Cell classification by applying graphlets on HiC data

Behnam Rasoolian Liangliang Xu Zheng Zhang

1 Motivation

In this study, we plan to find dissimilarities between normal cells and cancerous cells, in terms of the 3D structure of their chromosomes. This is done through investigating HiC contact maps. We suspect that there are systematic differences between how chromosomes are structured between normal cells and cancerous cells.

2 Background

The cell of a eukaryotic species forms a multi-granularity genome structure in order to compactly store a very long genomic DNA sequence in its small nucleus. The following describes the genome structure in the order of decreasing granularity (1) A **nucelotide** is the building block of DNA. There are 4 types of nucleotides: C, G, A and T. (2) Each pair of nucleotides in the DNA are called a **base**. A kilo-base is a group of 1000 bases. (3) thousands of bases join together to form **gene loci**. (4) A number of loci then fold into a large independent physical structure called **chromosome**.
One or more chromosomes interact to constitute the dynamic **three-dimensional (3D) conformation** of the entire genome of a cell.
Ideally, it is desirable to compare these 3D conformation 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 but there has been efforts at its characterization: In other to find dissimilarities in the 3D structure of chromosomes, we use HiC dataset. The HiC method 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.*
Graphlet comparison is a novel method used to compare large networks in order to find local similarities in them. Given a graph G, **fragment** are connected subgraphs G. **Motifs** are fragments that occur with a frequency much higher than that occuring in a randomly generated graph. Given a graph $G(V, E)$ and $S \subseteq V$, then $G'(S, E')$ is an **induced graphs** iff $E' = \{(u, v)|u, v \in V$ and $(u, v) \in E \rightarrow (u, v) \in E'\}$. **Graphlets** are arbitrary, induced fragments. An edge is the only two-node graphlet. **Orbits** are sets of all nodes in a graphlet that can be swapped with each other while not changing the graph. Orbits are "topographically similar" to each other.
A **signature** or **signature vector** of a node in graph G is 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 in G.

3 Data and Methods

We have HiC itra- and inter-chromosomal interaction matices for 4 types of cells, one of which is normal and the other three are cancer cells. We will extract signature vectors of size 73 for each loci in each chromosome. leading to a 23×73 matrix for each loci. Assuming there are n_i loci in each chromosome (i = 1 ... 23), it would lead to an $N \times P \times M$ matrix where N is $\sum_{i=1}^{23} n_i$ and P and M are 23 and 73 respectively. By estimating n_i to be 100 on average, we expect the size of the input (N) to be 2000. We will use the normal cell data and two of the cancer cell data in training phase, and used the fourth cell type to validate the results. Thus we have 3 categories: normal, cancer type I, and cancer type II. The problem is given a $23 * 73$ contact graphlet matrix, does it belong to a normal cell or cancerous cell. In terms of methods, we plan to start with standard VGG-Network. We will have to go towards more complicated networks in the case of poor results. We will be open to exploring various architectures and pick one that fits the data better.