HiC Contact Map Comaprison Using Graphlet Approach

Behnam Rasoolian *[1] and Debswapna Bhattacharya[1]

[1]Department of Computer Science and Software Engineering, Auburn University

Abstract

Chromosal Conformational Capturing (3C) data, particularly the Hi-C data is widely used for investigating the three-dimensional (3D) chromatin conformation inside the nucleus and for identifying chromatin interactions as well as topologically associating domains (TADs). However, Hi-C based structural analysis for exploring local chromatin conformational preferences has not been performed yet. Here, for the first time, we present graphlet based local structural analysis for normal and malignant human cell types using chromosal conformational capturing (Hi-C) data. We first applied thresholding in Hi-C interaction frequency data for one normal cell line and four leukemic cells lines in order to covert them into unweighted graphs. We then mined the unweighted graphs to extract graphlet -based orbitial frequency distributions? . Subsequently , pairwise graphlet distances as well as pairwise graphlet distribution correlations for each pair of cells were computed and compared using statistical methods namely ANOVA and multi-variate ANOVA. Our results demonstrate that normal-cancer pairs have significantly ($F(1, 1654) = 20.49$, $p{-}value < 0.0001$) higher difference than cancer-cancer pairs. Furthermore, for certain orbits cancer-cancer pairs demonstrated higher correlations ($F(365, 7882) = 391$, $p{-}value < 0.00001$, and $Wilk's \quad \Lambda = 0456$) than normal-cancer pairs in terms of orbit frequency distributions. Source code of GrapHiClet is freely available at `https://github.com/rasoolianbehnam/watson`. Also, the database used for this project can be found at `https://graphiclet.localtunnel.me/secret/`

1 Introduction

Genome of a eukaryotic cell is organized into a complex three-dimensional (3D) structure, which leads to a network of chromosomal interactions [1]. The spatial organization of genome is one of the most important factors in determining its function [2] via interactions between spatially close genes [3]. Such interactions can be modeled using a network approach [3]. However, our knowledge of 3D conformation of genome is limited. Fluorescent In Situ Hybridization (FISH) technology, which can capture 3D conformation of multiple loci but cannot capture global layout of the genome. In order to address this, more advanced techniques of chromosome conformation capturing such as 3C [4], 4C [5] and 5C [6] has been introduced which are based on chromatin fragment fixation, using restriction-enzyme digestion and intra-molecular ligation. One of the most recent versions of chromosome conformation capturing is Hi-C [7, 8], which IF_{ij} captures interactions between chromosomal fragments resulting in a contact map or interaction frequency (IF) matrix, which captures the number of interactions detected in Hi-C dataset between each pair of loci.

Recently, Hi-C data has been mostly used to predict the 3D structure at the chromosome level by satisfying intra-chromosomal interactions or even at the full genome by leveraging both inter- and intra-chromosomal interactions [9, 10, 11, 12, 13, 14, 15, 16, 17]. These efforts, although different in approach, usually translate interaction frequencies in contact maps as simple inverse measures of distance. They then try to find the coordinates of the chromosomal loci that would result in that particular distance matrix using optimization. These methods vary in the optimization models and methodologies used. Also, efforts has been made recently to learn the inverse conversion parameters from contact maps to distance matrices [18]. The reconstructed 3D structures could be used in order to compare normal and malignant cell lines. However,

[*] Tel: +1 334 5212814; Email: bzr0014@auburn.edu

the main obstacle to doing so is that the process of reconstruction is computationally expensive. Almost all reconstruction methods rely on either deterministic or heuristic optimization approaches in order to find the best conformation that matches contact maps. Also, there is no experimental way of evaluating the accuracy of the reconstructed 3D conformations. One experimental way of partially validating correctness of reconstruction process is Fluorescence in situ hybridizaiton (FISH) method. However, this technique is limited to a few loci and cannot be extended to cover the whole structure of the chromosomes. Furthermore, the process of reconstructing 3D chromatic structure is computationally expensive, hindering their applicability at the genome-wide scale.

Over the past decade, graphs have been used extensively in order to model domain-specific phenomena such as computer security, gene regulatory networks and social networks. These graphs are usually in the form of very large networks. It is often necessary to perform statistical analysis between graphs in order to compare their structure. Graphlets and orbits can be used in order to probe large graphs in order to find global and local similarities [19, 20, 21]. This can be done by counting the number of occurrences of a each graphlet and/or orbits for each node in the whole graph and the comparing them [22, 23]

Here, we present local structural analysis between 4 sets of Hi-C data. All four data sets are sequences from the same cell line, with one of them being a normal cell and the other three sequenced from three types of leukemic cells. In order to achieve this, we used graphlets, given their strength in capturing information about local structures of a graph. We first thresholded Hi-C interaction frequency matrices in order to convert in to an unweighted undirected graph adjacency matrix. We the extracted counts of the first 73 orbits for each node, identifying each node with a signature vector of size 73. We then applied graphlet distance metrics proposed in [22] together with statistical methods in order to find difference between cell lines. Our results show that difference of local structure between normal cells and leukemic cells are significantly larger that difference between cancer-cancer cells.

1.0.1 Notations

We denote the set of all contact maps in cell line K with $\mathbb{C}^K$. If no particular cell line is addressed, the subscripts are dropped. Any arbitrary member of $\mathbb{C}$ is denoted by $\mathbf{C}_{ij}$, where i and j ($j \geq i$) represent the two chromosomes involved. In human cells this set contains a total of 276 contact maps, 23 of which are intra-chromosomal and the rest are inter-chromosomal. For ease of representations, intra-chromosomal contact maps are distinguished by a single superscript, so we have $\mathbf{C}_{i,i} = \mathbf{C}_i$.

We denote the number of loci in a chromosome i by N_i. The set of all loci involved in contact map C_{ij} is denoted by $\mathbb{V}_{ij}$. In intra-chromosomal contact maps, $\mathbb{V}_{i,i}$ containts only the loci of that particular chromosome ($|\mathbb{V}_i| = N_i$), while in inter-chromosomal contact maps $\mathbb{V}_{ij}$ contains the loci in the both of chromosomes involved ($\mathbb{V}_{ij} = \mathbb{V}_i \cup \mathbb{V}_j$).

