Garaphlet Analysis for HiC data

Behnam Rasoolian

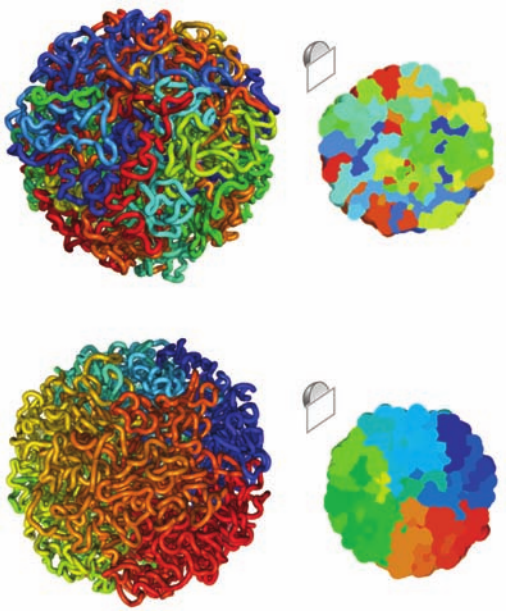
Auburn University

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.

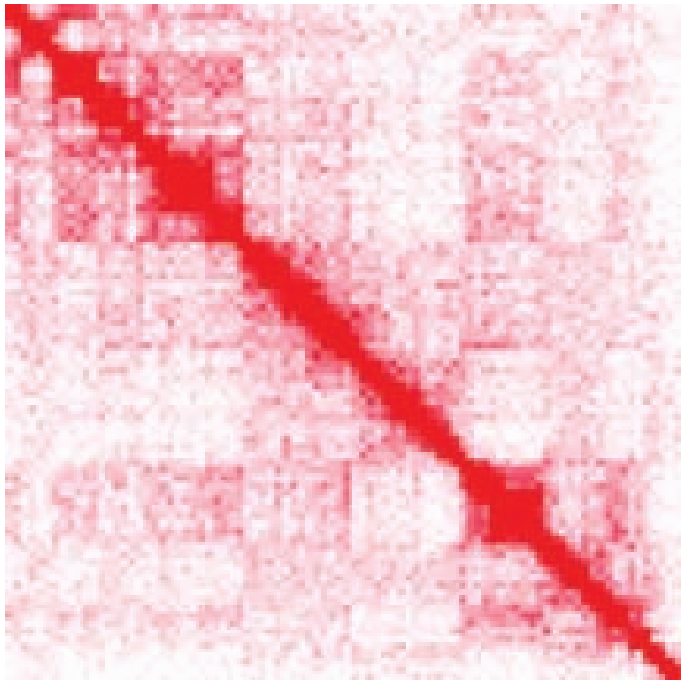


(Lieberman-Aiden et al., 2009)

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.

