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

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.

Conceptually probably the most simple method in this regard is HiC-Reg, proposed by Zhang et. al in 2019 [7]. It is a machine learning approach, using random forest regression to predict interactions between two genomic “windows”, using ChIP-seq data of 13 chromatin features, i. e. transcription factors and histone modifications, as well as DNase-seq data 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 [8, 9].

Another machine-learning approach for direct prediction of interaction matrices was presented by Farrè et al. [10] in 2018. 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) to predict interaction matrices. The weights of the network are trained on experimental, distance-normalized Hi-C matrices. The published results for drosophila melanogaster chromosomes look again interesting, since the predicted matrices resemble the general structure of real matrices quite well.

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] 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.

[8] 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.

[9] 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.

[10] 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.

Acronyms

ChIP-seq chromatin immunoprecipitation followed by sequencing.

CNN convolutional neural network.

DamID DNA adenine methyltransferase identification.

DNN dense neural network.