Computational Analysis of Methylome Sequencing Data

Master Thesis Bioinformatics

Till Helge Helwig

Eberhard-Karls-University Tübingen
Wilhelm-Schickard-Institut für Informatik
&
Max Planck Institute for Developmental Biology

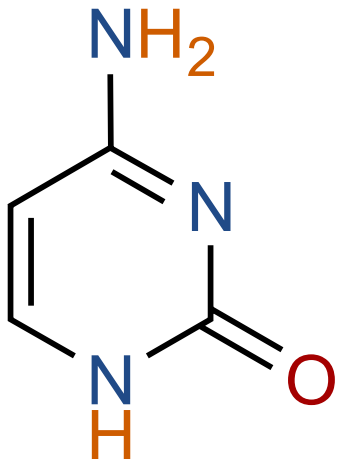
February 22, 2011

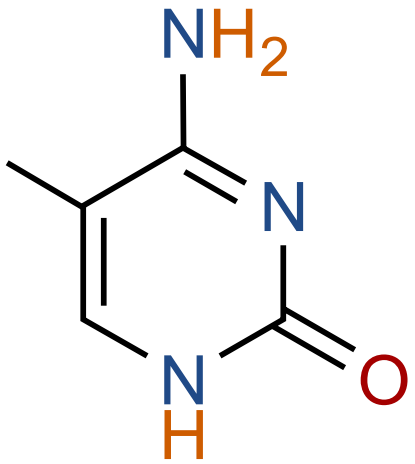
Outline

Outline

The Methylome

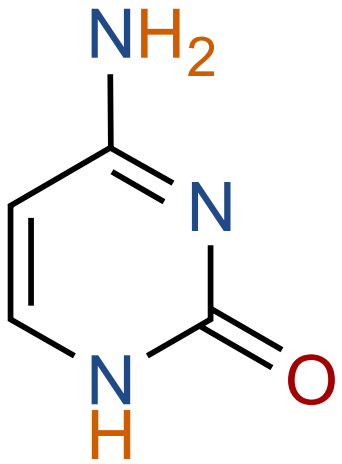
- Entirety of **methylated nucleotides** (e.g. cytosines) in the *DNA*
- Addition of a methyl group converts cytosine into **5-methylcytosine**

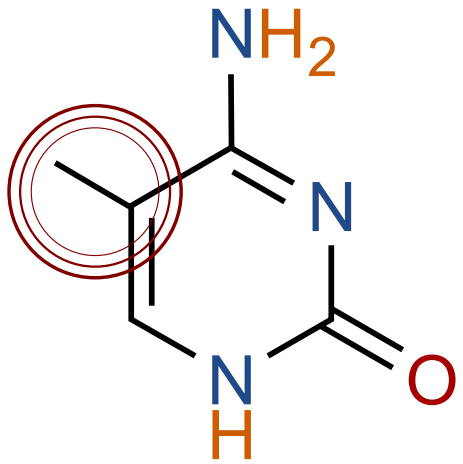




The Methylome

- Entirety of **methylated nucleotides** (e.g. cytosines) in the *DNA*
- Addition of a methyl group converts cytosine into **5-methylcytosine**





Properties of the Methylome

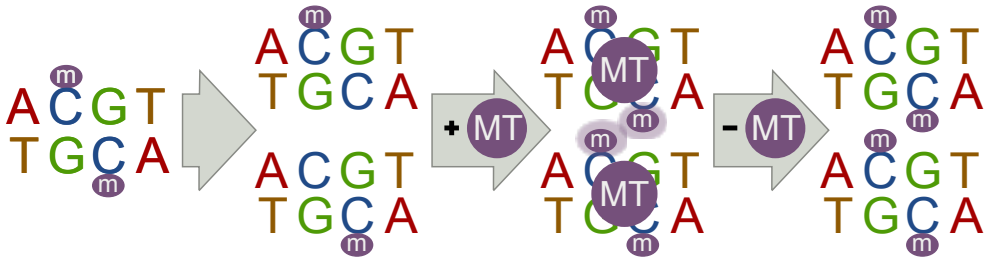
- **Additional layer of information** within the DNA

Properties of the Methylome

- **Additional layer of information** within the DNA
- Methylations are created by **methyltransferases**