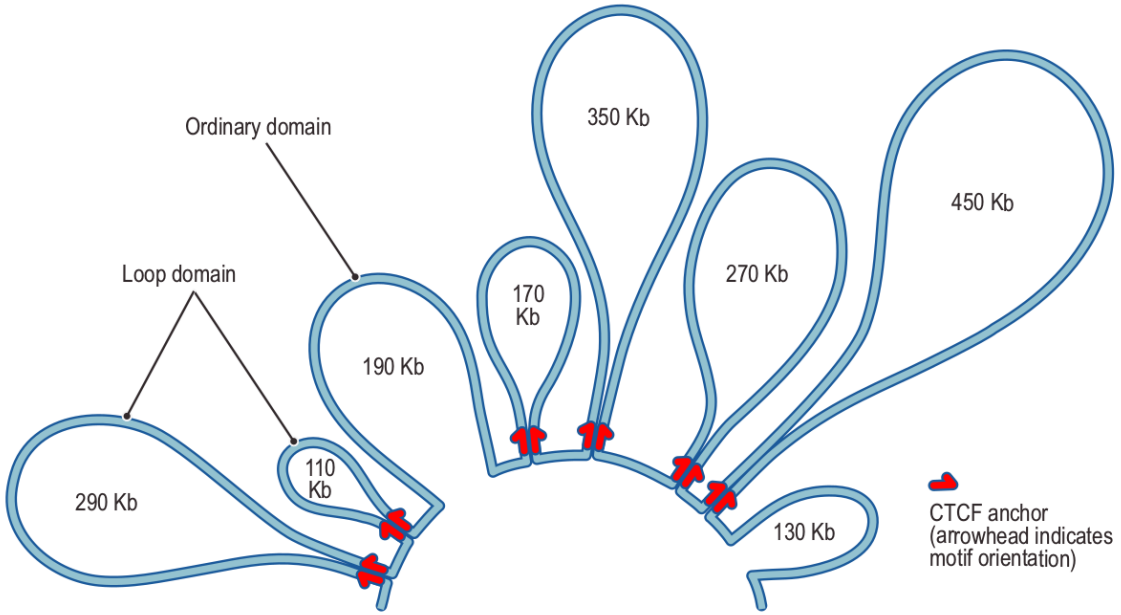


(Lieberman-Aiden et al., 2009)

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)

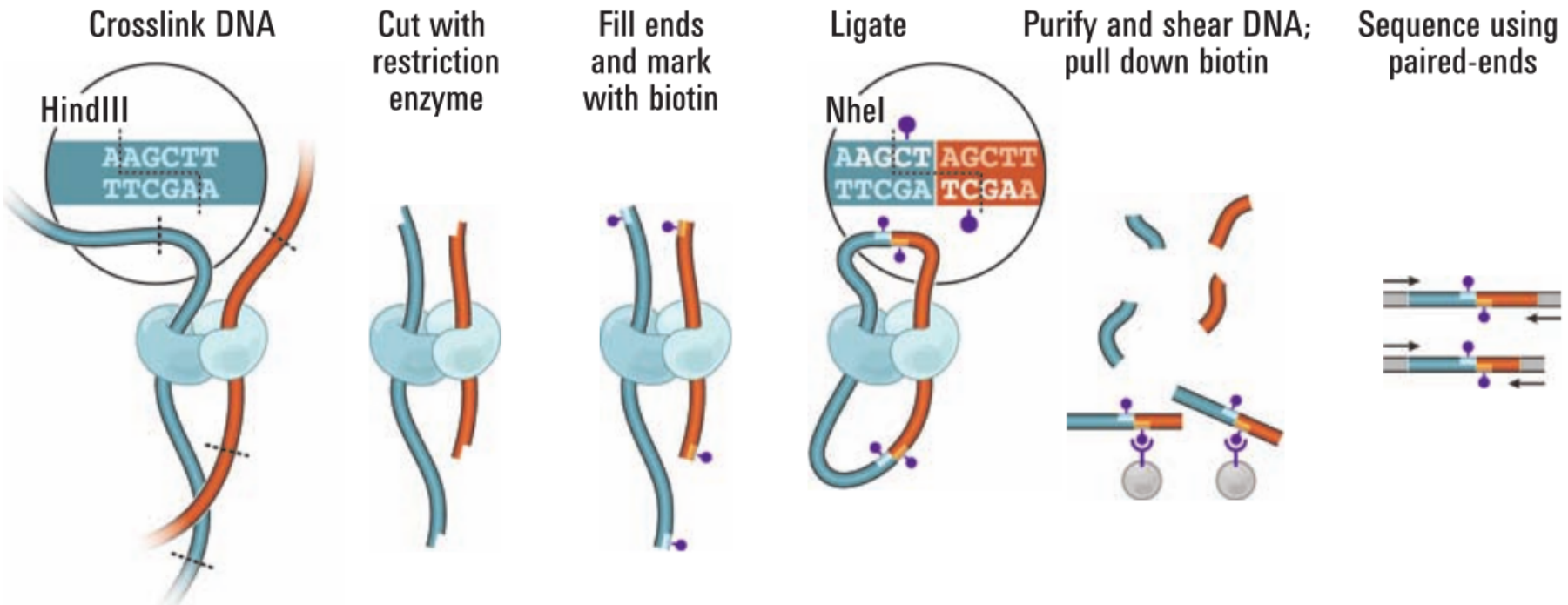
- **Nucleotide**: The monomer units that comprise DNAs. There are 4 types of nucleotides: (C, G, A, and T)
- **Base**: Each pair of nucleotides in the DNA are called a base.
  A kilo-base resolution is a resolution that corresponds to 1000 pairs of nucleotides in DNA.
- **Nucleosome**: A basic unit consisting 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.

