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 (6) 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.

(7) 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 (8), 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.

Implementations of algorithms of extracting graphlets:

- GraphCrunch: `http://www0.cs.ucl.ac.uk/staff/natasa/graphcrunch2/usage.html`
- PGD: `http://nesreenahmed.com/graphlets/`
- ORCA: Graphlet and orbit counting algorithm
  `https://CRAN.R-project.org/package=orca`
  This package is in R. In order to install it, type `install.packages("orca")`.

The authors of (9) generalized the idea of graphlets to ordered graphs were the nodes are labeled in ascending order. These graphlets are illustrated in Figure **??**. 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

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 the nodes represent.

Thresholding is achieved by thresholding the contact maps. The result of the thresholding procedure would be 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.

In order to go about the process of thresholding, 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 proved result 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 and some very larges values that correspond t interactions along and 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 and picking small numbers will lead to capturing too many *insignificant* interactions.

In order to threshold the matrix so that both global and local patterns are kept, 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 that is larger than the average of non-zero intensities in its *neighborhood*.

RESOURCES

Publications related to Hi-C:

1. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2858594/`
2. `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0058793`
3. `http://nar.oxfordjournals.org/content/42/7/e52.full`
4. `http://bioinformatics.oxfordjournals.org/content/`

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

early/2015/12/31/bioinformatics.btv754.abstract?keytype=ref&ijkey=A97WhKqBiEIcuzd

5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4417147/

6. http://www.pnas.org/content/112/47/E6456.full

7. http://www.pnas.org/content/113/12/E1663.full

Hi-C Datasets:

1. Original Datasets: https://bcm.app.box.com/v/aidenlab/folder/11234760671

2. Including cancerous cells: http://sysbio.rnet.missouri.edu/T0510/tmp_download/link_to_download_genome_data/

3. Chromosome3D project: http://sysbio.rnet.missouri.edu/bdm_download/chromosome3d/

Contact Matrix Analysis:

1. https://omictools.com/contact-matrix-normalization-category

2. http://hifive.docs.taylorlab.org/en/latest/

Labs working on 3D Human Genome:

1. http://mirnylab.mit.edu

2. http://dostielab.biochem.mcgill.ca

3. http://www.aidenlab.org/

4. http://web.cmb.usc.edu/people/alber/index.htm

5. http://calla.rnet.missouri.edu/cheng/nsf_career.html

Resources related to Graphlet:

1. https://en.wikipedia.org/wiki/Graphlets

2. https://academic.oup.com/bioinformatics/article/23/2/e177/202080/Biological-network-comparison-using-graphlet

3. http://www0.cs.ucl.ac.uk/staff/N.Przulj/index.html

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. Zheng Wang, Renzhi Cao, Kristen Taylor, Aaron Briley, Charles Caldwell, and Jianlin Cheng. The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types. *PloS one*, 8(3):e58793, 2013.
4. J. Gross. Automorphisms. http://www.cs.columbia.edu/ cs6204/files/Lec5-Automorphisms.pdf, April 2010.
5. Tuan Trieu and Jianlin Cheng. Mogen: a tool for reconstructing 3d models of genomes from chromosomal conformation capturing data. *Bioinformatics*, 32(9):1286–1292, 2015.
6. Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.
7. Tijana Milenkoviæ and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6:257, 2008.
8. 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.
9. 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.
10. Soheil Feizi, Daniel Marbach, Muriel Médard, and Manolis Kellis. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology*, 31(8):726–733, 2013.
11. Hai-Ping Sun, Yan Huang, Xiao-Fan Wang, Yang Zhang, and Hong-Bin Shen. Improving accuracy of protein contact prediction using balanced network deconvolution. *Proteins: Structure, Function, and Bioinformatics*, 83(3):485–496, 2015.
12. Hossein Azari Soufiani and Edo Airoldi. Graphlet decomposition of a weighted network. In *Artificial Intelligence and Statistics*, pages 54–63, 2012.
13. Cedric E Ginestet and Andrew Simmons. Statistical parametric network analysis of functional connectivity dynamics during a working memory task. *Neuroimage*, 55(2):688–704, 2011.
14. Hamed Daneshpajouh, Hamid Reza Daneshpajouh, and Farzad Didehvar. A metric on the space of weighted graphs. *arXiv preprint arXiv:0906.2558*, 2009.
15. Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In *Artificial Intelligence and Statistics*, pages 488–495, 2009.
16. Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC genomics*, 13(1):436, 2012.