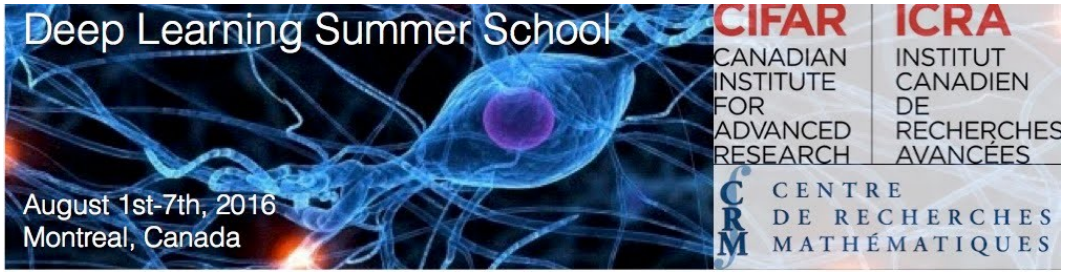


Deep siamese neural network for prediction of long-range interactions in chromatin

by **Davide Chicco**, Michael M.Hoffman
davide.chicco@gmail.com

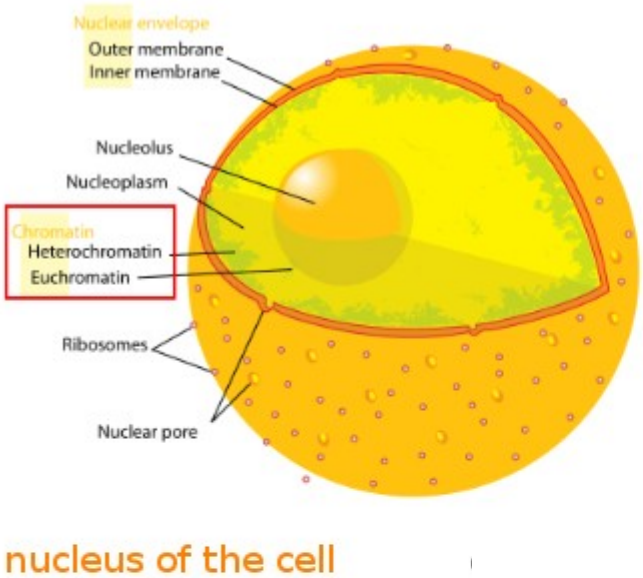
6th August 2016

Princess Margaret Cancer Centre
University Health Network
UHN

UNIVERSITY OF TORONTO

Biological problem

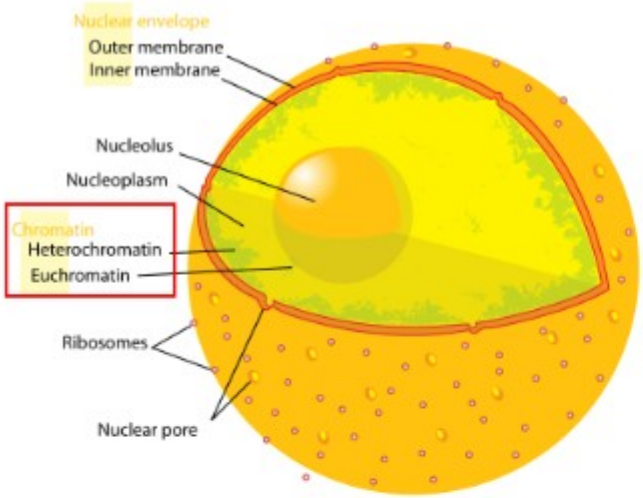
- Chromatin is the combination of DNA and proteins that form chromosomes within the nucleus of eukaryotic cells

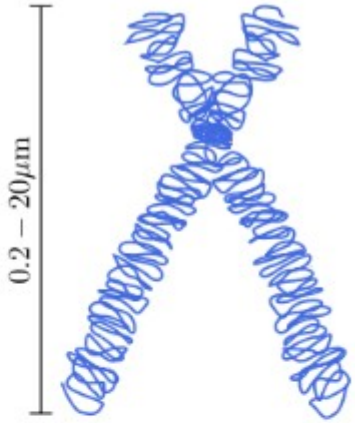


Nucleus if the cell - image from: Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Diagram_human_cell_nucleus.svg)

Biological problem

- Chromatin is the combination of DNA and proteins that form chromosomes within the nucleus of eukaryotic cells



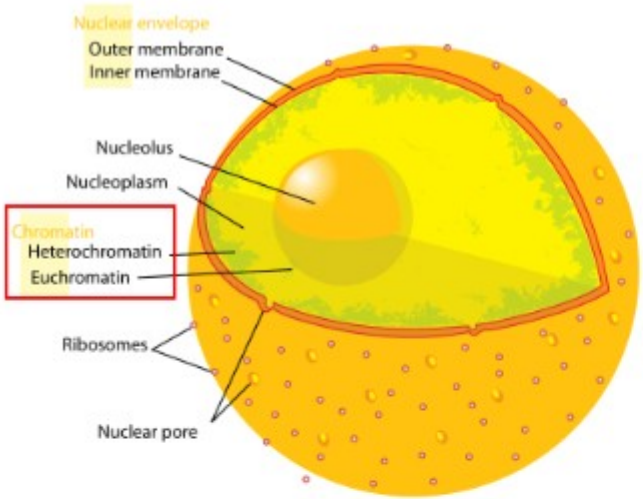


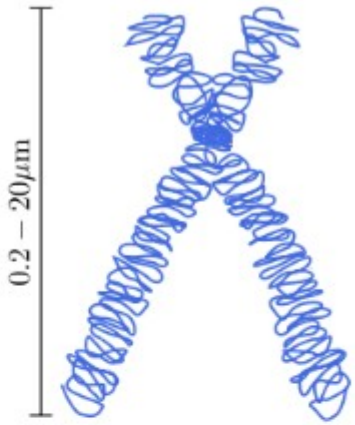
nucleus of the cell during mitosis, contains many metaphase chromosomes

Nucleus if the cell - image from: Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Diagram_human_cell_nucleus.svg)
Chromosome – image from: Wikim edia Commons (https://commons.wikimedia.org/wiki/File:Chromosome.svg)

Biological problem

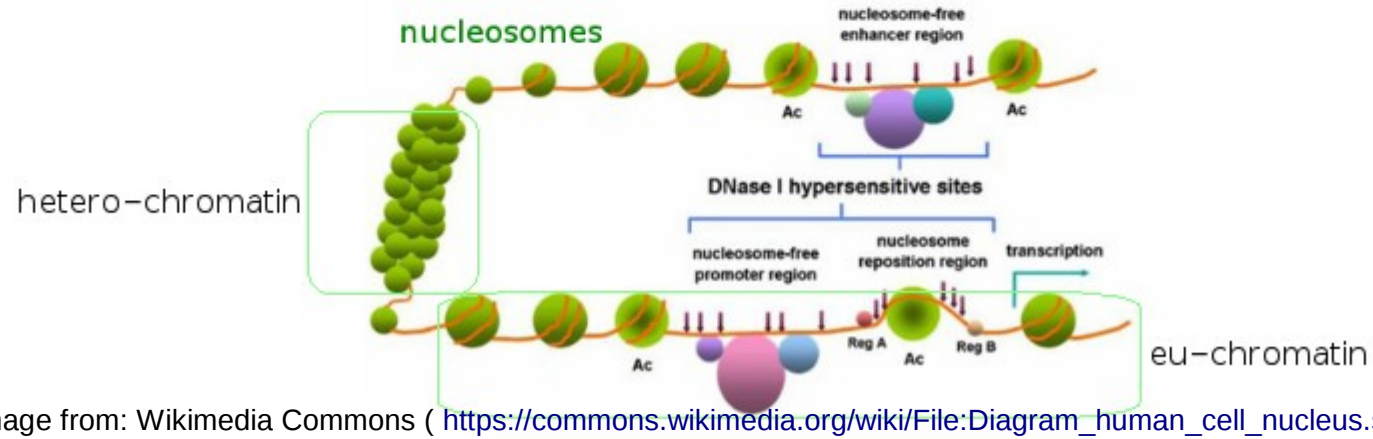
- Chromatin is the combination of DNA and proteins that form chromosomes within the nucleus of eukaryotic cells





nucleus of the cell during mitosis, contains many metaphase chromosomes

metaphase chromosomes are made by **Chromatin**



Nucleus if the cell - image from: Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Diagram_human_cell_nucleus.svg)
Chromosome – image from: Wikim edia Commons (https://commons.wikimedia.org/wiki/File:Chromosome.svg)
Chromatin - image from: "*Correlation Between DNase I Hypersensitive Site Distribution and Gene Expression in HeLa S3 Cells*". PLoS ONE, 2012

Biological problem

- Chromatin is the combination of DNA and proteins that form chromosomes within the nucleus of eukaryotic cells.

- Chromatin structure can significantly affect gene regulation and in transcriptional regulation. When it's more open, there is a higher chance that it might be experiencing a gene transcriptional phase.

- Transcriptional regulation depends on physical interactions between regulatory elements like enhancers and promoters, that are often not adjacent in a linear sense, even if they might be adjacent in a 3D sense.

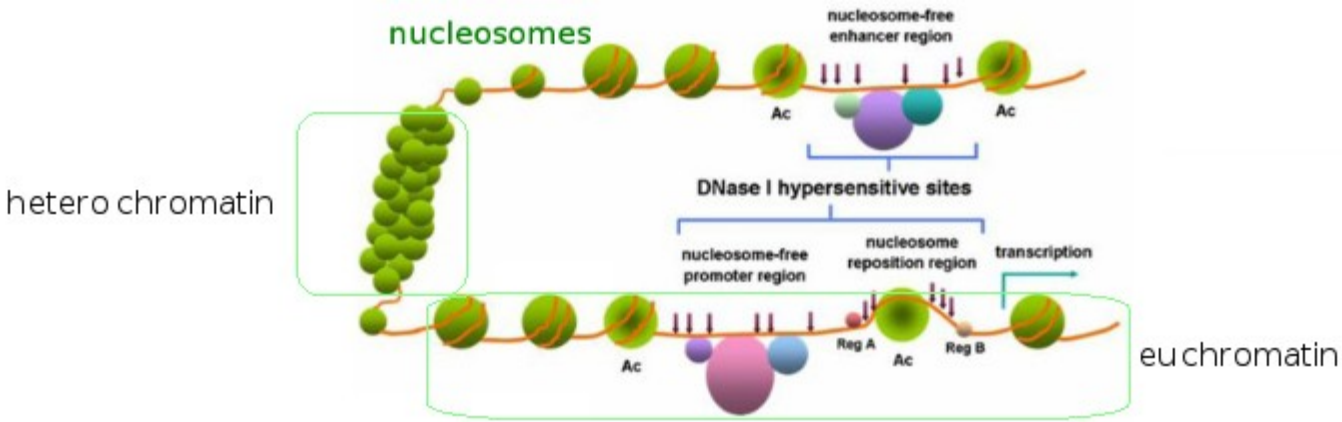


Image from: Wang Y-M, et al. "*Correlation Between DNase I Hypersensitive Site Distribution and Gene Expression in HeLa S3 Cells*". PLOS ONE, 2012

Tech problem

- Originally, former technologies used by biologists to understand genome organization were not able identify individual physical interactions, like those between enhancers and promoters.

