Briefings in Bioinformatics, 2022, **23(4)**, 1–11

https://doi.org/10.1093/bib/bbac168

Advance access publication date: 10 May 2022

Problem Solving Protocol

Persistent spectral simplicial complex-based machine learning for chromosomal structural analysis in cellular differentiation

Weikang Gong, JunJie Wee, Min-Chun Wu, Xiaohan Sun, Chunhua Li and Kelin Xia

Corresponding authors: Kelin Xia, Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, E-mail: xiakelin@ntu.edu.sg; Chunhua Li, College of Life Science and Chemistry, Faculty of Environmental and Life Sciences, Beijing University of Technology, China 100124, E-mail: chunhuali@bjut.edu.cn; Weikang Gong, College of Life Science and Chemistry, Faculty of Environmental and Life Sciences, Beijing University of Technology, China 100124, E-mail: weikanggong@emails.bjut.edu.cn and Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, E-mail: N2007046B@e.ntu.edu.sg

Abstract

The three-dimensional (3D) chromosomal structure plays an essential role in all DNA-templated processes, including gene transcription, DNA replication and other cellular processes. Although developing chromosome conformation capture (3C) methods, such as Hi-C, which can generate chromosomal contact data characterized genome-wide chromosomal structural properties, understanding 3D genomic nature-based on Hi-C data remains lacking. Here, we propose a persistent spectral simplicial complex (PerSpectSC) model to describe Hi-C data for the first time. Specifically, a filtration process is introduced to generate a series of nested simplicial complexes at different scales. For each of these simplicial complexes, its spectral information can be calculated from the corresponding Hodge Laplacian matrix. PerSpectSC model describes the persistence and variation of the spectral information of the nested simplicial complexes during the filtration process. Different from all previous models, our PerSpectSC-based features provide a quantitative global-scale characterization of chromosome structures and topology. Our descriptors can successfully classify cell types and also cellular differentiation stages for all the 24 types of chromosomes simultaneously. In particular, persistent minimum best characterizes cell types and Dim (1) persistent multiplicity best characterizes cellular differentiation. These results demonstrate the great potential of our PerSpectSC-based models in polymeric data analysis.

Keywords: Hi-C data, Hodge Laplacian, persistent spectral simplicial complex, chromosomal featurization, machine learning

Introduction

The genome's three-dimensional (3D) architecture within the cell nucleus is thought to play a crucial role in many biological processes, including gene regulation, cell replication and cell differentiation [1–11]. Recently, various studies have shown that changes in chromosomal structure at specific genomic regions and under certain conditions are associated with cell development and differentiation [12–15]. Thus, a comprehensive understanding of the chromosomal 3D structure is of fundamental significance to the decryption and interpretation of genetic information and has become one of the most important topics in genomics and epigenetic research.

Chromosome conformation capture (3C) [1, 18] and its derived techniques, including chromosome conformation capture-on-chip (4C) [19], chromosome conformation capture carbon copy (5C) [20] and high-throughput chromosome conformation capture (Hi-C) [2], have become widely used to generate genome-wide chromatin interaction maps [21]. The element in chromosome contact matrices is the frequency of contacts between pairs of chromosome loci within a population of cells. The matrices are usually visualized as a square heatmap, then analyzed to identify chromatin structures [22]. As increasing numbers of contact matrices are published in different cell types and various stages of differentiation, the Hi-C technique provides

Weikang Gong is a PhD student from the Beijing University of Technology, China. His research interests are chromosome structure analysis, geometry and topological data analysis (TDA), machine learning and protein/RNA/protein-RNA complex structure-dynamics relationship. From 2021 to 2022, he was a visiting student at the Nanyang Technological University in Singapore.

JunJie Wee is a PhD student from Nanyang Technological University, Singapore. His research interests are molecular data analysis, geometry and TDA and machine learning.

Min-Chun Wu is a postdoctoral fellow from Nanyang Technological University, Singapore. His research interests are TDA and machine learning.

Xiaohan Sun is a PhD student from the Beijing University of Technology, China. Her research interests are protein-RNA recognition and interaction.

Chunhua Li is a professor at the Beijing University of Technology, China. Her research interests include protein dynamics, protein folding and protein–RNA interaction.

Kelin Xia is an assistant professor at Nanyang Technological University, Singapore. His research interests are TDA, molecular-based mathematical biology and machine learning.

Received: February 21, 2022. **Revised:** April 12, 2022. **Accepted:** March 13, 2022

crucial information necessary for distinguishing the 3D shape of chromosomes or genomes.

To understand the polymer mechanisms underlying the complex spatial organization of chromosomes, various polymer models and computational methods have been developed [2, 4, 12, 23–34]. Xu *et al.* [23] developed an accurate and fast method, called FastHiC, to detect long-range chromatin interactions based on a novel implementation of the simulated field approximation from Hi-C data. More importantly, Sauerwald *et al.* [25] adapted the Gaussian Network Model (GNM) [35] to model chromatin dynamics using Hi-C data, which accesses the structural basis of genome-wide observations. Furthermore, Zhang *et al.* [27] followed GNM to reveal differences in the intrinsic spatial dynamics of the chromatin across different cell lines. More specifically, in order to study the variation in chromosomal structures between different cell types, Zhou *et al.* [26] described a single-cell clustering algorithm, called scHiCluster, for Hi-C contact matrices which is based on imputations using linear convolution and random walk. Recently, Sauerwald *et al.* [36] applied Topological Data Analysis (TDA) to study chromosomal structure through differentiation across three cell lines and identify persistent connected components and one-dimensional circles or loops topological features of chromosomes.

Here, we develop persistent spectral simplicial complex-based machine learning (PerSpectSC-ML) for chromosomal structural classification in cellular differentiation. We model the chromosomal structures and interactions as simplicial complexes. By using a filtration process, a series of simplicial complexes at various scales are systematically generated for a chromosomal structure. Based on them, we develop PerSpectSC-based chromosomal descriptors from the statistical and combinatorial properties of PerSpectSC. These descriptors are combined with *t*-SNE-assisted *k*-means for the classification of cell types. Our PerSpectSC-ML models have achieved great accuracy on 14 cell types representing various cell lines. Meanwhile, many of the patterns representing various stages of differentiation can be observed by chromosomal descriptors from PerSpectSC models.

Results

Persistent spectral simplicial complex

Different from the traditional graph- and network-based models, chromosomal structures and interactions are considered as simplicial complexes [37] in our PerSpectSC models. Mathematically, a simplicial complex, which is composed of simplices, can be viewed as a generalization of the graph into its higher-dimensional counterparts. A graph is composed of vertices (0-simplices) and edges (1-simplices), whereas a simplicial complex is made from 0-simplices, 1-simplices, 2-simplices (triangles), 3-simplices (tetrahedrons) and other higher-dimensional simplices (See MATERIALS AND METHODS for details). Physically, a graph characterizes pair-wise interactions through edges, whereas a simplicial complex can describe many-body interactions using simplices.

