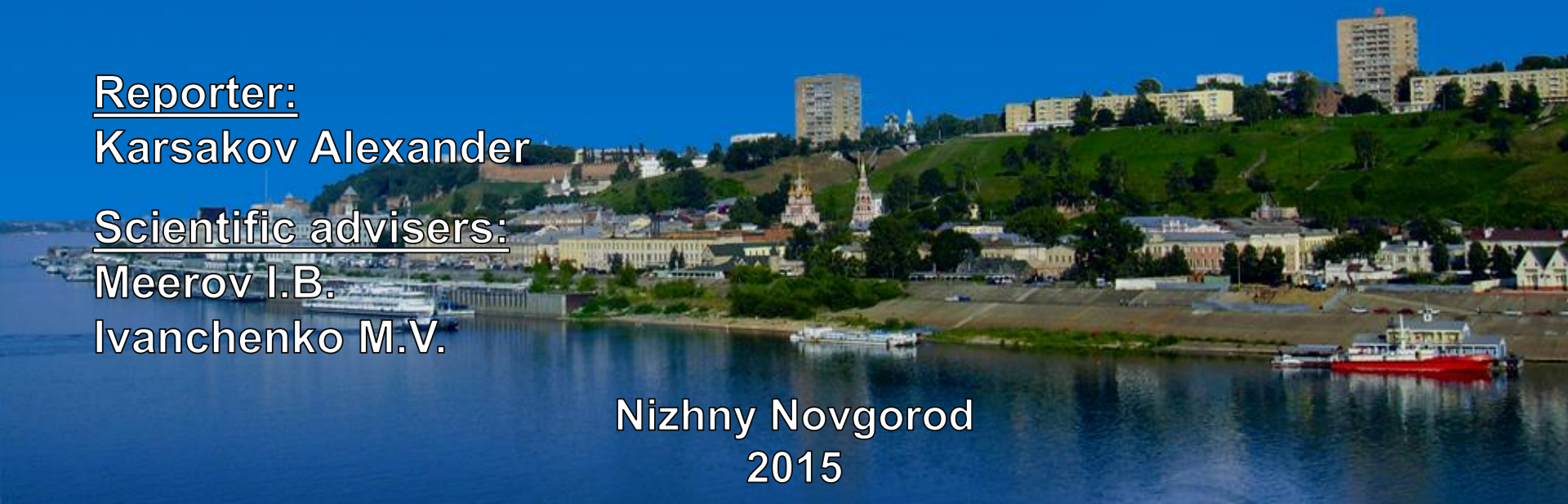
Lobachevsky State university of Nizhny Novgorod
Computational mathematics and cybernetics department

Network analysis of methylation data for cancer diagnostics



Reporter:
Karsakov Alexander

Scientific advisers:
Meerov I.B.
Ivanchenko M.V.

