Garaphlet Analysis for HiC data

Behnam Rasoolian

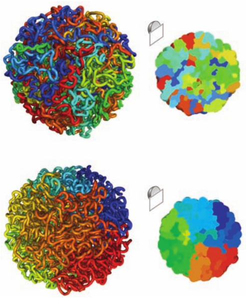
Auburn University

Biological Background

Purpose of this research

Introduction

- In this research we plan to find dissimilarities between normal cells and cancerous cells.
- Ideally, it is desirable to compare **3D confomation** of genomes in order to make such comparisons.

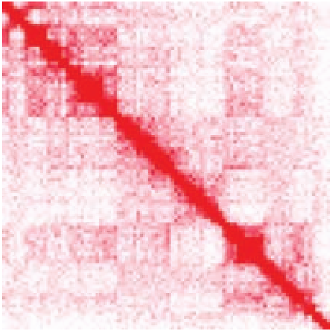


(Lieberman-Aiden et al., 2009)

Purpose of this research

Challenges

- We still don't have enough information regarding the exact configuration of a genome inside nucleus.
- However, we can map interactions in an *HiC contact map* (C).
- Rows and columns signify genome fragments.
- C_{ij} = Number/strength of interactions detected between fragment i and j.

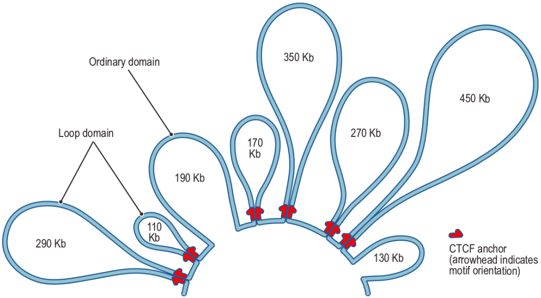


(Lieberman-Aiden et al., 2009)

Preliminaries

What is 3D conformation?

If you unfold the DNA inside one of your cells, it would measure 2 meters end to end. How is it folded up withing a nucleus which is only 6 micorns wide?



(Rao et al., 2014)

Preliminaries

Terminology (Wang et al., 2013)

- **Nucleotide**: The monomer units that comprise DNAs. There are 4 types of nucleotides: (C, G, A, and T)
- **Base**: Each pair of nucleotides in the DNA are called a base.
  A kilo-base resolution is a resolution that corresponds to 1000 pairs of nucleotides in DNA.
- **Nucleosome**: A basic unit consisting of 145-147 bases wrapped around a protein complex.
- **Chromatin Fiber**: Tens of nucleosomes are further collapsed into a larger dense structural unit of several kilobase (Kb) pairs.
- **Locus**: Multiple chromatin fibers form a large module of megabase pairs (Mb) DNA, which may be referred to as *domains, globules, gene loci, or chromatin clusters* in different contexts.
- **Chromosome**: A number of loci then fold into a large independent physical structure, chromosome.

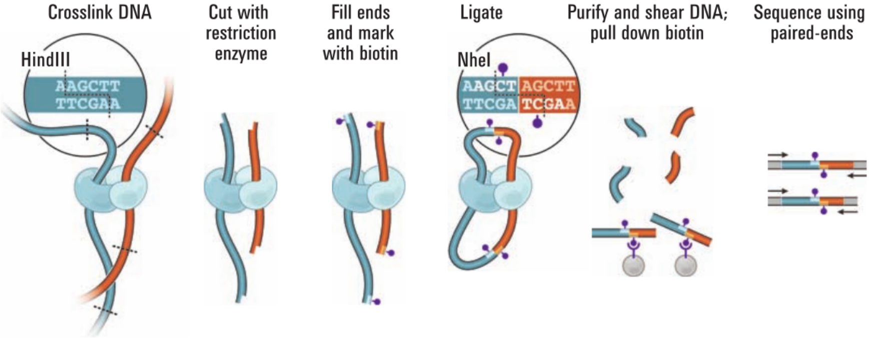
HiC Method

Procedure

1. Freeze the DNA in place.
2. Cut the genome in tiny pieces. Mark the ends using Biotin, and glue them together into diffused pieces of DNA. These diffused pieces is made up of two bits of the genome that are spatial neighbors.
3. Using DNA sequencing, the two parts of the diffused DNA are identified and a dataset is created where each cell corresponds to a pair.

