2Path: a terpenoid metabolic network modeled as graph database

Waldeyr Silva[†*], Marcelo Brígido[†]
[†]Biological Sciences Institute, UnB, Brasília, Brazil
[*]Federal Institute of Goiás, IFG, Formosa, Brazil
mendes@iscb.org, brigido@unb.br

Danilo Vilar, Daniel Souza,
Maria Emília Walter, Maristela Holanda
Computer Science Departament, UnB, Brasília, Brazil
{danilo.vilar,dssouzadan}@gmail.com, {mia, mholanda}@cic.unb.br

***Abstract*—Terpenoids have critical ecological, industrial and commercial relevance. Interactions as signaling for communication intra/inter species, signal molecules to attract pollinating insects, or defense against herbivores and microbes, are protagonized by terpenoids. Due to their chemical composition, many terpenoids possess vast pharmacological applicability in medicine and biotechnology. The biosynthesis of terpenes has been widely studied over the years and today it is known that they can be synthesized from two metabolic pathways: Mevalonate pathway (MVA) and Non-mevalonate pathway (MEP). However, genome-scale reconstruction of metabolic networks faces many challenges, of which include organizational data storage and data modeling to properly represent the complexity of the system biology. Recent NoSQL database paradigms have introduced new concepts of scalable storage and data queries as graph databases, which are versatile enough to cope with biological data. In this paper, we provide an overview of 2Path, a terpenoid metabolic network modeled and stored using a graph database, which aims to preserve important terpenoid biosynthesis characteristics and mitigate challenges in genome-scale storage metabolic networks.**

Keywords: NoSQL, graph database, terpenoid, metabolic, network.

I. Introduction

Computation has been a great allied of research in Molecular Biology in several instances: sequence alignment, assembly and mapping of biological sequences, implementation of statistical methods and data storage, among others. In the case of storage, databases play a key role. Metabolic networks are a collection of metabolic reactions. The reconstruction of metabolic networks is an important scientific target that seeks to understand the metabolism of an organism and its complex interactions and the data storage is equally important.

According to Have [1], graph databases are adequate for bioinformatics and can improve speedups over relational databases on selected problems. In this work, we proposed 2Path: a terpenoid metabolic network modeled as graph database. We modeled and build a core database with useful information of terpenoids metabolic pathways, including cellular compartimentalization, which can be used to explore metabolic characteristics of an organism of interest.

This article is organized as follows: Section II presents the biological problem to be modeled in our database. Section III presents a review of the use of databases for metabolic networks. Section IV presents the method used to construct our terpenoids metabolism database. Section V presents the results obtained, and after it, we present our conclusions and perspectives in Section VI.

II. Terpenoid metabolic network

The metabolism can be understood as the set of metabolic reactions and physicochemical processes that occur in an organism to keep it alive. Metabolic reactions are biochemical reactions that transform chemical compounds (substrates) into other chemical compounds (products) [2]. The metabolites are compounds chemically converted by metabolic reactions of biosynthesis or degradation [3]. However, certain criteria is required to classify a compound as a metabolite [4]:

- metabolites are recognized and affected by enzymes;
- the product of a reaction may be substrate for another reaction;
- metabolites have a finite existence, they do not accumulate in the cells;
- metabolites must have a biological role in cell, including regulation of its own metabolism.

Enzymes are proteins which catalyze metabolic reactions. Sometimes, besides of the enzyme, the reaction requires an additional molecule called a cofactor [5]. Cofactors can be organic or inorganic. The organic cofactors are also known as coenzymes [2]. The set of metabolic reactions that are essential for the organism, such as cellular respiration, comprising the primary metabolism [3]. The secondary metabolism is composed of a set of metabolic reactions inessential to the organism and tends to be specific to each species.

According to Keller [6], the classes of secondary metabolites are: Polyketides (PKS), Non-ribosomal peptides (NRP), Alkaloids and Terpenes. Among secondary metabolites, terpenoids, especially those produced by plants, act as a defense against microorganisms, insects and herbivores as well as a signal to attract insects, animal dispersers of seeds or fruits, and herbivorous insect predators [7].

