Comparison of Hi-C contact maps reveals difference of 3D conformation between normal and leukemic cells

Behnam Rasoolian, Debswapna Bhattacharya [[1]](#footnote-2)

# ABSTRACT

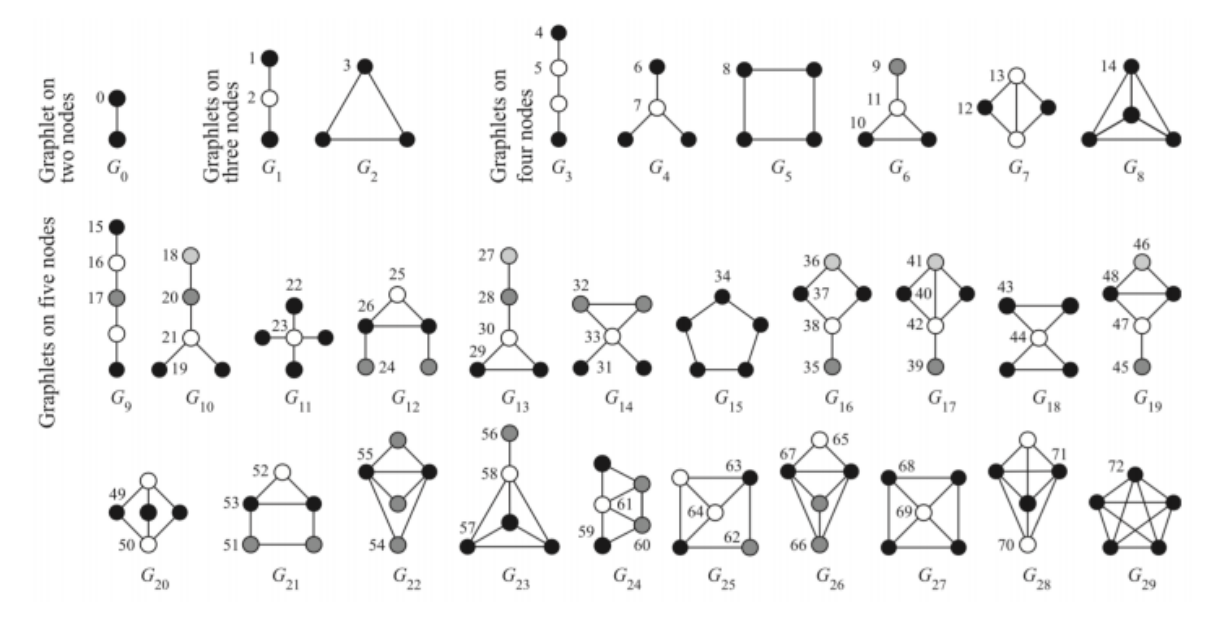
1. In this study, we investigated dissimilarities between normal cells and leukemic cells in terms of their three dimensional conformation. We first thresholed HiC data corresponding to one normal cell line and four leukemic cells lines. We then used the thresholded data to extract the first 72 orbits for all loci in each contact map. We then measured pairwise graphlet distances as well as pairwise graphlet distribution correlations for each pair of cells and compared them using statistical methods. Our results show that normal-cancer pairs have significantly higher dinstances from each other than cancer-cancer pairs. Also, we concluded that for certain orbits, cancer-cancer pairs demonstrate higher correlations than normal-cancer pairs in terms of orbit frequency distributions.**Introduction**
2. Studies has shown that genome of a eukaryotic cell is organized into a complex three-dimensional structure which leads to a network of chromosomal interactions [dekker2008gene]. Researches suggest that a genome's spatial organization is the most important factor in determining its function [fraser2007nuclear] by providing for interactions between spatially close genes [kagey2010mediator]. Such interactions can be modeled using a network approach [kagey2010mediator]. Ideally, it is desirable to compare 3D structures of cell in order to make such comparisons. However, the main challenge in this regard is that 3D structure of a cell is not readily available. Our knowledge of three-dimensional conformation of genomes started with the interphase fluorescent in situ hybridization (FISH) technology which can visualize 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[dekker2002capturing], 4C [simonis2006nuclear] and 5C [dostie2007mapping] has been introduced which are based on chromatin fragment fixation, using restriction-enzyme digestion and intra-molecular ligation. These efforts culminated in Hi-C method (rao20143d, lieberman2009comprehensive) which can capture interactions at a resolution of 1 megabase. These methods capture interactions between chromosomal fragments resulting in a *contact map* or *interaction frequency (IF)* matrix. A cell  in an interaction frequency matrix captures the number of interaction detected in HiC dataset between locus  and locus  in the genome. An interaction frequency matrices capture both inter- and intra-chromosomal interaction between loci.

