Published online 18 March 2017

Nucleic Acids Research, 2017, Vol. 37, No. 12 ***1–4***
doi:10.1093/nar/gkn000

HiC Contact Map Comaprison Using Graphlet Approach

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

[1] Auburn University

ABSTRACT

In this study, we plan to find dissimilarities between normal cells and cancerous cells, through investigating HiC contact maps. We suspect that there are systematic differences between how chromosomes are structured between normal cells and cancerous cells. we Ideally, it is desirable to compare 3D structures of cell in order to make such comparisons. However, the main challenge that we face is that 3D structure of a cell is not readily available. Based on (1), fluorescence in situ hybridizaiton (FISH) is used for investigating 3D configuration of chromosomes. However, this method can only be used locally and cannot map the whole structure of the chromosomes. In orther to find dissimilarities in the 3D structure of chromosomes, we used HiC dataset. The HiC method, which was developed by (2), captures interactions between chromosomal fragments in kilobase resolution. Based on HiC data, an *interaction frequency (IF)* matrix can be developed between *loci* at a desired resolution. A cell IF_{ij} in an interaction frequency matrix captures the number of interaction detected in HiC dataset between locus i and locus j in the genome. An interaction matrix can be used to develop both inter- and intra-chromosomal interaction matrices. We believe differences in interaction matrices can be found between normal cells and cancerous ones.

INTRODUCTION

Graphlet comparison is a novel method used to compare large networks in order to find local similarities in them. Authors of (3) 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.

(4) provide heuristics to compare two nodes based on some feature (or signature) vectors, which is a 73-dimensional vector $\mathbf{s}^T = [s_0, s_2, ..., s_{72}]$ where s_i denotes the number of nodes in the network that are part of an orbit i.
Important Result: Proteins with similar surroundings perform similar functions.

In (5), 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 i using an enrichment criteria $\frac{k}{|C_i|}$ where C_i is the cluster where protein i belongs and k is the number of cancer-causing proteins in C_i and $|C_i|$ is the size of C_i.

The authors of (6) 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, d_v^i denotes the number of orbit i touches node v. Each node, is then assigned a vector of length 14 [1] $(d_v^1, d_v^2, ..., d_v^{14})$ and similarity of two nodes in two contact maps can be compared by how geometrically close their corresponding vectors are.

MATERIALS AND METHODS

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 ith row of matrix M is denoted by $M_{i.}$ and its jth column is represented by $M_{.j}$.

We denote the set of all contact maps in cell line T with $\mathbb{C}^T$. If no particular cell line is addressed, the subscripts are dropped. Any arbitrary member of $\mathbb{C}$ is denoted by 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 $C_{i,i} = C_i$.

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 ($|\mathbb{V}_{ij}| = N_i + N_j$).

[1] number of orbits in graphlets of size 2 and 3

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

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.

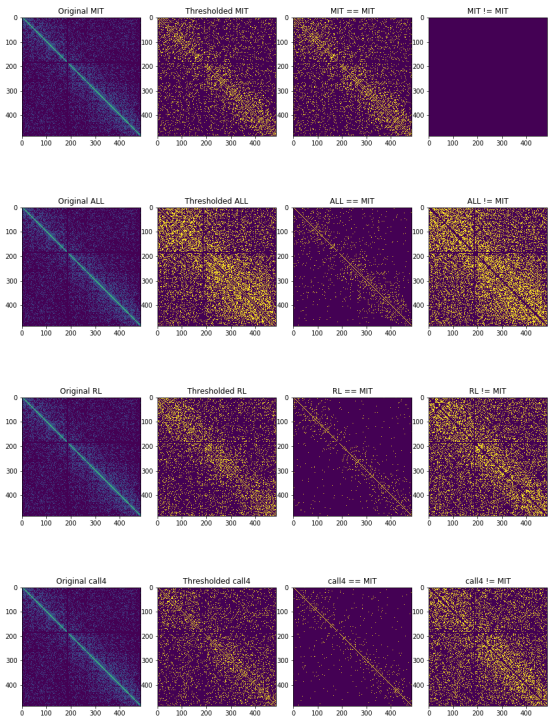


Figure 1. Result of thresholding interchromosomal contact map of chromosome 1 using a kernels of size 5×5 for all cell lines. The first row shows the thresholded maps. Second and third rows demonstrate pair-wise similarities and differences between contact maps respectively.

