University of Brasília

Institute of Biological Sciences
Department of Cellular Biology

In silico Reconstruction of Sesquiterpenes Metabolic Network of *Copaifera multijuga* Hayne

Reconstrução *in silico* da Rede Metabólica de Sesquiterpenos da *Copaifera multijuga* Hayne

Waldeyr Mendes Cordeiro da Silva

Thesis presented as a partial requirement for the conclusion of the Ph.D. in Molecular Biology

Advisor
Prof. Ph.D. Marcelo de Macêdo Brígido

Co-advisor
Prof. Ph.D. Maria Emília Machado Telles Walter

Brasília
2018

Ficha Catalográfica de Teses e Dissertações

Está página existe apenas para indicar onde a ficha catalográfica gerada para dissertações de mestrado e teses de doutorado defendidas na UnB. A Biblioteca Central é responsável pela ficha, mais informações nos sítios:

http://www.bce.unb.br
http://www.bce.unb.br/elaboracao-de-fichas-catalograficas-de-teses-e-dissertacoes

Esta página não deve ser inclusa na versão final do texto.

University of Brasília
Institute of Biological Sciences
Department of Cellular Biology

In silico Reconstruction of Sesquiterpenes Metabolic Network of *Copaifera multijuga* Hayne

Reconstrução *in silico* da Rede Metabólica de Sesquiterpenos da *Copaifera multijuga* Hayne

Waldeyr Mendes Cordeiro da Silva

Thesis presented as a partial requirement for the
conclusion of the Ph.D. in Molecular Biology

Prof. Ph.D. Marcelo de Macêdo Brígido (Advisor)
CEL/UnB

Prof. Ph.D. Georgios Joannis Pappas Júnior
CEL/UnB

Prof. Ph.D. Valdir Florêncio da Veiga Junior
Instituto Militar de Engenharia

Prof. Ph.D. João Batista Simão
Universidade Federal de Uberlândia

Ph.D. Roberto Coiti Togawa
Embrapa

Prof. Ph.D. Sônia Freitas
Coordenator of the Postgraduate Program in Molecular Biology

Brasília, 27 July 2018

Dedication

To my family, especially Suemilie Koch, Ana Clara Koch Mendes, Zulmira Mendes, Waldeylson Silva, Ana Mendes, Islara Mendes, Rudi and Claudete Koch, Arthur and Suellen Fachinetto.

Acknowledgements

Thanks to all the people and institutions that have helped to achieve this goal.

Thanks to professors Marcelo Brígido, Maria Emília, Maristela Holanda, Aletéia Araújo, Werner Treptow, Peter F. Stadler, Christoph Flamm, Fernanda Lima, Robert Miller, Sônia Freitas, João Batista Simão, Georgios J. Pappas Júnior, Valdir F. da Veiga Junior, Spartaco Astolfi Filho, Sérgio Lifschitz, Marlene De-Souza, Wagner Fontes, Nalvo Almeida, Ildinete Silva Pereira, Cynthia Maria Kyaw and many others.

Thanks to men and women of science Roberto C. Togawa, Tainá Raiol, J. L. Andersen, Gabriel S. C. Alves, and countless others.

Thanks to my colleagues and friends Daniel Souza, João Victor, Julien Jourde, André Lanna, Danilo Vilar, Polyane Wercelens, Daniel Saad, Emanuel Lôbo, Henrique Freitas, Gabriella Esteves, Andressa Valadares, Heidi Muniz, Ulisses Afonseca, Margaret Dick, Iasmini Lima, Marina Singarete, and many others.

Thanks to Instituto Federal de Goiás, CAPES, FAP-DF, University of Brasília, Leipzig Univertität, and Wien Universität including all the administrative staff.

Resumo

Comumente chamada de “copaíba”, a *Copaifera multijuga* Hayne (CmH) é uma planta do gênero *Copaifera* (*Leguminosae-Caesalpinoideae*) que ocorre na Amazônia brasileira. Extraído do tronco das árvores, o óleo-resina de *Copaifera spp.* é amplamente utilizado por povos indígenas da região amazônica na medicina popular e tem alto potencial associado a aplicações biotecnológicas como agente antimicrobiano, anti-inflamatório, antitumoral, antinociceptivo, antileishmanial e cicatrizante. O óleo-resina de *Copaifera spp.* é composto por ácidos resinosos e compostos voláteis, principalmente sesquiterpenos e diterpenos Neste trabalho, sesquiterpenos do óleo-resina da CmH, cenários biológicos para sua biossíntese e seus mecanismos químicos foram coletados de vários estudos. Com base nessa coleta de dados, em dados de um transcritoma da CmH e em métodos e ferramentas computacionais, foi reconstruída *in silico* uma rede metabólica de sesquiterpenos de *Copaifera multijuga* Hayne (CmH). Esta rede metabólica é uma compilação de reações enzimáticas cobrindo mecanismos de ciclização, compostos preditos e cenários biológicos para a biossíntese. Os resultados foram convenientemente armazenados em um banco de dados em grafos projetado especificamente para esta finalidade, tornando-se localizáveis, acessíveis, interoperáveis e reutilizáveis. O *workflow* utilizado para a reconstrução *in silico* funciona para múltiplos organismos, bem como pode ser adaptado para diferentes tipos de mecanismos químicos alterando o conjunto de regras de gramática de grafos.

Palavras-chave: *Copaifera multijuga*, rede metabólica, sesquiterpenos, banco de dados

Abstract

Ordinarily named “copaiba”, the *Copaifera multijuga* Hayne (CmH) is a plant of *Copaifera* genus (*Leguminosae-Caesalpinoideae*) occurring in the Brazilian Amazon. Exuded from the trunk of trees, the oil-resin of *Copaifera spp.* is widely used by indigenous people from the Amazon region for healing and in folk medicine, and it has high associated potential biotechnological applications, such as antimicrobial, anti-inflammatory, antitumor, antinociceptive, antileishmanial and healing. The oil-resin of *Copaifera spp.* is composed of resinous acids and volatile compounds, mainly sesquiterpenes and diterpenes In this study, a range of CmH oil-resin sesquiterpenes, biological scenarios for their biosynthesis, and its chemical mechanisms were collected from several studies. Based on this data collection, on CmH transcriptome data, and on computational methods and tools, an *in silico* sesquiterpene metabolic network of *Copaifera multijuga* Hayne (CmH) was reconstructed. The resulting sesquiterpene metabolic network of CmH is a compilation of reactions covering cyclization mechanisms, predicted compounds, and biological scenarios for the biosynthesis. These results were suitably stored in a graph database designed for it, and they became findable, accessible, interoperable, and reusable. The workflow for the *in silico* reconstruction can be used for multiple organisms as well as graph grammar rules can be added or removed to achieve different types of chemical mechanisms.

Keywords: *Copaifera multijuga*, metabolic network, sesquiterpenes, database

Contents

List of Figures

List of Tables

Acronyms

AFIR Artificial Force-Induced Reaction.

cDNA complementary DNA.

CmH *Copaifera multijuga* Hayne.

CPP Copalyl Diphosphate.

CRUD Create, Read, Update, and Delete.

DBMS Database Management Systems.

DFBA Dynamic Flux Balance Analysis.

DMAPP Dimethylallyl Pyrophosphate.

DNA Deoxyribonucleic Acid.

DPO Double Pushout.

EC Enzyme Commission Code.

FAIR Findability, Accessibility, Interoperability, and Reuse of digital assets.

FBA Flux Balanced Analysis.

FPP Farnesyl Diphosphate.

GGPP Geranylgeranyl Diphosphate.

GML Graph Modeling Language.

GO Gene Ontology.

GPP Geranyl Diphosphate.

GRAPHED Graph Description Diagram for Graph Databases.

HoC height of chest.