This research is an effort to compare the three-dimensional structure of cell genomes. To the best of our knowledge little research has dealt with quantitative comparison of spatial conformation of chromosomes. In order to find dissimilarities in the 3D structure of chromosomes, we used HiC dataset and have found statistically significant differences between interaction matrices of normal cells and cancerous ones. 

Recently, Hi-C data has been mostly used in efforts to predict the 3D conformation of genome. These efforts, although different in approach, usually translate interaction frequencies in contact maps as an inverse measure of distance. For example, developed a method, Chromosome3D which uses distance geometry simulated annealing (DGSA) in order to reconstruct genome structure. These reconstructed 3D structures could be used in order to compare normal and malignant cell lines. However, 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 real way of verifying whether the reconstruction results are accurate. The only method of verifying such results is Fluorescence in situ hybridizaiton (FISH) which can capture the actual 3D configuration of chromosomes. However, this method can only be used locally and cannot map the whole structure of the chromosomes.

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 ⊆ V×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 ⊆ V, then G’(S, E’) is a graphlet if and only if it is connected E’ = {(u, v) | u, v ∈ V and (u, v) ∈ E → (u, v) ∈ E’}. There a total of 30 graphlets of size 2, 3, and 5. Figure 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, we can see that G3 can be partitioned into 2 sets of nodes, the middle ones (while) 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.

raphlets and orbits can be used in order to probe large graphs in order to find global and local similarities [shervashidze2009efficient, borgs2006counting, bondy1977graph, borgs2006counting]. 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 [prvzulj2007biological, prvzulj2004modeling].

  
Figure 1





### The purpose of this paper is to investigated local structural differences between 4 set 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. Our work is novel since it is one the first efforts at characterizing and comparing structures of the genomes at local level. Here, we apply the graphlet distance approach presented in [ prvzulj2007biological]. 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 apply graphlet distance metrics proposed in [prvzulj2007biological] 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.

# Materials and Methods

### Notations

### In this paper, matrices and vectors are represented using bold capital and bold small letters respectively. Matrix rows and columns are represented by a *dot* notation. For example, the th row of matrix is denoted by and its th column is represented by .

We denote the set of all contact maps in cell line  with . If no particular cell line is addressed, the subscripts are dropped. Any arbitrary member of  is denoted by , where  and  () 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 .

We denote the number of loci in a chromosome  by . The set of all loci involved in contact map  is denoted by . In intra-chromosomal contact maps,  contains only the loci of that particular chromosome (), while in inter-chromosomal contact maps  contains the loci in the both of chromosomes involved ().

# **Hi-C contact maps:**

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 ith and jthlength.

We re-used Leukemic Hi-C libraries created in [ wang2013properties]. 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 [wang2013properties], We used normal B-cell line (GM068990) from as benchmark for our comparisons. The three datasets generated in [wang2013properties ] were valid since 98% of the contact generated in [wang2013properties ] were identical to that of [lieberman2009comprehensive] and 83% of contacts in [lieberman2009comprehensive] were also present in [wang2013properties]. 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 [servant2015hic,]. Normalization is necessary since Hi-C data usually contains different biases due to GC content, mappability and effective fragment length [yaffe2011probabilistic, hu2012hicnorm,]. 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 [feizi2013network]. In order to remove the effect of indirect interactions, We also performed and extra normalization by performing network deconvolution [feizi2013network], which uses eigenvalue decomposition and infinite series sums in order to reverse the bias posed by indirect relationships.