- However, they have defeated these limitations in recent years with a series of molecular techniques based on chromatin conformation capture (3C) and Hi-C.

- Very useful, but unfortunately are also very expensive, in both money and research time. In addition, they involve difficult techniques that few laboratories have the resources or skills to complete.

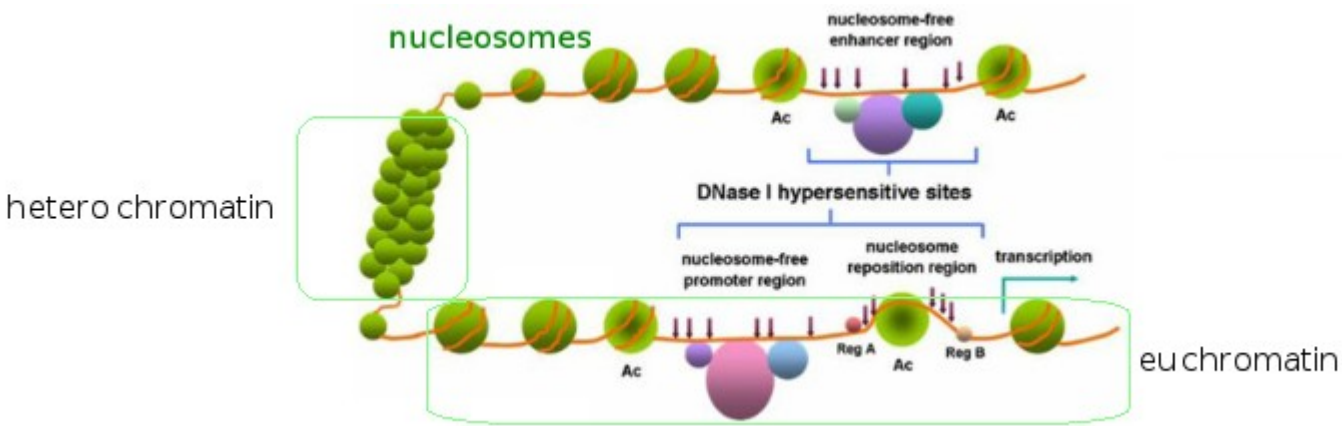


Image from: Wang Y-M, et al. "*Correlation Between DNase I Hypersensitive Site Distribution and Gene Expression in HeLa S3 Cells*". PLOS ONE, 2012

Biological problem

- DNase I hypersensitive sites (DHSs) are regions of chromatin that are sensitive to cleavage by the DNase I enzyme, and where chromatin is more open.

- So we use DNase hypersensitivity as a measurement of the level of openness of the chromatin

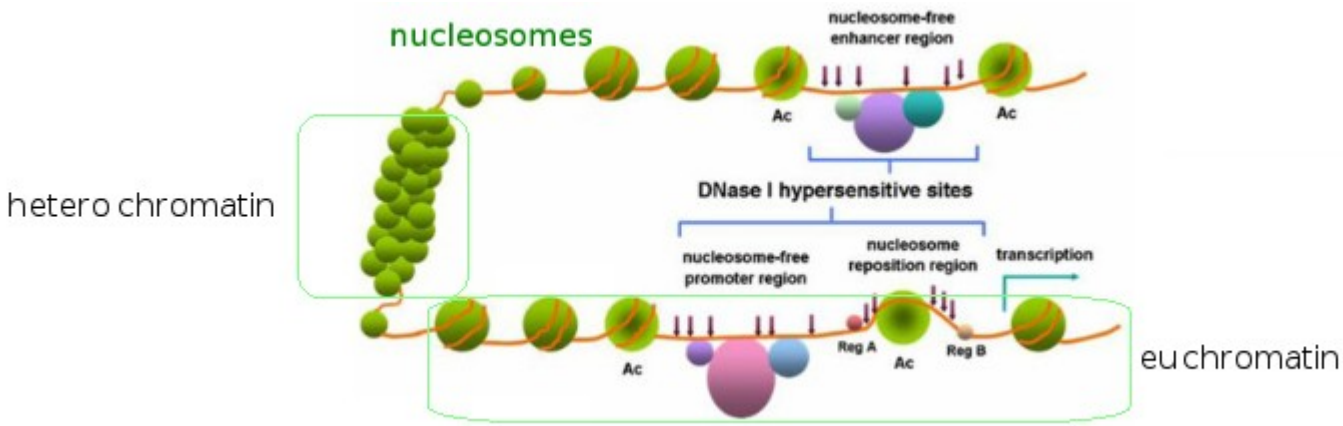


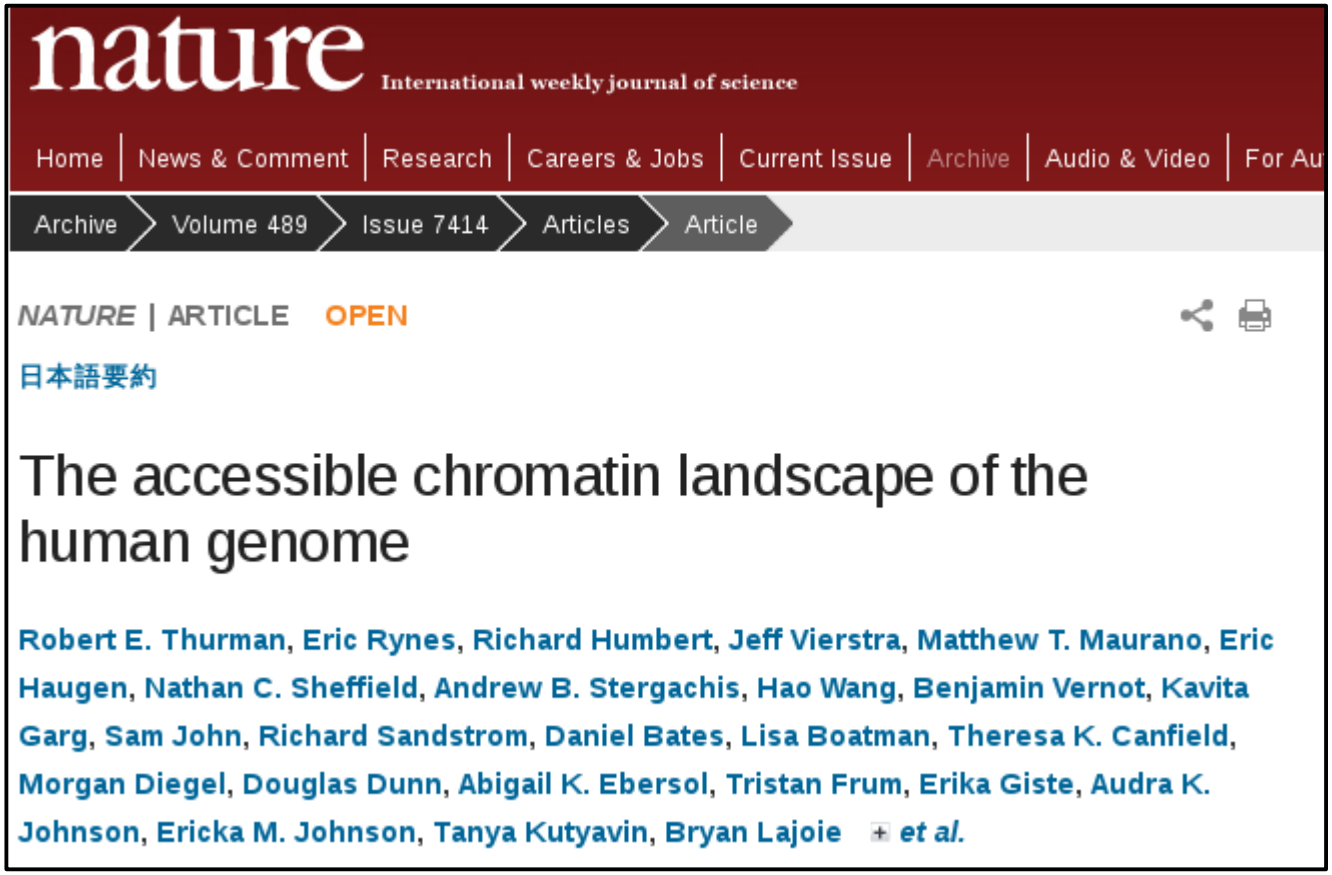
Image from: Wang Y-M, et al. "*Correlation Between DNase I Hypersensitive Site Distribution and Gene Expression in HeLa S3 Cells*". PLOS ONE, 2012

Algorithm problem

- To address these issues, scientists have recently developed new methods that rely on correlations between functional genomics assays (e.g. DNase-seq, CAGE-seq, ChIP-seq) to find chromatin interactions:

- PreSTIGE, IM-PET, RIPPLE, EpiTensor, TargetFinder

- "The accessible chromatin landscape of the human genome", by Thurman, et al., Nature 2012, highlighted first DNase datasets

