Proc. Annual Meeting of the Austrian Society for Biomedical Engineering 2021
DOI: 10.3217/978-3-85125-826-4-26

Purely Sequence based prediction of contact maps and classification of chromosomal compartments with DDA-DNA

X. F. D. Lainscsek[1] and L. Taher[1]

[1]Institute of Biomedical Informatics, Graz University of Technology, Austria

xenia.lainscsek@tugraz.at

Abstract *- The emergence of genome wide chromosome conformation capturing techniques such as HiC has enabled researchers to investigate the crucial role of chromatin folding in gene regulation. DNA folding forms distinct multiscale patterns which become visible in contact maps generated by such experiments. The abundance of information about chromatin architecture contained in the nucleotide sequence alone is still not well understood. Here we present a purely sequence based computational approach DDA-DNA that sifts out the sequence dependencies of genome architecture at 1Mb resolution.*

Keywords— *Hi-C, A/B Compartments, delay differential analysis, nonlinear dynamics*

Introduction

Advances in high-throughput chromosome conformation capture assays such as Hi-C has enabled the cataloging of genome-wide interaction maps in various cell types. How strongly DNA sequence signatures alone and at which scales they reflect this hierarchical organization remains unknown. The genome has various levels of organization. At megabase (Mb) resolution, chromosomes are organized into two types of chromatin called A and B compartments which correspond to open and closed chromatin [1]. The A/B compartments have been found to be cell-type specific and contribute to cell-type-specific gene expression [2].

Here we applied Delay Differential Analysis (DDA) [3] to extract dynamical properties of the DNA sequence that contribute to its conformation in 3D space. This method has been shown to achieve excellent classification and prediction performance in various data types [4, 5, 6]. The key difference to machine learning (ML) is that DDA uses a sparse feature set of only 4 terms compared to the typically huge parameter sets in ML. DDA does not utilize a typical training/testing approach, but rather a structure selection step where the model and the two fixed parameters that best represent the overall dynamics of the system are searched for. This makes DDA robust to overfitting and easily generalizable to new data [3].

We hypothesize that a substantial contribution of chromatin organization, at least at the 1Mb scale, arises from the sequence itself.

Methods

Background

Genome-wide chromosome conformation capturing techniques (Hi-C) is a type of next generation sequencing (NGS) method which produce contact frequency maps that depict the degree of interaction between two loci in the genome. The contact matrix is highly self-similar, a hallmark for a chaotic process, and can thus be understood as a recurrence map. It has been found that the contact frequency between two genomic regions i and j follows the power scaling law as

$$p(|i-j|) \sim |i-j|^{-\beta} \quad (1)$$

The scaling exponent β has been typically found to be slightly below 1. This is in good agreement with the predictions made by the so called “fractal” globule model of DNA [1, 7, 8].

Construction of contact maps from HiC-assay data

The Hi-C contact maps were derived using the publicly available Hi-C raw sequencing data set of a fetal lung fibroblast cell-line (GEO accession GSM862724). The reads were mapped using bowtie2 [9] and contact maps were generated with hicexplorer [10].

Construction of contact maps from nucleotide sequence

The DDA-contact maps were generated in a three-step process: 1) conversion of the nucleotide sequence into a numerically differentiable signal; 2) structure selection on chr14; and 3) testing on chr13,15,16, 17. A crucial fact of DDA ist that its functional form and parameters are never updated as in traditional ML methods. DDA does not learn but rather captures the important underlying macroscopic features of a dynamical system. A DDA model associates the numerical derivative of a signal, in this case a spatial sequence, with its delayed versions [3]. We used a cubic 3-term DDA model with two parameters τ_1 and τ_2

$$\dot{x}(t) = a_1 x_{\tau_1}^2 + a_2 x_{\tau_1}^2 x_{\tau_2} + a_3 x_{\tau_2}^3 + \rho \quad (2)$$

where $x_{\tau_i} = x(t-\tau_i)$ is the converted version of the DNA sequence delayed by τ_i. The coefficients a_1, a_2, a_3 and the least squares error ρ are estimated from the over-determined system of

Published by Verlag der TU Graz
Graz University of Technology

DOI: 10.3217/978-3-85125-826-4-26

equations with singular value decomposition (SVD) [11] and used as classifying features.

