Skoltech

Skolkovo Institute of Science and Technology

Assembling CRISPR arrays from PCR data

I. Oseledets

Plan

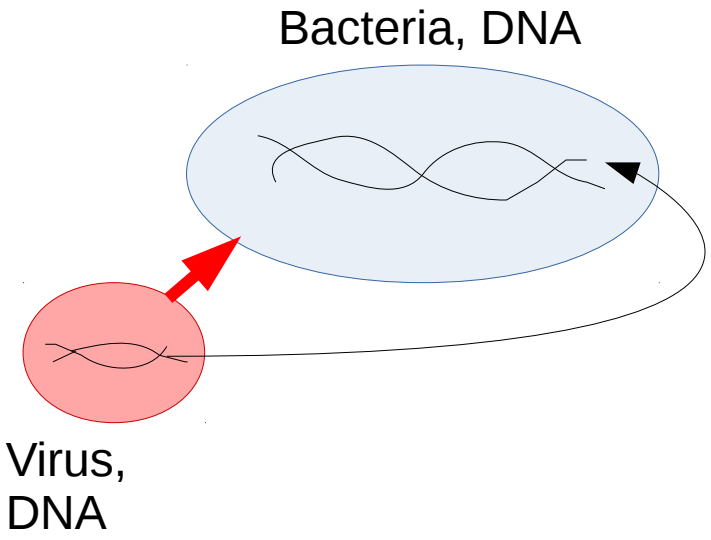
Skoltech
Skolkovo Institute of Science and Technology

- Intro to the problem
  - Setup
  - Data and Pipelines
- Naive approaches
  - Thresholding by occurrences
  - Greedy approach
- ML approaches
  - PCR models
  - U-net
  - Standard classification
  - GNN
- Results
- In development
- Conclusions & Plans

Intro

Skoltech
Skolkovo Institute of Science and Technology

- Bacteria possess CRISPR-Cas immunity systems
- CRISPR arrays consistis of short spacers separated by identical repeats and cas genes.
- Bacteria can fight against phages if contain CRISPR spacer complementary to phage genome.

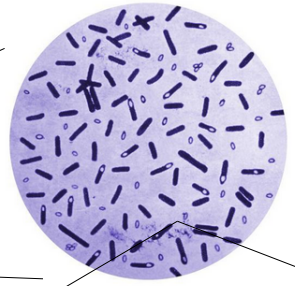


- Our goal is to reconstitute CRISPR arrays in complex bacterial populations
- So we will be able predict which phages will be able to escape CRISPR immunity and, thus, represent a promising tool for phage therapy.

Setup

Skoltech

Skolkovo Institute of Science and Technology



Bacteria DNA

CRISPR arrays (2 to 10 in each bacteria)

CRISPR cassette

Spacers (33 bp in *E.coli*) can be acquired from genomic DNA of phages or plasmids

Repeats (28 bp *in E.coli*)

Full Pipeline

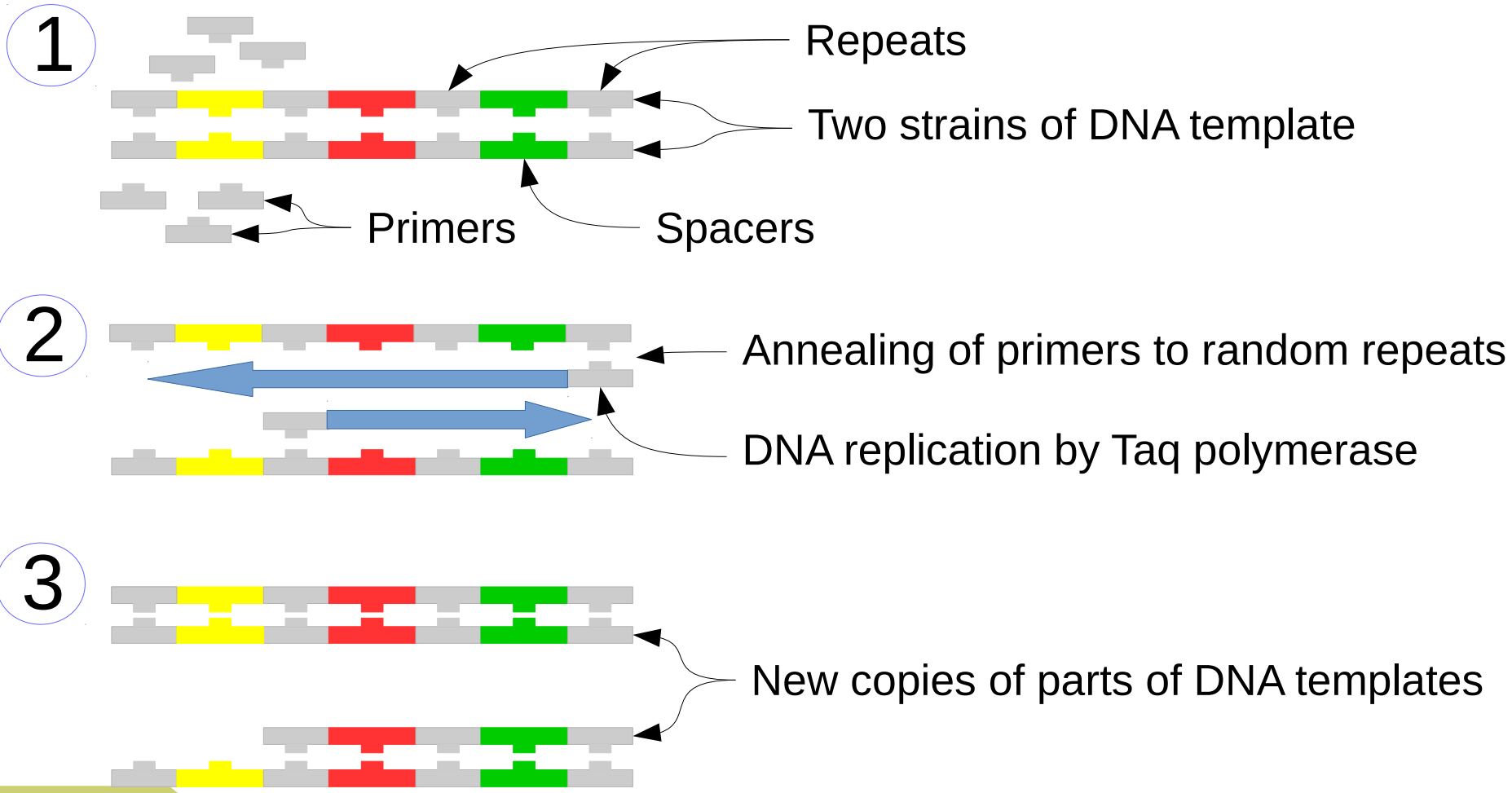
Skoltech
Skolkovo Institute of Science and Technology

1. Collecting DNA templates
2. PCR amplification of CRISPR arrays with primers complementary to repeats
3. High throughput sequencing
4. Preprocessing of obtained data
5. Restoration of CRISPR arrays

PCR

Skoltech
Skolkovo Institute of Science and Technology

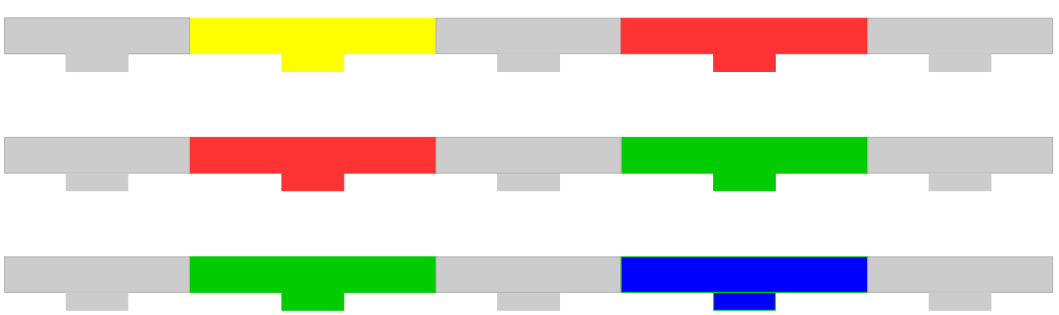
PCR is a process of amplification of a several copies of DNA to millions of copies.



Data

Skoltech
Skolkovo Institute of Science and Technology

Sequencing procedure returns reads which contain no more than two spacers.



