OXFORD

Briefings in Bioinformatics, 2023, **24(4)**, 1–12
https://doi.org/10.1093/bib/bbad198
Advance access publication date 1 June 2023
Problem Solving Protocol

Predicting chromosomal compartments directly from the nucleotide sequence with DNA-DDA

Xenia Lainscsek and Leila Taher

Corresponding author: Leila Taher, Institute of Biomedical Informatics, Graz University of Technology, Stremayrgasse16/I, 8010, Graz, Austria. Tel.: +43 316 873 - 5380; E-mail: leila.taher@tugraz.at

Abstract

Three-dimensional (3D) genome architecture is characterized by multi-scale patterns and plays an essential role in gene regulation. Chromatin conformation capturing experiments have revealed many properties underlying 3D genome architecture, such as the compartmentalization of chromatin based on transcriptional states. However, they are complex, costly and time consuming, and therefore only a limited number of cell types have been examined using these techniques. Increasing effort is being directed towards deriving computational methods that can predict chromatin conformation and associated structures. Here we present DNA-delay differential analysis (DDA), a purely sequence-based method based on chaos theory to predict genome-wide A and B compartments. We show that DNA-DDA models derived from a 20 Mb sequence are sufficient to predict genome wide compartmentalization at the scale of 100 kb in four different cell types. Although this is a proof-of-concept study, our method shows promise in elucidating the mechanisms responsible for genome folding as well as modeling the impact of genetic variation on 3D genome architecture and the processes regulated thereby.

Keywords: 3D genome architecture, chromosomal compartments, nonlinear dynamics, chaos theory, delay differential analysis, Hi-C

INTRODUCTION

Three-dimensional (3D) genome architecture allows linearly distal genomic loci to interact with one another, thereby impacting genome function. Chromosome conformation capturing techniques, in particular high-throughput chromosome conformation capture (Hi-C) [1, 2], have enabled us to systematically catalog genomic interactions and features of 3D genome architecture in various cell types.

Hi-C data are typically summarized in a contact map, a matrix that estimates the probability of interaction between any two loci in the genome. Such maps are characterized by a plaid pattern reflecting enrichment or depletion of Hi-C interactions. This was observed already by early Hi-C studies, which proposed to segregate the loci into two sets of compartments, and arbitrarily termed them “A” and “B” [1]. Loci in A compartments preferentially interact with other loci in A compartments, while loci in B compartments tend to interact only with other loci in B compartments. Additionally, loci in A compartments are associated with transcriptionally active euchromatin, are gene-rich, and are centrally located in the nucleus [3]. In contrast, loci in B compartments are in transcriptionally inactive heterochromatin, and tend to be gene-poor and occupy the nuclear periphery [3]. A and B compartments have been shown to be associated with distinct histone acetylation and methylation patterns that reflect their transcriptional activity, and more refined subcompartmentalizations have been suggested on the basis of the observed chromatin states [2–5]. Compartmentalization has been found to be evolutionary conserved across species [6–8]. Nevertheless, it can differ substantially between cell types [3–5] and sequence variation between individuals has also been shown to underlie changes in 3D genome architecture, in many cases with pathological consequences [9–11].

Hi-C assays have revolutionized our understanding of 3D genome architecture, but they are expensive, time-consuming and expertise-demanding. Therefore, Hi-C data are only available for a limited number of human cells, and substantial effort has been put into deriving predictive computational models. Here we present DNA-DDA, a computational method that is based on the principles of chaos, ergodic and embedding theory to predict A/B compartments from the DNA sequence alone.

Chaos is widespread in biological systems [12–14]. It manifests itself as the seemingly random behavior of a deterministic process which is hypersensitive to fluctuations in initial conditions [15]. A deterministic dynamical system can be described by its current *state* (i.e. the system variables’ current values) and a system of differential equations (i.e. *rules*) that govern the evolution of the system (i.e. the sequence of states it passes through) [16, 17]. The set of all possible states, i.e. the solutions to the system of differential equations for every possible set of initial conditions, is known as the system’s *state space* [18, 19]. A *trajectory* in the state space is a sequence of states resulting from a particular set of initial conditions. For most chaotic systems, there exists an ”attractor” [20, 21], i.e. a point or a set of points towards which trajectories from *almost any* set of initial conditions will approach, and that represents the long-term behavior—the dynamics—of the underlying system. Two initially infinitesimally close trajectories of such a system diverge exponentially and yet,

Xenia Lainscsek is a research assistant and doctoral candidate at the Institute of Biomedical Informatics at Graz University of Technology, Austria
Leila Taher is a professor of bioinformatics at the Institute of Biomedical Informatics at Graz University of Technology, Austria.
Received: November 16, 2022. **Revised:** April 18, 2023. **Accepted:** May 08, 2023

bounded by their attractors, they will be similar in a topological sense; i.e. they can be deformed into each other continuously by stretching and folding [22]. Due to this counterintuitive property of "deterministic chaos," the global structure of the state space can be investigated temporally, e.g. by studying the rate at which neighboring trajectories with similar initial conditions diverge as the system evolves [23], or spatially, e.g. by determining a trajectory's fractal dimension [24]. A bridge between these two perspectives is ergodic theory [25–27], which studies the statistical properties of a dynamical system. Trajectories of an ergodic system will eventually cover the entire state space so that under certain conditions, the time average of a function along a trajectory is related to the space average for almost all initial conditions [18].

The analysis of DNA sequences in the context of nonlinear dynamics (ergodic theory [28, 29] and chaos theory [14]), information theory [30, 31] and time series analysis (signal processing, spectral analysis [32–34]) has a long standing history and their concepts and ideas are interconnected [35]. These methods aim to gain insight into the macroscopic behavior of genomic control systems without having access to their innumerable variables and governing equations (*states* and *rules*). Variables underlying 3D genome architecture include, for example, the DNA sequence, histone modifications [36], DNA methylation [37] or the interaction of the DNA polymer with the surrounding environment [38]. Inspired by many of the aforementioned concepts and ideas, we adapted a nonlinear time series classification framework, delay differential analysis (DDA) [39], for the prediction of A/B compartments from the DNA sequence.

The fundament of DDA is given by Takens embedding theorem, which states that under certain conditions, the measurement of a single variable of a high dimensional dynamical system providing good or global observability of the system, is sufficient to reconstruct the system's state space [40–42]. DDA relates delay and derivative embeddings of a measured variable in a (nonlinear) functional form, and uses the fitting coefficients as classifying features. A particular flavor of DDA, dynamical ergodicity-DDA (DE-DDA) [43] is used for assessing dynamical similarity. We hypothesize that the DNA sequence and the interaction frequencies between genomic loci obtained from a Hi-C assay, are variables which are highly observable of 3D genome architecture, and that sequences in close proximity in 3D space will share certain dynamical properties. In accordance with the ergodic hypothesis [25], we compare the ensemble and time averages of dynamical information inherent in the numerical representation of the DNA sequences as described by DE-DDA, to infer their proximity in 3D space and, in turn, to predict A/B compartments.

A Hi-C map can be understood as a 2D projection of the n-dimensional state space of the system, i.e. a recurrence plot [44]. A recurrence plot is a method from nonlinear data analysis obtained by recording the instances t, when a trajectory visits the immediate proximity of a state it has visited in the past. Analogously, a Hi-C map visualizes how often the genomic locus at position t is involved in interactions with a subsequent locus in the DNA sequence (i.e. how often it is revisited). The patterns which arise in Hi-C maps are a hallmark for an underlying chaotic process, and DNA-DDA sifts out its dynamical signatures by mapping the DNA sequence onto an embedding space. DDA has been applied most extensively in epilepsy research [45–47] and our study confirms its potential to be extended to the field of genomics.

METHODS

Pre-processing of Hi-C data sequencing data

Raw FASTQ files from Hi-C experiments involving four cell lines were downloaded from the Gene Expression Omnibus database (Table 1; [2, 48, 51]) and mapped to the human reference genome (GRCh38/hg38) using `bowtie 2` (v.2.4.1; [53]) with options `–reorder` and `–very-sensitive-local`. The deepest sequenced data set which we considered was the "primary" GM12878 Hi-C data set comprising 3.6 billion sequence reads followed by the the K526 Hi-C data set with a library of 1.4 billion sequence reads. The older hESC and IMR90 data sets comprised 0.3 and 0.4 billion reads respectively.

Hi-C contact maps

The contact map of each autosome was generated from the mapped reads using the *HiCExplorer* pipeline (v.3.7.2; [54–56]). "`hicBuildMatrix`" was called with parameters "`–binSize 5000 –minMappingQuality 10 –restrictionSequence RS –danglingSequence DS,`" where `DS` and `RS` are the restriction and dangling sequences listed in Table 1. The resulting Hi-C matrix was balanced with the algorithm introduced by Knight and Ruiz [57] using "`hicCorrectMatrix correct`"; the "`–filterThreshold`" parameter was chosen based on the histogram produced by "`hicCorrectMatrix diagnostic_plot`" (Supplementary Table S1). From this matrix, a contact map at the resolution of 100 kb was derived using "`hicMergeMatrixBins`" with parameter "`–numBins 20.`"

For comparability, the contact map obtained in this manner was scaled to the 0 to 1 range with "`hicNormalize –normalize norm_range`". An entry in the resulting matrix $\mathbf{H} = (h_{i,j})$ represents the contact frequency between genomic bins i and j. Bins enclosing contromere locations (obtained from the UCSC table browser [50]) as well as low coverage bins (below 10% of overall contact probability) were excluded from analyses (Supplementary Tables S5 -S8).