The biosynthesis of secondary metabolites is a highly coordinated process that includes the formation of a metabolome and a secondary metabolic network. This metabolic network may also require different cell types utilizing a range of cell compartmentalization features, particularly in plants, to ensure specific biosynthesis and to prevent interference from extraneous molecules during the process [8].

The classical pathway of terpenoid formation is the mevalonate pathway (MVA), which was first identified in yeast and in mammals in the 1950's [9]. Up until the 1990's, it was assumed that MVA served as the sole metabolic pathway responsible for Isopentenyl pyrophosphate (IPP) training and Dimethylallyl pyrophosphate (DMAPP) from Acetyl-CoA in organisms. In the 1990's, was discovered in bacteria and plants, a non-mevalonate pathway (MEP) able to generate IPP and DMAPP from D-glyceraldehyde 3-phosphate (G3P) and pyruvate [9].

In addition to evidence of cell partitioning in terpenoid biosynthesis [10], [11], [12] both intra and inter organisms, monoterpenoids can also be derived from long chain degraded terpenoids [13].

Environmental factors influence the production of terpenoids. In the absence of light, under effects of shading, plants mainly produce isoprenoid derivatives of MVA, or sterols, which are necessary for rapid growth.s. In the presence of light, isoprenoid synthesis in plastids is increases, while the sterol synthesis progressively reduces. Another factor that influences terpenoid synthesis is the presence of pathogens [14].

Volatile terpenoids are stored in specialized tissues of vascular plants, or the inner sheets and peripheral tissues of stems and roots [15]. In a strategy called indirect defense, plants damaged by herbivore parasites release volatile terpenoids stored in glandular trichomes, as a signal to attract its predators [16].

Researches have increasingly suggested to an update in the methods, techniques and tools to reconstruct and store metabolic networks. An example is the metabolic network reconstruction of the plant *Arabidopsis thaliana* with information of cellular compartmentalization and tissue specificity [17].

III. Metabolic Network Databases

One of the first approaches to the reconstruction of metabolic networks is the so-called factographic database on enzymes and metabolic pathways [18]. In 1992, an integrated database to support research on *Escherichia coli* was designed by Baehr *et al.* using logic programming [19]. Later, Karp *et al.* [20], concerned with creating a more appropriate model to capture the details of this complex area, proposed the EcoCyc Project. According to Karp and Riley [20], the representation of the metabolism should include mechanisms that allow to distinguish enzyme classes from individual enzymes, since there is not a one-to-one mapping from enzymes to the reactions they catalyze. Moreover, the species' variation of metabolism must be represented. Besides, approaches have since been developed as Petri net representations of metabolic pathways [21].

To date, other projects addressing network databases have been launched. Several of these projects include: KEGG [22], a knowledge base for the systematic analysis of gene functions, connecting genomic information with higher order functional information; ERATO Systems Biology Workbench [23], an integrated environment for multiscale and multitheoretic simulations in systems biology; PathFinder [24], a tool for the dynamic visualization of metabolic pathways based on annotation data. Parallel to these projects is PathwayTools [25], which supports the creation of new PGDBs[1] using its PathoLogic [26] component. BioCyc uses the Pathway Tools software to predict metabolic pathways in sequenced and annotated genomes. This database contains two data fields to support pathway inference: the expected taxonomic range of each pathway, and a list of key reactions for pathways. These two fields have significantly improved the predictive accuracy of PathoLogic [26].

New methods and tools have been designed to account for specific types of organisms. Employing a two-stage approach based on groups of orthologs proteins and the KEGG [22] and MetaCyc [26] dabatses, FungiPath [27] serves as a tool for reconstructing a more specific metabolic pathway. In addition, there are tool-free protocols for generating a high-quality genome-scale metabolic reconstruction, e.g. the one proposed by Thiele and Palsson [28].

To store a metabolic network, it is necessary to design a comprehensive and consistent data model. Databases of metabolic pathways have been constructed since 1989 [29], with many proposed methods to store metabolic networks as evidenced in Table I.

TABLE I
STRATEGIES OF STORAGE/ORGANIZATION OF THE MAIN METHODS RELATED TO METABOLIC NETWORKS.

