HiC Contact Map Comaprison Using Graphlet Approach

Behnam Rasoolian \*1, Debswapna Bhattacharya1

1Department of Computer Science and software engineering, Auburn, Alabama, 36846, United States

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

**ABSTRACT**

In this study, we investigated dissimilarities between normal cells and cancerous cells, through analyzing HiC contact maps. Our results show that certain orbit distributions have significanly higher correlation between Leukemic cells.

# INTRODUCTION

Graphlet comparison is a novel method used to compare large networks in order to find local similarities in them. Authors of provide a new measure of PPI network comparison based on 73 constraints. This is used in order to compare two large networks in order to detect similarities.

provide heuristics to compare two nodes based on some feature (or signature) vectors, which is a 73-dimensional vector where denotes the number of nodes in the network that are part of an orbit .  
Important Result: Proteins with similar surroundings perform similar functions.

In , the same author investigates cancer-causing genes to find similarities in their signatures. After clustering the genes based on signature similarity criteria, some clusters contain a lot of cancerous genes. They use 4 different clustering methods with varying parameters to cluster the proteins. They then predict the cancer-relatedness of a protein using an enrichment criteria where is the cluster where protein belongs and is the number of cancer-causing proteins in and is the size of .

The authors of generalized the idea of graphlets to ordered graphs were the nodes are labeled in ascending order. As can be viewed, there are a total of 14 orbits for graphlets of size 2 and 3 since the label of graphlets is also included in toplogy. In the new definition, denotes the number of orbit touches node . Each node, is then assigned a vector of length 14 [[1]](#footnote-2) and similarity of two nodes in two contact maps can be compared by how geometrically close their corresponding vectors are.

### Notations

In this paper, matrices and vectors are represented with 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, containts only the loci of that particular chromosome (), while in inter-chromosomal contact maps contains the loci in the both of chromosomes involved ().

# MATERIALS AND METHODS

We re-used Leukemic Hi-C libraries created in These libraries we sequenced 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). Just as , we used normal B-cell line (GM068990) from for our comparisons. We created contact maps of resolution 500kb and normalized it using the iced package in python developed by .

## 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.

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 1 demonstrates result of this thresholding approach for intra-chromosomal contact maps of chromosome 1. Refer to supplementary material for all 23 interchromosomal thresholding results.

