EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE FAKULTÄT

Centre for Bioinformatics Tuebingen (ZBIT)

Bioinformatics tools and research

Clemens Wrzodek, Johannes Eichner and Andreas Zell

Linking the epigenome to the genome: Tissue-specific DNA methylation predictions

DNA methylation of CpG islands plays a crucial role in the regulation of gene expression. In this study, known and novel genomic and non-genomic features were evaluated with respect to their degree of predictivity for tissue-specific DNA methylation. Furthermore, the performance of diverse well-established machine learning methods was assessed. To this end, various binary classifiers were trained and evaluated by cross-validation on a dataset comprising methylation data for 190 CpG islands in five tissues. For all CpG islands, a total of 960 features were computed for each tissue separately. These features were subdivided into 16 categories and ranked by their predictive performance. We achieved an accuracy of up to 91% with an MCC of 0.8 using ten-fold cross-validation. Whole-genome predictions for five tissues have been generated and a webservice to compare CpG island methylation status using the USCS genome browser is provided at http://www.ra.cs.uni-tuebingen.de/software/dna-methylation/

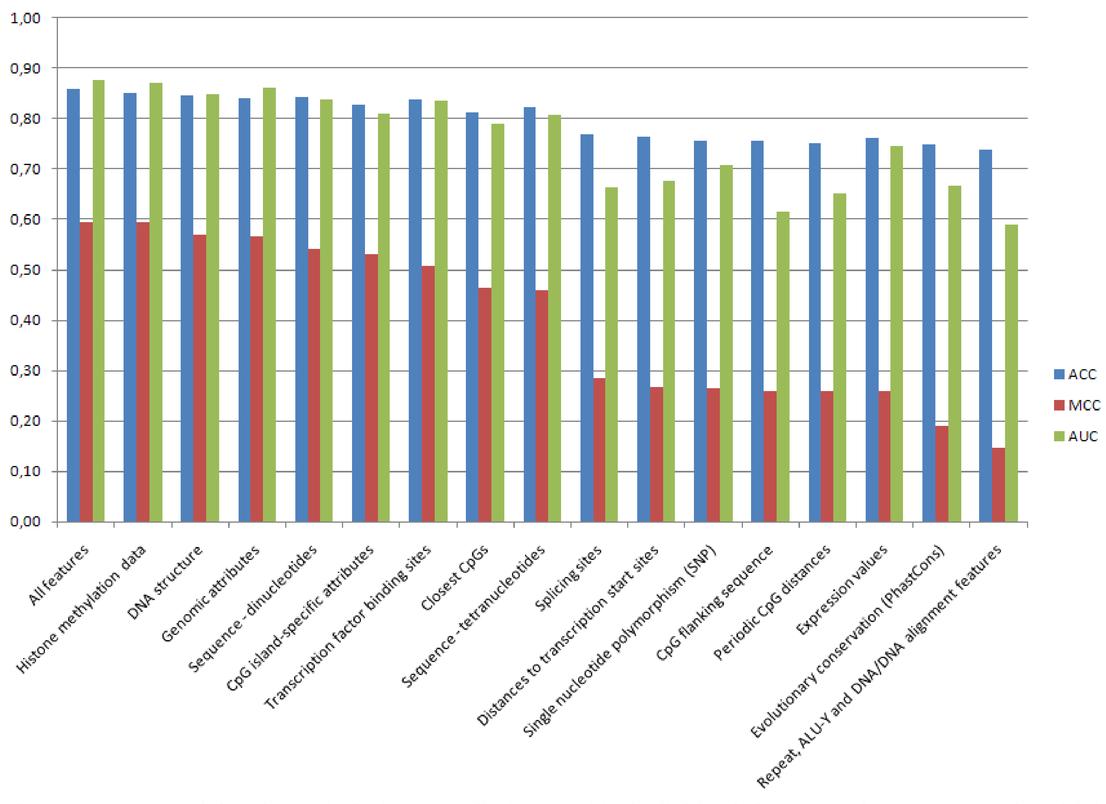


Figure 1: CpG island methylation predictions with individual feature classes reveal, which genomic and non-genomic imprints are correlated to the epigenome.
Each value is an average of a ten-fold cross-validation with ten repetitions. The Figure shows the accuracy (ACC), Matthews correlation coefficient (MCC), and the area under the receiver operating characteristic curve (AUC) and is sorted by average MCC.