	Factographic database	Structured files	Graphs	Relational Databases	Petri Network	Logic Programming	Others	GEM [30]
Factographic data bank [18]	X							
Integrated database [19]						X		
Metabolic knowledge [20]		X						
Petri net [21]					X			
Using incomplete information [31]						X		
KEGG [22]							X	
ERATO [23]		X						
PathFinder [24]			X					
Optstrain [32]							X	
Ab initio reconstruction [33]							X	
Genome-scale reconstruction [34]		X						
Petri Net in systems biology [35]					X			
Yeast reconstruction [36]								X
FUNGIpath [27]				X				
FARM [37]		X		X				

The development of big data, social networks, cloud services and other business and scientific applications, has generated massive amounts of data. This surplous of diverse data demands increased scalability and flexibility [38]. Recently,

[1]Pathway Genome Databases

NoSQL (Not Only Structured Query Language) databases emerged as an alternative, due to their scalability and flexibility. NoSQL databases do not require specific design schemas. However, a data model can be used and can contribute to the consistency of the data[39].

A graph database is a type of NoSQL database [40]. Graph theory is an area with many applications in computer science, genetics, chemistry, engineering, industry, business and social sciences since many years [41]. In contrast, graph databases is a recent area [42], which aims to store data in a graph, capable of representing many types of data in a highly accessible way [43]. Conceptually, instead of storing the data in tables, graph databases using vertices and edges that represent the relationships between the vertices [38]. The main elements of a database on graphs are vertices, relationships and properties [44]. According to Dominguez [45] the database graph is a data structure for a scheme and its instances modeled as graphs or its generalizations where manipulation of the data is expressed by operations oriented graphs and type constructors.

Graph databases have been effectively used to store, manage and update scientific data and relationships between them [46]. Seemingly, the use of NoSQL in Bioinformatics began in 2010 [47] and has since grown as shown in Figure 1.

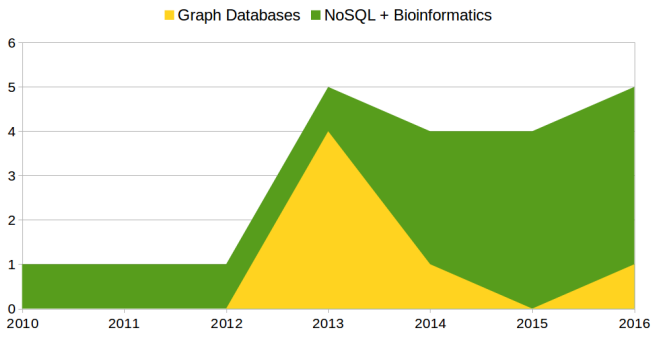

Fig. 1. Bioinformatics publications that have used NoSQL databases. Green is the NoSQL annual publications. Yellow Represents the specific annual Graph databases. There are 21 publications between 2010 and 2016, 6 of them using databases in graphs. We consulted PubMed, *Google Schoolar*, *Science Direct* and IEEE *Explorer*.

Almost a third of publications about NoSQL and bioinformatics are graph databases as shown in the Figure 1. In 2013, all published works related to Bioinformatics and NoSQL were about graph databases [1] [48] [42] [38]. Em 2014, Bonnici [49] used OrientDB to reconstruct and visualize non-coding regulatory networks in human. Now in 2016, the BioGraphDB [40] was published a graph database built on the OrientDB with related data to genes, microRNA (miRNA), proteins, pathways and diseases from ten online public resources.

IV. Method

The method presented here describes the construction of the database called 2Path. It is a terpenoids metabolism network database using the graph databases technology and is loaded with KEGG [22] data. Databases of biological content, such as KEGG[2] and BioCyc[3] have vast content of terpenoids metabolism. This is why we chose KEGG data to start our database.

A. 2Path

A graph database model is necessary for better management of graphs [50]. We built a graph database model structured to represents metabolics networks with rich biological details as cellular compartmentalization: the 2Path.

The 2Path model, shown in Figure 2, was constructed using the notation proposed by Erven [51] and contains the following data: enzymes, reactions, cofactors, compunds, cellular locus and organisms represented as nodes. The relationships between these nodes are represented by edges. In both nodes and edges, there are properties where its information is stored.