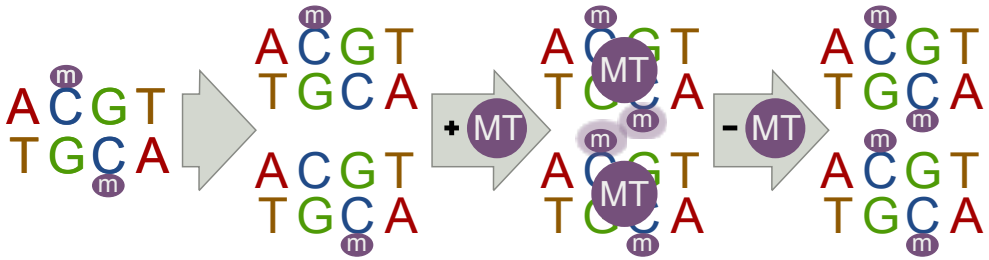
Properties of the Methylome

- **Additional layer of information** within the DNA
- Methylations are created by **methyltransferases**
- Maintenance of methylations after **transcription**



Properties of the Methylome

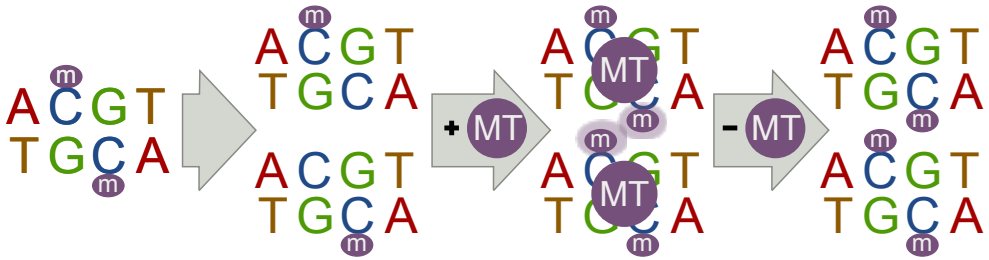
- **Additional layer of information** within the DNA
- Methylations are created by **methyltransferases**
- Maintenance of methylations after **transcription**



- **Environmental factors** influence the methylome

Properties of the Methylome

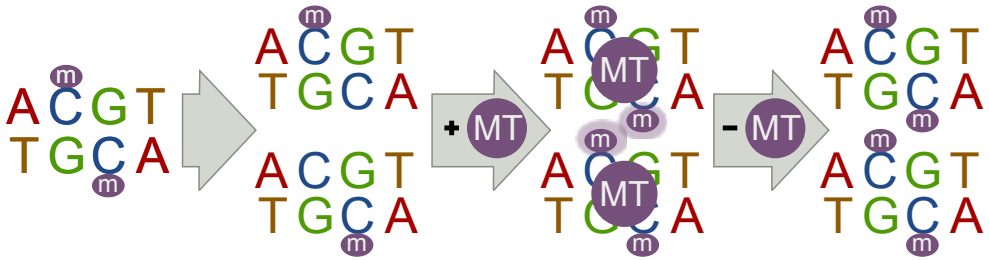
- **Additional layer of information** within the DNA
- Methylations are created by **methyltransferases**
- Maintenance of methylations after **transcription**



- **Environmental factors** influence the methylome
- The methylome is **highly variable**...
  - ...between different **species**

Properties of the Methylome

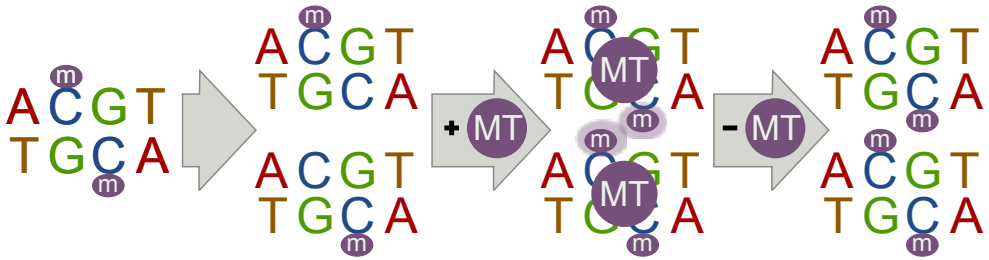
- **Additional layer of information** within the DNA
- Methylations are created by **methyltransferases**
- Maintenance of methylations after **transcription**



- **Environmental factors** influence the methylome
- The methylome is **highly variable**...
  - ...between different **species**
  - ...between **organisms** of the same species

Properties of the Methylome

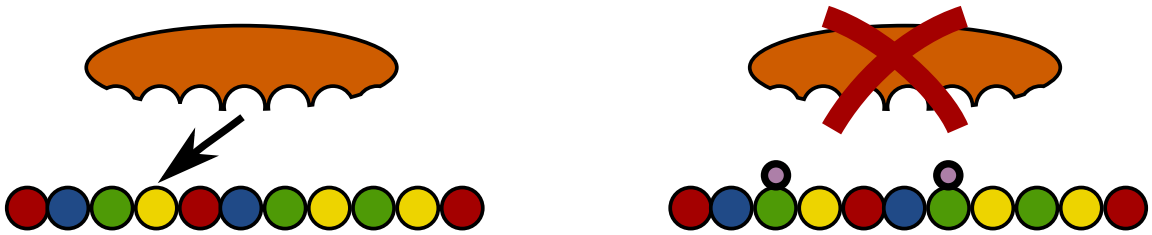
- **Additional layer of information** within the DNA
- Methylations are created by **methyltransferases**
- Maintenance of methylations after **transcription**