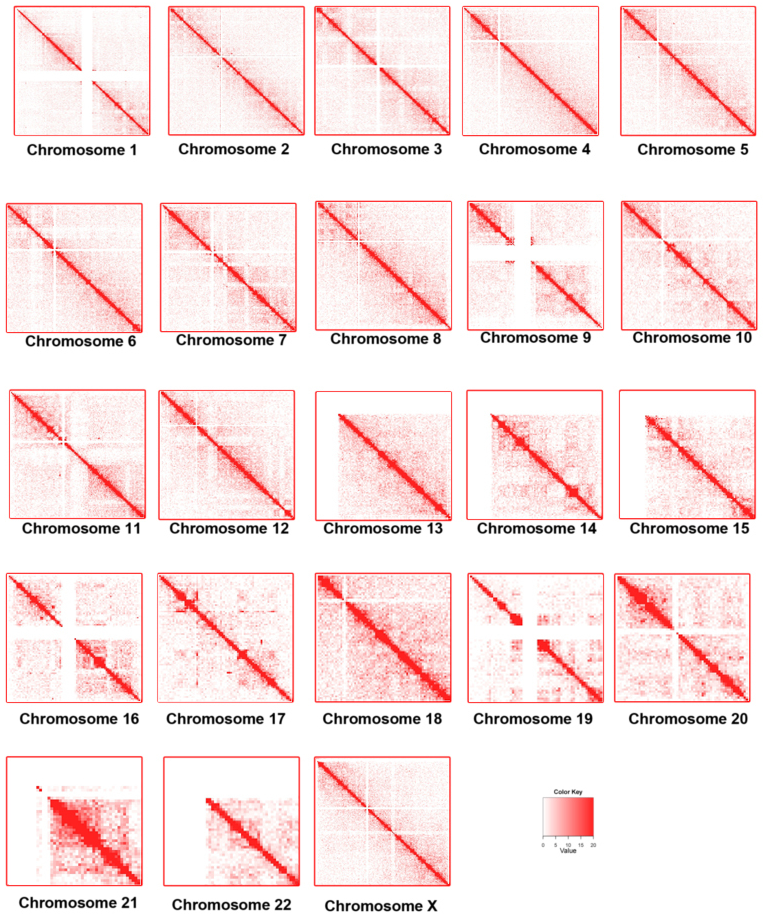
HiC Method
Illustration



(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.



(Wang et al., 2013)