Reads

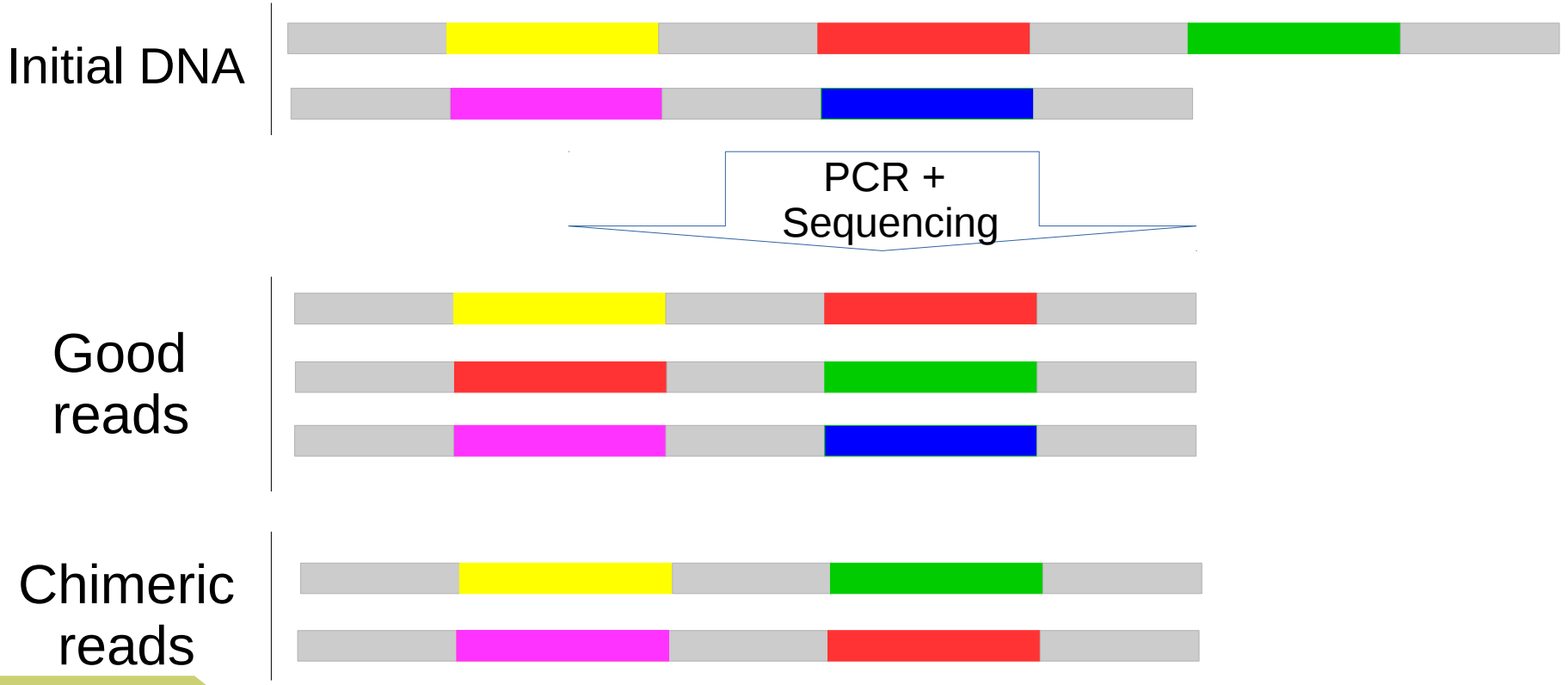
Our job

Restored array

Difficulties

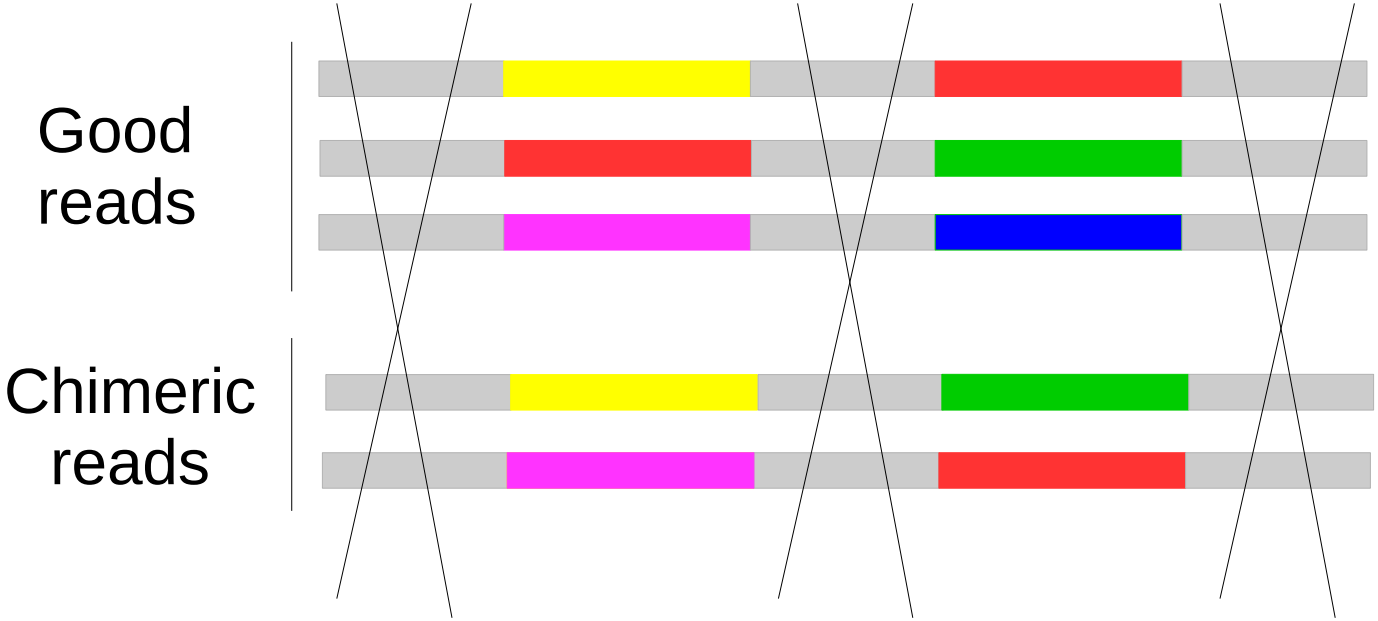
Skoltech
Skolkovo Institute of Science and Technology

Main difficulty is chimeric reads – reads where spacers from different parts of array or different array occur together.



Preprocessing

Skoltech
Skolkovo Institute of Science and Technology



Preprocessing steps:
1. Split by repeats
2. Cluster similar pairs of spacers together with hierarchical clustering procedure and count amount of pairs in each cluster

Finally

Skoltech
Skolkovo Institute of Science and Technology

Counts

Good Pairs

N_1

N_2

Chimeric pairs

N_k

Goal:

- Classify all pairs as good\chimeric
- Restore arrays from correct pairs

Real data

Skoltech
Skolkovo Institute of Science and Technology

At the moment:
8 datasets for two different bacterias

C. Difficile

3 datasets:

- 1x 1 bacteria strain*, known answer
- 2x 2 bacteria strains, known answer

- 70k to 200k pairs per dataset
- 1% of Chimeric pairs
- ~1% probability of mistake in each letter

E. Coli

5 datasets:

- 2x 1 bacteria strains, known answer
- 2x 2 bacteria strains, known answer
- 1x lots of strains and no answer

- 70k to 200k pairs per dataset
- 15% of Chimeric pairs
- ~1% probability of mistake in each letter

*Strain – bacteria type, with specific CRISPR arrays

Standard assembling methods

Skoltech
Skolkovo Institute of Science and Technology

We checked one of standard genome assemblers - SPAdes.

It does not return original arrays.

Probable reasons:

- short reads
- redundancy of CRISPR repeats
- chimeric reads

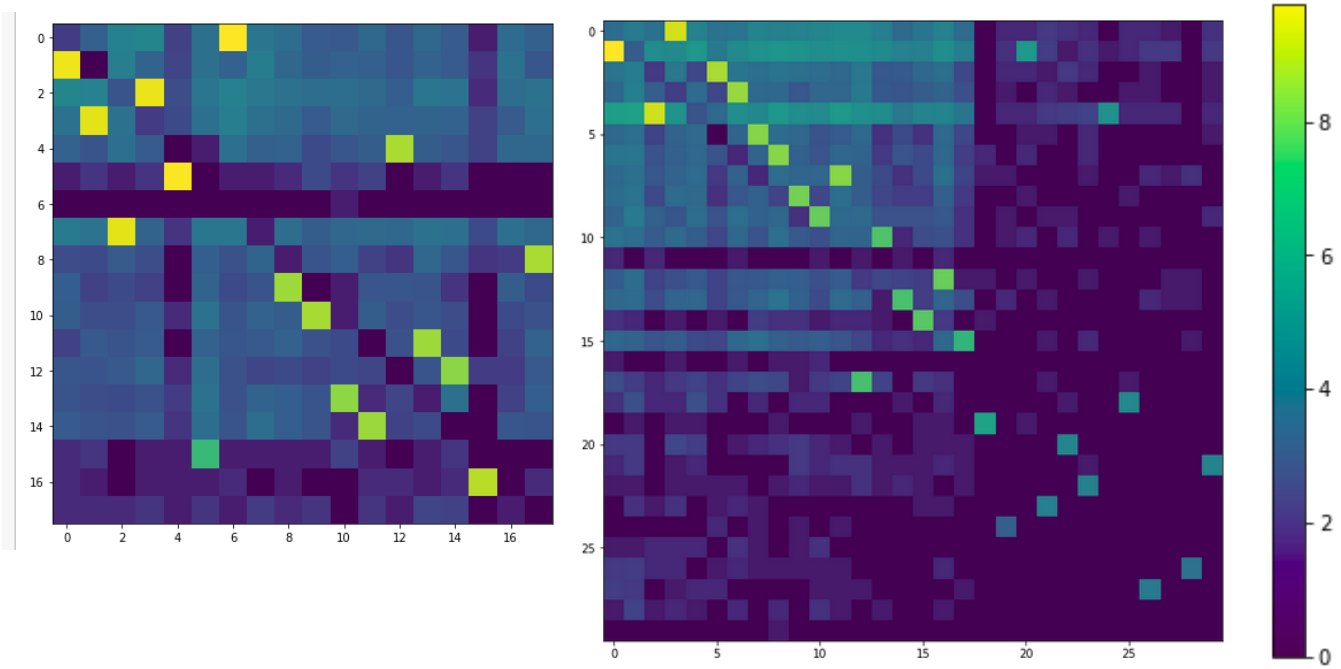
Graph representation I

Skoltech
Skolkovo Institute of Science and Technology

Graph representation:
- Spacers in nodes
- Edge weight = pair count

Logarithm of Adjacency matrix for single strain E.coli and two strains of E.coli

Yellow pixels probably represent correct pairs, light blue represents chimeric reads



E.Coli, single strain(left) and two strains(right)

Naive approach

Skoltech
Skolkovo Institute of Science and Technology

Threshold pairs by counts.