DNA-DDA

Delay differential analysis in general

Let $x(t)$ be a dynamical sequence of length L where t represents increments in time or space. A nonlinear DDA model has the following general functional form ([43] and citations therein):

$$\dot{x} = \sum_{k=1}^{K} a_k \prod_{n=1}^{N} x_{\tau_n}^{m_{n,k}} + \rho \tag{1}$$

where $\dot{x} = \dot{x}(t)$ is the derivative of the original time (or space) series $x = x(t)$ and $x(t-\tau)$ is the value of the series shifted by τ steps $()x_{\tau_n} = x(t-\tau_n)$. The model parameters are: K, the number of monomials; N, the number of delay embeddings x_{τ_n} contained in each monomial; $m_{n,k} \in \mathbf{N}_0$, the order of nonlinearity of the nth delay embedding in the kth monomial. Finally a_1, a_2, a_3 are the fitting coefficients, and ρ is the least-square error of the model:

$$\rho = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\dot{x}(t_l) - \sum_{k=1}^{K} a_k \prod_{n=1}^{N} x(t_l)_{\tau_n}^{m_{n,k}}\right)^2} \tag{2}$$

A full list of all possible three term DDA models up to cubic nonlinearity and two delay pairs ($K \in \{1,2,3\}$, $m \in \{1,2,3\}$, $n \in \{1,2\}$) can be found in Supplementary Table S4.

Table 1. DNA-DDA structure selection and testing data sets

Cell type	Hi-C data		ChIP-seq		RS	DS
GM12878	GSE63525 (primary)	[2], [48]	GSE29611; GSM733772	[49]	GATC	GATC
K562	GSE63525	[2], [48]	GSE31755; GSM788085	[50]	GATC	GATC
hESC	GSE35156	[51]	GSE29611; GSM733687	[49]	AAGCTT	GATC
IMR90	GSE35156	[51]	GSE103589; GSM2775001	[52]	AAGCTT	GATC

ChIP-seq: of histone mark H3K4me1 (GM12878, K562, IMR90) or H4K20me1 (hESC); RS: restriction sequence; DS: dangling sequence.

For a dynamical sequence $x(t)$, Equation 1 can be written as the over-determined system of equations

$$\dot{\vec{x}} = \mathbf{M}\vec{\alpha} \tag{3}$$

where each element in $\dot{\vec{x}}$ is the center derivative at each time/space point $(\dot{x}(t), \dot{x}(t+1), \ldots, \dot{x}(t+L))$, $\mathbf{M}$ a matrix where each column represents a monomial (delay embedding $x_{\tau_n}^{m_{n,k}}$) at a certain time/space point (row), and $\vec{\alpha} = (a_1, \ldots, a_K)$ are the fitting coefficients of the model which are estimated for the input data using singular value decomposition (SVD), and together with the least square error (Equation 2) make up the classifying feature set $F = (\vec{\alpha}, \rho)$. Furthermore, Equation 1 can be solved for I dynamical input sequences $\vec{x}(t) = (x_1(t), x_2(t), \ldots, x_I(t))$ either individually (single trial (ST) DDA) or simultaneously (cross trial (CT) DDA) (Figure 1) by extending the over-determined system of equations of Equation 3 to

$$\begin{pmatrix} \dot{\vec{x}}_1 \\ \dot{\vec{x}}_2 \\ \vdots \\ \dot{\vec{x}}_I \end{pmatrix} = \begin{pmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \\ \vdots \\ \mathbf{M}_I \end{pmatrix} \vec{\alpha}. \tag{4}$$

ST and CT DDA features have recently been combined by Lainscsek et. al [43] in a way which allows testing for dynamical similarity between two dynamical input sequences $x_i(t)$ and $x_j(t)$. Here the mean of the ST error is representative of the temporal average, and the CT error is representative of the ensemble average. In accordance with the ergodic hypothesis [25], if $x_i(t)$ and $x_j(t)$ are dynamically similar, the mean of the ST errors $\overline{\left(\rho_{s_i}, \rho_{s_j}\right)} = \frac{\rho_{s_i}+\rho_{s_j}}{2}$ and the CT error $\left(\rho_{c_{i,j}}\right)$ will also be similar and therefore, their quotient close to one. DE-DDA $\mathscr{E}_{i,j}$ is defined as

$$\mathscr{E}_{i,j} = \left| \frac{\overline{\left(\rho_{s_i}, \rho_{s_j}\right)}}{\left(\rho_{c_{i,j}}\right)} - 1 \right|. \tag{5}$$

The smaller is $\mathscr{E}_{i,j}$, the more similar are the dynamics of the two signals under investigation, i and j.

DDA for genomic sequence data

Numeric representation of DNA sequences

For DNA-DDA, t corresponds to the genomic position of a nucleotide and L is the resolution or number of nucleotides in each bin. We modeled the DNA sequence as a 1D random walk (DNA walk). Specifically, for every genomic bin, the walker starts at $x(t=1) = 0$ and progresses along the DNA sequence taking a step upwards ($x(t+1) = x(t) + 1$) if the nucleotide at position $t+1$ is C or G and down ($x(t+1) = x(t) - 1$) if the nucleotide at position $t+1$ is A or T [58]. The DNA walk of five exemplary nonempty bins covering the genomic region chr1:800001-1300001 is depicted in Supplementary Figure S1. A time delay in a DNA-DDA model is a shift in genomic coordinates. A delay embedding of the DNA sequence relates the value of the DNA walk at the genomic coordinate t to its "previous" value at $x(t-\tau)$. We are trying to gain access to (1) the interaction frequency of two genomic loci by considering only (2) the DNA nucleotide sequence within these loci. We achieve this by exploiting the concept of embedding theory, which suggests that certain variables of nonlinear systems are coupled and entail information about one another.

DE-DDA classifying feature set computation

The sequence of the GRCh38/hg38 assembly of the human genome was partitioned into 100 kb non-overlapping bins, and the sequence of each bin was represented as a 1D DNA walk. For a pair of bins i, j, we computed the ST- and CT-errors ρ_{s_i}, ρ_{s_j} and $\rho_{c_{i,j}}$ using a `C` implementation of DDA provided by the author [43]. This executable takes as input a time series and parameters listed in Supplementary Table S2 and outputs ST- and CT- classifying feature set (a_1,a_2,a_3,ρ). We combined the errors ρ_{s_i}, ρ_{s_j} and $\rho_{c_{i,j}}$ into $\mathscr{E}_{i,j}$ (Equation 5), which we propose as an estimation of the contact probability between the two pairs of genomic bins. We repeated this process for all bin pairs to obtain the DNA-DDA contact map $\mathbf{D} = (d_{i,j})$ of a given chromosome.

Matrix post-processing

We hypothesize that bins with higher contact probability will have similar certain dynamical properties. Thus, we inverted $\mathbf{D}$ so that the highest values are mapped to the lowest and vice versa. The logarithm of each non-zero value in $\mathbf{D}$ was taken. The matrices were saved in the file format of `HOMER` [59] and then converted to `h5` with *HiCExplorer*'s "`hicConvertFormat`" function. The resulting matrices were normalized to the 0 to 1 range with "`hicNormalize -normalize norm_range`". Bins that were excluded in the Hi-C contact maps $\mathbf{H}$ were also excluded in the DNA-DDA contact
maps $\mathbf{D}$.

Compartment calling

To call compartments, we first derived the Pearson correlation matrix $\mathbf{C_H} = (c_{\mathbf{H}_{i,j}})$ from the normalized Hi-C matrix $\mathbf{H}$ as described by Lieberman, and then applied principal component analysis (PCA) to $\mathbf{C_H}$ using MATLAB®'s "`pca`" function. For each principal component (PC), values larger than three scaled median absolute deviations (MAD) from the median were considered extreme outliers and replaced with nearest value that was not an outlier using MATLAB®'s "`filloutliers(PC,'nearest')`" function. The PCs were then normalized to zero mean and unit variance. Lastly, values greater than 0 were assigned to the A compartment and scaled to $[0, 0.5]$; values below 0 were assigned to the B compartment and scaled to $[0.5, 1]$. We used ChIP-seq data for H3K4me1 or H4K20me1, two epigenetic modifications associated with open chromatin [60], to determine which PC defines compartments (Table 1). More specifically, among the first four PCs, we selected the one with the largest absolute Pearson correlation

