Development of computational algorithms linking epigenetic features and three-dimensional organization of chromatin

Daniil Fishman

Novosibirsk State University

23.04.2020

Task description: Why it is important?

It is assumed that genome has a loop organization, so the far in linear structure parts of the genome appear close to each other in space. A number of studies have shown that changes in the 3D-contacts of the specific parts of genome during chromosomal rearrangements can lead to the genetic diseases.
Existing methods for determining 3D organization of the genome implies a series of time-consuming experiments. Therefore, prediction of 3D-contacts of normal and mutated genomes is highly important for clinical diagnostics.

Goals and tasks

Goals: The main goal of the following work is to predict contacts between different regions in DNA.
Tasks: To achieve that goal we are going to develop an algorithm, using the experimental information about DNA structure and DNA-protein interactions and applying machine learning techniques.

Task description: Mathematical problem definition

Lets present DNA of length $genome\_size$ (which means DNA consist of $genome\_size$ letters) as a stretch of segments of size $dist\_bin$, $1 <= dist\_bin <= genome\_size$.
Let i and j be indexes (coordinates) of two DNA segments of length $dist\_bin$, $1 < i, j < genome\_size/dist\_bin$.
Let S be symmetric matrix, each value S_{ij} correspond to experimental measure reflecting Euclidean distance between DNA segments i and j. **We will call S_{ij} contact between i and j.**
Let A be experimentally measured DNA-protein interaction matrix, where A_{kp} is experimentally measured interaction between protein k and DNA segment p of length 1, $k = 1, \ldots, N$;
$p = 1, \ldots, genome\_size$.
Let $B = B_1 .... B_{genome\_size}$ be a vector of categorical variables of length $genome\_size$, with each element $B_k \in \{A, T, G, C, N\}$ representing experimentally measured DNA sequence.
Task: For each given A, B, i, j and $dist\_bin$ satisfying $|i - j| * dist\_bin < 1.5e^7$ predict S_{ij}.

Approach 1

Use existing algorithm:

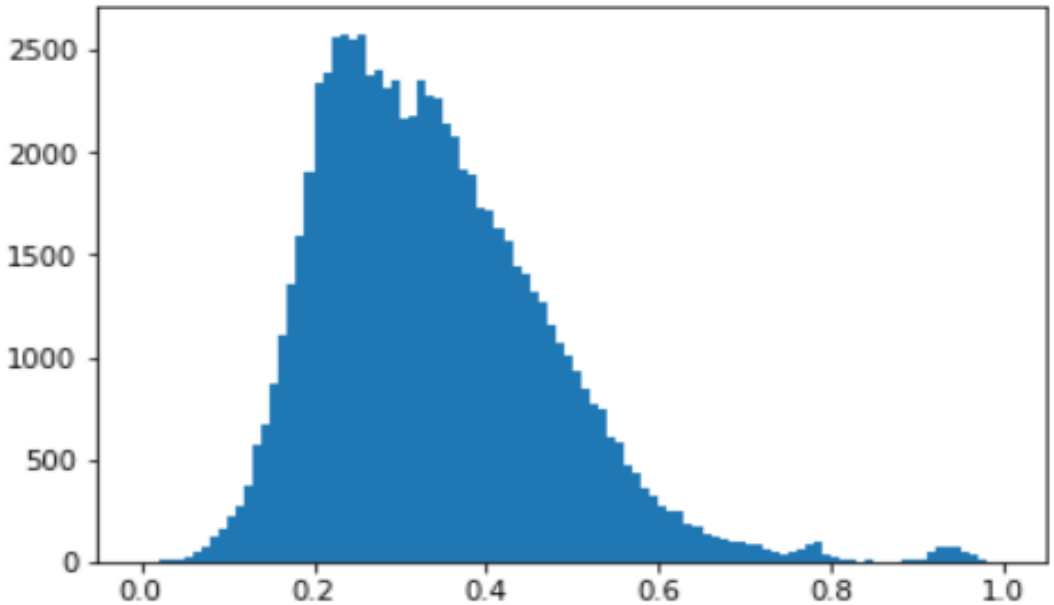

Figure 1: Histogram of S_{ij} values

Sean Whalen, Rebecca M Truty, Katherine S Pollard, Nature Genetics 2016 "Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin" (TargetFinder). Nature Genetics, Impact factor 27.125.