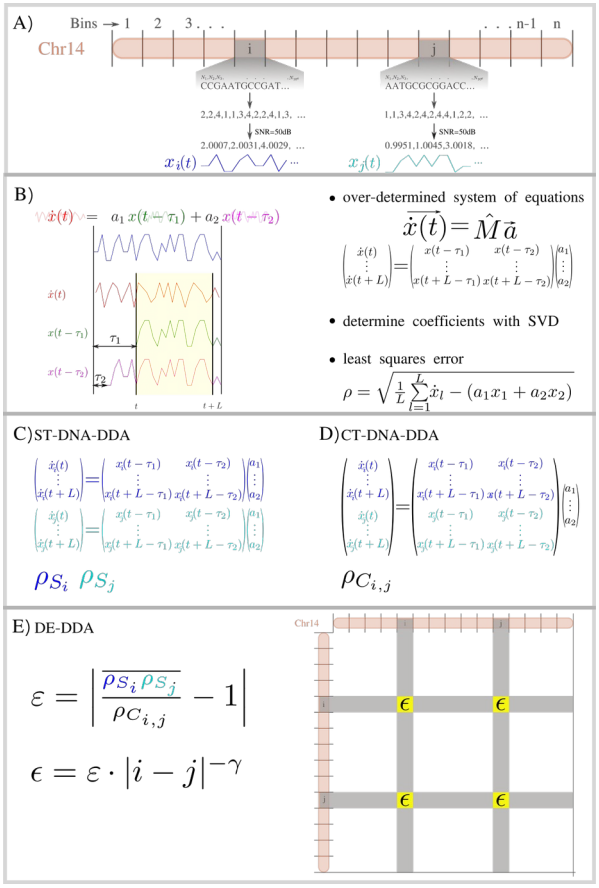

Figure 1: A) conversion of the nucleotide sequence B) DDA (example for a simple linear 2-term DDA model) C) ST DDA D) CT DDA E) DE-DDA

1) conversion of the nucleotide sequence

For DNA-DDA (see Fig. 1 B)), each nucleotide A, C, T and G in the sequence of the human genome (hg38 assembly) was encoded as 1, 2, 3 and 4 respectively, and the entire sequence was split into 1Mb long bins. To make the signal suitable for DDA, a small amount of signal-to-noise ratio (SNR) of 50 dB Gaussian noise was overlain to the signal of each bin $x_n = s + cN$ where s is the encoded DNA sequence, N are numbers drawn from the standard normal distribution and $c = \sqrt{\mathrm{var}(s) \cdot 10^{-(SNR/10)}}$.

2) structure selection

The converted DNA signal x_n of each bin was inputted into Eq. (2) and the features a_1, a_2, a_3 and ρ were calculated for each delay pair (τ_1, τ_2) in a probe list consisting of 870 delays between 1 and 30.

The individual calculation of DDA features is called single-trial(ST) DDA. Data windows of multiple time series can be combined in cross-trial(CT) DDA where features are computed simultaneously by including them in the over-determined system of equations given by Eq. (2) [6, 12].

Each feature may be considered separately or combined. For instance for two 1Mb genomic regions i and j, we can compute the ST DDA features ρ_{S_i}, ρ_{S_j} as well as the CT DDA feature $\rho_{C_{i,j}}$. The CT errors and mean of the ST errors should be similar if the analyzed time series have similar dynamics and their quotient will be close to one.

The dynamical ergodicity (DE) DDA ε [12] is defined by the quotient

$$\varepsilon = \left| \frac{\overline{\rho_{S_i} \rho_{S_j}}}{\rho_{C_{i,j}}} - 1 \right| \quad (3)$$

Thus the lower ε, the more dynamically similar these two 1Mb stretches of sequence are to one another. Motivation of this feature comes from ergodic theory [13] which is concerned with the statistical properties of a dynamical system.