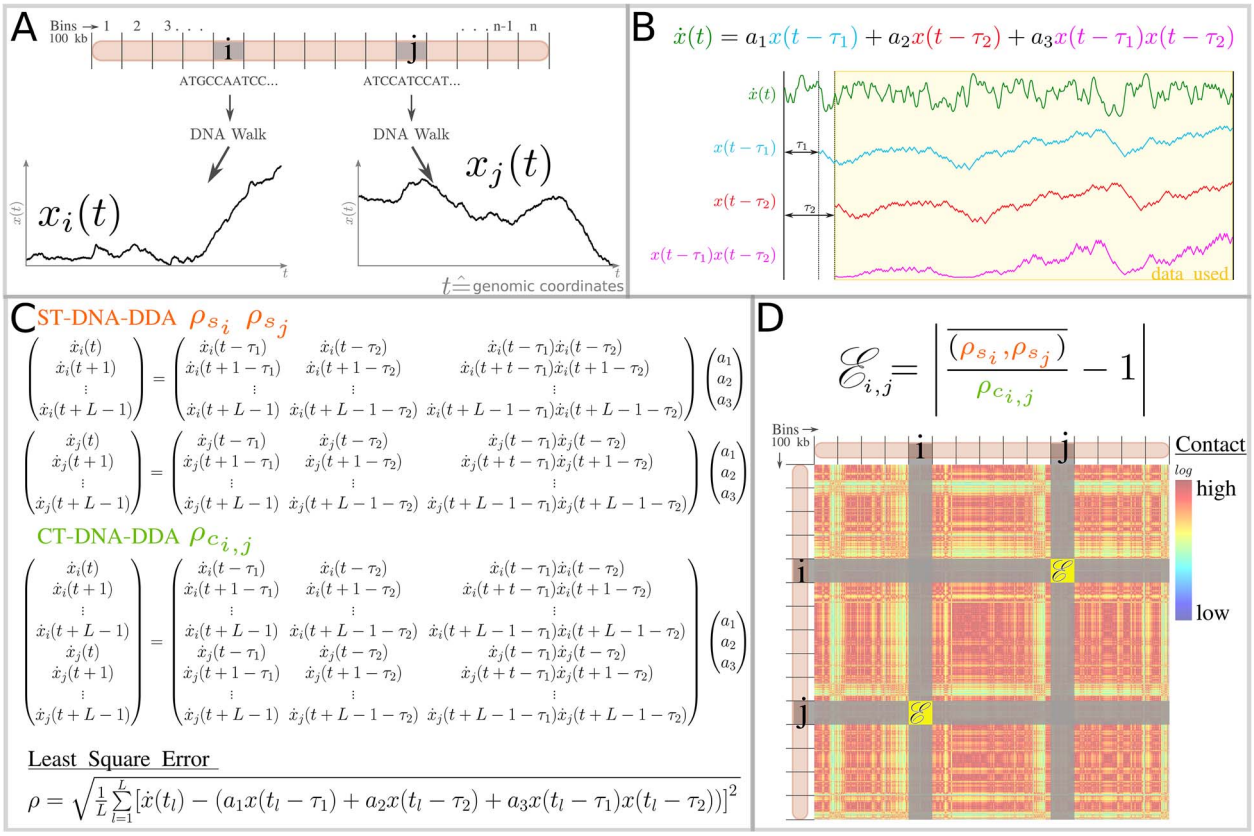

Figure 1. Steps taken to derive the entries of a DNA-DDA contact matrix **D**. (A) The DNA sequences of two exemplary bins *i* and *j* located on one chromosome are represented as *DNA walks* $x_i(t)$ and $x_j(t)$. Each walk starts at $x(t = 1) = 0$ and takes a step up (+1) if the nucleotide at position $t + 1$ is C or G and down (-1) if the nucleotide at position $t+1$ is A or T. (B) The DNA-DDA model and visualization of one DNA walk $x(t)$ and its respective derivative $\dot{x}(t)$ and delay embeddings $x(t - \tau_1)$ and $x(t - \tau_2)$. The yellow window indicates the data points used in estimation of the DDA features (a_1, a_2, a_3, ρ). (C) Overdetermined system of equations for ST and CT DDA. The coefficients (a_1, a_2, a_3) are determined by singular value decomposition (SVD) separately for bin i and *j* in ST DDA and in a single step in CT DDA. Least square errors are computed for ST DDA (ρ_{s_i}, ρ_{s_j}) and CT DDA $(\rho_{c_{i,j}})$. (D) The ST DDA least square errors (ρ_{s_i}, ρ_{s_j}) and CT DDA least square error $(\rho_{c_{i,j}})$ are combined to dynamical ergodicity $\mathscr{E}$ [43]. Note this figure is a detailed description of how the DNA-DDA matrix is determined in the workflow of DNA-DDA (see Supplementary Figure S3).

coefficient with the ChIP-seq profile to be representative of A/B compartments (Supplementary Table S9). ChIP-seq profiles were generated from the corresponding bed files; an empty vector of the same length as the PC was generated and +1 was added to each bin for each called peak falling into the bin's respective genomic region. If the Pearson correlation coefficient was negative, the PC was multiplied by −1. In two exceptional instances (chr1 for Hi-C contact maps of GM12878 and IMR90), a different PC with approximately the same Pearson correlation coefficient was deemed more likely to be associated with the compartments upon inspection, and used instead. From here on, we refer to the PC obtained from $\mathbf{C_H}$ as $PC_{Hi\text{-}C}$.

Computationally derived DNA-DDA matrices **D** were processed in the same manner with one exception, their corresponding PC's were smoothed with a sliding window of 5 using MATLAB®'s "`movmean()`" function before computing the Pearson correlation coefficients with the ChIP-seq profiles. We further refer to the resulting DNA-DDA Pearson correlation matrices as $\mathbf{C_D} = (c_{\mathbf{D}_{i,j}})$ and to the corresponding PCs defining A/B compartments as $PC_{DNA\text{-}DDA}$.

Saddle plot analysis

We performed saddle plot analysis on the contact maps predicted by DNA-DDA and Hi-C respectively in the K562 and IMR90 cell lines. For each chromosome, values of $PC_{Hi\text{-}C}$ and $PC_{DNA\text{-}DDA}$ were split into 30 equisized bins ($PC_{Hi\text{-}C\text{-}S}$ and $PC_{DNA\text{-}DDA\text{-}S}$). The expected interaction value at distance *d* of the balanced [57] Hi-C contact map **H** was computed as the sum of its *d*th diagonal $diag_d(\mathbf{H})$, divided by the number of elements in diagonal *d*, $n(diag_d(\mathbf{H}))$. Then we sorted the values of the expected/observed Hi-C matrix $\mathbf{H}_{oe}$ and the DNA-DDA matrices **D** according to $PC_{Hi\text{-}C\text{-}S}$ and $PC_{DNA\text{-}DDA\text{-}S}$ respectively. Finally, we quantified compartment strength from the interaction values that were allocated to the highest 25% of the PC values (implying *AA* and *BB* interactions) and lowest 25% of PC values (implying *AB* or *BA* interactions) as $S = \frac{AA+BB}{AB+BA}$. Note that all PC values were scaled to the interval [−1, 1], as often done (e.g. in [4]), and not as explained in the *Compartment calling* section, as this would have resulted in a large number of bins with no allocated PC values.

Correlation with CG content

We compared DNA-DDA's performance to predict compartments with a baseline of AT/CG percentage over 100 kb windows in all four cell types.

Structure selection and testing

A key difference to traditional machine learning-based approaches is that DDA models are not updated iteratively (i.e. they do not learn). Instead, an exhaustive sweep is performed over a list of

putative models to search for those best suited to discriminate the dynamics of interest; this step is called *structure selection*.

The functional form of a DDA model is dictated by the overall system and obtained data type. The data type most extensively studied using DDA is EEG, for which a particular functional form has been established [45–47, 61, 62]. Typically, most terms in Equation 1 are set to zero to reduce the chances of overfitting (e.g. $K \in \{1, 2, 3\}$, $m \in \{1, 2, 3\}$ $n \in \{1, 2\}$). This work is a proof of concept to explore the feasibility of applying DDA to genomics data. Thus, we decided to use a simple, symmetric model with only quadratic degree of nonlinearity (DDA model number 2 in Supplementary Table S4):

$$\dot{x} = a_1 x_{\tau_1} + a_2 x_{\tau_2} + a_3 x_{\tau_1} x_{\tau_2} + \rho \tag{6}$$

Symmetric models require half the computational effort of non-symmetric ones to compute the classifying feature set $\{a_1, a_2, a_3, \rho\}$ for a range of delays between r_1 and r_2; $\tau_1, \tau_2 \in (r_1, r_2)$ ($r_i \in [1 : 50]$ in this study).

Next, we searched for the delay pair τ_1, τ_2 that best captured A/B compartments in each of the four cell types. Specifically, this was done using a 20 Mb region on chr22 (chr22:16200000-36200001, Supplementary Figure S2). The chromosome was chosen arbitrarily and the region thereon corresponded to the one with the lowest compartment agreement between the four cell types considered in this study. The performance of each delay pair was measured as the Pearson correlation coefficient between $PC_{Hi\text{-}C}$ and $PC_{DNA\text{-}DDA}$ in this region.

The four obtained DNA-DDA models were tested on all 100 kb genomic loci of all human autosomes with various performance measures (Pearson correlation coefficient r_{PC} between $PC_{Hi\text{-}C}$ and $PC_{DNA\text{-}DDA}$, area under the receiver operating characteristic (ROC) curve AUC, accuracy ACC and F1-score F1 for classifying A/B compartments). We would like to emphasize that once we determined the DNA-DDA model for each cell type (Supplementary Table S3), all subsequent analyses and results were performed on never before seen data (with the small exception of chr22:16200000-36200001).

Comparison to other methods

We compared DNA-DDA to three other methods that can predict A/B compartments from sequence-based features or the sequence directly (Table 2). The "Sequence-based Annotator of chromosomal compartments by Stacked Artificial Neural Networks" (SACSANN [8]) predicts A/B compartments based on features derived from GC content, transposable elements (TE), and putative transcription factor binding sites (TFBS). It identifies the 100 most important species/cell-type specific features with a random forest predictor and then trains two stacked artificial neural networks (ANN) to classify 100 kb-long genomic bins into either the A or the B compartment.

The "A/B Compartment Network" (ABCNet [63]) is a deep convolutional artificial neural network (CNN) that takes a one-hot encoding of the sequence within 100 kb-long bins and extracts features by passing them through two-layer convolutional kernels, an average pooling layer, and a fully connected dense layer to output a single value representing the predicted PC value of each bin, classifying it either as an A or B compartment.

