

DIPLOMA THESIS ASSIGNMENT

Student: **Mukhiddin Yusupov**

Study programme: Open Informatics
Specialisation: Artificial Intelligence

Title of Diploma Thesis: **Utilization of methylation data in phenotype molecular models**

Guidelines:

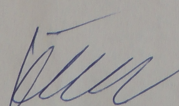
1. Learn about gene expression (GE) and the role of DNA methylation.
2. Get familiar with the existing research in the field of integrative analysis of genomic data.
3. Do literature recherche on integrative DNA methylation and mRNA expression analysis.
4. Empirically study the correlation between mRNA expression and methylation of corresponding CpG sites.
5. Use the known interactions between mRNA probes and methylation CpG sites in phenotype (disease) models.
6. Compare the reached results with mRNA, DNA methylation and concatenated phenotype models.
7. Formalize the method suggested ad 5, implement it as a new miXGENE functional plugin (R, Python).

Bibliography/Sources:

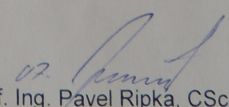
Alon, U.: An Introduction to Systems Biology: Design Principles of Biological Circuits. Taylor and Francis, 2006.
Jaenisch R., Bird A.: Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003 Mar, 33 Suppl: 245-54.
Lim, D. H. K., Maher E. R.: DNA methylation: a form of epigenetic control of gene expression. The Obstetrician and Gynaecologist 01/2011, 12(1):37-42.
Sass, S., Buettner, F., Mueller, N. S., Theis, F. J.: A modular framework for gene set analysis integrating multilevel omics data. Nucleic Acids Res. 2013 Nov, 41(21):9622-33.
Klema, J., Zahalka, J., Andel, M., Krejčík, Z.: Knowledge-Based Subtractive Integration of mRNA and miRNA Expression Profiles to Differentiate Myelodysplastic Syndrome. In: Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms. Porto: SciTePress - Science and Technology Publications, 2014, pp. 31-39.

Diploma Thesis Supervisor: doc.Ing. Jiří Kléma, Ph.D.

Valid until the end of the summer semester of academic year 2015/2016


doc. Ing. Filip Železný, Ph.D.
Head of Department




prof. Ing. Pavel Ripka, CSc.
Dean

Prague, October 31, 2014

Declaration of Originality

I declare that this thesis is my own work and the information (literature, research projects, SW and others) derived from the work of others has been acknowledged.

In Prague 11.05.2015

A handwritten signature in blue ink, consisting of a stylized first letter and a surname.

Czech Technical University
Faculty of Electrical Engineering



Utilization of methylation data in phenotype molecular models

(Master's thesis)

Mukhiddin Yusupov

Supervisor: doc. Ing. Jiří Kléma, PhD.

Study programme: Open Informatics

Specialization: Artificial Intelligence

May 2015

Abstract

Recent advances in technology have allowed scientists to conduct research in a systematic way by incorporating additional knowledge and to gain deeper insight on various biological processes in a domain of their study. One of the active research works where a great demand on the knowledge integration based approaches can be observed is the construction of phenotype molecular models or in particular biomedical disease models. Several studies have successfully applied statistical learning techniques by using gene expression (GE) profiles for building phenotype classification models. A lot of other works showed that DNA methylation (DNAm) brings new information and can be the main factor of GE regulation. Focusing on a dataset of patients with Myelodysplastic Syndrome (MDS) disease, we propose a prediction model that integrates DNAm with GE data in order to discover new signatures and improve the classification performance. Two types of integration, blind and smart integration, were employed and their performances were compared in the work. We show that for most of the disease associated treatment response types integration of DNAm improves the predictive power. We demonstrate that both GE and DNAm profiles contain specific patterns that contribute to classification of phenotype and the model with knowledge based integration outperforms the remaining models.

Abstrakt

Nedávný technologický rozvoj umožnil systematické využití apriorní znalosti v biologickém výzkumu. Cílem je hlubší vhled do podstaty biologických procesů. Jednou z nejčastějších oblastí využití znalostních přístupů je oblast tvorby fenotypových molekulárních modelů, kdy fenotyp často odpovídá různým typům a stupňům konkrétní nemoci. Dřívější úspěšné studie byly založeny zejména na statistickém učení založeném na profilech genové exprese (GE). Pozdější práce ukázaly, že i DNA metylační (DNAm) data přinášejí novou informaci a mohou být jedním z hlavních faktorů GE regulace. V této práci se zaměřujeme na konkrétní data pacientů s myleodysplastickým syndromem (MDS). Navrhujeme prediktivní model, který integruje DNAm s GE daty s cílem vylepšit prediktivní přesnost MDS klasifikátoru, popřípadě nalézt konkrétní prediktivní signatury. Porovnáváme dva typy integrace DNAm s GE daty. První je slepá, tj. neinformovaná, druhá je založená na znalostech interakcí mezi oběma typy dat. V práci demonstrujeme, že DNAm data napomáhají zvýšení přesnosti MDS klasifikace. Dále ukazujeme, že jak GE, tak i DNAm data, generují užitečné prediktivní vzory a znalostní modely svou přesností překonávají zbylé přístupy.

Acknowledgements

I would like to thank to my supervisor Jiří Kléma for giving me the opportunity to work on this exciting and challenging project, for guiding and supporting me and for showing me the right direction when I could not find it myself.

My thanks to my parents, family members and friends who continuously assisted me in everything they can and helped me to be patient and strong.

In the end, I am grateful towards all professors, teachers and instructors who have taught me the subjects and motivated me with the concepts in the area of Artificial Intelligence.

Contents

Abstract	iv
Abstract (Czech)	v
Acknowledgements	vi
List of Tables	x
List of Figures	xi
List of Abbreviations	xiii
1 Introduction	1
2 Phenotype molecular models	3
2.1 Biological background	3
2.1.1 Gene expression and regulation	3
2.1.2 DNA methylation	4
2.1.3 Detection methods	6

2.2	DNA methylation analysis	9
2.2.1	Exploratory analysis	9
2.2.2	Quality control and preprocessing	11
2.2.3	Differential methylation analysis	12
2.2.3.1	Single site differential analysis	13
2.2.3.2	Area based differential analysis	14
2.2.3.3	DNA methylation and gene expression integrative analysis	15
2.3	Phenotype prediction task	17
2.3.1	Challenges	17
2.3.2	Dataset	18
2.4	Goal of the work	19
3	Existing approaches	20
3.1	Feature selection/extraction	20
3.2	Model design and selection of classification algorithm	22
4	Proposed model	23
4.1	General design	23
4.1.1	Single Data types and Blind Integration based models	25
4.1.2	Smart Integration based models	25
4.2	Parameters specification	27

5 Experiments	28
5.1 Methodology	28
5.1.1 Experimental setup for differential analysis	28
5.1.2 Specification for integration study	29
5.1.3 Setup for model construction	29
5.2 Results	30
5.2.1 Data preparation	30
5.2.2 Findings from differential and integration study	33
5.2.3 Model evaluation	39
6 Conclusion	44
References	46
Appendix A - List of Software	51
Appendix B - Contents of the CD	52

List of Tables

2.1	Number of samples for each type of response treatment group	19
5.1	Best average true class probability performance for blindly integrated (BlindModFS) and two single data type based models (ExpModFS, MethModFS)	41
5.2	Number of features (probes) employed to obtain the best performance for BlindModFS, ExpModFS and MethModFS models	41
5.3	Best average true class probability performance for four smart integration based models: SmartModFe, SmartModFs, SmartModFeSEG and SmartModFsSEG	41
5.4	Number of features (probes) employed to obtain the best performance for four smart integration based models	42

List of Figures

2.1	Higher-order order chromatin structure is formed by wrapping of the double stranded DNA around histone proteins. Methylation (Me) can affect the formation of the chromatin structure, which in turn regulates DNA transcription and thus activates or inactivates genes. (Source: Nature 2008; 454: 711–755)	5
5.1	Cluster dendrogram for methylation data of MDS disease dataset . . .	30
5.2	Scatter plot with dimensionality reduction by PCA and MDS	31
5.3	Density plots of Beta - values for four types of treatment (PD, SD, PR, CR) versus reference - normal group samples	32
5.4	Density plots of M-values for four types of treatment (PD, SD, PR, CR) versus reference - normal group samples	32
5.5	Plots for PD and Normal samples before preprocessing. Left. Box plot of CpG-site intensity separately for two channels. Right. Density plot.	33
5.6	Plots after quantile normalization. Left. Density plot of M-values for four disease groups versus Normal samples. Right. Density plot of CpG-site intensity for PD and Normal samples	33
5.7	Box plot of CpG-site intensity for PD and Normal samples after quantile normalization (shown for two channels separately).	34

5.8	Histogram of p-values for the T-test with population variance equality assumption (Left) and for Welch’s test without such assumption(Right)	35
5.9	SAM plot fitting the expected results to the observed ones. This is a graph from two group experiment: PD type samples versus all remaining samples	36
5.10	Volcano plot. Finding significant probes by linear modeling. Two group experiment: <i>PD vs Normal</i> type of samples	37
5.11	Heat map of the first DMR location, corresponding gene and a sample information are shown in this graph	38
5.12	Showing one found region from PD vs NR experiment. Left: Plot of methylation sites. Two of these sites are differentially methylated. Right: Depiction of the difference in methylation levels	39
5.13	Box plot for two differentially expressed genes that have significant methylation regions	40
5.14	Scatter plots of methylation statuses for two CpG islands and their corresponding GE levels	40
5.15	Evaluation of SmartModFsSEG model. Plots are for each one of four groups of response types. Number of features is in horizontal axis and the average probability is in vertical axis	43

List of Abbreviations

BlindModFS	Blind integration and FS based Model
cDNA	Complementary DNA
DMR	Differentially Methylated Region
DMS	Differentially Methylated Site
DNA	Deoxyribonucleic acid
DNAm	DNA Methylation
ExpModFS	GE and FS based Model
FDR	False Discovery Rate
FE	Feature Extraction
FS	Feature Selection
GE	Gene Expression
MDS	Myelodysplastic Syndrome
MethModFS	DNAm and FS based Model
mRNA	Messenger RNA
NGS	Next Generation Sequencing
RNA	Ribonucleic acid

SEG	Significantly Expressed Gene
SmartModFe	Smart integration and FE based Model
SmartModFeSEG	Smart integration, FE and extra SEG enrichment based Model
SmartModFs	Smart integration and FS based Model
SmartModFsSEG	Smart integration, FS and extra SEG enrichment based Model
SNP	Single Nucleotide Polymorphism

Chapter 1

Introduction

DNA methylation is one of the most widely studied epigenetic mechanisms in human cells. Since it was proposed in 1975 that DNAm might be responsible for a stable maintenance of a particular gene expression pattern through mitotic cell division [18] a lot of research work provided evidences to support this concept and interest continues to grow rapidly resulting on expansion of a new areas of research [8, 27]. Now It is known fact that regulatory elements of one individual can be adjusted to outward conditions and transferred to its descendants. DNA methylation plays a crucial role in development, differentiation and diseases such as diabetes, schizophrenia, aging, and multiple forms of cancer.

Over the last few years with the advances in biotechnology massive amounts of data have been generated. Microarray and Next Generation Sequencing (NGS) technologies allow researchers to monitor and study the whole genome transcription level, mutation, DNAm, DNA copy number change, microRNA and protein expressions in a systematic way. This makes the integration of different types of data an indispensable component for biomedical research.