In the database model, the ENZYME nodes represents the known enzymes with catalytic activity. Its properties include the Enzyme Number (EC), its main name, Unipro ID and other cross references. An enzyme may be related to another enzyme, indicating that they are isozymes, namely, despite being different proteins have the same catalytic activity. The REACTION nodes stores the description of the reaction catalyzed by a particular enzyme, reaction reversibility and cross references. Reaction nodes are connected to the ENZYME and COMPOUND nodes. COMPOUND nodes can be substrate or a product of a REACTION. Its properties include the name of the compound, links to KEGG, BioCyc and Chebi and other cross-references. COFACTOR nodes, connected with REACTION nodes, represents the coenzymes that can be found covalently linked to the enzyme molecule, when consisting of a prostethic group of proteins, or as a free molecule that joins to REACTION only at the moment of catalysis, for example a $NAD+$ molecule. The ORGANISM contains all organisms taxonomically described by the NCBI[4]. CELLULARLOCUS node allows to store information about the location in the cell (organelles like chloroplast or cytosol) in which occur metabolic reaction. Relationships between nodes allow you to save information about the dynamics of metabolic reactions and compounds and enzymes.

We write a Java program to retrieve and process the data from KEGG [22] and cross references via Web service. Then, we use Talend ETL Software version 6.2.1[5] to load the database. The Talend ETL was used to treat and insert xml data in Neo4J [43] version 3.0.6, according to the model previously described.

V. Results

Through use of a graph databse, 2Path organizes all data relevant to terpenoids metabolism from KEGG [22]. Data is flexibly arranged to allow for manual curation. This is an important feature of the database because the cure is performed through manual review of the literature and focuses primarily

[2]http://www.kegg.jp
[3]http://biocyc.org
[4]https://www.ncbi.nlm.nih.gov/guide/taxonomy
[5]https://www.talend.com/download/talend-open-studio

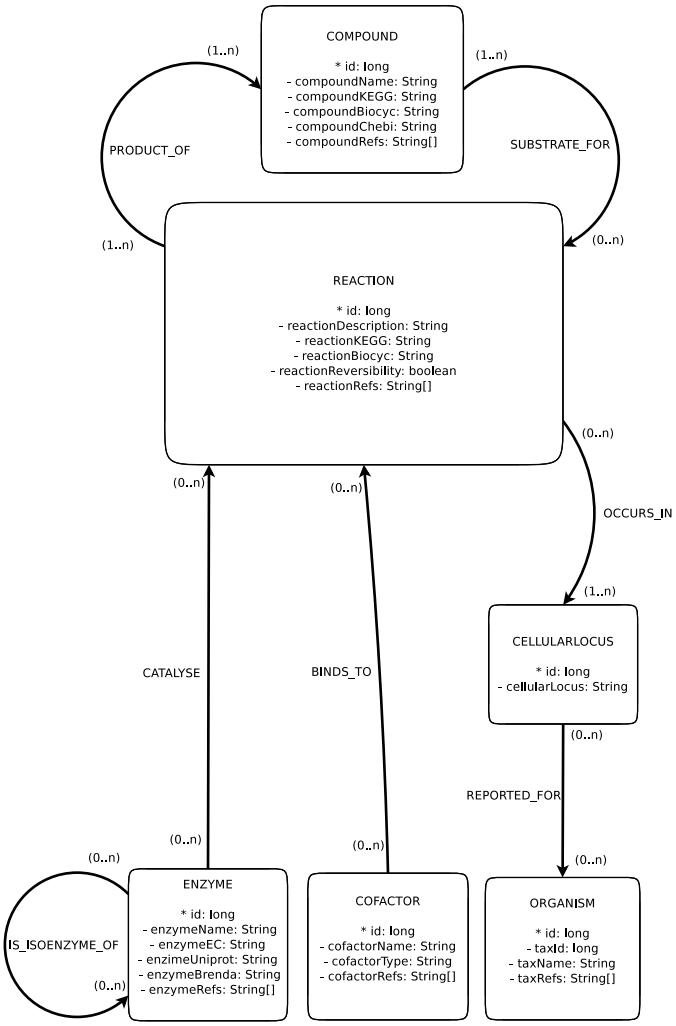

Fig. 2. 2Path database model. Graph database of pathway core containing the following data: enzymes, reactions, cofactors, compounds, cellular locus and organisms. The edges' multiplicity represents the minimum and maximum nodes that can exist in a relationship.

on the cell compartmentalisation of reactions reported for a given organism.