Approach 1

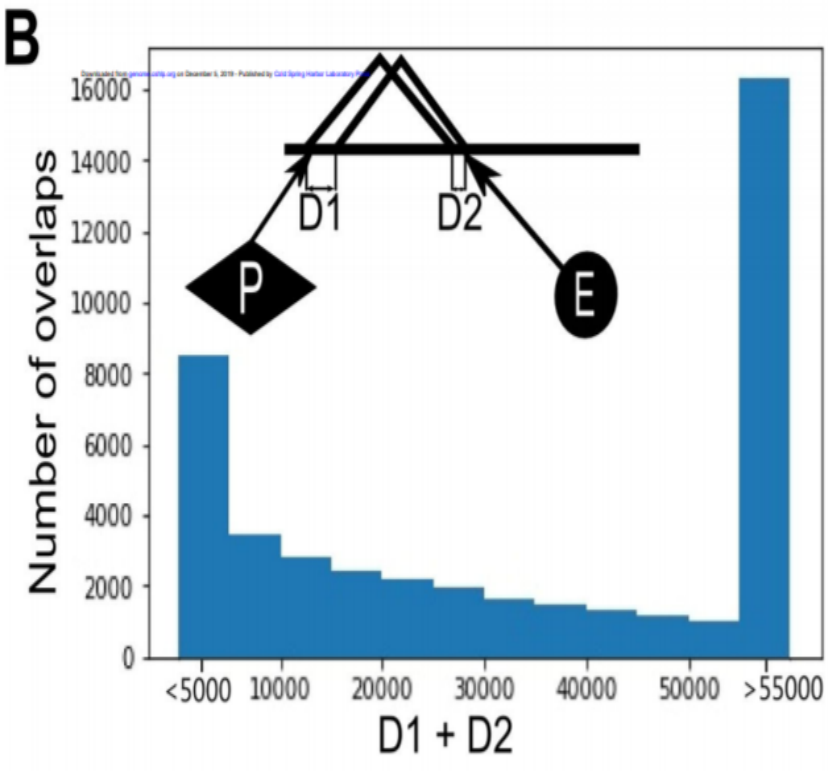

Figure 2: Intersection distribution.