One of the core elements for our PerSpectSC model is the Hodge-Laplacian-based spectral model. Chromosome structures can be modeled as different topological representations, including graphs, simplicial complexes and hypergraphs. Based on them, Hodge Laplacian matrices at different dimensions can be constructed. Here, we consider the simplicial complex representation. The k-dimensional Hodge Laplacian matrix (k is a non-negative integer), denoted as $\mathbf{L}_k$, describes topological connections between k-simplices. From these matrices, the spectral information, i.e. eigenvalues and eigenvectors, can be calculated and further used for the characterization of the structural properties.

The other core element for our PerSpectSC model is the filtration process. During a filtration process, a series of multiscale topological representations are generated [38]. At a lower filtration value, points (bins) are not 'connected', i.e. separated from each other, indicating a higher resolution focusing on local details. At a higher filtration value, points (bins) are 'overlapped' with each other and 'connect' into a topological structure, indicating a lower resolution that captures global information. With the increase (or decrease) of filtration value, there is a continuous change of scales, thus a multiscale representation can be achieved.

PerSpectSC models focus on persistence and variations of spectral information during a filtration process. Persistent attributes, which are defined as the statistic and combinatorial properties of the eigenvalues during the filtration process, are used as quantitative features for structural characterization. The PerSpectSC-based features can be combined with both unsupervised and supervised models for the data analysis.

PerSpectSC-based Hi-C data characterization

Here, we develop PerSpectSC-based descriptors for chromosome structural classification for the first time. The essential idea is to model chromosome structures as simplicial complexes and use persistent attributes to characterize their intrinsic structure properties.

Hodge Laplacian matrix can be generated from the simplical complexes. We denote $\mathbf{L}_k$ as kth dimensional Hodge Laplacian matrix. For Dim(0), at the very start of the filtration, there are only 30 vertices (0-simplex), and a 30*30 all-zero $\mathbf{L}_0$ matrix is generated according to Eq. (4) (see MATERIALS AND METHODS). As the increase of filtration value, the size of $\mathbf{L}_0$ matrices remains unchanged, while more and more entries with −1 value appear at its off-diagonal part. When the filtration value is large enough, a complete graph is obtained, and a full $\mathbf{L}_0$ matrix, i.e. all diagonal entries are 29 and all off-diagonal entries are −1, is generated according to Eqs. (4, 5). For Dim(1), at the early stage of filtration, there exists no edges (1-simplices) thus no $\mathbf{L}_1$ matrix. With edges

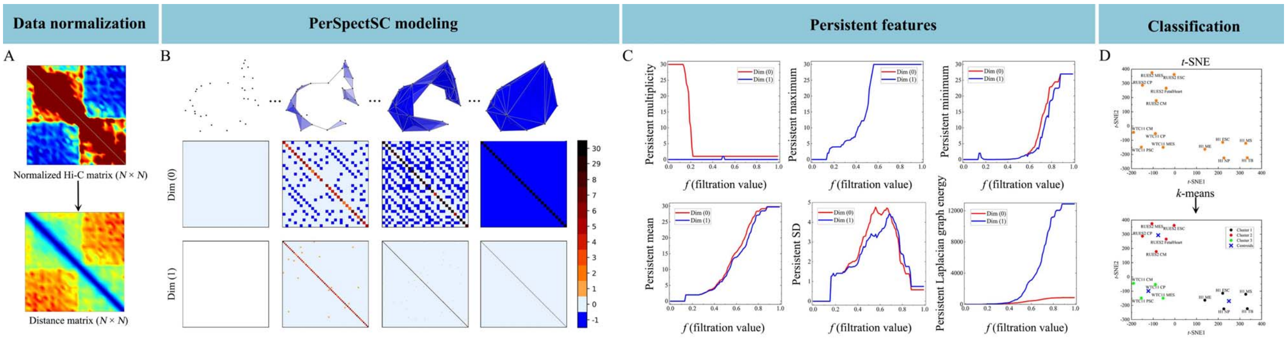

Fig. 1. An illustration of filtration process and PerSpectSC models for 30 bins range from 47.3 to 50.2 Mb of chromosome 22 of RUES2 CM. **(A)** Normalized Hi-C matrix is converted to distance matrix. Normalized Hi-C matrix is computed from the reads through the HiC-Pro pipeline [16], and sample is tested for quality at 100 kb resolution. *N* is genomic bins. Distance matrix is computed through Eq. (8). **(B)** Distance matrix is converted to point cloud by classical multidimensional scaling (CMDS) [17], and the nested sequence of simplicial complexes is constructed. From distance matrix, Hodge Laplacian matrices ($\mathbf{L}_0$ and $\mathbf{L}_1$) as in Eq. (4) can be systematically obtained at each filtration value. **(C)** Six persistent attributes obtained from the PerSpectSC models. Note that persistent multiplicity is equivalent to persistent Betti numbers. **(D)** Results of *t*-SNE and *t*-SNE-assisted *k*-means based on Dim (1) persistent multiplicity of chromosome Y.

emerging as the filtration value increases, $\mathbf{L}_1$ matrices are generated. Different from Dim(0) case, the size of $\mathbf{L}_1$ matrices increases with the number of edges. Off-diagonal entries can be 1 and −1 depending on the edge orientation as in Eq. (4, 6). When the filtration value is large enough, all edges will be either upper adjacent or not lower adjacent (The definitions of 'upper' and 'lower adjacent' see Eq. (6)); thus, $\mathbf{L}_1$ matrix becomes a diagonal matrix with all its diagonal entries as 30. Mathematically, higher-dimensional Hodge Laplacian matrices can also be generated.

Persistent attributes can be obtained from the filtration process. Figure 1C shows the persistent multiplicity, persistent maximum, persistent minimum, persistent mean, persistent SD and persistent Laplacian graph energy for 47.3 to 50.2 Mb of chromosome 22 of RUES2 CM. It can be seen that these persistent attributes change with the filtration value. Each variation of persistent attributes indicates a certain change of the simplicial complexes. Note that the persistent multiplicity is equivalent to the persistent Betti number or Betti curve. In this way, the persistent homology (PH) [37] information is naturally embedded into persistent multiplicity. At filtration size 1.00, a complete two-dimensional simplicial complex is achieved, i.e. any 3 vertices can form a 2-simplex. The corresponding $\mathbf{L}_0$ has eigenvalues 0 and 30. The size for the corresponding $\mathbf{L}_1$ is 435*435, and its eigenvalues are all 30. Note that $435 = C_{30}^2$.

PerSpectSC-ML for classification of different cell types

To validate the efficiency of our PerSpectSC-based features, we develop PerSpectSC-based machine learning models and use them for chromosome structural clustering. A general machine learning approach, i.e. 'dimensionality reduction + *k*-means', is considered. Figure 1 illustrates the general procedure of this ML model.

The classification of different cell types is a key step in an important regulator of gene expression [12, 39]. An accurate classification requires a better characterization of geometric structures of Hi-C data. Here, Figure 2 compares the performance of our PerSpect SC-ML for each persistent attribute per chromosome for 14 cell types classification. From the perspective of chromosomes, it can be seen that the PerSpectSC-ML performance of chromosome Y can achieve state-of-the-art results. More importantly, from the perspective of persistent attributes, it can be seen that PerSpectSC-ML performance of persistent minimum can achieve state-of-the-art results, which is better than the results of topological information (Dim (0) and Dim (1) persistent multiplicity) [36]. It is worth mentioning that the performance of persistent maximum is slightly better than that of Dim (1) persistent multiplicity. Meanwhile, the performances of persistent generalized mean graph energy, persistent Laplacian graph energy, persistent mean, persistent moment (second-order) and persistent standard deviation are comparable to that of Dim (1) persistent multiplicity.