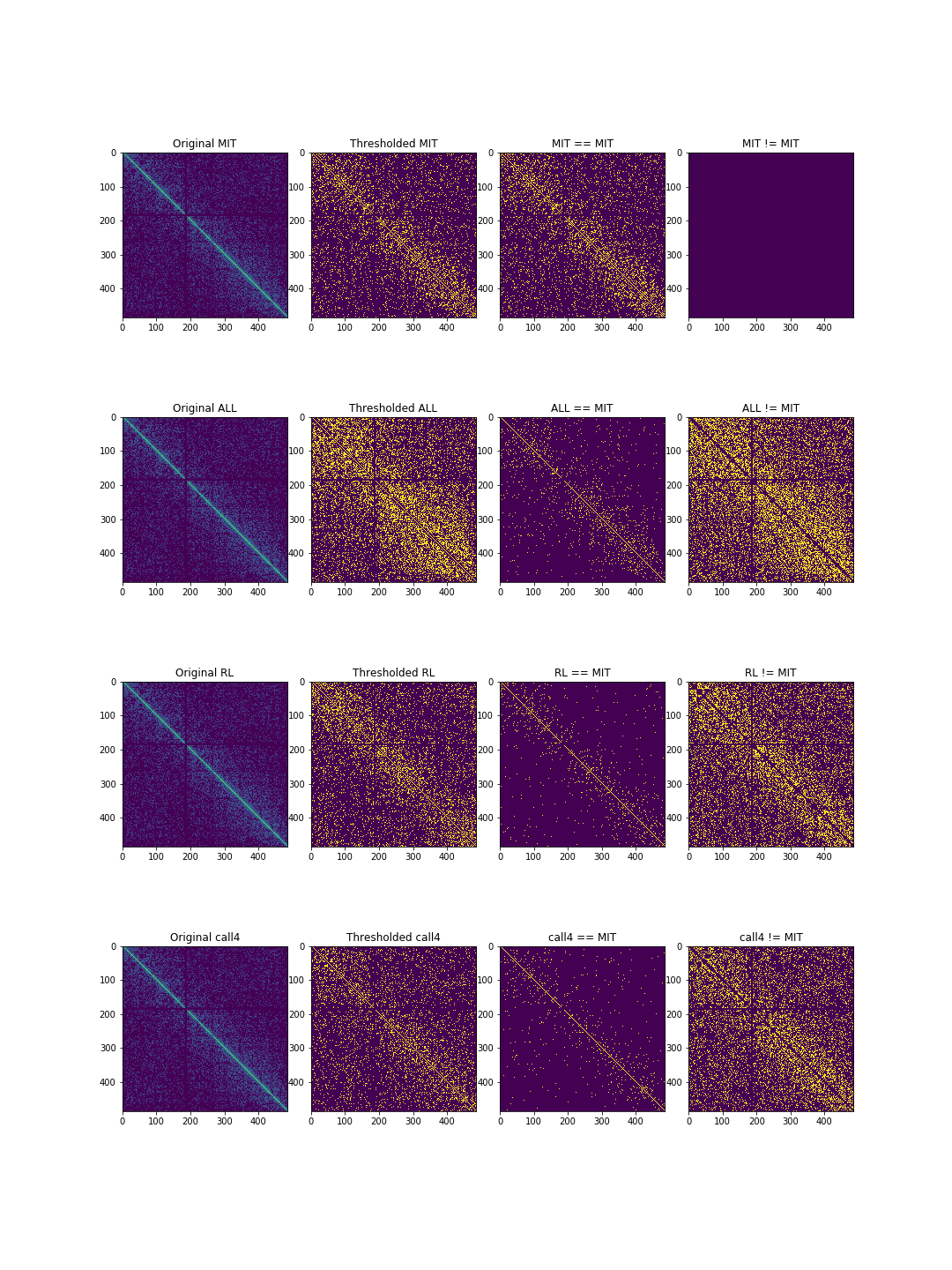


Figure 1. Result of thresholding interchromosomal contact map of chromosome 1 using a kernels of size for all cell lines. The first row shows the thresholded maps. Second and third rows demonstrate pair-wise similarities and differences between contact maps respectively.