HTS High-Throughput Sequencing Technologies.

IBGE Institute of Geography and Statistics.

IPP Isopentenyl Pyrophosphate.

IUBMB International Union of Biochemistry and Molecular Biology.

KEGG Kyoto Encyclopedia of Genes and Genomes.

MACiE Mechanism, Annotation and Classification in Enzymes.

MEP Methylerythritol phosphate pathway.

MVA Mevalonate pathway.

NCBI National Center for Biotechnology Information.

NGS Next Generation Sequencing.

NoSQL Not Only Structured Query Language.

NPP Nerolidyl Diphosphate.

NRP Non-ribosomal peptides.

OPP Diphosphate.

PEVS System of Plant Extraction and Silviculture.

PKS Polyketides.

PUBMED U.S. National Library of Medicine.

SMILES Simplified Molecular-Input Line-Entry System.

SQL Structured Query Language.

TPS Terpene synthase.

TPSs Terpene synthases.

Chapter 1

Introduction

Copaifera genus plants (*Leguminosae-Caesalpinoideae*), ordinarily named "Copaíba", grow abundantly in Brazil and several other countries in South America. *Copaifera multijuga* Hayne (CmH) is one of the *Copaifera* genus plants, being a widespread native species in the Amazon, not endemic to Brazil, but occurring throughout its northern region. Exuded from the trunk of trees, the oil-resin of *Copaifera spp.* is widely used by indigenous people from the Amazon region for healing, and in folk medicine (Junior and Pinto, 2002).

In addition, the oil-resin has associated biotechnological potential for applications such as antimicrobial (Dos Santos et al., 2008; Mendonça and Onofre, 2009a; Pacheco et al., 2006), antifungal (Deus et al., 2011, 2009), anti-inflammatory (Brito et al., 2005; de Matos Gomes et al., 2010; Veiga et al., 2006b, 2007), antitumor (Gomes et al., 2008; Lima et al., 2003b), antinociceptive (de Matos Gomes et al., 2010; Gomes et al., 2007), antileishmanial (Santos et al., 2008) and healing (Westphal et al., 2007). The oil-resin of *Copaifera spp.*, including CmH, is composed of resinous acids and volatile compounds, mainly sesquiterpenes and diterpenes (Leandro et al., 2012).

Terpenes are a large and varied group of natural products playing significant ecological roles, such as defense and communication, in addition to various applications in industry and medicine. They are produced by a range of organisms as plants, fungi and bacteria, through metabolic reactions catalyzed by Terpene synthases (TPSs) (Dewick et al., 2002).

The isoprene units (five carbons; $C5$), Isopentenyl Pyrophosphate (IPP) and Dimethylallyl Pyrophosphate (DMAPP) are the main substrates underlying the entire terpene diversity (Vattekkatte et al., 2018). Isoprene units ($C5$) give rise to Geranyl Diphosphate (GPP) ($C10$), Farnesyl Diphosphate (FPP) ($C15$), and Geranylgeranyl Diphosphate (GGPP) ($C20$), using chain elongation (Vattekkatte et al., 2018).

GPP is the substrate for monoterpenes, FPP for sesquiterpenes and $GGPP$ for diterpenes. However, FPP and $GGPP$ can also be dimerized to form the precursors of $C30$ and $C40$ terpenes (Wink, 2010). Depending on the amount of $C5$ isoprene units, terpenes are named as monoterpenes ($C10$), sesquiterpenes ($C15$), diterpenes ($C20$), sesterterpenes ($C25$), triterpenes ($C30$), tetraterpenes ($C40$) and polyterpenes ($\geq C40$) (Wink, 2010).

Terpene synthases (TPSs) exhibit a wide catalytic range and they may yield a variety of products from the same substrate (Schifrin et al., 2016; Tholl et al., 2005). There are two different classes of TPSs: Class I and Class II, defined by catalytically essential amino acid motifs (Chen et al., 2011) (Liu et al., 2014). FPP is the pivotal precursor of sesquiterpenes through the action of TPSs I (Zhang et al., 2016).

Regardless of the product, the biosynthesis of a sesquiterpene from FPP begins with the cleavage of Diphosphate (OPP), which is protonation-dependent on prevalently Mg^{2+} (Zhang et al., 2016). The resulting FPP cation may directly lead to the formation of terpenes, while another path leads to a rotation of farnesyl cation and a new OPP addition, forming cisoid or transoid Nerolidyl Diphosphate (NPP) (Tholl, 2006). The NPP can have the OPP cleaved, and the resulting NPP cation also can lead to the formation of terpenes (Tholl, 2006).

The mechanisms of sesquiterpenes synthesis involve $C-C$ bonds formation, cationic intermediates, Wagner-Meerwein rearrangements, carbocation capture by water and hydride, methyl, or allyl shifts caused by conformational changes of intermediate cations (Degenhardt et al., 2009; Schifrin et al., 2016; Tholl, 2006). In spite of all the variety of compounds resulting from the many combinations of cyclizations, it is known that all of them come from four initial cyclization groups: $C1-C10$, $C1-C11$, $C1-C6$, and $C1-C7$ (Christianson, 2017).

In silico metabolic networks are computational models comprising the biosynthesis reactions performed by a cell (Bazzani, 2014). They are the core of Systems Biology, and aim to model the interactions that occur during the biosynthesis of metabolites by an organism in a computationally manageable way. Computational techniques, literature review, and omics[1] data are employed to achieve this goal (Caspi et al., 2009; Wang et al., 2017).

The reconstruction of *in silico* metabolic networks is dependent on the amount and quality of available omics data (Le Novere, 2015). In general, *in silico* methods for reconstruction of metabolic networks infer a hypothetic metabolome from genome data and existing metabolic reactions databases (Wang et al., 2017). In non-model organisms, where the genome or transcriptome data are not so abundant, metabolic profiling is an alternative for metabolic networks reconstructions (Kell et al., 2005).

Another method for metabolic networks reconstruction is the prediction of compounds and reactions from computational simulations. Currently, there are approaches for generating these chemical networks through computational simulations such as the Artificial

[1] genomics, transcriptomics, metabolomics, among others

Force-Induced Reaction (AFIR) (Maeda et al., 2016) and ‘Modelling Pathways as Integer Hyperflows’ (Andersen et al., 2017).

Concerning the focus and level of detail, a metabolic network can be implemented for qualitative, quantitative or both objectives. Quantitative simulations in a metabolic network can estimate metabolite amounts, in which case one of the most used methods is the Flux Balanced Analysis (FBA) (Orth et al., 2010). The expected results of a qualitative perspective include, but are not limited to, the identification of enzymes and reactions, as well as other conditions that may influence the formation of a metabolome, like cellular or tissue location, and interaction with other biomolecules (Bazzani, 2014). The component detail level in a metabolic network can vary. Metabolic networks comprising chemical mechanisms can represent the steps of each reaction and its initial, intermediate and final compounds.

Metabolic networks can be stored in various file formats such as BioPax (Demir et al., 2010), RDF (RDF, 2014), and SBML (Hucka et al., 2003). Although structured files are versatile and functional, databases allow the management of more extensive and complex collections. There are many databases for metabolism such as KEGG (Kanehisa and Goto, 2000) or MetaCyc (Caspi et al., 2014). However, the reactions in these databases are in fact enzymatic reactions with specific multi-step chemical mechanisms (Andersen et al., 2013). While there is a considerable amount of knowledge about these multi-step chemical mechanisms in the literature, as in Christianson *et al.* (Christianson, 2017), Dickschat *et al.* (Dickschat, 2016), and Zhang *et al.* (Zhang et al., 2016), there are few initiatives with available specific repositories. Examples of such repositories are Jacob Blog (Jacobson, 2017), which brings a catalog of cyclization schemes and MACiE (Holliday et al., 2012), where the coverage of lyases[2] is approximately 6%.

