Graph Databases in Molecular Biology

Waldeyr M. C. da Silva[1,2](✉), Polyane Wercelens[2],
Maria Emília M. T. Walter[2], Maristela Holanda[2], and Marcelo Brígido[2]

[1] Federal Institute of Goiás, Formosa, Brazil
waldeyr.mendes@ifg.edu.br
[2] University of Brasília, Brasília, Brazil

Abstract. In recent years, the increase in the amount of data generated in basic social practices and specifically in all fields of research has boosted the rise of new database models, many of which have been employed in the field of Molecular Biology. NoSQL graph databases have been used in many types of research with biological data, especially in cases where data integration is a determining factor. For the most part, they are used to represent relationships between data along two main lines: (i) to infer knowledge from existing relationships; (ii) to represent relationships from a previous data knowledge. In this work, a short history in a timeline of events introduces the mutual evolution of databases and Molecular Biology. We present how graph databases have been used in Molecular Biology research using High Throughput Sequencing data, and discuss their role and the open field of research in this area.

Keywords: Graph databases · Molecular Biology · Omics
Contributions

1 Introduction

The development of Molecular Biology precedes the development of the modern Turing-machine based Computer Sciences. However, from the moment they meet up, this close and cooperative relationship has intensified and accelerated advances in the field. In 1953, the structure of DNA was described by Watson and Crick [39] paving the way for the central dogma of Molecular Biology, which has been continuously enhanced by the discoveries of science. In the same year, International Business Machines (IBM) launched its first digital computer, the IBM 701. In the decade that followed, the genetic code was deciphered with the triplet codon pattern identification [12], which was almost wholly cracked in the following years [21]. Meanwhile, the history of modern databases began in 1964 at General Electric, when the first considered commercial Database Management System (DBMS) was born and named IDS - Integrated Data Store [3], [2]. From then on, other initiatives appeared, such as Multivalue [15], MUMPS [25], and IMS [13], which was designed for the Apollo space program in 1966.

The 1970s brought in the relational model [8], a very significant database model that, according to the site DB-Engines, even today, it is the most widely

R. Alves (Ed.): BSB 2018, LNBI 11228, pp. 50–57, 2018.
https://doi.org/10.1007/978-3-030-01722-4_5

used throughout the world. The first DNA sequencing was completed in 1971 [41], and in 1975, Sanger's method [28] for DNA sequencing, led to a generation of methods [18]. Due to the possibility of using computers to analyze DNA sequence data, in 1976 the first COBOL program to perform this type of analysis was published [22] and could be considered the birth of Bioinformatics, even though the name had yet to be coined.

In the 1980s, discussions about the human genome naturally emerged with the advances in DNA sequencing which were due to the affordable costs [29]. Throughout the 1990s, the Human Genome Project conducted the sequencing of the human DNA, which in 2001, culminated in the publications of the two competitors in this assignment [19,38]. Also, in the 1990s, the modern Internet emerged, and in the second half of the 1990s, the world experienced the Internet bubble phenomenon [5].

The efforts of the genome projects have promoted new technologies for sequencing, such as the High-Throughput Sequencing technologies (HTS), which have been used in laboratories worldwide. Nowadays, biological data has increased intensely in volume and diversity, becoming known as omics (genomics, transcriptomics, epigenomics, proteomics, metabolomics, and others). Projections on omics are impressive, and it is estimated that in 2025, genomics will generate one zetta-bases per year, which enables us to characterize the omics as a Big Data science [32]. NoSQL databases have played a significant role in managing large volumes of data, and as in other areas, the omics recently became a target of the NoSQL movement.

Although the NoSQL movement does not have a consensual definition, the literature points out that NoSQL is an umbrella term for non-relational database systems that provide mechanisms for storing and retrieving data, and which has modeling that is an alternative to traditional relational databases and their Structured Query Language (SQL). According to Corbellini [10], there are different types of NoSQL database models commonly classified as key-value, wide column or column families, document-oriented, and graph databases. Despite this classification, the NoSQL databases may be hybrids, using more than one database model.