HiC Method

Illustration

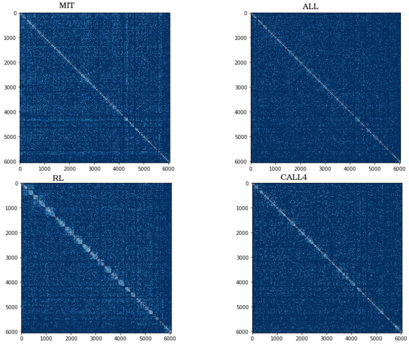

Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

(Lieberman-Aiden et al., 2009)

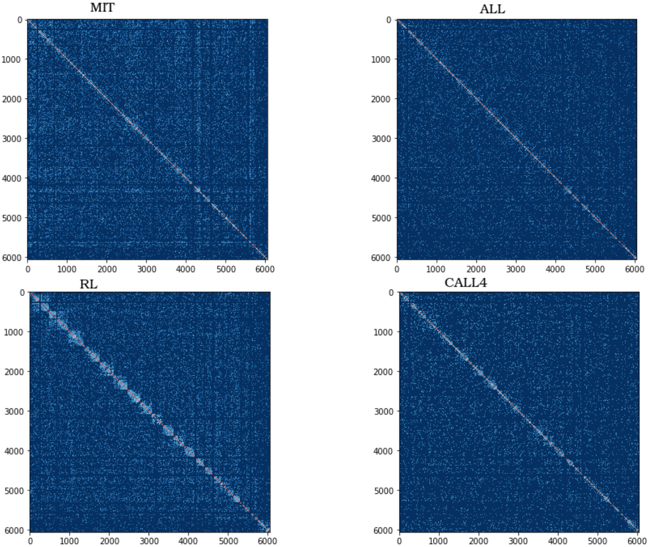
HiC Method

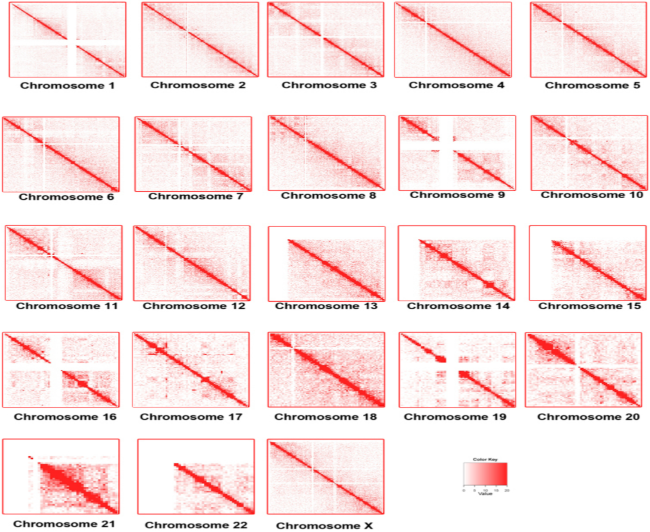
Contact Maps

- The whole genome is then divided into sections of certain length (i.e. 500kB or 1MB) and interactions are aggregated over them.
- Contact maps can be used to develop both inter- and intra-chromosomal interaction matrices.



HiC Data: An Overal Picture





(Wang et al., 2013)

Strategies

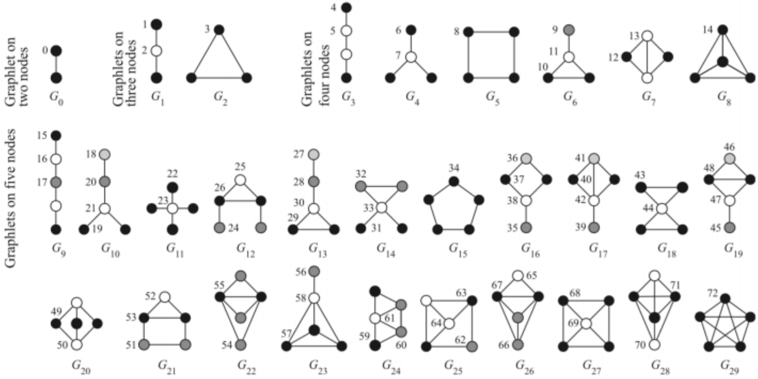
Graphlets

Definitions

Graphlet comparison, introduced by Pržulj (2007), is a novel method used to compare large networks in order to find local similarities in them.

- **Fragment:** A connected subgraph.
- **Motifs:** Fragments that occur with a frequency much higher than that occuring in a randomly generated graph.
- **Graphlets:** An arbitrary, induced fragment. An edge is the only two-node graphlet.
- **Induced graphs:** Given a graph $G(V, E)$ and $S \subseteq V$, then $G'(S, E')$ is a graphlet iff
  $E' = \{(u, v) | u, v \in V \text{ and } (u, v) \in E \rightarrow (u, v) \in E'\}$
- **Orbits:** Set of all nodes in a graphlet that can be swapped with each other while not changing the graph.

