Tuck *BMC Research Notes* (2022) 15:45
https://doi.org/10.1186/s13104-022-05912-9

BMC Research Notes

DATA NOTE

Open Access

A cancer graph: a lung cancer property graph database in Neo4j

David Tuck[*]

Abstract

Objectives: A novel graph data model of non-small cell lung cancer clinical and genomic data has been constructed with two aims: (1) provide a suitable model for facilitating graph analytics within the Neo4j framework or through tools which can interact through existing Neo4j APIs; and (2) provide a base model extensible to other cancer types and additional datasets such as those derived from electronic health records and other real world sources.

Data description: Clinical and genomic data integrated with a novel property graph database schema from publicly available datasets and analyses based on The Cancer Genome Atlas lung cancer datasets augmented by with subgraphs patient-patient social network from similarity and correlation as well as individual based biological networks.

Keywords: Non-small cell lung cancer, Property graph database, The Cancer Genome Atlas

Objective

The pathobiology of cancer involves the coordinated dysregulation of multiple processes across molecular, cellular, tissue, and organism scales [1]. Somatic mutations and genomic aberrations are crossed and intertwined with an individual patient's clinical, social, and medical histories. The complex interrelationships among all of these factors determine disease origin, trajectory, and outcomes of interventions [2]. Strategies that allow operation directly on the topology of the graph structures defined by these relationships are enabled by the development and growing maturity of native graph databases such as TigerGraph and Neo4j [3–5].

This note describes a representation of non-small cell lung cancer in Neo4j, a property graph database platform which natively stores and processes graph data models. A version is available in a GPL3-licensed open-source community edition [6].

Non-small cell lung cancer is the most common cause of cancer deaths worldwide [7, 8], and it has plentiful publicly available genomic, clinical, and molecular data [9–11]. The lung cancer graph database provides an analytic framework for integrating modeling of disease mechanisms on a genome-scale and clinical data from clinical electronic health records, diagnostic studies, therapeutic interventions, and molecular assays. This project utilizes several publicly available open data sources and extends these with calculated variables defining relationships to create a novel graph schema and nested set of subgraphs comprising the Neo4j database. Clinical, demographic, diagnostic, therapeutic, and multiple genomic measures are obtained from TCGA LUAD and LUSC datasets [9, 10]. Multiple analyses have extended available attributes for immunologic, and biologic signaling pathway profiles [11–15] enabling the creation of a graph structure at different scales based on relations among cancer cases, relations among biological molecules, relations of biological networks and processes within individual patients.

By adopting graph database technology, this data resource aims to provide a platform to explore the utility of integrative graph-based systems biology analyses to decode the molecular and clinical underpinnings of complex diseases.

*Correspondence: david.tuck@va.gov; dptuck@gmail.com
VA Boston Healthcare System, 150 South Huntington Avenue, Boston, MA 02130, USA

Data description

All data files and datasets are deposited in the Harvard Dataverse repository in dataset "A Cancer Graph: A Lung Cancer property graph using Neo4j" [16]. A file containing the entire graph database is provided as a binary database dump (Data file 1 in Table 1). The schema for the property graph is described in Dataset 2 which contains a graphic image of the schema and a json file containing the schema with all entities, attributes, and relationships among all the different entities. Data file 3 contains the commands for generating documentation of the schema, and indexes, for loading the binary file into a Neo4j instance. Data file 3 also provides example commands in the cypher language used by Neo4j which describes how the database was originally generated from input files. These individual input data source files are provided (Dataset 4 in Table 1) as comma separated value formatted files.

The property graph database consists of (a) publicly available open access data of patients with non-small cell lung cancer and (b) derived variables augmented by relationships defining different subgraphs. The database contains data from > 1000 patients from the Cancer Genome Atlas (TCGA) which contain clinical, diagnostic, and therapeutic data (chemotherapy, radiation, immunotherapy), as well as multiple genomic measures (gene expression, somatic, mutations, copy number, epigenetics). Additional attributes are derived from independent published analyses based on these data, providing signatures related to immunologic, DNA repair, molecular portrait subtypes, and profiles from a variety of biological pathways [11–15]. The dataset also incorporates relevant portions of precedent native graph representations of biological and biomedical systems including Hetio [17, 18] and Reactome [19], both of which use Neo4j platform to represent complex biological networks. This existing framework is supplemented by pathway, genomic and various calculated variables including graph kernels, embedded vector representation of somatic gene mutations, and computed pathway activations.