In this review, we summarize the current NoSQL graph databases contributions to Molecular Biology, limited to their use in omics data from HTS, from the time they were first reported in the literature. We approach the contributions of NoSQL graph databases to the different fields of Molecular Biology exploring technical characteristics for the efficient storage of data. Finally, we conclude by discussing the role of NoSQL graph databases and the open field of research in this area.

2 NoSQL Graph Databases

Graphs naturally describe problem domains, and graph databases assemble simple abstractions of vertices and relationships in connected structures, making it possible to build models that are mapped closer to the problem domain. Often, data sets and their relationships are represented by graphs, and the importance of the information embedded in relationships has prompted an increase in the number of graph database initiatives [1]. This occurs due to various factors, such as the interests in recommending systems, circuits in engineering, social media, chemical and biological networks, and the search and identification of criminal cells [31].

Graph databases are Database Management Systems (DBMS) with Create, Read, Update, and Delete (CRUD) methods, which can store graphs natively or emulate them in a different database model [27]. The schema in graph databases can store data in vertices and, depending on the database, can also be stored in edges [30].

A significant aspect of graph databases is the way they manage relationships making it possible to establish them between entities. It is similar to storing pointers between two objects in memory. In addition, indexes can make the data retrieve of queries more efficient. However, there are some restrictions for types, as the BLOB type (Binary Large Object), which is not yet supported by graph databases.

3 Graph Databases Applied to Omics Data

In this section, we present works in which the NoSQL graph databases bring contributions to the Molecular Biology using omics data.

With the advent of NoSQL databases, a fundamental question loomed: would the NoSQL databases be ready for Bioinformatics? Have and Jensen [16] published a paper answering this question for NoSQL graph databases. In their work, they measured the performance of the graph database Neo4J v1.8 and the relational database PostgreSQL v9.05 executing some operations on data from STRING [36]. They found, for example, that the graph database found the best scoring path between two proteins faster by a factor of almost 1000 times. Also, the graph database found the shortest path 2441 times faster than the relational database when constraining the maximal path length to two edges. The conclusion was that graph databases, in general, are ready for Bioinformatics and they could offer great speedups on selected problems over relational databases.

Bio4j [26] proposes a graph-based solution for data integration with high-performance data access and a cost-effective cloud deployment model. It uses Neo4J to integrate open data coming from different data sources considering the intrinsic and extrinsic semantic features. Corbacho *et al.* [9] used the Bio4J graph database for Gene Ontology (GO) analyzes in *Cucumis melo*.

ncRNA-DB [6] is a database that integrates ncRNAs data interactions from a large number of well-established online repositories built on top of the OrientDB. It is accessible through a web-based platform, command-line, and the ncINetView, a plugin for Cytoscape[1], which is a software for analyses and visualization of biological networks. Another Cytoscape plugin is the cyNeo4j [33],

[1] www.cytoscape.org.

designed to link Cytoscape and Neo4j and enable an interactive execution of an algorithm by sending requests to the server.

Henkel *et al.* [17] used the Neo4J to integrate the data from distinct system biology model repositories. This database offers curated and reusable models to the community, which describe biological systems through Cypher Query Language - the native query language of Neo4J.

Lysenko *et al.* [20] used a graph database to provide a solution to represent disease networks and to extract and analyze exploratory data to support the generation of hypotheses in disease mechanisms.

EpiGeNet [4] uses the Neo4J to storage genetic and epigenetic events observed at different stages of colorectal cancer. The graph database enhanced the exploration of different queries related to colorectal tumor progression when compared to the primary source StatEpigen[2].

The Network Library [34] used Neo4J to integrate data from several biological databases through a clean and well-defined pipeline.

2Path [30] is a metabolic network implemented in the Neo4J to manage terpenes biosynthesis data. It used open data from several sources and was modeled to integrate important biological characteristics, such as the cellular compartmentalization of the reactions.

Biochem4j [35] is another work that seeks integration of open data from different sources using Neo4J. It goes beyond a database and provides a framework starting point for this integration and exploration of an ever-wider range of biological data sources.

GeNNet [11] is an integrated transcriptome analysis platform that uses Neo4J graph database to unify scientific workflows storing the results of transcriptome analyses.