Finally, "Orca" [64] is a multiscale prediction model composed of two CNNs, a hierarchial multi-resolution sequence encoder and a cascading series of sequence decoders. The encoder takes up to 256 Mb one-hot-encoded sequence as input, and generates a series of decreasing resolution sequence representations,

Table 2. Comparison of DNA-DDA models to other methods

Name	Method	Input	Target	Training/ structure selection data	validation data	test data	nr. of data-sets (models)
SACSANN	Random forest, ANN	Number of TEs, TFBSs, GC content	Compartments	21 chromosomes†	21 chromosomes†	1 chromosome	8
ABCNet	CNN	Sequence	Compartments	90% of 21 chromosomes	10% of 21 chromosomes	1 chromosome	28
Orca	CNN	Sequence	Genome-wide contacts, compartments	chr1-7 chr11-22	chr8	chr9-10	3
DNA-DDA	DDA	Sequence	Genome-wide contacts, compartments	20 Mb on chr22	20 Mb on chr22	chr1-22	4

ANN: artificial neural network; **CNN**: convolutional neural network; **Nr. of data sets**: number of Hi-C data sets used to generate models. †: Training/validation split not stated.

centered around the input, with a convolutional architecture. The decoders then each predict interactions of up to 256 Mb at 1024 kb resolution at the top level and interactions within 1 Mb at 4 kb resolution at the bottom level.

Both SACSANN and ABCNet model architecture were applied to numerous data sets (Table 2) in terms of feature selection, training/testing procedures. Resulting models were evaluated using a chromosome-wise leave-one-out cross validation. We compared the performance of DNA-DDA on the hESC data set (GSE35156) to that reported by SACSANN's and ABCNet's authors on the same data set. In addition, we compared the *overall* performance of the methods, i.e. the average *AUC* or *ACC* across various data sets and models. In the case of SACSANN, we considered all examined data sets ("Summary (ROC AUC score) SACSANN" in supplemental_file_S3.xls at https://github.com/BlanchetteLab/SACSANN/tree/master/supplemental_files/, last accessed in July 2022). For ABCNet, we restricted the comparison to the *ACC* reported for human data sets (i.e., average μ across all "secondary" data sets in Table III in Kirchof [63].

Orca was trained on two of the highest resolution micro-C (an improvement of Hi-C) data sets available for the H1 human embryonic stem cell line (H1-ESC) and Human foreskin fibroblast cell line (HFF) [65]. We compared to Orca models which predict interactions at the closest resolution that DNA-DDA currently operates on (128 kb and 100 kb respectively). To increase comparability, we considered an Orca model trained on the same cell line (H1-ESC 4DNES21D8SP8); it takes 32 Mb as input and predicts 128 kb interactions within this region. Therefore, we split one of the Orca hold-out chromosomes, chr9, into three 32 Mb regions (Supplementary Table S11), derived a Pearson correlation matrix from the predicted Orca contact maps, performed a PCA in MATLAB® and computed the Pearson correlation coefficient r_{PC} of the resulting PCs to those identified by the HiC-data for hESC GSE35156 at 128 kb.

RESULTS

DNA-DDA is a statistical approach derived from DDA, a method based on nonlinear dynamics and traditionally used for time series data, which predicts A/B compartments from the reference sequence (Figure 1). DDA models relate the numerical derivatives of the input data to their time-delayed versions in a nonlinear functional form, of which the fitting coefficients and model error $\{a_1, a_2, a_3, \rho\}$ are used to access information about the underlying dynamical system. With simplicity and efficiency in mind, we chose the functional form given in Equation 6. In principle the analysis can be summarized in three steps: (1) segment the reference sequence into 100 kb long bins and represent the sequence in each bin as a 1D DNA walk, (2) determine the best suited model parameters (delay pair τ_1, τ_2) in each cell type by supervised structure selection with Hi-C derived compartment labels and (3) apply the model to the rest of the genome to predict A/B compartments.

Structure selection

To find the delay pairs τ_1, τ_2 in Equation 6 that best capture A/B compartments in each cell type, we tested model performances for all possible combinations of values for τ_1, τ_2 on a 20 Mb-long region of chr22 (Methods). Briefly, we partitioned this region into 200 100 kb-long bins, estimated the contact probability between each pair of bins i and j as $\mathscr{E}_{i,j}$ (Equation 5), built a DNA-DDA contact map $\mathbf{D}$, and applied PCA to the respective DNA-DDA Pearson correlation matrix $\mathbf{C_D}$ to call A/B compartments. For each cell type, we then chose the delay pair that resulted in the highest absolute Pearson correlation coefficient r_{PC} between $PC_{DNA\text{-}DDA}$ and $PC_{Hi\text{-}C}$ (Supplementary Figure S4). This resulted in four different delay pairs, one for each cell type (Supplementary Table S3). Pearson correlation coefficients between $PC_{DNA\text{-}DDA}$ and $PC_{Hi\text{-}C}$ ranged between 0.21 (hESC) and 0.49 (GM12878) and AUCs ranged between 0.61 (hESC) and 0.77 (GM12878).

Testing

The DNA-DDA Pearson correlation matrices $\mathbf{C_D}$ exhibited strikingly similar global patterns to the experimentally obtained Hi-C Pearson correlation matrices $\mathbf{C_H}$ (Figure 2 and Supplementary Figures S9 to S11). In general, supervised methods are limited by the uncertainty of the labels derived from the experimental data. It is known that identifying A and B compartments by PCA of the Pearson correlation matrix derived from the contact map is suboptimal and misses many features, especially at higher resolutions. This is particularly apparent in the IMR90 data set (Supplementary Figure S7) or particular chromosomes (e.g., chr22). Nevertheless, DNA-DDA achieves exceptional performances on never before seen genome, with $\overline{AUC} = 0.81$, $\overline{ACC} = 0.74$, $\overline{F1} = 0.72$, $\bar{r}_{PC} = 0.54$ over all chromosomes and cell lines (Table 3), especially considering the mere 20 Mb region that was used to determine the model/delay pair combination. The model/delay pair combination achieves the highest performance in the K562 cell line ($\overline{AUC} = 0.84$, $\overline{ACC} = 0.76$, $\overline{F1} = 0.75$ and $\bar{r}_{PC} = 0.60$), and the lowest performance in the IMR90 cell line ($\overline{AUC} = 0.75$, $\overline{ACC} = 0.70$, $\overline{F1} = 0.67$ and $\bar{r}_{PC} = 0.45$).

Saddle plot analysis

Saddle plot analysis revealed that DNA-DDA derived matrices had overall stronger saddle strengths strengths S than those derived from the expected/observed Hi-C matrix $\mathbf{H}_{oe}$ (Supplementary Figure S15). The mean saddle strength across all chromosomes for each cell line was $\overline{S}_{DDA} = 6.28$ and $\overline{S}_{HiC} = 2.27$ for K652, $\overline{S}_{DDA} = 4.69$ and $\overline{S}_{HiC} = 1.95$ for IMR90 and $\overline{S}_{DDA} = 5.91$ and $\overline{S}_{HiC} = 1.53$ for hESC. However, this analysis should be interpreted with caution as the Hi-C and DNA-DDA contact maps are of very different origins.

Correlation with CG content

CG content is known to correlate with the compartment signal. Nonetheless, DNA-DDA predicted compartments better for 100 kb windows than the corresponding percentage of AT/CG content in each window. The mean Pearson correlation coefficients across all chromosomes was between 0.45 and 0.60 as opposed to the means between 0.45 and 0.56 observed for the GC-content-based prediction. Specifically, when examining individual chromosomes, DNA-DDA performed significantly better for GM12878, K562 and hESC ($p < 0.05$; Wilcoxon signed-rank test) and exhibited no difference for IMR90, the cell line for which DNA-DDA performed the worst overall. These results suggest that DNA-DDA and GC content are complementary methods.

Comparison to other methods

We compared DNA-DDA to SACSANN [8], ABCNet [63] and Orca [64]. To our knowledge, these are the only methods that are directly comparable to DNA-DDA, as they rely only on the DNA sequence or sequence-based features to predict A/B compartments. DNA-DDA competes well with other methods predictive of A/B compartments when assessed on the same Hi-C data set (hESC GSE35156) as well as overall (Table 4). Indeed, all four methods showed similar scores measured by *AUC*, *ACC* or r_{PC}. Note the vast difference in the size of the "*training*" (pendant to