## 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 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 [local\_thresholded\_chr1\_chr1] demonstrates result of this thresholding approach for intra-chromosomal contact maps of chromosome 1. Refer to supplementary material for all 23 interchromosomal thresholding results.

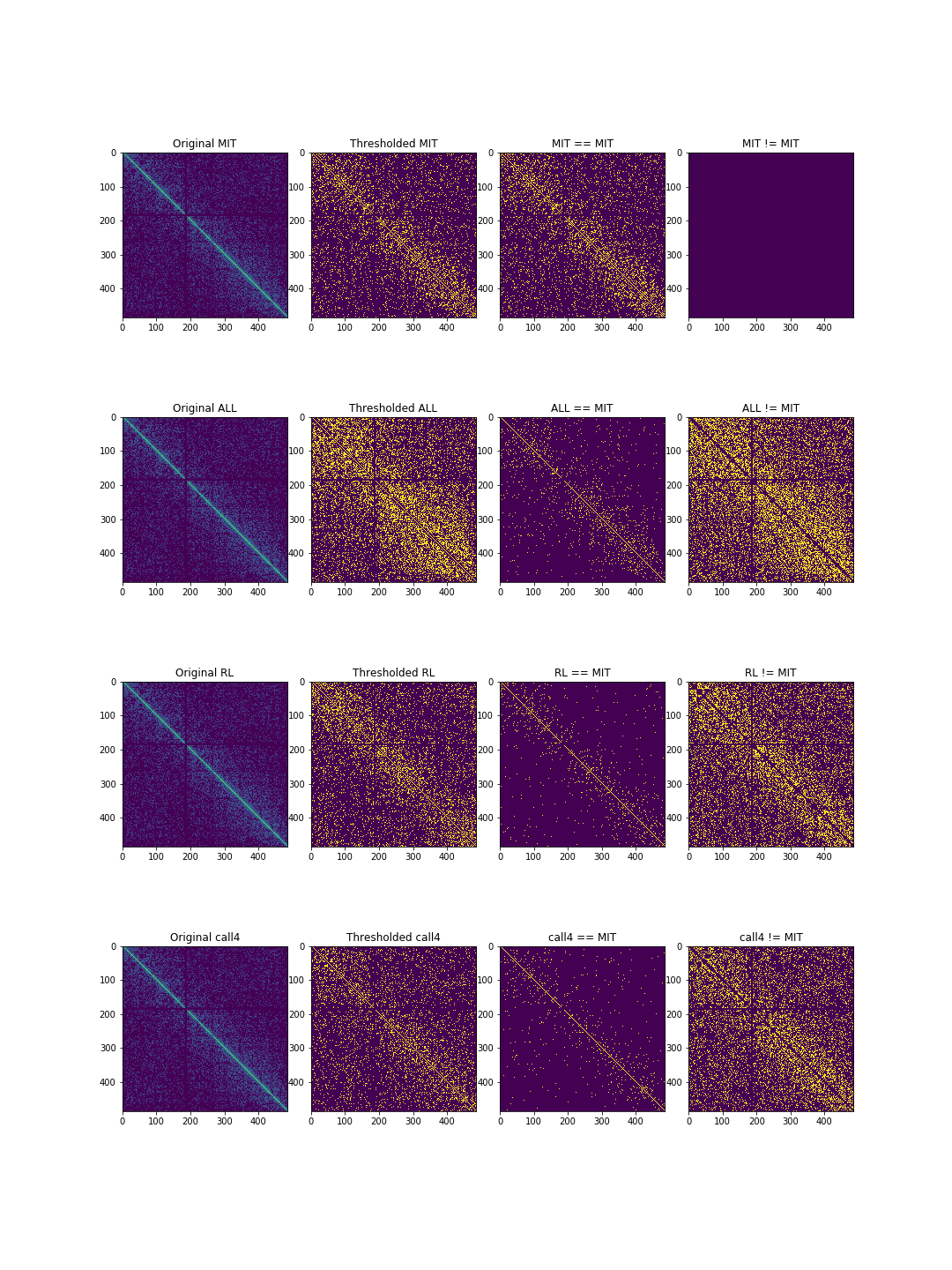


Figure 2: Result of thresholding interchromosomal contact map of chromosome 1 using a kernels of size 5x5 for all cell lines. The first row shows the thresholded maps. Second and third rows demonstrate pair-wise similarities and differences between contact maps respectively.