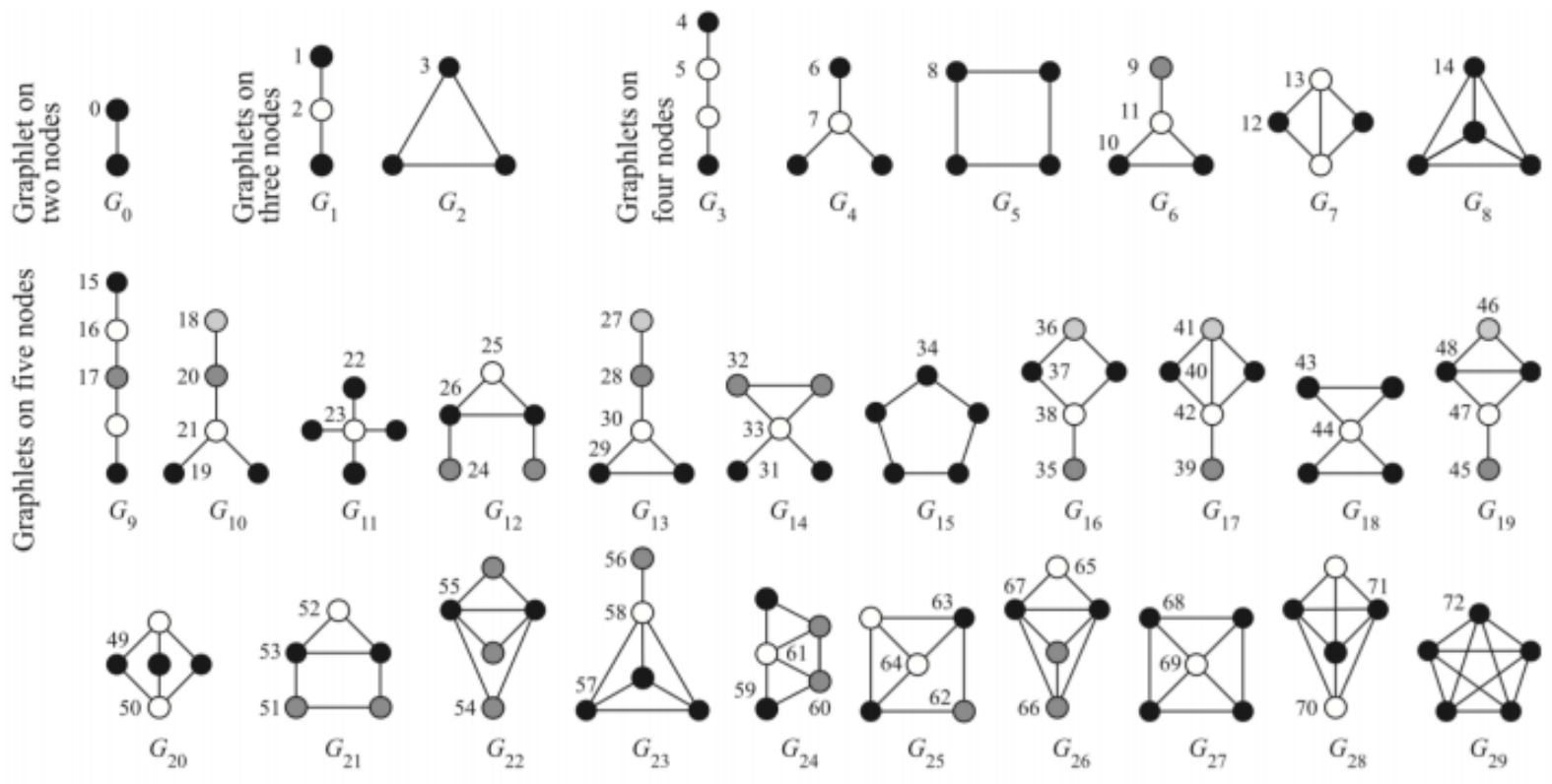
Strategies

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

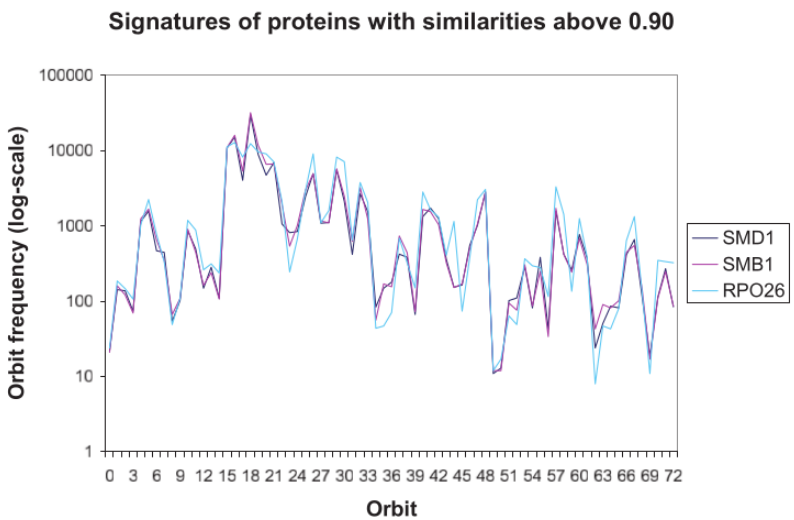
All 30 undirected two- to five-node graphlets with 73 orbits (Pržulj, 2007)

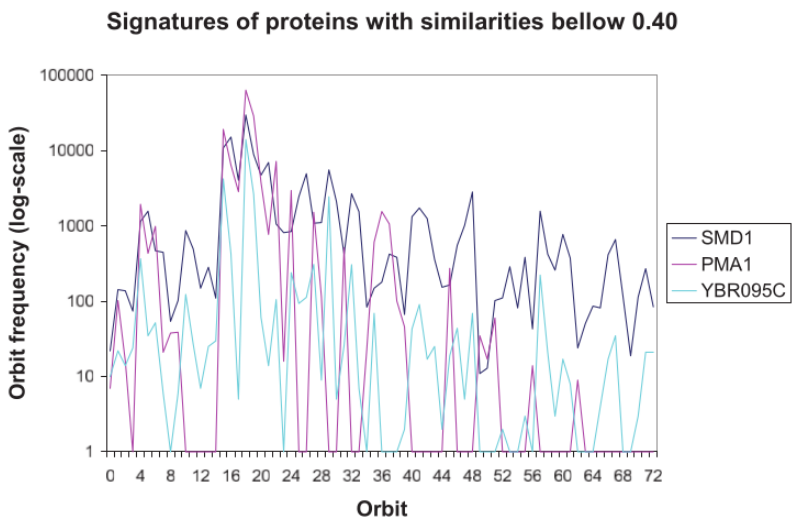
Graphlets
Applications

Milenkoviæ and Pržulj (2008):