- **Environmental factors** influence the methylome
- The methylome is **highly variable**...
  - ...between different **species**
  - ...between **organisms** of the same species
  - ...between different **cell types** of the same organism

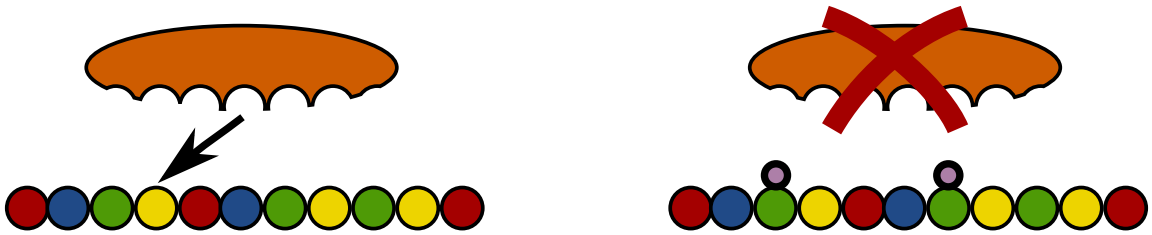
Transcription Inhibition

- Methylated nucleotides can **inhibit** the transcription



Transcription Inhibition

- Methylated nucleotides can **inhibit** the transcription



- Relevance for different research fields:
  - **Developmental biology** (e.g. for association studies)
  - **Medicine** (e.g. for tumorgenesis)
  - **Ecology** (e.g. documentation of environmental changes)
  - ...