Genetic and epigenetic studies have provided important new insights in the underlying pathology of diseases such as Acute Myeloid Leukemia (AML) [16, 37, 22], Breast cancer [34, 3], MDS [5] and many others. Some types of experiments where GE data have been employed to identify sets of genes accomplished an accurate prediction

rates for disease subtypes or phenotypes of interest. For instance, some subtypes of AML like t(8;21), t(15;17), inv(16) or CEBRA can be predicted very well using GE profiles. However, for the same AML it is much more difficult to predict subtypes FLT3 and NPM1 in a similar manner using GE profiles [39]. This phenomenon occurs in many other cell type classification tasks and it is an indication that GE profiles do not always contain features that are sufficiently discriminative to distinguish experimental study groups. In another kind of studies [25, 26] DNAm data alone was used to construct prediction models and when their performance was compared to the ones build with GE data [35] they demonstrated the results more or less the same nature.

Insights from previous research works indicate that DNAm can have complementary impact when exploited in phenotype models and, specifically, in disease outcome/subtype classification models. Our aim in this work is to propose a method for phenotype model construction which is based on integration of DNAm and GE data. We suggest that identification of DNAm patterns which are either interact with phenotype predictor genes or are key predictors themselves and their utilization in phenotype molecular models can allow us to obtain better results than with models built on only one type of data (GE or DNAm) or combination of two without taking into account their interplay.

In this study to demonstrate our approach we used dataset of patients with MDS disease of different risk levels. In the next chapter the relevant biological notions, general pipeline for differential expression and methylation analyses as well as integrative study of GE and DNAm are introduced. Chapter 3 covers techniques for phenotype model construction and proposed model is described in Chapter 4. Chapter 5 illustrates experimental results and a comparison of a designed models.

Chapter 2

Phenotype molecular models

In this chapter, we provide the relevant biological background, some important data analysis procedures and the task definition. Brief definition of necessary biological terms facilitates the understanding of goals and ideas behind the further notions and computational methods presented in the work. However, for the detailed description it is advised to refer to the following studies [23, 20, 29, 30]. We present the general pipeline for DNAm data analysis. Finally, chapter ends with description and aim of the task.

2.1 Biological background

2.1.1 Gene expression and regulation

GE is the process by which the genetic code, the nucleotide sequence, of a gene is used in the synthesis of a functional gene product. These products are often proteins, transfer RNAs, ribosomal RNAs and other functional units. Proteins then perform essential functions as enzymes, hormones and receptors. GE is one of the most im-

portant processes in all known life (eukaryotes, prokaryotes and viruses). It usually includes steps such as transcription, RNA splicing, translation and post-translational modification of a protein. The amount of an end product, its structure and function is determined by gene regulation that is comprised by a wide range of mechanisms used by cells. GE can be regulated in transcription, post-transcription, translation or mRNA degradation stages. The most extensively utilized one is transcription.

Transcription is the production of messenger RNA (mRNA) by the enzyme RNA polymerase and the processing of the resulting mRNA molecule. Regulation of transcription is either as a result of interaction of some control factor with the gene or transcription machinery, or non-sequence changes in DNA structure after transcription. The former kind of regulation is known as genetic or modulation and the latter as epigenetic.

Regulation at the post-transcriptional stage is controlled by importing and exporting proteins that influence the transport of RNA in and out of the nucleus.

Translation is the use of mRNA to direct protein synthesis. Regulation of GE at translation stage is less common. Translational regulation is used by antibiotics and toxins.

2.1.2 DNA methylation

DNAm is a widespread form of epigenetic regulation of GE. It has its roles in heritable transcription silencing and transcription regulation.

Methylation of DNA is one of the epigenetic mechanisms that cells use to control expression. DNAm refers to the addition of a methyl group (-CH₃) covalently to the base cytosine (C) in the dinucleotide 5'-CpG-3' where CpG refers to the base cytosine (C) linked by a phosphate bond to the base guanine (G) in the DNA nucleotide sequence. Unmethylated CpGs are grouped together and form the 'CpG islands'. Their location in humane genome is in promoter region. In this region we can observe the transcription of a particular gene [23].

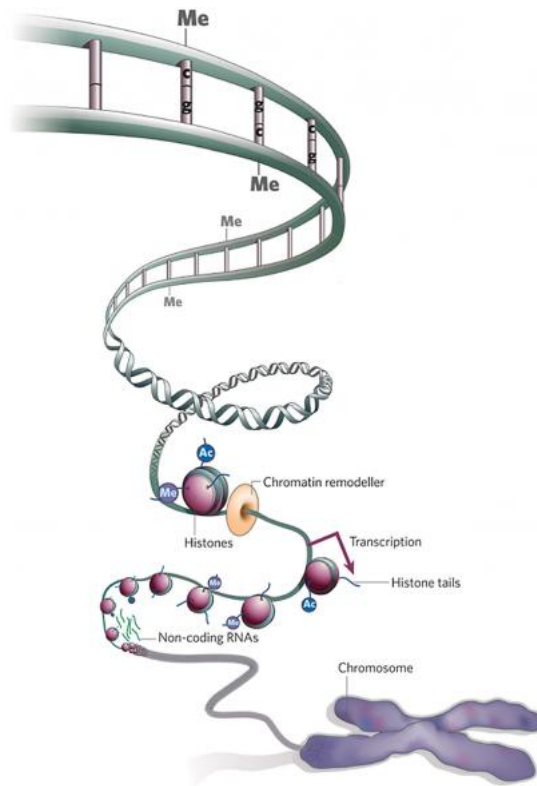


Figure 2.1: Higher-order order chromatin structure is formed by wrapping of the double stranded DNA around histone proteins. Methylation (Me) can affect the formation of the chromatin structure, which in turn regulates DNA transcription and thus activates or inactivates genes. (Source: Nature 2008; 454: 711–755)

DNAm can result transcriptional silencing of a gene. This occurs due to GE inhibitory potential of DNAm which is accomplished by preventing the binding of transcription factors to the promoter region. As a result genes go into so called “off” state. When located at gene promoters, DNAm is usually a repressive mark. However, CpG DNAm is increased in the gene bodies of actively transcribed genes in plants and mammals.

Epigenetic change of human genome by DNAm is heritable. It is essential this process to occur during embryonic development. Furthermore, aberrant methylation patterns have been associated with many types of human diseases. This can happen due to either hyper- or hypo-methylation. Hypermethylation takes place at CpG

islands in the promoter region and is associated with gene inactivation. A lower level of leukocyte DNAm is linked with many types of cancer. Global hypo-methylation has also been implicated in the development and progression of cancer through different mechanisms. Typically, there is hyper-methylation of tumor suppressor genes and hypo-methylation of oncogenes.

2.1.3 Detection methods

There are a lot of procedures for detecting and determining the abundance of GE and methylation sites. In practice it is easier to detect mRNA than detecting the final gene product for the measurement of expression. Some of the popular methods for quantifying the levels of mRNA are Northern blot, Reverse Transcription quantitative real-time Polymerase Chain Reaction or RT-qPCR, hybridization microarrays and RNA sequencing (RNA-Seq).

Northern blotting or Northern analysis presents several advantages over other techniques. It is the most easiest method to determine transcript size, identify alternatively spliced transcripts and multi-gene family members. It can also be used to directly compare the relative abundance of a given message between all the samples on a blot. Despite these advantages, there are some limitations associated with Northern analysis. The slight sample degradation can extremely affect the quality of data and ability of expression quantification. Other disadvantages of standard Northern procedure are least sensitivity of the reviewed techniques and difficulty associated with multiple probe analysis. However, there are improvements on standard Northern blotting which increased to substantial degree its sensitivity.

Another approach for measuring mRNA abundance is RT-qPCR. In this technique reverse transcription is followed by quantitative PCR. First a complementary DNA (cDNA) is generated from mRNA. Then cDNA is amplified exponentially by PCR. qPCR can produce an absolute measurement of the number of copies of original mRNA. RT-PCR is th most sensitive method of mRNA detection available. PCR based approaches have became widespread for quantifying individual mRNAs partly due to a reason that individual mRNA species are expressed in small quantities that

makes detection difficult.

One of the widely used approaches for expression quantitation is the hybridization microarray. DNA microarrays can be used to measure the expression levels of large numbers of genes simultaneously which enabled scientists to accomplish many genetic tests in parallel in a one microarray experiment. DNA microarrays can be used to measure changes in expression levels, to detect single nucleotide polymorphisms (SNPs) or to genotype. Two technologies oligonucleotide arrays and cDNA arrays have emerged for the construction of DNA microarrays. Oligonucleotide arrays, or also called one-color arrays, are variable efficient in hybridization than cDNA arrays. They can synthesize multiple probes complementary to each gene of interest by using short oligonucleotide sequences of length 30-60 each designed to represent a single gene. cDNA arrays, or two-color arrays, are typically hybridized with cDNA prepared from two samples to be compared (control versus treatment samples) and that are labeled with two different fluorophores. Absolute levels of GE can be determined with cDNA array, however, in practice or in data analysis relative differences in expression is a more preferred choice.

RNA-seq is the approach that uses the capabilities of NGS. This approach permits to produce vast quantities of sequence data that can be matched to a reference genome. Although, methods based on NGS are time-consuming, expensive and resource-intensive, they provide more information and data of whole genome giving rise to their utilization and exploitation for various kinds of analyses.

For the last few years the most often used in practice array based technology for DNAm processing and profiling is Illumina BeadArray Technology. Illumina adapted its BeadArray technology for genotyping to recognise bisulphite-converted DNA [7]. The Illumina BeadArray assays use oligonucleotides conjugated to bead types to measure specific target sequences, measuring multiple beads per bead type. The bead types are summarized by the average signal for methylated (M) and unmethylated (U) alleles, and are used to measure the methylation level. Two methods have been proposed to do this measurement. One is called Beta-value and it is computed according to the following formula:

$$Beta = \frac{Max(M, 0)}{Max(M, 0) + Max(U, 0) + a} \quad (2.1)$$

A Beta-value of 0 equates to an unmethylated site and 1 to a fully methylated CpG site and it can be interpreted as the percentage of methylation. Here a – a small additive constant. Genome Studio uses for this constant 100. However, it is shown in [13, 14] that Beta-value based methods have severe heteroscedasticity in the low and high methylation range, which imposes serious challenges in applying many statistic models. Another method for measuring the methylation level is referred as M-value method. M-value is the \log_2 ratio of methylated and unmethylated probe intensities as depicted in (2.2).

$$M = \log_2 \frac{M}{U} \quad (2.2)$$

M-value method is approximately homoscedastic in the entire methylation range. As a result, it is more statistically valid in differential and other statistic analysis [13, 14].

Illumina has developed three platforms for array-based assessment of DNAm: GoldenGate, Infinium Human Methylation27 or 27k array design and the Infinium HD 450K or 450k array design, which all use two fluorescent dye colors but differ in the chemistries used to recognize the bisulphite-converted sequence. Now, most of the studies focus on the Infinium arrays, as the GoldenGate array has been phased out from production. Illumina Infinium assay utilizes a pair of probes, methylated and unmethylated probes, to measure the intensities of methylated and unmethylated alleles at the interrogated CpG site. As the name suggests it allows the user to map single methylation resolution for 27,578 CpG sites across over 14,000 genes. 450k array design measures more than 450,000 CpG positions and it is employed in most of the recent experiments.

2.2 DNA methylation analysis

Several different methods for quality control, preprocessing and statistical analysis of microarray data were presented in a last few years. However, distinction has to be made among them when methylation data is concerned for analysis. Initially, most of the methods were based on the assumptions and characteristics of GE. Depending on the specific traits of assessment, nature and distribution of DNAm intensities certain adjustments to the available techniques have been made and also separate procedures have been evolved. In this section we try to address the differences and challenges of applying such procedures for DNAm analysis.

2.2.1 Exploratory analysis

Exploratory data analysis is a way of visually analyzing data to summarize its main characteristics. It is often the first step used in most of the data analysis experiments and it can be performed in a several different ways depending on the goal of visualization. One may want to know if there are substantial amount of noise or measurement errors accompanied with the data, or whether some additional information can be discovered. Here we focus our attention on the methods commonly used for biological and particularly for microarray datasets.