2 Materials and Methods

Chromosomes inside the nuclei are made up of pairs of nucleotides called a base. As a result of the Hi-C process, a database is generated which provides interaction counts found by the method between a number of such bases. The minimum number of bases that can be captured is called the resolution of that database. These counts are then binned so that counts are aggregated for every equal-sized length of the chromosome (e.g. 1 Mb pairs) leading to an N * N interaction frequency matrix. Each cell (i, j) in the matrix is the aggregate count of all interactions found between loci i^{th} and j^{th} length. We re-used Leukemic Hi-C libraries created in [24]. These libraries were sequenced using Illumina HiSeq 2000 for cases of primary human B-acute lymphoblastic leukemia (B-ALL or ALL), the MHH-CALL-4 B-ALL cell line (CALL4), and the follicular lymphoma cell-line (RL) for which high-quality paired-end reados of 39M, 79M and 33M were obtained respectively As in [24], We used normal B-cell line (GM068990) from as benchmark for our comparisons. The three datasets generated in [24] were valid since 98% of the contact generated in [24] were identical to that of [8] and 83% of contacts in [8] were also present in [24]. We created contact maps with bin sizes of 500 kilo-base and normalized them using normalization provided HiC-Pro(`iced` package in python) developed by [25]. Normalization is necessary since Hi-C data usually contains different biases due to GC content,

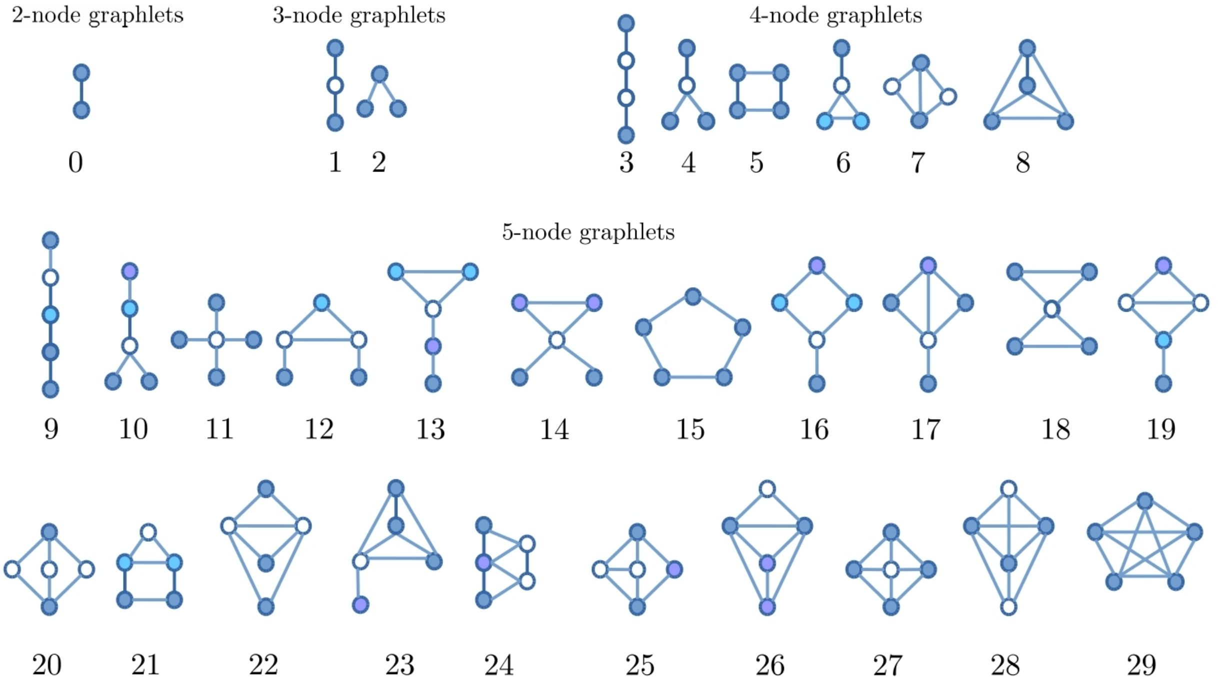


Figure 1

mappability and effective fragment length [26, 27]. The normalization proivdedi in the iced package is based on the Sinkhorn-Knopp algorithm which is a simple, parameter-free and capability to correct unknown biases. The edges in a network usually include indirect dependencies because correlations are transitive; that is, if there is a strong realationship between nodes 1 and 2, and also a strong realationship between nodes 2 and 3, it is highly likely that nodes relationship between nodes 1 and 3 is exaggerated in the network [28]. In order to remove the effect of indirect interactions, We also performed and extra normalization by performing network deconvolution [28], which uses eigenvalue decomposition and infinite series sums in order to reverse the bias posed by indirect relationships.

2.1 Graphlets and Orbits

Recently, graphlet comparison has emerged as a novel method for comparing large networks in order to find local similarities in them. A graph G is a pair (V, E), where V is a set of vertices and $E \subseteq V \times V$ is a set of edges. A connected graph is one where there is a path between every pair of vertices. Given a graph $G(V, E)$ and $S \subseteq V$, then $G'(S, E')$ is a graphlet if and only if it is connected $E' = \{(u, v)|u, v \in V \quad \& \quad (u, v) \in E \rightarrow (u, v) \in E'\}$. There a total of 30 graphlets of size 2, 3, and 5. Figure 1 demonstrates these graphlets. As can be seen there is only one graphlet of size 2 which is equivalent to an edge; that is, the number of graphlet 1 in a graph is the same as the number of edges on the graph. There are also a total of 2 graphlets of size 2, 6 of size 4 and 20 of size 5. The nodes of each graphlet can be partitioned into sets of topographically equivalent nodes called orbits. For example, in Figure 1, we can see that G3 can be partitioned into 2 sets of nodes, the middle ones (white) and the outer nodes (black). For the same set of graphlets, there are 73 orbits that are also illustrated in figure Figure using nodes of difference shades.

2.2 Thresholding contact maps

In order to be able to extract graphlets, HiC contact maps should be modeled as unweighted graphs where the nodes represent the loci and an edge between two nodes represent a *significant* interaction between the loci. This can be achieved by thresholding the contact maps. The result of the thresholding procedure is a binary matrix which also can serve as an adjacency matrix for an unweighted, undirected graph. The graph can then be used for orbit extraction.

When thresholding contact maps, it is necessary to make sure that both global and local features are maintained. We could consider thresholding the contact maps by simply setting values above a fixed value to one and the rest to zero; However, in practice, this method resulted in graphs that capture the local structure of the contact maps poorly. This is because intensities follow an exponential distribution with a mean close to zero with a few very larges values that correspond to interactions along or close to the main diagonal of the contact maps. Thus, picking relatively large numbers would result in ignoring interactions that are far from the main diagonal while picking small values will lead to capturing too many (insignificant) interactions.

To the best of our knowledge, little work has dealt with the task of thresholding HiC contact maps. There has been some statistical approaches developed on similar data in other fields. For example, authors of [29] developed Statistical Network (SPN) analysis where the choice of thresholding value is made by statistical inference. This method, although very robust, works within the framework of design of experiments where the same network can be extracted for different individuals under different treatments. Thus a relatively large set of different contact maps need to be available in order for this method to be applicable towards our end.

