# A Recurrent Neural Network for the Detection of Structure in Methylation Levels along Human Chromosome

Wim De Mulder  Rafel Riudavets  Martin Kuiper
wim.demulder@ugent.be

Norwegian University of Science and Technology, Trondheim, Norway

NTNU

IARIA

# Resume of the presenter

- Current position
  - Postdoc at the Eindhoven University of Technology (The Netherlands)
- Previous positions
  - Postdoc at the Norwegian University of Science and Technology (Norway)
  - Scientific researcher at KU Leuven (Belgium)
  - Scientific researcher at Ghent University (Belgium)
- Research experience
  - Machine learning
  - Bioinformatics
  - Statistics

# Outline

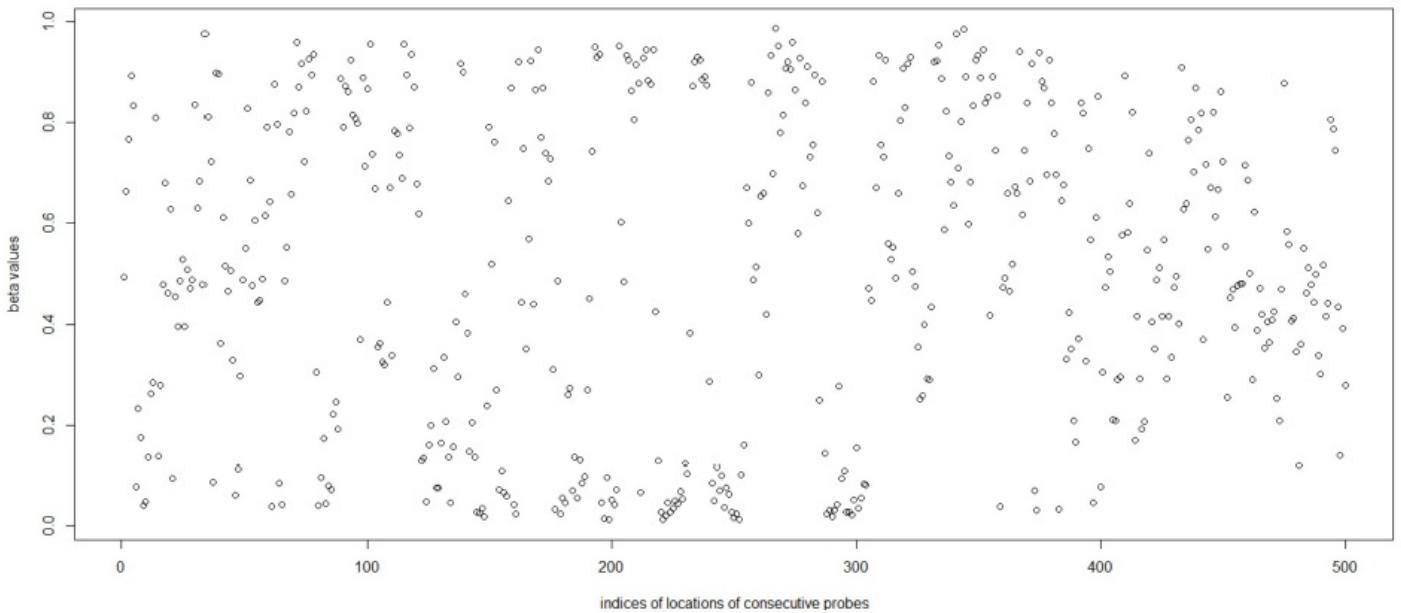
# Biological background

## Methylation

- Important form of epigenetic modification
- Regulatory mechanism to how specific genes in the genome are expressed
- Measuring methylation status
  - Microarray-based Illumina Infinium methylation assays
  - Methylation level is expressed as a value ranging from 0 to 1, called the beta value
  - Beta values can be interpreted as the percentage of methylation

# Goal of the paper

- Goal of the paper
  - Methylation levels along a given chromosome have similarities to a time series
    - Role of time is substituted by the location of the probes with respect to the DNA
  - Does sequence of beta values along a chromosome exhibit non random behavior?
    - Visual inspection suggests erratic pattern



# Method

## Recurrent neural networks (RNNs)

- Frequently used in applications with time series or sequential data
- In contrast to feedforward neural networks, RNNs have memory
- Consequently, all previous values in a sequence are used in predicting new value

## Application of RNNs for our case study

- Feed consecutive methylation values to a RNN
- Can the next value be predicted?
  - If so, conclude that structure is present in sequence of methylation values
  - If not, conclude that sequence of methylation values has random behavior

# Construction of examples

- Input vector: subsequence containing a predefined number of methylation values
  - Predefined number denoted by $w$
- Target output: subsequence shifted by one position
- Example of first 3 input-output pairs $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ with $w = 3$:

$$
\begin{aligned}
x_1 &= (a_1, a_2, a_3) \\
y_1 &= (a_2, a_3, a_4) \\
x_2 &= (a_4, a_5, a_6) \\
y_2 &= (a_5, a_6, a_7) \\
x_3 &= (a_7, a_8, a_9) \\
y_3 &= (a_8, a_9, a_{10})
\end{aligned}
$$

# Data set

- Examples are constructed from a data set collected from the Cancer Genome Atlas
  - More specifically, from a study involving Breast Invasive Carcinoma
- Contains beta values for 24 chromosomes
  - But Y chromosome not considered, because very few values
- For 1095 patients in 2 conditions (normal and tumor)
- Due to missing data, only for 96 patients data is available for normal tissue condition
- To limit computation time, we consider the 96 patients in normal condition and the first 96 patients in tumor condition
- On average, the data set contains about 17 000 measured beta values per chromosome