The first version of 2Path is comprised of thousands of secondary metabolism reactions. 2Path currently has 17,606 compounds, 7,069 enzymes and 1,811 reactions. A 2Path view can be seen in Figure 3. Of this total, 1,067 compounds, 679 enzymes and 1,001 reactions have been related to secondary metabolism. The compartmentalization cellular data will be loaded manually for the next two years, but some data, especially for *Arabidopsis thaliana* are already in the database. An example of stored metabolic pathway in 2Path is shown in the Figure 4.

2Path was modeled in an effort to answer several biological questions such as:

- Given the metabolites $x_1,\ x_2,\ \ldots\ x_m$ as input, are there biosynthetic pathways lead to the production of metabolites $y_1,\ y_2,\ \ldots\ y_n$?
- For which organisms?

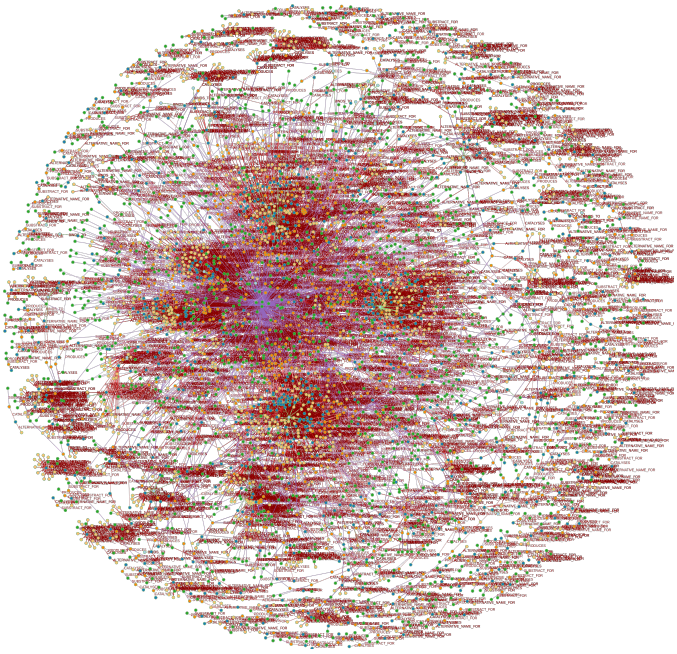

Fig. 3. Visualization of 2Path. In this example only enzymes, reactions and compounds are shown.

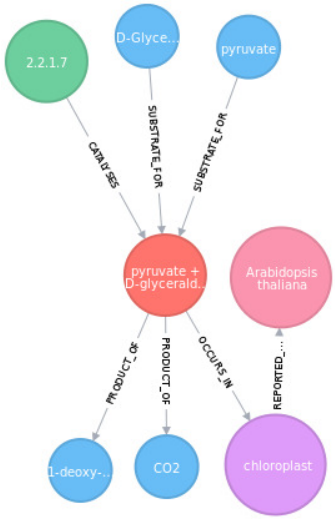

Fig. 4. This 2Path example represents the first reaction of MEP (non-mevalonate pathway). The nodes pyruvate and glyceraldehyde 3-phosphate are the substrates to a reaction that produces 1-Deoxy-D-xylulose 5-phosphate (DOXP). This reaction is catalysed by the DXP-synthase enzyme, whose EC is 2.2.1.7. This reaction is reported in the literature for *Arabidopsis thaliana* in the chloroplast.

- Based on taxonomy, what is the probable cellular location in which a specific reaction occurs?
- What is the probable tissue where a particular metabolite is stored?