After obtaining the data it is important to observe how samples or features are distributed. This can tell us what kind of filtration, adjustment or normalization techniques are better to be applied or are not needed at all. Usually, when we work with microarray data, certain assumptions are taken into considerations. Some of them can be the data source, type of experiment and the dimensionality of dataset. We may want to know if the data comes from human beings, plants or animals, how many people have conducted the experiment, in what way the tasks were distributed, what type of technology is used for measurements and many other relevant information. Being informed about these characteristics of our experiments may narrow down several other considerations to a great extent and form more accurate statistical hypothesis in a further stages.

Since microarray data have a large dimensionality, we can not simply plot and view its distribution, for instance, with scatter plots. In practice, often, dimensionality reduction based techniques are used for biomedical data. First, many researchers start the analysis with unsupervised learning technique. Clustering the the data into groups of genes/methylation sites or classes of samples is the most frequently employed technique and, generally, here the main point is not to discover a new type of class, which is sometimes the case, but rather with the aim of understanding the data, such as detection of genes' patterns, outliers, batch effects and others. In addition to clustering, scatter plots of dimensionality scaled approaches are also popular in this domain.

In this work, we considered hierarchical clustering and two of the dimensionality scaling based plots: Principal Component Analysis (PCA) and Multidimensional Scaling (MDS). With hierarchical clustering, by joining the genes which have similarity in their expression where similarity is defined by distance metrics we can identify such groups of probes that are “turned on/off” in response to the same experimental factor. Briefly, it is a recursive approach and clusters can be constructed either in a bottom-up (agglomerative) or top-down (divisive) way. In bottom-up approach we go by merging the similar probes and then clusters of probes. In top-way case, we split iteratively the set of probes until only one probe is left in a set. At the end we have a tree or dendrogram, and partitions result by cutting it in some level. Distance between two probes of GE/DNA_m can be computed with Euclidean, Manhattan distance or with Pearson correlation. Similarity among clusters of GEs are identified in one of following ways:

- *Single linkage*: Two most similar expressions are concerned;
- *Complete linkage*: Two most distant objects are concerned;
- *Average linkage*: Average pair distance is computed;
- *Centroid*: Distance of central elements is used for similarity

In Results section we show an outcome of applying hierarchical clustering to our DNA_m data.

Another way of visualizing the data is by using scatter plots, but, first, producing the transformation of data from higher to lower feature space is required, since each sample has tens of thousands features. For this purpose, scientists usually employ PCA or MDS plots. PCA is a way to describe samples with an artificial smaller number of variables, called principal components, that account for most of the variance in the observed variables. The principal components serve as a predictor in a further representation of dataset. MDS provides a visual representation of the pattern of similarities or dissimilarities among a set of objects. In a MDS scatter plot samples that are perceived to be similar to each other are placed near each other and those that are perceived to be different are placed far away from each other. The more detailed description of PCA and MDS can be found in [32, 10]. Application of these methods to our data and demonstration of scatter plots are provided in Results section.

2.2.2 Quality control and preprocessing

Quality assessment and preprocessing methods for DNAm datasets are specifically designed by taking into considerations the methylation specific distribution of the data. It is common in almost all types of quality control experiments to observe the distribution of data keeping in mind the process we are analyzing. When we draw a density plot of DNAm we expect more frequency of methylation status with higher values in experimental group than in control group. However, if we try to view this in a density graph drawn for Beta-values it might not be much observable than when we do the same for M-values. The reason for this is explained in [13, 14]. These studies report that for differential methylation analysis it is recommended to conduct experiments with M-values. Therefore it is always advisable to be aware of the consequences of choosing the type of measurement on the outcomes to be obtained from the studies. Since we are going to perform differential analysis of DNAm data and density plot is the first step to visualize this phenomenon we further base our experiments on M-values.

We cannot claim a sample has quality issue if its distribution is quite different

from others. Red and green colors are used by Illumina to label the final extended base following the hybridization of methylated or unmethylated probes. However there is a difference in labeling efficiency and scanning properties of two color channels and this might result the intensities measured in two color channels become imbalanced. We can not ignore color imbalance, because color effects may not necessarily be linear and most of the cases they are not. Therefore, we need to check color balance as it is also one of the indicators of samples' quality.

We can employ a method for sample quality assessment which is based on the across sample distribution and CpG-site intensity. CpG-site intensity is defined as sum of the methylated and unmethylated probe intensities. It is in proportion to the total copies of CpG sites, because only one probe can be bound on a particular CpG site. Also, it can be deduced that methylation levels (log ratio of methylated and unmethylated intensities) changes should not have substantial affect on CpG-site intensity.

It is not appropriate to directly apply normalization methods developed for GE microarray, because many assumptions used for expression data are not valid for methylation. One option is to use quantile normalization method. There are variants of this method for both GE and DNAm data. When applied to methylation probes it considers the sparsity of distribution.

2.2.3 Differential methylation analysis

The task of gene differential expression analysis has become popular mainly after being able to get access to the large amount of data in a form of microarray. There are different approaches that can be considered in this kind of analysis. Most of the used and successfully applied ones are statistical tests and their modifications for GE data. The overview of these methods for microarray experiments is provided in [11].

Although most of the classic and newly developed statistical approaches for microarray GE data can be used for DNAm profiles as well, some researchers proposed that detecting significantly methylated regions might be more suitable than just sin-

gle methylation sites. Searching for significantly methylated regions can have more biological interest and can be more associated with outcome of interest. Moreover, various authors have noted that methylation levels are strongly correlated across the genome [15, 19] and many reported relevant discoveries have been generally associated with genomic regions rather than single CpGs. In this subsection we concentrate our attention on identification of Differentially Methylated Sites (DMSs) and Differentially Methylated Regions (DMRs). We first describe conventional parametric and non-parametric methods which can also be used for finding Significantly Expressed Genes (SEG) and DMSs. Then we provide recently developed and widely used methods for region based analysis.

2.2.3.1 Single site differential analysis

The common statistical issues in identifying differential expression/methylation are test statistics, sample size, replicate structure and statistical significance. Moreover, the fact that usually microarray experiments are carried out with a number of samples which are much more less than the number of genes under investigation, often in the scale of thousands, adds up an extra load to the testing task.

We can employ classical parametric t-test or its modifications such Welch test. However, in practice, the more robust methods used for microarray experiments are usually non-parametric test statistics. They are appropriate when normality cannot be assumed and less sensitive than parametric methods for detecting significant changes. More importantly by using such methods we do decision not only based on the significance levels (p-values) but also we can incorporate other statistics, like fold change. The order of this fold change values corresponds to the order of rank these methods utilize, therefore they are also called rank based methods.

In microarray analysis it is important to control false positive rates, as we perform thousands of tests in parallel which is also referred as multiple testing. For multiple testing tasks people commonly use False Discovery Rate (FDR) that is defined as an expected proportion of Type I errors among the rejected hypotheses, including cases where no hypothesis are significant. It is less conservative than the second error

measure – Family Wise Error Rate (FWER). FWER is defined as the probability of at least one Type I error and since it is very stringent criterion for microarray gene expression analysis most people employ FDR in their methods. Here we can note another advantage of non-parametric testing methods that they are specifically designed for multiple testing and the reduction of type I error (false positive) rate.

There are some popular tools or packages in Bioconductor software project, like Significance Analysis Microarrays (SAM) [36] and limma [33], which are designed for microarray experiments. SAM, in addition to just finding differentially expressed genes, separates results into up- and down-regulated genes and makes useful graphs. limma makes use of linear modeling technique and it is one of the widely used approaches in bioinformatics, because of the interpretability of the model parameters. Moreover, usually in practice we have few replicates per sample which makes it difficult to estimate the gene-specific variances that are used in statistical testing. One way to deal with this issue is using Empirical Bayes approach that employs a global variance estimator computed on the basis of all genes' variances and the resulting test statistics is a moderated t-statistic. With limma package we can select significant probes in a same manner, by controlling FDR and a fold-change, as with SAM tool. We provide results obtained by applying these tools in a Results section.

2.2.3.2 Area based differential analysis

If we have a reason to believe for DNAm to have an effect on GE a region of the genome needs to be affected, not just a single CpG. Therefore, we should look beyond single sites. *First* idea is, first, to find all DMSs and then some combination or grouping technique is applied to these found sites. One strategy is to perform Wilcox test and identify DMSs. Next we can define DMRs by looking for regions for each chromosome where most measured CpGs are differentially methylated. Here definition of DMR can be based on a minimum number of sites in a region, minimum and maximum gaps for probes to be considered in or out of a region, threshold of methylation level difference between experimental conditions and a threshold for significance value (p-value). This and similar to this kind of approaches have been used in a various research works and they are published their approaches implemented

as packages or software tools that can be used by biologists or other researchers. For example, `methyAnalysis` [12] and `COHCAP` [41] packages in Bioconductor conduct their analysis more or less in above described manner.

Second approach also developed specifically for epigenomic data considers the probe spatial information and addresses the batch effect issues. The idea for this approach is based on a topic broadly discussed in a statistical literatures referred to as bump hunting and a group of researchers by adapting this idea created a package in Bioconductor called `bumphunter` [21]. As described by the authors of the package, `bumphunter` estimates regions for which a genomic profile deviates from its baseline value. The main concept of the procedure is that we first perform regression and obtain an estimate for the coefficient of interest. These estimates are then can be smoothed in some clusters of locations that are close enough where the maximum gap for closeness is specified by a user. This gives us an estimate of a genomic profile that is 0 when uninteresting. Candidate regions are taken by thresholding the estimate value of a region. Statistical uncertainty is attached by performing permutations to create and then test a null distributions for the candidate regions.

2.2.3.3 DNA methylation and gene expression integrative analysis

Integration study can be viewed as a work of finding associations between two or more types of data. This kind of analysis can be performed for various purposes such as finding genes, exons or genomic regions which have a crucial effect on or involved in cell differentiation, signaling pathways, cell cycle regulation, disease development or pathogenesis and many others. Depending on the type of application and data types to be integrated different approaches are developed in this area of analysis. In practice, people usually make associations between gene expression and methylation, SNP and gene expression or copy number and gene expression data sources.

Many studies have already demonstrated that making use of additional data source and conduct experiments in a systematic way increases our insight on studying the biological processes under interest. One of such areas where active research works are being carried out by integrated study is medical genomics and especially in cancer

studies. As we noted in Biological background section, hypo- and hyper-methylation is linked with many types of cancer. This fact and other findings motivates researchers even more to study genome scale influence of epigenetic processes. Hence, different types of strategies have arisen. Here we only describe some of them.

In [2] authors pay attention to the fact that other than miRNA expression and DNA methylation affecting on genes they regulate each other in both directions. Doing the study on glioblastoma and ovarian cancer data, they proposed that there might be a cyclic association among these three data types and carried out correlation based analysis. They discretized expression and methylation level values and computed correlation for methylation and miRNA and for GE and miRNA. Their study resulted on finding hsa-miR-142 signature that high correlation with significantly expressed genes and methylated sites. This and many other studies show that correlation is good way of discovering new terms or processes. Comprehensive examination of breast cancer data was done in [28] by integrated analysis of genome-wide DNA methylation and GE data. Authors of the work established, also with correlation analysis, that methylation of not only in promoter but also in different other genomic regions, like CpG shores, first exon, first intron, can be a key for subtype-specific effects.

We found out that several questions need to be addressed in order to detect and define significant associations between data types. Since our work is based on methylation and GE data, here we focus on these two data types. Investigations demonstrate that for DMSs/DMRs and GE we are analyzing on *there should not necessarily be a significant correlation*, because of the methylation occurring not in a region where transcription factors bind or due to many other levels of gene regulation acting simultaneously in a gene. Sometimes it may be helpful also to add non-coding RNAs for the analysis. Besides the biological explanation for not having the correlation between the data types, there might be a statistical or technical issue on designing our experiment. For instance, if for some methylation sites fold change is 2 and p-value is 0.07 these sites will not be called significant, but it is still does not mean that there is an evidence to support that the sites have not changed. In this case it is a lack of evidence for differential methylation rather than an evidence supporting unchanged differential methylation.

