node responsible for the classification. The CpG is located on chromosome 5 in an intron of the gene called aryl-hydrocarbon receptor repressor (AHRR) which is known to participate in the aryl-hydrocarbon receptor (AHR) signalling cascade which mediates dioxin toxicity. This is intriguing given that tobacco smoke also contains dioxins and polycyclic aromatic hydrocarbons from incomplete combustion. In this context, the AHRR counteracts the activation of the AHR.

Based on completely independent datasets this specific CpG in AHRR was previously described in a number of publications:

Zeilinger et al. [PLOS ONE 2013] wrote: “…methylation-specific protein binding patterns were observed for cg05575921 within AHRR, which had the highest level of detectable changes in DNA methylation associated with tobacco smoking…”

Boisen et al. [Thorax 2017] described the effect as: “AHRR (cg05575921) hypomethylation, a marker of smoking behaviour, provides potentially clinical relevant predictions of future smoking-related morbidity and mortality…”

Andersen et al. [Journal of Insurance Medicine 2019] concluded: “Decreased methylation at cg05575921, an emerging epigenetic biomarker for smoking, was associated with early mortality in a longitudinal study of adults after accounting for the impact of major demographic and clinical risk factors for all-cause mortality. This approach may be useful in clinical research or actuarial assessments.”

Figure 4: AH-receptor signalling. Interaction between AHR and AHRR
CONCLUSION

We provide an example of how data mining techniques can be successfully applied to analyse blood-derived epigenetic methylation data. Machine learning strategies provide a straightforward approach to address complex and "big" datasets originating from real world situations.

In the example presented here, the obtained classifier is both suitable to categorize individuals correctly into specific categories (smokers vs never-smokers) and delineate the underlying biological functions with high precision.

The potential of machine learning approaches go far beyond this example. A number of precision medicine approaches which incorporate epigenetic disease biomarkers are now possible by taking advantage of machine learning-enabled acceleration of discovery processes. Epigenetic classifiers will be of great value for disease classification, and the discovery of new mechanisms will enable more precise clinical decision pathways.

REFERENCES