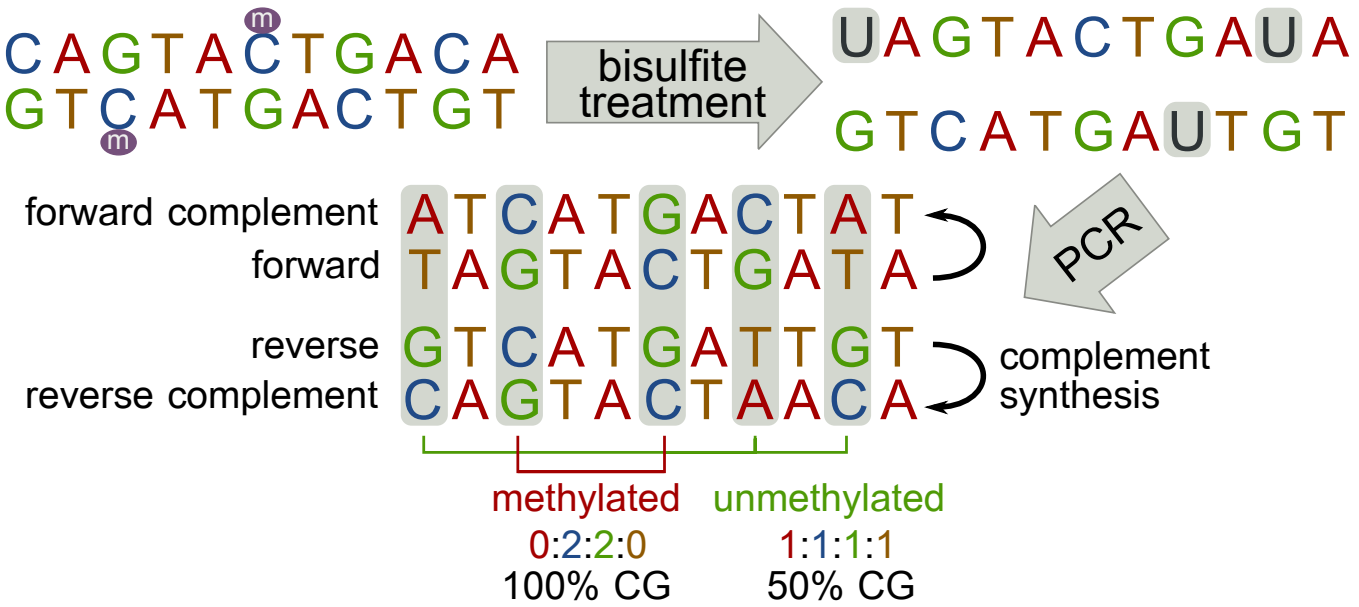
Outline

Making the Methylome Visible

- **Standard sequencing** can not identify methylations

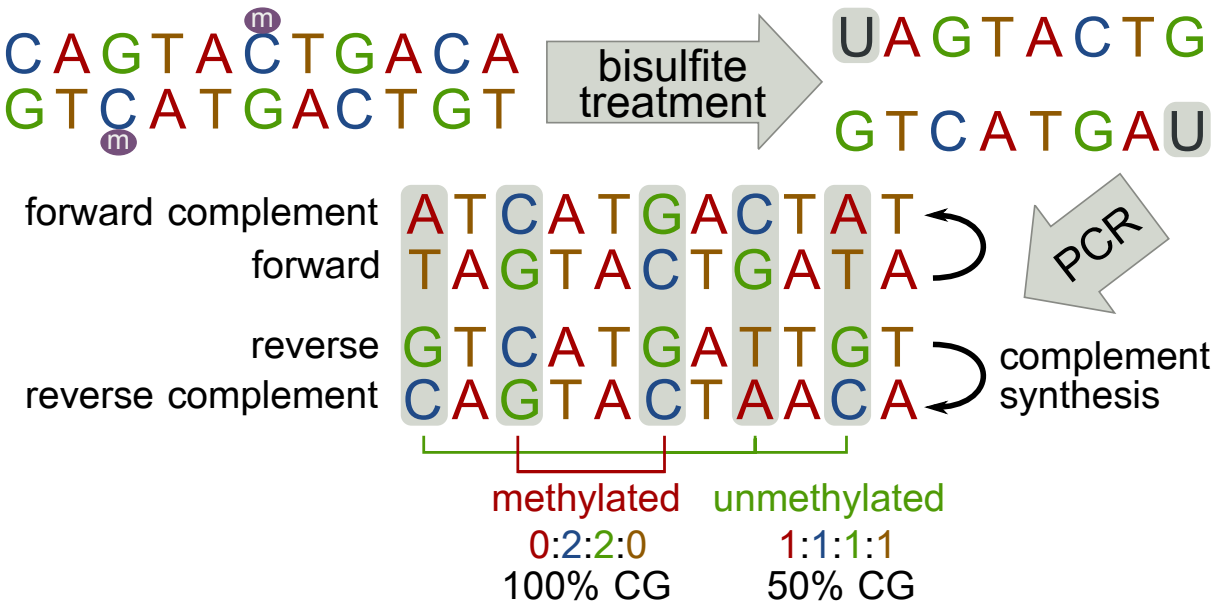
Making the Methylome Visible

- **Standard sequencing** can not identify methylations
- **Bisulfite treatment** makes methylations visible:



Making the Methylome Visible

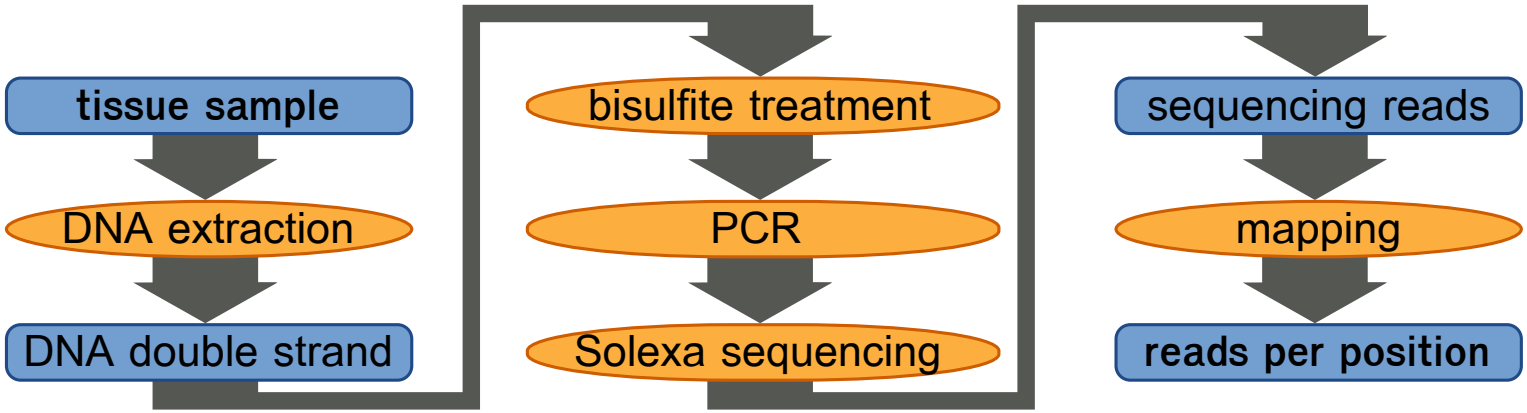
- **Standard sequencing** can not identify methylations
- **Bisulfite treatment** makes methylations visible:



- Sequencing now reports only **methylated cytosines**

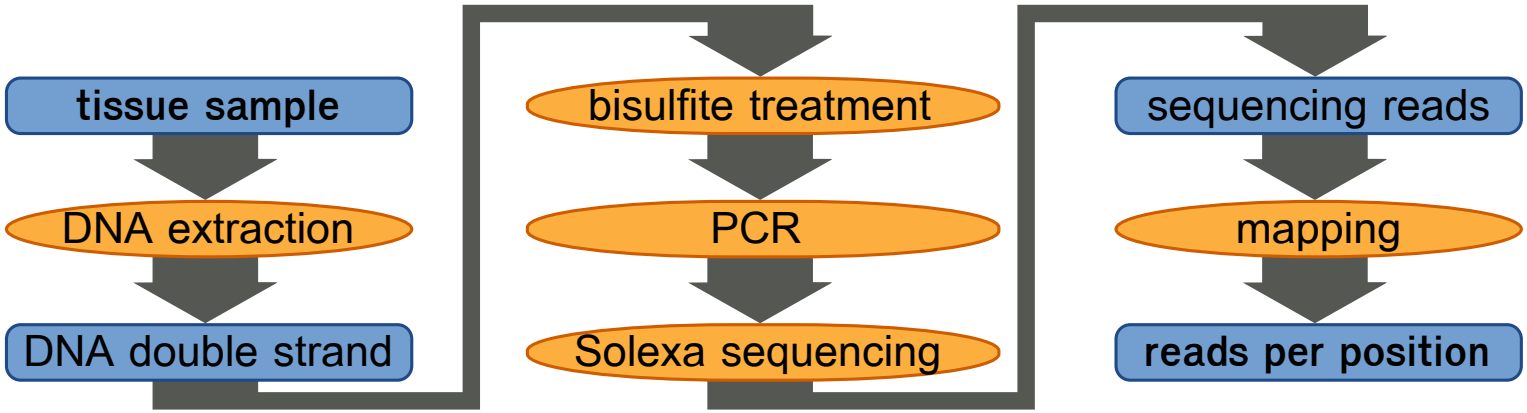
Sequencing Protocol

- Bisulfite treatment inserted into the **sequencing protocol**



Sequencing Protocol

- Bisulfite treatment inserted into the **sequencing protocol**



- **Methylation rates** calculated from the read counts per position

Methylome Sequencing is Imprecise

Bisulfite treatment

Has a significant conversion error rate.

$\Rightarrow$ **Can be estimated from the mitochondrium DNA.**

Methylome Sequencing is Imprecise

Bisulfite treatment

Has a significant conversion error rate.

$\Rightarrow$ **Can be estimated from the mitochondrium DNA.**

PCR

Might contain a preference for certain strands.

$\Rightarrow$ **Difficult to take into account.**

Methylome Sequencing is Imprecise

Bisulfite treatment

Has a significant conversion error rate.

⇒ **Can be estimated from the mitochondrium DNA.**

PCR

Might contain a preference for certain strands.

⇒ **Difficult to take into account.**

Sequencing

Reports wrong nucleotides sometimes.

⇒ **Accuracy value is reported as well.**

Methylome Sequencing is Imprecise

Bisulfite treatment

Has a significant conversion error rate.
⇒ **Can be estimated from the mitochondrium DNA.**

PCR

Might contain a preference for certain strands.
⇒ **Difficult to take into account.**

Sequencing

Reports wrong nucleotides sometimes.
⇒ **Accuracy value is reported as well.**

Mapping

Problematic due to repetetive regions and reduced sequence complexity.

Outline

Improvement via Machine Learning

- Methyltransferases need some form of **binding sites**

Improvement via Machine Learning

- Methyltransferases need some form of **binding sites**
- Binding sites are **patterns** in the DNA nucleotide sequence

Improvement via Machine Learning

- Methyltransferases need some form of **binding sites**
- Binding sites are **patterns** in the DNA nucleotide sequence
- Patterns can be **learned** in order to be recognized in new data

Improvement via Machine Learning

- Methyltransferases need some form of **binding sites**
- Binding sites are **patterns** in the DNA nucleotide sequence
- Patterns can be **learned** in order to be recognized in new data

Idea

Use machine learning to obtain an additional **confidence measure** based on sequence patterns.

Requirements

- Needs to handle **full genomes**

Requirements

- Needs to handle **full genomes**
- Will be used on **newly sequenced** genomes

$\Rightarrow$ Should not rely on more than the **nucleotide sequence**

Requirements

- Needs to handle **full genomes**
- Will be used on **newly sequenced** genomes

⇒ Should not rely on more than the **nucleotide sequence**

- Quantification of the **likelihood** for candidate nucleotides to be methylated

⇒ **Confidence score** between 0.0 and 1.0

Dataset for Training and Test

Problem

No dataset available with confirmed methylations.