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 $|V^{ij}|\times73$, where V^{ij} is the number of loci involved in contact map between chromosomes i and j. Figure 2 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 (3):

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

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:

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

This process is illustrated in Figure 3. 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

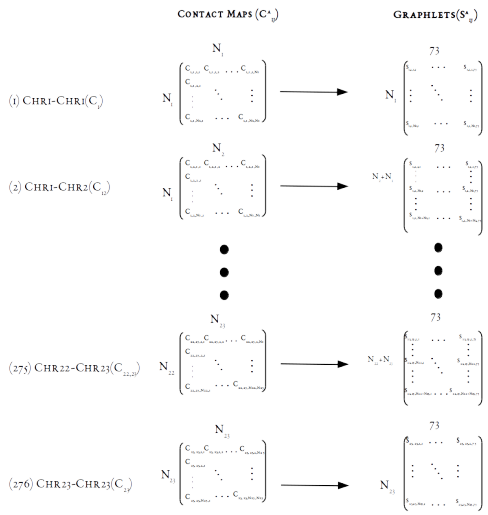


Figure 2. Graphlet extraction for the four cell lines. For each loci in each contact map between chromosomes i and j, the signature vectors of length 73 are extracted, resulting in a *signature matrix* of size $|V^{ij}| \times 73$, where V^{ij} is the number of loci involved.

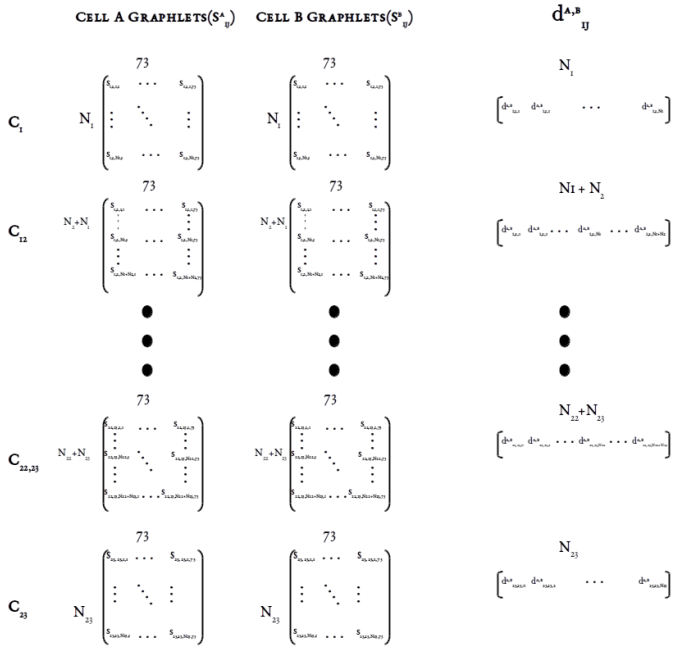


Figure 3. Calculating pair-wise loci distances. For each loci (row) in each contact map in MIT cell line, its distance is calculated based on equation **1** with the corresponding loci in leukemic cells. The result of this process is a *signature distance vector* of size $|V^{ij}| = N_i + N_j$ for each contact map.