Instead, in order to threshold the matrix so that both global and local patterns are captured, we borrowed the concept of *adaptive thresholding* from image processing context. In this method, in order to be set, a pixel should have an intensity larger than the average of non-zero intensities in its *neighborhood.* The neighborhood is defined by a sliding kernel that passes through the contact map with the pixel at its middle at each step. Figure 2 demonstrates result of this thresholding approach for intra-chromosomal contact maps of chromosome 1. Refer to supplementary material for all 23 interchromosomal thresholding results.

2.3 Orbit Extraction

Once the thresholded contact maps are obtained, graphlets and orbits can be extracted. We used the `orca` package in `R` programming language to extract the graphlets. As a result of graphlet extraction, For each loci in each contact map, a *signature vector* of size 73 is created. Thus for each cell line, we would have 276 *signature matrices* of size $|V^{ij}| \times 73$, where V^{ij} is the number of loci involved in contact map between chromosomes i and j. Figure 3 illustrates the process and results of signature matrix extraction schematically.

For a particular $\mathbf{C}_{ij}$, we denote $\mathbf{S}_{ij}$ as its *signature matrix.* Each cell S_{ijlo} in $\mathbf{S}_{ij}$ captures how many times loci l in $\mathbf{C}_{ij}$ occured as part of orbit o.

We consider two measures of *difference* when comparing contact map graphlets across cell lines. The first measure is *signature distance vectors* between each contact map of two cell lines. For a pair of cells A and B, let $\mathbf{S}^A_{ij}$ and $\mathbf{S}^B_{ij}$ be their signature matrices. The *signature distance* of contact map $\mathbf{C}_{i,j}$ between A and B is denoted by $\mathbf{d}^{A,B}_{ij}$. $\mathbf{d}^{A,B}_{ij}$ is a vector of size $|V_{i,j}|$ and its elements $d^{A,B}_{i,j,l}$ are calculated using the following formula from [22]:

$$d^{A,B}_{i,j,l} = \frac{1}{73}\sqrt{\sum_{o=0}^{72} t^2_{lo}} \tag{1}$$

where elements of $t_{i,j,l,o}$ is the distance between each loci (row) l in $\mathbf{S}^A$ and the the same loci in $\mathbf{S}^B$ for orbit o as is calculated as below:

$$t_{lo} = w_o \times \frac{log(S^A_{ijlo}+1) - log(S^B_{ijlo}+1)}{log(max(S^A_{ijlo}, S^B_{ijlo}) + 2)} \tag{2}$$

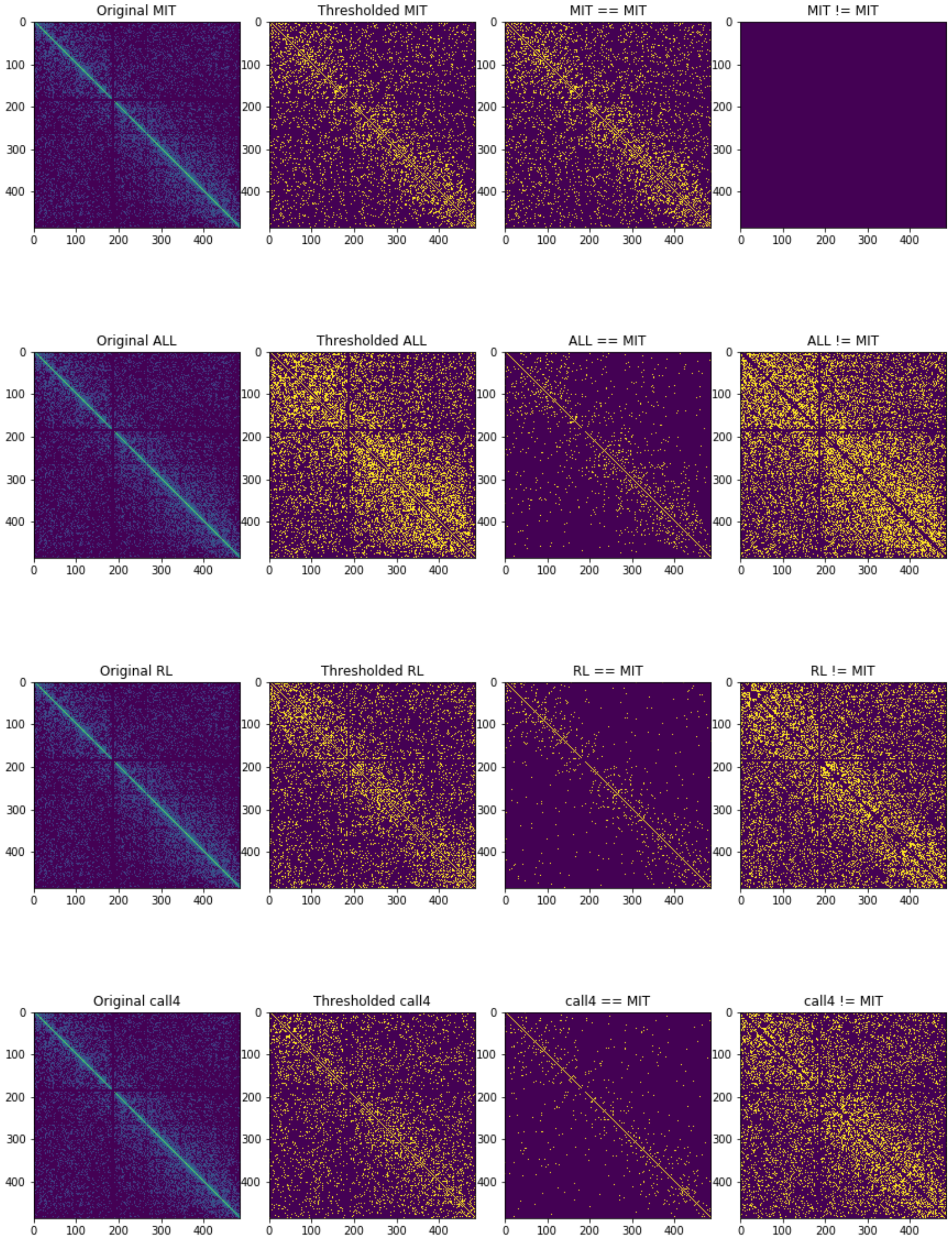


Figure 2: Result of thresholding interchromosomal contact map of chromosome 1 using a kernels of size 5×5 for all cell lines. The first row shows the thresholded maps. Second and third rows demonstrate pair-wise similarities and differences between contact maps respectively.