We hypothesize that DE DDA is correlated with the proximity of two 1Mb stretches of sequence in 3D space. Hence, we predict the contact probability between two DNA sequences i and j as

$$\epsilon = \varepsilon |i - j|^{-\gamma} \quad (4)$$

Where $|i - j|$ is the distance between genomic bin i and bin j and γ is the scaling exponent and was set to -0.77.

Calling A/B Compartments

We generated the pearson correlation matrices from the contact matrices to call A/B compartments. The HiC- and DDA-contact maps (Fig. 2 A) were normalized with Toeplitz normalization using the 4D nucleome Analysis toolbox [14] before being converted to a correlation matrix as described by [1]. The principal components which determine A/B compartments were derived using matlabs [15] pca function. The A and B compartments correspond arbitrarily to PC>0 or PC<0 respectively.

3) testing

Model performance was assessed with the stratum-adjusted correlation coefficient (SCC) given by HiCRep [19], the mean square error (MSE), Pearson's R of the resulting first or second PCs of the Pearson correlation matrices (r_{PC}), and lastly the area under the ROC curve of the compartment classification (AUC). Before calculating testing measures, both matrices were normalized to 0 and 1 whilst ignoring the main diagonal which was subsequently set to 1.

It is worth noting that the ordinary Pearson correlation coefficient is not sufficient for comparison of matrices of such type. The SCC statistic [19] takes spatial features such as domain structure and distance dependence into account. An averaging filter of size 2 was applied to HiC and

Published by Verlag der TU Graz
Graz University of Technology

DOI: 10.3217/978-3-85125-826-4-26

DDA maps using matlabs fspecial function before calculating the SCC (smoothing parameter=0) .

Results

The interaction probability between two genomic regions is not simply a matter of linear distance (Fig. 2A). Each chromosome has a unique and characteristic structure. Chr14 was arbitrarily chosen for the structure selection process. Based on the aforementioned performance measures the feature/delay pair combination that resulted in the highest performance was found to be ρ and the mean of the DDA-maps for $\tau_{1,2} = (5,3);(26,12)$. We tested this model-parameter combination on chr13,15, 16, 17 on which DNA-DDA shows promising performances (Tab. 1).

Table1: Performance of DNA-DDA for delays $\tau_{1,2} = (5,3);(26,12)$ on chr13-17

ChrNr	SCC	r_{PC}	AUC	MSE
13	0.74	0.78	0.84	0.06
14	0.62	0.74	0.80	0.02
15	0.68	0.61	0.78	0.06
16	0.74	0.71	0.82	0.06
17	0.73	0.71	0.82	0.06

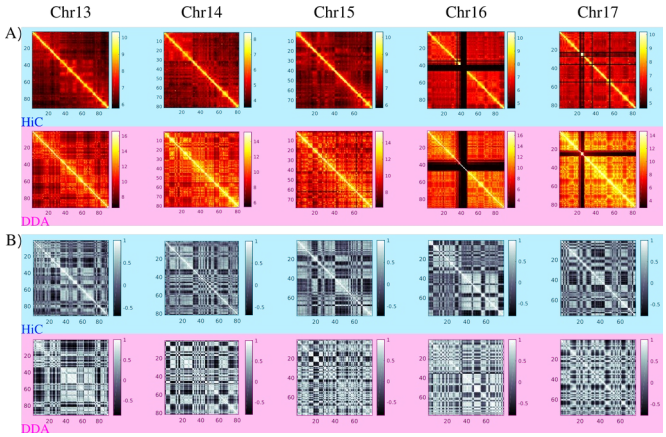

Figure 2: A) log-transformed HiC- and DDA-contact maps (ϵ^{-1}) for chromosomes 13-17. One value in the map corresponds to a genomic region of 1Mbp. B) Corresponding Pearson correlation HiC- and DDA-maps for chromosomes 13-17.