Additionally, Figure 3 visualizes the results of PerSpect SC-ML models of Dim (0) persistent multiplicity, Dim (1) persistent multiplicity and persistent minimum for chromosomes 13, 21 and Y, respectively. From the perspective of chromosome 13, there are 7 cells (WTC11 CM, RUES2 MES, WTC11 MES, H1 NP, RUES2 FetalHeart, H1 ME and H1 ESC) that are misclassified by Dim (0) persistent multiplicity. Similarly, there are also 7 cells that are misclassified by Dim (0) persistent multiplicity based on PCA-assisted *k*-means (see Figure S1A, see Supplementary Data available online at http://bib.oxfordjournals.org/). There are 5 cells (H1 NP, RUES2 CP, RUES2 ME, RUES2 ESC and RUES2 CM) that are misclassified by Dim (1) persistent multiplicity. There are only 2 cells (WTC11 CP and RUES2 ESC) which are misclassified by persistent minimum. From the perspective of chromosome 21, there are 4 cells (RUES2 CM, RUES2 MES, WTC11 MES and WTC11 CM) that are misclassified by Dim (0) persistent multiplicity. There are 6 cells (RUES2 FetalHeart, WTC11 PSC, WTC11 CP, H1 ESC, RUES2 CM and H1 NP) which are misclassified by Dim (1) persistent multiplicity, whereas

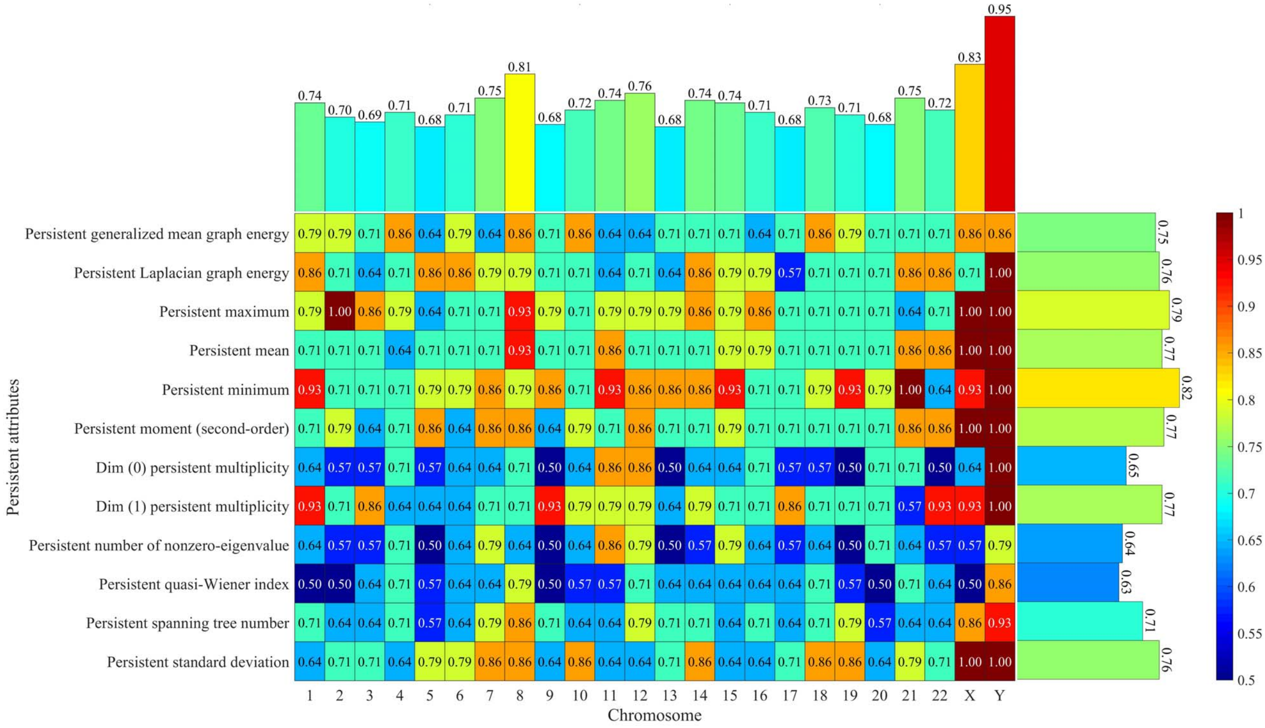

Fig. 2. Performance comparison of PerSpectSC-ML's results for each persistent attribute per chromosome for 14 cell types classification. The above histogram represents the average value over 12 persistent attributes. The right histogram represents the average value over 24 chromosomes.

persistent minimum of chromosome 21 can completely separate three different cell types. These results further demonstrate the performance of persistent minimum is better than the results of topological information (Dim (0) and Dim (1) persistent multiplicity). Furthermore, Dim (0) persistent multiplicity, Dim (1) persistent multiplicity and persistent minimum of the chromosome Y can also completely separate three different cell types. Based on PCA-assisted *k*-means, the persistent minimum of the chromosome Y can also completely separate three different cell types (see Figure S1B, see Supplementary Data available online at http://bib.oxfordjournals.org/). Meanwhile, we also compare *t*-SNE-assisted *k*-means's performance based on persistent minimum of chromosome Y when the perplexity hyperparameters of *t*-SNE are 2, 3, 4 and 5, respectively (see Figure S2, see Supplementary Data available online at http://bib.oxfordjournals.org/). These four *t*-SNE-assisted *k*-means's results also separate three cell types fully. Similarly, this result further confirms that the PerSpectSC-ML performance of chromosome Y can achieve state-of-the-art results. Other than chromosomal structures, the PerSpectSC-ML model can be used in the analysis of molecular structure from proteins, DNAs and RNAs [40]. It is suitable for structure representation.

PerSpectSC-ML for classification of H1 cells different stages

In order to further distinguish different stages of H1 cells, normalized areas of persistent attributes were compared for all chromosomes (see Figure 4 and Figure S3, see Supplementary Data available online at http://bib.oxfordjournals.org/) through Eq. (9). The normalized area of Dim (0) persistent multiplicity can not distinguish the 5 stages of H1 (see Figure 4A). Excluding ESC and ME stages, the normalized area of persistent generalized mean graph energy can well distinguish TB, NP and MS stages of H1 (see Figure 4B). More interestingly, Dim (1) persistent multiplicity is consistent with the result of persistent generalized mean graph energy (see Figure 4C). This may be due to the large difference in the number of loops per chromosome [41]. Excluding persistent minimum and persistent number of nonzero-eigenvalue, normalized areas of other persistent attributes distinguish stages of H1 to some extent (see Figure S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). These results further demonstrate that Dim (1) persistent multiplicity obtained from the PerSpectSC model at the genome-wide scale can characterize differentiation stages. There are huge differences in the number of loops of chromosomes in different differentiation stages. Thus, compared with the bottleneck distance which quantifies the difference between two persistence diagrams [36], the normalized area of persistent attributes in our PerSpectSC models can identify the general global topological changes during different stages of H1 cells for the entire 24 chromosomes.