2.3 Phenotype prediction task

The task as we called in a general form as a phenotype prediction can have diverse characteristics depending what specifically it is going to predict or for what type of data we are defining the task. In Medical applications physicians and patients may want to know survival time, risk factor, symptoms, drug response or sub-type of some disease. Knowing this information is very crucial for diagnosis and other purposes.

Decisions of which methods and data types to choose, how organize the experiments and process the data usually comes after determining the exact type of application area. For instance, for cancer medicine, now most of the works incorporate GE and methylation data into their study. Applications from evolutionary biology field can be based on mutation or SNP data sources. In addition, techniques and approaches for data processing varies for each these data types.

In this work we do not focus on some specific type of the application our model would be used in, however, *we suppose that most of the times applications would be related to diseases where epigenetic process like methylation plays important role.* We propose a method for constructing a phenotype prediction model that utilizes a methylation data. Data set of patients with MDS disease is used and the model predicts the risk factor or treatment response type of the patients.

2.3.1 Challenges

No general method exist for predicting phenotypes. Most of the procedures are proposed for some specific domain of the tasks. This is explained that we still need a much more understanding of biological processes. As an example we can simply take the process of gene expression regulation. Despite the fact that with advances of technology we can observe this process at any time and condition, its dynamicity and dependence on so many factors leaves several questions being unanswered. Nevertheless, after being able to read the whole genome scientists are getting more insight for various processes.

Now one of the big challenges researchers have after generating huge amount of data with microarray and sequencing technologies is to extract a knowledge from the generated data sources that later would be used for model construction. This can only be accomplished by doing systematic study where at the same time we use several kinds of data such as GE, DNA methylation, SNP, mutation and copy number. In this way we may find patterns common to all data types, but are not viewed when each data source is analyzed separately. Therefore, the integration has become a key step for vast majority of biological studies.

One distinctive character of biomedical data is that the number of features is far more than the number of samples. When we do integrated study of two or more data types the number of samples might decrease further, since not always the measurements for different kind of data are obtained from the same individuals. This may cause a problem known In statistical learning as curse of dimensionality and over-fitting.

Another also can be related with data. Since biological experiments are not always conducted under the same conditions, several issues like batch effect, variability of technical and biological replicates and measurement errors may arise that can reduce the statistical power of the findings or even make them insignificant.

2.3.2 Dataset

In this work we used patients data for MDS disease. The data were provided by our collaborative lab at the Institute of Hematology and Blood Transfusion in Prague. MDS or often referred to as a “bone marrow failure disorder” is a group of diverse bone marrow disorders in which the bone marrow does not produce enough healthy blood cells. Sometimes for some patients diagnosed with MDS, this type of bone marrow failure syndrome will progress to acute myeloid leukemia (AML).

The dataset contains scans for 27578 methylation sites and 31426 GEs obtained from matching samples of 30 individuals for these two feature measurements. These samples represent the patients with five different responses on treatments: Complete

remission (CR), Partial remission (PR), Complete remission with incomplete marrow recovery (Cri), Progressive disease (PD) and Stable disease (SD). In terms of risk associated with these response rates, PD is the most severe and has high risk, CR, Cri and PR can viewed as least severe with lowest risk and SD as intermediate not meeting criteria for all of the rest. In addition to that we have control samples from normal individuals (NR) and one sample for which there is no response type specified. We excluded from prediction model construction experiments the sample with Cri response type and from all analysis sample with unknown type of response.

Response type	Number of samples
CR	4
PR	5
SD	8
PD	5
NR	6
Cri	1
Unknown	1

Table 2.1: Number of samples for each type of response treatment group

2.4 Goal of the work

The main goal of this work is to perform comprehensive integrated study of DNA methylation as epigenetic event that has its effect on the regulation of gene expression, identification of global signature changes regulation of which causes important changes for all types of MDS response treatments and, finally, prediction of phenotype or given a sample of individual determination of group of response treatment this individual can be associated with. We propose a model for phenotype prediction (in our example dataset prediction of response treatment for MDS diseases) that utilizes knowledge obtained from integrated study in a form of significant features for model construction.

Chapter 3

Existing approaches

In this chapter, we would like to give some of the approaches to the problem of phenotype prediction we have found from the current active research in this area. Two important aspects of phenotype molecular model construction – Feature Selection (FS)/Feature Extraction (FE) and the selection of prediction algorithm were addressed in the review of existing methods.

3.1 Feature selection/extraction

FS/FE has become a prerequisite in building a model for biomedical data. Current microarray technology allows to measure simultaneously hundreds of thousands levels of methylation or gene expression for one array or replicate. Working with all of these features may cause a serious problems, two of the main being large input dimensionality and small sample size, and other problems as noted before. That is why, in practice some technique is applied for choosing the most relevant features which improve the model performance in comparison to scenario when all or some other subset data were used.

FS in contrast to extraction only selects significant probes based on some criteria without altering representation and preserving the semantics of variables. In-

interpretability is the main advantage selection methods offer. Review of the feature selection techniques specially designed for the tasks in bioinformatics can be found in [31]. FE methods apply the dimensionality reduction techniques and project the data from higher to lower feature space.

One of the illustrative works that employs feature selection is [40] where authors illustrate that model concentrated on features of integrated GE and DNAm improves the predictive power compared to classifiers trained on GE or methylation data alone. In this study logistic regression classifier with lasso regularization selects features by enforcing sparsity. Similar kind of work and prediction performances were published in [24] where main difference lies in the chosen classification algorithm. What generalizes the [40] and [24] is that they both exploit embedded technique for feature selection. Embedded methods have the advantage that they consider the interaction of variables and include this step with classification model. In other words chosen algorithm accomplishes both feature selection and classification tasks. There are works where we can note filtering and wrapper selection techniques. Some representatives of filtering techniques, like t-statistic, information gain and sum of variances, were chosen for a comparative study [42] and their effectiveness was evaluated. Study concludes that all selected techniques improve the classification outcomes, but some methods' efficiency depends on data. This fact of filtering methods' varied efficiency is explained in [31] as that this kind of selection techniques ignore the interaction with the classifier and hence dependency or correlation among the variables.

In a research studies published in [4, 42] some feature extraction approaches are applied in cancer disease prediction model. The main idea of the methods in both works is that they first with some approach select subset of genes and then for this subset apply space projection techniques. Semi-supervised approach proposed in [4] is designed for predicting survival of patients based on expression profile and survival times of previous patients. Authors identify a list of genes using the clinical data. Genes which have a correlation greater than some threshold are selected for the next processing. Then with the application of unsupervised technique final subset of genes for the prediction is selected. We can consider this approach as mixed, since here both selection by correlation and then feature space transformation are used. Method used

in [42] basically is also of this kind of approach with slight modifications that can be seen on selection of the first list of genes by incorporating different criteria.

3.2 Model design and selection of classification algorithm

Various kinds of models were proposed by researchers for the phenotype classification. Here we highlight some of the techniques often deployed in most of the models and give a brief description.

In [35], for designing the DNA methylation and GE integrated model, authors perform two stage classification scheme. In first stage, both feature selection and training of classifier is done using logistic regression with lasso regularization in a 5-fold cross-validation. Selected features with classifier from each 5-fold training are used in second stage to train the an additional classifier – nearest mean classifier (NMC) that uses the posterior probabilities of GE and methylation logistic regressors as feature space. At the end, output of NMC is evaluated on a validation set. The model from this approach achieves good results and the one reason for this can be a smooth integration step in a second stage of a method where no standardization is required, because regressors of GE and methylation models are in a same scale.

One of the important aspect for the classification problem is the selection of prediction algorithm. What I have observed by reviewing the recent research works is that if feature selection/extraction step is performed efficiently and the genes with good discriminating factor between experimental groups are selected, then even simple prediction methods like k-nearest neighbors or logistic regression can give high performance rates. However, studies [6, 9, 1] show that one of the methods giving constantly high classification rates in a wide variety of applications is SVM [38]. Also, it is worth note that the performance of models with linear kernel SVM is same or sometimes better than radial kernel SVM.

Chapter 4

Proposed model

In this chapter we describe the general design of the two major types of models. First type of models are built on single data types (DNAm and GE) and on blind integration based approach. Second class includes four classes of models constructed using knowledge based methods, hence they are called smart integration incorporated models. We in some parts refer to the first class of models as simple and to the second as intelligent or smart. We cover the used techniques and steps followed in a model construction. In addition, specification of parameters for different methods is illustrated in the chapter.

4.1 General design

Building the proposed model for this work includes the following main steps:

1. Data preprocessing
2. Identification of significant features for each data type
3. Integration of data types and detection of signatures
4. Generation of feature sets

5. Classification model construction
6. Evaluation

First and last two steps are identical for all types of models and these steps are enough for us to design two of the simple models. In the next subsections we discuss the special moments for each kinds of models.

The *first step* of the work is to check the quality of the GE and methylation data by employing exploratory tools often used in bioinformatics. Then we normalize the data with quantile normalization algorithm. For GE we applied general quantile normalization developed for microarray GE data and specially modified version of an algorithm for DNAm where modification considers distribution distinctive for DNAm.

In a *next step*, we found SEG and DMSs. For this purpose we use linear model based moderated t-statistic. This statistic is robust and it is most frequently used one for differential analysis.

SVM classification method with linear classifier is chosen for devising the prediction model for all types of models in a *fifth step*. One versus the rest or one class of samples versus all the other types scheme is used for training and testing. To assure unbiased measurements of the performance of the classifier we followed 4-fold outer and 3-fold inner CV schemes. Outer CV is intended for training and evaluating the models and with inner we performed the tuning of the classifier parameters.

To observe the change of prediction performance and to determine best model several classifier with varying number of features were constructed. Each time one new feature is added to build a new classifier. Finally, models were evaluated, *in a last step*, with average probability assigned to the correct class. We draw a graph where we can see how increasing the number of features affect the model efficiency. Ideally we want as less features as possible, but we also need to account overfitting issue.

4.1.1 Single Data types and Blind Integration based models

DNAm and GE data based approaches basically follows the above described general steps. We can add that for simple models initially specified 100 number of top features were preselected for model construction. Ranking of the features were done with the same moderated t-statistic we used for differential analyses.

Step 2, Identification of significant features, for the our blind integration based model is performed after Step 3, integration of data types. Simply speaking, we disregarded any “relationships” between the data types and just merged them in integration step.

Then we performed standardization of feature values, since M-values and GE levels differs and it is necessary to have all features in a similar range of values, so that for some features not being overstressed in a model construction. Significant features for this merged and standardized dataset is found in a similar way as it is done for single data types.

4.1.2 Smart Integration based models

For our intelligent models we need devise a feature set which is based on integration of two data types. There are different ways of doing this and we chose the procedure which is provided in 2.2.3.3. The main idea is the association strength between found DMRs and their corresponding GE levels.

For computing the correlation we need to represent DMRs for the computation, as DMR according to our definition is the region with at least two DMSs. We called this representation as FE, since the meaning of our features changes after transformation.

Feature Extraction: We need to define clearly what do we mean by FE. In our work it is very simple procedure. All features (probes or methylation sites’ levels) located in a region are averaged and described as one feature (or CpG site).

Having obtained DMRs as features we compute a correlation, perform test of

significance for the correlation value and identify significant relationships. Then all DMRs and their corresponding GE profiles from significant relationships are added to the new dataset. Since usually we have less such signatures and this result to have in the dataset very few features, we included all other found DMRs. Significance of DMRs not being correlated with GE profiles on various biological processes was demonstrated in many number of research studies, so it makes sense to include all DMRs and results from our experiments show this as well.