## 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, where  is the number of loci involved in contact map between chromosomes and . Figure 2 illustrates the process and results of signature matrix extraction schematically.

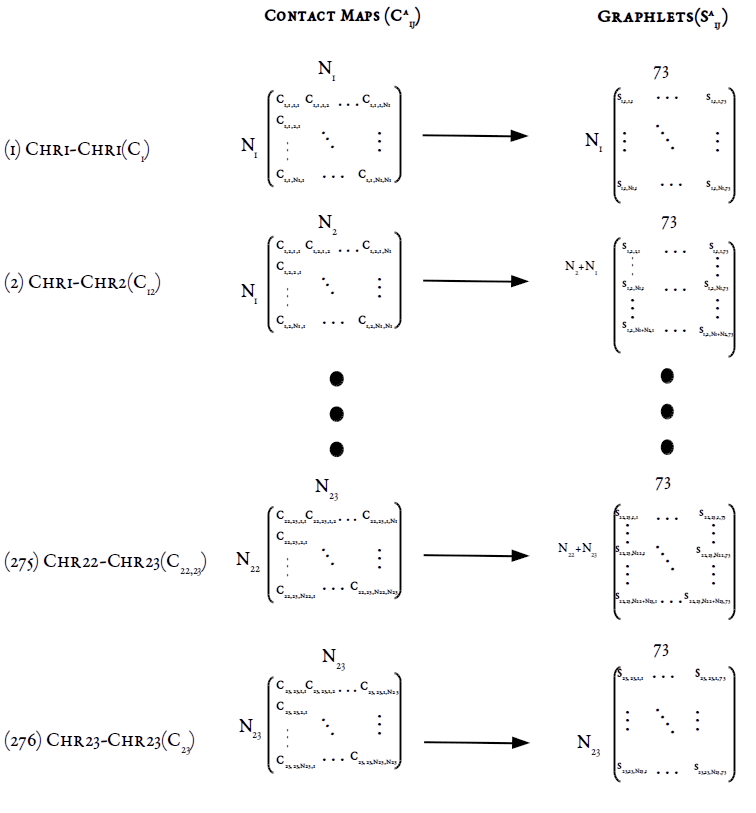


Figure 3: Graphlet extraction for the four cell lines. For each loci in each contact map between chromosomes i and j, the signature vectors of length 73 are extracted, resulting in a signature matrix of size , where  is the number of loci involved.

For a particular , we denote  as its *signature matrix*. Each cell  in  captures how many times loci  in  occurred as part of orbit .