Discussion

The present comparative study of the intrinsic properties of chromosomes in a series of cell lines using

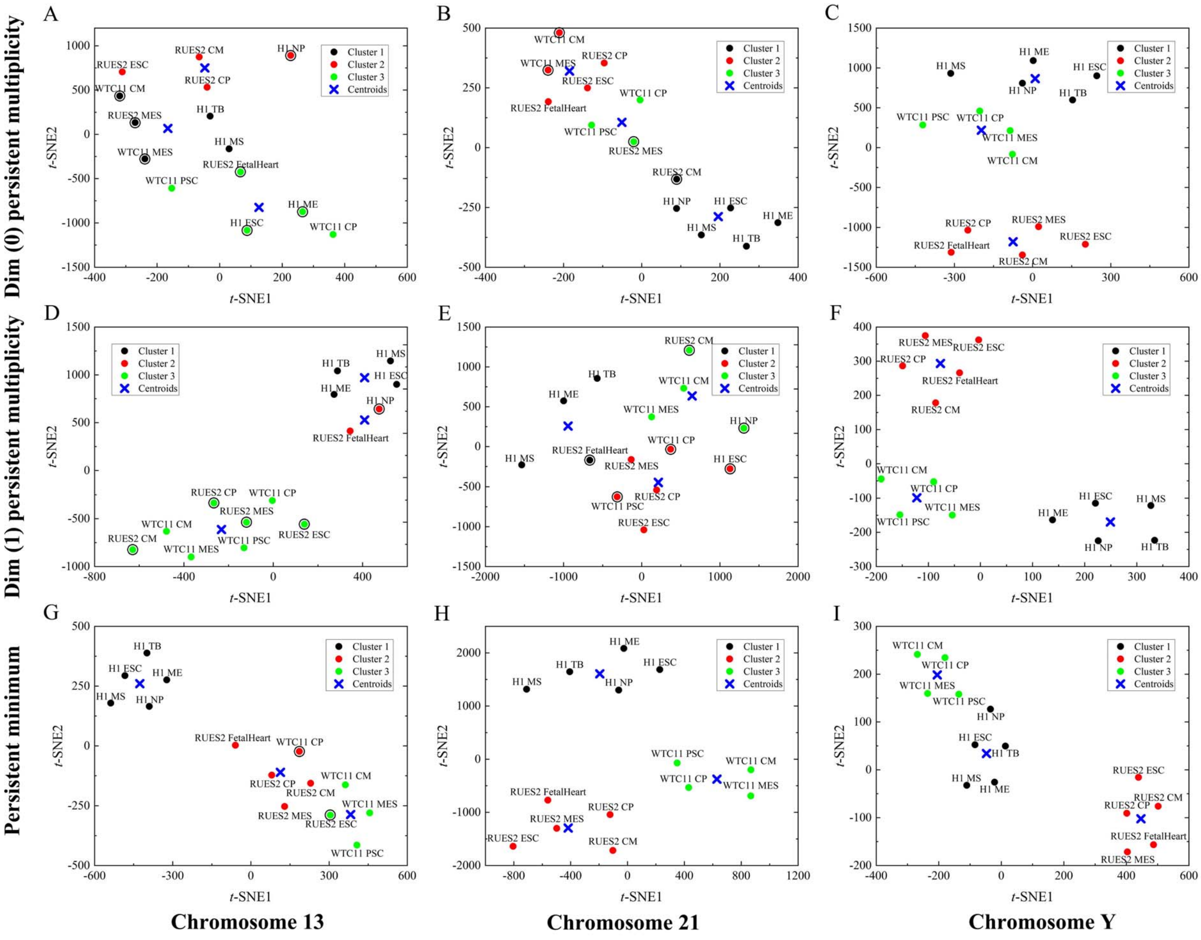

Fig. 3. Performance comparison of PerSpectSC-ML's results obtained on 3 persistent attributes of 3 chromosomes. **(A-C)** represent the results of Dim (0) persistent multiplicity for chromosome 13, chromosome 21 and chromosome Y, respectively. **(D-F)** represent the results of Dim (1) persistent multiplicity for chromosome 13, chromosome 21 and chromosome Y, respectively. **(G-I)** represent the results of persistent minimum for chromosome 13, chromosome 21 and chromosome Y, respectively. Misclassified cells are circled.

corresponding Hi-C data in the PerSpectSC shed light on several persistent attributes, including the topological information. A combinational *t*-SNE-assisted *k*-means for adapting PerSpectSC is constructed and the state-of-the-art results are obtained. By analyzing the results of PerSpectSC-ML of 14 samples from various cell lines and stages of differentiation, we identify generative principles of chromosomal structure. Based on the PerSpectSC-ML model, our descriptors have the best classification power for cell lines on chromosome Y. Over the 12 persistent attributes, the persistent minimum has the best overall performance on all the 24 types of chromosomes. Dim (1) persistent multiplicity, rather than Dim (0) persistent multiplicity, can characterize cell lines and stages of differentiation. PerSpectSC model shows promise for further analysis of Hi-C data, especially as computational limitations are overcome, permitting analysis of higher dimensional features at higher resolution.

For a more in-depth discussion, the proposed PerSpectSC model provides a highly efficient and effective way for chromosomal representation. Compared with traditional featurization, the PerSpectSC model has several advantages. First, the PerSpectSC model can not only calculate the topological invariants (Betti number) of chromosomes but also obtain other spectral information. Second, the PerSpectSC model captures the spectral information from various different scales through an expansion process. Thus, spectral information has multiscale properties. Third, the PerSpectSC model considers the physics of higher-order interactions within chromosome systems, such as simplicial complexes. Last, the PerSpectSC model is convenient to combine with machine learning for structural data analysis.

Materials and Methods

Dataset

The Hi-C datasets used in this study are obtained from the Kingsford Group [36]. The detailed Hi-C data information can be found in Table 1, including accession codes. Cell samples from all 14 conditions included two replicates each. All of the samples were processed from raw reads to normalized contact matrices at 100 kb using the HiC-Pro pipeline [16] and iterative correction and