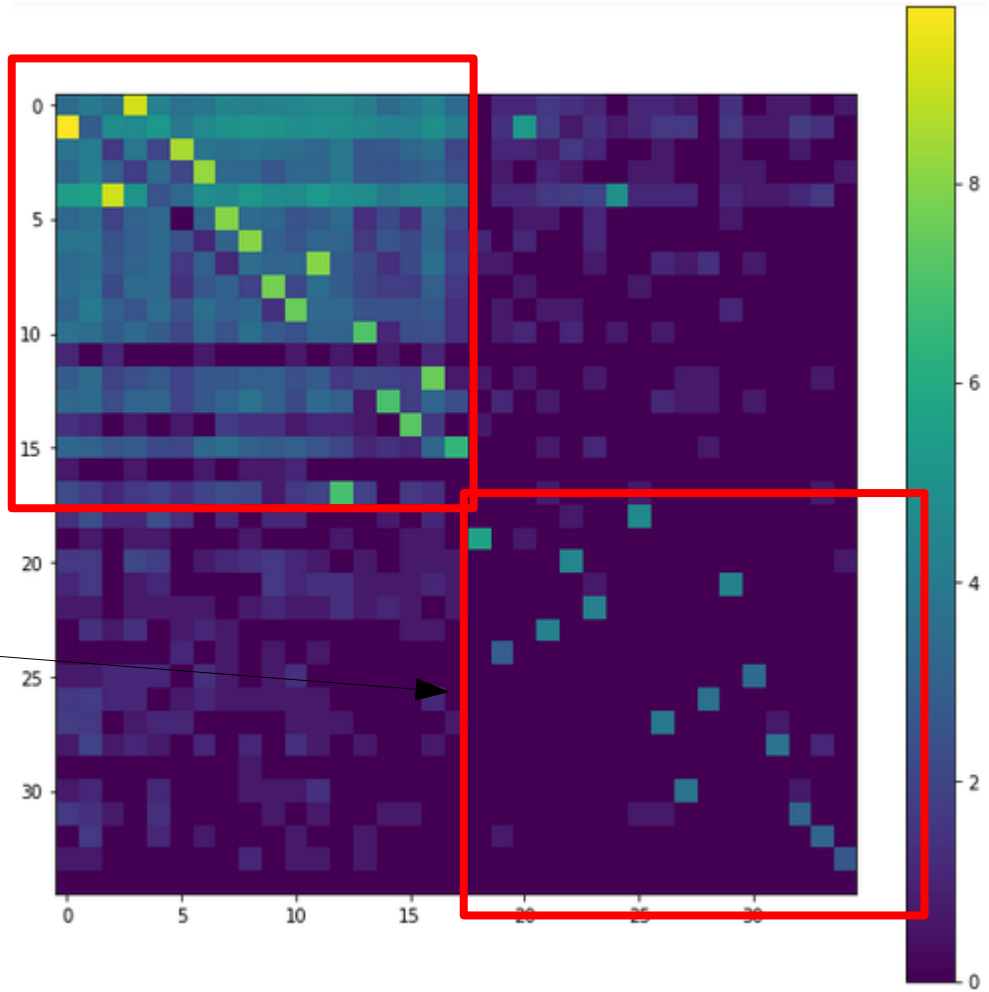
+:
- Easy to implement
- Works for single strain\balanced strains

-:
- Does not work for unbalanced strains
- Setting threshold is unclear

Two unbalanced strains.

Chimeric reads in better presented strain have larger counts than true reads in less presented strain

E.Coli example



Greedy Approach

Skoltech
Skolkovo Institute of Science and Technology

Mark pairs as correct starting from most frequent to less frequent with restrictions:

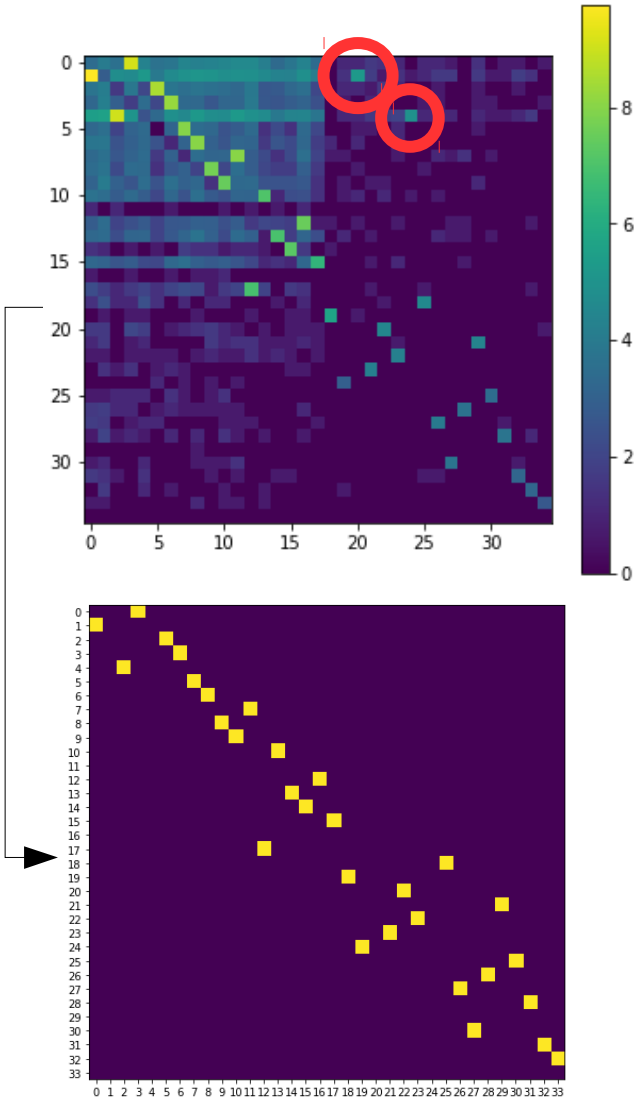
- No cycles
- No undirected cycles
- No tree structure

+:

- Easy to implement
- Works for unbalanced strains

-:

- Misses tree structures if they exist
- Setting threshold is unclear



ML approach

Skoltech
Skolkovo Institute of Science and Technology

To further improve our method we have to tune a lot of thresholds, for different types of pairs.

This task actually looks like ML problem

Problem:
Not enough data.

Stochastic PCR model

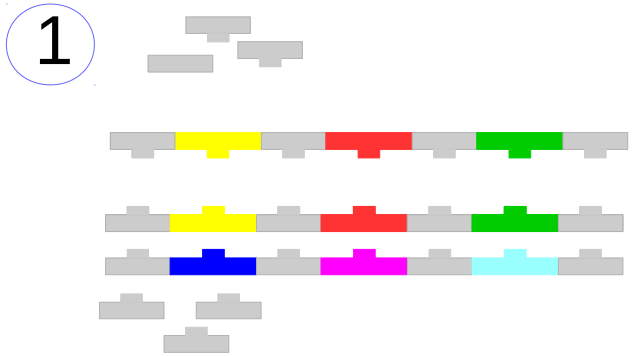
Skoltech
Skolkovo Institute of Science and Technology

We developed 3 different stochastic PCR models

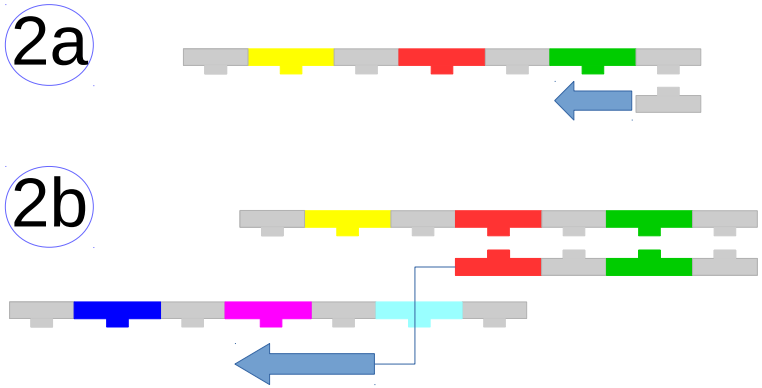
	CRISPR Array representation	Initial DNA population	Source of errors	Other features
Naive	List of integers. Each integer repr. spacer	10k random initial arrays sampled from CRISPR E.Coli database	Random switching from one array to another during replication	nothing
Different repeats model	Array of integers. Integers repr. spacers and repeats		Repeats-to-repeats annealing.	Annealing between different repeat-repeat and repeat-pimer combinations is different
Fast no repeats model	Array of integers. Each integer repr. spacer		Repeats-to-repeats annealing.	Fast model. Allows to set initial DNA to primers ratio, as in real experiments

Chimeric reads

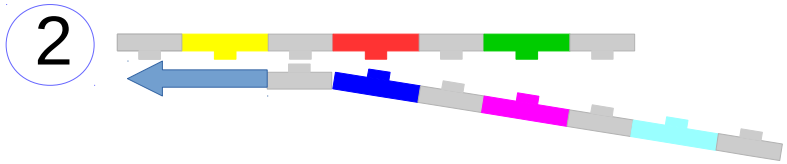
Skoltech
Skolkovo Institute of Science and Technology

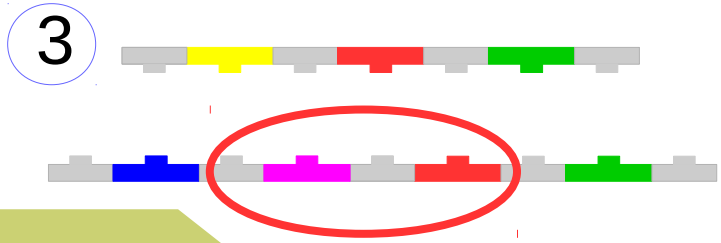


Source of errors for Naive model

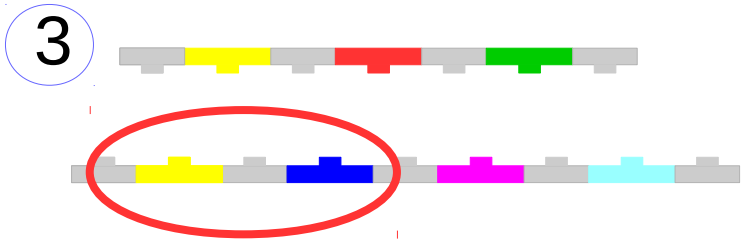


Source of errors for Fast and Different repeats models





Chimeric read



ML approach

Skoltech
Skolkovo Institute of Science and Technology

For all ML methods we currently use ***Different repeats*** simulation model and the following validation scheme:

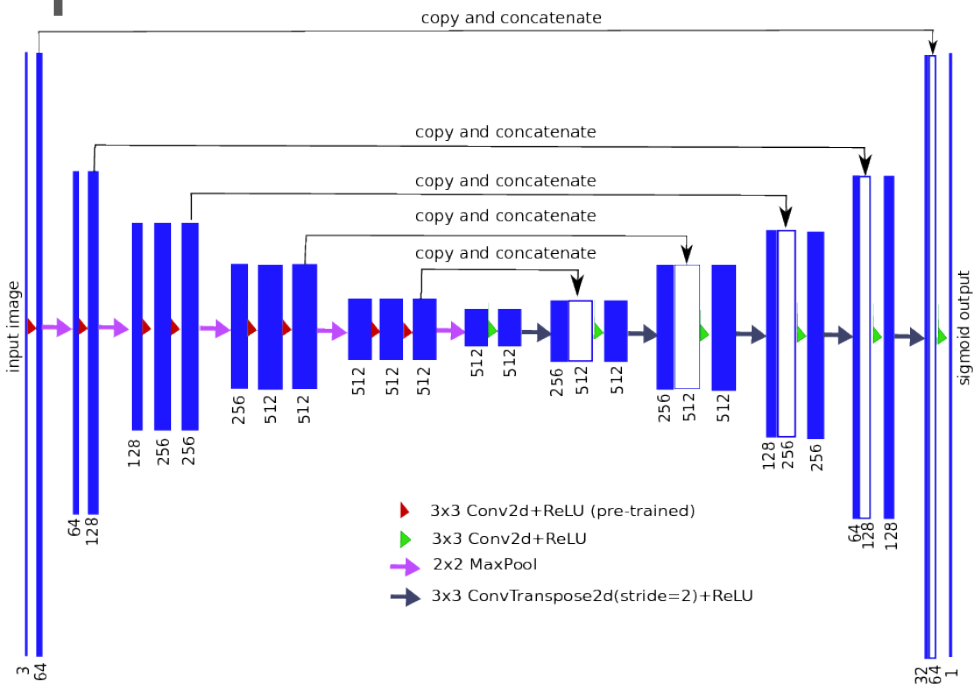
- Train set: 800 simulations
- Test set 1: 200 simulations
- Test set 2: real data

How not to make deeplearning on graphs

Skoltech

Skolkovo Institute of Science and Technology

- Train U-net on normalized adjacency matrix to predict existing edges
- Test on part of synthetic data
- Apply to real data



Two variations of input for U-net:

U-net-D:

$$Input: D_i^{-\frac{1}{2}} A D_i^{-\frac{1}{2}}, D_i^{-1} A, A D_i^{-1}, D_o^{-\frac{1}{2}} A D_o^{-\frac{1}{2}}, D_o^{-1} A, A D_o^{-1}$$

U-net-M:

$$Input: D_{mi}^{-1} A, A D_{mo}^{-1}$$

Where D_i, D_o are diagonal matrices with total incoming and outgoing weights, and D_{mi}, D_{mo} are diagonal matrices with maximum incoming and outgoing weights.

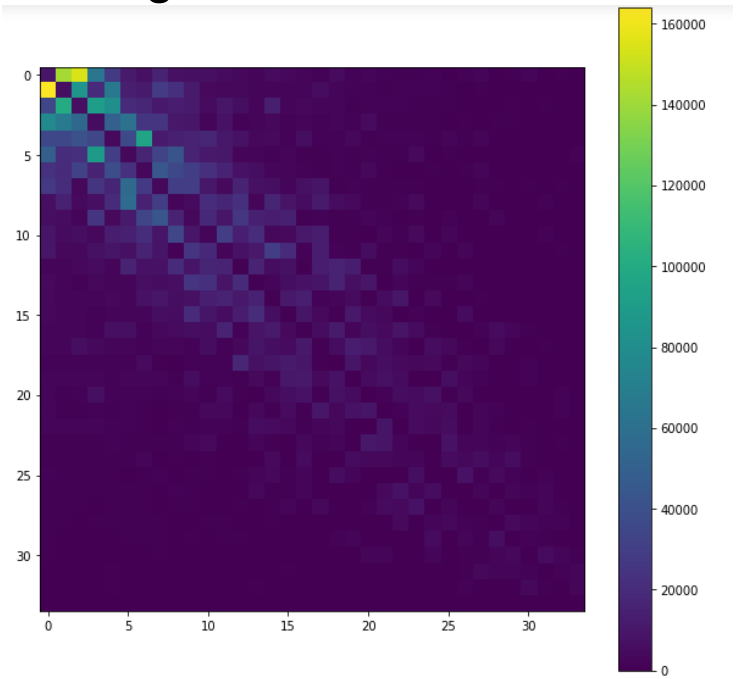
Preparing adjacency matrix

Skoltech
Skolkovo Institute of Science and Technology

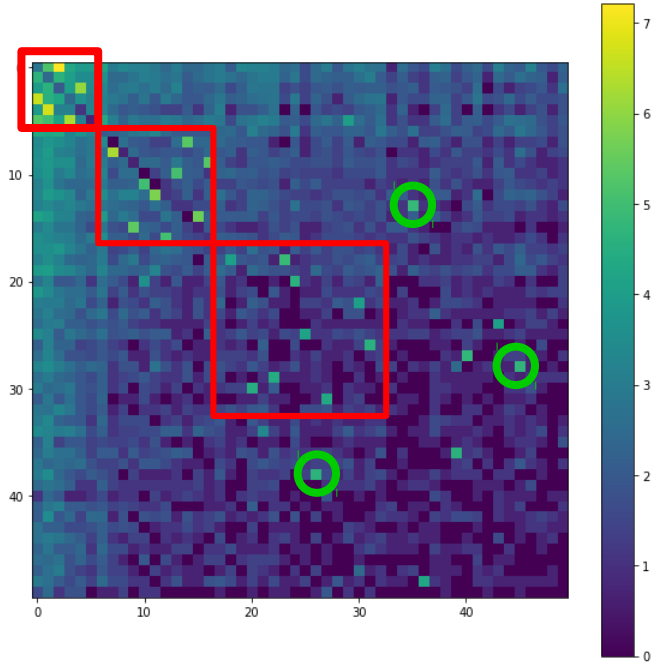
- Rearrange vertexes in the graph according to:

$$max\left(incoming\,weights, outgoing\,weights\right)$$

- Spacers from same arrays are grouped together
- Weight in adjacency matrix are grouped around main diagonal
- Edges between arrays are separated from background noise aside from main diagonal



Sum of all adjacency matrices of generated data



Part of adjacency matrix for Elephant

U-net results

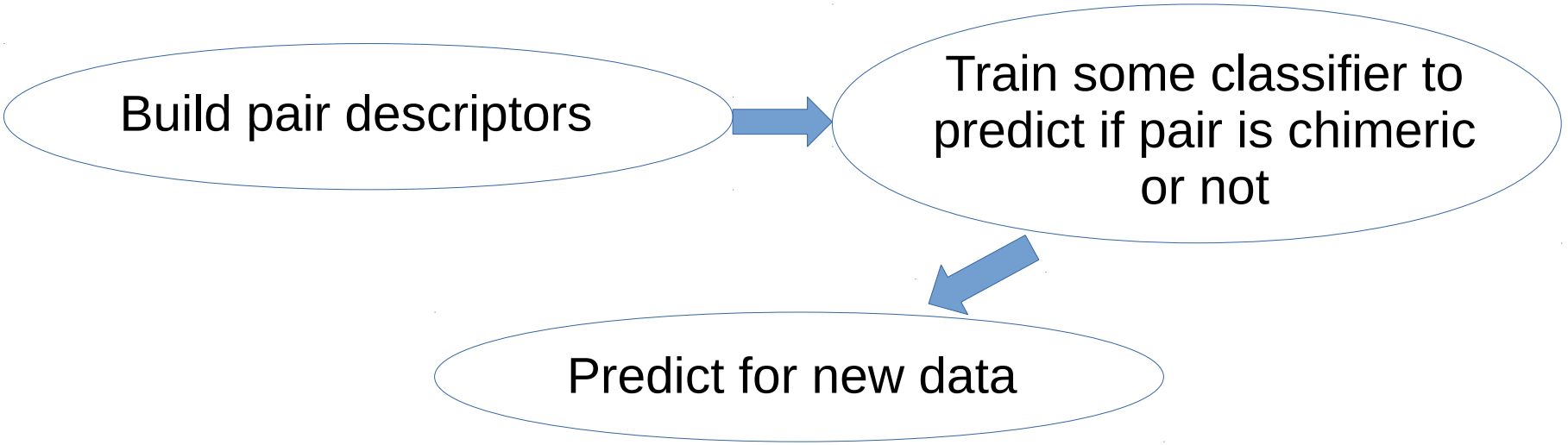
Skoltech
Skolkovo Institute of Science and Technology

U-Net shows good performance on synthetic set:
ROC AUC : 0.986
However it fails on real data with ROC AUC: 0.8

Also U-net is not based on the nature of the data

Pair descriptors

Skoltech
Skolkovo Institute of Science and Technology



For each pair (i, j) we build the following descriptor:

$$d = [\frac{w_{ij}}{\sum_i w_{ij}}, \frac{w_{ij}}{\sum_i w_{ji}}, \frac{w_{ij}}{\sum_i w_{ij}}, \frac{w_{ij}}{\sum_j w_{jj}}]$$

where w_{ij} is a count for pair (i, j)

Classifiers

Skoltech
Skolkovo Institute of Science and Technology

We split synthetic data in 800 training samples and 200 test samples and train several models.

Models are validated on synthetic test set and real data.