Table 3. Performance of DNA-DDA

chr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	μ	σ^2
GM12878																								
r_{PC}	0.68	0.52	0.72	0.63	0.53	0.69	0.45	0.54	0.62	0.41	0.70	0.61	0.60	0.69	0.42	0.66	0.52	0.65	0.64	0.66	0.58	0.49	0.59	0.01
AUC	0.88	0.81	0.91	0.85	0.82	0.90	0.79	0.80	0.82	0.76	0.86	0.87	0.84	0.87	0.76	0.85	0.76	0.88	0.83	0.85	0.86	0.76	0.83	0.00
ACC	0.79	0.73	0.82	0.79	0.73	0.81	0.69	0.75	0.77	0.68	0.82	0.75	0.77	0.81	0.67	0.80	0.73	0.79	0.80	0.80	0.73	0.70	0.76	0.00
F1	0.76	0.70	0.79	0.75	0.69	0.80	0.64	0.70	0.78	0.64	0.82	0.73	0.74	0.82	0.66	0.79	0.77	0.77	0.83	0.80	0.72	0.75	0.75	0.00
K562																								
r_{PC}	0.66	0.67	0.69	0.53	0.45	0.74	0.60	0.67	0.52	0.58	0.56	0.69	0.74	0.72	0.47	0.59	0.38	0.75	0.54	0.76	0.66	0.32	0.60	0.02
AUC	0.89	0.89	0.90	0.82	0.74	0.91	0.86	0.87	0.74	0.85	0.86	0.89	0.90	0.92	0.79	0.85	0.72	0.93	0.80	0.92	0.86	0.66	0.84	0.01
ACC	0.79	0.79	0.80	0.75	0.65	0.83	0.77	0.80	0.68	0.74	0.75	0.80	0.84	0.80	0.68	0.76	0.65	0.83	0.74	0.85	0.79	0.59	0.76	0.00
F1	0.75	0.77	0.77	0.70	0.65	0.82	0.74	0.78	0.69	0.73	0.72	0.78	0.83	0.79	0.66	0.75	0.67	0.83	0.76	0.85	0.77	0.65	0.75	0.00
IMR90																								
r_{PC}	0.63	0.39	0.53	0.50	0.42	0.22	0.30	0.29	0.49	0.32	0.56	0.24	0.68	0.56	0.31	0.62	0.36	0.51	0.44	0.60	0.55	0.30	0.45	0.02
AUC	0.86	0.75	0.78	0.81	0.76	0.64	0.67	0.67	0.78	0.72	0.80	0.64	0.86	0.83	0.67	0.85	0.70	0.77	0.73	0.81	0.80	0.69	0.75	0.00
ACC	0.78	0.66	0.75	0.73	0.69	0.60	0.65	0.64	0.70	0.65	0.74	0.61	0.80	0.74	0.63	0.77	0.66	0.72	0.70	0.76	0.74	0.64	0.70	0.00
F1	0.72	0.61	0.68	0.69	0.60	0.60	0.54	0.58	0.68	0.68	0.74	0.62	0.79	0.72	0.59	0.75	0.69	0.70	0.70	0.73	0.69	0.69	0.67	0.00
hESC																								
r_{PC}	0.42	0.40	0.77	0.59	0.42	0.41	0.67	0.36	0.73	0.34	0.66	0.80	0.54	0.41	0.46	0.51	0.45	0.47	0.47	0.42	0.71	0.54	0.53	0.02
AUC	0.83	0.73	0.93	0.80	0.72	0.75	0.84	0.66	0.92	0.64	0.85	0.95	0.82	0.77	0.79	0.79	0.75	0.76	0.78	0.76	0.92	0.84	0.80	0.01
ACC	0.71	0.68	0.85	0.74	0.68	0.69	0.81	0.64	0.84	0.61	0.79	0.88	0.76	0.68	0.68	0.72	0.68	0.71	0.71	0.68	0.82	0.72	0.73	0.01
F1	0.74	0.66	0.82	0.73	0.63	0.65	0.77	0.60	0.81	0.60	0.76	0.85	0.74	0.64	0.64	0.69	0.66	0.71	0.75	0.68	0.80	0.72	0.71	0.01

μ: mean performance measure over all autosomes; σ^2: variance of performance measure over all autosomes.

Table 4. Comparison of DNA-DDA performance to other methods

	Performance measure	**hESC (GSE35156)**	μ_1
SACSANN	AUC	0.81	0.83
DNA-DDA	AUC	0.82	0.81
ABCNet	ACC	0.75	0.80
DNA-DDA	ACC	0.75	0.76
	Performance measure	**hESC (GSE35156) regions 1–3**	μ_2
Orca	r_{PC}	0.86, 0.87, 0.90	0.89
DNA-DDA	r_{PC}	0.76, 0.71, 0.74	0.74

hESC (GSE35156): average performance across all chromosomes on the hESC GSE35156 Hi-C data set. μ_1: average performance across multiple human Hi-C data sets (see Methods and Supplementary Table S10). **hESC (GSE35156) regions 1–3**: Pearson correlation r_{PC} of predicted PCs of DNA-DDA and Orca contact maps at 100 kb and 128 kb respectively with three 32 Mb regions on chr9 in hESC. μ_2: mean prediction performance across the three 32 Mb regions.

structure selection in DNA-DDA) and test data sets between the methods in Table 2. Furthermore it should be noted that although the Orca model we compared to was trained on the same cell line, a far more deeply sequenced HiC library was used in the training/validation procedure (0.3 billion vs 5.6 billion (4DNES21D8SP8) total reads).

DISCUSSION AND CONCLUSION

DNA-DDA is a nonlinear dynamics method for predicting A/B compartments based on the DNA sequence alone. We derived DNA-DDA models for four human cell types using only a 20 Mb-long region on chr22 and corresponding compartment labels defined by Hi-C data. DNA-DDA achieved mean AUCs across all held-out chromosomes between 0.75 and 0.84. The achieved performances on never-before-seen testing data demonstrate the potential of DDA, a method which has been shown to accurately classify time series data, in the field of genomics.

Many computational methods for predicting genome architecture have been developed in recent years [66], but only a few are able to predict A/B compartments from the sequence alone. ABCNet [63] and SACSANN [8] are both convolutional neural network (CNN) based models that can predict A/B compartments. While ABCNet relies only on the sequence, SACSANN [8] uses counts of sequence-derived features (e.g. TEs, TFBSs). Both achieve AUC scores close to 80%, but while ABCNet does not require any previous knowledge about the genome of interest, SACSANN derives its input features from annotation data sets that are not available for every species. The currently most comprehensive sequence-based approach for modeling 3D genome architecture is Orca [64]. Orca is also the first sequence based model able to predict truly long-range interactions (> 1 Mb). Orca models have been trained on two of the highest resolution microC data sets to date, and predict interactions simultaneously at different resolutions. The models' ability to capture sequence dependencies of 3D genome architecture has been experimentally validated. DNA-DDA competes well with these state-of-the-art methods which use a larger portion of the genome for training and/or input features other than solely the sequence (Table 4).

This study is a proof of concept meant to illustrate the potential in applying DDA-based approaches in the field of structural genomics. We want to emphasize that the DNA-DDA models presented here highly unlikely represent an optimum. Many aspects of our analysis could be evaluated and modified, including the use of a different numerical DNA sequence representation, other functional forms for the DDA models, and alternative targets to assess model performance.

We encoded the DNA sequence as a 1D DNA walk [58, 67], which is a common representation for analysis of DNA sequences in time series frameworks and has been the basis for numerous studies

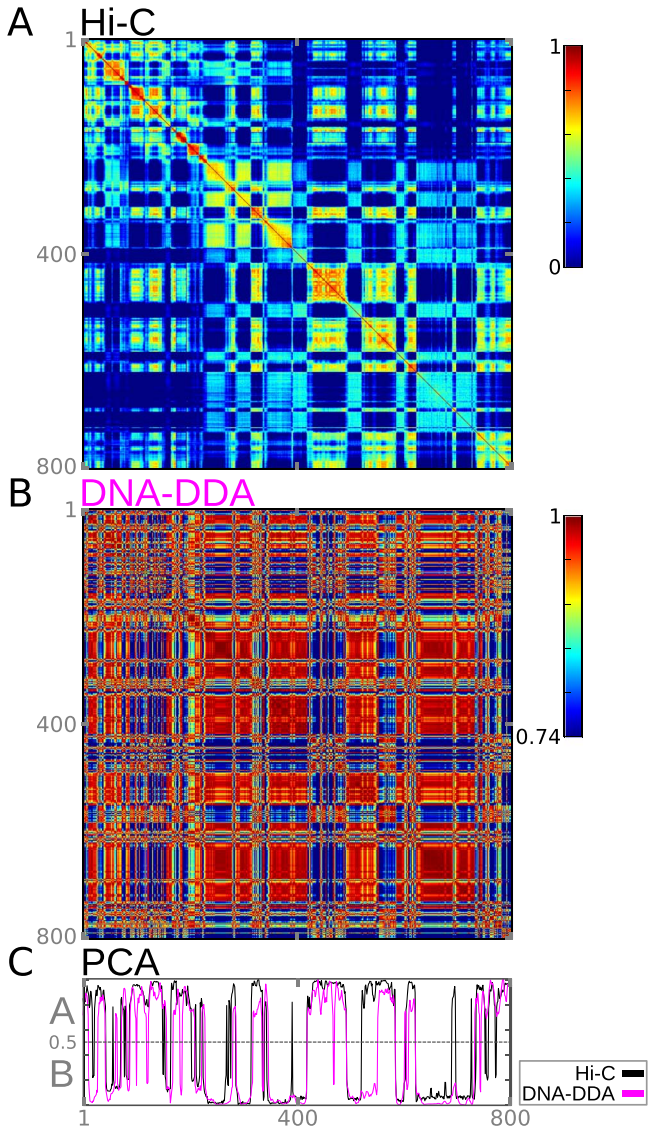

Figure 2. DNA-DDA predicts A/B compartments from the DNA sequence alone. (A) Hi-C ($\mathbf{C_H}$) and (B) DNA-DDA ($\mathbf{C_D}$) Pearson correlation matrices of the K562 cell line for an example hold out chr18 exhibit strikingly similar patterns. The color scale of the DNA-DDA Pearson matrix goes from its mean minus its variance ($\mathbf{C_{D_\mu}} - \mathbf{C_{D_{\sigma^2}}}$) to its maximum. (C) Resulting $PC_{Hi\text{-}C}$ (black) and $PC_{DNA\text{-}DDA}$ (magenta) used to define A/B compartments are in very strong correlation to one another ($r_{PC} = 0.73$).

that applied spectral analysis and signal processing methods such as discrete or Ramanujan fourier transform and, wavelet or fractal analysis for revealing high-level periodicities and patterns with biological significance [33, 34, 68, 69].