	Leukocytes	HEPG2	HEK293	Fibroblast	Trisomic fibroblast
LIBSVM	0.819 ±0.15	0.375 ±0.30	**0.758** ±0.15	**0.611** ±0.19	0.307 ±0.32
LIBLINEAR	**0.825** ±0.14	**0.382** ±0.24	0.743 ±0.17	0.564 ±0.18	**0.340** ±0.33
Random Decision Forest	0.765 ±0.18	0.363 ±0.28	0.667 ±0.19	0.333 ±0.32	0.022 ±0.15
kNN	0.683 ±0.21	0.383 ±0.27	0.654 ±0.19	0.407 ±0.28	0.253 ±0.33
Decision tree (J48)	0.629 ±0.26	0.204 ±0.24	0.526 ±0.20	0.214 ±0.30	0.077 ±0.25
K*	0.393 ±0.36	0.128 ±0.33	0.381 ±0.36	0.312 ±0.40	0.200 ±0.31
NaiveBayes	0.146 ±0.22	0.057 ±0.21	0.117 ±0.26	0.064 ±0.24	0.095 ±0.21

Table 1: Machine learning algorithm performance.
We measured Matthews correlation coefficient (MCC) for every algorithm and every tissue using all features. The values shown in this table are averaged across ten repetitions using ten-fold cross-validation.

Federal Ministry
of Education
and Research

marcar
towards novel biomarkers
for cancer risk assessment

Pipeline for the analysis of gene expression data

For the analysis of gene expression data (e.g., Affymetrix data) on a systems level, we implemented a pipeline which integrates several analysis steps from preprocessing (e.g., normalization, missing value imputation) to basic analysis methods (e.g., statistical tests and clustering) as well as tools for systems-level analysis (e.g. ínference of gene-regulatory networks and signaling pathways) of the biological process under study.

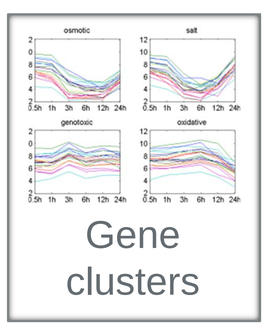
All analysis steps are publicly available as web tools (http://webservices.cs.uni-tuebingen.de/), which can be applied either individually or jointly as consecutive steps of custom-built workflows.

Overview of analysis steps

Raw data

Preprocessing
- Normalization
- Log transformation
- Missing value imputation

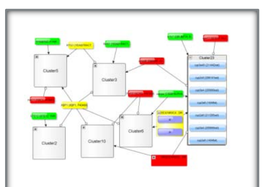
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS



Gene clusters

Basic analysis
- Clustering methods
- Statistical tests
- Gene set enrichment analysis

EDISA



Regulatory networks

Systems-level analysis
- Search for *cis*-regulatory modules
- Inference of gene-regulatory network
- Inference of signaling pathways

ModuleMaster

References

[1] Wrzodek C et al.: Linking the epigenome to the genome: Correlation of different features to DNA methylation of CpG islands. *PLoS ONE*, 7(4), 2012.

[2] Wrzodek C et al.: ModuleMaster: A new tool to decipher transcriptional regulatory networks. Biosystems 99, 79-81, 2010.

[3] Supper J et al.: EDISA: extracting biclusters from multiple time-series of gene expression profiles. BMC Bioinformatics 8(1), 2007.

[4] Supper J et al.: BowTieBuilder: modeling signal transduction pathways. BMC Systems Biology 3(1), 2009.

Center for Bioinformatics Tuebingen (ZBIT), Tübingen, Germany
Prof. Dr. Andreas Zell· Sand 1 · 72076 Tübingen
{Clemens.Wrzodek, Johannes.Eichner, Andreas.Zell} @uni-tuebingen.de