Signature vector: A 73-dimensional vector $\mathbf{s}^T = [s_0, s_2, ..., s_{72}]$ 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

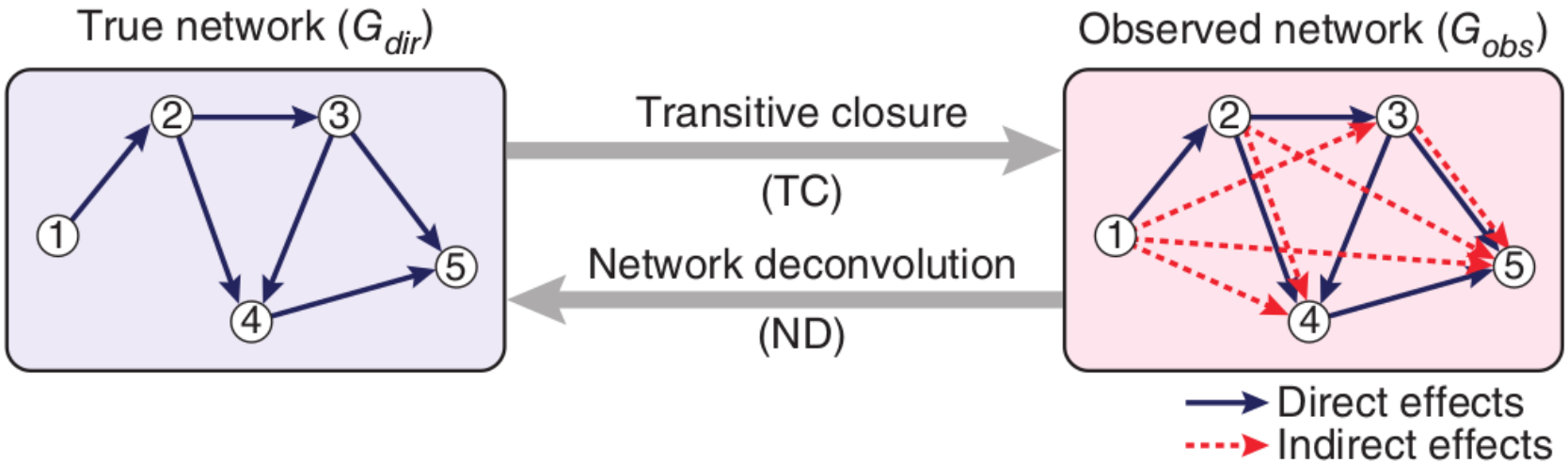
Graphlets
Challenges

1. HiC contact maps are noisy. How do we de-noise them?
2. Current applications of graphlets were on unweighted graphs, while HiC contact maps are weighted.

Balanced Network Deconvolution

Introduction

Proposed by Feizi, Marbach, Médard, and Kellis (2013), Balanced Network Deconvolution, is a method that can be used to remove *indirect effects* from a graph.



(Feizi et al., 2013)

Balanced Network Deconvolution

Details

They assume that:

$$G_{obs} = G_{dir} + G_{dir}^2 + G_{dir} + ... \tag{1}$$

They assume also that both G_{obs} and G_{dir} can be eigen-decomposed and they have the same eigen-vectors:

$$G_{dir} = X\Sigma_{dir}X^T \tag{2}$$

$$G_{obs} = X\Sigma_{obs}X^T \tag{3}$$

$$G_{obs} = X\Sigma_{obs}X^T = X(\Sigma_{dir} + \Sigma_{dir}^2 + ...)X^T \tag{4}$$

Balanced Network Deconvolution

Details

They also assume that eigen-values of the direct network are all between -1 and 1, i.e.