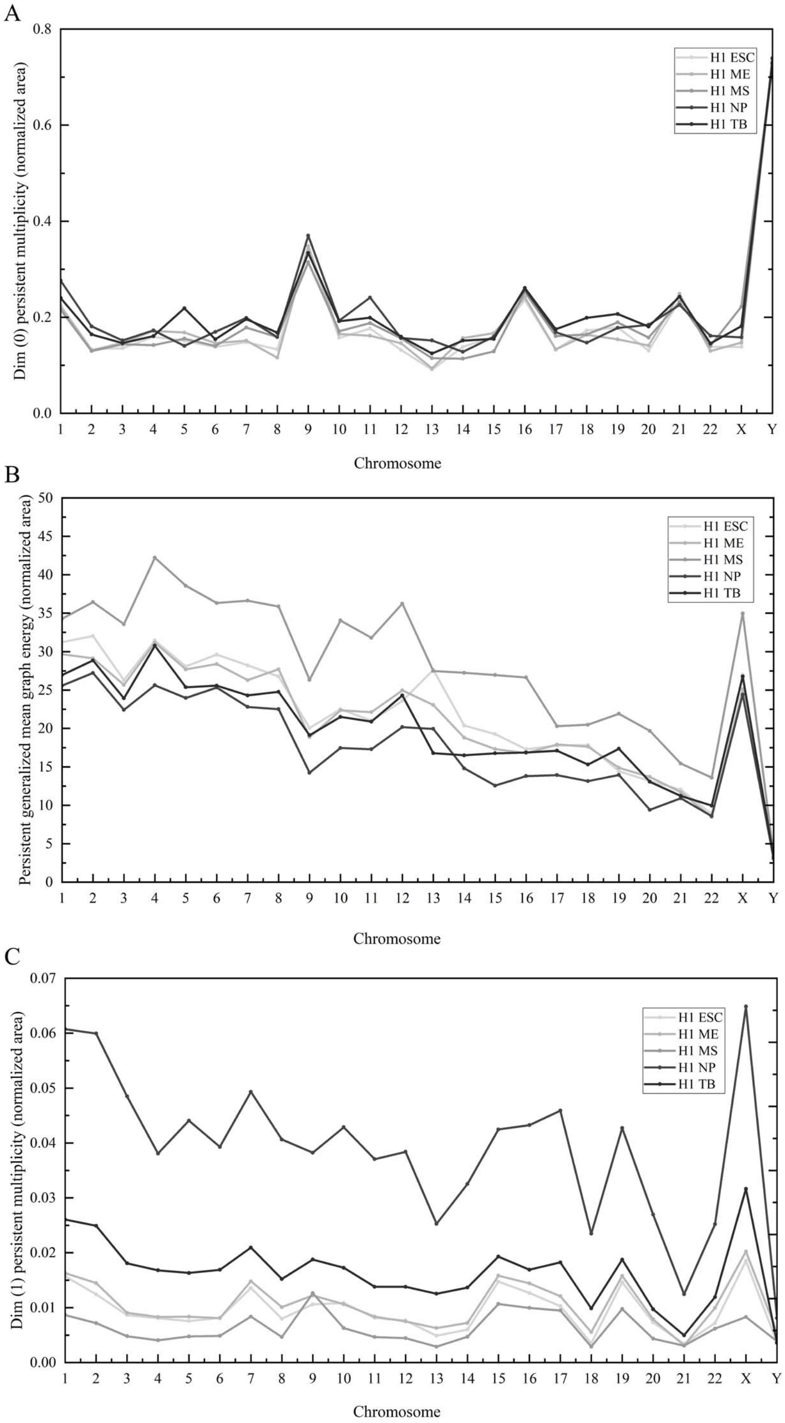

Fig. 4. Performance comparison of normalized areas obtained from Dim (0) persistent multiplicity **(A)**, persistent generalized mean graph energy **(B)** and Dim (1) persistent multiplicity **(C)** of 24 chromosomes for H1 ESC, H1 ME, H1 MS, H1 NP and H1 TB.

eigenvector decomposition (ICE) normalization [42]. To maximize coverage, all of the reads from replicates were combined to produce one Hi-C matrix per sample.

Simplicial complex

Simplicial complexes have been applied to map the real organization of various material, biological and chemical systems [43]. The simplicial complex is composed of simplices. Each simplex is a finite set of vertices and can be viewed geometrically as, a point (0-simplex), an edge (1-simplex), a triangle (2-simplex), a tetrahedron (3-simplex) and their k-dimensional counterpart (k-simplex). More specifically, a k-simplex $\sigma^k = \{v_0, v_1, v_2, \cdots, v_k\}$ is the convex hull formed by $k+1$

Table 1. All Hi-C datas used for this study

Cell type	Description	SRA Accessions
H1 ESC	embryonic stem cell	SRX378271, SRX378272
H1 ME	mesendoderm	SRX378273, SRX378274
H1 MS	mesenchymal stem cell	SRX378275, SRX378276
H1 NP	neural progenitor	SRX378277, SRX378278
H1 TB	trophoblast-like cells	SRX378279, SRX378280
RUES2 CM	cardiac myocyte	SRX3375353, SRX3375354
RUES2 CP	cardiac progenitor	SRX3375351, SRX3375352
RUES2 ESC	embryonic stem cell	SRX3375347, SRX3375348
RUES2 FetalHeart	fetal heart tissue	SRX3375355, SRX3375356
RUES2 MES	mesoderm	SRX3375349, SRX3375350
WTC11 CM	cardiac myocyte	SRX4958487, SRX4958488
WTC11 CP	cardiac progenitor	SRX4958485, SRX4958486
WTC11 MES	mesoderm	SRX4958483, SRX4958484
WTC11 PSC	pluripotent stem cell	SRX4958481, SRX4958482

affinely independent points $v_0, v_1, v_2, \cdots, v_k$ as follows:

$$\sigma^k = \left\{ \lambda_0 v_0 + \lambda_1 v_1 + \cdots + \lambda_k v_k \mid \sum_{i=0}^{k} \lambda_i = 1; \forall i, 0 \le \lambda_i \le 1 \right\}. \tag{1}$$

The ith dimensional face of σ^k ($i < k$) is the convex hull formed by $i+1$ vertices from the set of $k+1$ points $v_0, v_1, v_2, \cdots, v_k$. The simplices are the basic components for a simplicial complex.

A simplicial complex K is a finite set of simplices that satisfy two conditions. Firstly, any face of a simplex from K is also in K. Secondly, the intersection of any two simplices in K is either empty or a shared face. A kth chain group C_k is the free Abelian group generated by oriented k-simplices, which are simplices together with an orientation, i.e. one of the two classes of permutations of the vertex set of a simplex. The boundary operator ∂_k ($C_k \to C_{k-1}$) for an oriented k-simplex σ^k is defined by

$$\partial_k \sigma^k = \sum_{i=0}^{k} (-1)^i [v_0, v_1, v_2, \cdots, \hat{v}_i, \cdots, v_k]. \tag{2}$$

Here, $[v_0, v_1, v_2, \cdots, \hat{v}_i, \cdots, v_k]$ is an oriented $(k-1)$-simplex, generated by the original set of vertices except v_i. The boundary operator maps a simplex to its faces and it guarantees that $\partial_{k-1}\partial_k = 0$. The commonly used methods to define simplicial complexes are Vietoris-Rips (VR) complex, Čech Complex, Alpha complex, Clique complex, Cubic complex and Morse complex [37]. Among them, the VR complex is used in this study.

Spectral simplicial complex

The spectral simplicial complex theory characterizes the spectral properties of Hodge (or combinatorial) Laplacian matrices, which are constructed based on a simplicial complex [44, 45]. Computationally, for oriented simplicial complex, its kth boundary (or incidence) matrix $\mathbf{B}_k$ is defined as follows:

$$B_k(i,j) = \begin{cases} 1, & \text{if } \sigma_i^{k-1} \subset \sigma_j^k \sim \text{ and } \sim \sigma_i^{k-1} \sim \sigma_j^k \\ -1, & \text{if } \sigma_i^{k-1} \subset \sigma_j^k \sim \text{ and } \sim \sigma_i^{k-1} \nsim \sigma_j^k \\ 0, & \text{if } \sigma_i^{k-1} \not\subset \sigma_j^k. \end{cases} \tag{3}$$

Here, $\sigma_i^{k-1} \subset \sigma_j^k$ indicates that σ_i^{k-1} is a face of σ_j^k and $\sigma_i^{k-1} \not\subset \sigma_j^k$ indicates the opposite. The notation $\sigma_i^{k-1} \sim \sigma_j^k$ indicates that the given orientation of σ_i^{k-1} is the same that it has in the boundary of σ_j^k, and $\sigma_i^{k-1} \nsim \sigma_j^k$ indicates the opposite. With the relation of boundary operator $\partial_k \partial_{k+1} = 0$, these boundary matrices satisfy the condition that $\mathbf{B}_k \mathbf{B}_{k+1} = \mathbf{0}$.

