

Methylome-based Pan Cancer Classifier (MePCC) - pan-cancer classification based on methylomics data

Jan Bińkowski, Tomasz K. Wojdacz

Independent Clinical Epigenetics Laboratory, Pomeranian Medical University, Szczecin, Poland.

Regional Center of Digital Medicine, Pomeranian Medical University, Szczecin, Poland.

Keywords: epigenomics, machine-learning, classification, cancer

Background:

The human methylome undergoes extensive alterations during carcinogenesis. While some DNA methylation changes occurring during the neoplastic transformation are stochastic, acting similarly to the accumulation of passenger mutations, others are critical for cancer development. Methylome changes essential for neoplastic transformation are stably transmitted during cell division, making them potential candidates for diagnostic biomarkers. In this study, we investigated whether methylomics data obtained using the Illumina BeadChip array are sufficient to develop a highly accurate, machine-learning (ML) based pan-cancer classifier.

Methods:

To build the classifier we used methylation profiling data (Illumina BeadChip microarrays) from public repositories including Genomic Data Commons and Gene Expression Omnibus databases. Overall the dataset for the development of our classifier included 12 156 genome wide methylation profiles for 39 types of cancer and 11 healthy tissues. Following the generally accepted principles of development of ML-based solutions we randomly (using cluster-sampling strategy) divided the entire dataset into three independent subsets: training (n=7805 samples), testing (n=3264 samples), and validation (n=1087 samples). Three types of measurements: (1) methylation levels expressed as beta-values, (2) co-methylation levels expressed as delta-beta-values, and (3) CNVs inferred from normalized probe intensities, were extracted from each microarray. The new type of effect size metric adjusted for sample heterogeneity was used to select the microarray measurements that best discriminated between samples in training data set. We then trained, tuned, and evaluated seven commonly used ML models using nested cross-validation combined with grouped-stratified sampling. Additionally, to ensure the quality of predictions we implemented a local outlier factor algorithm to detect abnormal samples - anomalies (outliers or novelties) that could potentially affect the confidence of model predictions.

Results: The logistic regression model achieved the best average specificity and sensitivity of pan-cancer classification in training set during the cross-validation procedure. Importantly, in the testing procedure model achieved an impressive average sensitivity and specificity of 0.94, and this level of sensitivity and specificity was also achieved using independent validation dataset, confirming excellent classification power of the classifier.

Conclusions:

We demonstrated that despite the limited resolution of microarrays, combined genomic and epigenomic data extracted from Illumina BeadChip microarrays are a source of stable and reproducible cancer and tissue type specific biomarkers. Furthermore, our results showed that combining omics data with machine learning algorithms and precise data processing strategies can successfully address complex biological challenges, such as pan-cancer classification.

Funding:

OPUS22 grant from National Science Centre, grant ID: 2021/43/B/NZ2/02979.