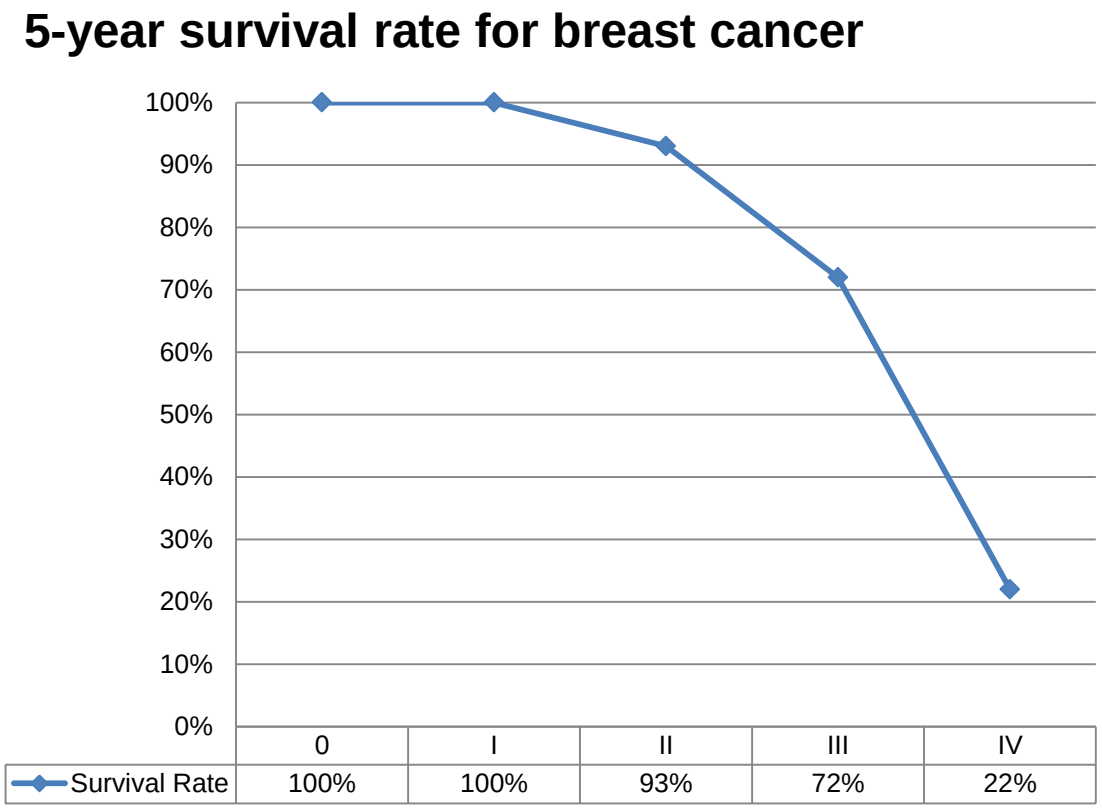
Nizhny Novgorod
2015

Agenda

- Introduction
- Methods for cancer diagnostics
- Data description
- Graph constructing procedure
- Analysis of graph topology
- Metrics for analyzing graph topology
- Experiment
- Conclusions

Introduction

- The most effective treatment is possible when cancer is diagnosed on early stages.



Methods of cancer diagnosis

- Clinical (examination, anamnesis)
- Instrumental methods (X-ray, CT, MRI)
- Laboratory methods (study of various biological materials)*

In this work we developed a method of analyzing the methylation data, based on the analysis of graph topology.

The method was firstly proposed in the paper:

Zanin M., Boccaletti S. “Complex networks analysis of obstructive nephropathy data”.

Data description

The data were obtained from an open archive National Human Genome Research Institute. It contains Methylation data, collected via the Illumina Infinium HumanMethylation450 platform, for 13 cancers.

- Data contains methylation values for >485k CpG sites (including 330k with known gene annotations, corresponding to on average 17 CpGs per gene).
- We calculated mean of methylation for each gene. Total number of genes is 15295.
- All values was normalized to [0,1] interval

Graph construction procedure

1. Split healthy subjects into two parts. One of the part is called control subjects.
2. For each cancer and healthy subject we build complete graph in which vertices corresponds to processing genes.

2.1. For each pair ***i*** and ***j*** of gene elements, we calculate mean and covariance of control subjects values:

$$\mu_{i,j} = \begin{matrix} E(X_i) \\ E(X_j) \end{matrix}, \quad S = cov(X_i, X_j),$$

