Albert-Ludwigs University Freiburg

Department of Computer Science

Bioinformatics Group

Master Thesis

Predicting Hi-C contact matrices using machine learning approaches

Author:
Ralf Krauth

Examiner:
Prof. Dr. Rolf Backofen

Second Examiner:
Prof. Dr. Ralf Gilsbach

Advisors:
Anup Kumar, Joachim Wolff

Submission date:
20.04.2021

Abstract

Harhar!

Zusammenfassung

Hohoho!

Declaration / Erklärung

Hiermit erkläre ich, dass ich die vorliegende Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe und alle Stellen, die wörtlich oder sinngemäß aus anderen Werken entnommen wurden, als solche kenntlich gemacht habe. Darüber hinaus erkläre ich, dass die eingereichte Masterarbeit weder vollständig, noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens war oder ist.

Contents

1 Introduction **6**

1.1 Hi-C process and contact matrices 6

1.2 Goal of the thesis 6

1.3 Structure of the thesis 6

2 Related work **7**

2.1 Methods for predicting DNA-DNA interactions and contact matrices 7

2.2 Generative techniques from computer vision 9

3 Advancing predictions of Hi-C interaction matrices **10**

3.1 Deep Neural Network approach 10

3.1.1 Basic network setup 10

3.1.2 Modifying the convolutional part of the network 10

3.1.3 Modifying the loss function 10

3.1.4 Modifying resolution and windowsize 10

3.1.5 DNA sequence as an additional network input branch 10

3.2 HiC-GAN approach 10

3.2.1 Pix2Pix as a generic generative network 10

3.2.2 Using a DNN for 1D - 2D conversion 10

3.2.3 Using a CNN for 1D - 2D conversion 10

4 Method details **11**

4.1 Deep Neural Network approach 11

4.1.1 Modifying kernel size, number of filter layers and filters 11

4.1.2 Modifying input resolution 11

4.1.3 Custom loss function based on TAD insulation score 11

4.1.4 Combination of mean squared error, perception loss and TV loss 11

4.2 HiC-GAN approach 11

4.2.1 Using a DNN for 1D-2D conversion 11

4.2.2 Using a pre-trained DNN for 1D-2D conversion 11

4.2.3 Using a CNN for 1D-2D conversion 11

5 Results **12**

5.1 Deep Neural Network approaches 12

5.2 HiC-GAN approaches 12

6 Discussion and Outlook **13**

References **15**

Acronyms **17**

1 Introduction

In recent years, the three-dimensional organization of DNA has been shown to be a key driver of important processes in molecular biology. However, even with the most advanced experimental methods like Hi-C, it remains comparatively expensive to study the spatial folding of DNA, so that current knowledge of three-dimensional DNA organization is still sketchy. In the last five years, several methods have been proposed to improve on this situation by determining DNA-DNA interactions *in-silico*, using existing experimental data.

However, most current in-silico approaches leave room for improvement, for example with regard to usability, the necessary type and amount of data or the pre- and postprocessing steps required. The goal of this master thesis is thus to provide an approach for the prediction of Hi-C interaction matrices which is easy to operate and requires only data from ChIP-seq experiments with minimal pre- and postprocessing.

1.1 Hi-C process and contact matrices

1.2 Goal of the thesis

1.3 Structure of the thesis

2 Related work

In the last five years, several approaches have been presented to determine DNA-DNA interactions *in silico*, using existing data from various experiments. subsection 2.1 gives an overview about these methods. Furthermore, some methods originally developed for image synthesis and similar tasks in computer vision might also be useful in the field of Hi-C matrix generation and are thus summarized in subsection 2.2.

2.1 Methods for predicting DNA-DNA interactions and contact matrices

As of 2020, there is quite a body of existing work in the field of predicting DNA-DNA interactions, using various approaches and different types of input data.

Two conceptually similar methods have been proposed by Brackley et al. in 2016 and MacPherson et al. in 2018 [1, 2]. In both approaches, DNA is modeled as a “beads-on-a-string” polymer, and simulation techniques are employed to find energy-optimal spatial structures of these polymers. Apart from constraints derived from the molecule’s DNA sequence itself, the models also consider spatial contact constraints derived from ChIP-seq experiments of chromatin factors which are known to mediate such DNA-DNA contacts. The interaction matrices derived from the simulations look interesting, but the paper from Brackley et al. [1] is unfortunately lacking a comparison with “true” experimentally measured Hi-C matrices, and the results from MacPherson et al. [2] seem inferior to most other ones presented in this section.