The primary value of the dataset come from calculated relationships which create subgraphs that serve as a substrate for the application of exploration and application of graph algorithms [20–22]. These occur primarily at two different scales: (1) patient-patient network with direct relationships among patients (or tumor samples) based on similarity scores or correlation for genomic features or signatures; (2) biological networks within single patient samples.

- CancerCase (Patient-based) networks provide graphs of the relationships between patients based on calculation of similarity and correlation scores of molecular signatures such as immune scores or DNA repair profiles.
- Intra-patient biological signaling activation networks InFlo [14] is a robust systems biology approach for integrative analysis of multi-omics data which can characterize complex biological signaling network activities in any given biological sample. InFlo was applied for individual samples from TCGA including the non-small cell lung cancer samples/Thus calculating a complete biological network activation state for each individual tumor sample.

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	lung-cancer-graph-neo4j-2021-07-15T022024.bin	binary dump file (bin)	Harvard Dataverse [16] https://doi.org/10.7910/DVN/RIXLG8
Data file 2	README.md	text/markdown	Harvard Dataverse [16] https://doi.org/10.7910/DVN/RIXLG8
Data file 3	ACancerGraphSchema.png	image (png)	Harvard Dataverse [16] https://doi.org/10.7910/DVN/RIXLG8
Data file 4	makeIndexesConstraints.cql	cypher (cql)	Harvard Dataverse [16] https://doi.org/10.7910/DVN/RIXLG8
Data file 5	ACancerGraphLoader.cql	cypher (cql)	Harvard Dataverse [16] https://doi.org/10.7910/DVN/RIXLG8
Data file 6	schema.json	json	Harvard Dataverse [16] https://doi.org/10.7910/DVN/RIXLG8
Data set 1	Input files (csv format) to create database	comma separated values (csv)	Harvard Dataverse [16] https://doi.org/10.7910/DVN/RIXLG8
Data file set 2	Input files (csv format) to create supportive data	comma separated values (csv)	Harvard Dataverse [16] https://doi.org/10.7910/DVN/RIXLG8

In summary, a novel graph data model has been constructed integrating clinical and molecular data of non-small cell lung cancer patients with aims: (1) a graph model for facilitating graph analytics within the Neo4j framework or through tools via the Neo4j application programming interface (API); and (2) exploratory basis extension to other tumor types or clinical datasets derived from electronic health records.

Limitations

- The database is limited in the number of variables, which may not satisfy specific needs.
- The schema of the database.
- TCGA is rich in omics but relatively poor in clinical details (comorbidity, frailty assessment, specific lab results, extended pharmacy).

And other sources with additional modifications (TCGA is rich in omics but relatively poor in clinical details (comorbidity, frailty assessment, specific lab results, extended pharmacy, etc.).

Abbreviations

TCGA: The Cancer Genome Atlas; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; API: Application Programming Interface.

Acknowledgements

Hetionet was originally developed by Daniel Himmelstein and collaborators in the Baranzini Lab at the University of California, San Fransisco, more details available at: https://het.io.

Authors' contributions

DPT conceived the study and wrote the manuscript. The author read and approved the final manuscript.

Funding

Funding information is not applicable/No funding was received.

Availability of data and materials

The data described in this Data note can be freely and openly accessed at Harvard Dataverse: https://doi.org/10.7910/DVN/ZU [16]. Please see Table 1 and references [16] for details. Original data sets are available in the Cancer Genome Atlas repository, https://tcga-data.nci.nih.gov/tcga/. Inflo was developed in the Varadan lab at Case Western : https://varadanlab.github.io/InFlo/.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have declared that no competing interests exist.