The kth Hodge Laplacian matrix can be expressed as follows:

$$\mathbf{L}_k = \begin{cases} \mathbf{B}_1 \mathbf{B}_1^T & \text{if } k = 0 \\ \mathbf{B}_k^T \mathbf{B}_k + \mathbf{B}_{k+1} \mathbf{B}_{k+1}^T & \text{if } 0 < k < n \\ \mathbf{B}_n^T \mathbf{B}_n & \text{if } k = n. \end{cases} \tag{4}$$

Note that n represents the highest order of the simplicial complex K. More specifically, the above Hodge Laplacian matrices can be explicitly described in terms of the simplex relations. $\mathbf{L}_0$ can be expressed as

$$L_0(i,j) = \begin{cases} d(\sigma_i^0), & \text{if } i = j \\ -1, & \text{if } i \neq j \sim \text{ and } \sigma_i^0 \widehat{\sigma}_j^0 \\ 0, & \text{if } i \neq j \sim \text{ and } \sigma_i^0 \not\frown \sigma_j^0. \end{cases} \tag{5}$$

From Eq. (5), $\mathbf{L}_0$ is exactly the graph Laplacian matrix. Furthermore, when $k > 0$, $\mathbf{L}_k$ can be expressed as

$$L_k(i,j) = \begin{cases} d(\sigma_i^k) + k + 1, & \text{ifi} = j \\ 1, & \text{if } i \neq j, \sigma_i^k \not\frown \sigma_j^k, \sigma_i^k \smile \sigma_j^k \sim \text{ and } \sigma_i^k \sim \sigma_j^k \\ -1, & \text{if } i \neq j, \sigma_i^k \not\frown \sigma_j^k, \sigma_i^k \smile \sigma_j^k \sim \text{ and } \sigma_i^k \nsim \sigma_j^k \\ 0, & \text{if } i \neq j, \sigma_i^k \widehat{\sigma}_j^k \sim \text{ or} \sim \sigma_i^k \not\smile \sigma_j^k. \end{cases} \tag{6}$$

Here, $d(\sigma_i^k)$ is the (upper) degree of a k-simplex σ_i^k, i.e. the number of $(k+1)$-simplices, of which σ_i^k is a face. $\sigma_i^k \widehat{\sigma}_j^k$ means that the two simplices are upper adjacent, i.e. they are faces of a common $(k+1)$-simplex, and

$\sigma_i^k \not\frown \sigma_j^k$ means the opposite. Notation $\sigma_i^k \smile \sigma_j^k$ means that the two simplices are lower adjacent, i.e. they share a common $(k-1)$-simplex as their face, $\sigma_i^k \not\smile \sigma_j^k$ means the opposite. Notation $\sigma_i^k \sim \sigma_j^k$ means that the two simplices have the same orientation, i.e. oriented similarly, and $\sigma_i^k \not\sim \sigma_j^k$ means the opposite. The eigenvalues of Hodge Laplacian matrices are independent of the choice of the orientation [44]. All the eigenvalues are non-negative. More importantly, the multiplicity of zero-eigenvalues, i.e. the total number of zero-eigenvalues, of $\mathbf{L}_k$ equals to the kth Betti number β_k. Geometrically, β_0 represents the number of connected components of the VR complex, β_1 represents the number of one-dimensional circles or loops, and β_2 represents the number of two-dimensional voids or cavities. All the positive eigenvalues characterize detailed structure properties. As an example, we consider an oriented simplicial complex K_1 as in Figure S4 (see Supplementary Data available online at http://bib.oxfordjournals.org/) and explain it accordingly.

Persistent spectral simplicial complex

PerSpectSC models characterize the intrinsic topological and geometric information of data. PerSpectSC models do not consider the eigenspectrum information of the simplicial complex, constructed from data at a fixed scale; instead, they focus on the variation of the eigenspectrum of these topological representations during a filtration process.

Physically, a filtration process generates a multiscale representation of complex systems [46]. For instance, a filtration operation on a distance matrix, i.e. a matrix with distances between any two vertices as its entries, can be defined by using a cutoff value as the filtration parameter. More specifically, for simplicial complex, if the distance between two vertices is smaller than the cutoff value, an edge (1-simplex) is formed between them. Furthermore, a triangle (2-simplex), a tetrahedron (3-simplex) and their k-dimensional counterpart (k-simplex) are formed. In this way, a systematical increase (or decrease) of the cutoff value will deliver a series of nested simplicial complexes, with the simplicial complex generated at a lower cutoff value as a subset (or a part) of the simplicial complex generated at a larger cutoff value. Nested simplicial complexes can be constructed by using various definitions of complexes, such as VR complex, Čech Complex, Alpha complex, Clique complex, Cubic complex and Morse complex.

Mathematically, a filtration process can naturally induce a nested series of simplicial complexes at different scales as follows:

$$K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m. \tag{7}$$

Here, the ith simplicial complex K^i is generated at filtration value f_i. Computationally, we can equally divide the filtration region (of the filtration parameter) into m intervals and consider a topological representation at each interval. Hodge Laplacian matrix series $\{\mathbf{L}_k^i|_{i=1,2,\dots,m;k=0,1,2,\dots}\}$ can be constructed from these simplicial complexes $\{K^i\}$. Note that the size of these Laplacian matrices may be different.

Applying PerSpectSC to single-cell Hi-C data

PerSpectSC models use a distance matrix that describes the distances between all point cloud data. Although Hi-C data is interpreted as describing the 3D distances between chromosomal segments, the value of a Hi-C matrix is contact counts rather than distance values, where a high contact count implies a low distance. A normalized Hi-C matrix $\mathbf{C}$ is converted to a distance matrix $\mathbf{D}$ as follows:

$$D(i,j) = 1 - \begin{cases} 1, & \text{if } i = j \\ \frac{1}{m}\mathbf{log}(C(i,j)+1), & \text{if } i \neq j. \end{cases} \tag{8}$$

Here, $m = 1.01\mathbf{max}_{i,j\le N}(\mathbf{log}(C(i,j))+1)$, and N is the number of rows in the contact matrix $\mathbf{C}$. A pseudo-count of 1 is added to all off-diagonal values in the Hi-C matrix to avoid taking a logarithm of zero, and the factor of 1.01 is included to ensure that all distances where $i \neq j$ are non-zero. GUDHI [47], a Python library for TDA, is used to construct simplicial complex series $\{K^i\}$ at different filtration value f_i based on these transformational distance matrices $\mathbf{D}$ at 100 kb resolution. Based on the simplicial complex series $\{K^i\}$, PerSpectSC models can be constructed.