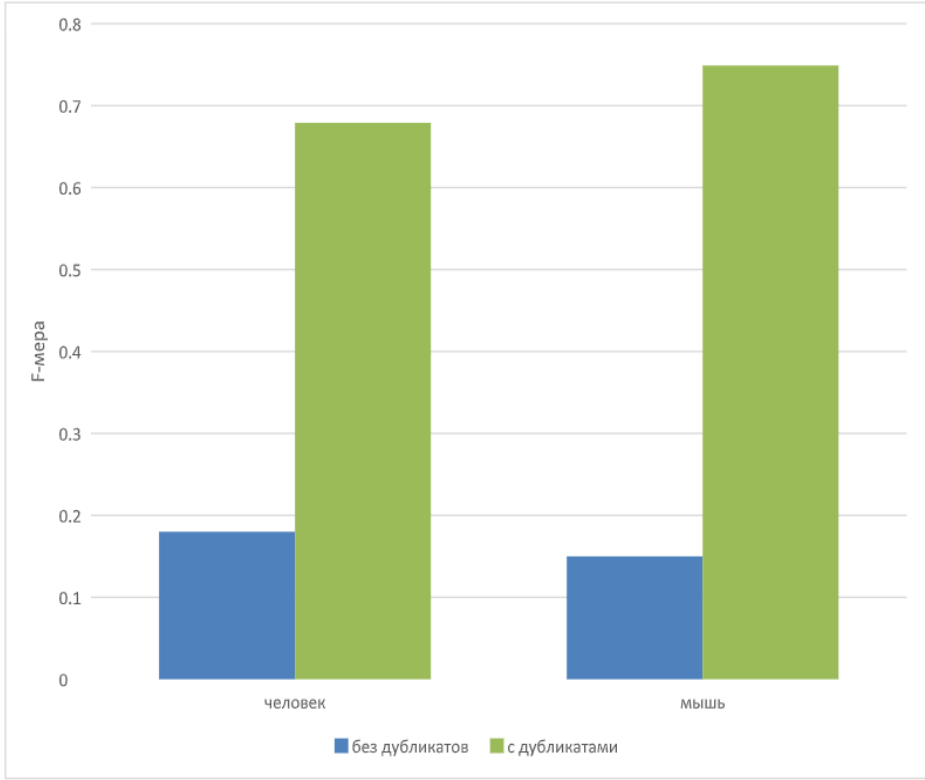

Figure 3: After and before removing duplicates.

Approach 2

Develop new method to predict 3D structure:
We will predict contacts not just for "contact-rich areas" but for all regions with a distance less than $1,5 * e^7$.

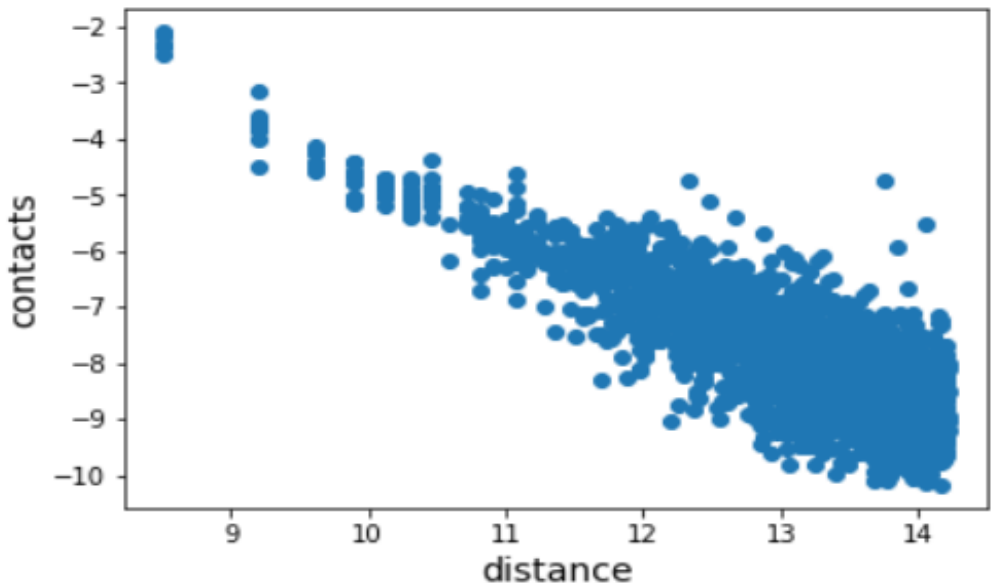

Figure 4: Contacts-distance dependence on logarithmic scale

Data structure and preparation

- $\sim$ 120000 objects in train
- $\sim$ 30000 objects in test
- Information about 15 proteins in the "window" between regions
- $5000, 10000, 15000, \ldots, 15000000$ possible window sizes
- Unprocessed values of proteins (vectors) considered as features

Methodologies

- Classical algorithms (Gradient boosting, linear regression) using statistical features.
- Neural networks using unprocessed signals.