In the context of the *in silico* reconstruction and storage of metabolic networks, a comprehensive and consistent data schema supports the FAIR Guiding Principles for scientific data management (Wilkinson et al., 2016). The scope of this work was defined with the purpose of supporting the generation and management of knowledge about the biosynthesis of sesquiterpenes by CmH. The method for the reconstruction of the CmH sesquiterpene metabolic network focuses on the mechanisms of FPP cyclizations using computational simulations, literature data, and a graph database to store and make it available.

[2] Lyases are the enzymes class to which most TPSs belong.

1.1 Problem

There was no available *in silico* metabolic network for the sesquiterpenes biosynthesis of *Copaifera multijuga* Hayne (CmH) comprising chemical mechanisms, and initial, intermediate and final compounds.

1.2 Objective

The objective of this thesis is to reconstruct, store and make available an *in silico* metabolic network for the biosynthesis of the sesquiterpenes present in the *Copaifera multijuga* Hayne (CmH) oil-resin comprising the chemical mechanisms.

The following specific objectives assist in achieving the primary objective:

- Generate a chemical network with the sesquiterpenes biosynthesis reactions and their chemical mechanisms, initial, intermediate and final compounds.
- Define and build a workflow for the *in silico* reconstruction of metabolic networks based on the generated chemical network.
- Define and implement a graph database to store the reconstructed metabolic network.
- Implement the workflow as a public and available computational tool.

Description of Chapters

Chapter 2 presents fundamental concepts and information on CmH oil-resin sesquiterpenes, terpene biosynthesis, *in silico* reconstruction of metabolic networks, generation of computational chemical networks, and graph databases. Chapter 3 describes the method used for the reconstruction and storage of the CmH metabolic network of sesquiterpenes. Chapter 4 presents the results and discussion where related works, limitations and perspectives are explored. Chapter 5 presents the conclusions and future work. Apendix I presents an expanded abstract in Portuguese.

Chapter 2

Background

This chapter presents the fundamental concepts driving this work. Section 2.1 introduces the *Copaifera multijuga* Hayne (CmH) and a summary of the identified and described sesquiterpenes in its the oil-resin. Section 2.2 describes the biological topics on sesquiterpenes biosynthesis. Section 2.3 presents some of the most important methods for *In silico* Metabolic Network Reconstruction. In addition, it introduces the generation of computational chemical networks based on graph grammar, and the NoSQL graph databases as an approach for storing biological data.

2.1 *Copaifera multijuga Hayne*

In Brazil and several other countries in South America, *Copaifera* genus plants from family *Leguminosae* subfamily *Caesalpinoideae*, ordinarily named 'Copaíba', grow abundantly (Costa, 2018; Veiga and Pinto, 2002). Copaíba plants are slow-growing trees living up to 400 years that can reach 25 to 40 meters in height, and their trunk is rough and dark (Junior and Pinto, 2002).

The taxonomic system for classification is not unified and Table 2.1 shows some different purposed taxonomies for *Copaifera* genus (Junior and Pinto, 2002).

Table 2.1: Taxonomic classifications of *Copaifera* genus.

	Engler	Cronquist	Redeflora
Family	*Leguminosae Juss.*	*Leguminosae Juss.*	*Fabaceae*
Subfamily	*Caesalpinodae Knt*	*Caesalpinodae R. Br.*	-
Genus	*Copaifera*	*Copaifera*	*Copaifera*

Copaifera multijuga Hayne (CmH) is a widespread native species in the Amazon, not endemic to Brazil, but it occurs throughout its northern region as shown in Figure 2.1.

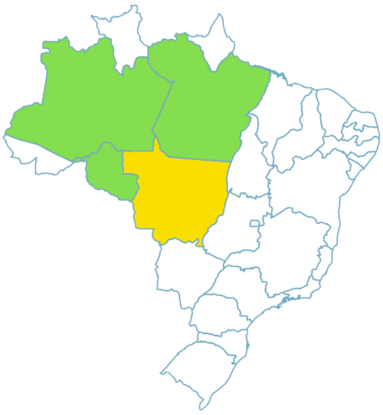


Figure 2.1: Regions where the *CmH* can be found in Brazil. (Source: Junior and Pinto (2002); Martins-da Silva (2006)).

According to the Manual of Seeds of the Amazon, Fascicle 9 (Brum et al., 2009), CmH can reach 60 m in height, with a diameter at the height of chest (HoC) of approximately 120 cm, which far exceeds the more typical diameter of 40 cm of HoC. The leaves measure 12 to 23 cm, with 8 to 20 elliptical and alternating leaflets. Trees are green most of the time except at the end of fruiting and before the next flowering. The leave shedding is often observed during the dry months (Brum et al., 2009). In Central Amazon, new leaves are produced at the same time as the flowering during the months of less sunshine and higher rainfall (Brum et al., 2009).

The flowers are white, mono-perianth, with four sepals measuring from 4 to 5 mm, with a red-rusty tone. They have ten stamens with long and glabrous fillets, inserted in the bud. Ovaries (up to 4 mm) are uniovulated, more or less globose, densely pilose, with a curved stylet and captioned stigma. The flowers exude nectar and produce large amounts of pollen (Brum et al., 2009). Bees, such as *Trigona spp.* and *Aphis mellifera* perform pollination (Rigamonte-Azevedo et al., 2004).

The fruits are vegetables, with an average size of 3.6 x 2.9 x 2.0 cm and an average weight of 8.5 g, initially yellow to red, becoming dark brown with maturation. They have a globoid shape, flattened laterally, with asymmetric base and a rounded an mucronate apex. The fruits open entirely through an invaginated longitudinal suture, and their pericarp is woody, slightly rough, and composed of two parts; internally it is yellowish-gray, fibrous and velvety (Brum et al., 2009).

Figure 2.1 shows a branch of CmH with fruits. Figure 2.3, adapted from (Brum et al., 2009), summarizes the phenology phases observed in the region of Manaus.

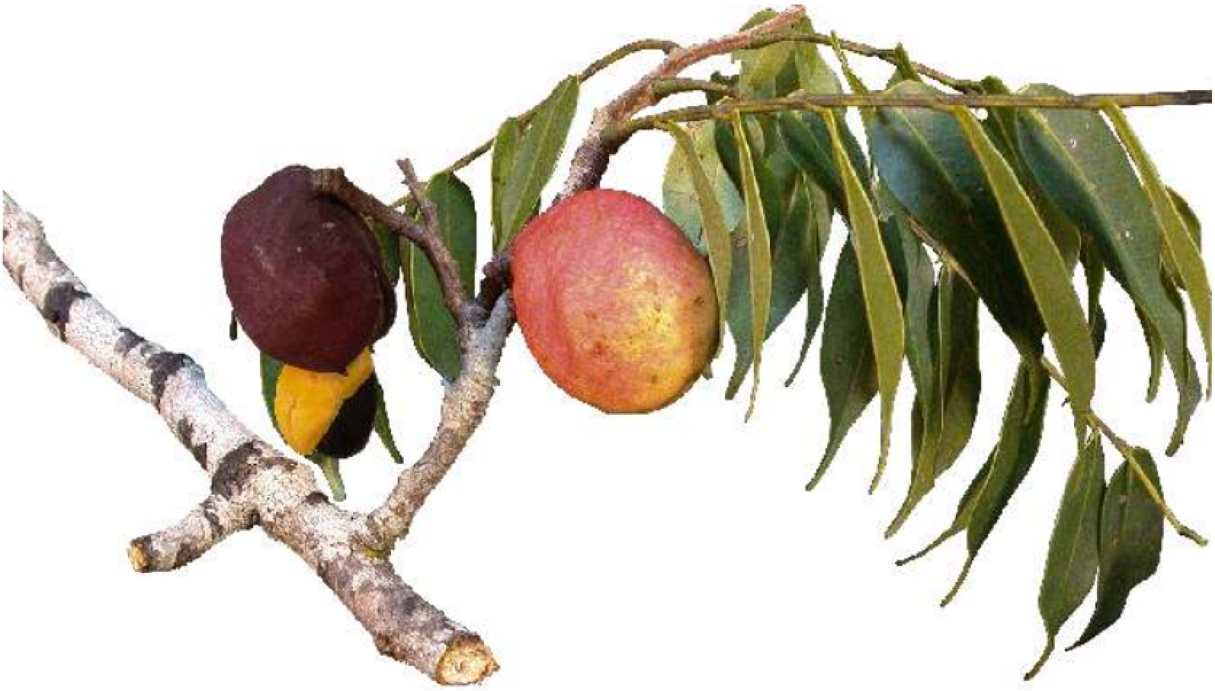


Figure 2.2: Branch of CmH with fruits. (Source: Brum et al. (2009)).