## 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.

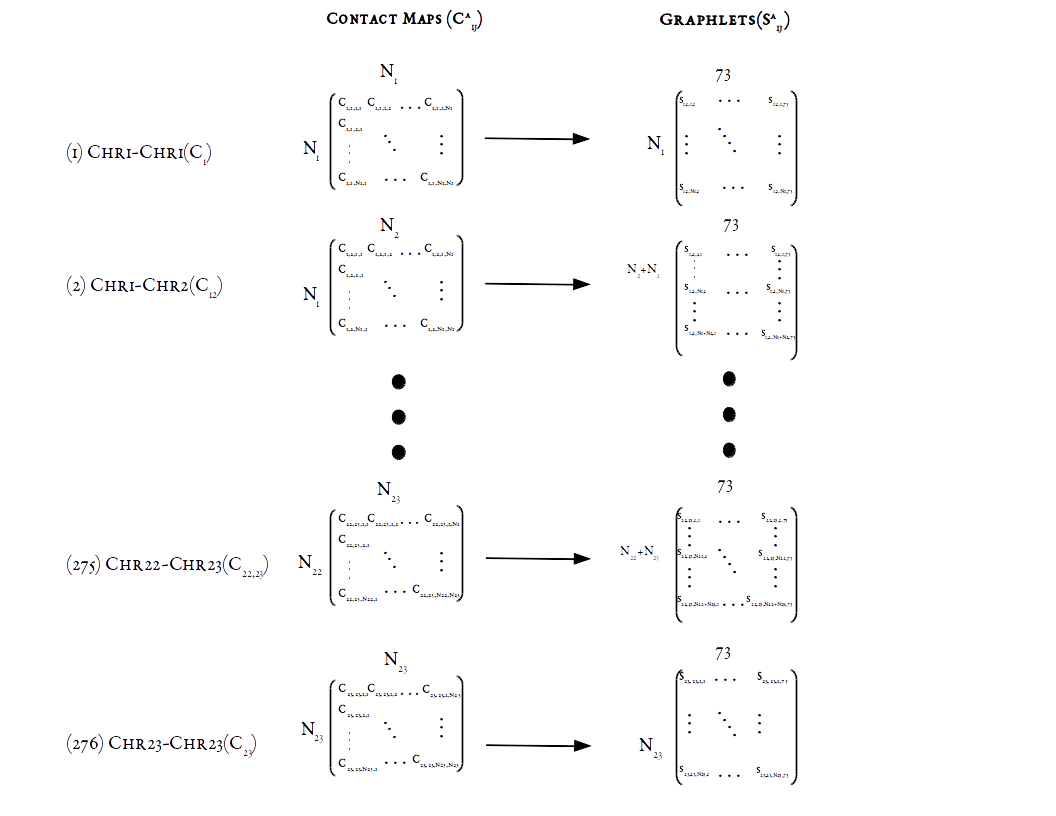
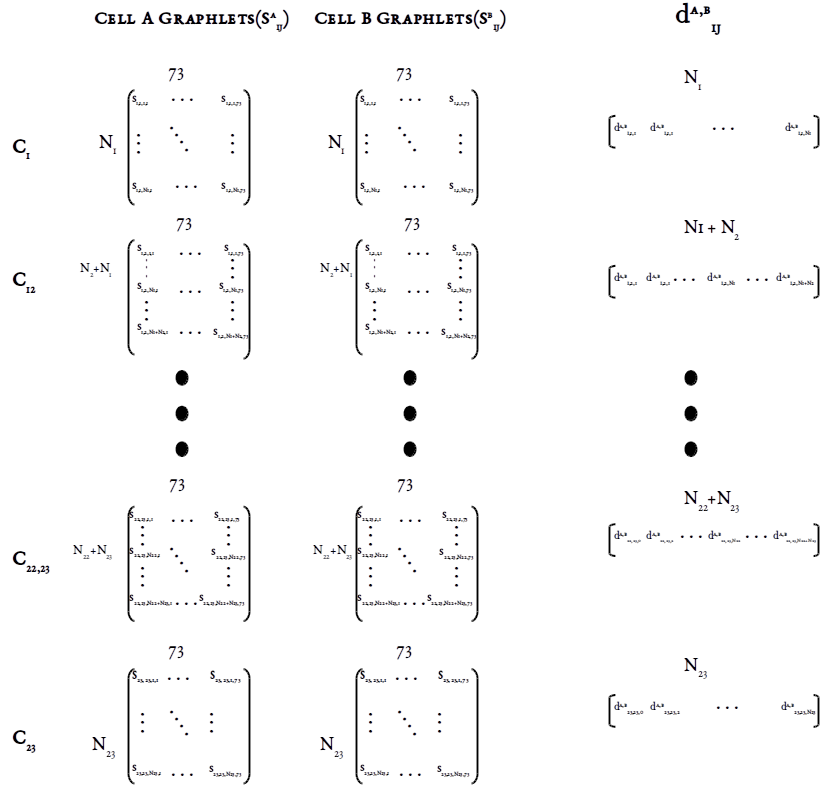
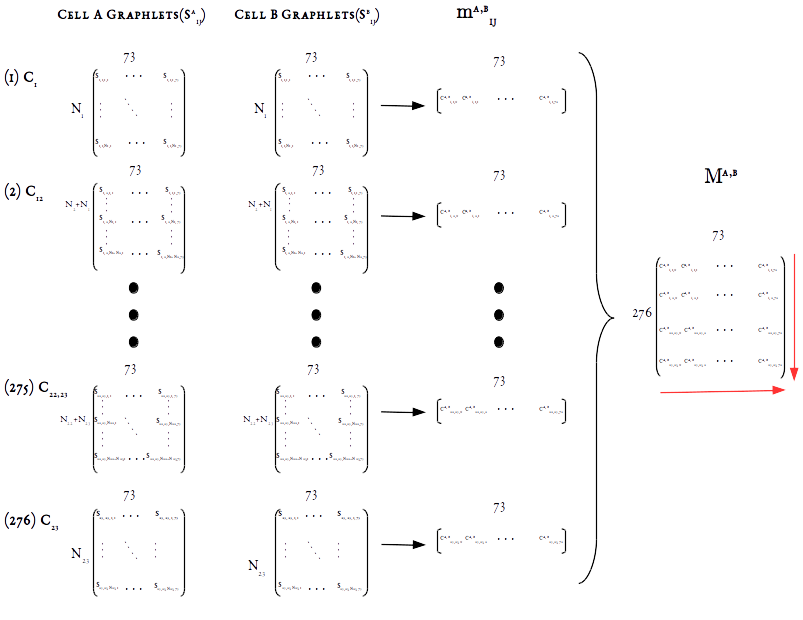