We compare the following standard algorithms:

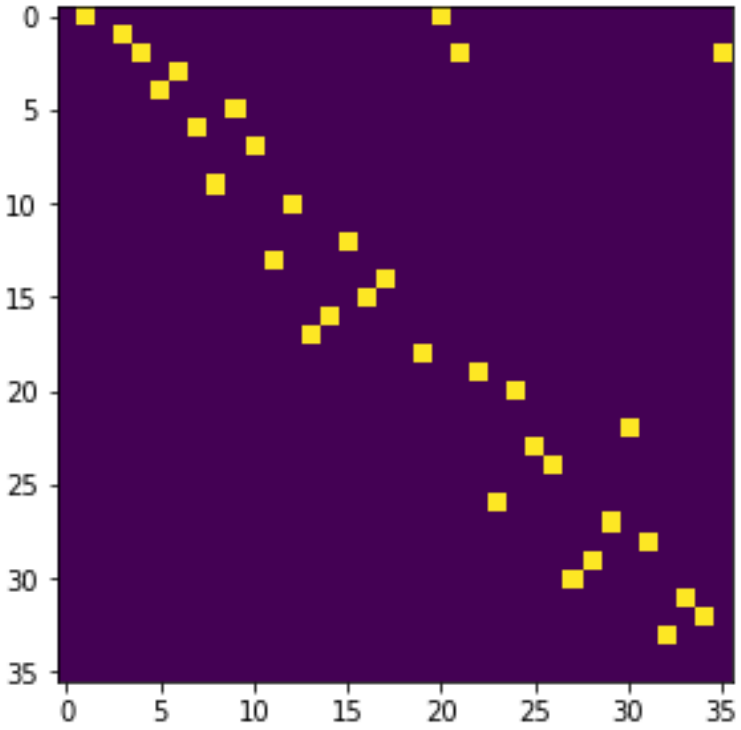
- Logistic Regression
- Kernel SVM
- Gradient Boosting Classifier

Classifiers results

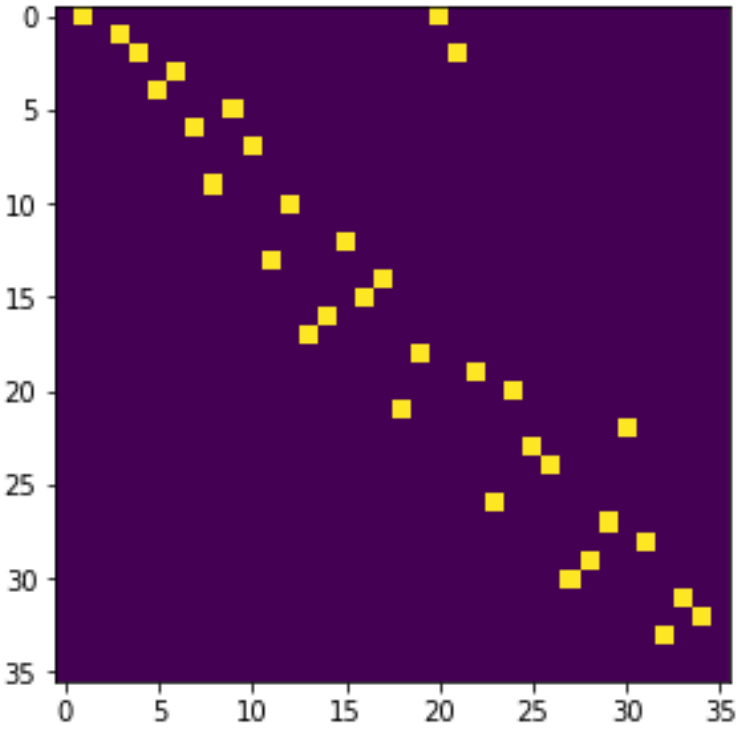
Skoltech
Skolkovo Institute of Science and Technology

Gradient Boosting classifier seems almost perfect.

On real data it makes only 1 mistake, classifying 1 edge as False Positive.



GB prediction



True edges

Graph representation II

Skoltech
Skolkovo Institute of Science and Technology

Our data has graph nature.
Our task is to classify existing edges in graphs.
Nodes classification investigated much better

This setup allows to classify nodes, thus we rebuild our graph representation in the following manner:

- Nodes represent pairs
- Edges are directed and exist between pairs with common spacer, like:

$$(A, B) \rightarrow (B, C)$$

- Thus we have obtain new graph with nodes descriptors

Gated GNN

Skoltech
Skolkovo Institute of Science and Technology

Gated GNN is model of neural message passing, where embeddings in vertexes are processed by GRU’s.

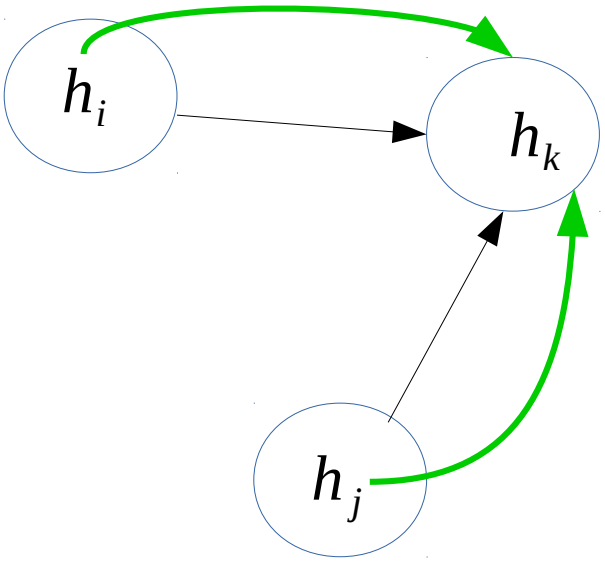
Gated GNN allows to classify nodes

Gated GNN

Skoltech
Skolkovo Institute of Science and Technology

Model works in two steps:

1. Propagation



$$h_v^{(0)} = [d_v^T, 0]^T$$

$$z_v^{(t)} = f_{nn}(h_v^{(t)})$$

$$x_v^{(t)} = A_v[z_1^{(t)T} \dots z_n^{(t)T}]$$

$$h_v^{(t+1)} = f_{gru}(x_v^{(t)})$$

1. Prediction

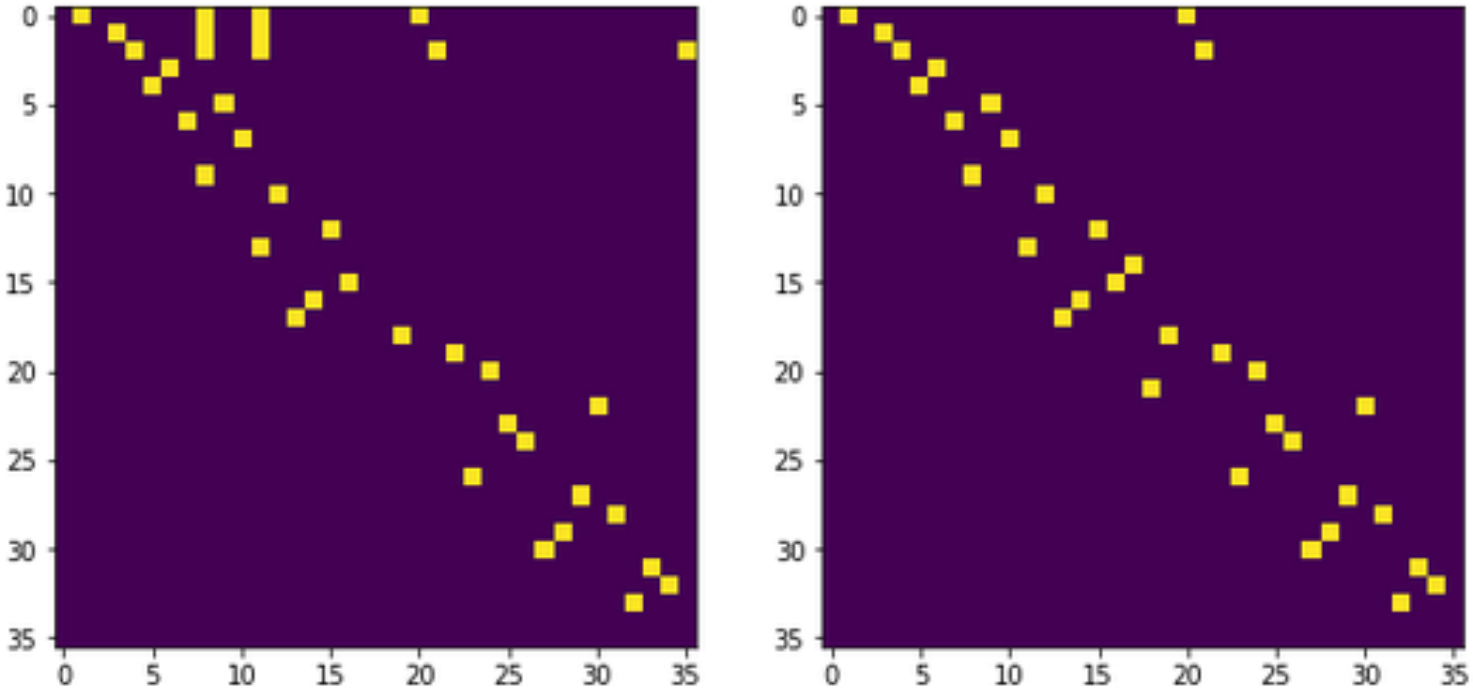
$$y_v = f_{pred}(h_v^{(T)})$$

Classifiers results

Skoltech
Skolkovo Institute of Science and Technology

Synthetic test set: ROC AUC 0.999

Real data: ROC AUC 0.97, tendency to False Positives



Real data example. *E.coli* with two strains.
Left – GNN predictions, *Right* – ground truth

Results

Skoltech
Skolkovo Institute of Science and Technology

	No ML		U-net	Edge Descriptors			GNN
	Naive	Greedy	U-net	LogReg	SVM	GB	GNN
Synthetic Data	0.83	0.973	0.986	0.999	0.995	0.999	0.999
E.Coli	0.90	0.968	0.8	0.983	0.960	0.999	0.968

Results

Skoltech
Skolkovo Institute of Science and Technology

- We are pretty good in classifying pairs

- We can not fully restore arrays:
  These two arrays:

  $a \to b \to c \to d$
  $a \to b \to e \to f$

  does not differ from these two

  $a \to b \to c \to d$
  $\quad\quad b \to e \to f$

  We restore second combination in both situations

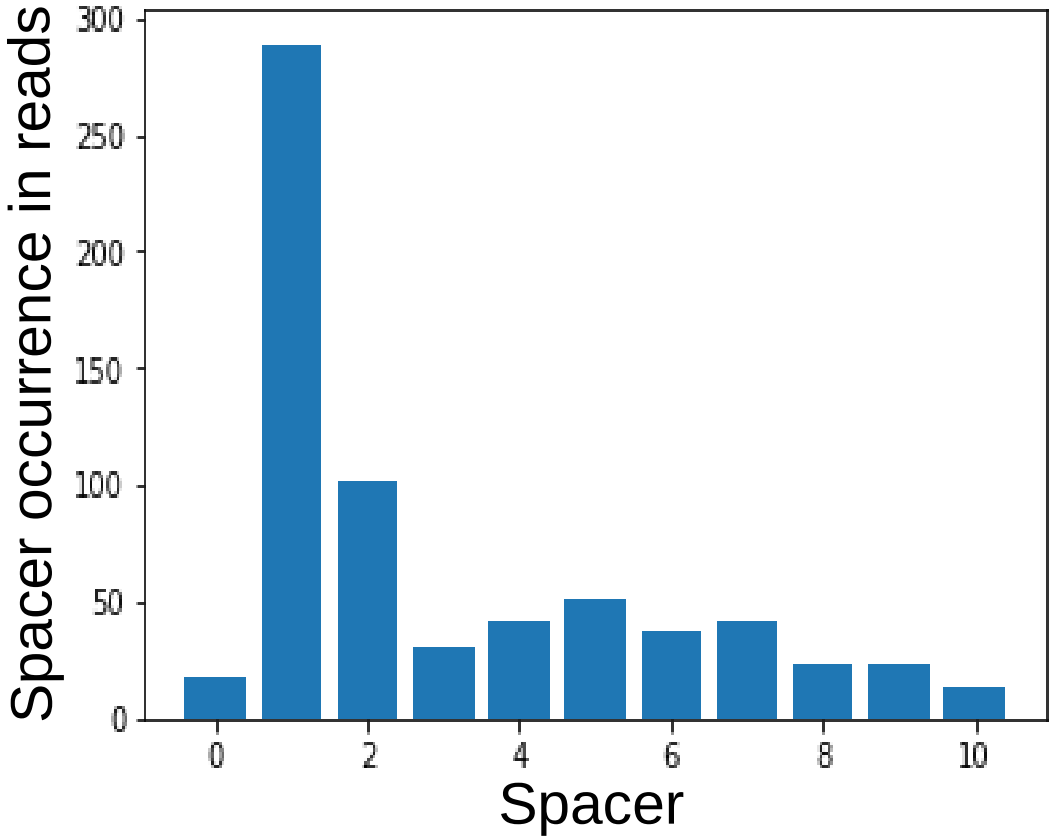
In development

Skoltech
Skolkovo Institute of Science and Technology

Spacers in one array are not equally amplified.

That is why we can not group equally amplified spacers to highlight arrays.

This prevents us from correct array restoration from pairs.



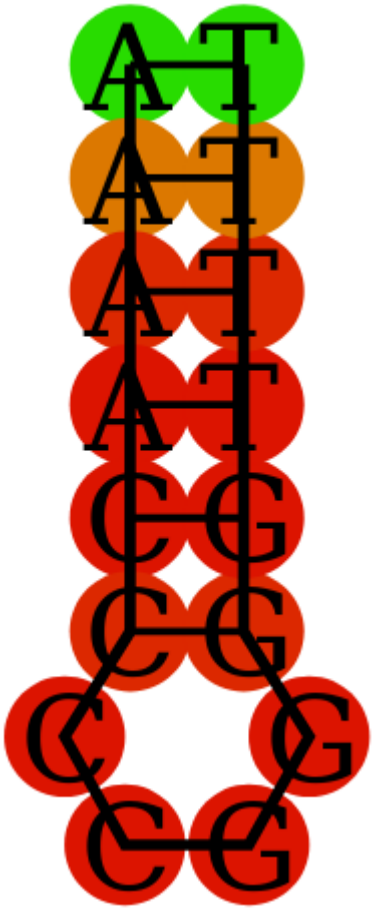
Spacers amplification in one array

In development

Skoltech
Skolkovo Institute of Science and Technology

Between denaturation and annealing steps of PCR, arrays can form secondary structures. This could be a reason for different amplification.

We calculated DNA secondary structures and tested several CNN and RNN architectures to predict spacer amplification by it's array secondary structure



Does not work at the moment, probably due to lack of data

Results

Skoltech
Skolkovo Institute of Science and Technology

At the moment:

- Done with classifying reads
- Can restore all parts of arrays from classified reads

Future plans:

- Check on more *E.Coli* data
- Predict amplification of different spacers
- Restore arrays with ratios in original mixture
- Develop full pipeline