In the 1D DNA walk using the *hydrogen bond energy (SW) rule* [58, 70], the walker starts at zero and continues along the linear chain of nucleotides taking a step up for strongly bonded pairs (C or G) and down for weakly bonded pairs (A or T). Thus, DNA-DDA models capture dynamical properties based mainly on GC-content. Of course, more complex representations of the DNA sequence have been proposed as well, some of which take all four nucleotides into account (overview and comparison in [68, 71, 72]). We initially considered the alternative and equally simple 1D mapping integer representation ($T = 0$, $C = 1$, $A = 2$, $G = 3$). However this method implies biologically irrelevant properties on the bases such that purines are weighted more than pyrimidines ($(A, G) > (C, T)$). In future work, we plan to resort to DNA representations that include information of all nucleotides and do not have such a bias such as the 2D DNA walk [73]. Naturally, the ergodicity measure has to be substantially modified to achieve this.

The functional form of the DNA-DDA model was chosen based only on simplicity and computational efficiency. Previous work has shown that the overall functional form of a DDA model tends to be specific to the data type used to measure the system of interest (e.g. EEG, ECG, DNA sequence), while the delay pairs are sensitive to the question we ask about the system [62]. A large-scale exhaustive sweep of model-delay-pair combinations such as described in Lainscsek *et al.* [45] could be implemented to optimize model-delay pair combinations.

To predict A/B compartments, DNA-DDA first constructs a (DNA-DDA) contact matrix. This matrix is then post-processed (e.g. filtering, logarithmic transformation, etc.) before being subjected to PCA, as proposed by Lieberman *et al* [1]. Although compartments are routinely identified using PCA, the binary classification into A and B compartments is most likely an oversimplification of 3D genome architecture [5]. In fact, Rao *et al* [2] showed that genomes segregate into at least six subcompartments, each exhibiting a distinctive pattern of genomic and epigenetic features. Furthermore, we chose the PC to be most representative of compartments based on its correlation with a histone mark associated with open chromatin in the respective cell type and genomic region of interest, as suggested by [4]. However, the chosen PC is clearly dependent on the region being considered and often, two or more PCs will often exhibit very similar correlation coefficients to the ChIP-seq signal. Naturally, this is not a limitation exclusive to DNA-DDA, but rather, of all methods calling compartments based on PCA. A large improvement in the structure selection step would be to select delay-pairs based on an alternative mutli-variate compartment classification method or the similarity of the DNA-DDA and Hi-C contact matrices directly. Since all relevant 3D structures could be extracted from interaction matrices (at various resolutions), we strongly believe that choosing the delay-pairs with a better suited target label will substantially boost cell-type specificity.

It is important to remember that a Hi-C contact map represents the average 3D genome architecture of a cell population – a multitude of systems (different cell types) at various initial conditions. DNA-DDA models were trained using a small portion of the reference genome under the assumption that the sequence-based mechanisms contributing to chromatin folding are location-independent and cell-type invariant. Recent studies comparing bulk RNA-Seq and single-cell RNA-Seq transcriptomic profiles have demonstrated the existence of distinct expression clusters corresponding to cell types sharing similar functions [74]. Thus, despite the hypersensitivity of chaotic systems to initial conditions, the state spaces of two cells of the same type can be assumed to be more similar than those of two different types, in the topological sense. Since in the present study the target labels were compartment calls derived from bulk Hi-C data, DNA-DDA is also expected to predict the compartments that reflect the most common 3D chromatin structures in a hypothetical cell pool.

DDA models are sparse and comprise a small number of features (typically 1-4), making them robust to overfitting and well generalizable to new data [39]. Large models such as deep learning networks run the risk of capturing irrelevant patterns or "noise" in the data. In contrast, small and simple models typically fail to capture dominant signatures in the data. Since DDA does not look for the most predictive, but rather the most discriminative model, most terms in Equation 1 are set to zero. This is the power

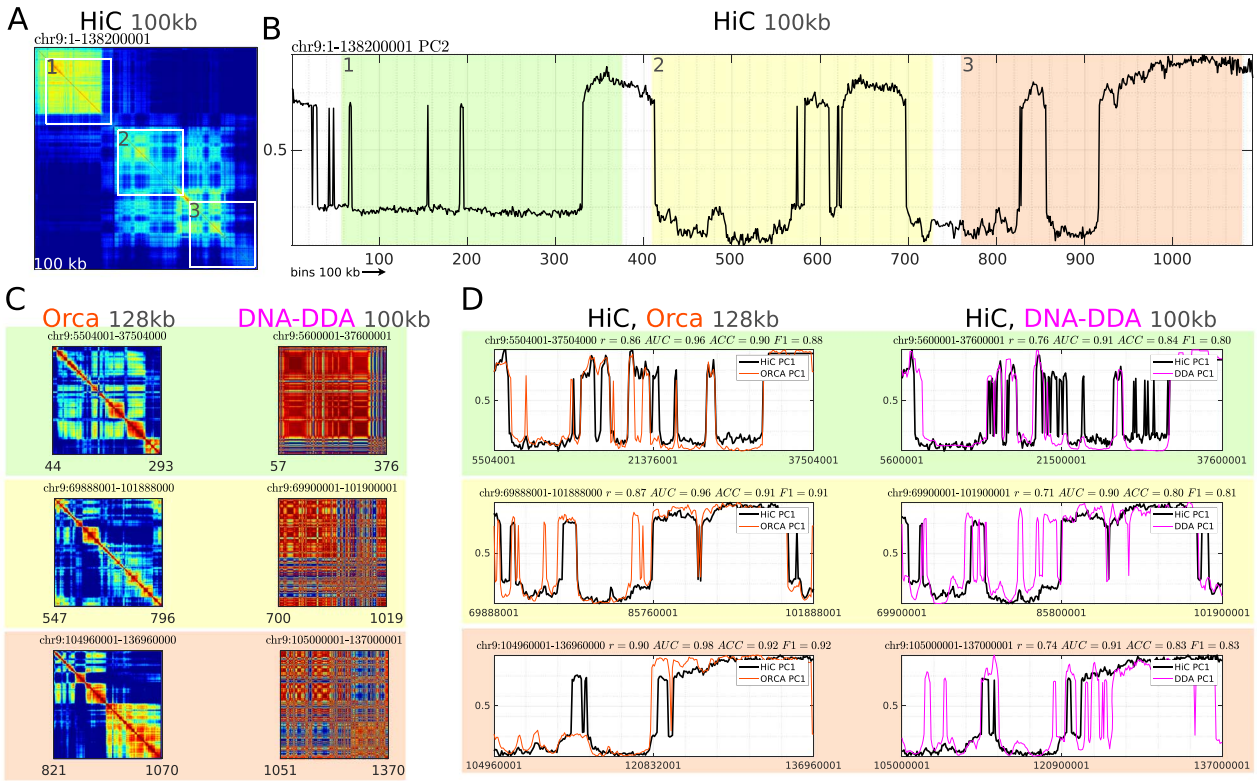

Figure 3. Comparison of DNA-DDA and Orca for the hESC cell line. (A) hESC Hi-C contact map of chr9 at 100 kb resolution and three highlighted regions (1, 2 and 3), which were used for comparison. The PC with the highest correlation to the H3K4me1 is PC2. (B) PCA on the entire Hi-C contact map in (A). (C) Orca and DNA-DDA Pearson correlation matrices for regions 1, 2 and 3. The regions were slightly increased to accommodate the Orca model, which predicts contacts at 128 kb resolution. (D) PCA analysis on regions 1, 2 and 3 individually for 128 kb resolution (left) and 100 kb resolution (right) for Hi-C data (black), Orca (orange) and DNA-DDA (magenta). In all cases and for all regions, the PC with the highest correlation to H3K4me1 was PC1. The correlation coefficients for different PCs (here, PC1 and PC2) are often very similar.

of DDA, it does not aim to model but rather capture dominant dynamical signatures in the data and 3-term DDA models have been proven sufficient for classifying complex biological data sets (eg. [45, 62, 75]). Still, one caveat of DNA-DDA concerns feature interpretability. Although models with fewer parameters are often more interpretable than the immense feature spaces that are typical of deep learning, this is not the case with DDA due to its foundation and motivation in embedding theory [40]. DDA has be related to spectral analysis: The estimated coefficients and delays of *linear* DDA models relate to the frequencies of a signal, and the estimated coefficients of *nonlinear* DDA models are connected to higher-order statistical moments [39]. In general, the delays of nonlinear DDA models are characterized by complex phase and frequency couplings which are not attributable to one particular system property. Nevertheless, due to the extremely low computational load of DNA-DDA, sequence-based mechanisms and the impact of genetic variants, such mutations in known or putative binding sites for regulatory proteins or genomic rearrangements, could easily be tested.