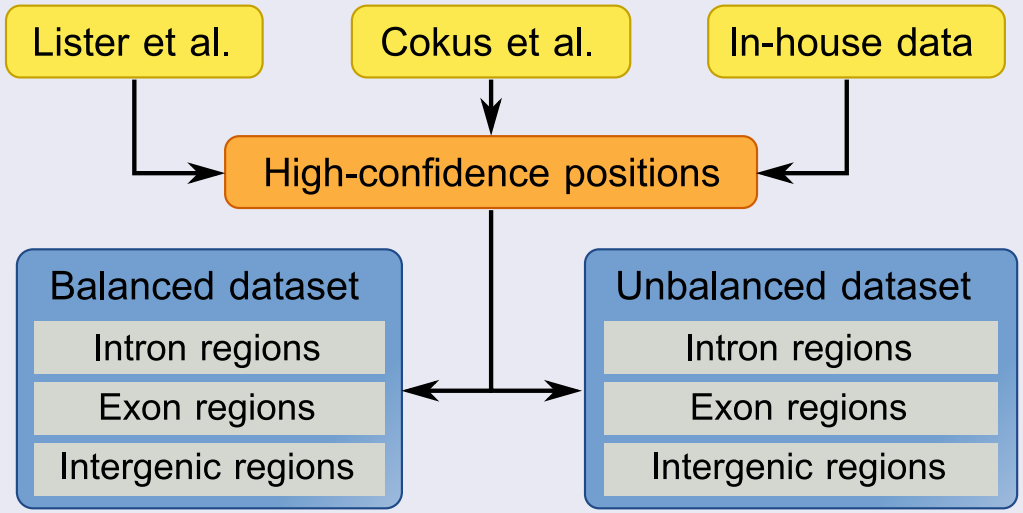
Dataset for Training and Test

Problem

No dataset available with confirmed methylations.

Solution

Manual creation of a **high-confidence dataset**



Experimental Setup

- **Support Vector Machines**

Experimental Setup

- **Support Vector Machines**
- 3 different **kernels**:

Experimental Setup

- **Support Vector Machines**
- 3 different **kernels**:
  - ***k*-Spectrum Kernel**
    (considers substring occurences in the input strings)

Experimental Setup

- **Support Vector Machines**
- 3 different **kernels**:
  - ***k*-Spectrum Kernel**
    (considers substring occurences in the input strings)
  - **Extension of the *k*-Spectrum Kernel**
    (considers additionally the position of the substrings)

Experimental Setup

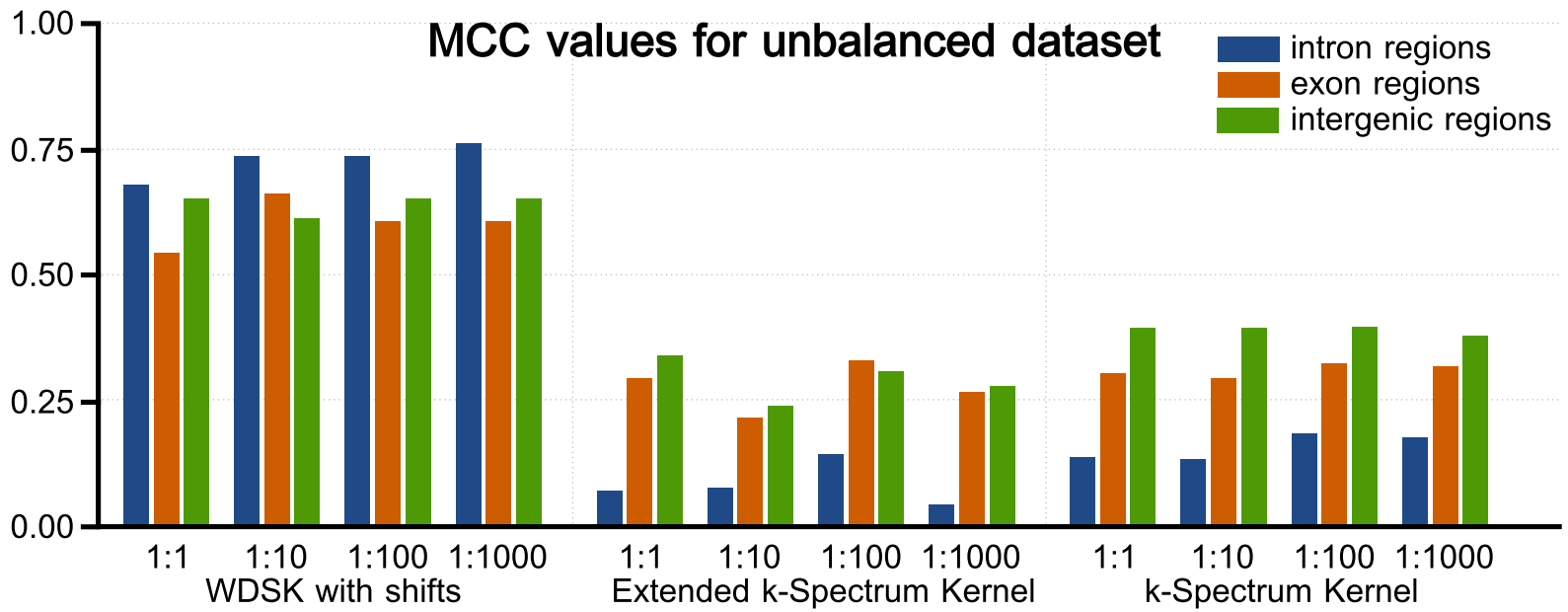
- **Support Vector Machines**
- 3 different **kernels**:
  - ***k*-Spectrum Kernel**
    (considers substring occurences in the input strings)
  - **Extension of the *k*-Spectrum Kernel**
    (considers additionally the position of the substrings)
  - **Weighted Degree String Kernel with shifts**
    (adds weights to account for substring shifts, substring lengths and substring positions)