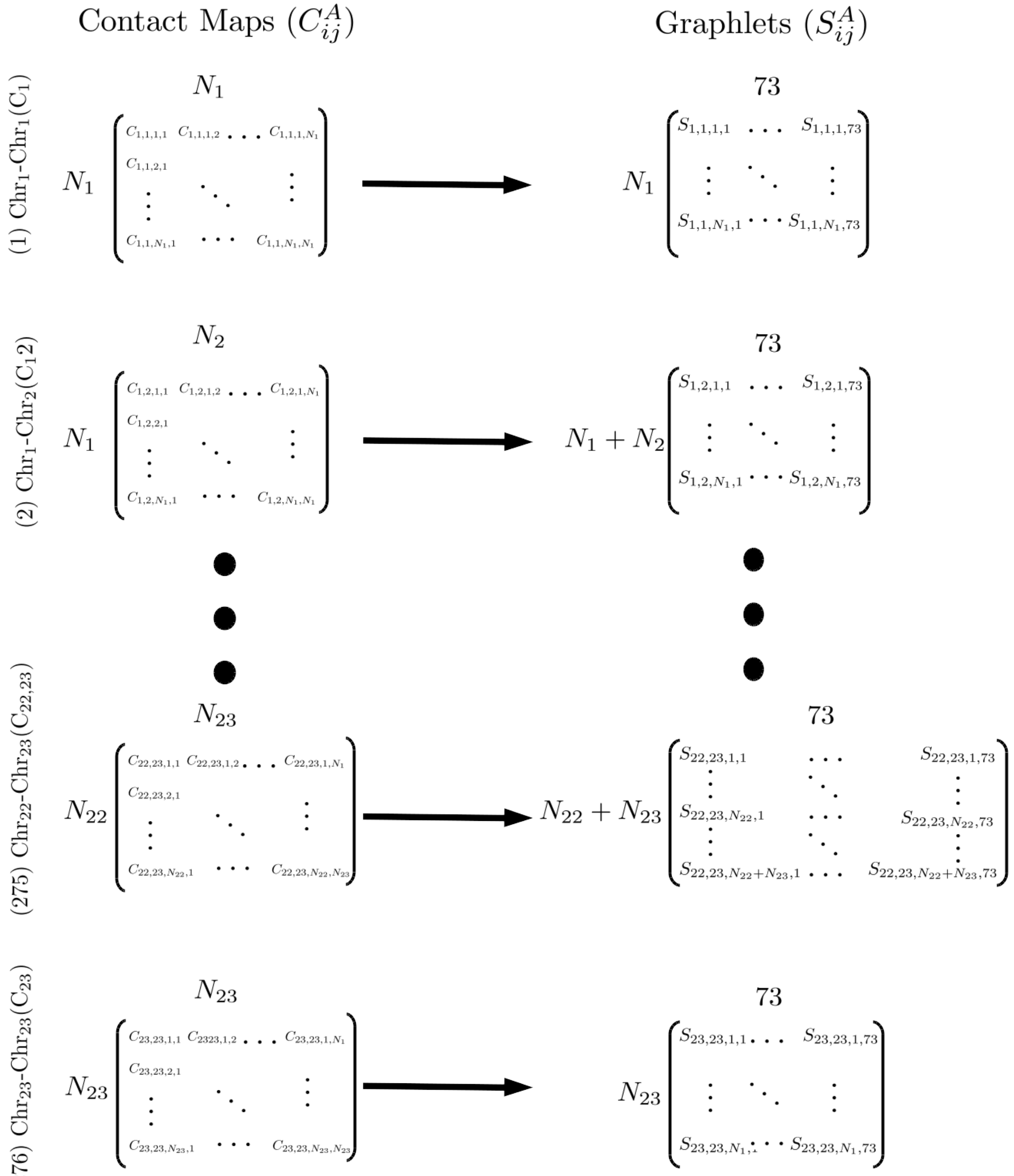

Contact Maps (C^A_{ij})
Graphlets (S^A_{ij})
(1) Chr$_1$-Chr$_1$(C$_1$)
N_1
N_1
73
N_1
(2) Chr$_1$-Chr$_2$(C$_1$2)
N_2
N_1
73
$N_1 + N_2$
(275) Chr$_{22}$-Chr$_{23}$(C$_{22,23}$)
N_{23}
N_{22}
73
$N_{22} + N_{23}$
(276) Chr$_{23}$-Chr$_{23}$(C$_{23}$)
N_{23}
N_{23}
73
N_{23}

Figure 3: Graphlet extraction for the four cell lines. For each loci in each contact map between chromosomes i and j, the signature vectors of length 73 are extracted, resulting in a *signature matrix* of size $|V^{ij}| \times 73$, where V^{ij} is the number of loci involved.

$$\text{Cell A Graphlets}(\mathrm{S}^A_{ij}) \qquad \text{Cell B Graphlets}(\mathrm{S}^B_{ij}) \qquad\qquad d^{A,B}_{ij}$$

$$\text{(1) Chr}_1\text{-Chr}_1(\mathrm{C}_1)\begin{pmatrix} S^A_{1,1,1,1} & \cdots & S^A_{i,j,1,73} \\ \vdots & & \vdots \\ S^A_{1,1,i,1} & \ddots & S^A_{1,1,i,73} \\ \vdots & & \vdots \\ S^A_{1,1,N_1,1} & \cdots & S^A_{1,1,N_1,73} \end{pmatrix}\begin{pmatrix} S^B_{1,1,1,1} & \cdots & S^B_{i,j,1,73} \\ \vdots & & \vdots \\ S^B_{1,1,i,1} & \ddots & S^B_{1,1,i,73} \\ \vdots & & \vdots \\ S^B_{1,1,N_1,1} & \cdots & S^B_{1,1,N_1,73} \end{pmatrix} \longrightarrow \begin{pmatrix} d^{A,B}_{1,1,0} \\ d^{A,B}_{1,1,1} \\ \cdots \\ d^{A,B}_{1,1,72} \end{pmatrix}_{N_1 \times 1}$$

$$\text{(2) Chr}_1\text{-Chr}_2(\mathrm{C}_12)\begin{pmatrix} S^A_{1,2,1,1} & \cdots & S^A_{1,2,1,73} \\ \vdots & & \vdots \\ S^A_{1,2,N_1,1} & \ddots & S^A_{1,2,N_1,73} \\ \vdots & & \vdots \\ S^A_{1,2,N_1+N_2,1} & \cdots & S^A_{1,2,N_1+N_2,73} \end{pmatrix}\begin{pmatrix} S^B_{1,2,1,1} & \cdots & S^B_{1,2,1,73} \\ \vdots & & \vdots \\ S^B_{1,2,N_1,1} & \ddots & S^B_{1,2,N_1,73} \\ \vdots & & \vdots \\ S^B_{1,2,N_1+N_2,1} & \cdots & S^B_{1,2,N_1+N_2,73} \end{pmatrix} \longrightarrow \begin{pmatrix} d^{A,B}_{1,2,0} \\ d^{A,B}_{1,2,1} \\ \cdots \\ d^{A,B}_{1,2,72} \end{pmatrix}_{(N_1 + N_2) \times 1}$$