Figure 2. Graphlet extraction for the four cell lines. For each loci in each contact map between chromosomes and , the signature vectors of length 73 are extracted, resulting in a signature matrix of size , where is the number of loci involved.

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



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



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. In order to test our second hypothesis, we calculated averages across contact maps (along the vertical red arrow) to test hypothesis [eq:h2a] and across orbits (along the horizontal red arrows) to test hypothesis [eq:h2b].

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 {d}^{ A,B}\_{ij}. {d}^{ A,B}\_{ij} is a vector of size and its elements d^{ A,B}\_{i,j,l} 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:

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 {m}^{ A,B}\_{i,j} 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 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.

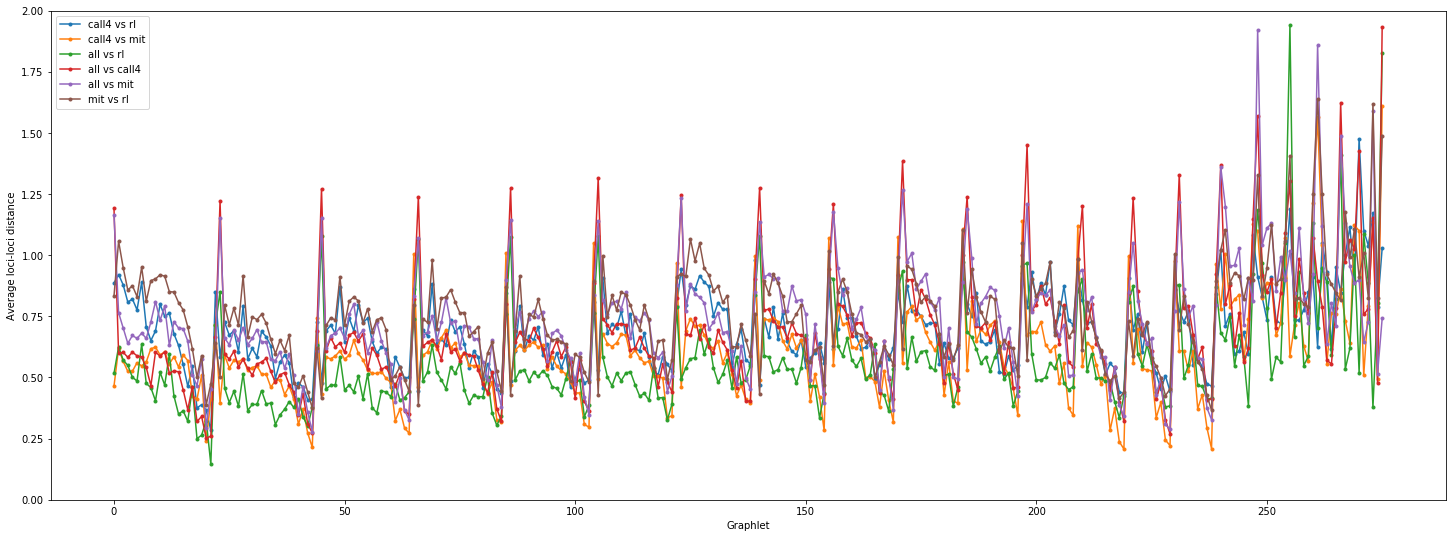
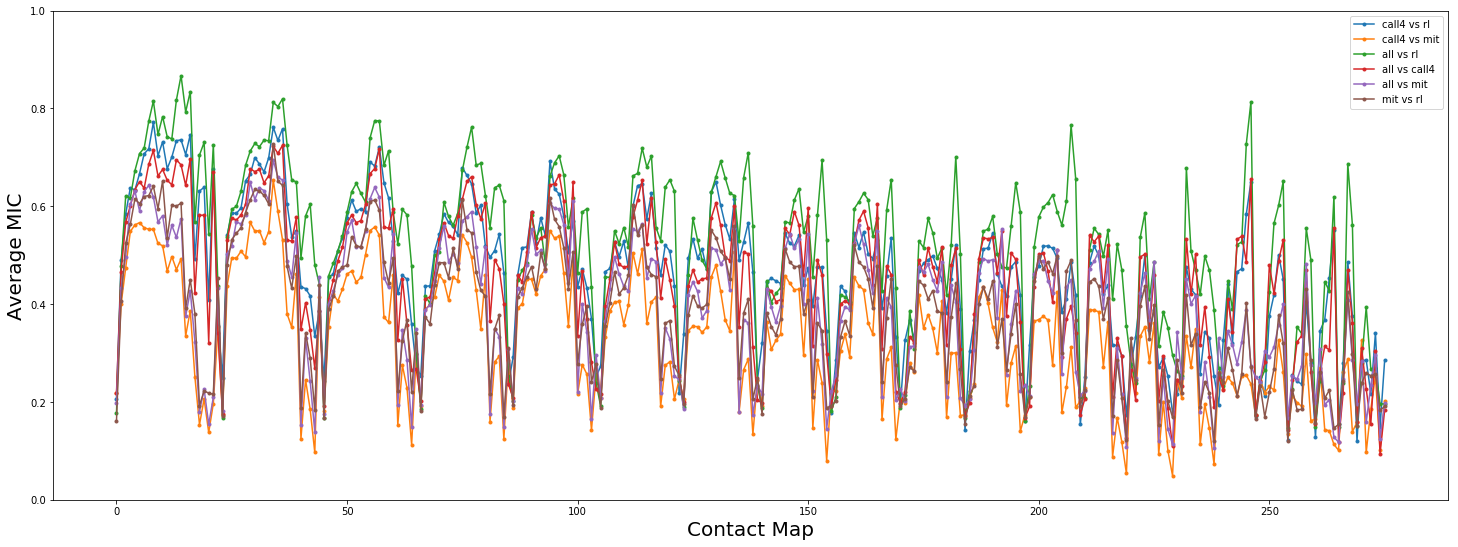
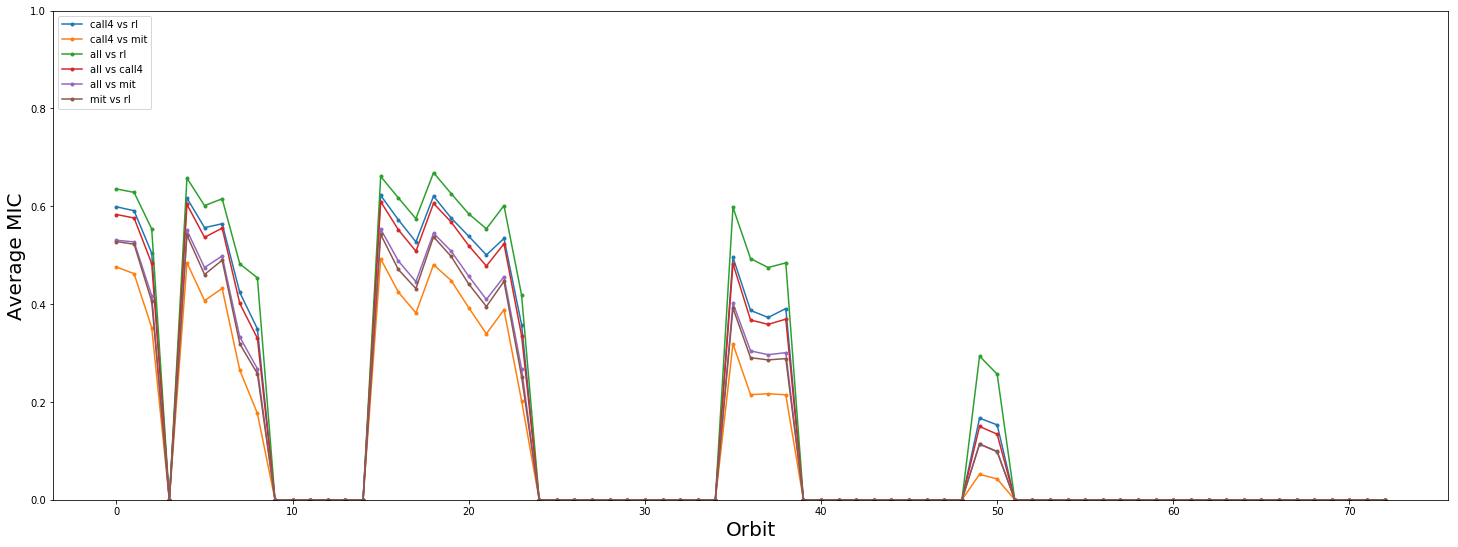
If MIC is used as correlation measure, each element of is calculated as below:

Alternatively, if we use Pearson criterion we would have:

# RESULTS AND DISCUSSIONS

Result of pair-wise contact map graphlet distances can is illustrated in Figure [fig:orbit-distances\_all]. Each point on the graph is the average of the graphlet distance vector () of the two cell lines specified in the legend. A one sided paired t-test was conducted in order to compare distances to zero and also to each other. The resulting p-values showed highly significant differeces from zero for all pairs of cells. Detailed results of t-tests for each contact map can be found in supplementary materials. For each contact map, each pair of cells are ordered based on whether they are statistially larger that the other or not. For example for contact map , the results of t-test is as follows:  
  
which means that is statistically larger than 0, but less than and so forth. We can also conclude that {d}^{ ALL, CALL4}\_{1,1} is not statistically different from { between graphlets extracted from all contact maps. Refer to supplementary material for results of all hypotheis tests.

We caclulated pair-wise MIC for each orbit in each of the 276 contact maps from MIT data and ALL, RL and CALL4 data separately. Figure [fig:contact\_maps\_correlations\_all] shows average orbit correlations across all contact maps, while figure [fig:orbits\_correlations\_all] demonstrates average correlations across all 72 orbits within each contact map. It is worth mentioning that inter chromosomal 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. We ignored these values when we calculated averages across orbits in figure [fig:contact\_maps\_correlations\_all] since they would result in a bias towards zero in averages. You can see the bias in figure [fig:orbits\_correlations\_all] where average correlations of orbits are close to zero. Actually, all correlations corresponding to these orbits are 0 except for the ones between the same chromosomes.

Figure ?? clearly illustrates certain orbits of Leukemic have higher correlation to each other than to the normal MIT cell. In fact our statistical analysis shows that for orbits 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.   

We have conducted one-sided t-test in order to test whether the average correlations across contact maps is equal to 1 and whether the average correlations across orbits is equal to 1. The results for both test showed that all values are significantly less than 1. Please refer to supplementary material for result of the full list of t-test results.

# 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. number of orbits in graphlets of size 2 and 3 [↑](#footnote-ref-2)