Another simulation-based method has been developed by di Pierro et al. in 2017 [3] and later extended by Qi and Zhang [4]. In both cases, a convolutional neural network (CNN) is trained to learn different “open” and “closed” chromatin states from 11 chromatin factors, and the predicted chromatin states are then taken as constraints for beads-on-a-string models. The difference between [3] and [4] lies mainly in the number of states considered and the simulation methods applied; the results are mathematically convincing in both cases.

A further approach using chromatin states is due to Farrè and Emberly [5]. Here, the conditional probability of two genomic regions being in contact, given their distance and the chromatin state around them, is estimated using Bayes’ rule. In this case, the chromatin state – reduced to active or inactive – is derived from DNA adenine methyltransferase identification (DamID) signals of 53 chromatin factors using probabilistic methods [6]. The conditional probabilities on the right side of Bayes’ rule are either computed from training data or estimated with different probabilistic approaches, too. While the predicted contact matrices do not look like real Hi-C matrices with this approach, highly interacting regions are still often well identifiable.

Three further approaches in the field make use of random forests. 3DEpiLoop by Bkhetan and Plewczynski and Lollipop by Kai et al., both from 2018, use a random forest classifier to predict DNA loops, but differ in input data and preprocessing [7, 8]. While 3DEpiLoop is using only ChIP-seq data of histone modifications and transcription factors [7], Lollipop additionally takes ChIA-PET-, RNA-seq- and DNase-seq-data, CTCF motif orientation and loop length as inputs [8]. Both approaches show good coincidence of predicted loops with experimental data, but their output is binary and rather sparse. Contrary to these two, the third random-forest-based approach, HiC-Reg by Zhang et al. from 2019, allows predicting real-valued Hi-C interaction matrices directly [9]. To this end, it employs random forest regression to predict interactions between two genomic “windows”, using ChIP-seq data of 13 chromatin factors and the genomic distance of the two windows. The published results for five human cell lines look interesting, but could not be reproduced in two study projects at the university of Freiburg for unknown reasons [10, 11]

Another recent method which investigated decision-tree-based algorithms is due to Martens et al. (2020) [12]. Here, gradient boosted decision trees, logistic regression and neural networks were used to predict highly interacting chromatin regions and TAD domain boundaries from histone modifications and CTCF ChIP-seq data. However, in this setup, the neural network approach yielded the best, overall acceptable results of the three approaches, but again in form of a binary classifications. A neural-network approach with comparable input data, but without the restriction on binary classifications has been presented by Farrè et al. in 2018 [13]. Here, a one-dimensional convolutional filter is used to convert ChIP-seq data from 50 chromatin factors into a one-dimensional chromatin vector, which is then processed by a dense neural network (DNN). This allows predicting real-valued Hi-C interaction matrices, which resemble the general structure of experimentally derived matrices quite well.

While the approaches discussed so far have either modeled DNA as a “beads-on-a-string” polymer or not used it explicitly at all, there are also several machine-learning approaches which directly consider DNA sequence without a need for polymer modeling. In 2019, Singh et al. presented SPEID [14], an approach to predict promoter-enhancer interactions from DNA sequence, using a combination of CNNs, a recurrent network (LSTM) and a DNN. The results match well with experimental data, but are limited to promoter and enhancer loci by design, disallowing predictions of complete Hi-C contact matrices. Other researchers have tried to design similar methods without such limitations. The work by Peng from 2017 [15] is an extension of SPEID, based on a 2016 preprint [16], additionally taking into account a “middle sequence” between enhancer- and promoter sequences, CTCF motif counts within the sequences and genomic distance between two sequence snippets. However, the network lacks generalization, i. e. the results are only good in training regions [15, figs. 4, 5]. A conceptually similar method to the one by Peng [15], but with a different neural network design has been presented by Schreiber et al., also in 2017, named Rambutan [17]. It accepts sequence, DNase-seq data and distance between two genomic loci as inputs and then uses a combination of CNNs and a DNN to predict whether the given two loci interact or not. Unfortunately, it is difficult to decide whether the results of Schreiber et al. are useful in practice, since the evaluation is done only by statistical means and no actual Hi-C matrices have been published. The original paper [17] also contains a known error and seems not to have appeared in a peer-reviewed journal in improved form yet. A probably more promising method working on DNA sequence, Akita, has been published by Fudenberg et al. in 2020 [18]. It is based on two rather involved convolutional neural networks. While the first one, “trunk”, processes one-dimensional, one-hot encoded DNA sequence input through convolutional filters, the second one, “head”, converts one-dimensional representations to 2D, further processes the data with convolutional filters and enforces symmetry. Although Fudenberg et al. initally seemed to focus on determining the influence of DNA modifications on spatial structure [19], predicting complete Hi-C matrices is an integral part of their work, and a large number of images of Hi-C matrices from the test set has been published alongside the article. The predicted matrices often hardly look like Hi-C matrices, but mostly indicate highly interacting regions quite well.