$$-1 < \lambda_i^{dir} < 1 \quad \forall 1 \leq i \leq n \tag{5}$$

$$\Sigma_{obs} = \Sigma_{dir} + \Sigma_{dir}^2 + \ldots \tag{6}$$

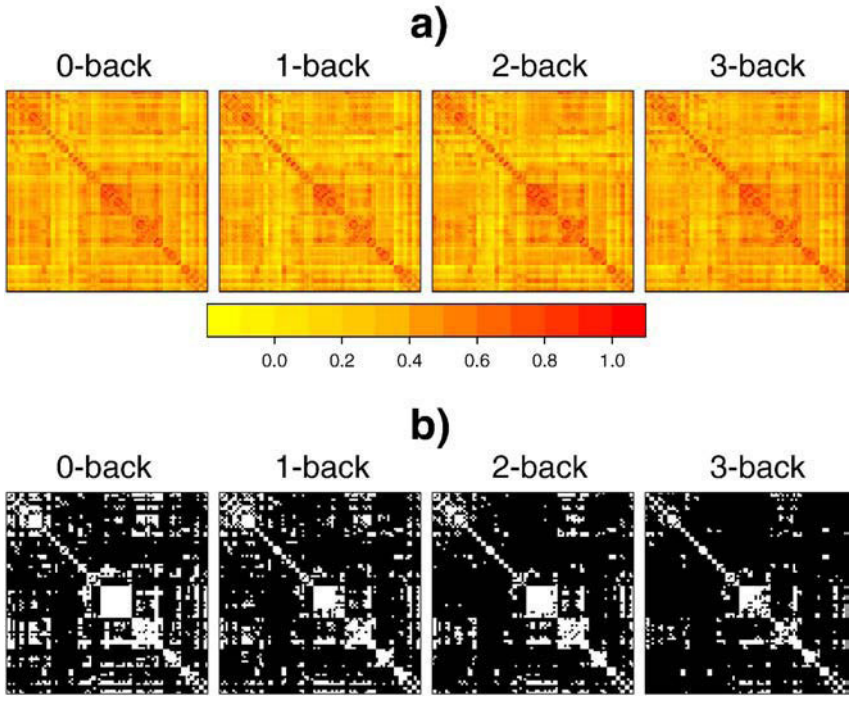
$$\lambda_i^{obs} = \sum_{j=1}^{\infty} \lambda_{ij}^{dir} \qquad \forall i = 1 \ldots n \tag{7}$$

$$\lambda_i^{obs} = \frac{\lambda_i^{dir}}{1 - \lambda_i^{dir}} \tag{8}$$

$$\lambda_i^{dir} = \frac{\lambda_i^{obs}}{1 + \lambda_i^{obs}} \tag{9}$$

Statistical Parametric Network (SPN)

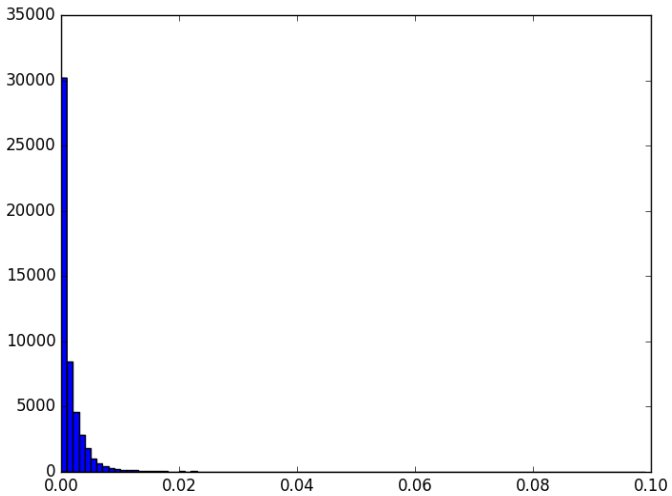
Developed by Ginestet and Simmons (2011).



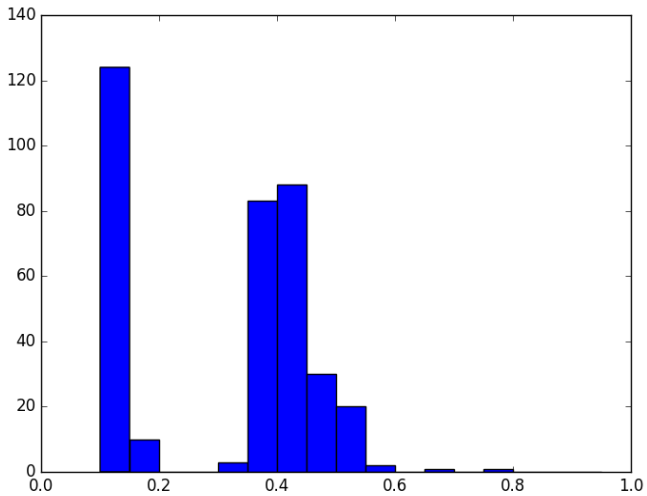
An study of neuron connectivity networks among 43 subjects for 4 different memory tasks (0-back, 1-back, 2-back and 3-back), and the resulting thresholded networks (Ginestet & Simmons, 2011)