BioKrahn [24] is a graph-based deductive and integrated database containing resources related to genes, proteins, miRNAs, and metabolic pathways that take advantage of the power of knowledge graphs and machine reasoning, to solve problems in the domain of biomedical science as interpreting the meaning of data from multiple sources or manipulated by various tools.

Messaoudi [23] evaluated the performance time needed for storing, deleting and querying biomedical data of two species: *Homo sapiens* as a large dataset and *Lactobacillus Rhamnosus* as a small dataset, using Neo4J and OrientDB graph databases. They found that Neo4J showed a better performance than OrientDB using 'PERIODIC COMMIT' technique for importing, inserting and deleting. On the other hand, OrientDB achieved best performances for queries when more in-depth levels of graph traversal were required.

Reactome [14] is a well-established open-source, open-data, curated and peer-reviewed database of pathways, which recently adopted the graph database as a storage strategy due to performance issues associated with queries traversing highly interconnected data. In this case, the adoption of graph database improved the queries reducing the average query time by 93%.

[2] http://statepigen.sci-sym.dcu.ie.

Arena-Idb is a platform for the retrieval of comprehensive and non-redundant annotated ncRNAs interactions [7]. It uses two different DBMS: a relational MySQL and the graph database Neo4J, which is applied to handle the construction and visualization of the networks on a web page.

Table 1 summarizes the contributions of each reported work in this review. Although there are many NoSQL graph databases available, so far only three of them (Neo4J, OrientDB and Grakn) have been reported in this field as shown in the Fig. 1.

Table 1. Contributions of graph-oriented databases for Molecular Biology

Graph database	Main contribution	Other contributions	Source
Neo4J	Biological networks	Protein-protein interaction	[16]
Neo4J	Gene annotation	GO analyses	[9]
OrientDB	Data integration	ncRNA interactions	[6]
Neo4J	Data integration	-	[17]
Neo4J	Data integration	-	[26]
Neo4J	Data visualization	-	[33]
Neo4J	Biological networks	Diseases association	[20]
Neo4J	Cancer	Epigenetic events	[4]
Neo4J	Data integration	-	[34]
Neo4J	Biological networks	Metabolic networks	[30]
Neo4J	Data integration	-	[35]
Neo4J	Transcriptome analyses	-	[11]
Grakn	Data integration	Biomedical analyses	[24]
Neo4J/OrientDB	Biomedical analyses	-	[23]
Neo4J	Biological networks	Metabolic networks	[14]
Neo4J	Biological networks	ncRNA interactions	[7]

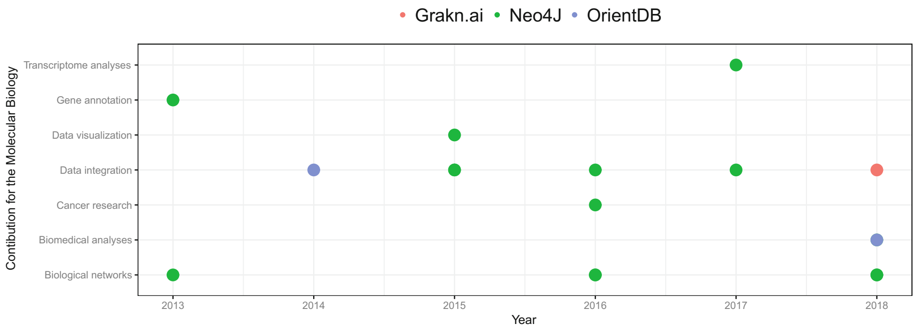

Fig. 1. Main contributions for the Molecular Biology using graph databases and omics data.

4 Discussion and Conclusion

In this work, we listed meaningful contributions of NoSQL graph databases to Molecular Biology using omics data. Performing queries across databases is routine activity in *in silico* biological studies, which, despite the available interfaces, is not a trivial task [35]. In this sense, the data integration is both a contribution to the field of Molecular Biology and Computer Science.