$$\vdots \qquad\qquad\qquad \vdots$$

$$\text{(275) Chr}_{22}\text{-Chr}_{23}(\mathrm{C}_{22,23})\begin{pmatrix} S^A_{22,23,1,1} & \cdots & S^A_{22,23,1,73} \\ \vdots & & \vdots \\ S^A_{22,23,N_{22},1} & \ddots & S^A_{22,23,N_{22}73} \\ \vdots & & \vdots \\ S^A_{22,23,N_{22}+N_{23},1} & \cdots & S^A_{22,23,N_{22}+N_{23},73} \end{pmatrix}\begin{pmatrix} S^B_{22,23,1,1} & \cdots & S^B_{22,23,1,73} \\ \vdots & & \vdots \\ S^B_{22,23,N_{22},1} & \ddots & S^B_{22,23,N_{22}73} \\ \vdots & & \vdots \\ S^B_{22,23,N_{22}+N_{23},1} & \cdots & S^B_{22,23,N_{22}+N_{23},73} \end{pmatrix} \rightarrow \begin{pmatrix} d^{A,B}_{22,23,0} \\ d^{A,B}_{22,23,1} \\ \vdots \\ d^{A,B}_{22,23,72} \end{pmatrix}_{(N_{22} + N_{23}) \times 1}$$

$$\text{(276) Chr}_{23}\text{-Chr}_{23}(\mathrm{C}_{23})\begin{pmatrix} S^A_{23,23,1,1} & \cdots & S^A_{23,23,1,73} \\ \vdots & & \vdots \\ S^A_{23,23,i,1} & \ddots & S^A_{23,23,i,73} \\ \vdots & & \vdots \\ S^A_{23,23,N_{23},1} & \cdots & S^A_{23,23,N_{23},73} \end{pmatrix}\begin{pmatrix} S^B_{23,23,1,1} & \cdots & S^B_{23,23,1,73} \\ \vdots & & \vdots \\ S^B_{23,23,i,1} & \ddots & S^B_{23,23,i,73} \\ \vdots & & \vdots \\ S^B_{23,23,N_{23},1} & \cdots & S^B_{23,23,N_{23},73} \end{pmatrix} \longrightarrow \begin{pmatrix} d^{A,B}_{23,23,0} \\ d^{A,B}_{23,23,1} \\ \vdots \\ d^{A,B}_{23,23,72} \end{pmatrix}_{N_{23} \times 1}$$

Figure 4: Calculating pair-wise loci distances. For each loci (row) in each contact map in MIT cell line, its distance is calculated based on equation 1 with the corresponding loci in leukemic cells. The result of this process is a *signature distance vector* of size $|V^{ij}| = N_i + N_j$ for each contact map.

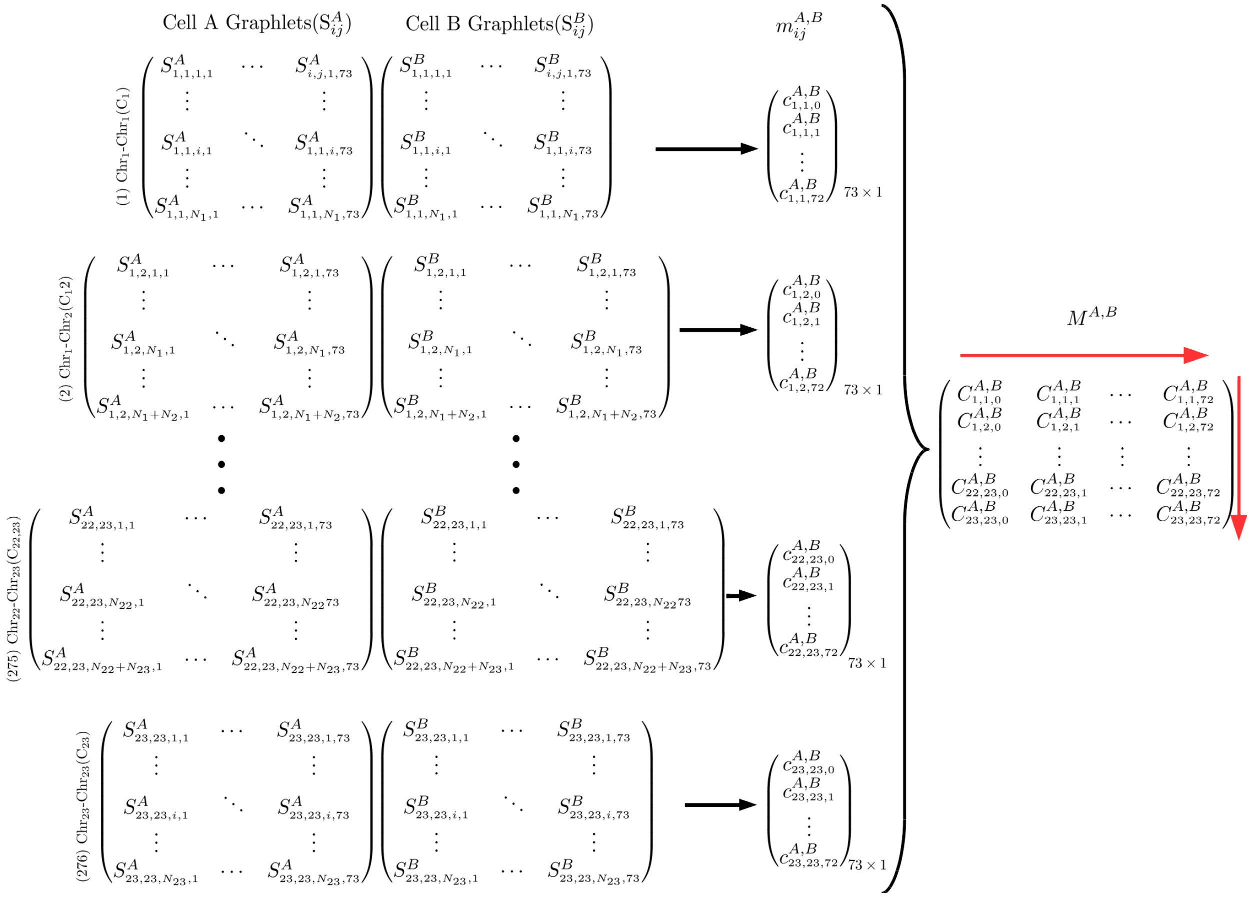

Cell A Graphlets(S_{ij}^{A})
Cell B Graphlets(S_{ij}^{B})
$m_{ij}^{A,B}$
$M^{A,B}$
(1) Chr_1-Chr_1(C_1)
(2) Chr_1-Chr_2(C_12)
(275) Chr_{22}-Chr_{23}($\mathrm{C}_{22,23}$)
(276) Chr_{23}-Chr_{23}(C_{23})
73×1
73×1
73×1
73×1

Figure 5: Calulating pair-wise orbit correlations. For each orbit (column) in each contact map in MIT cell line, its correlation with the same orbit in the same contact map in leukemic cells is calculated. The result of this process is a *signature correlation* vector of size 73 which captures how similar frequencies of two orbits are.