VI. Conclusion

In this work, we proposed and implemented a graph database, using Neo4J, to store a terpenoid metabolic network. Our database was modeled in a way that preserves important

biological characteristics. This work introduces a new way to store and access metabolic networks. The use of graph databases as an alternative to traditional methods of storage proved practical and efficient. In addition, the use of a careful modeling database was a success factor in the result of implementation.

The biological target of this study, terpenoids biosynthesis, is highly relevant, ecologically and economically. Thus, this work opens new perspectives to explore. Moving forward, we plan to expand the 2Path database with data from gene regulation and two-dimensional chemical structures. We also plan to explore and implement new features in upcoming releases of our database. Upon the next release of 2Path, we aim to incorporate the reconstruction of terpenoid metabolic network to a given organism by submitting its genome. At a later date, we hope to develop a semi-automatic update capability. Other data sources such as BioCyc [52] and Reactome [53] will still be utilized.

ACKNOWLEDGMENTS

Waldeyr Mendes Cordeiro da Silva and Daniel Silva Souza kindly thanks CAPES for the scholarship.

REFERENCES

[1] C. T. Have, L. J. Jensen, and J. Wren, "Are graph databases ready for bioinformatics?" *Bioinformatics*, vol. 29, no. 24, pp. 3107–3108, 2013.

[2] D. L. Nelson, A. L. Lehninger, and M. M. Cox, *Lehninger principles of biochemistry*. Macmillan, 2008.

[3] G. Michal and D. Schomburg, *Biochemical pathways*. John Willey & Sons Inc., 2012.

[4] E. D. Harris, *Biochemical Facts behind the definition and Properties of Metabolites*. Texas A&M University, 2013.

[5] J. D. Fischer, G. L. Holliday, and J. M. Thornton, "The CoFactor database: Organic cofactors in enzyme catalysis," *Bioinformatics*, vol. 26, no. 19, pp. 2496–2497, 2010.

[6] N. P. Keller, G. Turner, and J. W. Bennett, "Fungal secondary metabolism - from biochemistry to genomics," *Nature Reviews Microbiology*, vol. 3, no. 12, pp. 937–947, 2005.

[7] K. Jørgensen, A. V. Rasmussen, M. Morant, A. H. Nielsen, N. Bjarnholt, M. Zagrobelny, S. Bak, and B. L. Møller, "Metabolon formation and metabolic channeling in the biosynthesis of plant natural products," *Current Opinion in Plant Biology*, vol. 8, no. 3 SPEC. ISS., pp. 280–291, 2005.

[8] M. Wink, *Annual plant reviews volume 40, Biochemistry of plant Secondary Metabolites*, 2010, vol. 40.

[9] J. Lombard and D. Moreira, "Origins and early evolution of the mevalonate pathway of isoprenoid biosynthesis in the three domains of life," *Molecular Biology and Evolution*, vol. 28, no. 1, pp. 87–99, 2011.

[10] J. A. Bick and B. Lange, "Metabolic cross talk between cytosolic and plastidial pathways of isoprenoid biosynthesis: unidirectional transport of intermediates across the chloroplast envelope membrane," *Archives of Biochemistry and Biophysics*, vol. 415, no. 2, pp. 146–154, 2003.

[11] S. Bartram, A. Jux, G. Gleixner, and W. Boland, "Dynamic pathway allocation in early terpenoid biosynthesis of stress-induced lima bean leaves," *Phytochemistry*, vol. 67, no. 15, pp. 1661–1672, 2006.

[12] D. A. Nagegowda, "Plant volatile terpenoid metabolism: Biosynthetic genes, transcriptional regulation and subcellular compartmentation," {*FEBS*} *Letters*, vol. 584, no. 14, pp. 2965–2973, 2010.

[13] P. Sun, R. C. Schuurink, J.-C. Caissard, P. Hugueney, and S. Baudino, "My Way: Noncanonical Biosynthesis Pathways for Plant Volatiles," *Trends in Plant Science*, pp. –, 2016.

[14] E. Vranová, D. Coman, and W. Gruissem, "Structure and dynamics of the isoprenoid pathway network," *Molecular Plant*, vol. 5, no. 2, pp. 318–333, 2012.