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 $\mathbf{m}^{A,B}_{i,j}$ 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

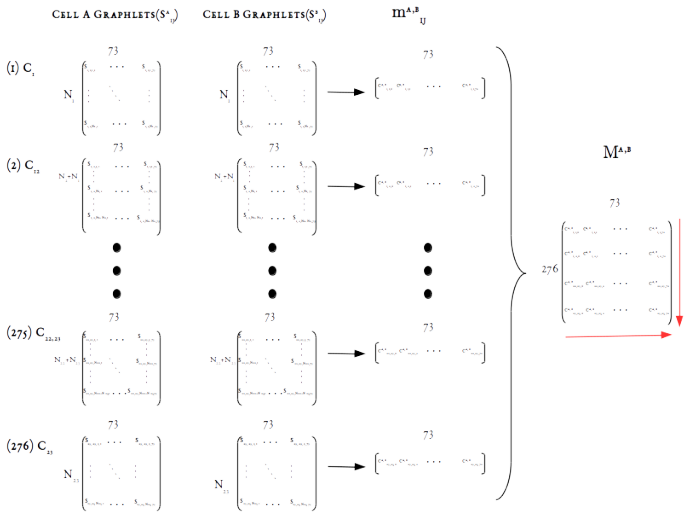


Figure 4. Calulating pair-wise orbit correlations. For each orbit (column) in each contact map in MIT cell line, its correlation with the same orbit in the same contact map in leukemic cells is calculated. The result of this process is a *signature correlation* vector of size 73 which captures how similar 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 **??** and across orbits (along the horizontal red arrows) to test hypothesis **??**.

not be able to capture non-functional relationships between distributions. As a result, we also used Maximal Information Coefficient (MIC) (7) 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 MI 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 $\mathbf{c}$ is calculated as below:

$$m^{A,B}_{ijo} = MIC(\mathbf{S}^A_{ij.o}, \mathbf{S}^B_{ij.o}) \quad \textbf{(3)}$$

Alternatively, if we use Pearson criterion we would have:

$$m^{A,B}_{ijo} = Pearson(\mathbf{S}^A_{ij.o}, \mathbf{S}^B_{ij.o}) \quad \textbf{(4)}$$

RESULTS

Result of pair-wise contact map graphlet distances can is illustrated in Figure 5. Each point on the graph is the average of the graphlet distance vector ($\bar{\mathbf{d}}^{A,B}_{i,j}$) of the two cell lines specified in the legend. A one sided paired t-test was conducted . 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 fount in supplementary materials. For each contact map, each pair of cells are order based on whether they are statistially larger that the other or not. For example for contact map $C_{i,j}$, the results of t-test is

as follows:

zero $<$ call4-mit $<$ all-rl $<$ mit-rl $<$ call4-rl $<$ all-mit $=$ all-call4

which means that $\mathbf{d}_{1,1}^{CALL4,MIT}$ is statistically larger than 0, but less than $\mathbf{d}_{1,1}^{ALL,RL}$ and so forth. We can also conclude that $\mathbf{d}_{1,1}^{ALL,CALL4}$ is not statistically different from $\mathbf{d}_{1,1}^{ALL,MIT}$ 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 6 shows average orbit correlations across all contact maps, while figure 7 demonstrates average correlations across all 72 orbits within each contact map. 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. We ignored these values when we calculated averages across orbits in figure 6a since they would result in a bias towards zero in averages. You can see the bias in figure 7a 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 which are illustrated in Figure 7b.

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 t-tests.