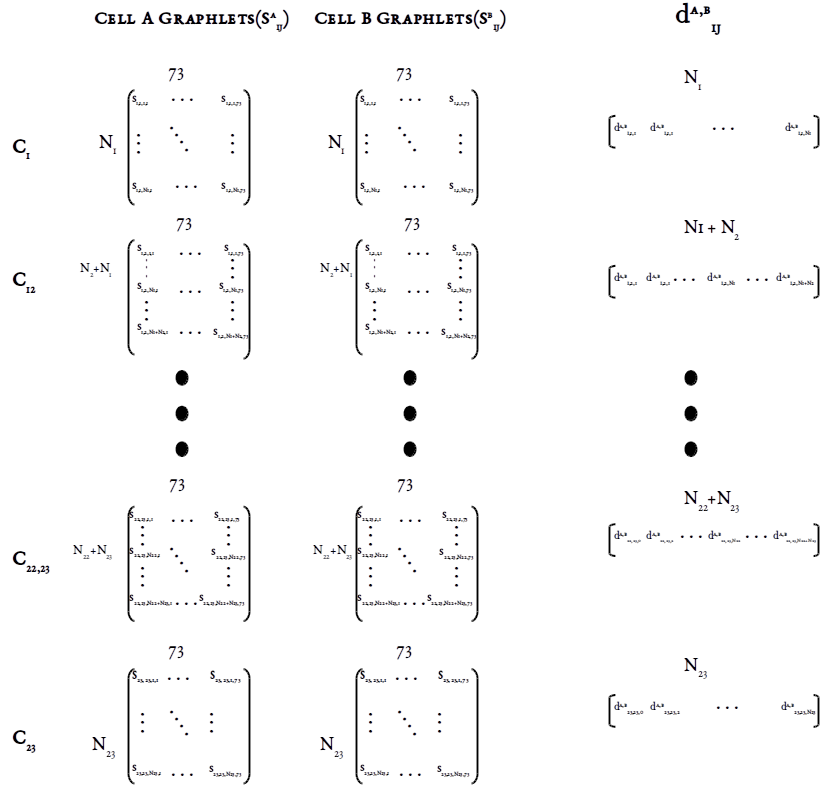


Figure 4: Calculating pair-wise loci distances. For each loci (row) in each contact map in MIT cell line, its distance is calculated based on equation 1 with the corresponding loci in leukemic cells. The result of this process is a signature distance vector of size  for each contact map.

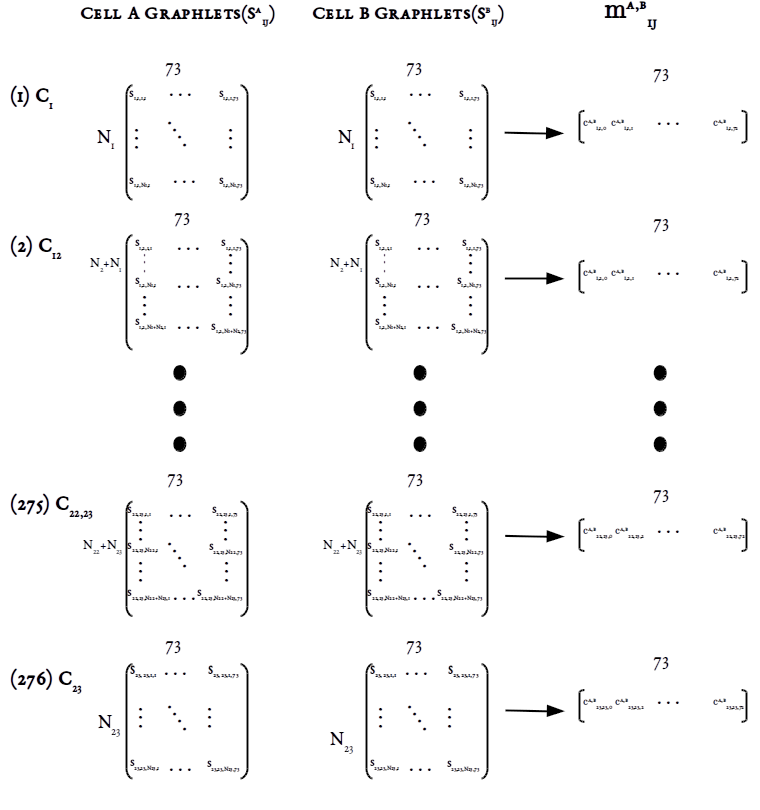


Figure 5: Calculating pair-wise orbit correlations. For each orbit (column) in each contact map in MIT cell line, its correlation with the same orbit in the same contact map in leukemic cells is calculated. The result of this process is a signature correlation vector of size 73 which captures how similar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

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  and  be their signature matrices. The *signature distance* of contact map  between A and B is denoted by.  is a vector of size  and its elements are calculated using the following formula from :

where elements of  is the distance between each loci (row)  in  and the the same loci in  for orbit  as is calculated as below:

 (Equation 1)

This process is illustrated in Figure [graphlet\_distance\_schema]. 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  can provide information regarding the *frequency distribution* of orbits throughout the contact map. We can find how similar these distributions are to each other using correlation measures. These correlations are denoted by  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) 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 dependent 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  is calculated as below:

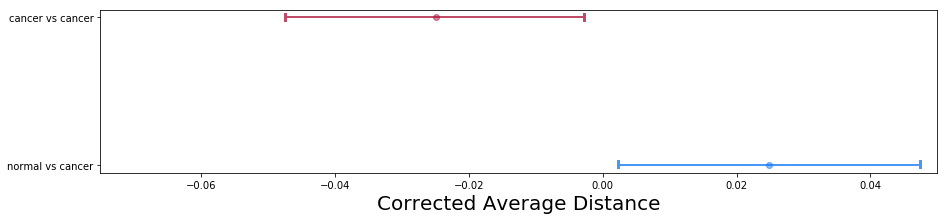


Alternatively, if we use Pearson criterion we would have:



# Results and discussions

## Contact Map Orbit Vector Distance



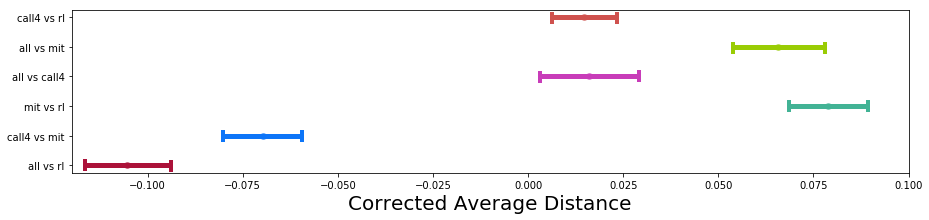


Figure 6

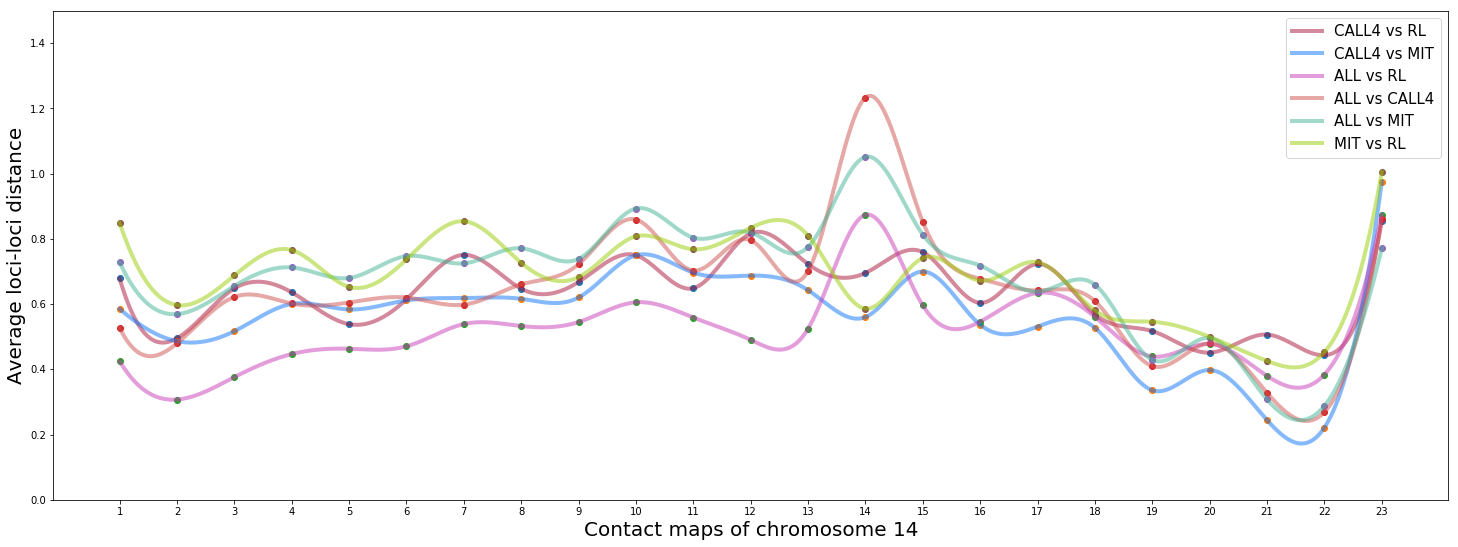


Figure 7

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.