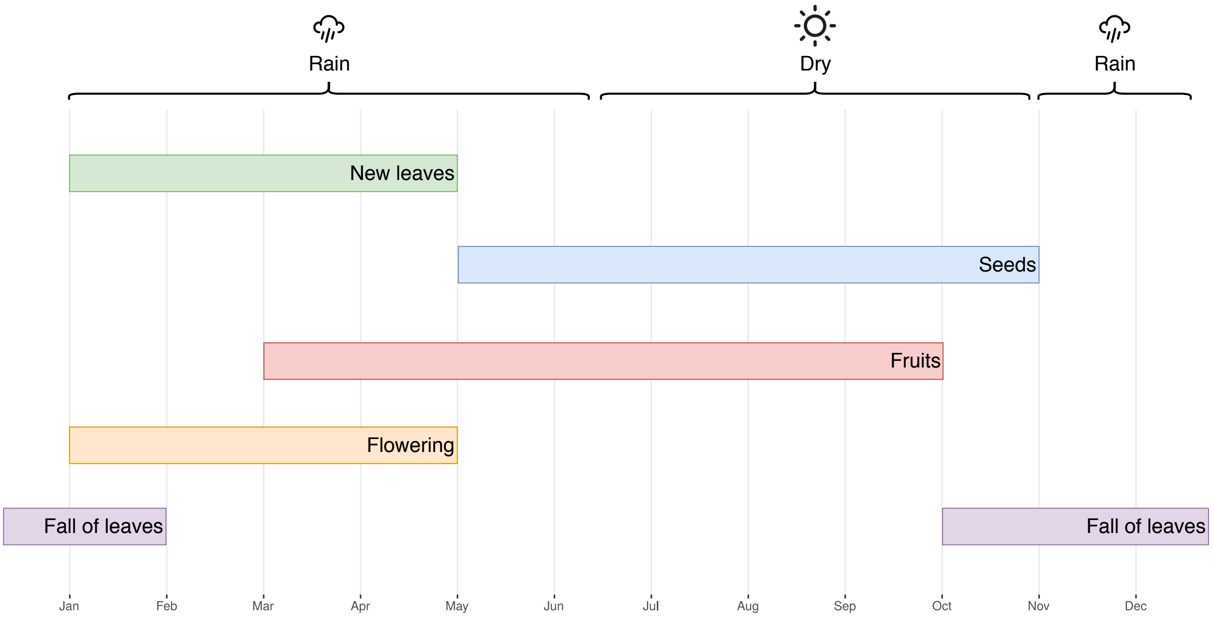


Figure 2.3: Phenology phases of CmH. (Source: Brum et al. (2009)).

The Brazilian Institute of Geography and Statistics (IBGE) provides statistical information on the quantity and value of the main products obtained through the process of exploitation of native forest resources through the System of Plant Extraction and Silviculture (PEVS) (IBGE, 2018). Figure 2.4 shows a time series of volume (in tonnes) and amount (in $R\$$) of Copaíba oil-resin extracted for commercial purposes from Brazilian forests according to PEVS. Looking at the data since 2013, although production is relatively stable, the value of the product has increased by approximately 74%.

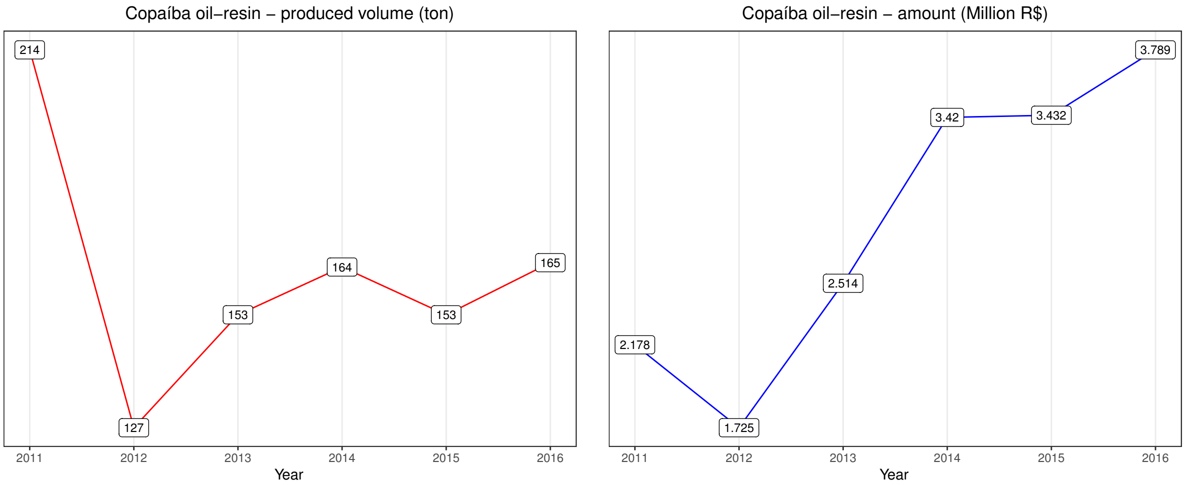


Figure 2.4: Produced volume (in tonne) and amount (in $R\$$) of Copaíba oil-resin extracted for commercial purposes from Brazilian forests. (Source: IBGE (2018)).

Sesquiterpenes of *Copaifera multijuga Hayne*

The oil-resin of *Copaifera spp.*, exuded from the trunk of the trees, is composed of resinous acids and volatile compounds mainly sesquiterpenes and diterpenes (Leandro et al., 2012). In its diversity, although similar, the *Copaifera spp.* oil-resin varies in composition and applications (Veiga et al., 2007), being widely used by indigenous people from the Amazon region for healing and folk medicine (Veiga et al., 2006b).

Also, the oil-resin properties have been the subject of research in areas as healing (Leandro et al., 2012), anti-inflammatory (Brito et al., 2005; de Matos Gomes et al., 2010; Veiga et al., 2006a, 2007), antimicrobial (Deus et al., 2011, 2009; Dos Santos et al., 2008; Mendonça and Onofre, 2009b; Pacheco et al., 2006), antitumor (Gomes et al., 2008; Lima et al., 2003b), antinociceptive (de Matos Gomes et al., 2010; Gomes et al., 2007), and antileishmanial applications (Santos et al., 2008).

In this work, the diversity and amount of the identified CmH sesquiterpenes in oil-resin from the results of several studies as (Lima et al., 2003a), (Veiga et al., 2006a), and (Junior et al., 2007) was summarized, as shown in Table 2.2. Figures 2.5, 2.6, 2.7, and 2.8 show the chemical structure and the average concentration percentages for respectively the polycycle, tricycle, bicycle, and mocycle sesquiterpenes identified in CmH oil-resin.