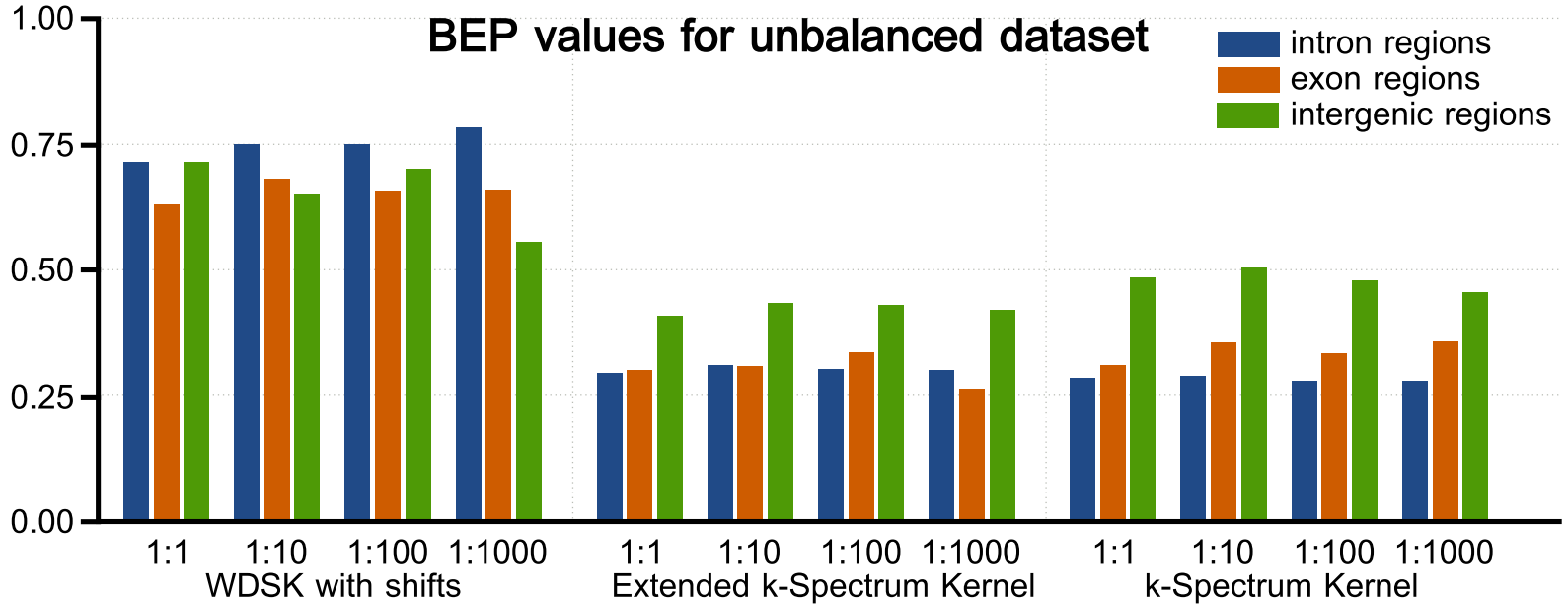
Experimental Setup

- **Support Vector Machines**
- 3 different **kernels**:
  - ***k*-Spectrum Kernel**
    (considers substring occurences in the input strings)
  - **Extension of the *k*-Spectrum Kernel**
    (considers additionally the position of the substrings)
  - **Weighted Degree String Kernel with shifts**
    (adds weights to account for substring shifts, substring lengths and substring positions)
- Prediction of methylations on **whole genome** with best classifiers