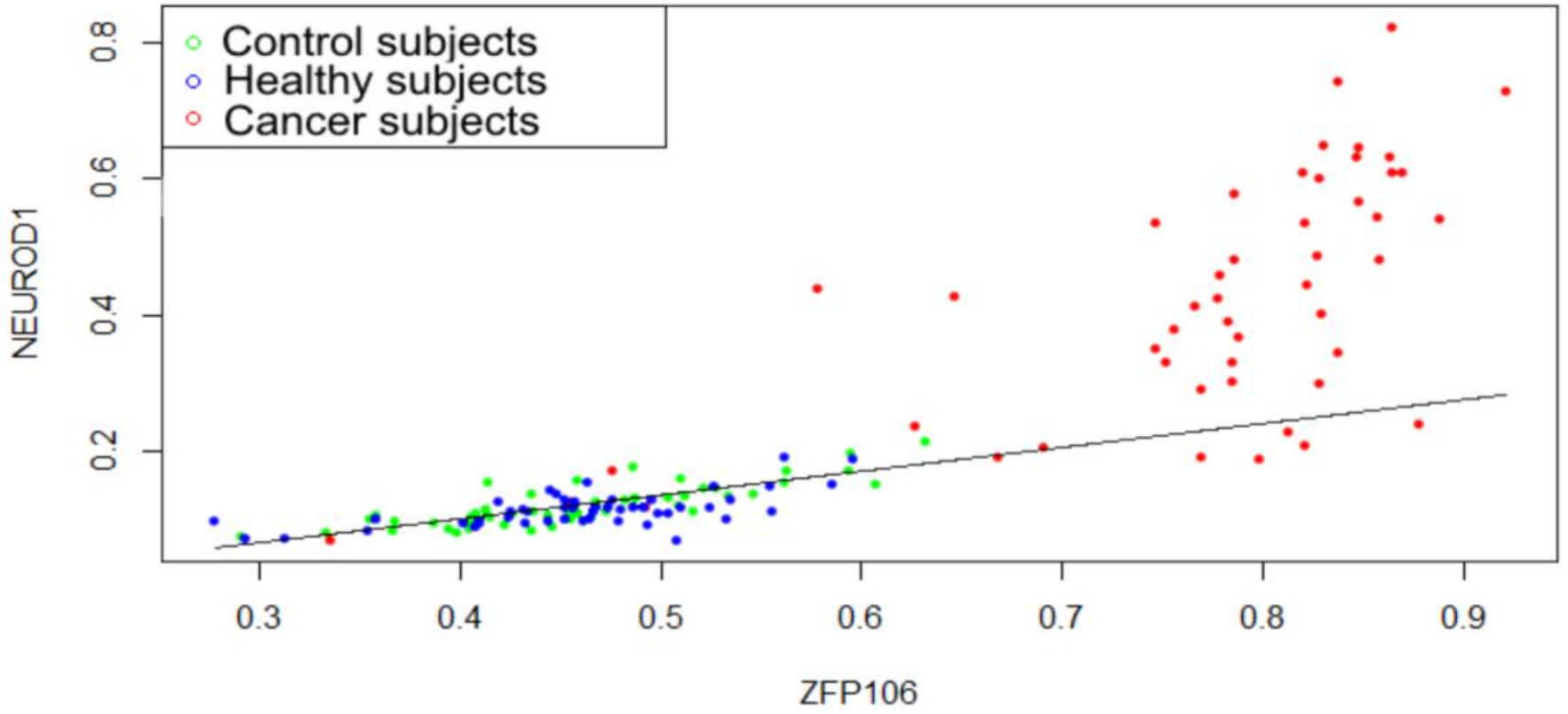
where X_i, X_j - methylation values for ***i*** and ***j*** genes respectively.