One last thing to note is that *we should not forget that DNAm provides us with additional knowledge information and the main insight of the questions we are studying still lies within GE data*. So, *we hypothesize* that when we have both additional patterns from integration study and main patterns from GE data we possibly can have the best result among all other models considered in this work.

Depending of the presence of either FE or FS and existence of additional SEGs enrichment step or not existence, there are four workflows involved for designing respective intelligent models:

- **Workflow 1:** Model design based on FE and addition of extra SEGs – Smart Model with FE and extra SEGs (SmartModFeSEG)
- **Workflow 2:** Model construction involving FS and additional SEGs - Smart Model with FS and extra SEGs (SmartModFsSEG)
- **Workflow 3:** Model built with FE and without additional SEGs - Smart Model with FE (SmartModFe)
- **Workflow 4:** Model designed with FS and without extra SEGs - Smart Model with FS (SmartModFs)

4.2 Parameters specification

In two parts in our work where we need to specify parameters. First, when we are looking for DMSs and DMRs and, the second, when we are training and testing the classifier.

During the detection of DMSs and DMRs we specify threshold or cut off parameters. We set 0.3 or 30 % of methylation level for the site to be considered differentially methylated. Both expression and methylation probes were considered significant if they met 5% significance level (p-value equal to 0.05) and 100% FDR. We did not obtain any result with lower rates of FDR.

For SVM with linear kernel method we need to specify only one - cost parameter. After internal CV for each training classifier we found that 0.1 value for this parameter is optimal one.

Chapter 5

Experiments

In this chapter we presented the methodology and results for our experiments. All experiments were carried out in R statistical language of version 3.2 using Bioconductor software project of version 1.18

5.1 Methodology

5.1.1 Experimental setup for differential analysis

For finding the differentially expressed and methylated probes for each experimental response groups of diseases first thing we need to do design our experiments. As represented in Dataset section the data set of patients with MDS diseases is divided into 5 groups of samples: CR, PR, SD, PD and NR. Based on these groupings and the severity type associated with patients of each group I devised 3 experiments for each group of response types. All four disease groups (CR, PR, SD and PD) were tested against NR and against all the rest types of groups. So, for instance, for CR group CR versus NR and CR versus PR, SD, PD, NR scenarios are created. In addition to these one more scenario is possible and it is created against a group(s) that has closest severity rank position and there is only one such kind of case: CR,

PR, SD versus PD. We associated findings from this experiment as belonging to all four disease groups. Significant GE profiles were detected using the limma package where linear model based moderated t-statistic is used as a main method. DMSs and DMRs are found with COHCAP package. We could also use limma for identification of DMSs, but chose COHCAP, because it facilitates the process of finding DMRs and also we need this package for integration analysis.

5.1.2 Specification for integration study

Pearson correlation method is used to calculate a association between a DMR and corresponding GE profile. Statistical significance for the correlation is computed by doing 1000 times repeated permutation test. Additionally, the result of Pearson correlation were verified with other rank based correlation calculation methods and the same outcomes were produced.

Here we again used the COHCAP package which also allows users to conduct integration study. It annotates each region (island) with corresponding gene, so that we do not need to specify the information about the annotation, but only need to provide GE data.

5.1.3 Setup for model construction

4-fold CV is used for training and evaluation the classifier. Moreover, classifier parameter is tuned in a separate inner 3-fold CV.

Experiments for building a prediction model were carried out using the CMA Machine Learning package in Bioconductor. This package makes CV and parameter tuning steps easy to implement and also evaluate a model with various evaluation criteria.

5.2 Results

5.2.1 Data preparation

We first look to the whole methylation status data by producing hierarchical clustering and plot the data by reducing dimensionality with PCA and MDS methods. Then we analyze the quality of this data by viewing density plots of methylation statuses for each group type of samples.

The cluster dendrogram was produced by complete hierarchical clustering. In Figure 5.1, we can view that although the separation is not completely right most of the samples of one type are in a same cluster.

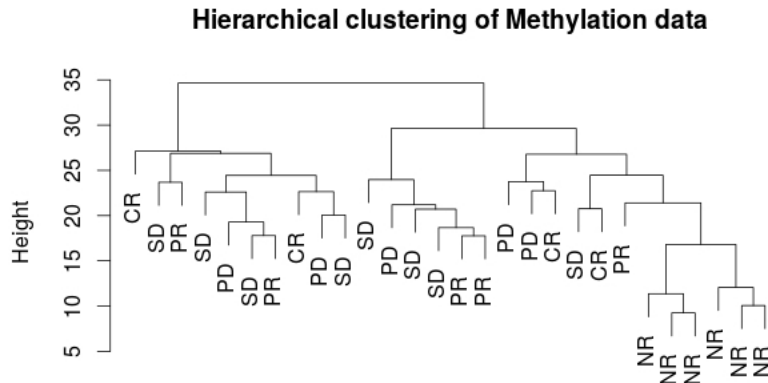


Figure 5.1: Cluster dendrogram for methylation data of MDS disease dataset

The PCA and MDS plots tell us about the same amount of information. Specifically in Figure 5.2 we can see that only samples in Normal (NR) group are placed close to each other and other types of samples are somewhat mixed.

When we draw a density plot of DNAm profiles we expect more frequency of methylation statuses with higher values in disease associated response groups than in Normal group. This fact is noticeable in our next figures. from the plots in Figure 3, but not as obviously as we can see in other methylation analysis works. Both Figure

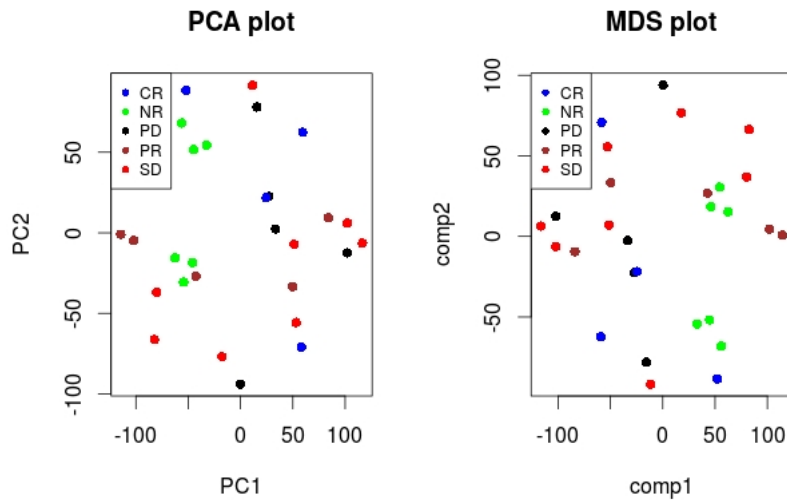


Figure 5.2: Scatter plot with dimensionality reduction by PCA and MDS

5.3 and Figure 5.4 demonstrate the same kind of graphs, but they are produced with different measurements of methylation levels. The phenomenon of methylation specific distribution and the necessity for preprocessing are more visible in graph generated with M-values than in Figure 5.3. The reason for this difference is briefly mentioned in 2.2.2 and more in [6,7].

Then we expect to see similar across sample difference on CpG-site intensity distribution of methylation for different conditions. Figure 5.5 shows box plot of two color channels and density plot for PD and Normal groups of samples.

Last graph shows clear color imbalance and a need for background color adjustment and data normalization. We chose quantile normalization method specifically designed for methylation data. We accomplished the task of color adjustment and data normalization by using the method from Bioconductor package lumi [6].

In a Figure 5.6 and Figure 5.7, we provide results after normalization. Graph of M-values for all four experimental groups against Normal one and also density of CpG-site intensity for PD and Normal samples can be seen in Figure 5.6. Figure 5.7 demonstrates box plot of CpG-site intensity for PD and Normal samples shown for two channels separately. We can see evident effect of normalization by comparing the

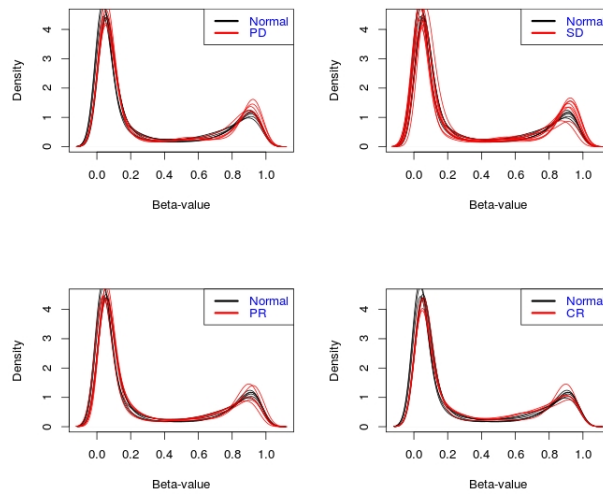


Figure 5.3: Density plots of Beta - values for four types of treatment (PD, SD, PR, CR) versus reference - normal group samples

graphs with previous ones for the cases before normalization.

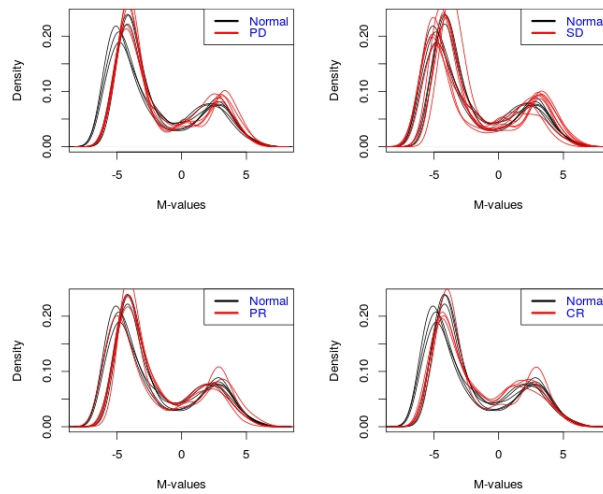


Figure 5.4: Density plots of M-values for four types of treatment (PD, SD, PR, CR) versus reference - normal group samples

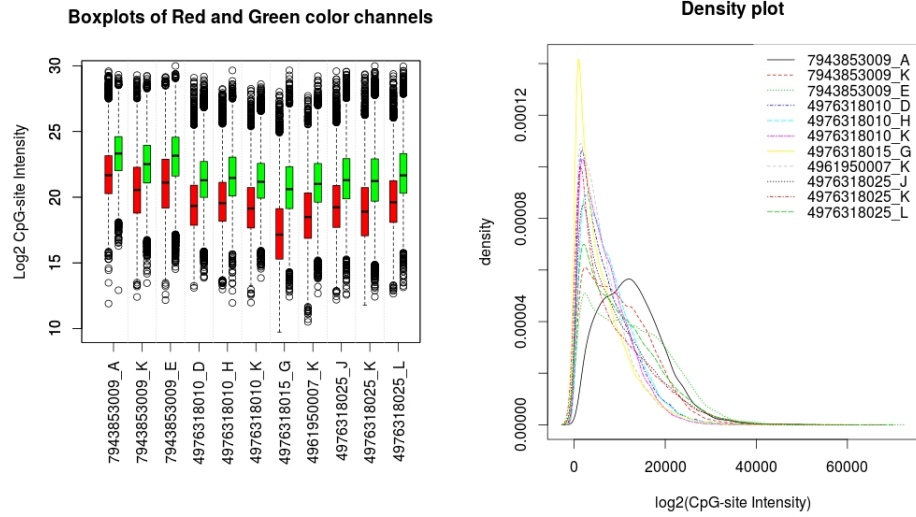


Figure 5.5: Plots for PD and Normal samples before preprocessing. Left. Box plot of CpG-site intensity separately for two channels. Right. Density plot.