Data integration and biological networks were the most significant fields where the graph databases were employed. Data integration intends to represent relationships from previously related data knowledge, while metabolic networks intend to infer knowledge from existing relationships. However, it seems there is a hierarchy within data integration where it is a root contribution from which the others are derived. Biological networks are intuitively represented as graphs, and the use of graph databases for this purpose was predictable.

Once the data has already been processed and entered into the graph database, the queries become very intuitive and fast because of the way the nodes can be traversed. The performance and intuitiveness of queries in graph databases seem to be the main reason for using them as discussed in [14]. Graph queries are more concise and intuitive compared to equivalent relational database SQL queries complicated by joins. In addition, the engine of the graph databases is different, which leads to another point of investigation regarding the relationship between an engine and performance.

Databases contribute to the efficient storage of data, helping to ensure essential aspects of information security such as availability and integrity. The lack of schema in NoSQL graph databases, despite offering flexibility, can also remove the interoperability pattern of the data [20]. Graph database schemas may positively influence the maintainability of the graph databases, and open an ample field to examine the best graph schema for the data and their relationships concerning the normalization of data. A significant point to explore here is the threshold where the granularity of the vertices negatively influences the complexity and performance of the queries. Graph Description Diagram for graph databases (GRAPHED) [37] offers rich modeling diagrams for this purpose.

Although the scientific production using NoSQL databases is growing fast, the non-mutual citation supposedly shows a not explicit collaborative network. In summary, the use of NoSQL graph databases to store general data has increased, and the main contributions are related to data integration and performance in searches with queries traversing complex relationships. graph databases can help reach these solutions following the FAIR Guiding Principles for scientific data management and stewardship, which aims to improve the findability, accessibility, interoperability, and reuse of digital assets [40].

Acknowledgements. W. M. C. S. kindly thanks CAPES and IFG. M. E. M. T. W. thanks CNPq (Project 308524/2015-2).

References

1. Angles, R., et al.: Benchmarking database systems for social network applications. In: First International Workshop on Graph Data Management Experiences and Systems, p. 15. ACM (2013)
2. Bachman, C.W.: Integrated data store. DPMA Q. **1**(2), 10–30 (1965)
3. Bachman, C.W.: The origin of the integrated data store (IDS): the first direct-access dbms. IEEE Ann. History Comput. **31**, 42–54 (2009)
4. Balaur, I., et al.: EpigeNet: a graph database of interdependencies between genetic and epigenetic events in colorectal cancer. J. Comput. Biol. **24**, 969–980 (2017)
5. Berners-Lee, T., et al.: World-wide web: the information universe. Internet Res. **20**(4), 461–471 (2010)
6. Bonnici, V., et al.: Comprehensive reconstruction and visualization of non-coding regulatory networks in human. Front. Bioeng. Biotechnol. **2**, 69 (2014)
7. Bonnici, V., et al.: Arena-Idb: a platform to build human non-coding RNA interaction networks, pp. 1–13 (2018)
8. Codd, E.F.: A relational model of data for large shared data banks. Commun. ACM **13**(6), 377–387 (1970)
9. Corbacho, J., et al.: Transcriptomic events involved in melon mature-fruit abscission comprise the sequential induction of cell-wall degrading genes coupled to a stimulation of endo and exocytosis. PloS ONE **8**(3), e58363 (2013)
10. Corbellini, A., et al.: Persisting big-data: the NoSQL landscape. Inf. Syst. **63**, 1–23 (2017)
11. Costa, R.L., et al.: GeNNet: an integrated platform for unifying scientific workflows and graph databases for transcriptome data analysis. PeerJ **5**, e3509 (2017)
12. Crick, F.H., et al.: General nature of the genetic code for proteins. Nature **192**(4809), 1227–1232 (1961)
13. Deen, S.M.: Fundamentals of Data Base Systems. Springer, Heidelberg (1977). https://doi.org/10.1007/978-1-349-15843-0
14. Fabregat, A., et al.: Reactome graph database: efficient access to complex pathway data. PLoS Comput. Biol. **14**(1), 1–13 (2018)
15. Fry, J.P., Sibley, E.H.: Evolution of data-base management systems. ACM Comput. Surv. (CSUR) **8**(1), 7–42 (1976)
16. Have, C.T., Jensen, L.J.: Are graph databases ready for bioinformatics? Bioinformatics **29**(24), 3107 (2013)
17. Henkel, R., Wolkenhauer, O., Waltemath, D.: Combining computational models, semantic annotations and simulation experiments in a graph database. Database **2015** (2015)
18. Hutchison III, C.A.: Dna sequencing: bench to bedside and beyond. Nucl. Acids Res. **35**(18), 6227–6237 (2007)
19. Lander, E.S.: Initial sequencing and analysis of the human genome. Nature **409**(6822), 860–921 (2001)
20. Lysenko, A., et al.: Representing and querying disease networks using graph databases. BioData Min. **9**, 23 (2016)
21. Martin, R.G., et al.: Ribonucleotide composition of the genetic code. Biochem. Biophys. Res. Commun. **6**(6), 410–414 (1962)
22. McCallum, D., Smith, M.: Computer processing of dna sequence data. J. Mol. Biol. **116**, 29–30 (1977)
23. Messaoudi, C., Mhand, M.A., Fissoune, R.: A performance study of NoSQL stores for biomedical data NoSQL databases: an overview, November 2017 (2018)

