NeoHiC: a web application for the analysis of Hi-C data

Daniele D'Agostino(1), Ivan Merelli(2), Marco Aldinucci(3), Pietro Liò(4)

(1) Institute for Applied Mathematics and Information Technologies "E. Magenes", National Research Council of Italy, Genoa, Italy, dagostino@ge.imati.cnr.it

(2) Institute for Biomedical Technologies, National Research Council of Italy, Segrate (MI), Italy, ivan.merelli@itb.cnr.it

(3) Computer Science Department, University of Torino, Italy, aldinuc@di.unito.it

(4)Computer Laboratory, University of Cambridge, UK, Pietro.Lio@cl.cam.ac.uk

Keywords: Hi-C, graph databases, web application, graph visualization.

Abstract. High-throughput sequencing Chromosome Conformation Capture (Hi-C) allows the study of chromatin interactions and 3D chromosome folding on a larger scale. A graph-based representation of Hi-C data is very important for a proper visualization of the spatial pattern they represent, in particular for comparing different experiments or for re-mapping omics-data in a space-aware context. But the size of these graphs can be very large, and this prevents the straightforward use of current available graph visualization tools and libraries. In this paper, we present the first version of NeoHIC, a web application for the progressive graph visualization of Hi-C data based on the use of the Neo4j graph database.

1 Scientific Background

The exploration of the 3D organization of chromosomes in the nucleus of cells is of paramount importance for many cellular processes related to gene expression regulation, including DNA accessibility, epigenetic patterns and chromosome translocations. In particular, High-throughput sequencing Chromosome Conformation Capture (Hi-C) allows the study of chromatin interactions and 3D chromosome folding on a larger scale [1]. The graph-based representation of Hi-C data produced, for example, by NuChart [2, 3] or CytoHic [4], which are software for representing the spatial position of genes in the nucleus, will be very important for creating maps where further omics data can be mapped, in order to characterize different spatially associated domains. This visualization is an effective complement of the traditional matrix-based representations, e.g. as produced by Juicer [1] or TADbit [2].

Moreover, using a gene-centric point of view to provide a map on which other omic data can be mapped. Contact matrices, or better their probabilistic models, allow to create representations that only involve two chromosomes, while graphs are able to describe the interactions of all the chromosomes together using a graph-based approach. This representation gives more importance to the physical proximity of genes in the nucleus in comparison to coordinate-based representations. This is the same problem that impairs representations based on Circos [3], which are able to characterize the whole genome in one shot, but fail to describe the physical proximity of genes.

In particular, we showed some interesting results relying on the possibility of creating metrics for defining how far two genes are one from the other, with possible applications

[1] https://github.com/aidenlab/juicer
[2] https://github.com/3DGenomes/TADbit
[3] https://circos.ca/

to cytogenetic profiling, to the analysis of the DNA conformation in the proximity of the nucleolus, and for describing the social behavior of genes [5, 6].

However, the typical size of a graph achieved through a Hi-C analysis is in the order of thousands of nodes and hundred of thousands edges, which makes difficult an effective exploration by users, at least using the tools actually available. We tested both esyN [7], a tool for the construction and analysis of networks for biological research, and Cytoscape [4] with a network composed by about 2,400 nodes and 175,000 edges and we found many difficulties in visualizing and analyzing such a huge network with these tools.

Furthermore, the adoption of Graph databases for storing and managing Hi-C data has been considered only in these last years, e.g. [8], while the most important repositories as STRING [9] or InterMine [10] are still based on relational databases.