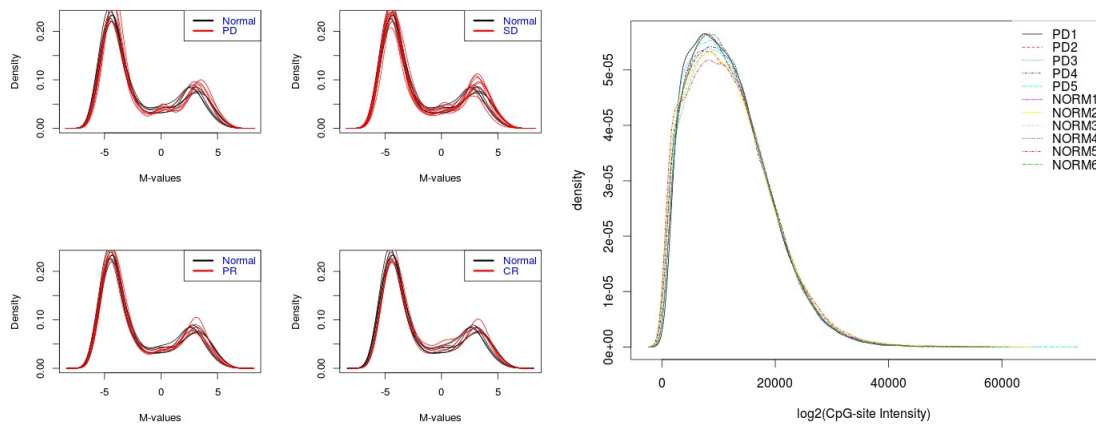


Figure 5.6: Plots after quantile normalization. Left. Density plot of M-values for four disease groups versus Normal samples. Right. Density plot of CpG-site intensity for PD and Normal samples

5.2.2 Findings from differential and integration study

We carried out differential analysis for both GE and DNAm data. Because methods for identifying DMSs can be almost the same, here, we provide results of differ-

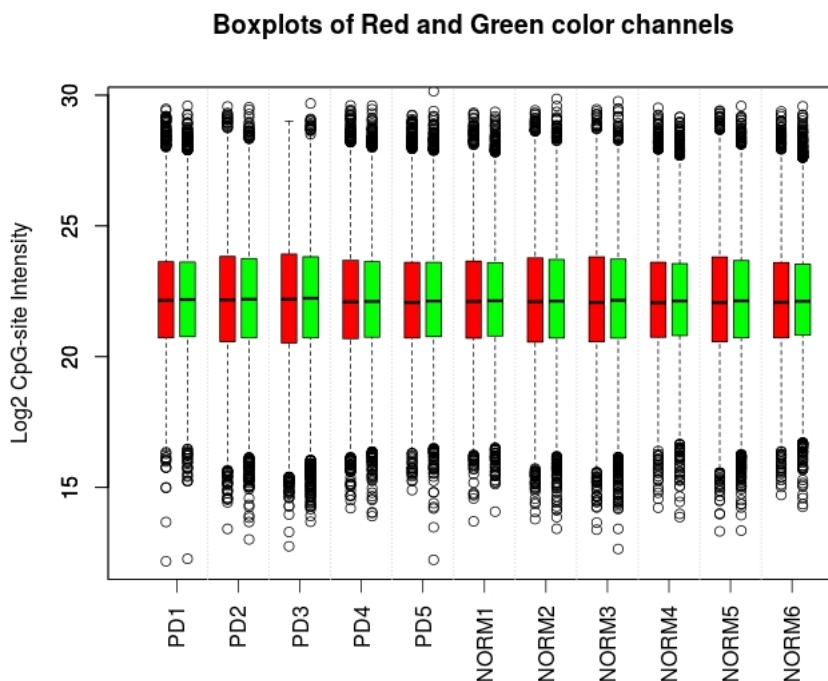


Figure 5.7: Box plot of CpG-site intensity for PD and Normal samples after quantile normalization (shown for two channels separately).

ential analysis for DNAm data. Moreover, when we want to find DMRs this would require special dealing with the problem which is not the same as for DMSs.

First we run t-test that assumes equality of population variances followed by a test using Welch - modification of t-test not bound to such assumption. The histogram of p-values for these two tests for finding DMSs in PD samples are presented in Figure 5.8.

1986 DMSs were found with simple T-test and 2435 with it's modification – Welch test under the significance level of 0.05 without adjusting the p-values. After adjusting the p-values using the method proposed by Benjamini and Hochberg only one DMS was left for T-test and 1098 for a Welch test with 50

Next by making use of SAM package in Bioconductor we conduct non-parametric

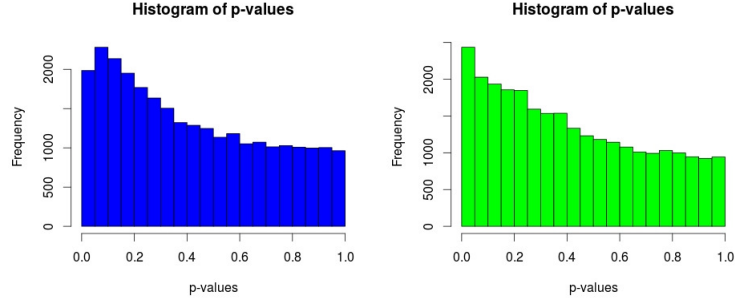


Figure 5.8: Histogram of p-values for the T-test with population variance equality assumption (Left) and for Welch's test without such assumption(Right)

statistical multiple testing. In these methods repeated permutations of the data are used to determine DMSs. We run testing for 1000 permutations. Wilcoxon test statistics and 20% FDR were used in the experiment. Table of significant DMSs were computed separately for sites that have positive and negative correlation (up- and down-expressed sites) with the condition of interest. For the two group experiment where one group contains PD and the other all remaining types of samples 4754 DMSs were obtained.

Methylation sites found significant in non-parametric Wilcoxon test is not much different from those we got using parametric t-test. However, it is much more robust and designed for multiple testing, so that we can control it by FDR and based on an experiment select significant probes. Also in addition to the FDR we can choose significant probes by fold change. SAM provides such functionality that results will be returned based on both statistics. As we can see one advantage of SAM is possibility to make decision or narrow down the list of results based on both error rates (significance values) and fold change. In our case we only used FDR, since when we incorporated fold change statistics no DMSs were found.

It is also useful to look at the SAM plot, shown in Figure 5.9, where methylation of CpGs corresponding to the observed scores of above or lower band are those that are found significant.

Further, we used limma package where using linear modeling and moderated t-

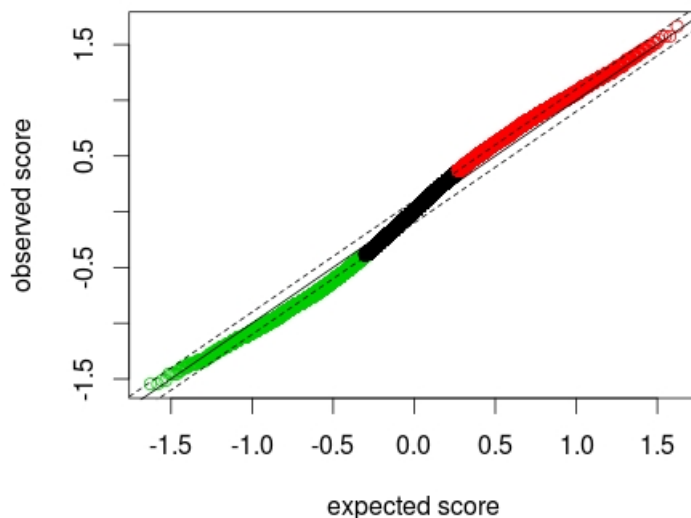


Figure 5.9: SAM plot fitting the expected results to the observed ones. This is a graph from two group experiment: PD type samples versus all remaining samples

statistics we identified DMSs. In Figure 5.10, significant probes for type PD samples were presented in a volcano plot with blue color. In the figure significant probes were obtained only by controlling FDR rate, since incorporating fold-change reveals no DMSs.

As definition of our features in a phenotype model is based on DMRs, further we provide results for identifying significant regions and corresponding genes that contain these regions.

We applied first approach described in 2.2.3.1 and found 34 DMRs for PD samples. Then we annotated them to know with what genes they overlap. In a Figure 5.11 heat map of methylation by gene CALCA is presented. This plot allows us visually verify the result of DMRs detection by specified gene.

In Figure 5.12, we present the result of found one DMR. To get this result we have exploited the second, bump hunting, approach described in 2.2.3.2. The in a graph there are two DMSs found in a two group experiment: PD versus Normal.

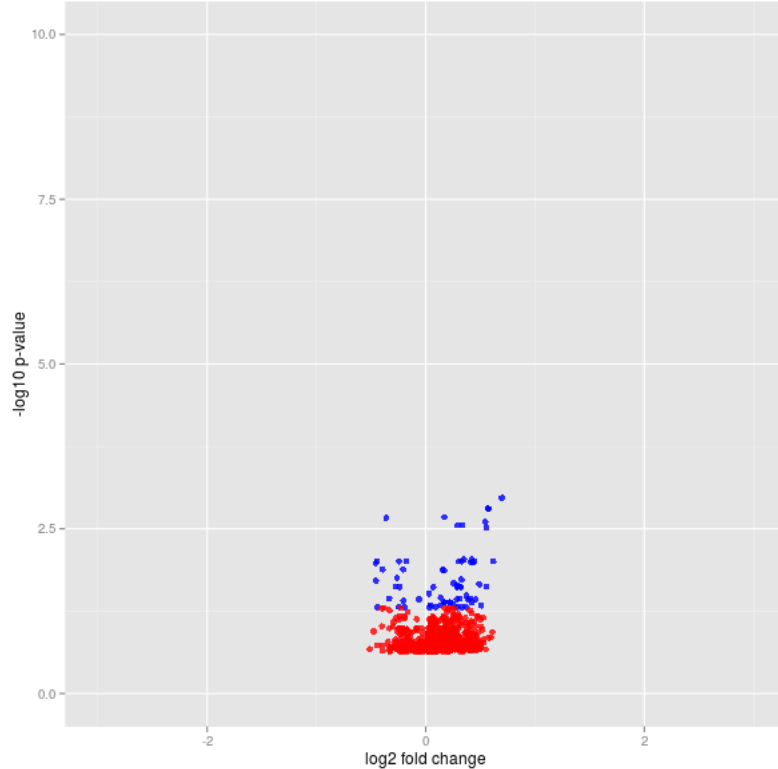


Figure 5.10: Volcano plot. Finding significant probes by linear modeling. Two group experiment: *PD vs Normal* type of samples

We compared the results of two approaches and concluded that in terms of performance there is not much difference between two strategies. In a later steps for identifying the DMRs we chose the first approach. This is because in this way we can perform both feature selection and feature extraction. In other words, with first method we are able to find DMRs and at the same time know which sites in the regions are differentially methylated and which are not. While with the second bump hunting method, though it is also possible, it requires additional step for filtering out significant features for using them later in our model.

