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

1.1 Hi-C process and contact matrices

1.2 Goal of the thesis

1.3 Structure of the thesis

2 Related work

The following sections will give an overwiew on existing work dealing with the prediction of DNA-DNA interactions, in particular the prediction of Hi-C contact matrices.

Additionally,

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 from various experimental data, using several different techniques. Most methods make use of known correlations between certain chromatin factors and DNA looping or between chromatin factors and topologically associating domain (TAD) boundaries, as shall be outlined below.

One method to predict DNA-DNA interactions is HiC-Reg, which has been proposed by Zhang et. al in 2019 [1]. It is using random forest regression to predict interaction matrices using ChIP-seq data of 14 transcription factors and histone modifications as well as DNase-seq data and genomic distance. The published results for five human cell lines are interesting, but could not be reproduced in two study projects at the university of Freiburg [2, 3].

2.2 Generative machine learning techniques

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] 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`.

[2] 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`.

[3] 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`.

Acronyms

ChIP-seq chromatin immunoprecipitation followed by sequencing.

TAD topologically associating domain.