DISCUSSION

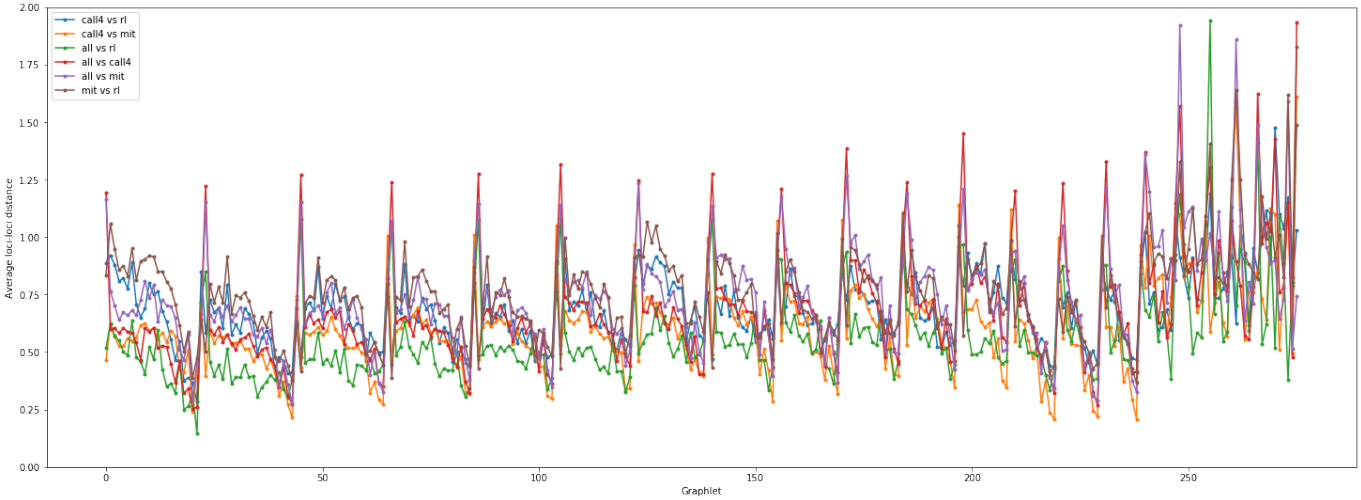
RESOURCES

Hi-C Datasets:

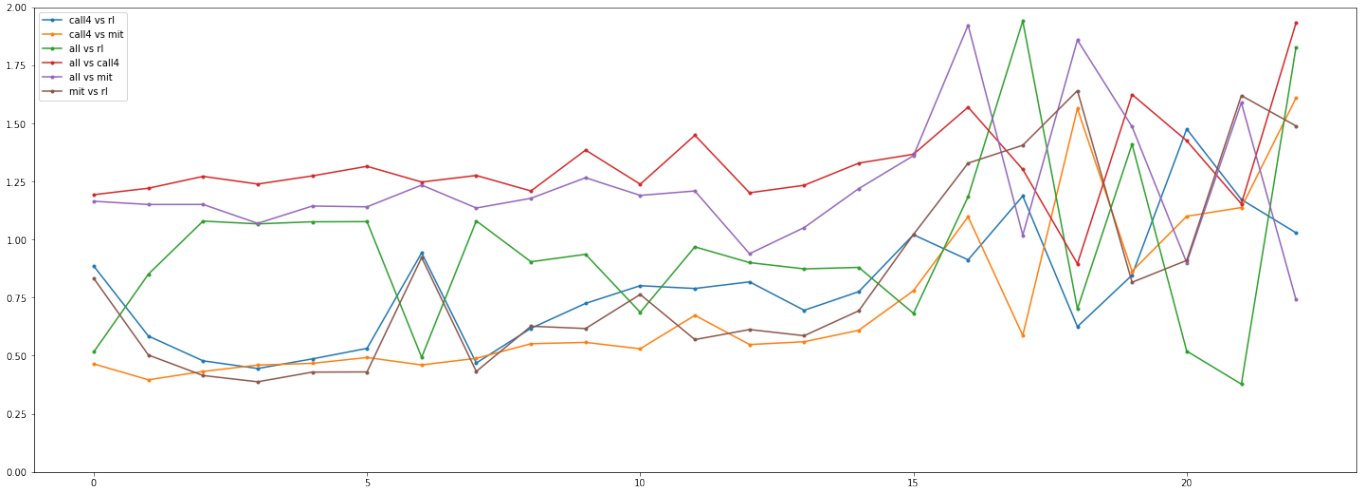
1. Code base for this article
2. Datasets including cancerous cells
3. Original Datasets

REFERENCES

1. 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.
2. 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.
3. Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.
4. Tijana Milenkoviæ and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6:257, 2008.
5. Tijana Milenković, Vesna Memišević, Anand K Ganesan, and Nataša Pržulj. Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. *Journal of the Royal Society Interface*, 7(44):423–437, 2010.
6. Pietro Di Lena, Piero Fariselli, Luciano Margara, Marco Vassura, and Rita Casadio. Fast overlapping of protein contact maps by alignment of eigenvectors. *Bioinformatics*, 26(18):2250–2258, 2010.
7. 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.