A further method by Schwessinger et al. [20] also makes use of DNA sequence and additional epigenetic data for its predictions, but is conceptually different from the ones presented so far. Here, ChIP-seq tracks are initially used to train a CNN on the relationship between sequence and the corresponding chromatin factors. The weights of this first network are then used to seed another convolutional neural network, which is responsible for predicting DNA-DNA interactions from DNA sequence. In this case, the results were blurry, but generally in good accordance with Hi-C experimental data. However, the neural network is large and thus requires long training times at small batch sizes.

Yet another machine-learning concept based on sequence data has been proposed by Nikumbh and Pfeifer in 2017 [21]. Here, a support vector machine is trained on 5C data, using a specific representation for DNA sequence called oligomer distance histograms. The results showed acceptable consensus with experimental Hi-C data and allowed some interesting conclusions about the importance of certain k-mers for DNA folding.

2.2 Generative techniques from computer vision

3 Advancing predictions of Hi-C interaction matrices

3.1 Deep Neural Network approach

3.1.1 Basic network setup

3.1.2 Modifying the convolutional part of the network

3.1.3 Modifying the loss function

3.1.4 Modifying resolution and windowsize

3.1.5 DNA sequence as an additional network input branch

3.2 HiC-GAN approach

3.2.1 Pix2Pix as a generic generative network

3.2.2 Using a DNN for 1D - 2D conversion

3.2.3 Using a CNN for 1D - 2D conversion

4 Method details

4.1 Deep Neural Network approach

4.1.1 Modifying kernel size, number of filter layers and filters

4.1.2 Modifying input resolution

4.1.3 Custom loss function based on TAD insulation score

4.1.4 Combination of mean squared error, perception loss and TV loss

4.2 HiC-GAN approach

4.2.1 Using a DNN for 1D-2D conversion

4.2.2 Using a pre-trained DNN for 1D-2D conversion

4.2.3 Using a CNN for 1D-2D conversion

5 Results

5.1 Deep Neural Network approaches

5.2 HiC-GAN approaches

6 Discussion and Outlook

References

[1] Chris A. Brackley et al. “Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models”. In: *Genome Biology* 17.1 (Mar. 2016). DOI: 10.1186/s13059-016-0909-0.

[2] Quinn MacPherson, Bruno Beltran and Andrew J. Spakowitz. “Bottom–up modeling of chromatin segregation due to epigenetic modifications”. In: *Proceedings of the National Academy of Sciences* 115.50 (Nov. 2018), pp. 12739–12744. DOI: 10.1073/pnas.1812268115.

[3] Michele Di Pierro et al. “De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture”. In: *Proceedings of the National Academy of Sciences* 114.46 (Oct. 2017), pp. 12126–12131. DOI: 10.1073/pnas.1714980114.

[4] Yifeng Qi and Bin Zhang. “Predicting three-dimensional genome organization with chromatin states”. In: *PLOS Computational Biology* 15.6 (June 2019). Ed. by Jian Ma, e1007024. DOI: 10.1371/journal.pcbi.1007024.