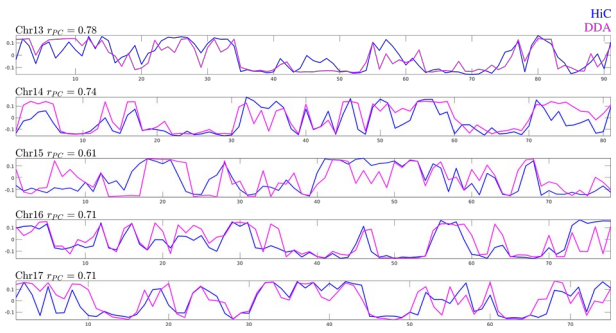

Figure 3: PCs of HiC (blue) and DDA (magenta) correlation maps (Fig 2 B))

Highly interacting regions predicted by DNA-DDA match very well those of the HiC-contact maps achieving a mean SCC of 0.72 on hold out chromosome contact maps (Fig. 2B). DNA-DDA-based compartment prediction was also remarkably accurate (Fig. 3) achieving a mean AUC of 0.82 on hold out chromosomes.

Discussion

We present DDA-DNA, a novel method based in nonlinear dynamics and ergodic theory that predicts the folding of chromatin inside the nucleus using the nucleotide sequence alone. Being able to infer structural changes in the genome could immensly aid in understanding disease pathology and be of clinical use in the long run. Hierarchial organization of the genome is crucial for nuclear activity such as transcription, DNA replication as well as for cellular processes and development.

Current approaches for modeling genome organization are based on machine learning or polymer chemistry and physics. The former typically rely on epigenomic information as input (eg [2, 16]), which are not able to model the effects of genetic variation. However recently, some deep learning sequence based multi-scale models including DeepC, Akita, and Orca for chromatin architecture have emerged [17, 18, 20]. DeepC is a transfer-learning based neural network which like Akita, predict interactions within Mb-scale loci. The training/testing and validation sets were split based on chromosomes. DeepC uses two training procedures the first of which used chr11 and chr12 for validation and chr15-17 for testing, the second used tha same validation chromosomes but only chr16 and chr17 for testing. GPU support was needed for training and the final models had ~60M parameters. DeepC models were trained on seven human and one mouse data sets and cross-validation accross all chromosomes acheived an average distance stratified Pearson's R of ~0.36 on raw skeleton data and ~0.57 when applying a smoothing filter to the discrete and noisy skeleton [17]. Akita uses a convolutional neural network that predicts interaction contacts up to 1Mbp. They divided the human genome into ~1Mb sequences and used a 80/10/10 random split for training, testing and validation sets (~262kb in the training set and ~524kb for the validation and test sets). The resulting model has 746,149 trainable parameters. Training and prediction were conducted on 5 high quality Hi-C and Micro-C data sets and acheived performances of MSE=0.14 and distance stratified Pearson's R=0.61. Akita currently makes predictions for 1Mb long windows and will need to be extended to make prediction on more distal pairs of genomic loci to obtain chromatin features such as A/B compartmentalization [18]. Currently in preprint, Orca is the first sequence

DOI: 10.3217/978-3-85125-826-4-26

based model that predicts chromatin architecture from kp to whole-chromosome scale. The model takes 1Mb-256Mb as input and predicts interactions from 4kb-256Mb. On holdout test chromosomes 9 and 10, the model achieves a Pearson correlation of 0.78-0.84 and 0.72-0.79 consistently across all scales for the two micro-C datasets. Based upon additional analysis on sequence effects on A/B compartments, they proposed that compartment A formation is driven by TSS sequences whereas compartment B requires sequences of > 6-12kp without compartment A activity, is AT-enriched, and may be the “default” state established on all sequences not belonging to compartment A [20].