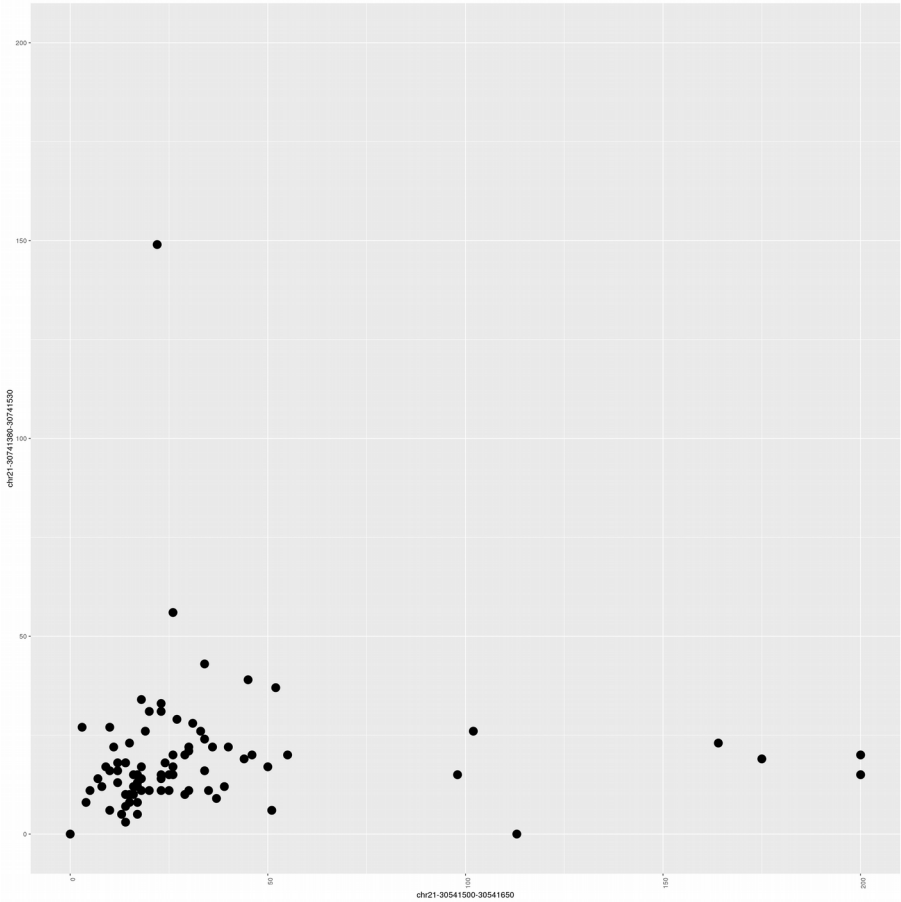


Thurman 2012 algorithm

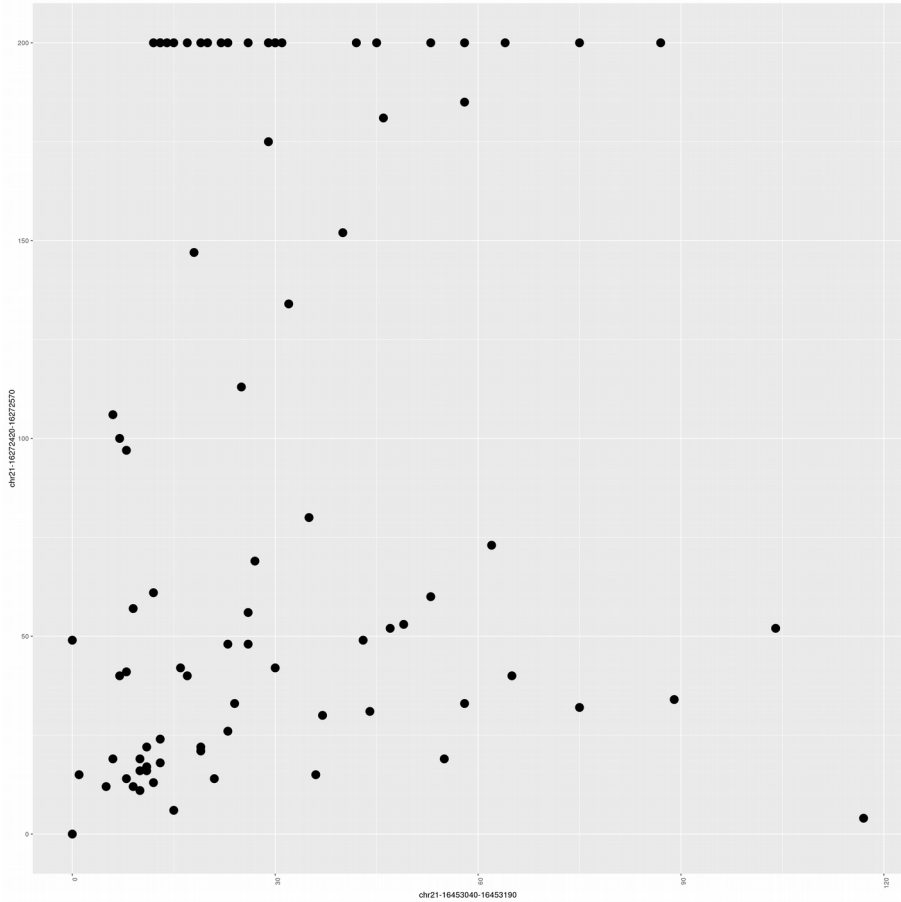
- Our goal is to compare the predictions made through our model with the interactions discovered by Thurman 2012 algorithm.

- Thurman and colleagues highlighted that the correlation between DNase I signal profiles might show the existence of an interaction between chromosome regions

true interaction
chr21-30541500-30541650 and chr21-30741380-30741530



false interaction
chr21-16453040-16453190 and chr21-16272420-16272570

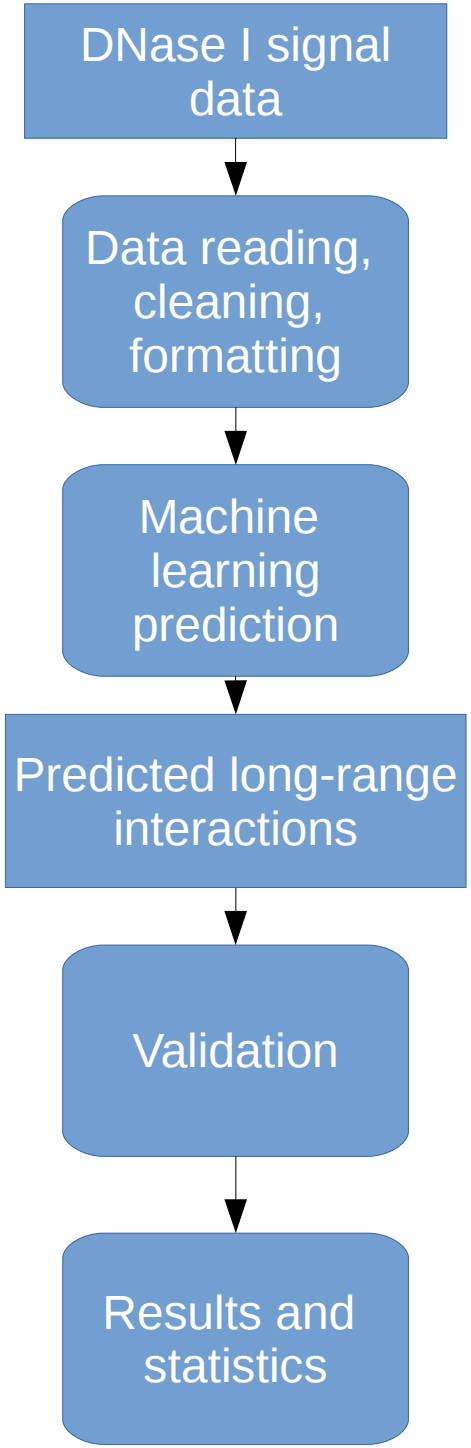


Thurman 2012 algorithm

- This method uses simple statistics correlation measures, and can somehow just analyze the existing situation, without making predictions

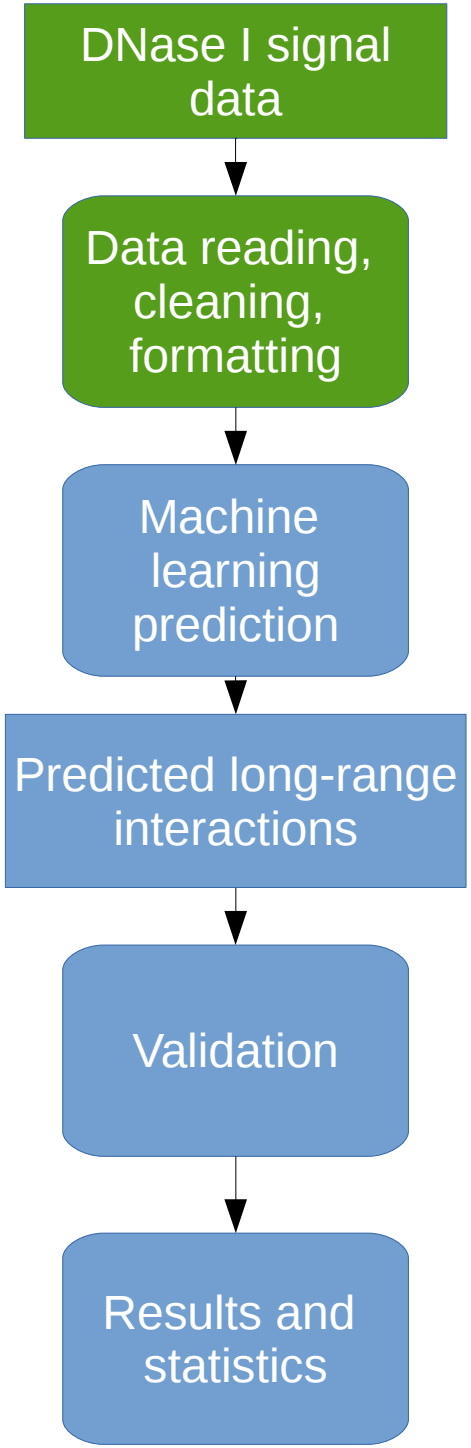
- Also, only few interactions predicted with this method were later found in the recent Hi-C datasets (current gold standard) released by Lieberman-Aiden lab (~0.1% for each chromosome).

- Since we want not only to analyze the current datasets, but also to make predictions, and possibly to integrate multiple data sources in our pipeline, a machine learning algorithm might be more suitable for this task