Techniques

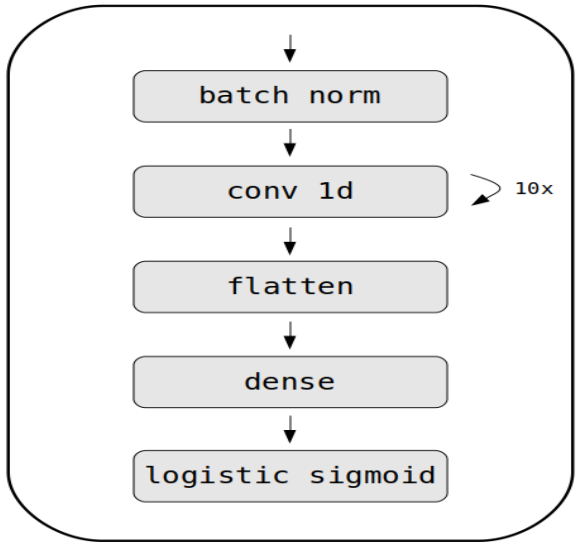

Figure 5: Model architecture

Training process

- SGD optimizer ($\mathrm{lr} = 3 * 10^{-6}$)
- Batch normalization
- log contacts
- sigmoid activation function on last layer
- cos lr sheduler
- 100 epochs (best usually is about 30-40)

Results

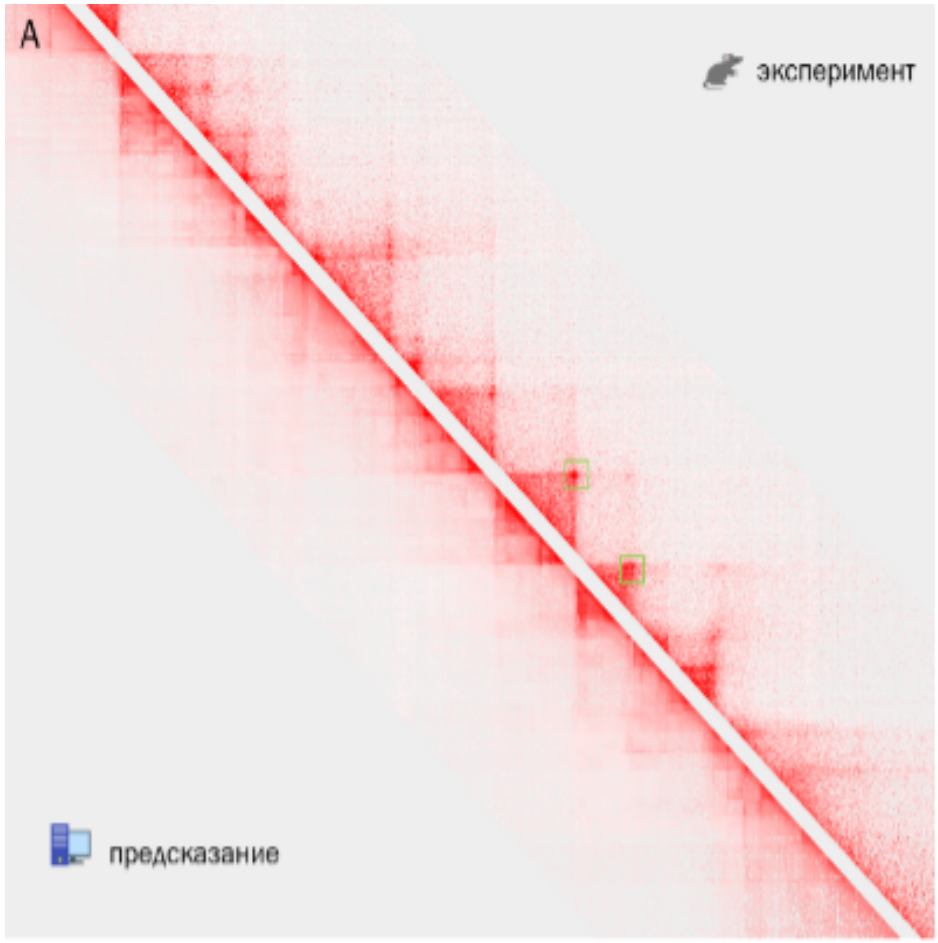

Figure 6: Predicted/train

Results

Algorithm	MSE
Linear regression	$4.5 * e^{-5}$
Gradient Boosting	$3.1 * e^{-5}$
Neural Network	$7.4 * e^{-7}$

Table 1: Mean Squared Error for different algorithms