2.2. Weight of the edge connecting ***i*** and ***j*** vertices, is defined as follows:

$$w_{i,j}^{k} = \sqrt{(x_{i,j} - \mu_{i,j})^{T} S^{-1} (x_{i,j} - \mu_{i,j})}$$

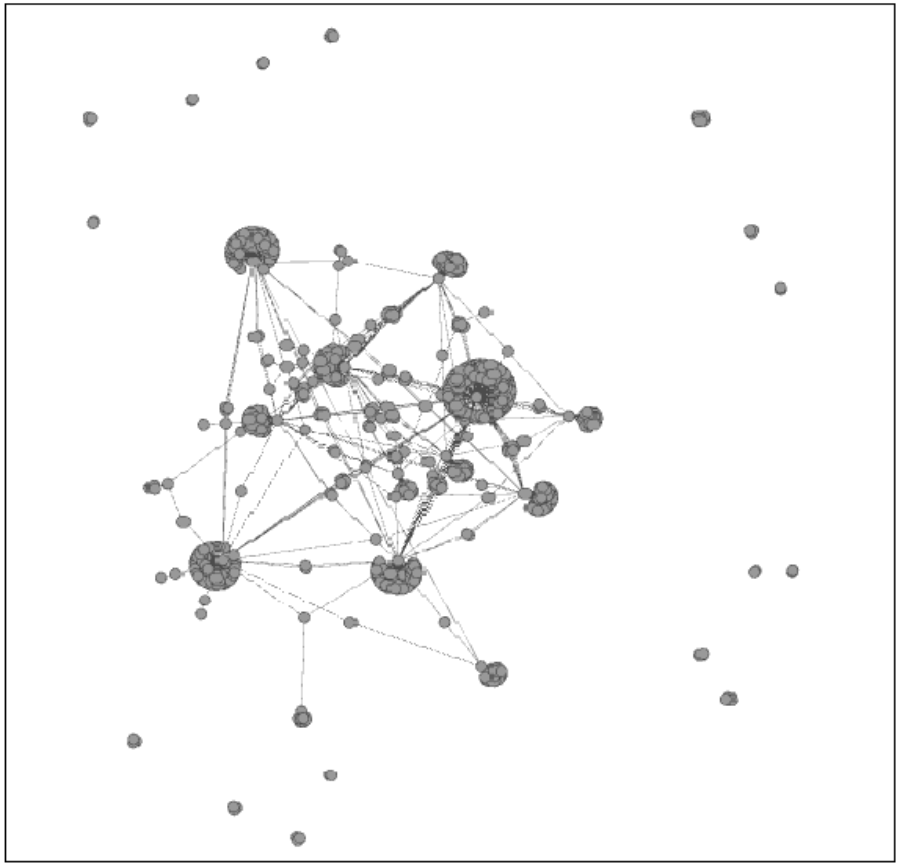
Graph construction procedure (cont.)

- Example: calculating weights for edge between NEUROD1 and ZFP106 genes.