The power of methods like DNA-DDA and Orca lies in their intermediate step of predicting the contact matrix before calling A/B compartments, since all relevant 3D structures could be extracted from them in the same manner as Hi-C maps. Although Orca currently predicts contact maps for higher resolutions, DNA-DDA shows promising results whilst tackling the problem from a different angle and requiring far less "training data." We chose a relatively low resolution (100 kb) for this proof-of-concept study to (1) efficiently test the many possible parameters and aspects of our approach; and (2) allow for a better comparison with SACSANN and ABCNet, which both operate at 100 kb, and together with Orca are the only other purely sequence based methods to predict compartments of which we are aware. We acknowledge that our current DNA-DDA contact maps capture global large-scale patterns rather than subtle cell-type specific changes. Nevertheless, we believe that the use of a different numerical DNA sequence representation and/or other DDA model forms, and alternative targets to assess model performance will greatly improve cell-type specificity and result in more diverse DNA-DDA contact maps (Figure 4). Moreover, our results show that DDA models can be derived from a small amount of supervised data, which would enable the prediction of genome-wide interactions in other cell types from a very limited amount of chromosome conformation capturing data (e.g. generated with 5C). With some optimization, we are convinced that the method might be able to do this as well as deep learning algorithms, but using just a miniscule fraction of the volume of data required by such approaches.

The DNA sequence plays a fundamental role in the formation and maintenance of 3D genome architecture, which in turn is a central orchestrator of the gene regulatory network. It is indisputable that the sequence contains key underlying features contributing to genome folding, however, to what extent it alone can be used to gain access to all 3D interactions and relevant structures remains an open question. Our findings strongly support the hypothesis that the DNA sequence represents a highly observable variable of chromatin architecture and that chromosomal

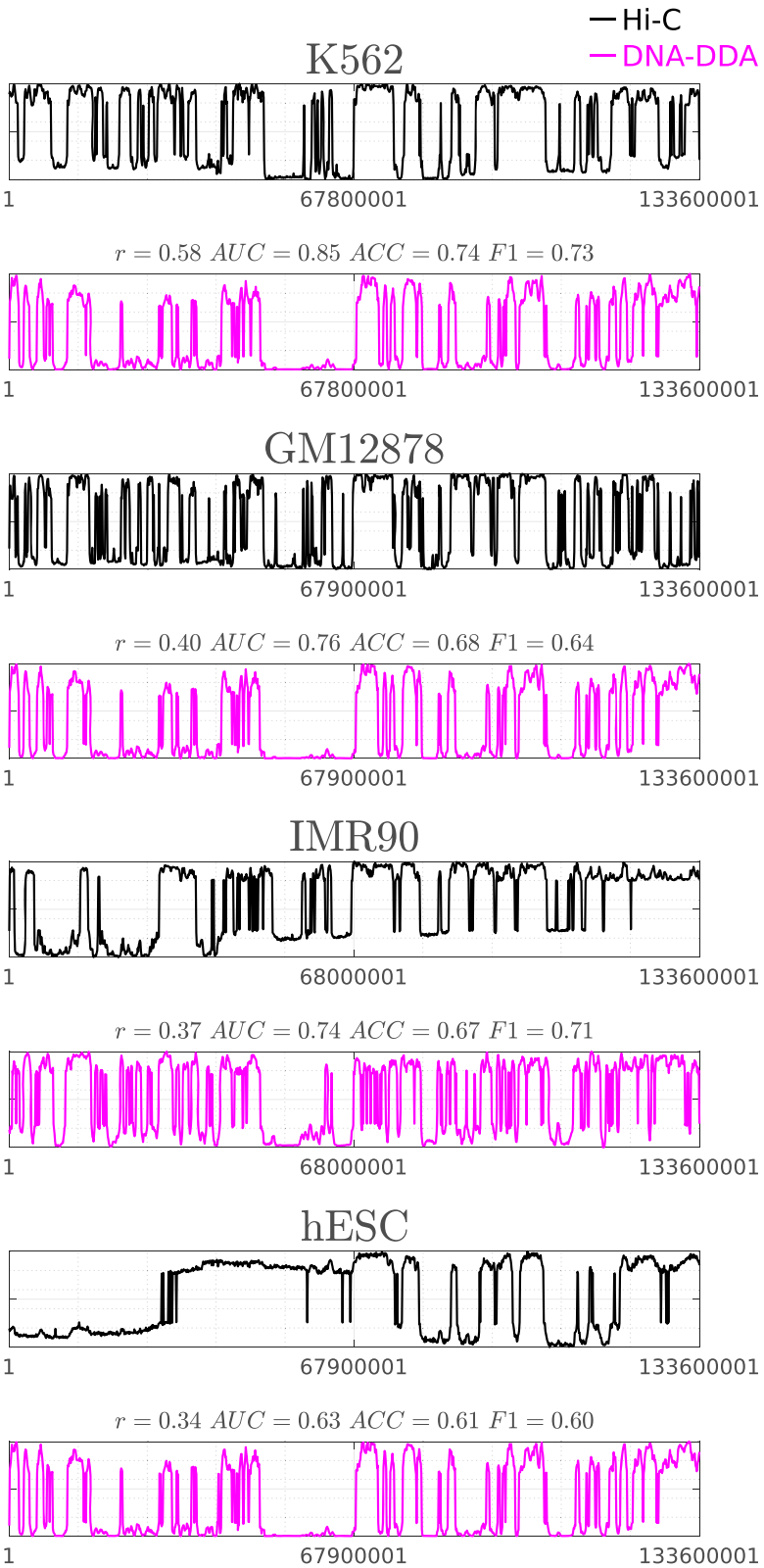

Figure 4. Comparison of DNA-DDA models in for four cell types. Principal components $PC_{DNA\text{-}DDA}$ (magenta) and $PC_{Hi\text{-}C}$ (black) are shown for chr10 of each cell line (K562, GM12878, IMR90, hESC). The overall Pearson correlation coefficient between the $PC_{Hi\text{-}C}$s across all cell lines is $\bar{r}_{HiC} = 0.24$ where GM12878 and K562 are the most similar ($r^{GK}_{HiC} = 0.69$), and K562 and hESC are the most different ($r^{Kh}_{HiC} = -0.19$). The overall Pearson correlation coefficient between the $PC_{DNA\text{-}DDA}$s across all cell lines is $\bar{r}_{DDA} = 0.81$ where GM12878 and hESC are the most similar ($r^{Gh}_{DDA} = 0.99$), and IMR90 and GM12878 are the most different ($r^{IG}_{DDA} = 0.62$).

compartments can be predicted solely from the DNA sequence. This opens up a variety of possibilities such as discovering novel sequence signatures imperative to structural genome function and how disruption of 3D genome architecture relates to human disease.

Key Points

- Substantial information about 3D genome architecture can be uncovered from from solely the DNA sequence.
- DNA-DDA models derived from a fraction of the typical amount of required target labels, already show high predictive power in classifying A/B compartments.
- Delay differential analysis, a technique with foundations in chaos theory, is suited for analyzing genomic sequence data in the context of structural genomics.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

ACKNOWLEDGEMENTS

We thank Dr Claudia Lainscsek for the fruitful discussions and sharing the C implementation of DDA, which was used as a basis for DNA-DDA.

DATA AVAILABILITY

Exemplary code used to predict contacts at 100 kb resolution of chr22 can be found at https://github.com/xX3N1A/DNA-DDA.