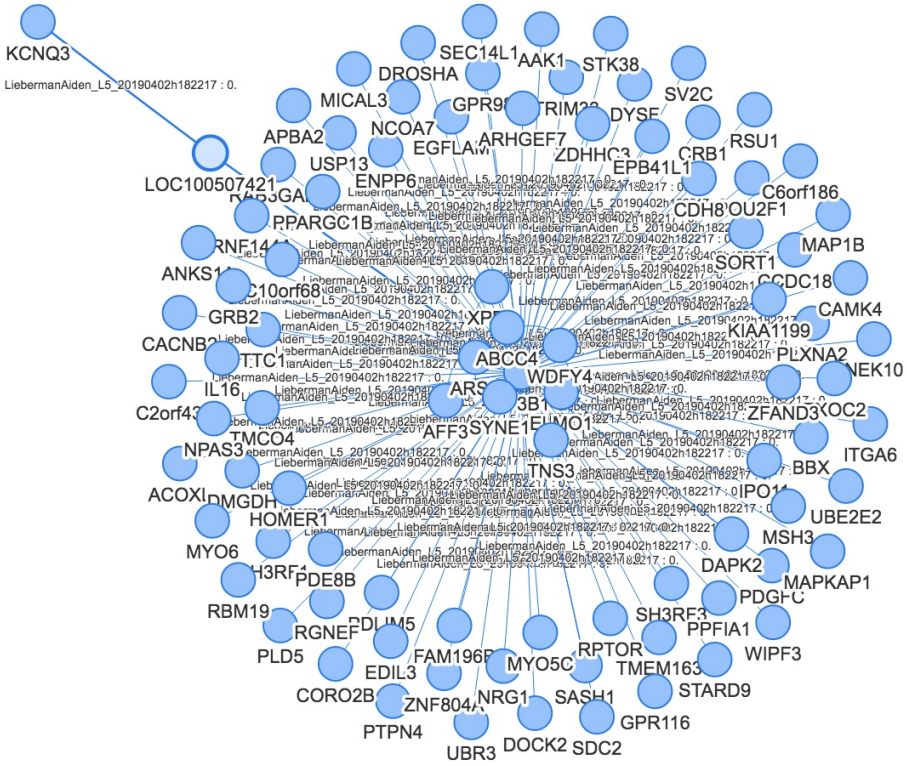

Figure 1: The initial visualization of the same network after selecting starting gene using NeoHiC.

2 Materials and Methods

We developed a first version of NeoHiC, a web application relying on the Neo4j graph databases and related modern web technologies, such as the Node.js JavaScript framework and the Neovis.js[5] visualization library, in order to manage and analyze graphs produced by investigating Hi-C data.

Graph databases are part of the NoSQL databases created to address the limitations of the existing relational databases. While the graph model explicitly lays out the dependencies between nodes of data, the relational model and other NoSQL database models link the data by implicit connections.

In relational databases, references to other rows and tables are indicated by referring to primary key attributes via foreign key columns. Joins are computed at query time

[4]https://cytoscape.org/

[5]https://github.com/neo4j-contrib/neovis.js/

by matching primary and foreign keys of all rows in the connected tables. These operations are compute-heavy and memory-intensive and have an exponential cost. Moreover, When many-to-many relationships occur in the model, you must introduce a JOIN table (or associative entity table) that holds foreign keys of both the participating tables, further increasing join operation costs.

Unlike other database management systems, relationships are of equal importance in the graph data model to the data itself. This means we are not required to infer connections between entities using special properties such as foreign keys, This is the reason why graph databases, by design, allow simple and fast retrieval of complex hierarchical structures that are difficult to model in relational systems.

3 Results

NeoHiC is based on the same approach adopted by STRING, where a protein-protein interaction network is expanded one step at a time by clicking on one of the visible nodes. Examples of visualization are provided in Fig. 1 and Fig. 2.

The present version of the Web application allows users to select a starting gene and inspecting, step by step, the network described by the Hi-C data. It is also possible to filter the edges on the basis of their value. We are working for providing the possibility to compare two or more different Hi-C experiments, in order to statistically highlight differences in the chromatin conformation of different cells.