Persistent attributes

In order to reveal the principles of chromosomal structure in cellular differentiation, we propose a set of persistent attributes using statistical and combinatorial properties of PerSpectSC models. These persistent attributes are of the same sizes and can be used as chromosomal descriptors. More specifically, we consider 12 chromosomal descriptors as follows:

- persistent generalized mean graph energy $\left(\sum_{i=1}^{p}|\lambda_i - \bar{\lambda}|\right)$
- persistent Laplacian graph energy $\left(\sum_{i=1}^{p}\lambda_i\right)$
- persistent maximum ($\mathbf{max}\{\lambda_1, \lambda_2, \cdots, \lambda_p\}$)
- persistent mean $\left(\frac{1}{p}\sum_{i=1}^{p}\lambda_i\right)$
- persistent minimum ($\mathbf{min}\{\lambda_1, \lambda_2, \cdots, \lambda_p\}$)
- persistent moment (second-order) $\left(\sum_{i=1}^{p}\lambda_i^2\right)$
- Dim (0) persistent multiplicity (of zero-eigenvalue)
- Dim (1) persistent multiplicity (of zero-eigenvalue)
- persistent number of non-zero-eigenvalue
- persistent quasi-Wiener index $\left(\sum_{i=1}^{p}\frac{p+1}{\lambda_i}\right)$
- persistent spanning tree number $\left(\mathbf{log}\left(\frac{1}{p+1}\cdot\prod_{i=1}^{p}\lambda_i\right)\right)$

- persistent standard deviation $\left(\sqrt{\frac{1}{p-1}\sum_{i=1}^{p}(\lambda_i - \overline{\lambda})^2}\right)$.

Here, $\lambda_i > 0$ and p is the total number of all non-zero eigenvalues. In particular, persistent multiplicity is the multiplicity of zero eigenvalues and is equivalent to the persistent Betti number or Betti curve. Note that other than the persistent multiplicity, all persistent attributes are calculated from Dim (0) Laplacians.

In our PerSpectSC models, the distance value is considered as the filtration parameter. In filtration parameter discretization process, Hodge Laplacian matrices are generated for each chromosome of each cell. Computationally, the filtration parameter goes from 0.00 to 0.99 with a step of 0.01. A total of 12 chromosomal descriptors, as stated above, are considered. An example can be found in Figure 1 (30 bins range from 47.3 to 50.2 Mb of chromosome 22 of RUES2 CM). For each contact matrix, a total of 100 Laplacian matrices are generated. The normalized area of the kth persistent attribute f^k is defined as follows:

$$A_{rea}(f^k) = \frac{\Delta_x}{N}\sum_{i=1}^{N} f^k(x_i). \tag{9}$$

Computationally, the filtration region is equally divided into N intervals with grid spacing Δ_x. More specifically, we use filtration region $[0, 1]$, $N = 100$ and $\Delta_x = 0.01$.

PerSpectSC-ML for classification of cell types

In our PerSpectSC-based machine learning (PerSpectSC-ML) models, the t-distributed stochastic neighbor embedding (t-SNE) [48] is used for chromosomal descriptors dimensionality reduction. In fact, 'dimensionality reduction +k-means' models are very general approaches for data clustering [49]. Here, we consider t-SNE model. More specifically, we use one of the 12 chromosomal descriptors of each chromosome of each cell as the input (14 [cell types]×100 [filtration values]) of t-SNE. Its hyperparameter settings can be found in Table S1 (see Supplementary Data available online at http://bib.oxfordjournals.org/). Note that, for each pair of data points, their Euclidean distance in the original high-dimensional space will be different from the Euclidean distance in the reduced 2D space (after t-SNE). However, the reduced 2D space is still a metric space. Then, we apply the k-means clustering to the classification dataset (14 [cell types]×2) by setting the number of clusters equal to the number of the real label (here, the number of clusters is 3). In the end, in each cluster, we take the data with the dominant label as the test for all samples and then calculate the k-means clustering accuracy for the whole-cell types.

We denote the training set as $\left\{(\mathbf{X}_i, \mathbf{Y}_i) \mid \mathbf{X}_i \in \mathbb{R}^m, \mathbf{Y}_i \in \mathbf{C}_k\right\}_{i=1}^{n}$ with $\mathbf{C}_k = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_k\}$. Here, n, m and k represent the number of samples, the number of features $\{\mathbf{X}_i\}$ and the number of labels $\{\mathbf{Y}_i\}$, respectively. We set the number of clusters equals to the number of labels k. After applying the k-means clustering, we get k different clusters $\{\mathbf{c}_j\}_{j=1}^{k}$. In each cluster, we define the predictor of the k-means clustering in the cluster $\mathbf{c}_j$ to be

$$\hat{\mathbf{Y}}\,(\mathbf{c}_j) = \max\{F_j(\hat{\mathbf{Y}}_1), \cdots, F_j(\mathbf{Y}_k)\}. \tag{10}$$

Here, $\{\hat{\mathbf{Y}}_i\}$ are predicted labels and $F_j(\hat{\mathbf{Y}}_1), \cdots, F_j(\mathbf{Y}_k)$ are the appearance frequencies of each label in the cluster $\mathbf{c}_j$. Then, the clustering accuracy can be defined as follows:

$$\text{Accuracy} = \frac{1}{n}\sum_i \chi\,(\mathbf{Y}_i, \hat{\mathbf{Y}}_i). \tag{11}$$

Note that the function $\chi(\mathbf{Y}_i, \hat{\mathbf{Y}}_i) = 1$ only when $\mathbf{Y}_i = \hat{\mathbf{Y}}_i$, otherwise it is always zero. It is used to count the number of correctly predicted data points.

Key Points

Our main contributions in this paper are as follows:

- Persistent spectral simplicial complex (PerSpectSC)-based chromosomal structural descriptors are developed.
- PerSpectSC-based chromosomal descriptors can successfully classify cell types and also cellular differentiation stages for all the 24 types of chromosomes simultaneously.
- Our PerSpectSC model provides a powerful chromosomal representation that can be widely used in polymeric data analysis.

Code and data availability

The PerSpectSC-ML models can be found in https://github.com/Weikang-Gong/PerSpectSC-ML. Additional data or code would be available upon reasonable request.

Supplementary Data

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

Authors' contributions statement

K.X. designed research; W.G. performed research; K.X., J.W., M.W., X.S., C.L. and W.G. analyzed data and K.X. and W.G. wrote the paper.

Acknowledgments

We thank the National Supercomputing Centre, Singapore (NSCC) for providing the computing resource (https://www.nscc.sg).

Funding

Nanyang Technological University Startup Grant (grant no. M4081842.110); Singapore Ministry of Education Academic Research fund (grant nos. Tier 1 RG109/19, Tier 2 MOE2018-T2-1-033, Tier 2 MOE-T2EP20120-0013); National Natural Science Foundation of China (grant no. 31971180); China Scholarship Council (grant no. 201906540026).