REFERENCES

1. Lieberman-Aiden E, van Berkum NL, Williams L, *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**(5950):289–93.
2. Rao SS, Huntley MH, Durand NC, *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**(7):1665–80.
3. Liu Y, Nanni L, Sungalee S, *et al.* Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotype. *Nat Commun* 2021;**12**(1): 2439.
4. Fortin J, Hansen K. Reconstructing a/B compartments as revealed by hi-C using long-range correlations in epigenetic data. *Genome Biol* 2015;**16**(180): 180.
5. Nichols MH, Corces VG. Principles of 3D compartmentalization of the human genome. *Cell Rep* 2021;**35**(13): 109330.
6. Corbo M, Damas J, Bursell MG, Lewin HA. Conservation of chromatin conformation in carnivores. *PNAS* 2022;**119**(9): e2120555119.
7. Feurtey A, Lorrain C, Croll D, *et al.* Genome compartmentalization predates species divergence in the plant pathogen genus Zymoseptoria. *BMC Genomics* 2020;**21**:588.
8. Prost J, Cameron C, Blanchette M. SACSANN: identifying sequence-based determinants of chromosomal compartments. *bioRxiv* 2020. https://doi.org/10.1101/2020.10.06.328039.
9. Krijger PHL, de Laat W. Regulation of disease-associated gene expression in the 3D genome. *Nat Rev Mol Cell Biol* 2016;**17**(12): 771–82.
10. Gorkin DU, Qiu Y, Hu M, *et al.* Common DNA sequence variation influences 3-dimensional conformation of the human genome. *Nat Rev Mol Cell Biol* 2019;**20**(1): 255.
11. Krumm A, Duan Z. Understanding the 3D genome: emerging impacts on human disease. *Semin Cell Dev Biol* 2019;**90**:62–77.
12. Degn H, Holden AV, Olsen LF. Chaos in Biological Systems. In: Degn H, Holden AV, Olsen LF (eds). *NATO Advanced Research Workshop on "Chaos in Biological Systems" December 8–12, 1986*. St. Nicholas, Cardiff, U. K.: Dyffryn House, 1987, 1–171.
13. Letellier C. *Chaos in Nature*, 2nd edn. Singapore: WORLD SCIENTIFIC, 2013.
14. Hewelt B, Li H, Jolly MK, *et al.* The DNA walk and its demonstration of deterministic chaos-relevance to genomic alterations in lung cancer. *Bioinformatics* 2019;**35**(16):2738–48.
15. Lorenz E. Deterministic nonperiodic flow. *J Atmospheric Sci* 1963;**20**(2):130–41.
16. Dias F, Iooss G. Chapter 10 - Water-Waves as a Spatial Dynamical System. In: Friedlander S, Serre D (eds). *Handbook of Mathematical Fluid Dynamics. vol. 2 of Handbook of Mathematical Fluid Dynamics*. North-Holland, 2003, 443–99.
17. Morfu S, Marquié P, Nofiele B, *et al.* Nonlinear Systems for Image Processing. *Adv Imaging Electron Phys* 2008 01; https://doi.org/10.1016/S1076-5670(08)00603-4.
18. Poincaré H. Sur le problème des trois corps et les équations de la dynamique. *Acta Math* 1890;**13**:1–270.
19. Poincaré H. *Méthodes nouvelles de la mécanique céleste*. Paris: Gauthier-Villars et fils, 1854-1912.
20. Ruelle D. Strange attractors. *Math Intell* 1980;**2**:126–37.
21. Grebogi C, Ott E, Pelikan S, Yorke JA. Strange attractors that are not chaotic. *Phys D: Nonlinear Phen* 1984;**13**(1):261–8.
22. Lefranc M. The topology of deterministic chaos: stretching, squeezing and linking. *Phys Theor Comput Sci* 2007;**01**(7):71–90.
23. Lyapunov AM. *The General Problem of the Stability of Motion* PhD thesis. University of Moscow, 1892.
24. Mandelbrot BB. *Les objets fractals: forme, hasard et dimension*. Paris: Flammarion, 1975.
25. Boltzmann L. *Vorlesungen über Gastheorie. Bd. 2*. Leipzig: Verlag von Johann Ambrosius Barth, 1889.
26. Birkhoff GD. Proof of the ergodic theorem. *Proc Natl Acad Sci* 1931;**17**(12):656–60.
27. Neumann J. Proof of the quasi-ergodic hypothesis. *Proc Natl Acad Sci* 1932;**18**(1):70–82.
28. Shields PC. String matching: the ergodic case. *Ann Prob* 1992;**20**(3):1199–203.
29. Falconnet M, Gantert N, Saada E. Ergodicity of some dynamics of DNA sequences. *arXiv*. 2019.
30. Shannon CE. *An algebra for theoretical genetics* PhD thesis. Massachusetts Institute of Technology - Department of Mathematics, 1940.
31. Chanda P, Costa E, Hu J, *et al.* Information theory in computational biology: where we stand today. *Entropy* 2020;**22**(6).
32. Lobzin VV, Chechetkin VR. Order and correlations in genomic DNA sequences. *The Spectral Approach Phys-Uspekhi* 2000;**43**: 55–78.
33. Weighill D, Macaya-Sanz D, DiFazio SP, *et al.* Wavelet-based genomic signal processing for centromere identification and hypothesis generation. *Front Genet* 2019;**10**:487.
34. Yin C, Chen Y, Yau STS. A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering. *J Theor Biol* 2014;**359**:18–28.

35. S Vinga. Information theory applications for biological sequence analysis. *Brief Bioinform* 2014;**15**:376–89.
36. Yoo J, Park S, Maffeo C, *et al.* DNA sequence and methylation prescribe the inside-out conformational dynamics and bending energetics of DNA minicircles. *Nucleic Acids Res* 2021;**49**(20): 11459–75.
37. Regan K, Dotterweich R, Ricketts S, Robertson-Anderson RM. Diffusion and conformational dynamics of single DNA molecules crowded by cytoskeletal proteins. *J Undergraduate Rep Phys* 2018;**28**(1): 100005.
38. Nishio T, Yoshikawa Y, Yoshikawa K. Higher-order structure of DNA determines its positioning in cell-size droplets under crowded conditions. *PloS One* 2021;**16**(12): e0261736.
39. Lainscsek C, Sejnowski TJ. Delay differential analysis of time series. *Neural Comput* 2015;**23**(3):594–614.
40. Takens F. Detecting strange attractors in turbulence. In: Rand D, Young LS (eds). *Dynamical Systems and Turbulence, Warwick 1980*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1981, 366–81.
41. Aguirre I, Letellier C. Investigating observability properties from data in nonlinear dynamics. *Phys Rev E Stat Nonlin Soft Matter Phys* 2020;**83**.
42. Gonzalez CE, Lainscsek C, Sejnowski TJ, Letellier C. Assessing observability of chaotic systems using delay differential analysis. *Chaos* 2020;**30**:103113.
43. Lainscsek C, Cash SS, Sejnowski TJ, Kurths J. Dynamical ergodicity DDA reveals causal structure in time series. *Chaos* 2021;**31**:103108.
44. Eckmann JP, Oliffson Kamphorst S, Ruelle D. Recurrence plots of dynamical systems. *Europhys Lett (EPL)* 1987;**4**(9): 973–7.
45. Lainscsek C, Weyhenmeyer J, Cash SS, Sejnowski TJ. Delay differential analysis of seizures in multichannel Electrocorticography data. *Neural Comput* 2017;**29**(12):3181–218.
46. Lainscsek C, Gonzalez CE, Sampson AL, *et al.* Causality detection in cortical seizure dynamics using cross-dynamical delay differential analysis. *Chaos* 2019;**29**:101103.
47. Lainscsek C, Rungratsameetaweemana R, Cash SS, Sejnowski TJ. Cortical chimera states predict epileptic seizures. *Chaos* 2019;**29**:121106.
48. Sanborn AL, Rao SS, Huang SC, *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* 2015;**112**(47): E6456–65.
49. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**(7414): 57–74.
50. Karolchik D, Hinrichs AS, Furey TS, *et al.* The UCSC table browser data retrieval tool. *Nucleic Acids Res* 2004;**32**(32): 493D–6.
51. Dixon JR, Selvaraj S, Yue F, *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**(7398):376–80.
52. Parry AJ, Hoare M, Bihary D, *et al.* NOTCH-mediated non-cell autonomous regulation of chromatin structure during senescence. *Nat Commun* 2018;**9**(1): 1840.
53. Ben L, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012;**9**:357–9.
54. Wolff J, Rabbani L, Gilsbach R, *et al.* Galaxy HiCExplorer 5: a web server for reproducible hi-C, capture hi-C and single-cell hi-C data analysis, quality control and visualization. *Nucleic Acids Res* 2020;**48**(W1): W177–84.
55. Wolff J, Bhardwaj V, Nothjunge S, *et al.* Galaxy HiCExplorer: a web server for reproducible hi-C data analysis, quality control and visualization. *Nucleic Acids Res* 2018;**46**(W1): W11–6.
56. Ramírez F, Bhardwaj V, Arrigoni L, *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* 2018;**9**:189.
57. Knight AP, Ruiz D. A fast algorithm for matrix balancing. *IMA J Numer Anal* 2007;33.
58. Buldyrev SV, Goldberger AL, Havlin S, *et al.* Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 1995;**51**(5):5084–91.
59. Heinz S, Benner C, Spann N, *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;**38**:576–89.
60. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. *Epigenet Chromatin* 2014;**7**(33).
61. Lainscsek C, Hernandez ME, Weyhenmeyer J, *et al.* Non-linear dynamical analysis of EEG time series distinguishes patients with Parkinson's disease from healthy individuals. *Front Neurol* 2013;**4**:4.
62. Sampson AL, Lainscsek C, Gonzalez CE, *et al.* Delay differential analysis for dynamical sleep spindle detection. *J Neurosci Methods* 2019;**316**:12–21.
63. Kirchhof M, Cameron CJ, Kremer SC. End-to-end chromosomal compartment prediction from reference genomes. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, 50–7.
64. Zhou J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genetics* 2022;**54**:725–34.
65. Krietenstein N, Abraham S, Venev S, *et al.* Ultrastructural details of mammalian chromosome architecture. *Mol Cell* 2020;**78**(3): 554–565.e7.
66. Belokopytova P, Fishman V. Predicting genome architecture: challenges and solutions. *Front Genet* 2021;**11**:617202.
67. Peng CK, Buldyrev SV, Goldberger AL, *et al.* Long-range correlations in nucleotide sequences. *Nature* 1992;**356**:168–70.
68. Mendizabal-Ruiz G, Román-Godínez I, Torres-Ramos S, *et al.* On DNA numerical representations for genomic similarity computation. *PloS One* 2017;**12**(3):1–27.
69. Haimovich AD, Byrne B, Ramaswamy R, Welsh WJ. Wavelet analysis of DNA walks. *J Comput Biol* 2006;**13**(7): 1289–98.
70. Berger JA, Mitra SK, Carli M, Neri A. Visualization and analysis of DNA sequences using DNA walks. *J Franklin Inst* 2004;**341**(1): 37–53.
71. Kwan HK, Arniker SB. *Numerical representation of DNA sequences*. In: IEEE International Conference on Electro-Information Technology, 2009, 307–10.
72. Kumar GS. *DNA Sequence Representation methods*. In: ISB Proceedings of the International Symposium on Biocomputing, 2009, 1–4.
73. Zhang L, Jiang Z. Long-range correlations in DNA sequences using 2D DNA walk based on pairs of sequential nucleotides. *Chaos Solitons Fractals* 2004;**22**(4):947–55.
74. Karlsson M, Zhang C, Méar L, *et al.* A single-cell type transcriptomics map of human tissues. *Sci Adv* 2021;**7**(31).
75. Lainscsek C, Weyhenmeyer J, Hernandez ME, *et al.* Non-linear dynamical classification of short time series of the rössler system in high noise regimes. *Front Neurol* 2013;**4**:182.