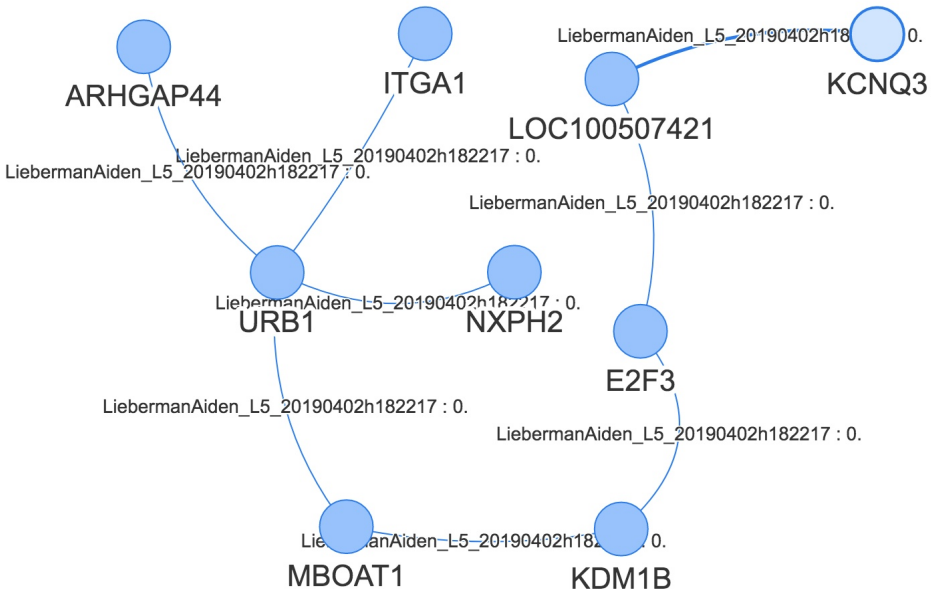

Figure 2: The visualization of a path in the network after five steps.

4 Conclusion

The NeoHiC web application represents a first step in the development of an extensible science gateway for the sharing and analysis of Hi-C data. To this extent a further development direction is represented by the integration of 1D and 2D information on the Hi-C graphs in order to correlate the 3D conformation of the genome with regulatory and expression patterns.

Acknowledgments

This work has been funded by the Short-term 2018 Mobility Program (STM) of the National Research Council of Italy (CNR).

References

[1] Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., ... & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289-293. doi:https://doi.org/10.1126/science.1181369. PubMed: 19815776.

[2] Merelli, I., Lio', P. & Milanesi, L. (2013). NuChart: an R package to study gene spatial neighbourhoods with multi-omics annotations. PLoS One, 8(9), e75146.

[3] Tordini, F., Drocco, M., Misale, C., Milanesi, L., Lio', P., Merelli, I., Torquati, M. & Aldinucci, M. (2017). NuChart-II: the road to a fast and scalable tool for Hi-C data analysis, International Journal of High Performance Computing Applications", vol. 31, iss. 3, pp. 196-211.

[4] Shavit, Y. & Lio', P. (2013). CytoHiC: a cytoscape plugin for visual comparison of Hi-C networks. Bioinformatics, 29(9), 1206-1207.

[5] Merelli, I., Tordini, F., Drocco, M., Aldinucci, M., Lio', P. & Milanesi, L. (2015). Integrating multi-omic features exploiting Chromosome Conformation Capture data, Front. Genet., 6:40.

[6] Tordini, F., Aldinucci, M., Milanesi, L, Lio', P. & Merelli, I. (2016). The genome conformation as an integrator of multi-omic data: the example of damage spreading in cancer, Front. Genet., 7:194.

[7] Bean, D.M., Heimbach, J., Ficorella, L., Micklem, G., Oliver, S.G. & Favrin, G. (2014) esyN: Network Building, Sharing and Publishing. PLoS ONE 9(9): e106035.

[8] Have, C. T. & Jensen, L. J. (2013). Are graph databases ready for bioinformatics? Bioinformatics, 29(24), 3107.

[9] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., ... & Kuhn, M. (2014). STRING v10: proteinprotein interaction networks, integrated over the tree of life. Nucleic acids research, 43(D1), D447-D452.

[10] Smith, R. N., Aleksic, J., Butano, D., Carr, A., Contrino, S., Hu, F., ... & Stepan, R. (2012). InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics, 28(23), 3163-3165.