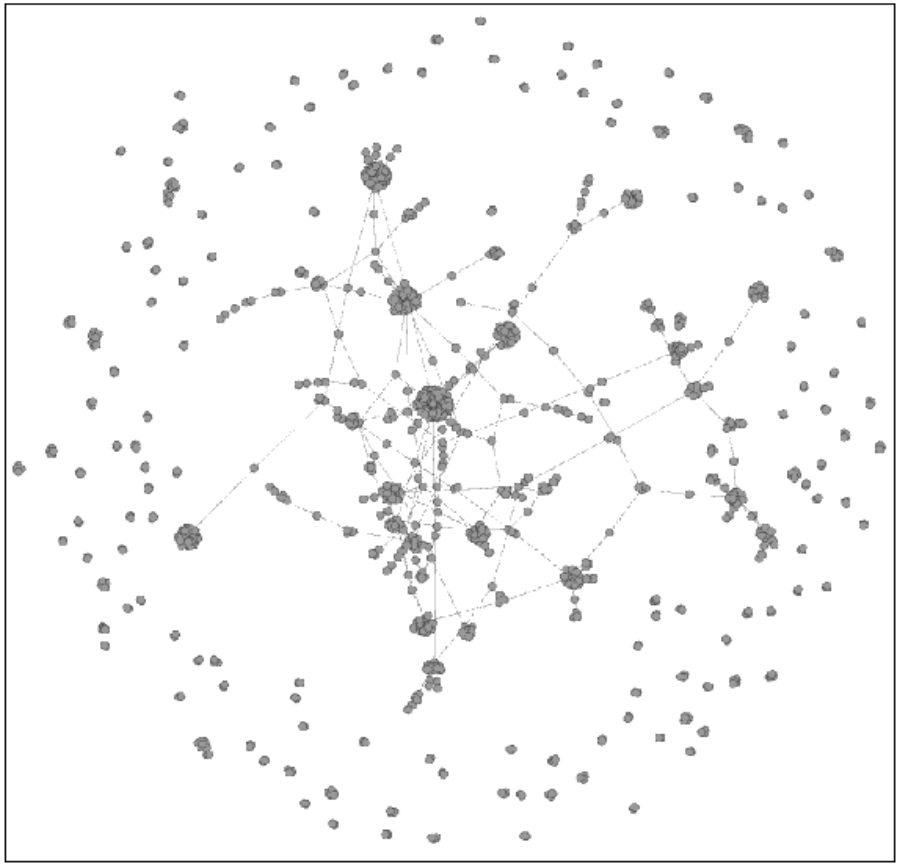


Analysis of graph topology

Cancer subject



Healthy subject



Metrics for analyzing graph topology

- Mean, variation and max value of edge weight
- Mean, variation and max value of vertex degree
- Mean, variation of shortest path weights
- Diameter of graph
- Degree centralization
- Efficiency
- Betweenness centralization

Metrics for analyzing graph topology (cont.)

- Node degree for weighted graph:

$$\deg(v) = 2 * \sum_{e \in E} w(e)$$

- Holds for analogous of Handshake lemma

Metrics for analyzing graph topology (cont.)

- Degree centrality of node - ratio of node degree to overall number of nodes:

$$C_D(v) = \frac{\deg(v)}{|V|}$$

- Degree centralization - measure of how central its most central node is in relation to how central all the other nodes are:

$$C_D(G) = \frac{\sum_{j=1}^{|V|} |C_D(v^*) - C_D(v_i)|}{H} = \frac{\sum_{j=1}^{|V|} |C_D(v^*) - C_D(v_i)|}{(|V|-1)*(|V|-2)},$$

Metrics for analyzing graph topology (cont.)

- Graph *efficiency* defined graph as follows:

$$E_C(G) = \frac{\sum_{y \neq x} \frac{1}{d(x,y)}}{|V| * (|V| - 1)}$$

Metrics for analyzing graph topology (cont.)

- Betweenness centrality of node - number of times a node acts as a bridge along the shortest path between two other nodes:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Computing graph metrics

To compute all specified metrics we need to:

- Compute degrees for all vertices
- Compute and traverse shortest paths between all pairs of vertices

- We developed a tool for graph processing (building and metrics calculation), which was parallelized using TBB library.