[15] M. E. Maffei, "Sites of synthesis, biochemistry and functional role of plant volatiles," *South African Journal of Botany*, vol. 76, no. 4, pp. 612–631, 2010.

[16] G. ichiro Arimura, C. Kost, and W. Boland, "Herbivore-induced, indirect plant defences," *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, vol. 1734, no. 2, pp. 91 – 111, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S138819810500048X

[17] S. Mintz-Oron, S. Meir, S. Malitsky, E. Ruppin, A. Aharoni, and T. Shlomi, "Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 1, pp. 339–44, 2012.

[18] E. Selkov, I. Goryanin, N. Kaimatchnikov, E. Shevelev, and I. Yunus, "Factographic data bank on enzymes and metabolic pathways," *Studia Biophysica*, vol. 129, no. 2-3, pp. 155–164, 1989.

[19] a. Baher, G. Dunham, a. Ginaburg, R. Hagstrom, D. Joerg, T. Krazik, H. Matsuda, G. Michaels, R. Overbeek, C. Smith, R. Taylor, K. Yoshida, and D. Zawada, "Integrated database to support research on Escherichia coli." *Argonne Technical Report*, vol. ANL92/1, no. JANUARY, 1992.

[20] P. D. Karp and S. M. Paley, "Representations of metabolic knowledge: pathways," *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB)*, vol. 2, pp. 203–211, 1994.

[21] V. N. Reddy, M. L. Mavrovouniotis, and M. N. Liebman, "Petri net representations in metabolic pathways." *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 1, no. August, pp. 328–336, 1993.

[22] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[23] M. Hucka, A. Finney, H. Sauro, H. Bolouri, J. Doyle, and H. Kitano, "The ERATO Systems Biology Workbench: An Integrated Environment for Multiscale and Multitheoretic Simulations in Systems Biology," in *Foundations of Systems Biology*. Cambridge, MA: The MIT Press, 2001, ch. 6, pp. 125–143.

[24] A. Goesmann, M. Haubrock, F. Meyer, J. Kalinowski, and R. Giegerich, "PathFinder: reconstruction and dynamic visualization of metabolic pathways." *Bioinformatics (Oxford, England)*, vol. 18, no. 1, pp. 124–129, 2002.

[25] P. D. Karp, S. Paley, and P. Romero, "The Pathway Tools software," *Bioinformatics*, vol. 18, no. Suppl 1, pp. S225–S232, Jul. 2002.

[26] P. D. Karp, M. Latendresse, and R. Caspi, "The pathway tools pathway prediction algorithm," *Standards in genomic sciences*, vol. 5, no. 3, pp. 424–9, Dec. 2011.

[27] S. Grossetête, B. Labedan, and O. Lespinet, "FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology." *BMC genomics*, vol. 11, no. Table 1, p. 81, 2010.

[28] I. Thiele and B. O. Palsson, "A protocol for generating a high-quality genome-scale metabolic reconstruction." *Nature protocols*, 2010.

[29] E. Selkov, I. Goryanin, N. Kaimatchnikov, E. Shevelev, and I. Yunus, "Factographic data bank on enzymes and metabolic pathways," *Studia Biophysica*, vol. 129, no. 2-3, pp. 155–164, 1989.

[30] K. Arakawa, Y. Yamada, K. Shinoda, Y. Nakayama, and M. Tomita, "GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes." *BMC bioinformatics*, vol. 7, p. 168, Jan. 2006.

[31] T. Gaasterland and E. Selkov, "Reconstruction of metabolic networks using incomplete information." *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 3, pp. 127–135, 1995.

[32] P. Pharkya and C. Maranas, "Optstrain: a Hierarchical Metabolic Pathway Discovery and Design Framework for Microbial Production Systems," *Projects.Csail.Mit.Edu*, no. Eppstein 1994, 2002.

[33] F. Boyer and A. Viari, "Ab initio reconstruction of metabolic pathways," *Bioinformatics*, vol. 19, 2003.

[34] J. Förster, I. Famili, and P. Fu, "Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network," *Genome Research*, pp. 244–253, 2003.

[35] J. W. Pinney, D. R. Westhead, and G. a. McConkey, "Petri Net representations in systems biology." *Biochemical Society transactions*, vol. 31, no. iv, pp. 1513–1515, 2003.