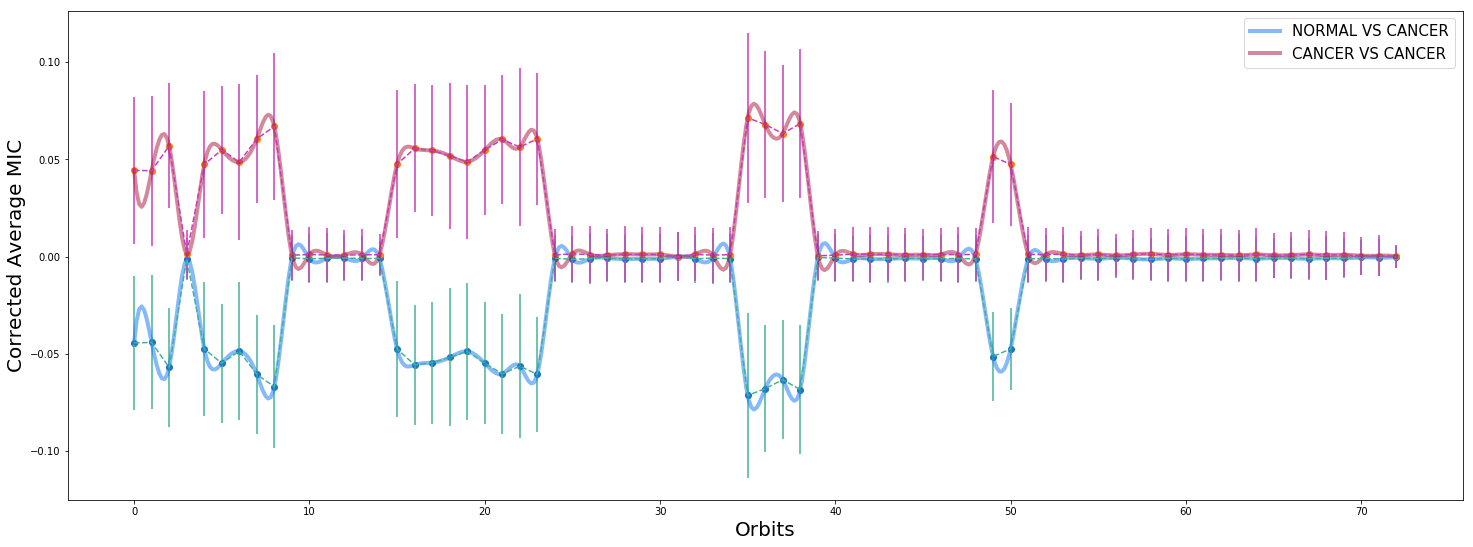
We can compare pairs of contact maps in terms of their closeness to each other. As an example, in figure 6, all pairs of cells are compared to each other in terms of their distance for contact maps involving chromosome 14. We can see that for the first 13 inter-chromosomal contact maps, ALL and RL cell 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. (). As illustrated in figure 5a, 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 5b. We can see that ALL-RL pair are closest to each other while ALL-MIT and MIT-RL are most distant. We found statistically 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 these tests can be found in supplementary material.

The results in figure 5b is also in keeping with what we see in figure 6. 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 6 for majority of contact maps.