Experiment

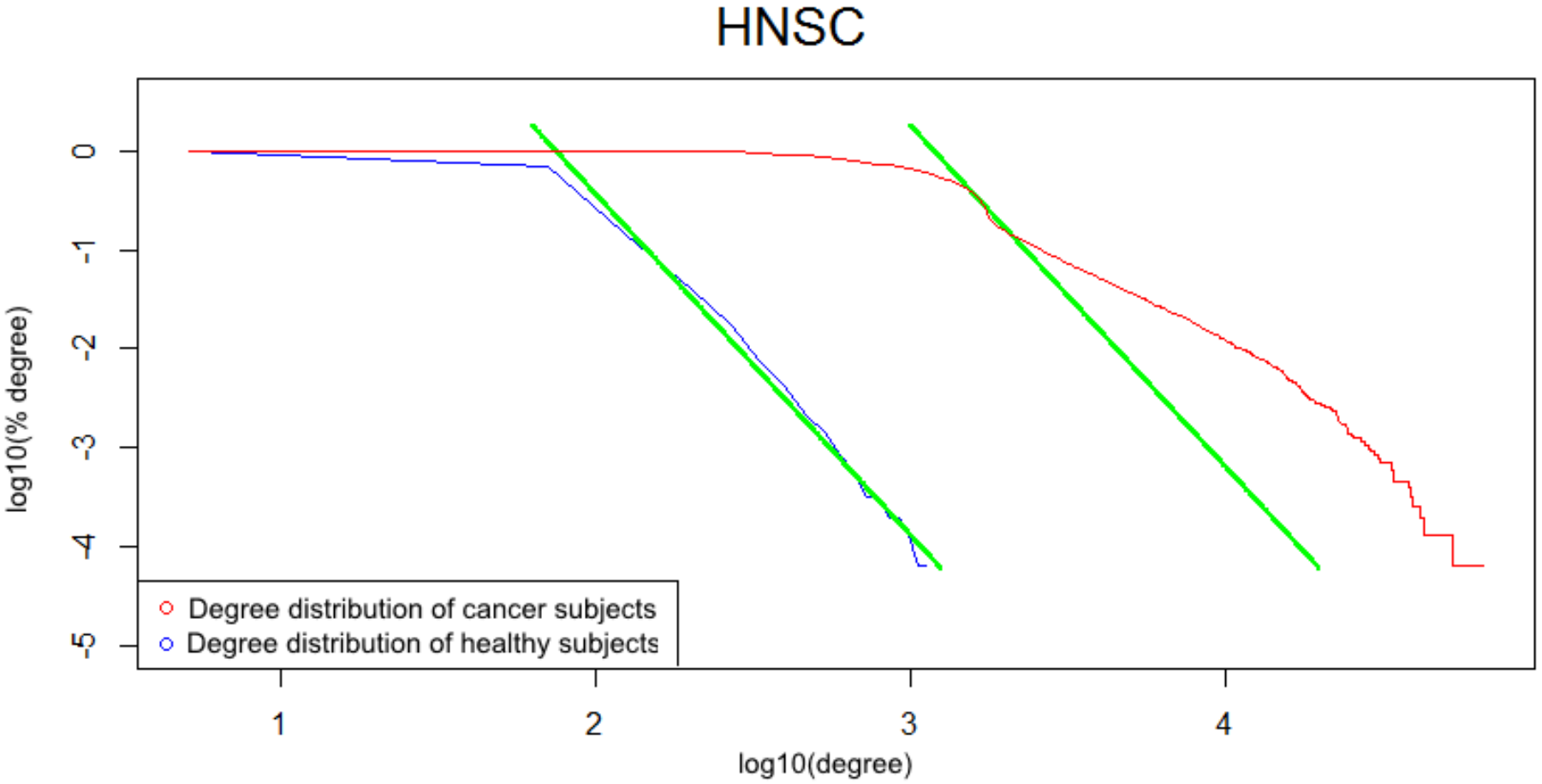
- For all patients (healthy and cancer) we constructed graphs using algorithm described previously.
- For each graph calculated topology metrics
- Perform binary classification: cancer vs. healthy.

Results

Cancer type	Classification rate using topology metrics	Classification rate using methylation of all genes
BLCA	100%	96.6%
BRCA	91.84%	98.343%
COAD	94.35%	99.623%
HNSC	100%	98.246%
KIRC	98.5%	99.655%
KIRP	100%	98.615%
LIHC	93.26%	96.531%
LUAD	99.43%	99.704%
PRAD	81.15%	99.659%
THCA	83.96%	94.464%
UCES	100%	97.537%