Finally, after differential analysis where we found in average 4-15 DMRS for each response group associatsed with diseases we carried out integration analysis. This experiment is done in a way described in 4.1 for all disease related groups and some important signatures were revealed from the analysis. Genes ***CRMP1*** and

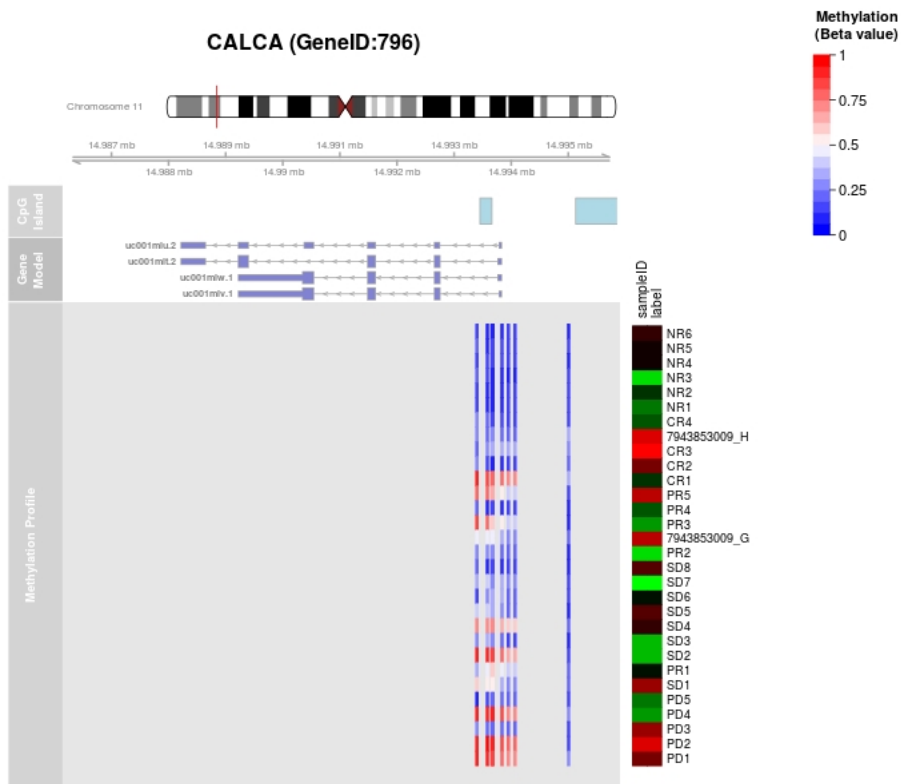


Figure 5.11: Heat map of the first DMR location, corresponding gene and a sample information are shown in this graph

VAMP5 showed downstream expression regulation for CR, PR and SD response groups. Furthermore, same regulation is observed for the gene *EDNRB* in CR and for genes *ZNF154* and *ZNF540* in SD groups. For the most severity type of response group PD no signature was detected. Besides these findings, GE profiles for genes *MGC15523* and *WT1* displayed significant expression in almost all experiments and in all cases where some DMRs were found for some experimental group there was also one DMR associated with these genes.

In Figure 5.13 we provide box plots showing the status of methylation for two DMRs and then in Figure 5.14 scatter plots depicting the correlations of two DMRs with the gene expression levels from the integration study.

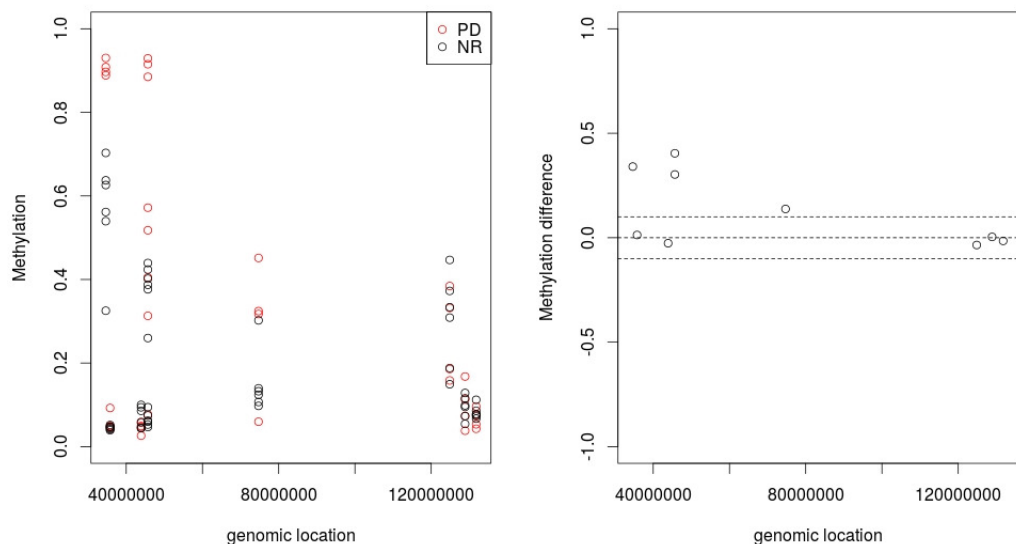


Figure 5.12: Showing one found region from PD vs NR experiment. Left: Plot of methylation sites. Two of these sites are differentially methylated. Right: Depiction of the difference in methylation levels

5.2.3 Model evaluation

Here present the evaluation of our model. In Tables 5.1 – 5.4 we provide the evaluation results and the number of features used to obtain the “best model”. The definition of the best model is optimistically biased, since we are selecting the one which has highest average probability assigned to the true class and it might not necessarily be the case that it is the best model. However, we believe that it is the case and we define our way of choosing the model as such. As we hypothesized model before, highest performance model was found as the one which is built with both SEGs and additional pattern knowledge gained by DNAm and GE interaction analysis and between the two feature set construction techniques the FS based one was superior than FE dependent model. In three of the group types (PD, SD and CR) this model beats remaining models. In Figure 5.15 we show the plots of this best model where average probability progression assigned to correct class is in vertical axis and in horizontal axis the number of used features are depicted.

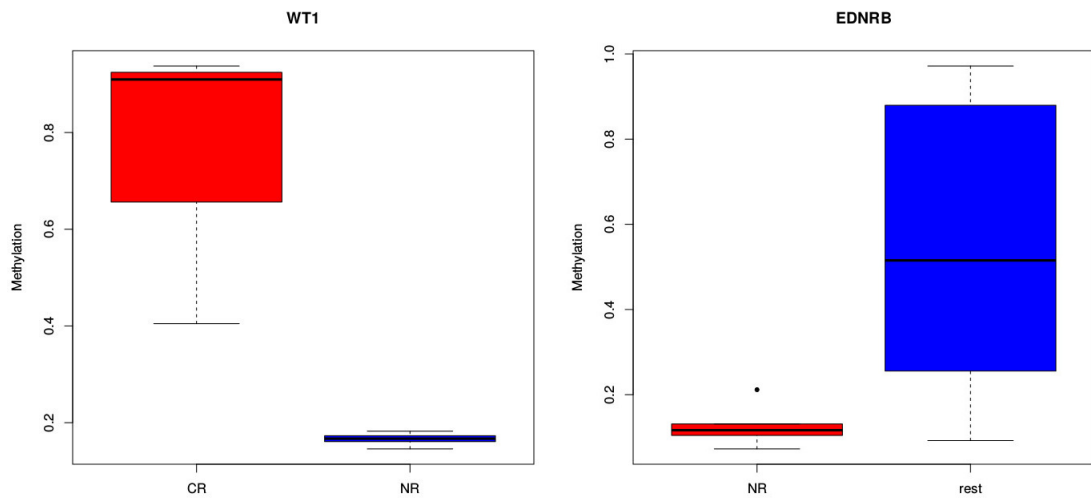


Figure 5.13: Box plot for two differentially expressed genes that have significant methylation regions

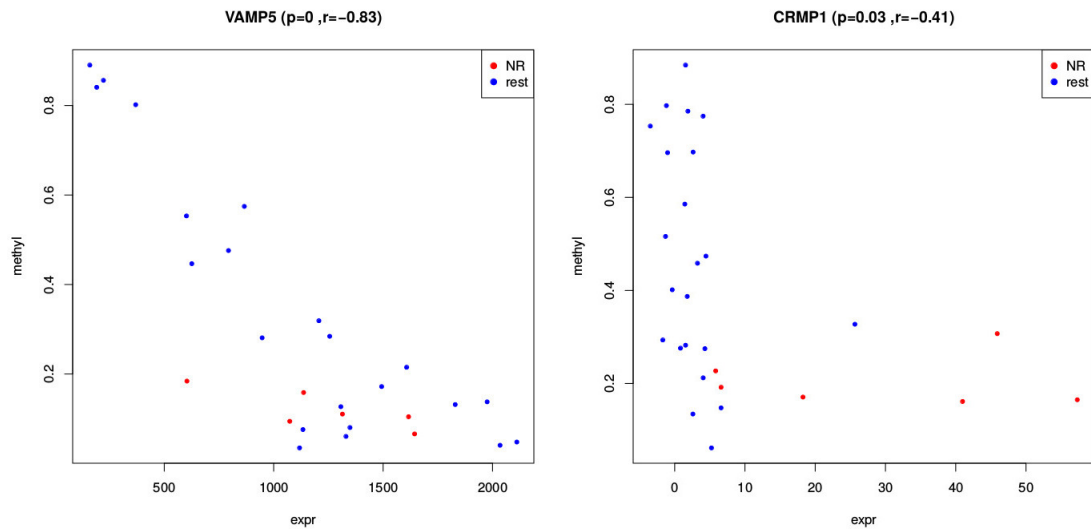


Figure 5.14: Scatter plots of methylation statuses for two CpG islands and their corresponding GE levels

Class types	Average true class probability		
	BlindModFS	ExpModFS	MethModFS
PD	0.72	0.74	0.74
SD	0.63	0.61	0.63
PR	0.71	0.75	0.69
CR	0.82	0.79	0.82

Table 5.1: Best average true class probability performance for blindly integrated (BlindModFS) and two single data type based models (ExpModFS, MethModFS)

Class types	Number of used features (probes) to get optimistically biased best performance		
	BlindModFS	ExpModFS	MethModFS
PD	7	5	3
SD	13	6	3
PR	100	85	89
CR	1	18	3

Table 5.2: Number of features (probes) employed to obtain the best performance for BlindModFS, ExpModFS and MethModFS models

Class types	Average true class probability			
	SmartModFe	SmartModFs	SmartModFeSEG	SmartModFsSEG
PD	0.72	0.77	0.84	0.90
SD	0.74	0.73	0.73	0.74
PR	0.72	0.76	0.87	0.82
CR	0.75	0.79	0.83	0.84

Table 5.3: Best average true class probability performance for four smart integration based models: SmartModFe, SmartModFs, SmartModFeSEG and SmartModFsSEG

Class types	Number of used features (probes) to get optimistically biased best performance			
	SmartModFe	SmartModFs	SmartModFeSEG	SmartModFsSEG
PD	4	14	21	51
SD	16	21	18	33
PR	3	3	4	11
CR	16	10	15	4

Table 5.4: Number of features (probes) employed to obtain the best performance for four smart integration based models