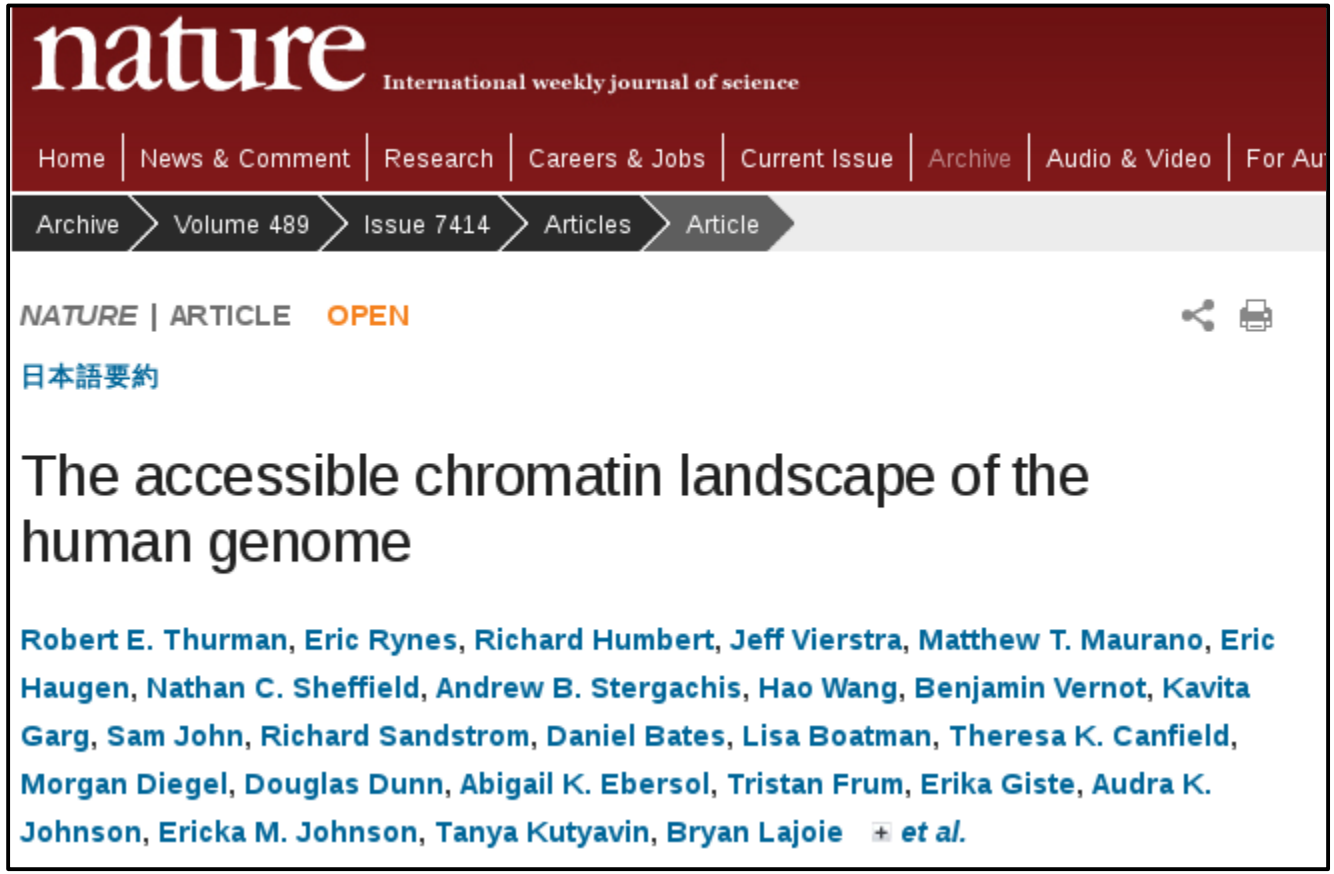
Our idea

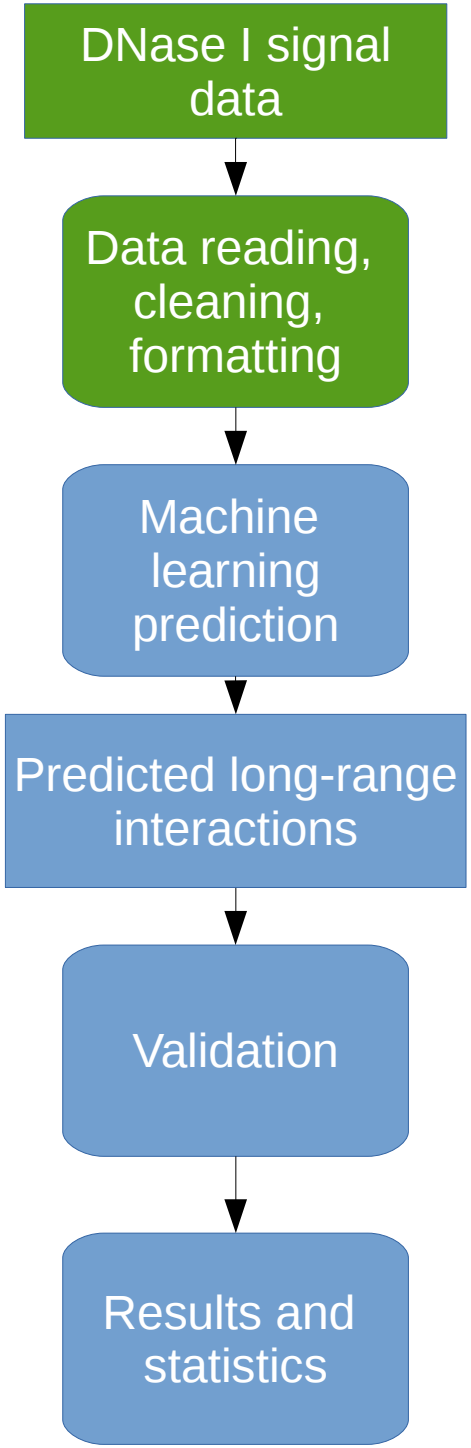
We decided to create a computational machinery, based on a machine learning method, and able to predict long-range interactions



Data reading and setting up

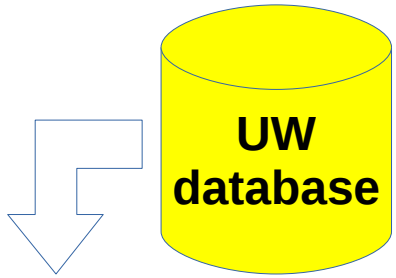
The software reads DNase I hypersensitivity peak calls used by Stamatoyannopoulos lab at University of Washington, in the Thurman et al. "The accessible chromatin landscape of the human genome", *Nature*, 2012





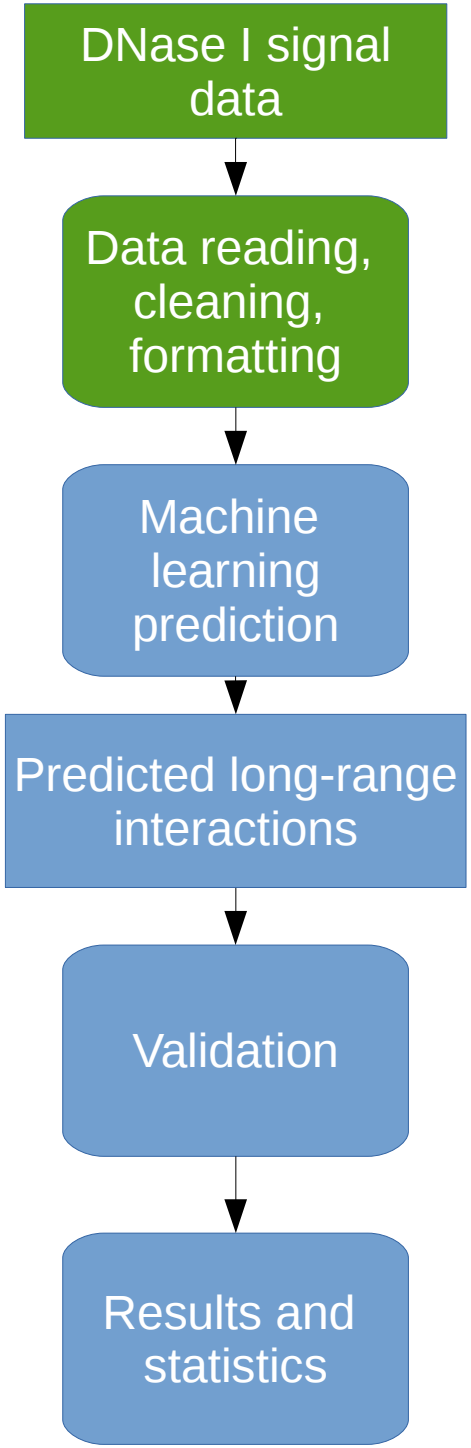
Data reading and setting up

Original matrix:
chr21: 32,692 rows * 82 columns



	A549	AG10803	A0AF	...	TROPHO BLAST	VHMEC
chr1-66660-66810	0.00	0.00	2.83	...	0.00	0.85
chr1-564520-564670	15.63	4.55	57.78	...	5.81	101.68
chr1-568060-568210	17.91	3.70	15.96	...	4.10	31.04
chr1-568900-569050	41.70	7.46	28.40	...	8.52	44.66
....	....	....	....	....	....	....

rows = chromosome regions
columns = cell types
entries = DNase I hypersensitivity (DHS) peak intensity



Data reading and setting up

As gold standard, we use the Hi-C interactions discovered by Liberman-Aiden lab and resealed with the paper Rao, Huntley, et al. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping", *Cell*, December 2014

Article

Cell

A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping

Suhas S.P. Rao,[1,2,3,4,10] Miriam H. Huntley,[1,2,3,4,5,10] Neva C. Durand,[1,2,3,4] Elena K. Stamenova,[1,2,3,4] Ivan D. Bochkov,[1,2,3] James T. Robinson,[1,4] Adrian L. Sanborn,[1,2,3,6] Ido Machol,[1,2,3] Arina D. Omer,[1,2,3] Eric S. Lander,[4,7,8,*] and Erez Lieberman Aiden[1,2,3,4,9,*]
[1]The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA
[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
[3]Department of Computer Science, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA
[5]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[6]Department of Computer Science, Stanford University, Stanford, CA 94305, USA
[7]Department of Biology, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA
[8]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[9]Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA
[10]Co-first author
*Correspondence: lander@broadinstitute.org (E.S.L.), erez@erez.com (E.L.A.)
http://dx.doi.org/10.1016/j.cell.2014.11.021

Goal: couples of chromosome regions

Long range interactions are couplings of chromosome regions that are connected in the chromatin (validation data):

- chr1-202526940-202527090 and chr1-209946180-209946330 is an interaction in the Hi-C dataset
- chr1-202536400-202536550 and chr1-227709560-227709710 is an interaction in the Hi-C dataset
- chr1-202936600-202936750 and chr1-203322060-203322210 is an interaction in the Hi-C dataset
- ... ...

Goal: couples of chromosome regions

Long range interactions are couplings of chromosome regions that are connected in the chromatin (validation data):

- chr1-202526940-202527090 and chr1-209946180-209946330 is an interaction in the Hi-C dataset
- chr1-202536400-202536550 and chr1-227709560-227709710 is an interaction in the Hi-C dataset
- chr1-202936600-202936750 and chr1-203322060-203322210 is an interaction in the Hi-C dataset
- ... ...

Definition problem:

we know the what is an interaction in the biological sense,
but we **do not** know its definition in the statistical sense

Input interaction matrix with labels

We took the validation interaction list and assigned TRUE/FALSE labels to each possible couple:

- chr1-202526940-202527090 and chr1-209946180-209946330 is an interaction in the Hi-C dataset
- chr1-202526940-202527090 and chr1-227709560-227709710 is an interaction in the Hi-C dataset
- chr1-202526940-202527090 and chr1-203322060-203322210 is an interaction in the Hi-C dataset
- … … …

- chr1-202526940-202527090 chr1-209946180-209946330 TRUE
- chr1-202526940-202527090 chr1-227709560-227709710 TRUE
- chr1-202526940-202527090 chr1-203322060-203322210 TRUE
- … … …

- chr1-202526940-202527090 chr1-209946180-209946330 TRUE
- chr1-202526940-202527090 chr1-227709560-227709710 TRUE
- chr1-202526940-202527090 chr1-203322060-203322210 TRUE
- chr1-202526940-202527090 chr1-227709560-227709710 FALSE
- chr1-202526940-202527090 chr1-203322058-203322205 FALSE
- … … …

Input interaction matrix with labels

We replaced the chromosome region names with their real DHS data signals:

- chr1-202526940-202527090 chr1-209946180-209946330 TRUE
- chr1-202526940-202527090 chr1-227709560-227709710 TRUE
- chr1-202526940-202527090 chr1-203322060-203322210 TRUE
- chr1-202526940-202527090 chr1-227709560-227709710 FALSE
- chr1-202526940-202527090 chr1-203322058-203322205 FALSE
- ... ...