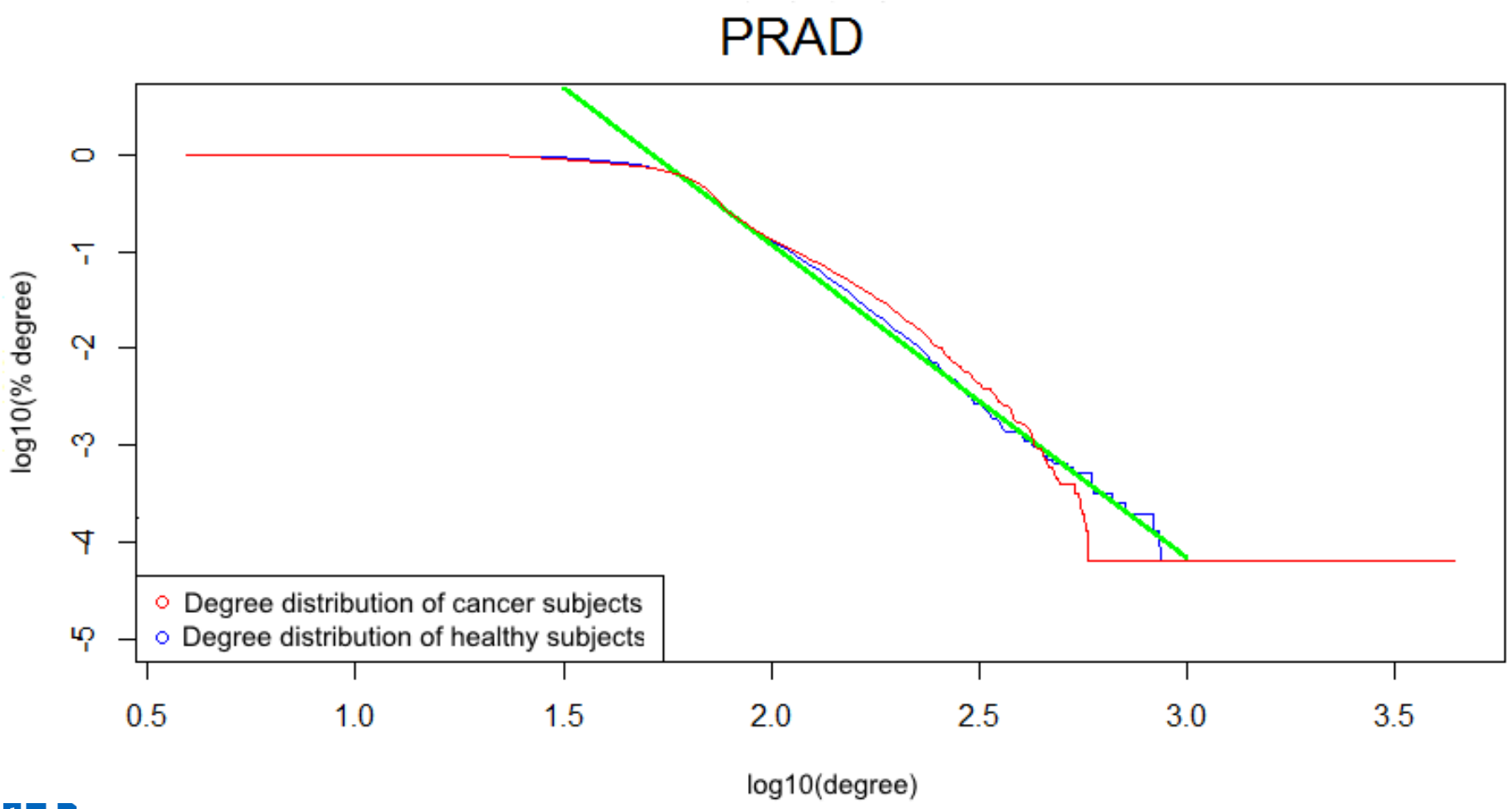
Further investigations

- Degree distribution for HNSC cancer



Further investigations

- Degree distribution for PRAD cancer



Conclusions

- The proposed modeling method is adequate. Using topology indices we can distinguish healthy and cancer tumors.
- Explanation based on degree distribution was given why we got insufficient result for two cancers.
- Also we can use topology metrics to determine which genes are responsible for cancer development (in progress).

Thanks for your attention

???