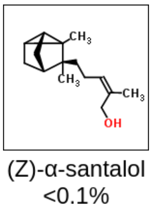


Figure 2.5: Policycle sesquiterpenes found in oil-resin of CmH supplemented by their average percentages.

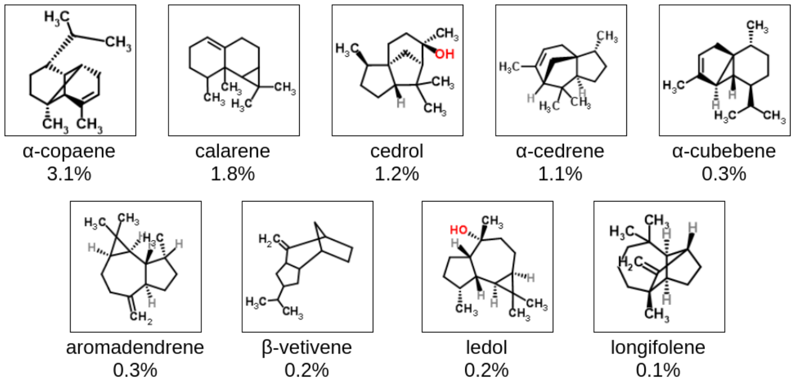


Figure 2.6: Tricycle sesquiterpenes found in oil-resin of CmH supplemented by their average percentages.

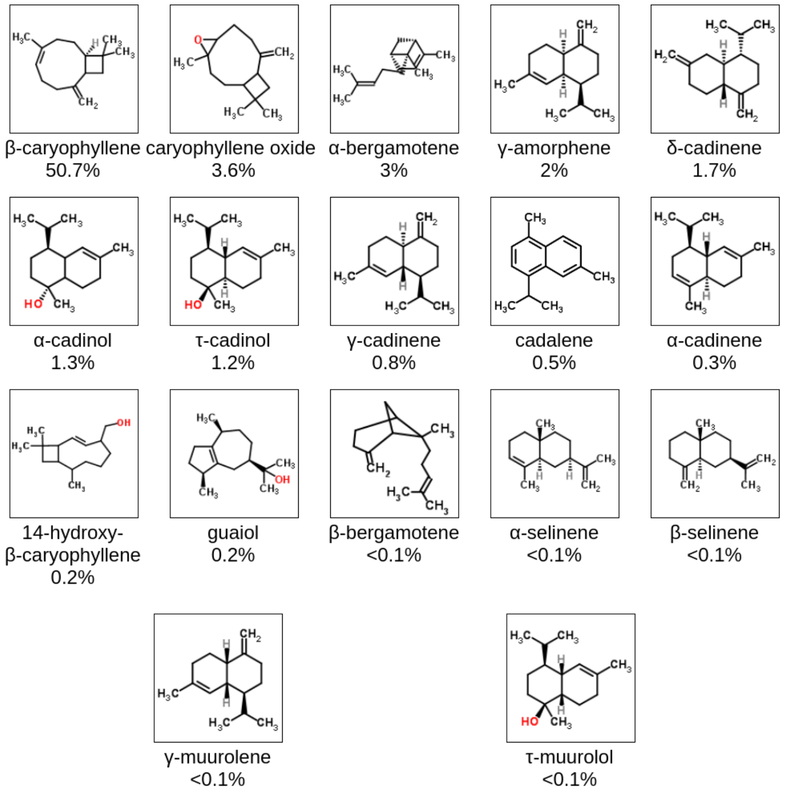


Figure 2.7: Bicycle sesquiterpenes found in oil-resin of CmH supplemented by their average percentages.

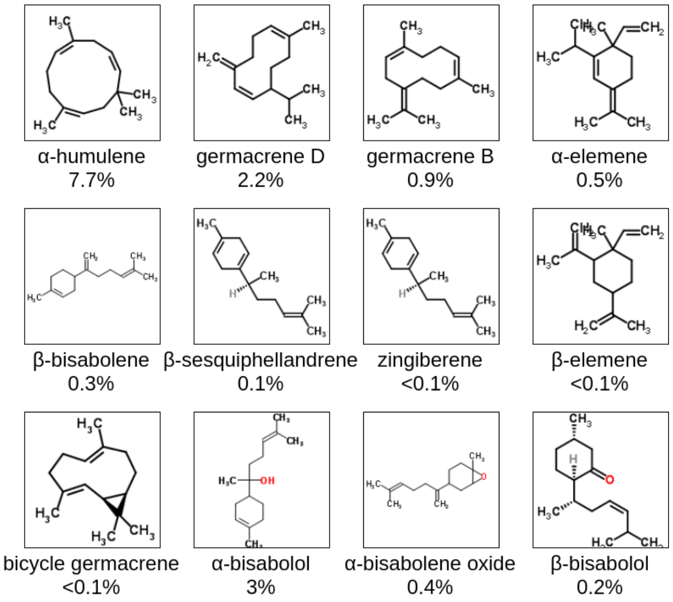


Figure 2.8: Monocycle sesquiterpenes found in oil-resin of CmH supplemented by their average percentages.

Table 2.2: Diversity and amount of the identified CmH sesquiterpenes in oil-resin.

Sesquiterpene	(Lima et al., 2003a)	(Veiga et al., 2006a)	(Veiga et al., 2006a)	(Junior et al., 2007)	Average percentage
β-caryophyllene	57.46	29.6	58.4	57.5	50.7
α-humulene	8.28	5.7	8.4	8.3	7.7
caryophyllene oxide	0.54	13	0.5	0.5	3.6
α-copaene	2.51	5	2.5	2.5	3.1
α-bergamotene	2.58	4.4	2.6	2.6	3.0
epi-α-bisabolol	0	3	0	0	3.0
germacrene D	2.42	1.5	2.5	2.4	2.2
α-caryophyllenol	0.74	5.8	0.8	0.7	2.0
γ-amorphene	1.88	2.3	1.9	1.9	2.0
calarene	0.3	5.3	1.1	0.3	1.8
δ-cadinene	1.67	1.9	1.7	1.7	1.7
α-cadinol	0	2.2	0.4	0	1.3
τ-cadinol	0	1.2	0	0	1.2
cedrol	0.37	3.6	0.4	0.4	1.2
α-cedrene	1.12	0	0	1.1	1.1
14-hydroxycaryophyllene	0	1	0	0	1.0
germacrene B	0.98	0.7	1	1	0.9
γ-cadinene	0.58	1.3	0.6	0.6	0.8
α-elemene	0.34	1	0	0.3	0.5
cadalene	0.36	0.9	0.4	0.4	0.5
α-bisabolene oxide	0.42	0	0	0.4	0.4
α-cadinene	0.24	0.7	0.2	0.2	0.3
α-cubebene	0.3	0.4	0.3	0.3	0.3
aromadendrene	0.15	0.6	0	0.2	0.3
β-bisabolene	0.33	0	0	0.3	0.3
β-bisabolol	0.09	0.7	0.1	0.1	0.2
acetoxy-caryophyllene	0.23	0	0	0.2	0.2
β-vetivene	0.12	0.5	0.1	0.1	0.2
ledol	0	0	0.2	0.2	0.2
guaiol	0.19	0	0.2	0.2	0.2
β-sesquiphellandrene	0.1	0.2	0.1	0.1	0.1
longifolene	0.11	0	0	0.1	0.1
Total of sesquiterpenes	84.41	92.5	84.4	84.6	92.47

2.2 Sesquiterpenes biosynthesis

The metabolites produced in a cell come from metabolic reactions of biosynthesis or degradation that transform chemical compounds (substrates) into other chemical compounds (products) (Nelson et al., 2008), (Michal and S., 2012). According to Harris (Harris, 2013), some criteria are required to classify a compound as a metabolite:

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

The genome encodes information to produce particular proteins that catalyze the metabolic reactions and which are called enzymes. Reactions often require a cofactor, which is an additional inorganic or organic molecule, in this last case, it is a coenzyme (Fischer et al., 2010; Nelson et al., 2008). The set of metabolic reactions that are essential for an organism, e.g., cellular respiration, comprises the primary metabolism (Michal and S., 2012). The secondary metabolism includes a set of metabolic reactions, which are not essentially necessary to the organism and tends to be specific over species.