This process is illustrated in Figure 4. Using this distance measure, we can quantify how two loci are close to each other in terms of local neighborhood between the two contact maps.

The second measure of comparison that we use captures how similar two orbits are in terms of their count frequencies across loci between two contact maps. Each column in S_{ij} can provide information regarding the *frequency distribution* of orbits throughout the contact map C_{ij}. We can find how similar these distributions are to each other using correlation measures. These correlations are denoted by $\mathbf{m}_{i,j}^{A,B}$ and can be calculate using any plausible correlation measure. In this study, for each contact map, we calculated similarity between orbit distributions using Pearson's r correlation, which is computationally efficient. However, pearson's r might not be able to capture non-functional relationships between distributions. As a result, we also used Maximal Information Coefficient (MIC) [30] in order to compare correlations. MIC calculates mutual information (MI) between two distributions, but utilizes dynamic programming in order adjust bin sizes and numbers in order to achieve highest MI. MIC values between two variables fall between 0 and 1, with 0 meaning the two variables are completely independent and 1 meaning one is dependant on the other. We used both Pearson's r and MIC in order to compare orbit frequencies. Although results from both approaches were more or less consistent, MIC showed higher robustness than Pearson's r method.

If MIC is used as correlation measure, each element of **c** is calculated as below:

$$m_{ijo}^{A,B} = MIC(\mathbf{S}_{ij.o}^{A}, \mathbf{S}_{ij.o}^{B}) \tag{3}$$

Alternatively, if we use Pearson criterion we would have:

$$m_{ijo}^{A,B} = Pearson(\mathbf{S}_{ij.o}^{A}, \mathbf{S}_{ij.o}^{B}) \tag{4}$$

3 results and discussions

3.1 Contact Map Orbit Vector Distance

By comparing signature distance vectors, one can find how contact maps differ from each other in terms of local structure. Contact maps can serve as measures of spatial proximity between loci. Graphlets capture certain patterns of interaction, or in other words, spatial neighborhood for each loci. Thus, if signature vectors of two loci are close, it can be inferred that they have similar spatial neighborhood.

We can compare pairs of contact maps in terms of their closeness to each other. As an example, in figure 7, all pairs of cells are compared to each other in terms of their distance for contact maps involving chromosome 14. We can se that for the first 13 inter-chromosomal contact maps, ALL and RL cell lines are closer to each other than to other cell lines.

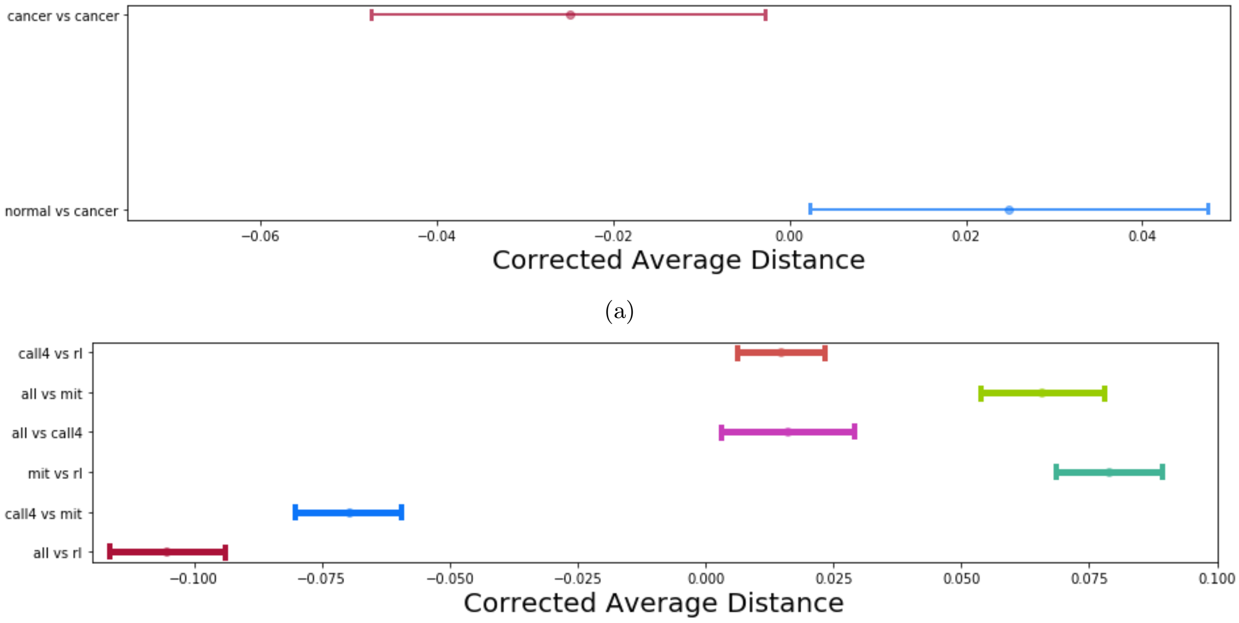
We performed one-way ANOVA statistical test to see if there are significant difference between cancer-normal and cancer-cancer pairs. We found that the difference statistically significant difference. ($F(1, 1654) = 20.49, p < 0.0001$). As illustrated in figure 6a, we can see that normal-cancer pairs have higher distance from each other than cancer-cancer pairs.

We then continued to investigate each pair separately to see if there is any significant difference between them. Again, our statistical tests (ANOVA) showed significant difference between pairs of cells. The results are shown in figure 6b. We can see that ALL-RL pair are closest to each other while ALL-MIT and MIT-RL are most distant. We found statistcally significant difference for difference between individual pairs except for ALL-CALL4 and CALL4-RL as well as MIT-RL and ALL-MIT. The results of thse tests can be found in supplementary material.

The results in figure 6b is also in keeping with what we see in figure 7. For example, as mentioned earlier, our results show that on average, ALL and RL cell lines are closer to each other than to other cell lines, which is also the case in figure 7 for majority of contact maps.

3.2 MIC Comparison

In addition to comparing cells in terms of their orbit distances, we can compare them by measuring how often certain graphlets occur in their contact maps. By doing so, we measure the frequency distribution of the



(a)

(b)

Figure 6: **Average Pair-wise graphlet signature difference over all 276 contact maps:** Each point on a graph is the result of averaging all the distances across all loci of all contact maps for each pair of cells. A 0.1 standard deviation error bar is also plotted for each point. ($\bar{\mathbf{d}}^{A,B}$).