Outline

Performances on Unbalanced Dataset





Obtaining the Confidence Value

- The **three best classifiers** (WDSK with shifts) used to predict for whole genome

Obtaining the Confidence Value

- The **three best classifiers** (WDSK with shifts) used to predict for whole genome
- **Balanced classifiers** performed badly (59% methylation rate)

Obtaining the Confidence Value

- The **three best classifiers** (WDSK with shifts) used to predict for whole genome
- **Balanced classifiers** performed badly (59% methylation rate)
- **Unbalanced classifiers** report 6% methylation rate (7% expected)

Obtaining the Confidence Value

- The **three best classifiers** (WDSK with shifts) used to predict for whole genome
- **Balanced classifiers** performed badly (59% methylation rate)
- **Unbalanced classifiers** report 6% methylation rate (7% expected)
- SVM calculates a **confidence value**

Obtaining the Confidence Value

- The **three best classifiers** (WDSK with shifts) used to predict for whole genome
- **Balanced classifiers** performed badly (59% methylation rate)
- **Unbalanced classifiers** report 6% methylation rate (7% expected)
- SVM calculates a **confidence value**
- **However:** Few reported methylated positions occur in original datasets

What Did We Learn?

Biology

- Methylation state to some degree reflected by the **neighboring nucleotides**

What Did We Learn?

Biology

- Methylation state to some degree reflected by the **neighboring nucleotides**
- **No unique patterns** identifying methylated positions

What Did We Learn?

Biology

- Methylation state to some degree reflected by the **neighboring nucleotides**
- **No unique patterns** identifying methylated positions
- Different properties of methylated positions in varying **genomic regions**

What Did We Learn?

Biology

- Methylation state to some degree reflected by the **neighboring nucleotides**
- **No unique patterns** identifying methylated positions
- Different properties of methylated positions in varying **genomic regions**

Bioinformatics

- Application of supervised learning methods requires **more reliable datasets**

What Did We Learn?

Biology

- Methylation state to some degree reflected by the **neighboring nucleotides**
- **No unique patterns** identifying methylated positions
- Different properties of methylated positions in varying **genomic regions**

Bioinformatics

- Application of supervised learning methods requires **more reliable datasets**
- **Unbalanced data** is more realistic but leads to additional complexity

A Look Into the Crystal Ball

- Research toward **validation** of methylome datasets

A Look Into the Crystal Ball

- Research toward **validation** of methylome datasets
- More extensive study using more **parameter values** and more **complex features**

A Look Into the Crystal Ball

- Research toward **validation** of methylome datasets
- More extensive study using more **parameter values** and more **complex features**
- **Relaxation** of confidence threshold in example selection

A Look Into the Crystal Ball

- Research toward **validation** of methylome datasets
- More extensive study using more **parameter values** and more **complex features**
- **Relaxation** of confidence threshold in example selection
- Thorough analysis of **methylome variability** between species, organisms and cell types

A Look Into the Crystal Ball

- Research toward **validation** of methylome datasets
- More extensive study using more **parameter values** and more **complex features**
- **Relaxation** of confidence threshold in example selection
- Thorough analysis of **methylome variability** between species, organisms and cell types
- **Unsupervised** learning methods

A Look Into the Crystal Ball

- Research toward **validation** of methylome datasets
- More extensive study using more **parameter values** and more **complex features**
- **Relaxation** of confidence threshold in example selection
- Thorough analysis of **methylome variability** between species, organisms and cell types
- **Unsupervised** learning methods
- **Recent research** promises methylome data as byproduct of standard sequencing

Thank you for your attention!

Acknowledgements

- Prof. Dr. Daniel Huson
- Prof. Dr. Detlef Weigel
- Dr. Karsten Borgwardt
- MLCB group
- WeigelWorld
- Jörg Hagmann

Most important sources:

S. J. Cokus, S. Feng, and S. E. Jacobsen.

Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.
Nature, 452(7184):215–219, 2008.

R. Lister, R. C. O'Malley, and J. R. Ecker.

Highly integrated single-base resolution maps of the epigenome in Arabidopsis.
Cell, 133(3):523–536, 2008.

G. Rätsch, S. Sonnenburg, and B. Schölkopf.

RASE: recognition of alternatively spliced exons in C. elegans.
Bioinformatics, 21(suppl 1):i369, 2005.

K. Schneeberger, J. Hagmann, and D. Weigel.

Simultaneous alignment of short reads against multiple genomes.
Genome Biology, 10:R98, 2009.

B. Schölkopf and A. J. Smola.

Learning with kernels: support vector machines, regularization, optimization, and beyond.
MIT Press, 2002.