The biosynthesis of secondary metabolites is a highly coordinated process that includes the formation of a 'metabolome'[1] and a secondary metabolic network. The metabolome is the quantitative complement of low-molecular-weight metabolites present in cells under a set of physiological conditions (Kell et al., 2005; Oliver et al., 1998). This network may also require different types of cells, using a range of cell compartmentalization features, particularly in plants, both to ensure specific biosynthesis and to prevent interference from extraneous molecules during the process (Wink, 2010).

According to Keller (Keller et al., 2005), the classes of secondary metabolites are Polyketides (PKS), Non-ribosomal peptides (NRP), Alkaloids and Terpenes. Among the secondary metabolites, terpenes are a large group of metabolites playing significant ecological roles, and 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 (Jørgensen et al., 2005).

The canonical metabolic pathway for the biosynthesis of terpenes is the Mevalonate pathway (MVA), which can produce IPP and DMAPP from acetyl-CoA (Lombard and

[1] First use of the term metabolome in the literature: (Oliver et al., 1998)

Moreira, 2011). MVA pathway was first identified in yeast and mammals in the 1950s (Lombard and Moreira, 2011). Since then, it was believed to be the only pathway for the formation of terpenes until the 1990s, when the Methylerythritol phosphate pathway (MEP), capable of generating IPP and DMAPP from D-glyceraldehyde-3-phosphate ($G3P$) and pyruvate, was discovered in bacteria and plants (Lombard and Moreira, 2011).

The MVA pathway generally occurs in eukaryotes, archaea, and the upper plant cytosol while the MEP pathway occurs in eubacteria, green algae (Lohr et al., 2012), and chloroplasts from higher plants (Kuzuyama, 2002). However, there are exceptions for both pathways, among them, a group of eukaryotes with plastids (*Apicomplexa* and some photosynthetic eukaryotes) which synthesize IPP and DMAPP from the MEP pathway (Lange et al., 2000). In addition, the MEP pathway may contribute to cytosolic biosynthesis of sesquiterpenes by allowing IPP to be shuttled from plastids to cytosol (Arimura et al., 2008).

The biosynthesis of terpenes can be understood in three stages. The first stage is the formation of the C_5 units of isoprene: IPP or DMAPP from MVA or MEP pathways. In the second stage, C_5 units of isoprene are condensed to generate three higher prenyl diphosphates, geranyl diphosphate (GPP: $C10$), farnesyl diphosphate (FPP: $C15$) and geranylgeranyl diphosphate ($GGPP$: $C20$). In the third stage, GPP, FPP, and $GGPP$ undergo Terpene synthases (TPSs) catalysis through a wide range of changes in bonding, hybridization, and rearrangements on the carbon skeletons during a cyclization cascade initiated by the formation and propagation of highly reactive carbocation intermediates (Christianson, 2006). Figure 2.9 presents the three stages for the biosynthesis of terpenes and a scheme for plant MEP and MVA pathways.

Depending on the amount of $C5$ units, terpenes are named as monoterpenes or iridioids ($C10$), sesquiterpenes ($C15$), diterpenes ($C20$), sesterterpenes ($C25$), triterpenes ($C30$), tetraterpenes ($C40$) and polyterpenes ($\geq C40$) (Wink, 2010). GPP gives rise to monoterpenes, FPP to sesquiterpenes and GGPP to diterpenes. However, FPP and GGPP can also be dimerized to form the precursors of C_{30} and C_{40} terpenoids.

TPSs are the enzymes that catalyze the reactions involved in the biosynthesis of terpenes. They exhibit a wide catalytic range and they may yield a variety of products from the same substrate (Schifrin et al., 2016; Tholl et al., 2005). The formation of multiple products can be related to many aspects as the substrate geometry (Hong and Tantillo, 2009; Vattekkatte et al., 2015) or minor changes in TPSs structures, which can result in different carbocation intermediates (Schifrin et al., 2016; Singh and Sharma, 2015). Even in closely related TPSs genes, their expression may be influenced by factors such as tissue-specificity (Chen et al., 2003, 2004; Ro et al., 2006; Tholl, 2006), or induced response (Mercke et al., 2004; Schnee et al., 2006; Yuan et al., 2008).

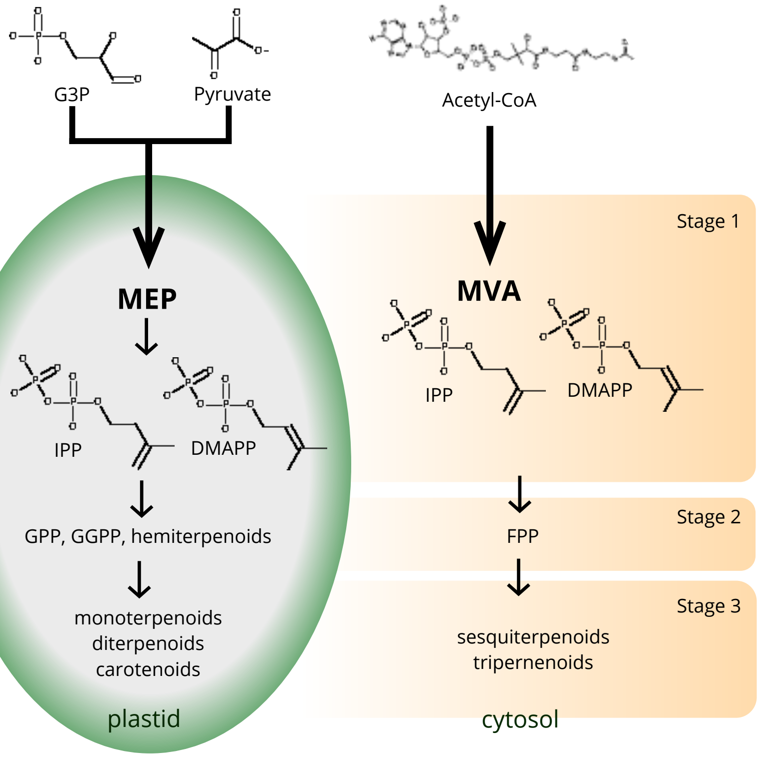


Figure 2.9: Three stages of plant MEP or MVA terpenoid biosynthesis pathways.

There are two different classes of TPSs: Class I and Class II, defined by catalytically essential amino acid motifs (Chen et al., 2011; Liu et al., 2014). TPSs I convert linear, all-trans, isoprenoids, geranyl (C10)-, farnesyl (C15)-, or geranylgeranyl (C20)-diphosphate into numerous varieties of monoterpenes, sesquiterpenes, and diterpenes. The TPSs I bind their substrate by coordination of a trinuclear divalent metal ion catalytic site (generally a Mg^{2+}), consisting of a central cavity formed by mostly antiparallel α-helices. This catalytic site has an aspartate-rich $DDxxD/E$ motif, and often another NSE/DTE motif in the C-terminal portion (Kempinski et al., 2015; Lesburg, 1997; Oldfield and Lin, 2012).

TPSs II act by triggering GGPP protonation which results in successive carbocations and cyclizations to form, for example Copalyl Diphosphate (CPP) (Liu et al., 2014; Oldfield and Lin, 2012). The $DxDD$ motif of the TPSs II (distinct from the TPS I $DDxxD/E$ motif) catalyzes the reaction, also using a Mg^{2+} cofactor to assist substrate binding and positioning (Gao et al., 2012).