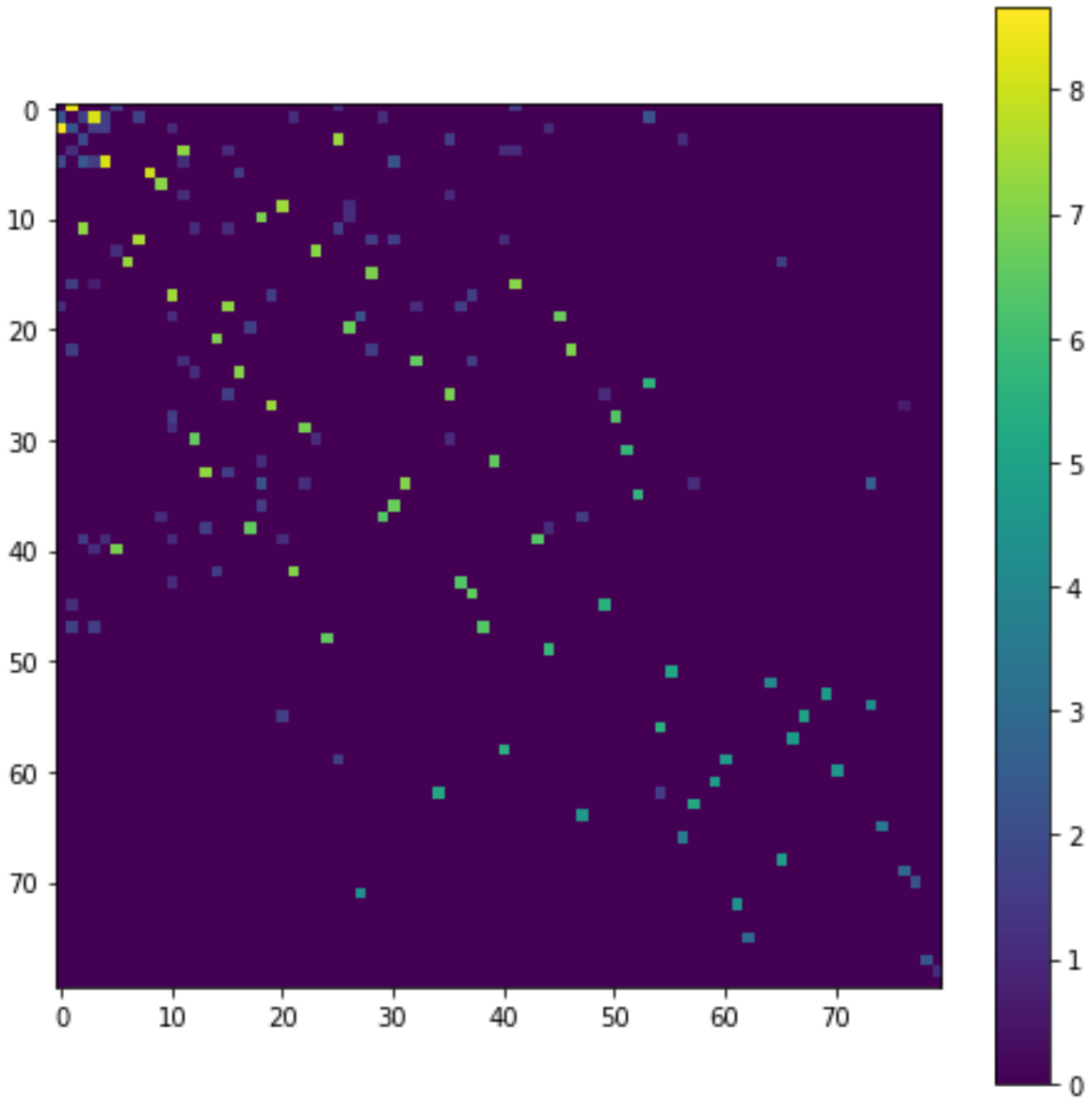
Skoltech

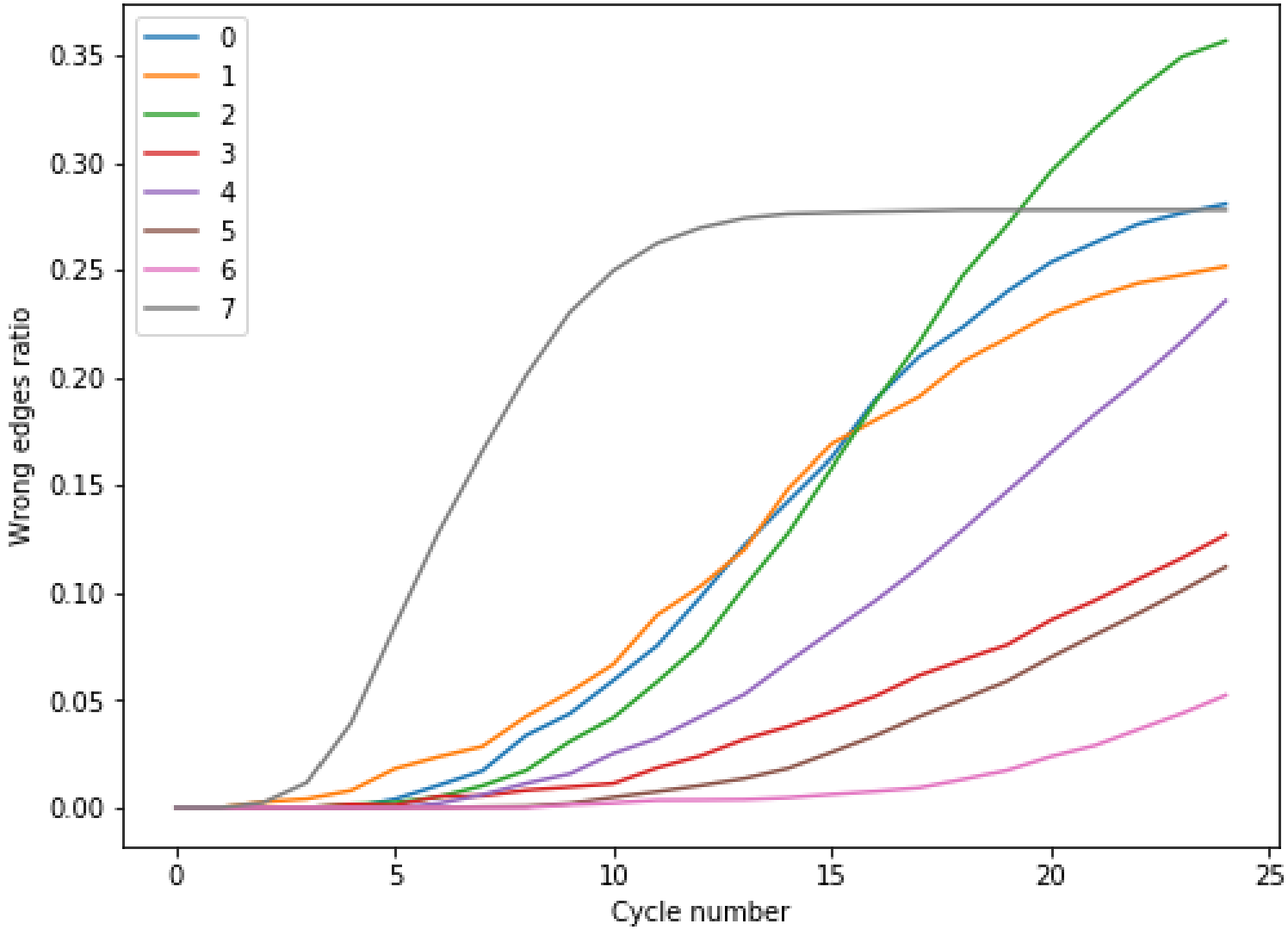
Skolkovo Institute of Science and Technology

Thank you for your attention

Skoltech

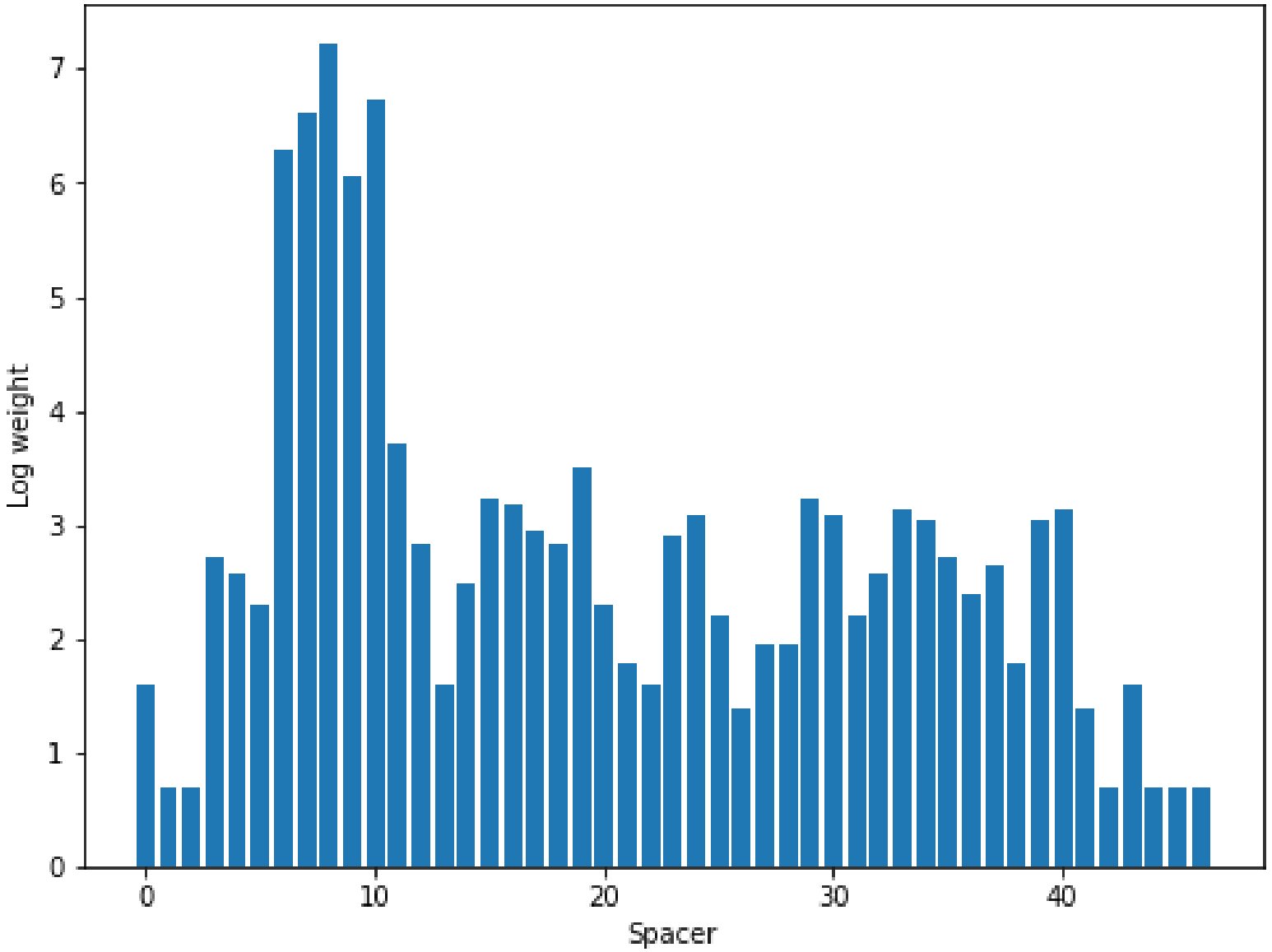
Skolkovo Institute of Science and Technology




0
1
2
3
4
5
6
7
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Wrong edges ratio
0
5
10
15
20
25
Cycle number