DHS signals:

- 4.94 7.76 0.17 16.59 ... 15.10 5.41 3.44 0.00 41.42 ... 1.83 TRUE
- 4.94 7.76 0.17 16.59 ... 15.10 0.58 3.53 1.07 1.40 ... 1.66 TRUE
- 4.94 7.76 0.17 16.59 ... 15.10 2.20 4.79 0.36 6.46 ... 1.65 TRUE
- 4.94 7.76 0.17 16.59 ... 15.10 35.08 27.16 1.27 20.43 ... 29.39 FALSE
- 4.94 7.76 0.17 16.59 ... 15.10 0.67 4.27 0.93 4.88 ... 1.11 FALSE
- ... ...

chr21: 32,692 chromosome regions * 82 cell types
(we only consider interactions < 500kbp distant)
number of possible interactions: 18,480,814

Supervised approach: deep neural network

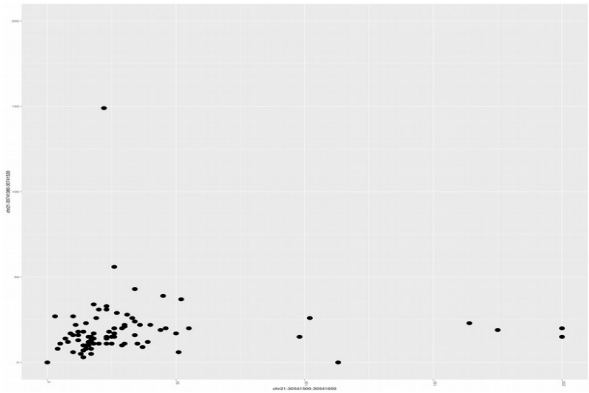
We can consider the real data submatrix as the input matrix of our neural network, and the final vector as the target array:

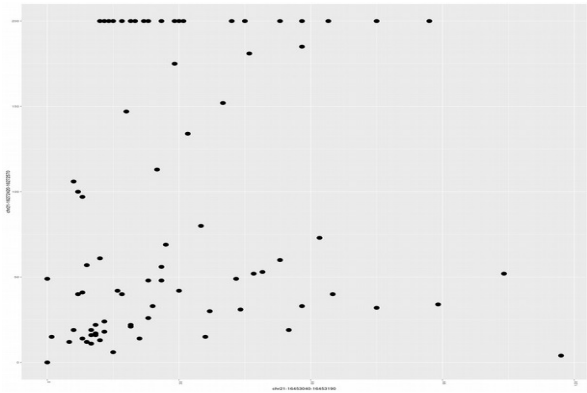
4.94	7.76	0.17	16.59	...	15.10	5.41	3.44	0.00	41.42	...	1.83	TRUE
4.94	7.76	0.17	16.59	...	15.10	0.58	3.53	1.07	1.40	...	1.66	TRUE
4.94	7.76	0.17	16.59	...	15.10	2.20	4.79	0.36	6.46	...	1.65	TRUE
4.94	7.76	0.17	16.59	...	15.10	35.08	27.16	1.27	20.43	...	29.39	FALSE
4.94	7.76	0.17	16.59	...	15.10	0.67	4.27	0.93	4.88	...	1.11	FALSE
... ...												

Input matrix target

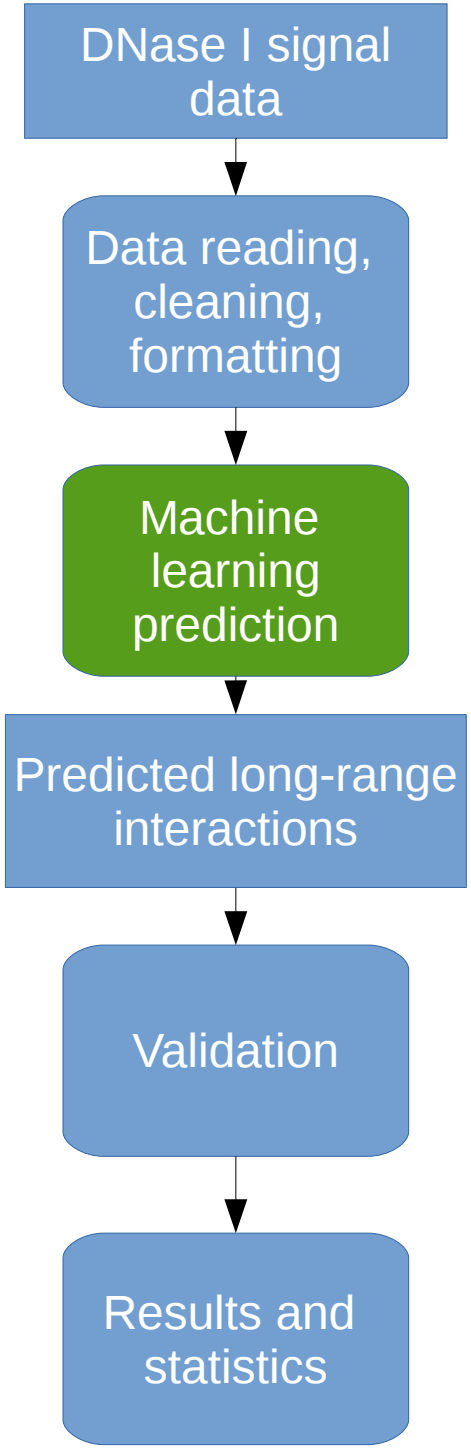
chr21: 32,692 chromosome regions * 82 cell types
(we only consider interactions < 500 kbp distant)
number of possible interactions: 18,480,814

true interaction
chr21-30541500-30541650
and
chr21-30741380-30741530





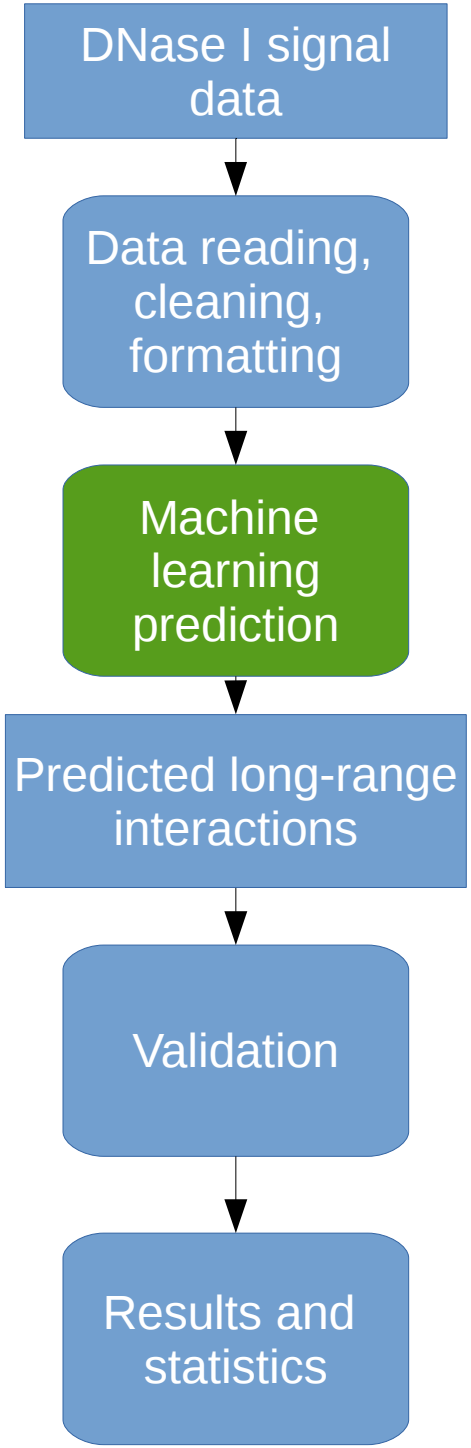
false interaction
chr21-16453040-16453190
and
chr21-16272420-16272570



Previous algorithms

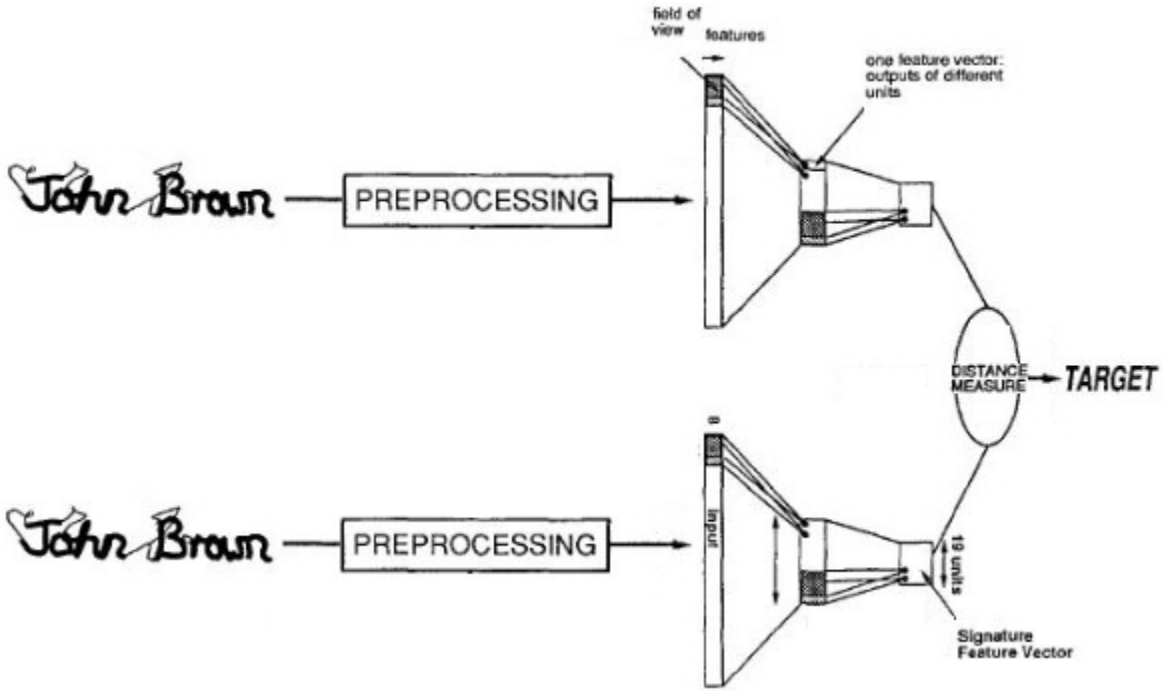
- The first machine learning method we tried is latent Dirichlet allocation (LDA), but it lead to bad results mainly because the concept of topic was adding too much complexity to this problem

- Then we tried k-means, but it lead to bad results mainly because the training was done trough geometrical coordinates of DHS, that were training the algorithm in the wrong direction