[36] M. J. Herrgård, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blüthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novère, P. Li, W. Liebermeister, M. L. Mo, A. P.

Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasić, D. Weichart, R. Brent, D. S. Broomhead, H. V. Westerhoff, B. Kirdar, M. Penttilä, E. Klipp, B. O. Palsson, U. Sauer, S. G. Oliver, P. Mendes, J. Nielsen, and D. B. Kell, "A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology." *Nature biotechnology*, vol. 26, no. 10, pp. 1155–1160, 2008.

[37] J. M. Dreyfuss, J. D. Zucker, H. M. Hood, L. R. Ocasio, M. S. Sachs, and J. E. Galagan, "Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus Neurospora crassa using FARM." *PLoS computational biology*, vol. 9, no. 7, p. e1003126, Jan. 2013.

[38] H. Huang and Z. Dong, "Research on architecture and query performance based on distributed graph database Neo4j," in *2013 3rd International Conference on Consumer Electronics, Communications and Networks*. Chongqing, China: IEEE, Nov. 2013, pp. 533–536.

[39] G. Manyam, M. A. Payton, J. A. Roth, L. V. Abruzzo, and K. R. Coombes, "Relax with CouchDB - Into the non-relational DBMS era of bioinformatics," *Genomics*, vol. 100, no. 1, pp. 1–7, 2012.

[40] A. Fiannaca, I. National, L. La, P. Universit, A. Urso, I. National, M. La, and R. Italian, "BioGraphDB: a New GraphDB Collecting Heterogeneous Data for Bioinformatics Analysis," in *The Eighth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies BioGraphDB:*, no. August, Palermo, Italy, 2016.

[41] V. I. Voloshin, *Introduction to Graph and Hypergraph Theory*. New York: NOVA Science Publishers Inc., 2009.

[42] S. Jouili and V. Vansteenberghe, "An Empirical Comparison of Graph Databases," in *2013 International Conference on Social Computing*. Mont-Saint-Guibert, Belgium: IEEE, Sep. 2013, pp. 708–715.

[43] Neo4J, "Neo4j manual v2.2.5," http://neo4j.com, 2015.

[44] R. Pinheiro, B. Aires, A. F. Araujo, M. Holanda, M. E. Walter, and S. Lifschitz, "Storing provenance data of genome project workflows using graph database," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 16–22.

[45] D. Dominguez-Sal, N. Martinez-Bazan, V. Muntes-Mulero, P. Baleta, and J. L. Larriba-Pey, "A discussion on the design of graph database benchmarks," in *Technology Conference on Performance Evaluation and Benchmarking*. Springer, 2010, pp. 25–40.

[46] R. Kaliyar, "Graph Databases : A Survey," in *International Conference on Computing, Communication and Automation (ICCCA2015)*, Galgotias University, India, 2015, pp. 785–790.

[47] B. D. O'Connor, B. Merriman, and S. F. Nelson, "SeqWare Query Engine: storing and searching sequence data in the cloud," *BMC Bioinformatics*, vol. 11, no. Suppl 12, p. S2, 2010.

[48] K. K.-Y. Lee, W.-C. Tang, and K.-S. Choi, "Alternatives to relational database: comparison of NoSQL and XML approaches for clinical data storage," *Computer methods and programs in biomedicine*, vol. 110, no. 1, pp. 99–109, 2013.

[49] V. Bonnici, F. Russo, N. Bombieri, A. Pulvirenti, and R. Giugno, "Comprehensive reconstruction and visualization of non-coding regulatory networks in human." *Frontiers in bioengineering and biotechnology*, vol. 2, no. December, p. 69, 2014.

[50] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2013.

[51] G. G. a. V. Erven, "MDG-NoSQL: Data models to NoSQL databases based in graphs," Master's thesis, Department of Computer Science, University of Brasilia, 2015, in Portuguese.

[52] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. a. Fulcher, T. a. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. a. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases," *Nucleic Acids Research*, vol. 42, no. D1, pp. 459–471, 2014.

[53] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 42, no. D1, pp. D472–D477, 2014.