Graphlets



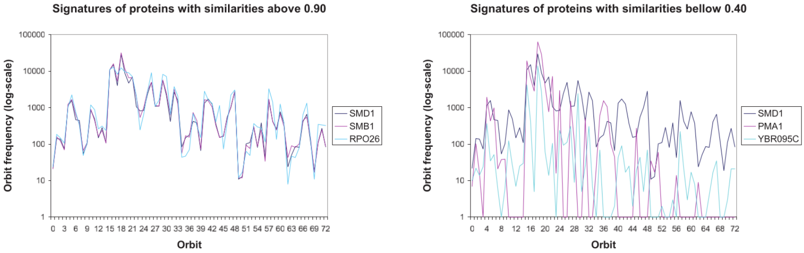
all 30 undirected two- to five-node graphlets with 73 orbits

All 30 undirected two- to five-node graphlets with 73 orbits (Pržulj, 2007)

Graphlets

Applications

Milenkoviæ and Pržulj (2008):
Signature vector: A 73-dimensional vector $\mathbf{s}^T = [s_0, s_2, \ldots, s_{72}]$ 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.



(Milenkoviæ & Pržulj, 2008)

Introduction

Milenković, Memišević, Ganesan, and Pržulj (2010): Investigate cancer-causing genes to find similarities in their signatures.

1. Cluster the genes based on *signature similarity* criteria. Some clusters contain a lot of cancerous genes.
2. Predict the cancer-relatedness of a protein i using an enrichment criteria $\frac{k}{|C_i|}$
    - C_i : the cluster where protein i belongs
    - k : the number of cancer-causing proteins in C_i
    - $|C_i|$: the size of C_i

Stage I: Thresholding

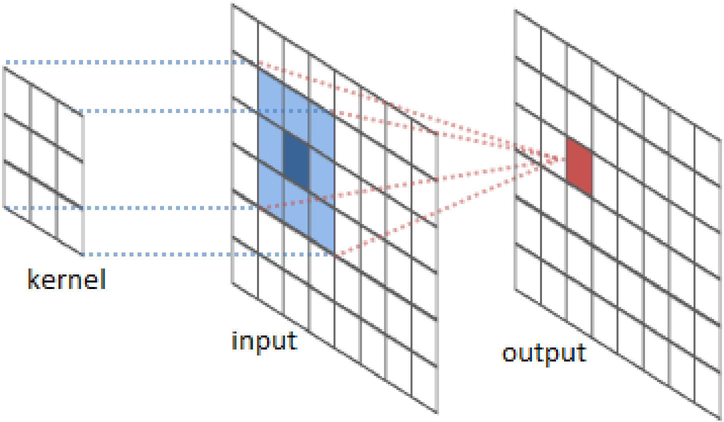
For thresholding, I use *local thresholding* methods.

1. Create a zero matrix with the same size as the contact map (M)
2. Slide a kernel through each pixel
3. If the pixel (i, j) satisfies a particular condition the set $A[i, j]$.

Conditions that I considered:

1. If the pixel is local maximum with respect to the pixels that fall in the kernel
   $$A_{ij} > A_{i',j'} \qquad (i', j') \in \{i-k, i+k\} \times \{j-k, j+k\}$$
2. If the pixel value is larger that some standard deviation from the mean of the values that fall inside the kernel.
   $$A_{ij} > mean(A_{i',j'}) + std(A_{i',j'})$$

Kernels



Extracting graphlets

1. I used the `orca` library in `R` programming language.
2. For each loci, a *signature vector* of length 73 is created.
3. For each chromosome, an $N_c \times 73$, matrix is returned. where N_c is the number of loci in chromosome c.

Finding distance between loci

In order to see how different two loci are in terms of local neighborhood, we need to compare their corresponding *signature values*. The distance measure proposed by Pržulj (2007):

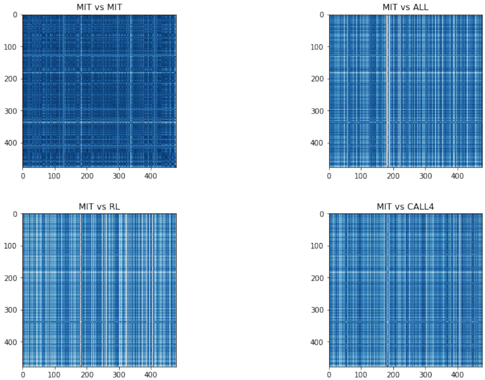
$$d_i = w_i \times \frac{log(S_i + 1) - log(S_i' + 1)}{log(max(S_i, S_i') + 2)} \tag{1}$$