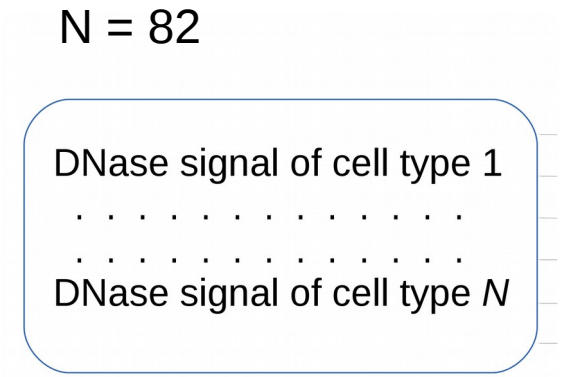
Supervised approach: deep siamese neural network

Rich Zemel suggested us to treat this like a Siamese Neural Network, first used by Yann LeCun in the paper entitled "Signature verification using a siamese time delay neural network" (NIPS 1994)



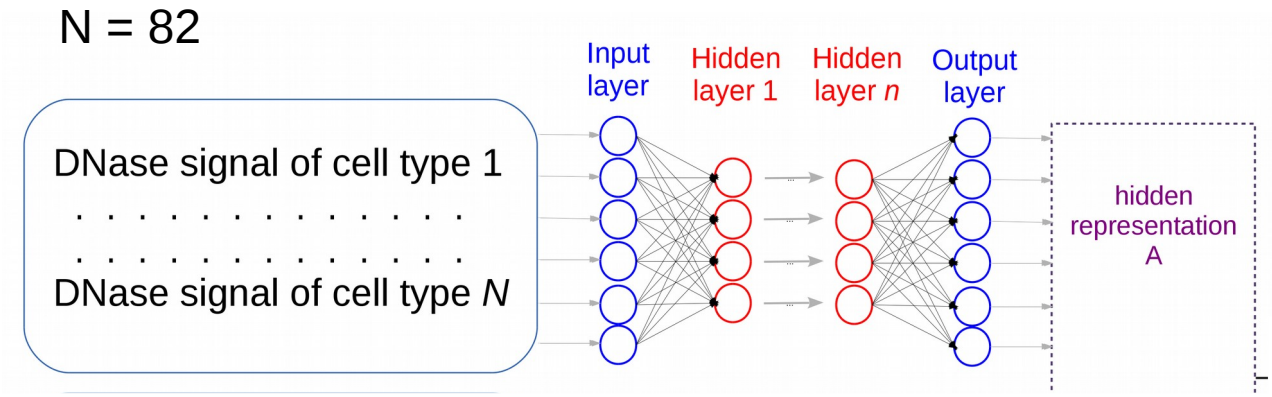
Supervised approach: deep siamese neural network

Siamese neural network architecture:



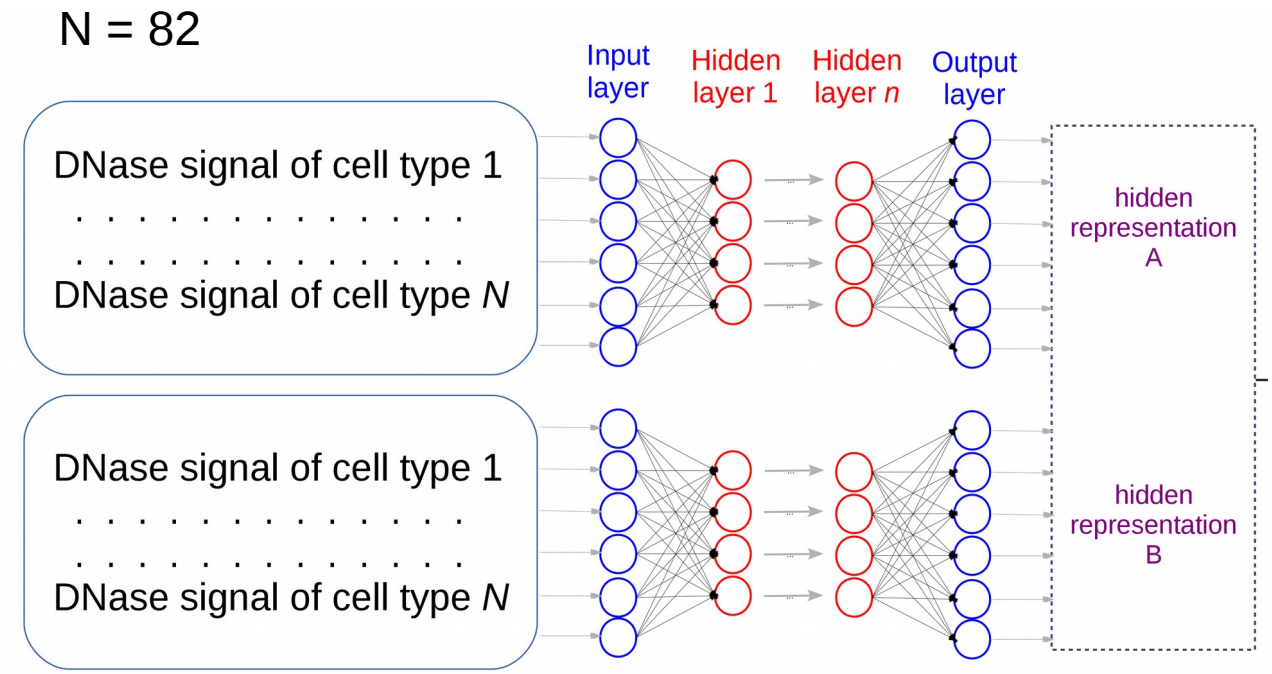
Supervised approach: deep siamese neural network

Siamese neural network architecture:



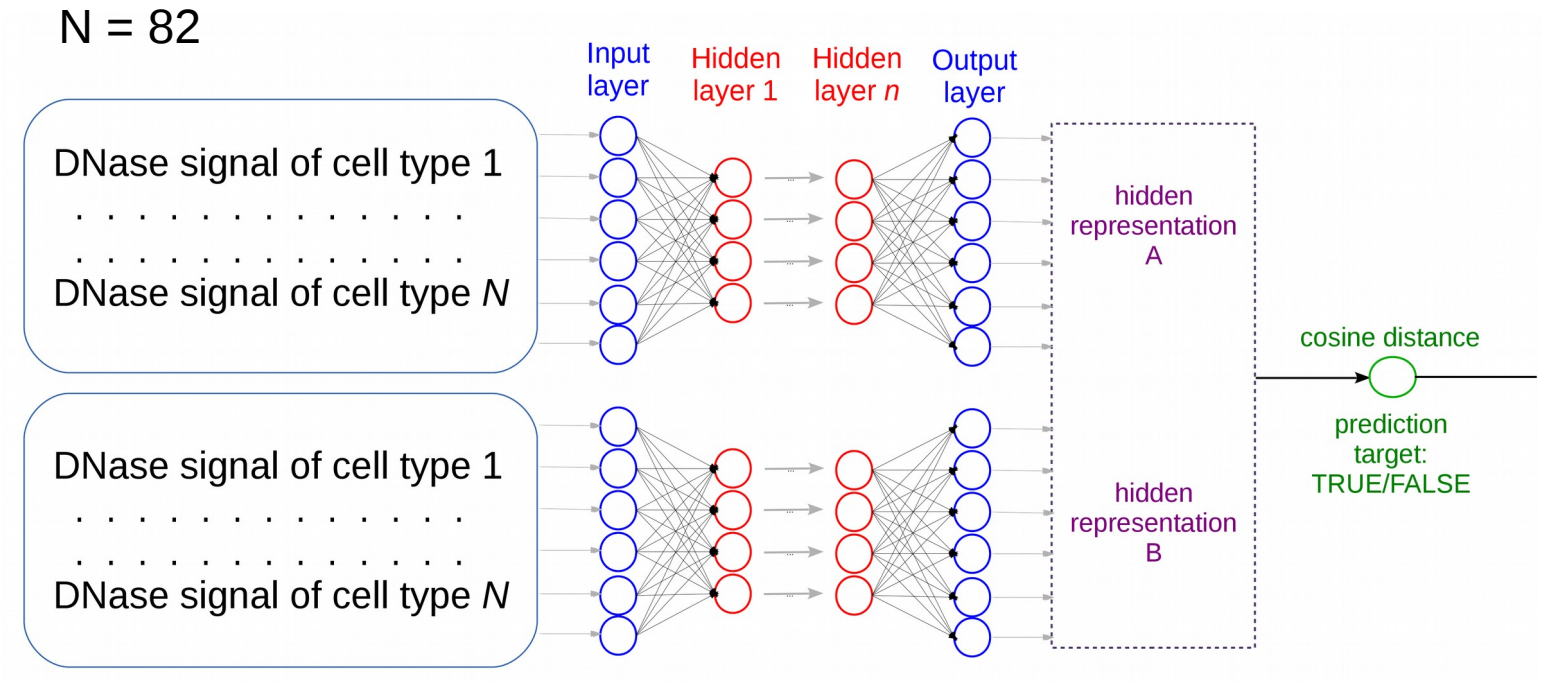
Supervised approach: deep siamese neural network

Siamese neural network architecture:



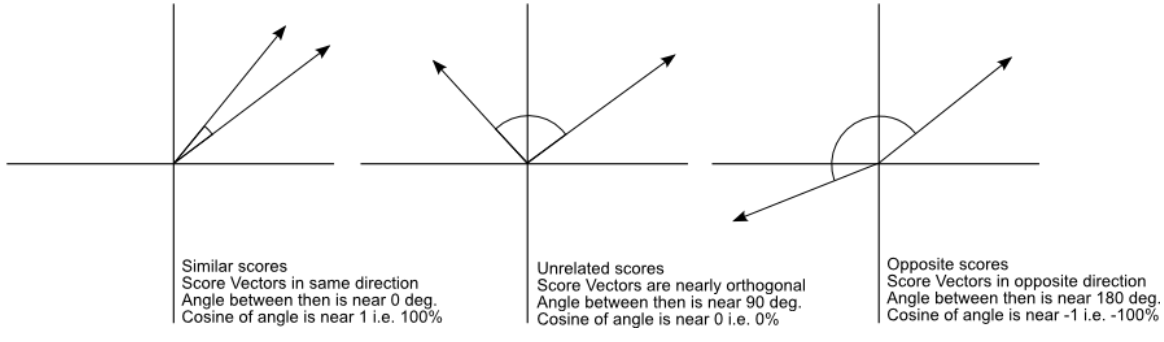
Supervised approach: deep siamese neural network

Siamese neural network architecture:



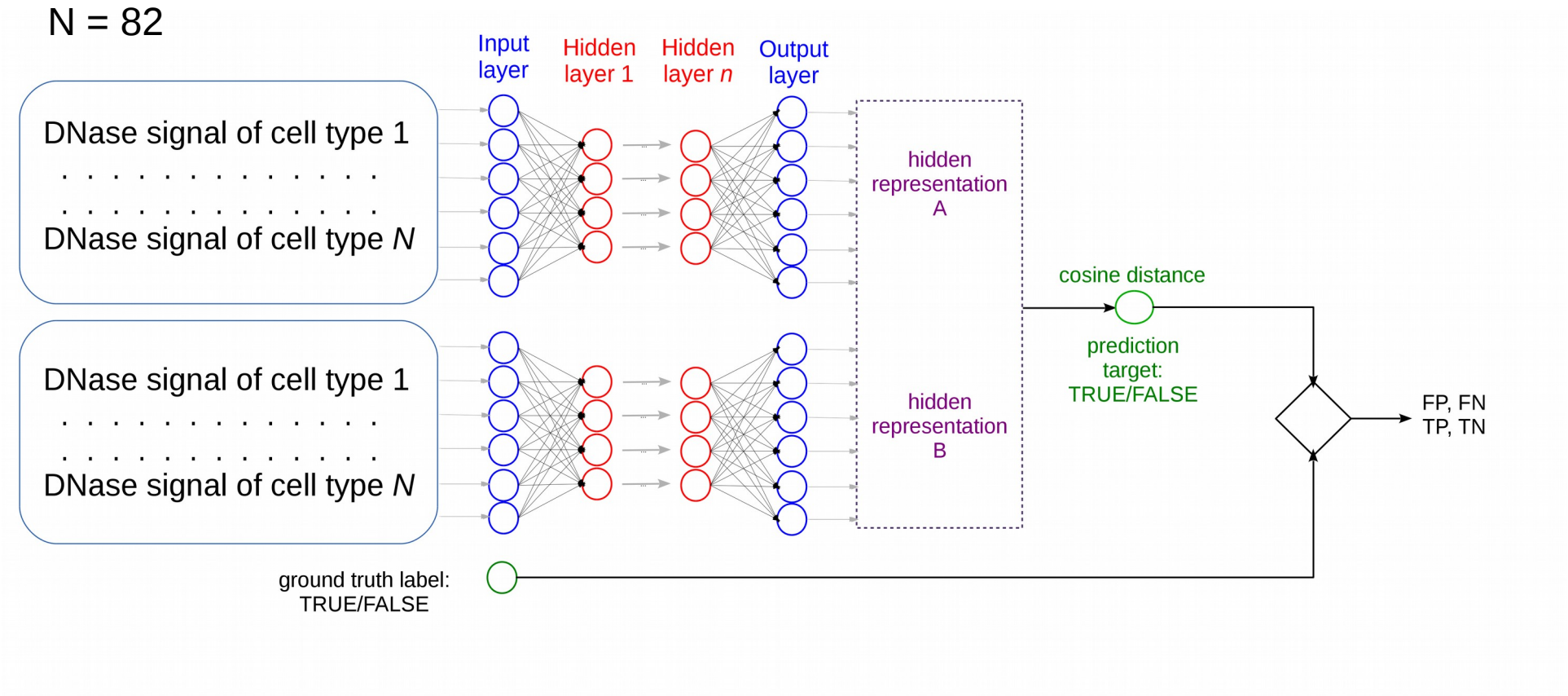
cosine distance:

Image from ChristianPerone.com

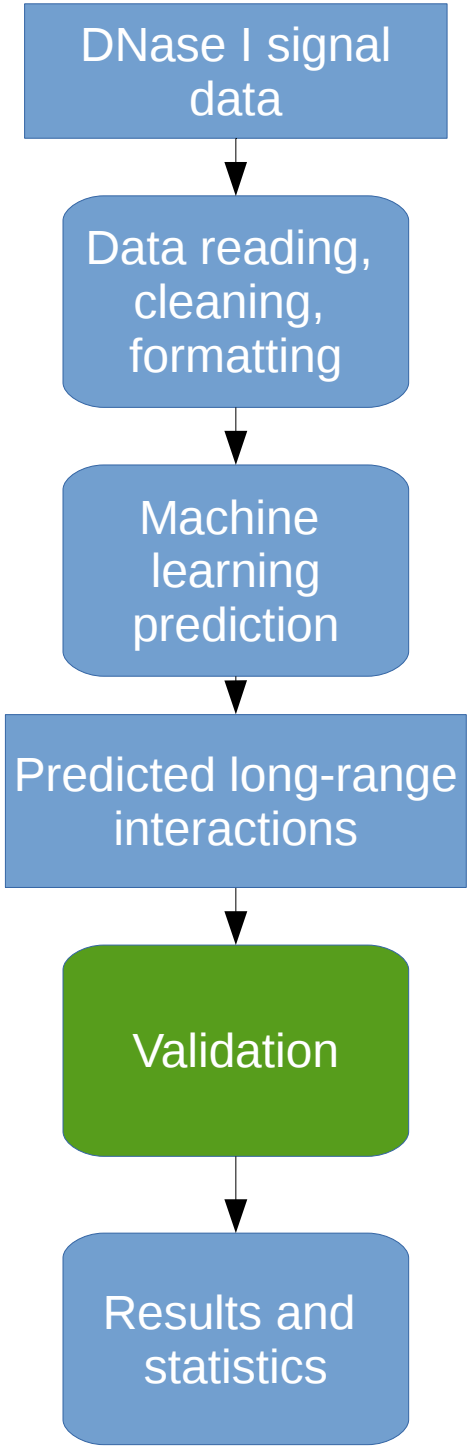


Supervised approach: deep siamese neural network

Siamese neural network architecture:



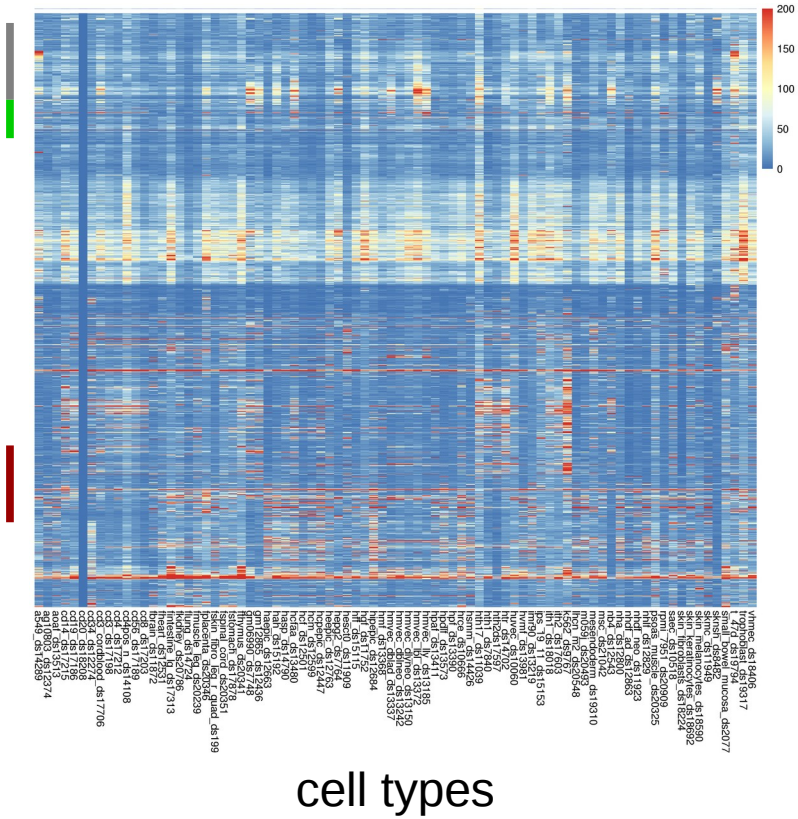
Optimization, training, validation, testing



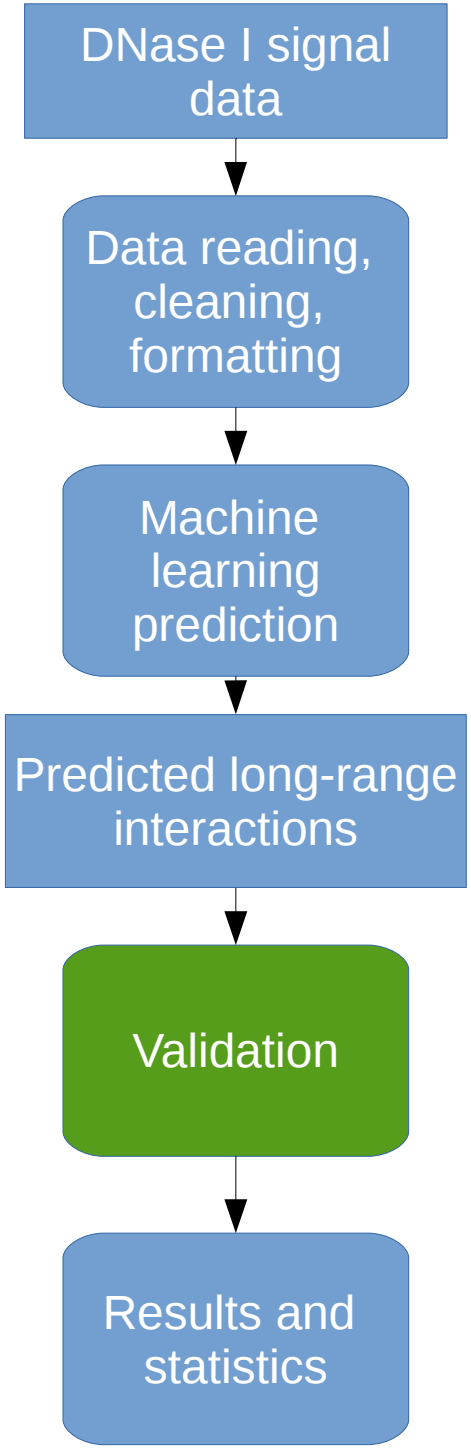
Our computational algorithm runs first an optimization phase, in which several siamese neural network models (with different hyper-parameters) are trained on a training set, and tested on a held-out validation set (each has 10,000 elements).