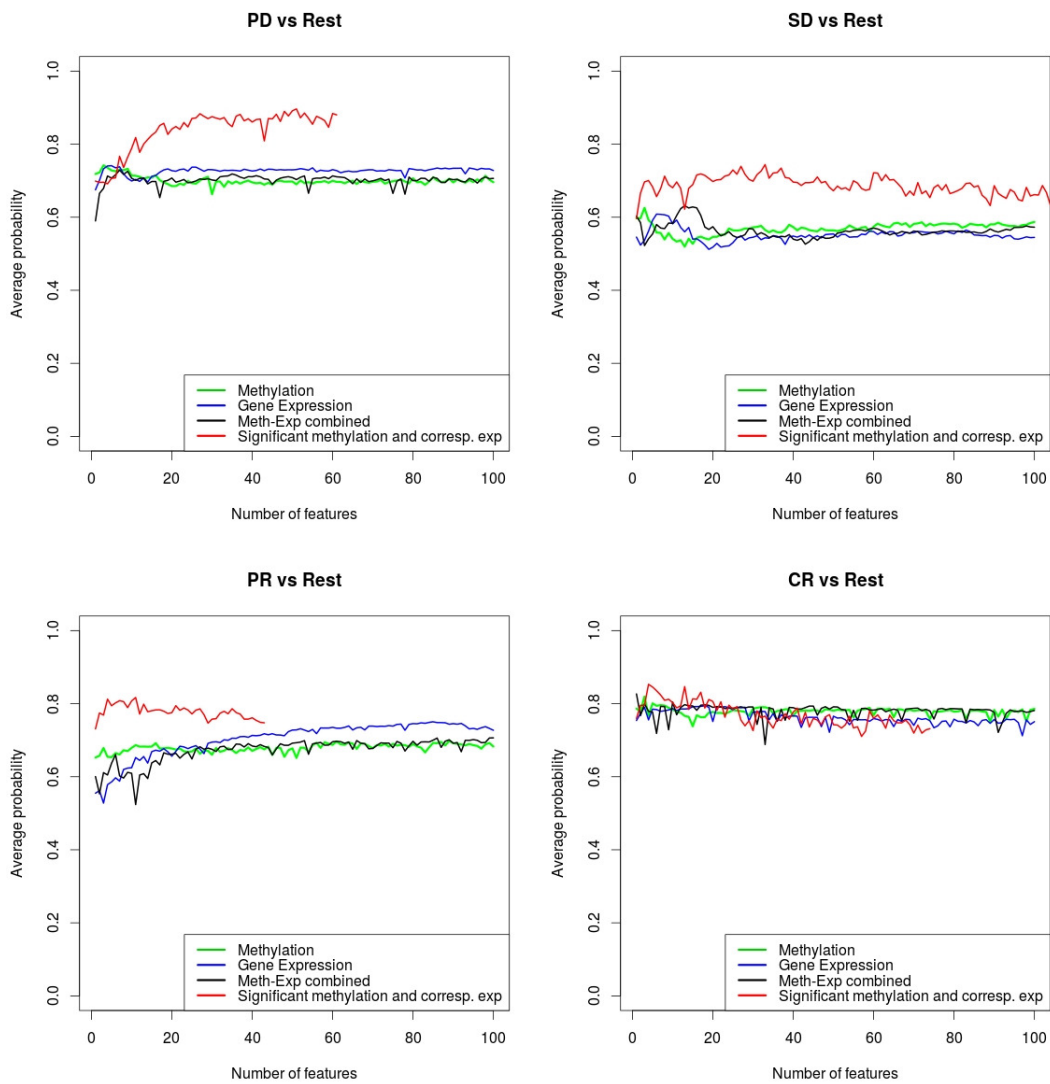


Figure 5.15: Evaluation of SmartModFsSEG model. Plots are for each one of four groups of response types. Number of features is in horizontal axis and the average probability is in vertical axis

Chapter 6

Conclusion

In this work we attempted to design a phenotype molecular model that can be used in an phenotype predictive applications where the most valuable information or knowledge can be obtained using GE and DNAm data types. Usually, exploitation of these data can be observed in biomedical domains with the main focus directed towards diseases, such as various kinds of cancer. After doing some research, we have found that almost all research works with the high success rates of outcomes, make use of two or more types of data.

Concentrating on GE and DNAm data types, we hypothesized that the phenotype model that is based on DNAm and GE profiles can has more potential to perform better than the models built on single data types. Furthermore, we made additional proposals stating that the model constructed not just on a merged dataset, but on a feature set which comprises only significant patterns found by integration study possibly tend to give even better results. We referred to these patters as the patterns that can complement or enrich the main signatures obtained by GE profiles.

The results of the evaluation of the models show that the models based on simple or blind integration that is with the feature set devised by just merging the two data sources has the varying performance and most of the time has more or less the same predictive power (depends on phenotype class type) as the models built with GE and DNAm data types alone. However, model designed with only primary GE patterns

or SEGs and extra knowledge gained by integration study between GE and DNAm profiles clearly outperforms all of the remaining design cases. Experiments show that FS is more appropriate for classification model training/testing, while region based integration analysis yields more valuable signatures than the technique based on single methylation sites.

We understand that our method of model evaluation is not much robust and optimistically biased. Nevertheless, we believe that the true model performance is about in a same level that we reported.

We also inform that our model is implemented in a way to be compatible and to be added as an additional feature to the the miXGENE [17] tool developed for learning from heterogeneous gene expression data using prior knowledge.

Bibliography

- [1] H. Abusamra. A comparative study of feature selection and classification methods for gene expression data of glioma. *Proceedings of the National Academy of Sciences*, 5(23):14, 2013.
- [2] B. Andreopoulos and D. Anastassiou. Integrated analysis reveals hsa-mir-142 as a representative of a lymphocyte-specific gene expression and methylation signature. *Cancer informatics*, 11:61, 2012.
- [3] J. Andrews, W. Kennette, J. Pilon, A. Hodgson, A. B. Tuck, A. F. Chambers, and D. I. Rodenhiser. Multi-platform whole-genome microarray analyses refine the epigenetic signature of breast cancer metastasis with gene expression and copy number. *PLoS One*, 5(1):e8665, 2010.
- [4] E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS biology*, 2(4):e108, 2004.
- [5] R. Bejar. Clinical and genetic predictors in myelodysplastic syndromes. *Haematologica*, 99:956–964, 2014.
- [6] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of computational biology*, 7(3-4):559–583, 2000.
- [7] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, J.-B. Fan, and R. Shen. High density dna methylation array with single cpg site resolution. *Genomics*, 98:288–295, 2011.

- [8] J. Bradbury. Human epigenome project up and running. *PLoS biology*, 1(3):e82, 2003.
- [9] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, 2000.
- [10] A. Buja, D. F. Swayne, M. L. Littman, N. Dean, H. Hofmann, and L. Chen. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472, 2008.
- [11] X. Cui and G. A. Churchill. Statistical tests for differential expression in cdna microarray experiments. *Genome Biology*, 4:210, March 2003.
- [12] P. Du and R. Bourgon. *methyAnalysis: an R package for DNA methylation data analysis and visualization*, 2015.
- [13] P. Du, G. Feng, W. A. Kibbe, and S. Lin. *lumi: a Bioconductor package for processing Illumina*, 2010.
- [14] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11:587, 2010.
- [15] F. Eckhardt, J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, et al. Dna methylation profiling of human chromosomes 6, 20 and 22. *Nature genetics*, 38(12):1378–1385, 2006.
- [16] M. E. Figueroa, S. Lugthart, Y. Li, C. Erpelinck-Verschueren, X. Deng, P. J. Christos, E. Schifano, J. Booth, W. van Putten, L. Skrabanek, F. Campagne, M. Mazumdar, J. M. Grealley, P. J. Valk, B. Löwenberg, R. Delwel, and A. Melnick. Dna methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell*, 17(1):13–27, january 2010.
- [17] M. Holec, V. Gologuzov, and J. Klema. mixgene tool for learning from heterogeneous gene expression data using prior knowledge. *Proceedings of the National Academy of Sciences*, pages 247–250, May 2014.

- [18] R. Holliday and J. E. Pugh. Dna modification mechanisms and gene activity during development. *Science*, 187:226–232, 1975.
- [19] R. A. Irizarry, C. Ladd-Acosta, B. Carvalho, H. Wu, S. A. Brandenburg, J. A. Jeddloh, B. Wen, and A. P. Feinberg. Comprehensive high-throughput arrays for relative methylation (charm). *Genome research*, 18(5):780–790, 2008.
- [20] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(5):245–254, 2003.
- [21] A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, 41(1):200–209, 2012.
- [22] E. Lee, H. Jung, P. Radivojac, J.-W. Kim, and D. Lee. Analysis of aml genes in dysregulated molecular networks. *BMC Bioinformatics*, 10:S2, 2009.
- [23] D. H. K. Lim and E. R. Maher. Dna methylation: a form of epigenetic control of gene expression. *The Obstetrician Gynaecologist*, 12(1):37–42, 2010.
- [24] M. List, A.-C. Hauschild, Q. Tan, T. A. Kruse, J. Mollenhauer, J. Baumbach, and R. Batra. Classification of breast cancer subtypes by combining gene expression and dna methylation data. *Journal of integrative bioinformatics*, 11(2):236, 2014.
- [25] F. Model, P. Adorjan, A. Olek, and C. Piepenbrock. Future selection for dna methylation based cancer classification. *Bioinformatics*, 17(Suppl 1):S157–S164, 2001.
- [26] J. Nordlund, C. L. Bäcklin, V. Zachariadis, L. Cavelier, J. Dahlberg, I. Öfverholm, G. Barbany, A. Nordgren, E. Övernäs, J. Abrahamsson, T. Flaegstad, M. M. Heyman, Ólafur G Jónsson, J. Kanerva, R. Larsson, J. Palle, K. Schmiegelow, M. G. Gustafsson, G. Lönnnerholm, E. Forestier, and A.-C. Syvänen. Dna methylation-based subtype prediction for pediatric acute lymphoblastic leukemia. *Clinical epigenetics*, 7:11, February 2015.

- [27] V. K. Rakyan, T. Hildmann, K. L. Novik, J. Lewin, J. Tost, A. V. Cox, T. D. Andrews, K. L. Howe, T. Otto, A. Olek, et al. Dna methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS biology*, 2(12):e405, 2004.
- [28] J.-K. Rhee, K. Kim, H. Chae, J. Evans, P. Yan, B.-T. Zhang, J. Gray, P. Spellman, T. H.-M. Huang, K. P. Nephew, et al. Integrated analysis of genome-wide dna methylation and gene expression profiles in molecular subtypes of breast cancer. *Nucleic acids research*, 41(18):8464–8474, 2013.
- [29] C. M. Roth. Quantifying gene expression. *Current issues in molecular biology*, 4(4):93–100, 2002.
- [30] D. S.A., K. K.R., R. R., O. G., R. K.S., I. H., M. Y., and A. G.K. Revolutionizing detection and expression analysis of genes. *Current Genomics*, 8(1):234–251, 2007.
- [31] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.
- [32] J. Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.
- [33] G. K. Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.
- [34] M. Szyf. Dna methylation signatures for breast cancer classification and prognosis. *Genome Med*, 4(3):26, 2012.
- [35] E. Taskesen, S. Babaei, M. M. Reinders, and J. de Ridder. Integration of gene expression and dna methylation profiles improves molecular subtype classification in acute myeloid leukemia. *BMC Bioinformatics*, 16(Suppl 4):S5, 2015.
- [36] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

- [37] P. J. Valk, R. G. Verhaak, M. A. Beijen, C. A. Erpelinck, S. B. v. W. van Doorn-Khosrovani, J. M. Boer, H. B. Beverloo, M. J. Moorhouse, P. J. van der Spek, B. Löwenberg, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *New England Journal of Medicine*, 350(16):1617–1628, 2004.
- [38] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.
- [39] R. G. Verhaak, B. J. Wouters, C. A. Erpelinck, S. Abbas, H. B. Beverloo, S. Lugthart, B. Löwenberg, R. Delwel, and P. J. Valk. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*, 94(1):131–134, 2009.
- [40] Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. Mayer, and H. W. Mewes. Gene selection from microarray data for cancer classification—a machine learning approach. *Computational biology and chemistry*, 29(1):37–46, 2005.
- [41] C. D. Warden, H. Lee, J. D. Tompkins, X. Li, C. Wang, A. D. Riggs, H. Yu, R. Jove, and Y.-C. Yuan. Cohcap: an integrative genomic pipeline for single-nucleotide resolution dna methylation analysis. *Nucleic acids research*, page gkt242, 2013.
- [42] T. Wilhelm. Phenotype prediction based on genome-wide dna methylation data. *BMC bioinformatics*, 15(1):193, 2014.

Appendix A

List of Software

R	Programming language and software environment for statistical computing
Bioconductor	Open Source and open development software project for the analysis of genomic data obtained from molecular biology experiments.
matplotlib	Python 2D plotting library, together with IPython provides a MATLAB-like environment for explorative programming and data visualisation. The convergence graphs were created using it.
teTeX	A complete TeX distribution for UNIX compatible systems.

Appendix B

Contents of the CD

The CD content is divided into the following directories:

data	Dataset of patients with MDS disease and data related files meant to be processed.
doc	Additional information related to the directory structure and notes to the actual implementation of the scripts.
latex	L ^A T _E X source codes of this text.
pdf	This text in the pdf format.
scripts	Relevant scripts in R used for experiments.