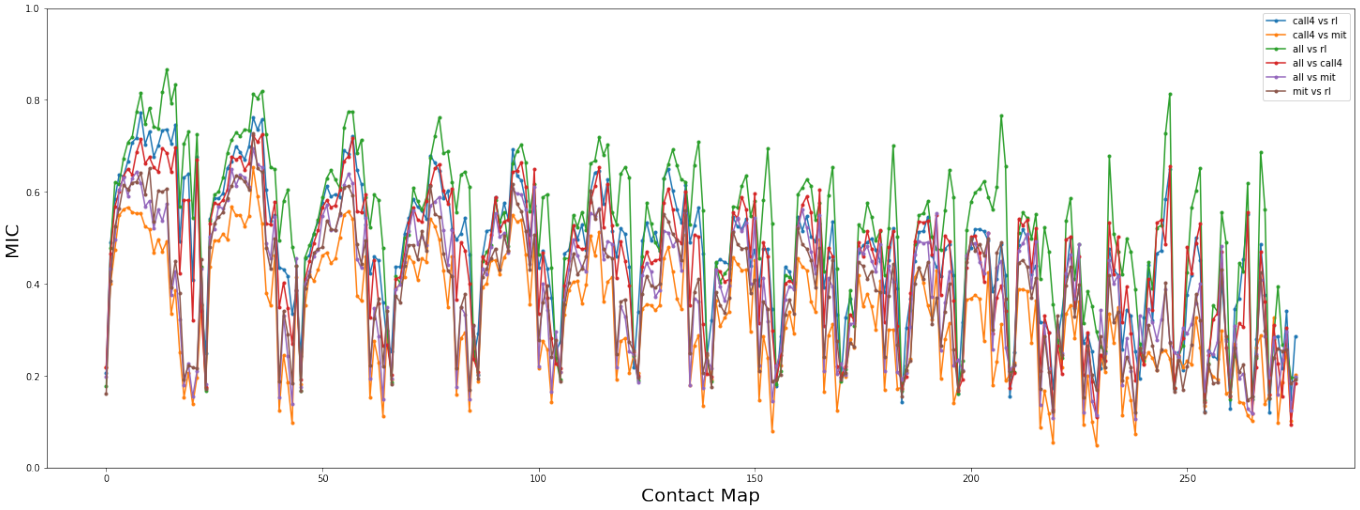
(a)