## MIC Comparison



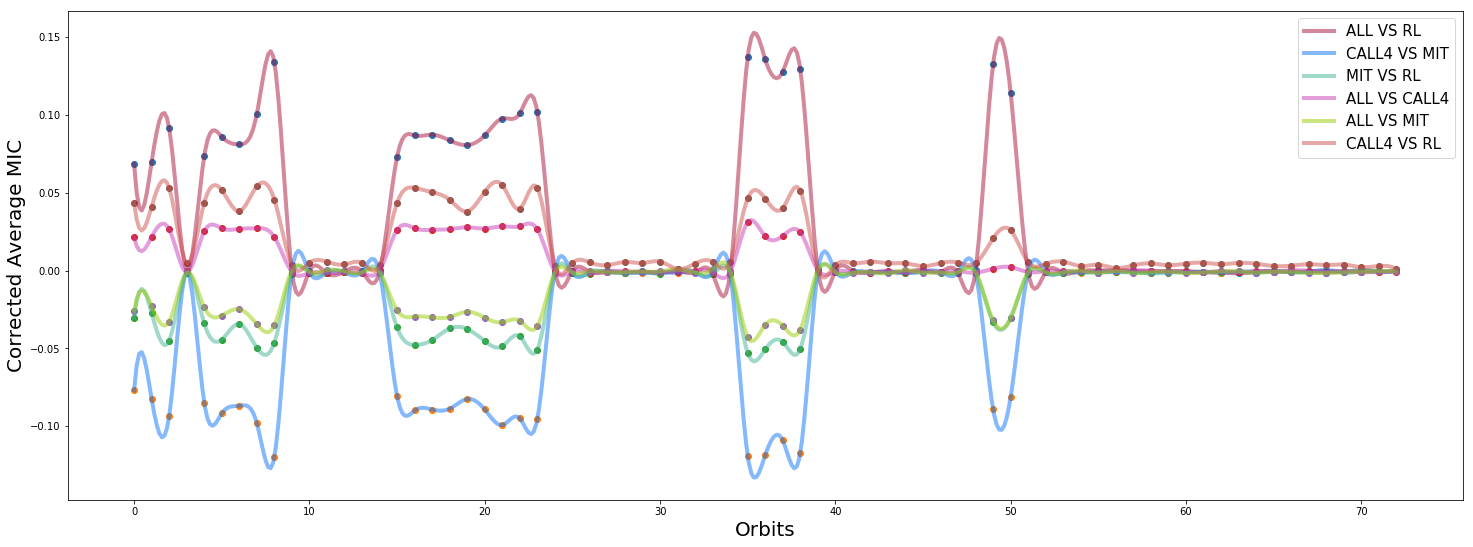


Figure 8

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 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 bipartite graph with the loci from each chromosome on one side. Due to this bipartite 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 7 where average correlations of orbits = {3, 9, 10-14, 20-34, 39-48, 51-72} are close to zero. In fact all correlations corresponding to these orbits are 0 except for the ones between the same chromosomes.

We calculated 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. (, , Wilk’s ) The difference is also illustrated in figure 7a, which plots corrected mean difference of MIC values for cancer-cancer and normal-cancer correlations. We performed statistical test to see if there is a significant difference correlation between individual pairs of cells. Figure 7b demonstrate such difference. Average correlation over all contact maps for normal-cancer pairs are smaller than cancer-cancer pairs. This is corroborated by the results of statistical test which verify that such difference is in fact significant. (, , Wilk’s ) Figures 7a 7b both show more details about this difference in correlation. As can be observed, for obits in , normal correlations are smaller than cancer correlations, while for the rest of the orbits, there is no difference.

Figure [fig:orbits\_correlations] 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, 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.

# Resources

**Hi-C Datasets:**

1. [Code base for this article](https://github.com/rasoolianbehnam/watson)
2. [Datasets including cancerous cells](http://sysbio.rnet.missouri.edu/T0510/tmp_download/link_to_download_genome_data/)
3. [Original Datasets](https://bcm.app.box.com/v/aidenlab/folder/11234760671)

1. Tel: +1 334 5212814; Email: bzr0014@auburn.edu [↑](#footnote-ref-2)