- training set
- validation set
- testing set

All their elements are selected through all the chromosome regions



Optimization, training, validation, testing

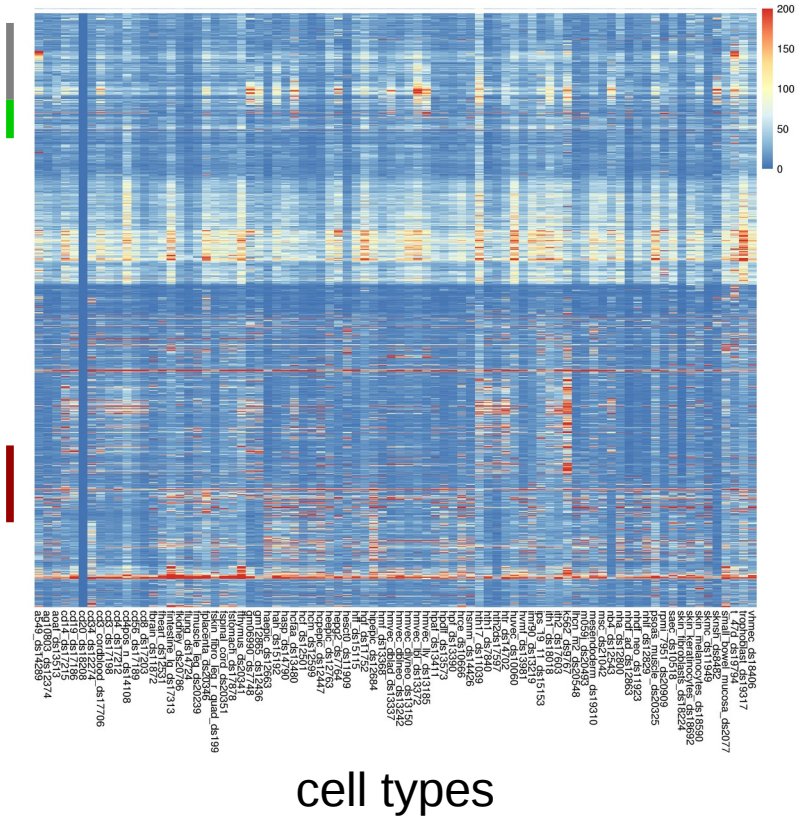


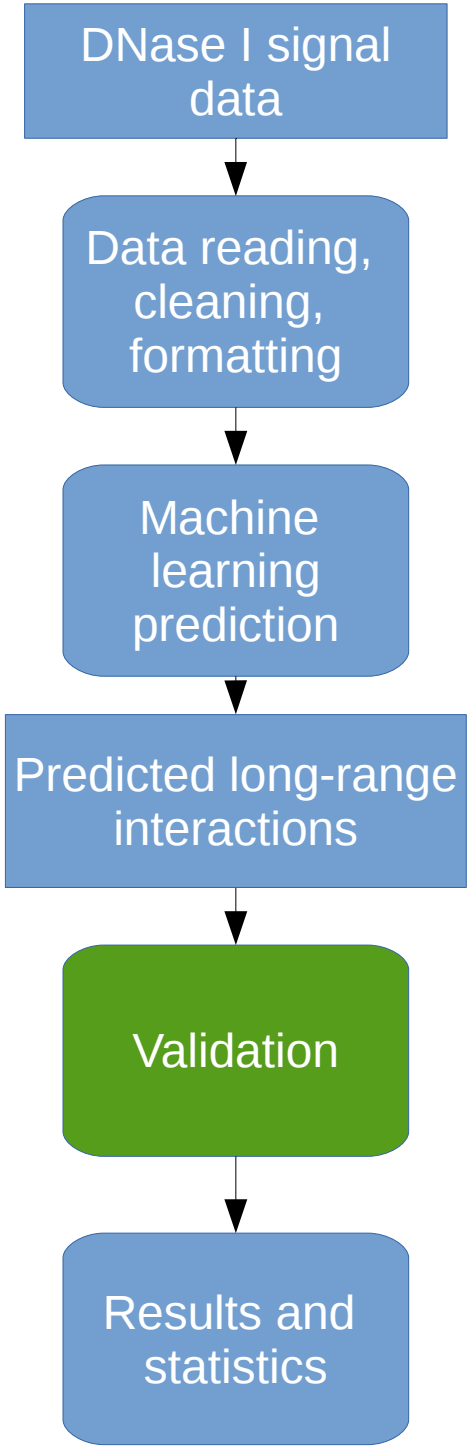
We enhanced the algorithm with momentum, dropout, Xavier initialization, and minibatches (size = 20)

Database in PostgreSQL
Software in Torch
Parallelized on Sun Grid Engine (SGE)

- training set
- validation set
- testing set

All their elements are selected through all the chromosome regions





Training & validation

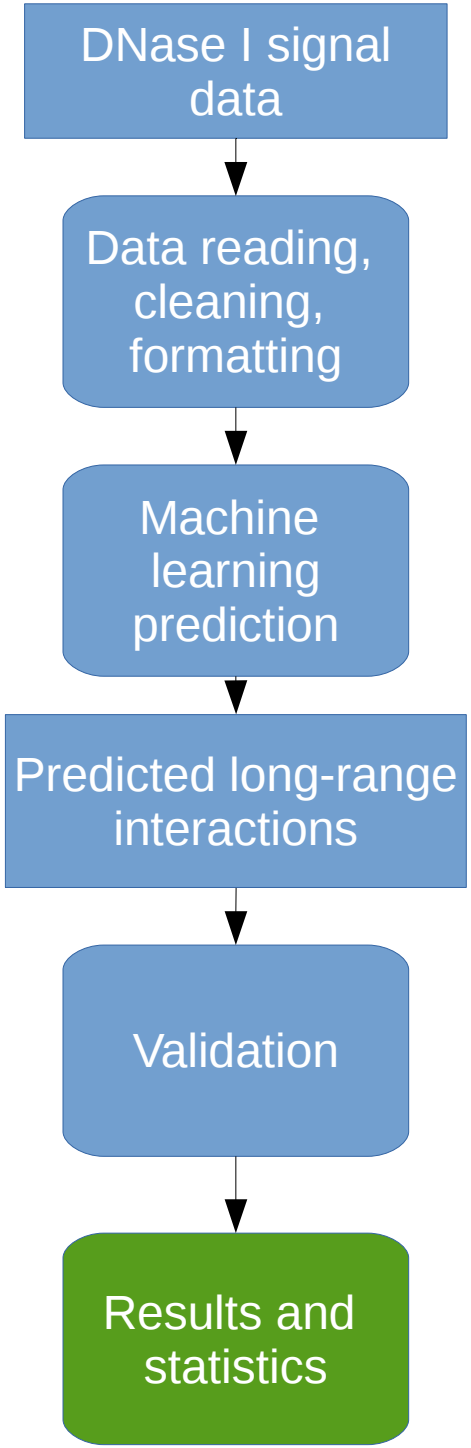
After the training (on each dataset fold or on the training set), the script tests the trained model on the left over test set.

In the test, we compute the Matthews correlation coefficient (MCC), instead of the receiver operating characteristic curve (ROC) area under the curve (AUC).

MCC is a balanced measure that takes into account the different sizes of the classification classes

$$MCC = \frac{TP * TN - FP * TN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

We use a prediction threshold $\tau \geq 0.5$, that corresponds to 0 in the cosine distance, where the [–1,0] interval means false, while the (0,+1] interval means true

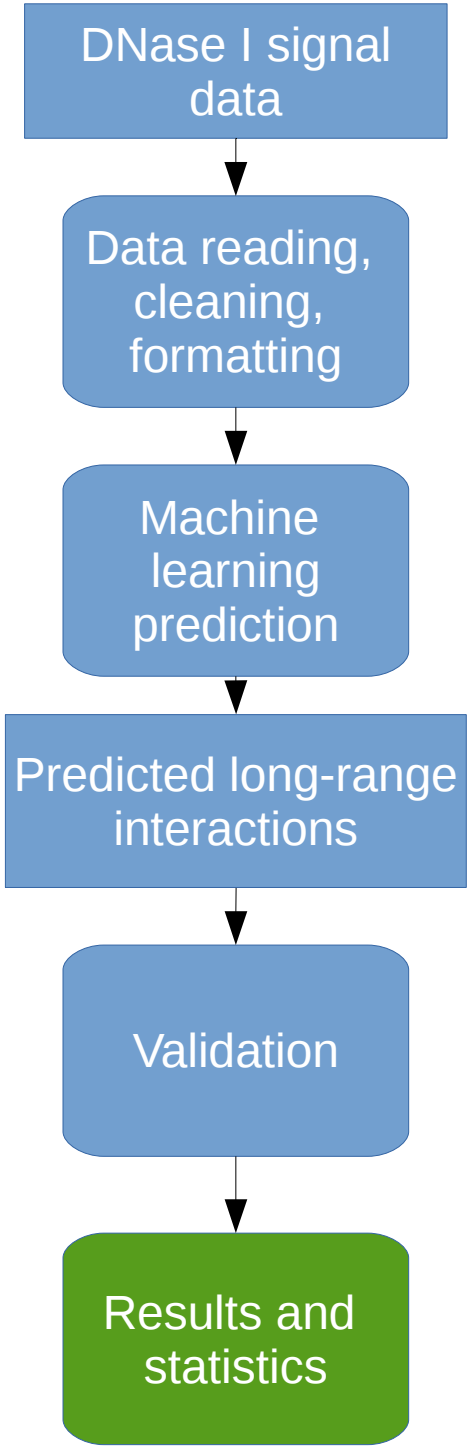


Validation set results

We were able to produce some preliminary results:

- balanced dataset for training: 20,000 elements (10,000 negatives and 10,000 positives)
- validation set: 2,000 elements
- model: 400 hidden units and 1 hidden layer (tests ongoing)

dataset	MCC	dataset	MCC
chr1	+0.16	chr13	+0.39
chr2	+0.23	chr14	+0.36
chr3	+0.25	chr15	+0.29
chr4	+0.31	chr16	+0.41
chr5	+0.25	chr17	+0.28
chr6	+0.21	chr18	+0.48
chr7	+0.31	chr19	+0.36
chr8	+0.25	chr20	+0.45
chr9	+0.24	chr21	+0.59
chr10	+0.26	chr22	+0.45
chr11	+0.27	chrX	+0.39
chr12	+0.29		



Future directions

We were able to obtaion some preliminary good results, and we're considering other approaches to enhance our algorithm:

- boosting technique to manage the imbalance of the datasets
- regularization technique to make the gradient update more stable in the neural network
- nested cross-validation

On the biological side, we want then to be able to make some cell-specific prediction (single-column)
- feature selection (Random Forests?)

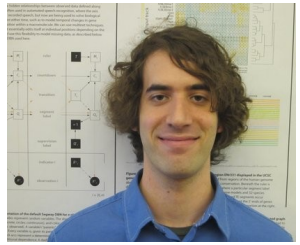
Any suggestion here is very appreciated!

Future goal

- We want to finally produce a significant comparison between the existing interactions found by Rao, Huntley et al, *Cell*, 2014, Thurman et al, *Nature*, 2012 and our method

	Thurman et al., *Nature*, 2012	Rao, Huntley et al., *Cell*, 2014	Our siamese neural network
interaction #1	yes / no	yes / no	yes / no
...	...	...	...
...	...	...	...
interaction #N	yes / no	yes / no	yes / no

The end: acknowledgments



- Michael M. Hoffman
  Princess Margaret Cancer Centre, University of Toronto

- Richard Zemel
  Department of Computer Science, University of Toronto

- Alexander Schwing
  Department of Computer Science, University of Toronto

- Coby Viner
  Department of Computer Science, University of Toronto

We're hiring!

PhD students, postdocs, research assistants, etc

www.HoffmanLab.org

Princess Margaret Cancer Centre
University Health Network
UHN

UNIVERSITY OF TORONTO