# Set of RNN architectures

- We consider RNNs with the following window sizes $w$:

$$w = \{10, 30, 50, 70, 90\}$$

- We try as number of hidden neurons, based on heuristic from literature

$$\begin{aligned} n_{h_1} &= \text{round}(2/3 \times 2 \times w) \\ n_{h_2} &= \text{round}(0.9 \times n_{h_1}) \\ n_{h_3} &= \text{round}(0.8 \times n_{h_1}) \\ n_{h_4} &= \text{round}(1.1 \times n_{h_1}) \\ n_{h_5} &= \text{round}(1.2 \times n_{h_1}) \\ n_{h_6} &= \text{round}(2/3 \times w) \end{aligned}$$

- In total we thus consider $5 \times 6 = 30$ different RNN architectures

# Training, validation and test sets

- For each combination of patient, condition and chromosome, we train a separate RNN
- The beta values for each triplet of the form (patient, chromosome, condition) are converted into examples as described earlier
- Each set of examples is then split into a training set, a validation set and a test set:
  - training set: first 60% of the examples
  - validation set: next 20% of the examples
  - test set: next 20% of the examples
- Use of the different sets:
  - training set: train the different RNN architectures for each (patient, chromosome, condition) triplet
  - validation set: used to select the best RNN architecture
  - test set: allows to evaluate the performance of the selected architecture

# Performance measures to evaluate RNNs

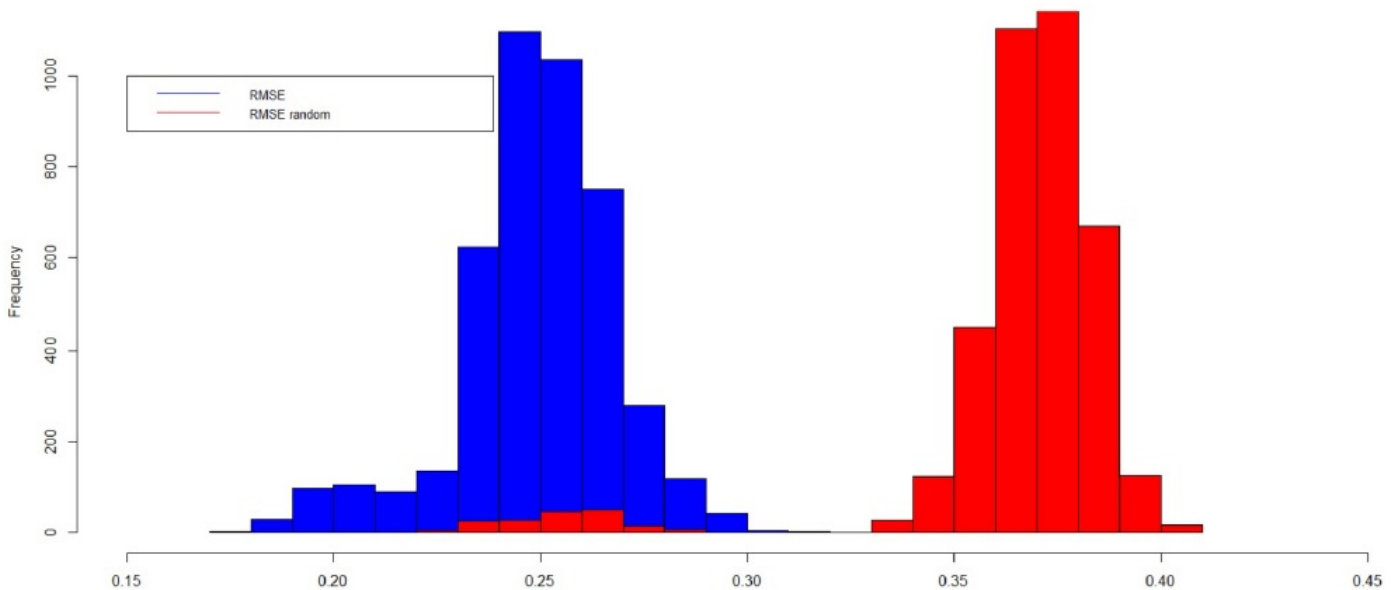
- Given test set containing examples $(x_1, y_1), \ldots, (x_m, y_m)$
- Outputs generated by the considered RNN: $\hat{y}_1, \ldots, \hat{y}_m$
- We use the notation $y_k(i)$ to refer to the $i$th component of $y_k$
- Root mean square error (RMSE):

$$RMSE = \sqrt{\sum_{k=1}^{m} \sum_{i=1}^{w} \frac{(y_k(i) - \hat{y}_k(i))^2}{mw}}$$

- Other performance measures described in the paper

# Results (1)

- Most suitable architecture was found to be
  - Window size: $w = 10$
  - Number of hidden neurons: $n_h = 7$
- Performance is evaluated with respect to a random permutation of the training sets.
- Result for RMSE:



# Results (2)

- Previous figure shows substantial difference between the results related to the permutated training sets and the non-permutated ones
- Beta values along a chromosome are non randomly distributed

# Conclusion

- Application of recurrent neural network analysis for the detection of structure in sequences of measured methylation levels along human chromosomes
- Our work demonstrates that structure is present in sequences of methylation levels
- Obtained results are relevant to both the machine learning and the biological community