Received: 27 October 2021 Accepted: 24 January 2022
Published online: 14 February 2022

References

1. Cancer Complexity Knowledge Portal. NIH National Cancer Institute-sponsored Cancer Systems Biology Consortium (CSBC) https://www.cancercomplexity.synapse.org/
2. Hochheiser H, Castine M, Harris D, Savova G, Jacobson RS. An information model for computable cancer phenotypes. BMC Med Inform Decis Making. 2016. https://doi.org/10.1186/s12911-016-0358-4.
3. Timón-Reina S, Rincón M, Martínez-Tomás R. An overview of graph databases and their applications in the biomedical domain. Database. 2021. https://doi.org/10.1093/database/baab026·.
4. TigerGraph: Graph Database | Graph Analytics Platform; https://www.tigergraph.com. Accessed 29 Dec 2021.
5. Neo4j Graph Platform – The Leader in Graph Databases Neo4j Graph Database Platform; https://neo4j.com. Accessed 29 Dec 2021.
6. Neo4j Graph Database Platform: Download Neo4j; https://neo4j.com/download. Accessed 29 Dec 2021.
7. Cancer.Net: lung cancer—non-small cell—Statistics Cancer.Net; https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/statistics. Accessed 29 Dec 2021.
8. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. Cancer J Clin. 2021. https://doi.org/10.3322/caac.21654.
9. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012. https://doi.org/10.1038/nature11404·.
10. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014. https://doi.org/10.1038/nature13385.
11. Singal G, Miller PG, Agarwala V, Li G, Kaushik G, Backenroth D, Gossai A, Frampton GM, Torres AZ, Lehnert EM, Miller VA. Association of patient characteristics and tumor genomics with clinical outcomes among patients with non-small cell lung cancer using a clinicogenomic database. JAMA. 2019. https://doi.org/10.1001/jama.2019.3241.
12. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, Shukla SA, Guo G, Brooks AN, Murray BA, The Cancer Genome Atlas Research Network. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet. 2016. https://doi.org/10.1038/ng.3564.
13. Zhang JC, Yao X, Devarakonda S, Deshpande A, Damrauer JS, Silva TC, Wong CK, Choi HY, Felau I, Robertson AG, et al. Whole-genome characterization of lung adenocarcinomas lacking alterations in the RTK/RAS/RAF pathway. Cell Rep. 2021. https://doi.org/10.1016/j.celrep.2021.108707.
14. Dimitrova N, Nagaraj AB, Razi A, Singh S, Kamalakaran S, Banerjee N, Joseph P, Mankovich A, Mittal P, DiFeo A, Varadan V. InFlo: a novel systems biology framework identifies cAMP-CREB1 axis as a key modulator of platinum resistance in ovarian cancer. Oncogene. 2016. https://doi.org/10.1038/onc.2016.398.
15. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Yang T-HO, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, et al. The immune landscape of cancer. Immunity. 2018. https://doi.org/10.1016/j.immuni.2018.03.023.
16. Tuck D. A cancer graph. 2021. Harvard Dataverse. https://doi.org/10.7910/DVN/RIXLG8.
17. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife. 2017. https://doi.org/10.7554/elife.26726·.
18. Hetionet—an integrative network of biomedical knowledge https://het.io. Accessed 29 Dec 2021
19. Fabregat A, Korninger F, Viteri G, Sidiropoulos K, Marin-Garcia P, Ping P, Wu G, Lincoln L. Biology. 2018. https://doi.org/10.1371/journal.pcbi.1005968.
20. Sugiyama M, Ghisu ME, Llinares-López F, Borgwardt K. graphkernels: R and Python packages for graph comparison. Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/btx602.
21. Panyam NC, Verspoor K, Cohn T, Ramamohanarao K. Exploiting graph kernels for high performance biomedical relation extraction. J Biomed Semant. 2018. https://doi.org/10.1186/s13326-017-0168-3·.
22. Qiangrong J, Guang Q. Graph kernels combined with the neural network on protein classification. J Bioinform Comput Biol. 2019. https://doi.org/10.1142/s0219720019500306.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.