Particularly for plants, and based on the homology of sequences, Bohlmann (Bohlmann et al., 1998) classified the TPSs into six subfamilies: *a*, *b*, *c*, *d*, *e*, and *f*. Nowadays, there are seven subfamilies, since the addition of TPS-g (Dudareva et al., 2003). This classification has been revised and updated as new discoveries come out, as in (Martin et al., 2004) and (Chen et al., 2011).

Under the evolutionary aspect, it is agreed that all plant terpene synthases share a common ancestor, and the TPSs speciation events are close related to the separation between angiosperms and gymnosperms (Martin et al., 2004; Singh and Sharma, 2015).

Like other proteins, the sequence of amino acid residues influences its three-dimensional structure (Liu et al., 2014). Mutations in these amino acids can affect their structure, and hence their function, causing changes in the efficiency, specificity, or concentration of their products (Hong and Tantillo, 2009; Keeling et al., 2008; Liu et al., 2014; Zhuang et al., 2012), (Chen et al., 2014).

In addition to conserved domains, the TPSs structure influences their specialization, since they may contain one or more of six main fold types: α, β, γ, δ, ϵ, and ζ, giving rise to structures of TPSs (Liu et al., 2014; Oldfield and Lin, 2012). Three of these domains, α, β and γ, are present in plant TPSs. A summary of the homology and structure based classifications of plant TPSs can be seen in Figure 2.10, where the TPS I family can have a,b,d, an e/f subfamilies. Also, TPS I family can have α or α and β structures. TPS II family have only subfamily c, but have α, β, and γ structures.

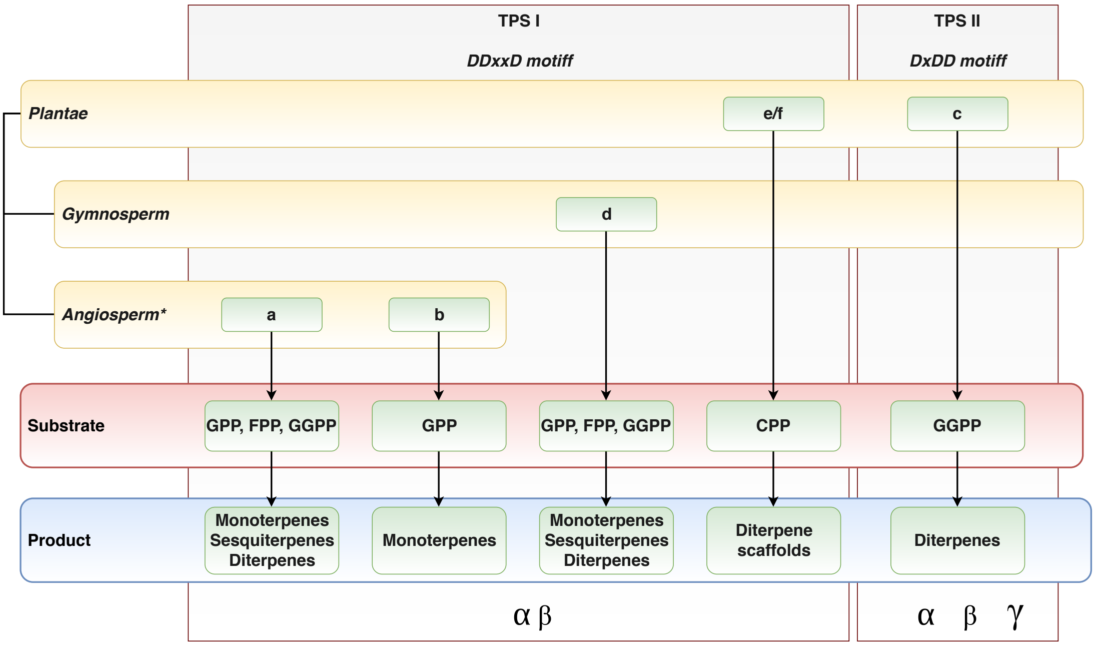
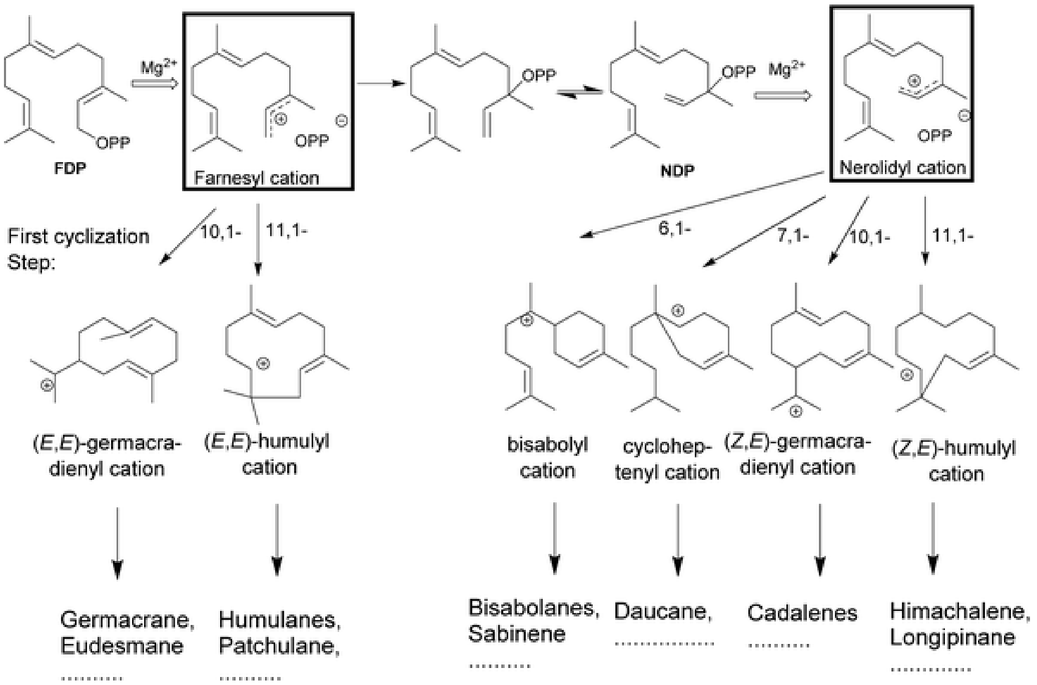


Figure 2.10: Plant terpene synthases classification.

FPP is the pivotal precursor of sesquiterpenes through of TPS I action (Zhang et al., 2016). Regardless of the product, the biosynthesis of a sesquiterpene from FPP begins with the cleavage of OPP, which is protonation-dependent pricipally on Mg^{2+} (Zhang et al., 2016). The resulting FPP cation may directly lead to the formation of terpenes, or it can pass through a rotation and a new OPP addition, forming cisoid or transoid NPP, from which the OPP is cleaved again (Tholl, 2006).

The mechanisms of sesquiterpenes synthesis formation involve $C - C$ bonds formation, cationic intermediates, Wagner-Meerwein rearrangements, carbocation capture by water and hydride, methyl, or allyl shifts caused by conformational changes of intermediates cations (Degenhardt et al., 2009; Schifrin et al., 2016; Tholl, 2006). In spite of all the variety of compounds resulting from the enormous amounts of combinations of cyclizations, it is known that all of them come from 4 initial cyclization groups: $C1 - C10$, $C1 - C11$, $C1 - C6$, and $C1 - C7$ (Christianson, 2017). The Figure 2.11 illustrates this scenario.