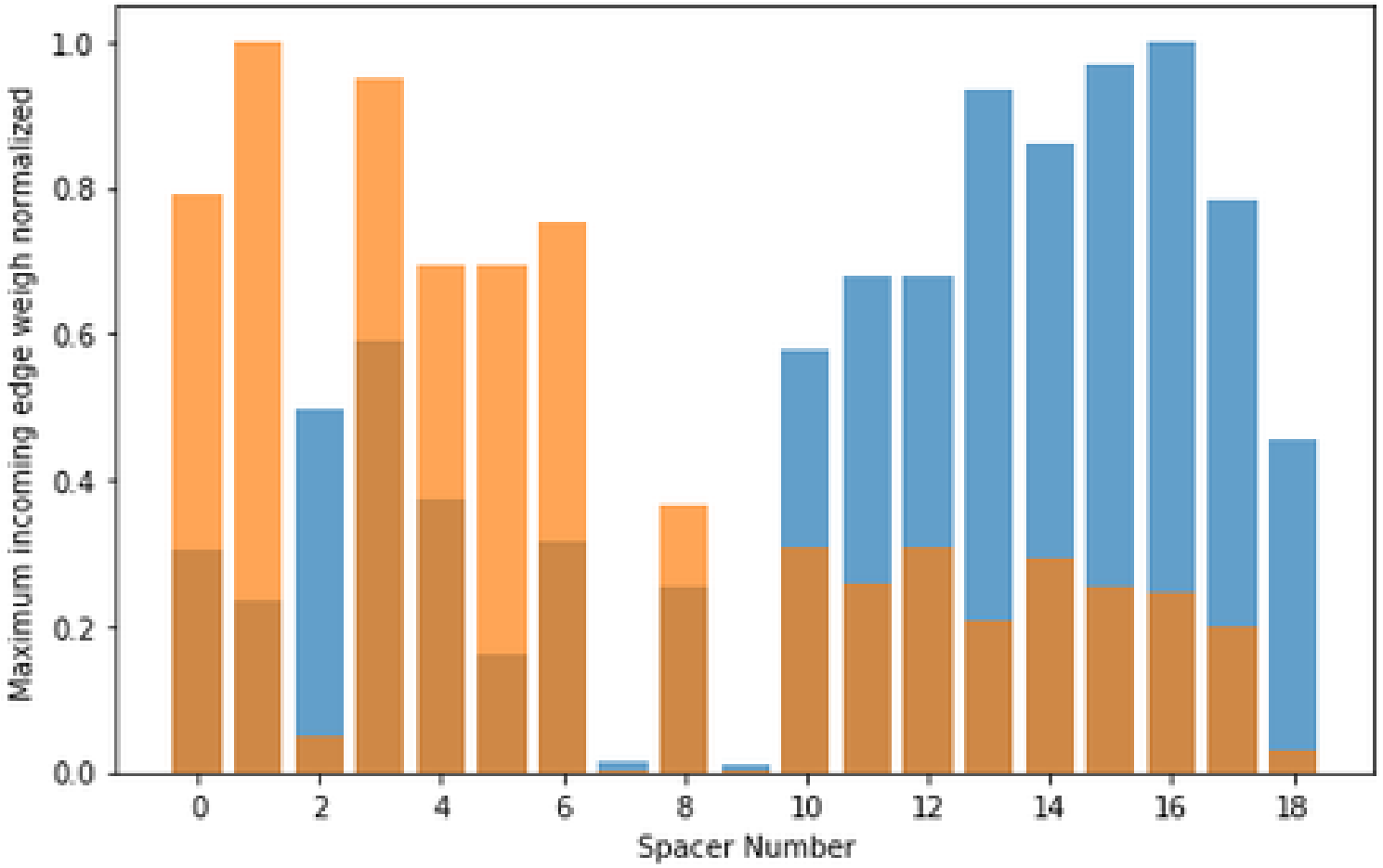
Skoltech

Skolkovo Institute of Science and Technology



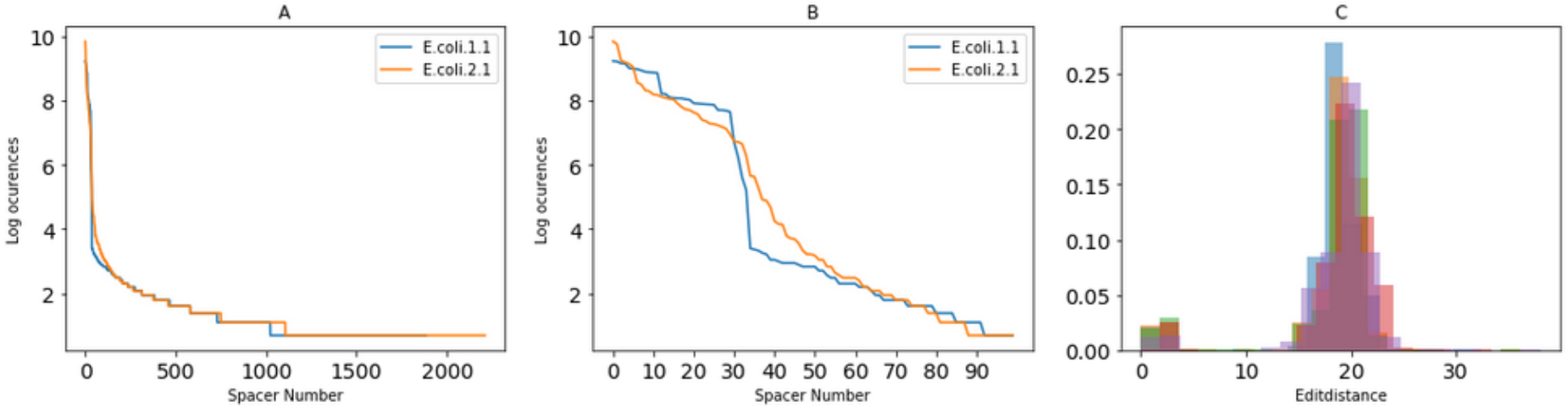
Skoltech

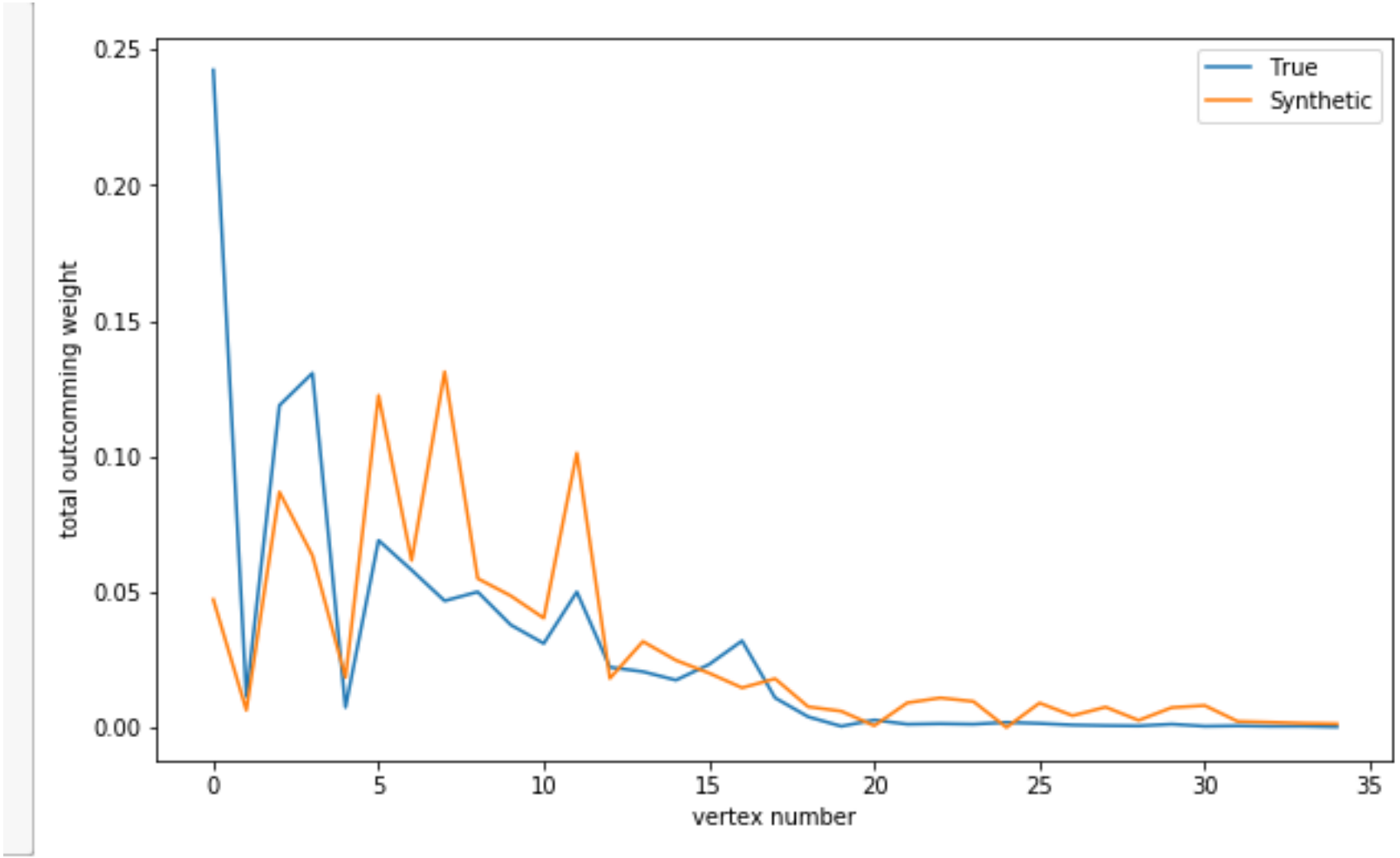
Skolkovo Institute of Science and Technology

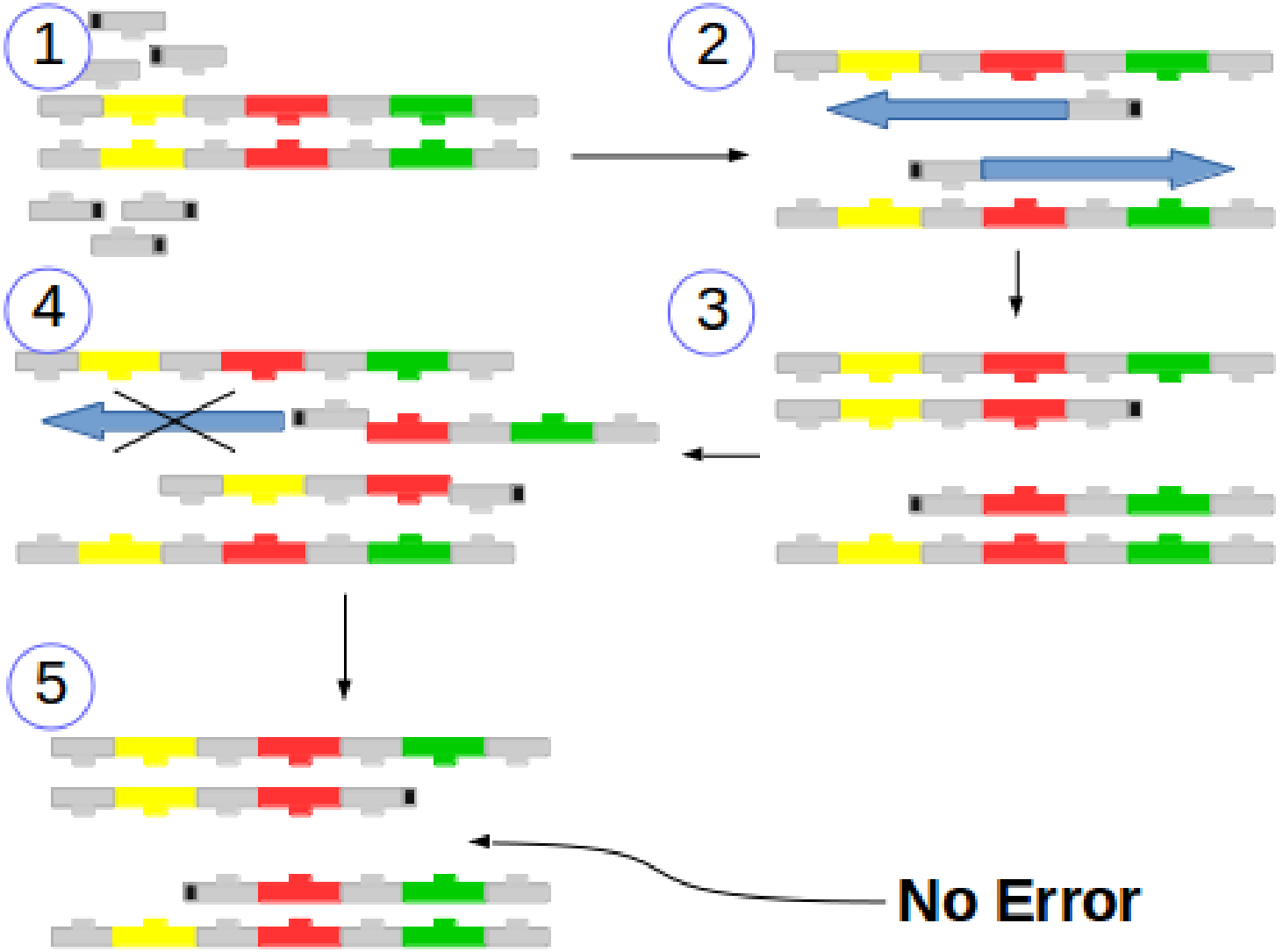


Skoltech

Skolkovo Institute of Science and Technology




True
Synthetic
total outcomming weight
vertex number
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
25
30
35


1
2
3
4
5
No Error