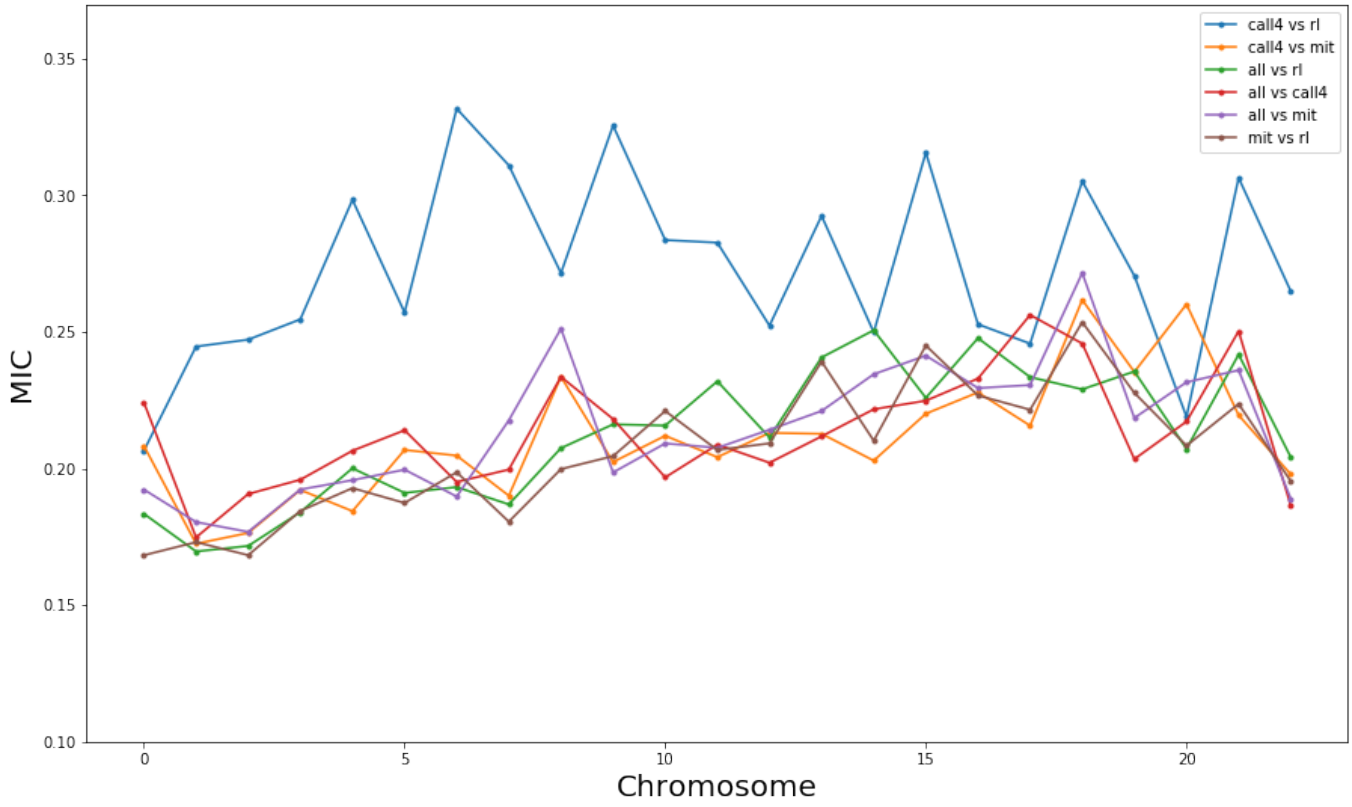
(b)

(c)

Figure 5. Pair-wise graphlet signature difference for (**a**) all 276 contact maps ($\bar{\mathbf{d}}^{A,B}_{i,j} \quad \forall i,j \in \{1...23\} \quad \& \quad j \geq i$). (**b**) only the 23 intra-chromosomal contact maps ($\bar{\mathbf{d}}^{A,B}_{i,i}$). (**c**) all 276 contact maps as well as corresponding standard errors. Each point on a graph is the result of averaging all the distances acroos all loci of that contact map. As can be viewd, although inter-chromosomal graphlets do not show significant differences, RL's inter-chromosomal graphlets seem to be closer to MIT's. Also in some contact maps ALL graphlets show greater distance from MIT graphlets than the other two cell lines.

