**Probabilistic model**

We aim to develop an approach to model gene expression levels as a function of DNA methylation and promoter sequence data. The goal is to determine the driver (transcriptional) regulators for a particular set of genes derived in a certain abnormality. The expectation is that the transcriptional regulators work via a common motif profile to influence the gene expression levels. In addition, it is known that expression levels can also be influenced by DNA methylation. We define a probabilistic graphical model that integrates gene expression, DNA methylation and DNA sequence motifs into a Bayesian model.

We start with a set of genes from which we want to determine how the expression levels are influenced by transcriptional regulators and/or change in DNA methylation levels. The set of genes can be derived by various phenotypical states, such as by comparing the gene expression levels of the group of interest versus the control samples using e.g. the student T-test (with *P* < alpha, after multiple testing), or by selecting a set of genes that represents a particular pathway or process which is relevant to be studied. As an example, genes can be selected that are present in a particular pathway from the e.g. molecular signature database (MsigDB), such as the WNT-pathway.

**Data.** The data that we will be used: gene expression profiles (Affymetrix hgu133p2), DNA methylation profiles (HELP assay) and the 615 unique (397 families) motifs (MsigDB).

***DNA methylation model.*** A key component that affects mRNA the expression level is the presence of DNA methylation in Gene Promoters. The DNA methylation measurements (real valued) will be discretized, e.g. hypermethylation and hypomethylation states are derived by using the median DNA methylation intensity of the control-group. Conditional probability table (CPT) that defined *P(methylation)* are created by using all available samples. An illustration of the discretized table is shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sample 1 | Sample 2 | … | Sample k |
| Gene 1 | hyper | hypo |  | hypo |
| Gene 2 | hyper | Hypo |  | hyper |
| … |  |  | … |  |
| Gene i | hyper | hyper |  | Hyper |

The second step is to count the frequency of hypermethylation and hypomethylation per gene, and normalize this ratio between [0,..1] as shown below:

|  |  |  |
| --- | --- | --- |
| Un-normalized | **Hyper** | **Hypo** |
| Gene 1 | 1/3 | 2/3 |
| Gene 2 | 2/3 | 1/3 |
| … | … | … |
| Gene i | 3/3 | 0 |

***Motif model.*** Another key component that regulates mRNA expression is the binding of transcriptional regulators to their motifs in the Gene Promoters (GP). Conditional probability table (CPT) that defined *P(motif)* is determined by the ratio of the binary-valued motifs*,* where *motif* is true if motif *i* appears in the promoter region of a gene. We model the presence of motifs in the Gene Promoter region by using the 397 transcriptional family names. For each transcriptional regulator, the probability of binding, given the gene promoter is specified as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Motif 1 | Motif 2 | … | Motif k |
| Gene 1 | absent | present |  | present |
| Gene 2 | absent | absent |  | Present |
| … |  |  | … |  |
| Gene i | present | present |  | present |

The absence and presence can then easily be transformed towards a ratio between [0,..,1] per motif.

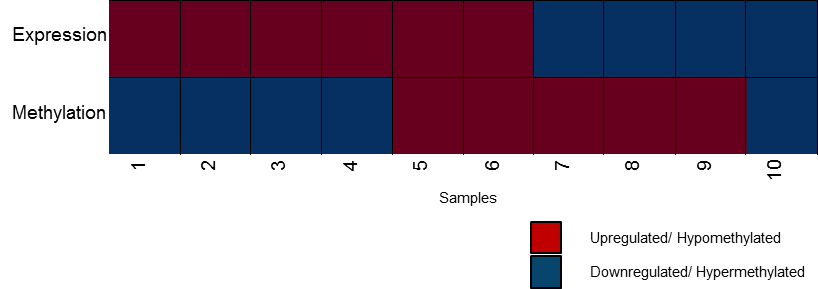
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Normalized | Motif 1 | Motif 2 | … | Motif k |
| Absent | 2/3 | 1/3 |  | 0 |
| Present | 1/3 | 2/3 |  | 1 |

This results in a higher probability of an active regulator if a motif is seen in multiple gene promoters. The Conditional probability table (CPT) per motif can now be used for each gene.

***Expression model.***Each probeset is modeled independently with the associated methylation probeset and the motif. The expression measurements (real valued) will be discretized, e.g. upregulated and downregulated states are derived by using the median expression intensity of the control-group. This results into mutually exclusive bins (under-expressed and over-expressed). An advantage of the discretizing step, instead of continues expression levels is that non-linear relationships can be detected. The CPT, that defines the happiness between the factors, *P(Expression|Methylation,Motif)* is determined in two steps:

1. The overlap between methylation-state and the expression-state (directly derived from the data).
2. Incorporation of the motif factor by the factor product: *Expressioni* ∩Methylation*i* \* *P(motif i).*

As an example, suppose we have the underneath 10 samples with their associated gene expression and DNA methylation profiles.



Step 1: From this data we can easily compute *Expressioni* ∩Methylation*i* for each of the states in the CPT (see table).

|  |  |  |  |
| --- | --- | --- | --- |
| Methylation | Motif | P(Expression=up) | P(Expression=down) |
| Hyper | Absent | 4 | 1 |
| Hyper | Present | 4 | 1 |
| Hypo | Absent | 2 | 3 |
| Hypo | Present | 2 | 3 |

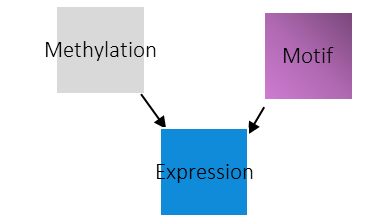
Step 2: Multiply by the motif factor and normalize between [0..1].

|  |  |  |  |
| --- | --- | --- | --- |
| Methylation | Motif | P(Expression=up) | P(Expression=down) |
| Hyper | Absent | 4 \* *P(motif k)* | 1 \* *P(motif k)* |
| Hyper | Present | 4 \* *P(motif k)* | 1 \* *P(motif k)* |
| Hypo | Absent | 2 \* *P(motif k)* | 3 \* *P(motif k)* |
| Hypo | Present | 2 \* *P(motif k)* | 3 \* *P(motif k)* |

**Model creation.** Each probeset expression is modeled with the associated DNA methylation probeset and the motif by means of a Bayesian network. The network is defined as:

*P*(Methylation,Motif,Expression) = *P*(Methylation) *P*(Motif) *P*(Expression|Methylation,Motif)

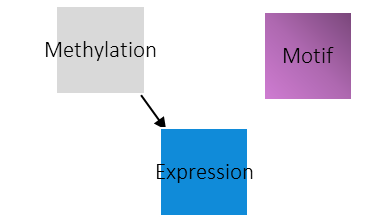
The Directed Acyclic Graph (DAG) for this model (model 1) will therefore look as follows:



To evaluate the effect of the motif in the network, thus “how much” the motif influences the expression level, we create an alternative network:

*P*(Methylation,Expression) = *P*(Methylation) *P*(Motif) *P*(Expression|Methylation)

The DAG of model 2, as shown below. The Bayes Factor between both models will indicate the effect of the motif. Thus, this approach allows elimination of motifs to finally relearn a structure DAG\*, that only indicates the expression as a function of methylation and motifs.



The Bayes Factor is computed as follows:

BF = *P*(model 1 | data) / *P*(model 2 | data) = *P*(Expression|Methylation, Motif) / *P*(Expression|Methylation)