DNA-DDA’s performance measures do very well when compared with these recent publications achieving a mean of SCC=0.72, MSE=0.06, AUC=0.82 and r_{PC}=0.70 across the test chromosomes. What sets this method apart from the others, is the vastly lower number of parameters and distinguishing features. Opposed to other methods, we use merely one chromosome (chr14) to fix the DDA-model and parameters ($\tau_{1,2} = (5, 3); (26, 12)$) and subsequently test it on four others (chr13, 14, 15, 16, 17). Furthermore, the final fixed DNA-DDA can be computed on CPUs on new chromosomes in minutes (chr13: ~23 minutes on 6 AMD Ryzen 9 3950X CPUs). There remain many possibilities of adjusting our analysis such as: conversion of the sequence to a time series signal and using a different DDA functional form.

Additional analysis is still needed to assess DNA-DDA’s robustness on all remaining human chromosomes and on the sub-megabase scale.

We hypothesize that DNA-DDA has the potential to detect cell-type specific structural differences. DDA applied to other systems such as the human brain is able to classify various disease states defined by a certain delay pair and thus we hypothesize that also here, a different delay pair will best characterize DNA structure in another cell type.

We believe tha DNA-DDA has high potential in helping to understand the mechanisms by which derangement in the hierarchial architecture of the genome causes diesease pehotypes. Implementing DNA-DDA to perturbed sequences could help predict the effects of various genetic mutations. This is of particular interest for understanding disease progression such as in cancer. Similarly, removal of certain sequence motifs can give us insight into the highest contributing sequence signatures and biological mechanisms of genome folding at various scales.

Acknowledgments

We would like to express our gratitude to Dr. Claudia Lainscsek (Salk Institute) for the valuable discussions and insight into novel DDA methods.

References

[1] Lieberman-Aiden, Erez et al., Comprehensive mapping of long range interactions reveals folding principles of the human genome, 2009

[2] Fortin, Jean-Philippe and Hansen, Kasper, Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data, 2015

[3] Lainscsek, Claudia and Sejnowski, Terrence J., Delay differential analysis of time series., 2015

[4] Lainscsek, Claudia and Sejnowski, Terrence J., Electrocardiogram classification using delay differential equations, 2013

[5] Lainscsek, Claudia et al. , Delay Differential Analysis of Seizures in Multichannel Electrocorticography Data, 2017

[6] Lainscsek, Claudia et al., Nonlinear dynamics underlying sensory processing dysfunction in schizophrenia, 2019

[7] Grosberg, Alexander et al., The role of topological constraints in the kinetics of collapse of macromolecules, 1988

[8] Grosberg, Alexander et al., Crumpled Globule Model of the Three-Dimensional Structure of DNA, 1993

[9] Langmead, Ben and Salzberg, Steven L, Fast gapped-read alignment with Bowtie 2, 2012

[10] Wolff, Joachim et al. , Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization, 2020

[11] Golub, Gene H. and Reinsch, Christian, Singular value decomposition and least squares solutions, 1970

[12] Lainscsek, Claudia , Dynamical Ergodicity, Soon to be submitted for publication

[13] Boltzmann, Ludwig, Vorlesungen über Gastheorie , 1898

[14] Seaman, Laura and Rajapakse, Indika, 4D nucleome Analysis Toolbox: analysis of Hi-C data with abnormal karyotype and time series capabilities , 2018

[15] The MathWorks, MATLAB and Statistics Toolbox Release 2018a,

[16] Mourad, Raphaël and Cuvier, Olivier , Predicting the spatial organization of chromosomes using epigenetic data, 2015

[17] Schwessinger, Ron et al. , DeepC: predicting 3D genome folding using megabase-scale transfer learning, 2020

[18] Fudenberg, Geoff et al., Predicting 3D genome folding from DNA sequence with Akita, 2020

[19] Yang, Tao et al., HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient, 2017

[20] Zhou, Jian, Sequence-based modeling of genome 3D architecture from kilobase to chromosome-scale, 2021

Published by Verlag der TU Graz
Graz University of Technology