The distance between S and S' can be calculated as follows:

$$D = \sum_{0}^{72} d_i^2 \tag{2}$$

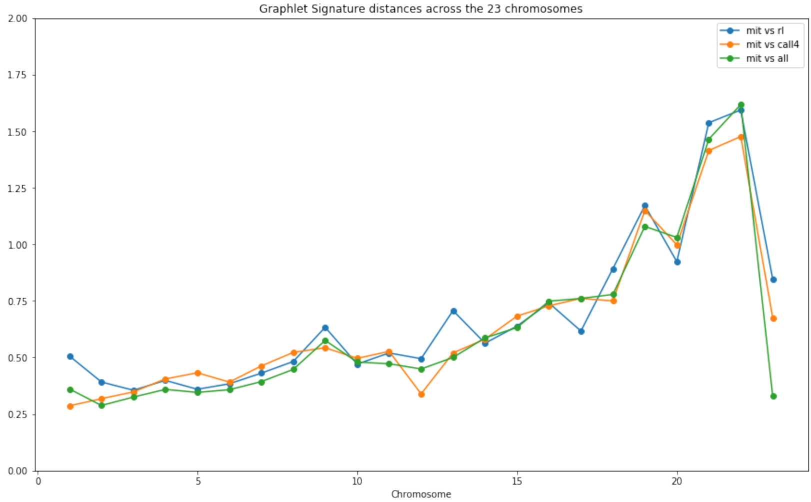
loci-loci distances between cell lines

I calculated loci-loci distance between cell lines using the formula above:

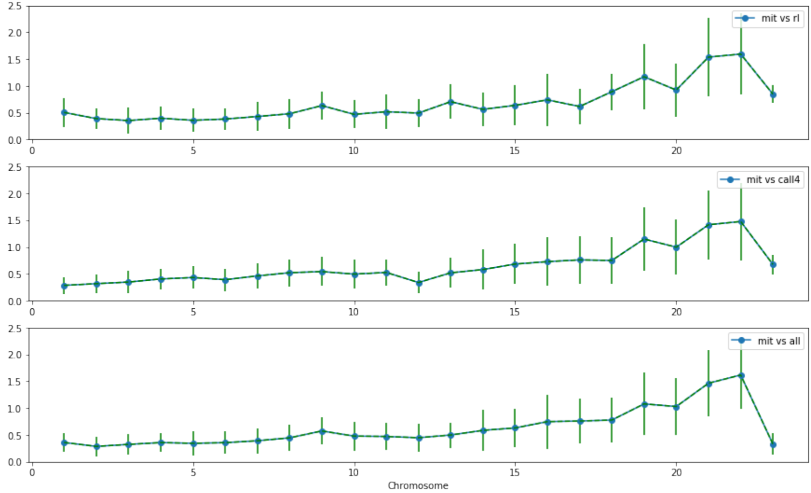


A_{ij} in matrices above denotes distance between loci i in row cell line and loci j in the column cell line.

Loci-loci distances

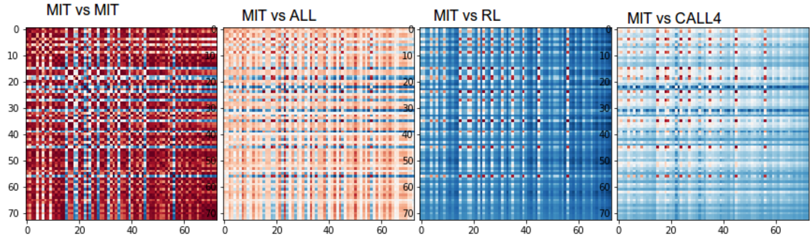


Loci-loci distances



cross-chromosomal signature distance with 1 standard deviation error bars

Orbit correlations



A_{ij} in matrices above denotes correlation between orbit i in row cell line and orbit j in the column cell line.

References I

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . others (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, *326*(5950), 289–293.

Milenkoviæ, T., & Pržulj, N. (2008). Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, *6*, 257.

Milenković, T., Memišević, V., Ganesan, A. K., & Pržulj, N. (2010). 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.

References II

Pržulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics*, *23*(2), e177–e183.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., . . . others (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680.

Wang, Z., Cao, R., Taylor, K., Briley, A., Caldwell, C., & Cheng, J. (2013). The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types. *PloS one*, *8*(3), e58793.