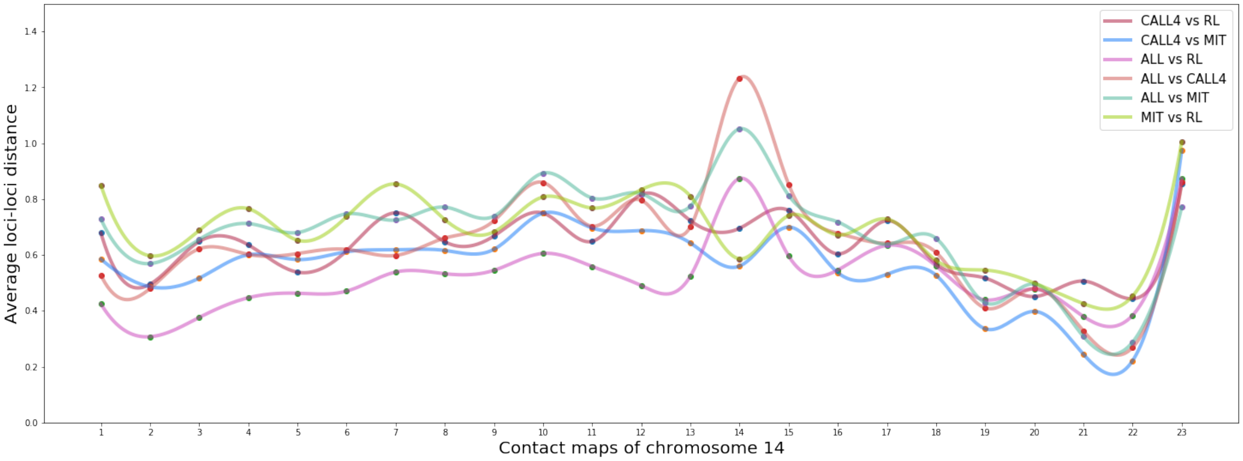
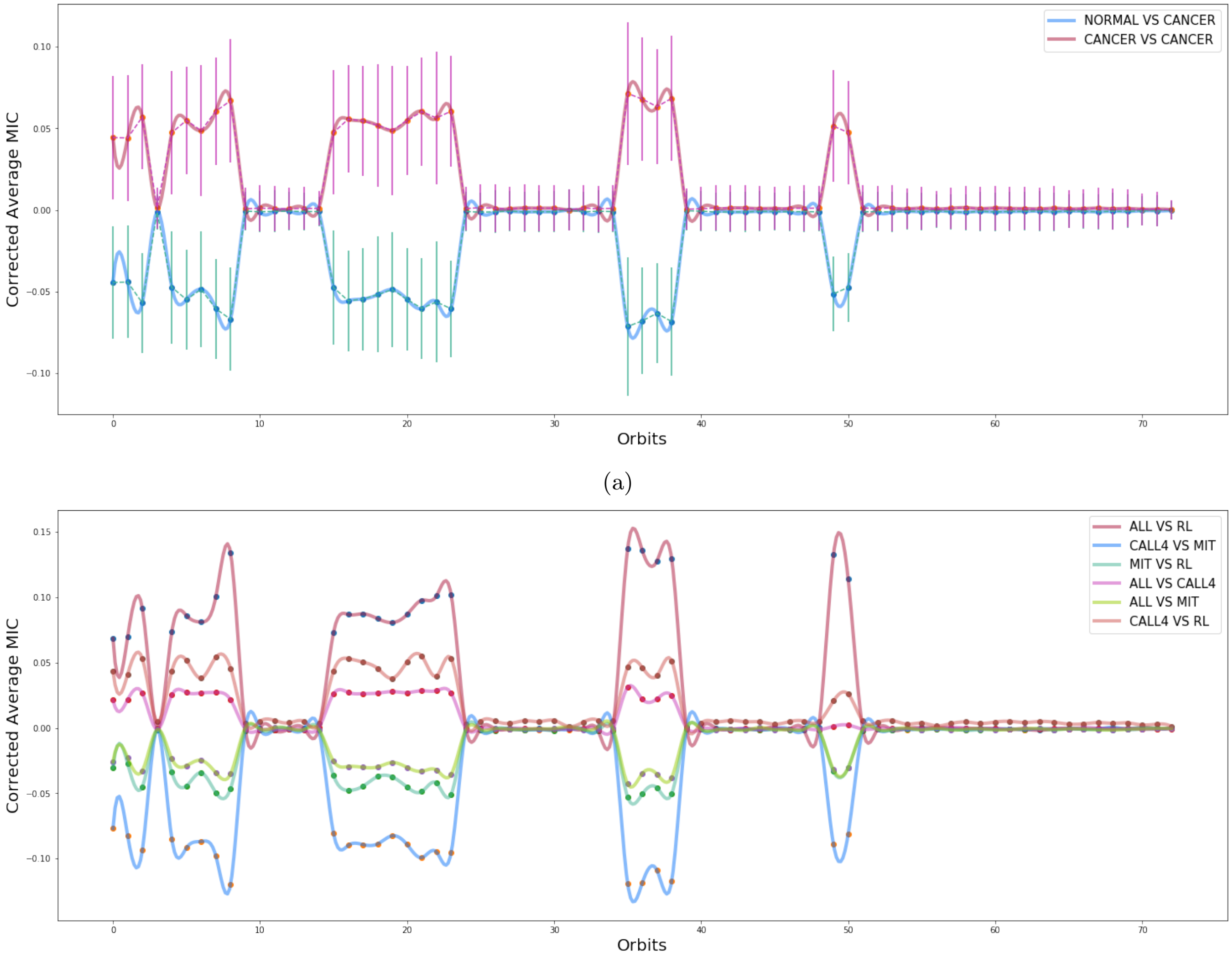


Figure 7: Pair-wise graphlet signature distances for all contact maps of chromosome 14. Warm colors as used for cancer-cancer pairs and cold colors are used for normal-cancer pairs. As can be seen cancer-cancer pairs tend to be close to each other than to the normal cell. This is specially true for the first 13 inter-chromosomal contact maps.



(a)

(b)

Figure 8: **Pair-wise average contact map orbit correlations for all contact maps:** $(\bar{\mathbf{m}}^{A,B}_{i,j} \quad \forall i,j \in \{1...23\} \quad \& \quad j \geq i$: average along the red *vertical* arrow in figure 5). These values are calculated by averaging over pairwise correlations of orbits of $\mathbb{Q}$ in a contact map. **Pair-wise average orbit correlations:** In figure 8b, each point in the graph is the result of averaging pair-wise orbit correlations over all contact maps $(\frac{1}{276}\sum_{i=0}^{23}\sum_{j=i}^{23} m^{A,B}_{i,j,o} \quad \forall o \in \{0,1,...,72\}$: average along the red *horizontal* arrows in figure 5). Counts for certain orbits are always zero in inter-chromosomal maps, leading to average value close to zero in Figure 8b. In this figure, cancer cells data points are depicted in warm colors while normal cells are depicted in cold colors in increased contrast. As can be seen orbit distributions are more similar to each other for cancer cells.

spatial structures represented by orbits in each contact map. In order to see how closely such structures are distributed, we can compare contact maps by calculating the correlation between their orbit distributions. A higher correlation for certain orbits would mean higher similarity in terms of that particular spatial structure between the loci involved.