24. Messina, A., Pribadi, H., Stichbury, J., Bucci, M., Klarman, S., Urso, A.: BioGrakn: a knowledge graph-based semantic database for biomedical sciences. In: Barolli, L., Terzo, O. (eds.) CISIS 2017. AISC, vol. 611, pp. 299–309. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-61566-0_28
25. O'Neill, J.T.: MUMPS language standard, vol. 118. US Department of Commerce, National Bureau of Standards (1976)
26. Pareja-Tobes, P., et al.: Bio4j: a high-performance cloud-enabled graph-based data platform. bioRxiv (2015)
27. Robinson, I., Webber, J., Eifrem, E.: Graph Databases. O'Reilly Media Inc, Sebastopol (2013)
28. Sanger, F., Coulson, A.R.: A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. **94**(3), 441IN19447–441IN20448 (1975)
29. Shreeve, J.: The Genome War: How Craig Venter Tried to Capture the Code of Life and Save the World. Random House Digital Inc., Manhattan (2005)
30. Silva, W.M.C.D., et al.: A terpenoid metabolic network modelled as graph database. Int. J. Data Min. Bioinform. **18**(1), 74–90 (2017)
31. Srinivasa, S.: Data, storage and index models for graph databases. In: Sakr, S., Pardede, E. (eds.) Graph Data Management, pp. 47–70. IGI Global, Hershey (2011)
32. Stephens, Z.D., et al.: Big data: astronomical or genomical? PLoS Biol. **13**(7), e1002195 (2015)
33. Summer, G., et al.: cyNeo4j: connecting neo4j and cytoscape. Bioinformatics **31**(23), 3868–3869 (2015)
34. Summer, G., et al.: The network library: a framework to rapidly integrate network biology resources. Bioinformatics **32**(17), i473–i478 (2016)
35. Swainston, N., et al.: biochem4j: Integrated and extensible biochemical knowledge through graph databases. PloS ONE **12**(7), e0179130 (2017)
36. Szklarczyk, D., et al.: The string database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucl. Acids Res. **45**(D1), D362–D368 (2017)
37. Van Erven, G., Silva, W., Carvalho, R., Holanda, M.: GRAPHED: a graph description diagram for graph databases. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) WorldCIST'18 2018. AISC, vol. 745, pp. 1141–1151. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-77703-0_111
38. Venter, J.C., et al.: The sequence of the human genome. Science **291**(5507), 1304–1351 (2001)
39. Watson, J.D., Crick, F.H.: A structure for deoxyribose nucleic acid. Nature **171**(4356), 737–738 (1953)
40. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3** (2016). https://doi.org/10.1038/sdata.2016.18
41. Wu, R., Taylor, E.: Nucleotide sequence analysis of DNA: II. Complete nucleotide sequence of the cohesive ends of bacteriophage λ DNA. J. Mol. Biol. **57**(3), 491–511 (1971)