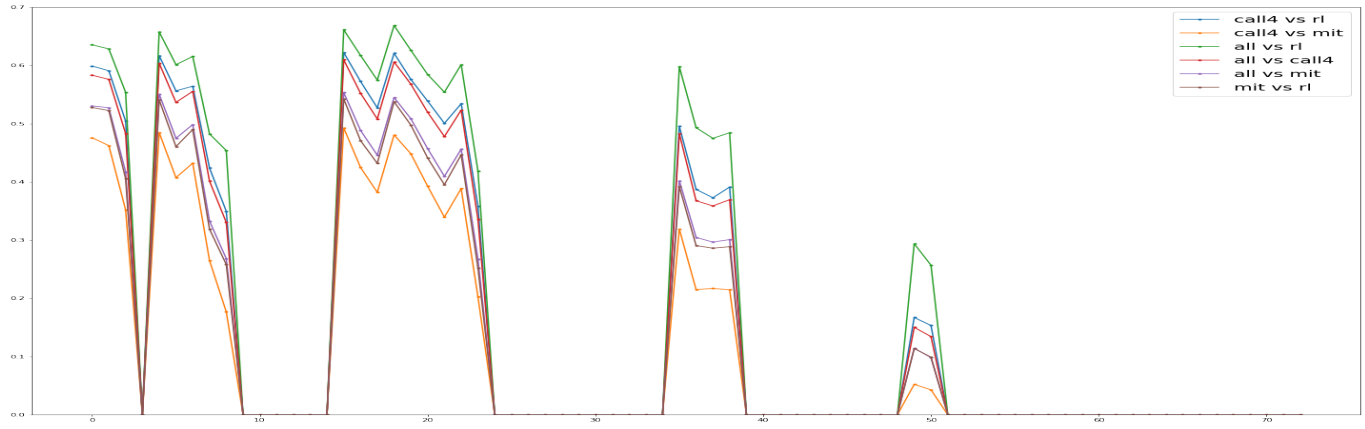


(a)

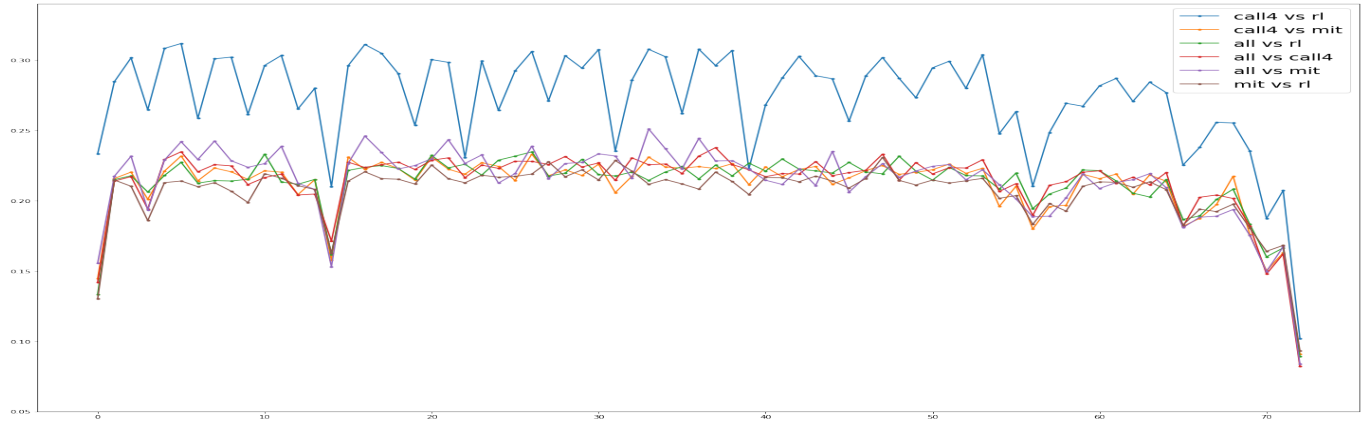


(b)

Figure 6. pair-wise average contact map orbit correlations (a) 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 4) (b) only for intra-chromosomal contact maps ($\bar{\mathbf{m}}^{A,B}_{i,i} \quad \forall i \in \{1...23\}$) These values are calculated by averaging over pairwise correlations of all orbits in a contact map.



(a)



(b)

Figure 7. pair-wise average orbit correlations. In figure 7a, each point in the graph is the result of averaging pair-wise orbit correlations over all contact maps ($\frac{1}{276}\sum_{i=0}^{23}\sum_{j=i}^{23} m_{i,j,o}^{A,B} \quad \forall o \in \{0,1,...,72\}$, average along the red *horizontal* arrows in figure 4), while each point in figure 7b are averaged only over intra-chromosomal contact maps ($\frac{1}{23}\sum_{i=0}^{23} m_{i,i,o}^{A,B} \quad \forall o \in \{0,1,...,72\}$). Counts for certain orbits are always zero in inter-chromosomal maps, leading to average value close to zero in 7a.