Before going on with the results, it is worth mentioning that interchromosomal thresholded contact maps represent a bipartide graph with the loci from each chromosome on one side. Due to this bipartide nature of the graphs in inter-chromosomal maps, count of certain orbits is always 0, resulting in a correlation values of 0 for them as well. You can see the bias in figure 8 where average correlations of orbits $\mathbb{Q} = \{3, 9, 10\text{-}14, 20\text{-}34, 39\text{-}48, 51\text{-}72\}$ are close to zero. In fact all correlations corresponding to these orbits are 0 except for the ones between the same chromosomes.

We caclulated pair-wise MIC values for each orbit in each of the 276 contact maps from MIT, ALL, RL, and CALL4 data separately. We found statistically significant difference between cancer-cancer and normal-cancer correlations. ($F(73, 1582) = 6.29$, $p < 0.00001$, Wilk's $\Lambda = 0.775$) The difference is also illustrated in figure 8a, which plots corrected mean difference of MIC values for cance-cancer and normal-cancer correlations. We performed statistical test to see if there is a significant difference correlation between individual pairs of cells. Figure 8b demonstrate such difference. Average correlation over all contac maps for normal-cancer pairs are smaller than cancer-cancer pairs. This is corroborated by the results of statical test which verify that such difference is in fact signifcant.($F(365, 7882) = 3.91$, $p < 0.00001$, Wilk's $\Lambda = 0456$) Figures 8a and 8b both show more details about this difference in correlation. As can be observed, for obits in $\mathbb{Q}$, normal correlations are smaller than cancer correlations, while for the rest of the orbits, there is no difference.

Figure 8 demonstrates that certain orbits of Leukemic cells have higher correlation to each other than to the normal MIT cell. In fact our statistical analysis shows that *for orbits NOT in $\mathbb{Q}$, intra-leukemic orbit correlations are significantly higher than leukemic-normal orbit correlations.* This implies there are significant differences between normal and leukemic cells in terms of their local structure.

4 Resources

Hi-C Datasets:

1. Code base for this article
2. Datasets including cancerous cells
3. Original Datasets

References

[1] Job Dekker. Gene regulation in the third dimension. *Science*, 319(5871):1793–1794, 2008.

[2] Peter Fraser and Wendy Bickmore. Nuclear organization of the genome and the potential for gene regulation. *Nature*, 447(7143):413, 2007.

[3] Michael H Kagey, Jamie J Newman, Steve Bilodeau, Ye Zhan, David A Orlando, Nynke L van Berkum, Christopher C Ebmeier, Jesse Goossens, Peter B Rahl, Stuart S Levine, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430, 2010.

[4] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *science*, 295(5558):1306–1311, 2002.

[5] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo De Wit, Bas Van Steensel, and Wouter De Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature genetics*, 38(11):1348, 2006.

[6] Josée Dostie and Job Dekker. Mapping networks of physical interactions between genomic elements using 5c technology. *Nature protocols*, 2(4):988, 2007.

[7] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[8] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[9] William Noble, Zhi-jun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, and C Anthony Blau. A three-dimensional model of the yeast genome. In *International Conference on Research in Computational Molecular Biology*, pages 320–320. Springer, 2011.

[10] Mathieu Rousseau, James Fraser, Maria A Ferraiuolo, Josée Dostie, and Mathieu Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling. *BMC bioinformatics*, 12(1):414, 2011.

[11] Ming Hu, Ke Deng, Zhaohui Qin, Jesse Dixon, Siddarth Selvaraj, Jennifer Fang, Bing Ren, and Jun S Liu. Bayesian inference of spatial organizations of chromosomes. *PLoS computational biology*, 9(1):e1002893, 2013.

[12] Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. A statistical approach for inferring the 3d structure of the genome. *Bioinformatics*, 30(12):i26–i33, 2014.

[13] Tuan Trieu and Jianlin Cheng. Large-scale reconstruction of 3d structures of human chromosomes from chromosomal contact data. *Nucleic acids research*, 42(7):e52–e52, 2014.

[14] ZhiZhuo Zhang, Guoliang Li, Kim-Chuan Toh, and Wing-Kin Sung. 3d chromosome modeling with semi-definite programming and hi-c data. *Journal of computational biology*, 20(11):831–846, 2013.

[15] Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 3d genome reconstruction from chromosomal contacts. *Nature methods*, 11(11):1141, 2014.

[16] Davide Baù, Amartya Sanyal, Bryan R Lajoie, Emidio Capriotti, Meg Byron, Jeanne B Lawrence, Job Dekker, and Marc A Marti-Renom. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature Structural and Molecular Biology*, 18(1):107, 2011.

[17] Badri Adhikari, Tuan Trieu, and Jianlin Cheng. Chromosome3d: reconstructing three-dimensional chromosomal structures from hi-c interaction frequency data using distance geometry simulated annealing. *BMC genomics*, 17(1):886, 2016.

[18] Oluwatosin Oluwadare, Yuxiang Zhang, and Jianlin Cheng. A maximum likelihood algorithm for reconstructing 3d structures of human chromosomes from chromosomal contact data. *BMC genomics*, 19(1):161, 2018.

[19] Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In *Artificial Intelligence and Statistics*, pages 488–495, 2009.

[20] Christian Borgs, Jennifer Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi. Counting graph homomorphisms. In *Topics in discrete mathematics*, pages 315–371. Springer, 2006.

[21] John Adrian Bondy and Robert L Hemminger. Graph reconstruction—a survey. *Journal of Graph Theory*, 1(3):227–268, 1977.

[22] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[23] Natasa Pržulj, Derek G Corneil, and Igor Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

[24] Zheng Wang, Renzhi Cao, Kristen Taylor, Aaron Briley, Charles Caldwell, and Jianlin Cheng. The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types. *PloS one*, 8(3):e58793, 2013.

[25] Nicolas Servant, Nelle Varoquaux, Bryan R Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. Hic-pro: an optimized and flexible pipeline for hi-c data processing. *Genome biology*, 16(1):259, 2015.

[26] Eitan Yaffe and Amos Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43(11):1059, 2011.

[27] Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S Liu. Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012.

[28] Soheil Feizi, Daniel Marbach, Muriel Médard, and Manolis Kellis. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology*, 31(8):726–733, 2013.

[29] Cedric E Ginestet and Andrew Simmons. Statistical parametric network analysis of functional connectivity dynamics during a working memory task. *Neuroimage*, 55(2):688–704, 2011.

[30] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.