[5] Pau Farré and Eldon Emberly. “A maximum-entropy model for predicting chromatin contacts”. In: *PLOS Computational Biology* 14.2 (Feb. 2018). Ed. by Alexandre V. Morozov, e1005956. DOI: 10.1371/journal.pcbi.1005956.

[6] Jian Zhou and Olga G. Troyanskaya. “Probabilistic modelling of chromatin code landscape reveals functional diversity of enhancer-like chromatin states”. In: *Nature Communications* 7.1 (Feb. 2016). DOI: 10.1038/ncomms10528.

[7] Ziad Al Bkhetan and Dariusz Plewczynski. “Three-dimensional Epigenome Statistical Model: Genome-wide Chromatin Looping Prediction”. In: *Scientific Reports* 8.1 (Mar. 2018). DOI: 10.1038/s41598-018-23276-8.

[8] Yan Kai et al. “Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features”. In: *Nature Communications* 9.1 (Oct. 2018). DOI: 10.1038/s41467-018-06664-6.

[9] Shilu Zhang et al. “In silico prediction of high-resolution Hi-C interaction matrices”. In: *Nature Communications* 10.1 (Dec. 2019). DOI: 10.1038/s41467-019-13423-8.

[10] Ralf Krauth. *Improving predictions of Hi-C matrices from ChIP-seq data.* Tech. rep. Albert-Ludwigs Universität Freiburg, 2020. URL: https://github.com/MasterprojectRK/HiCPrediction.

[11] Andre Bajorat. *Hi-C Predictions based on protein levels.* Tech. rep. Albert-Ludwigs Universität Freiburg, 2019. URL: https://www.bioinf.uni-freiburg.de/Lehre/Theses/TP_Andre_Bajorat.pdf.

[12] Laura D. Martens et al. “Identifying regulatory and spatial genomic architectural elements using cell type independent machine and deep learning models”. In: (Apr. 2020). DOI: 10.1101/2020.04.19.049585.

[13] Pau Farré et al. “Dense neural networks for predicting chromatin conformation”. In: *BMC Bioinformatics* 19.1 (Oct. 2018). DOI: 10.1186/s12859-018-2286-z.

[14] Shashank Singh et al. “Predicting enhancer-promoter interaction from genomic sequence with deep neural networks”. In: *Quantitative Biology* 7.2 (June 2019), pp. 122–137. DOI: 10.1007/s40484-019-0154-0.

[15] Rui Peng. *Predicting High-order Chromatin Interactions from Human Genomic Sequence using Deep Neural Networks*. Tech. rep. Carnegie Mellon University, 2017. URL: https://www.ml.cmu.edu/research/dap-papers/F17/dap-peng-rui.pdf.

[16] Shashank Singh et al. "Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks". In: *bioRxiv* (Nov. 2016). DOI: 10.1101/085241.

[17] Jacob Schreiber et al. "Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture". In: *bioRxiv* (Jan. 2017). DOI: 10.1101/103614.

[18] Geoff Fudenberg, David R. Kelley and Katherine S. Pollard. "Predicting 3D genome folding from DNA sequence with Akita". In: *Nature Methods* 17.11 (Oct. 2020), pp. 1111–1117. DOI: 10.1038/s41592-020-0958-x.

[19] Geoff Fudenberg, David R. Kelley and Katherine S. Pollard. "Predicting 3D genome folding from DNA sequence". In: *bioRxiv* (Oct. 2019). DOI: 10.1101/800060.

[20] Ron Schwessinger et al. "DeepC: Predicting chromatin interactions using megabase scaled deep neural networks and transfer learning". In: *bioRxiv* (Aug. 2019). DOI: 10.1101/724005.

[21] Sarvesh Nikumbh and Nico Pfeifer. "Genetic sequence-based prediction of long-range chromatin interactions suggests a potential role of short tandem repeat sequences in genome organization". In: *BMC Bioinformatics* 18.1 (Apr. 2017). DOI: 10.1186/s12859-017-1624-x.

Acronyms

ChIA-PET chromatin interaction analysis by paired-end tag sequencing.

ChIP-seq chromatin immunoprecipitation followed by sequencing.

CNN convolutional neural network.

DamID DNA adenine methyltransferase identification.

DNN dense neural network.

LSTM long-short-term memory.

TAD topologically associating domain.