Results Thus Far

Histogram of frequencies in HiC



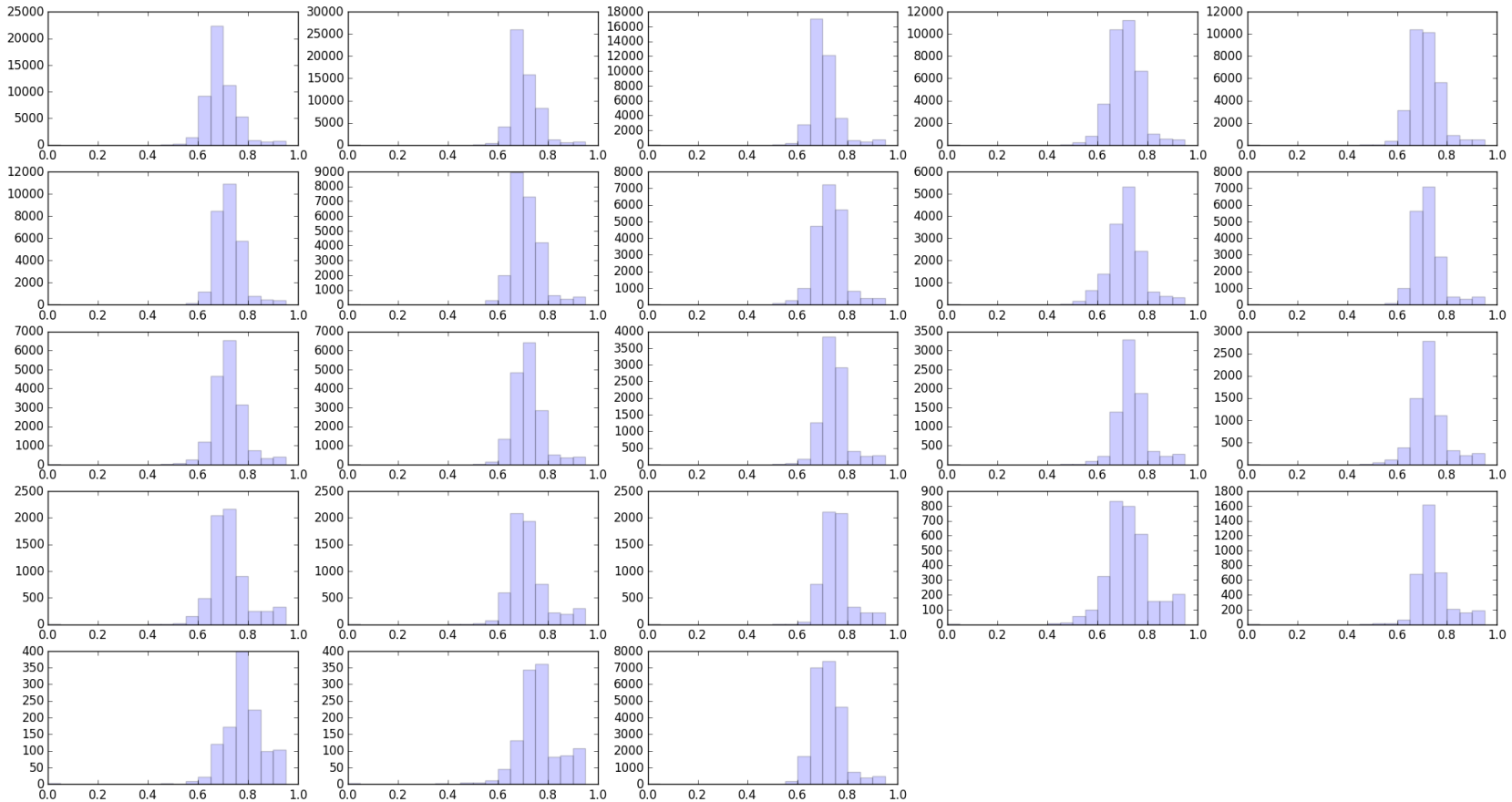
(a) Interval $[0, 0.1]$



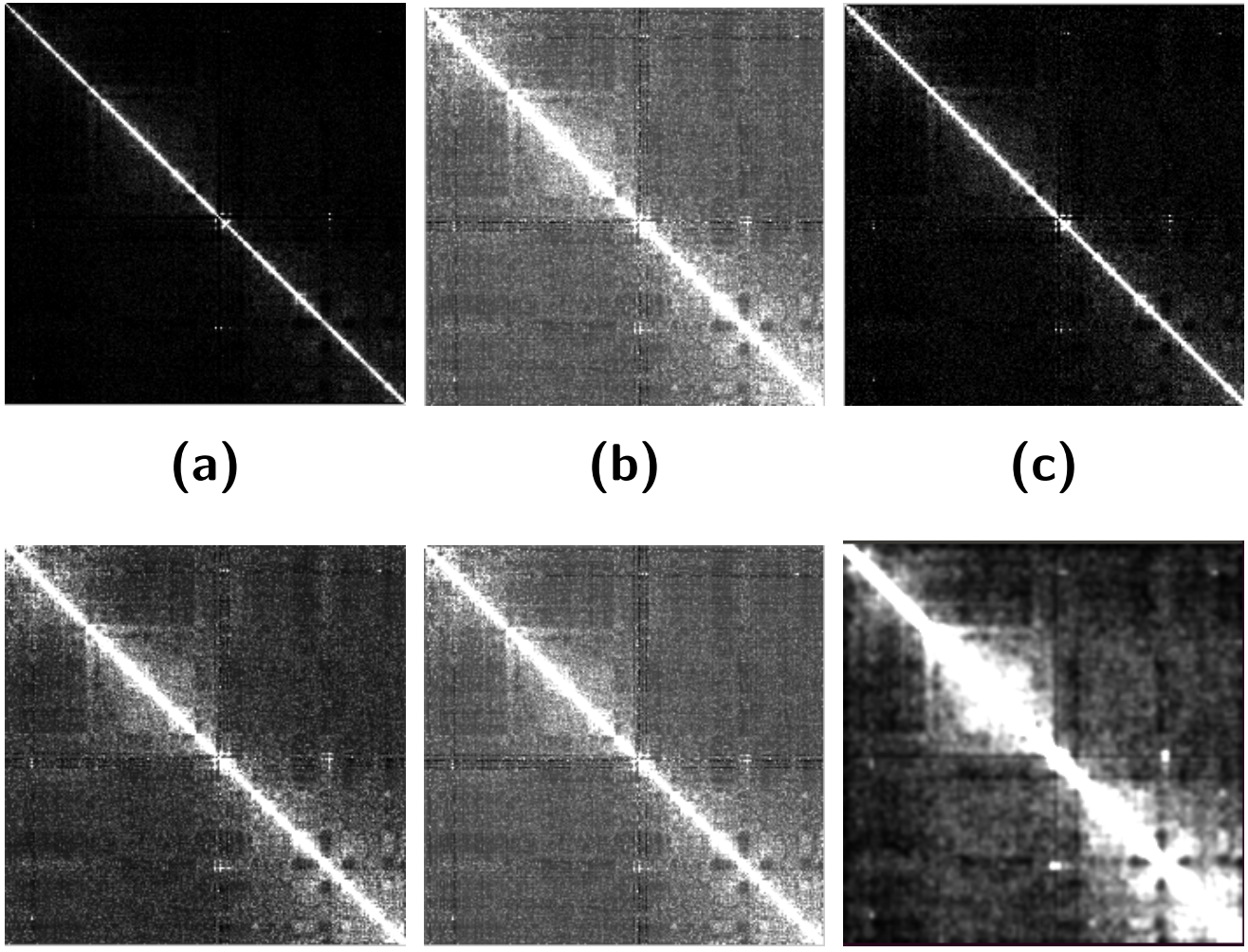
(b) Interval $[0.1, 1]$

Figure: 1a is a histogram of intensities from chromosome 1 As can be seen, the two histograms are orders of magnitude different in terms of frequency.

Histogram of logarithm



Different Normalization methods



(a) (b) (c)

(d) (e) (f)

References I

Feizi, S., Marbach, D., Médard, M., & Kellis, M. (2013). Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology*, *31*(8), 726–733.

Ginestet, C. E., & Simmons, A. (2011). Statistical parametric network analysis of functional connectivity dynamics during a working memory task. *Neuroimage*, *55*(2), 688–704.

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.

References II

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.

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.

References III

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.