References

1. Dekker J, Rippe K, Dekker M, *et al.* Capturing chromosome conformation. *Science* 2002;**295**:1306–11.
2. 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**:289–93.
3. Sexton T, Cavalli G. The role of chromosome domains in shaping the functional genome. *Cell* 2015;**160**:1049–59.
4. Zhang B, Wolynes PG. Topology, structures, and energy landscapes of human chromosomes. *Proc Natl Acad Sci U S A* 2015;**112**:6062–7.
5. Dekker J, Mirny L. The 3D genome as moderator of chromosomal communication. *Cell* 2016;**164**:1110–21.
6. Dekker J, Belmont AS, Guttman M, *et al.* The 4D nucleome project. *Nature* 2017;**549**:219–26.
7. Zhang H, Emerson DJ, Gilgenas TG, *et al.* Chromatin structure dynamics during the mitosis-to-G1 phase transition. *Nature* 2019;**576**:158–62.
8. Takei Y, Yun J, Zheng S, *et al.* Integrated spatial genomics reveals global architecture of single nuclei. *Nature* 2021;**590**:344–50.
9. Quinodoz SA, Jachowicz JW, Bhat P, *et al.* RNA promotes the formation of spatial compartments in the nucleus. *Cell* 2021;**184**:5775–5790.e30.
10. Takei Y, Zheng S, Yun J, *et al.* Single-cell nuclear architecture across cell types in the mouse brain. *Science* 2021;**374**:586–94.
11. Strom AR, Biggs RJ, Banigan EJ, *et al.* HP1α is a chromatin crosslinker that controls nuclear and mitotic chromosome mechanics. *Elife* 2021;**10**:e63972.
12. Dixon JR, Jung I, Selvaraj S, *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* 2015;**518**: 331–6.
13. Andrey G, Mundlos S. The three-dimensional genome: regulating gene expression during pluripotency and development. *Development* 2017;**144**:3646–58.
14. Bonev B, Mendelson Cohen N, Szabo Q, *et al.* Multiscale 3D genome rewiring during mouse neural development. *Cell* 2017;**171**:557–572.e24.
15. Cheng RR, Contessoto VG, Lieberman Aiden E, *et al.* Exploring chromosomal structural heterogeneity across multiple cell lines. *Elife* 2020;**9**:e60312.
16. Servant N, Varoquaux N, Lajoie BR, *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* 2015;**16**:1–11.
17. Mead A. Review of the development of multidimensional scaling methods. *J R Stat Soc Ser A: Stat* 1992;**41**:27–11.
18. Hakim O, Misteli T. SnapShot: chromosome conformation capture. *Cell* 2013;**148**:1068–e1.
19. Simonis M, Klous P, Splinter E, *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 2006;**38**:1348–54.
20. Dostie J, Richmond TA, Arnaout RA, *et al.* Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 2006;**16**:1299–309.
21. Oksuz BA, Yang L, Abraham S, *et al.* Systematic evaluation of chromosome conformation capture assays. *Nat Methods* 2021;**18**:1046–55.
22. Eagen KP. Principles of chromosome architecture revealed by Hi-C. *Trends Biochem Sci* 2018;**43**:469–78.
23. Xu Z, Zhang G, Wu C, *et al.* FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. *Bioinformatics* 2016;**32**:2692–5.
24. Forcato M, Nicoletti C, Pal K, *et al.* Comparison of computational methods for Hi-C data analysis. *Nat Methods* 2017;**14**: 679–85.
25. Sauerwald N, Zhang S, Kingsford C, *et al.* Chromosomal dynamics predicted by an elastic network model explains genome-wide accessibility and long-range couplings. *Nucleic Acids Res* 2017;**45**:3663–73.
26. Zhou J, Ma J, Chen Y, *et al.* Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. *Proc Natl Acad Sci U S A* 2019;**116**:14011–8.
27. Zhang S, Chen F, Bahar I. Differences in the intrinsic spatial dynamics of the chromatin contribute to cell differentiation. *Nucleic Acids Res* 2020;**48**:1131–45.
28. Li X, Feng F, Pu X, *et al.* scHiCTools: a computational toolbox for analyzing single-cell Hi-C data. *PLoS Comput Biol* 2021;**17**:e1008978.
29. Kos PI, Galitsyna AA, Ulianov SV, *et al.* Perspectives for the reconstruction of 3D chromatin conformation using single cell Hi-C data. *PLoS Comput Biol* 2021;**17**:e1009546.
30. Liu L, Zhang B, HyeonID C. Extracting multi-way chromatin contacts from Hi-C data. *PLoS Comput Biol* 2021;**12**:e1009669.
31. Lin X, Qi Y, Latham AP, *et al.* Multiscale modeling of genome organization with maximum entropy optimization. *J Chem Phys* 2021;**155**:010901.
32. Lin D, Sanders J, Noble WS. HiCRep.py: fast comparison of Hi-C contact matrices in Python. *Bioinformatics* 2021;**37**:2996–7.
33. Yu M, Abnousi A, Zhang Y, *et al.* SnapHiC: a computational pipeline to identify chromatin loops from single-cell Hi-C data. *Nat Methods* 2021;**18**:1056–9.
34. Wang J, Nakato R. HiC1Dmetrics:framework to extract various one-dimensional features from chromosome structure data. *Brief Bioinform* 2022;**23**:1–16.
35. Gong W, Liu Y, Zhao Y, *et al.* Equally weighted multiscale elastic network model and its comparison with traditional and parameter-free models. *J Chem Inf Model* 2021;**61**:921–37.
36. Sauerwald N, Shen Y, Kingsford C. Topological data analysis reveals principles of chromosome structure in cellular differentiation. In: *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, Vol. **23.** 2019, pp. 1–16.
37. Otter N, Porter MA, Tillmann U, *et al.* A roadmap for the computation of persistent homology. *EPJ Data Science* 2017; **6**:1–38.
38. Meng Z, Xia K. Persistent spectral-based machine learning (PerSpect ML) for protein-ligand binding affinity prediction. *Sci Adv* 2021;**7**:eabc5329.
39. Jacob A, Vedaie M, Roberts DA, *et al.* Derivation of self-renewing lung alveolar epithelial type II cells from human pluripotent stem cells. *Nat Protoc* 2015;**14**:3303–32.
40. wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* 2019;**47**:D520–8.
41. Lee H, Seo PJ. HiCORE: Hi-C analysis for identification of core chromatin looping regions with higher resolution. *Mol Cells* 2021;**44**:883–92.

42. Imakaev M, Fudenberg G, McCord RP, *et al*. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* 2012;**9**:999–1003.
43. Battiston F, Amico E, Barrat A, *et al*. The physics of higher-order interactions in complex systems. *Nat Phys* 2021;**17**: 1093–8.
44. Horak D, Jost J. Spectra of combinatorial Laplace operators on simplicial complexes. *Adv Math* 2013;**17**:1093–8.
45. Schaub MT, Benson AR, Horn P, *et al*. Random walks on simplicial complexes and the normalized Hodge 1-Laplacian. *SIAM Review* 2020;**62**:353–91.
46. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete Comput Geom* 2002;**28**: 511–33.
47. Maria C, Boissonnat JD, Glisse M, *et al*. The gudhi library: simplicial complexes and persistent homology. In: *International Congress on Mathematical Software*, Vol. **8592**. 2014, pp. 167–74.
48. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
49. Hozumi Y, Wang R, Yin C, *et al*. UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. *Comput Biol Med* 2021;**131**:104264.