Mg2+
OPP
FDP
OPP
Farnesyl cation
OPP
OPP
Mg2+
NDP
OPP
Nerolidyl cation
First cyclization Step:
10,1-
11,1-
6,1-
7,1-
10,1-
11,1-
(E,E)-germacra-dienyl cation
(E,E)-humulyl cation
bisabolyl cation
cyclohep-tenyl cation
(Z,E)-germacra-dienyl cation
(Z,E)-humulyl cation
Germacrane, Eudesmane ....
Humulanes, Patchulane, ...
Bisabolanes, Sabinene ....
Daucane, ...
Cadalenes ....
Himachalene, Longipinane ....

Figure 2.11: Initial cyclizations from FPP. (Source: Vattekkatte et al. (2018)).

2.3 *In silico* Reconstruction of Metabolic Networks

In silico metabolic networks are the core of Systems Biology, and aim to model the interactions that occur during the biosynthesis of metabolites by an organism in a computationally manageable way. Computational techniques, literature review, and omics[2] data are employed to achieve this goal (Caspi et al., 2009). The reconstruction of *in silico* metabolic networks is dependent on the amount and quality of available omics data (Le Novere, 2015).

[2] genomics, transcriptomics, metabolomics, among others

A metabolic network comprising all possible reactions performed by a cell, is called genome-scale metabolic network (Bazzani, 2014). In general, methods for reconstruction of metabolic networks infer a hypothetical metabolome using genome data and existing repositories of metabolic reactions. These qualitative results have relevance for exposing the metabolic diversity among the organisms. Examples of these methods are PathFinder (Goesmann et al., 2002), Optstrain (Pharkya et al., 2004), *ab initio* reconstruction of metabolic pathways (Boyer and Viari, 2003), FUNGIpath (Grossetête et al., 2010) and PathwayTools (Caspi et al., 2014), among others.

After a qualitative reconstruction of a metabolic network, quantitative analysis can be performed under the chosen availability of external nutrients simulating the active part of the network (Bazzani, 2014). The objective of these simulations is to verify how efficiently the metabolic targets are synthesized assuming a steady-state where all produced metabolites need to be consumed in the cell. This approach simplifies the computational complexity of the mathematical problem since generally, simulations do not take into account the time during which the cell metabolism is a quasi-steady-state (Bazzani, 2014). Also, it is possible to introduce 'perturbations' in the system to compare the results (Bazzani, 2014). Examples of perturbations are the limitation of external nutrients, knock-out of genes encoding enzymes, and the introduction of inhibitors. Quantitative simulations in metabolic networks can estimate metabolite amounts, in which case one of the most commonly used methods is the Flux Balanced Analysis (FBA) (Orth et al., 2010). Exceeding the FBA approach, others initiatives such as Dynamic Flux Balance Analysis (DFBA), have been studied, enabling future developments towards whole-cell models, including spatio-temporally varying and multicellular systems (Øyås and Stelling, 2017).

In non-model organisms, where genomic or transcriptomic data are not so abundant, the use of metabolic profiling is an alternative for reconstructing metabolic networks qualitatively. Metabolic profiling is cheaper and has higher throughput than proteomics and transcriptomics, in addition, it favours the analysis of large numbers of samples (Kell et al., 2005). Changes in the metabolome vary widely relative to changes in the transcriptome and proteome, and are arguably more tractable numerically (Oliver et al., 1998).

Beyond the methods, it is necessary to design a comprehensive and consistent data model which copes with the challenge of efficiently storing a metabolic network, while also achieving the FAIR principles. Databases of metabolic pathways have been constructed since 1989 (Selkov et al., 1989b) through distinct methods, but only recently for storing metabolic networks as in (Silva et al., 2017) and (Fabregat et al., 2018). There are distinct approaches for storing metabolic networks, as summarized in Table 2.3.

The level of detail of a metabolic network varies with regard to the range and amount of data. As for metabolic reactions, for example, a network may or may not contain the chemical mechanisms and intermediate compounds of a reaction. MACiE (Holliday et al., 2012) is an example of a database of the chemical mechanisms of enzymatic reactions. Another issue is how to computationally reproduce the details of enzymatic reactions, leading to an exploration of Biochemical Networks.

Table 2.3: Strategies of storage of some of the main methods or databases related to metabolic networks.

	Factographic database	Structured files	Graphs	Relational Databases	Petri Network	Logic Programming	Not identified	Graph databases
Factographic data bank (Selkov et al., 1989a)	X							
Integrated database (Baher et al., 1992)						X		
Metabolic knowledge (Karp and Paley, 1994)		X						
Petri net (Reddy et al., 1993)					X			
Gaasterland and Selkov (Gaasterland and Selkov, 1995)						X		
KEGG (Kanehisa and Goto, 2000)							X	
ERATO (Hucka et al., 2001)		X						
PathFinder (Goesmann et al., 2002)			X					
Optstrain (Pharkya et al., 2004)							X	
Ab initio reconstruction (Boyer and Viari, 2003)							X	
Genome-scale reconstruction (Förster et al., 2003)		X						
Petri Net in systems biology (Pinney et al., 2003)					X			
Yeast reconstruction (Herrgård et al., 2008)		X						
FUNGIpath (Grossetête et al., 2010)				X				
FARM (Dreyfuss et al., 2013)		X		X				
2Path (Silva et al., 2017)								X
Reactome (Fabregat et al., 2018)								X

2.3.1 Biochemical Networks

Molecules as Undirected Graphs

A convenient and natural way to model chemical molecules is using undirected graphs. Substrates or products of metabolic reactions are molecules, which can be represented as undirected graphs where the vertices denote atom types, and the number of edges between the vertices denotes bond types.

An undirected graph G can be defined as an ordered pair $(V,\ E)$ consisting of (Bondy et al., 1976):

(i) a nonempty set of vertices V;

(ii) a set of edges E disjoint of V;

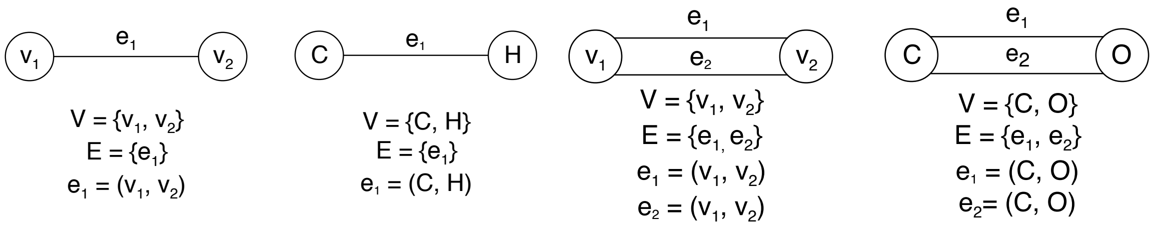
An edge can be represented by $e_i = (v_j,\ v_k)$. Example of undirected graph:

$$V = \{v_1,\ v_2,\ v_3\}$$

$$E = \{e_1, e_2\}$$

$$e_1 = (v_1,\ v_2), e_2 = (v_2,\ v_3)$$

Figure 2.12 shows an example of this abstraction for representing molecules as undirected graph. Figure 2.12a shows the set of vertices V composed by v_1 and v_2. The set of edges E is composed only by e_1. A Carbon (C) and a Hydrogen (H) atoms can be represented labeling the elements of V. The single bond between C and H is represented by e_1. Figure 2.12b shows the set of vertices V composed by v_1 and v_2. The set of edges E is composed by e_1 and e_2. A Carbon (C) and Oxygen (O) atoms represented labeling the elements of V. The double bond between C and O is represented by the edges e_1 and e_2.



(a) Single edge representing a single bond. (b) Two edges representing double bonds.

Figure 2.12: Abstraction for representing molecules as undirected graph.

As a concrete example, Figure 2.13 shows a molecule of FPP represented as an undirected graph. Note that in this example the vertices are not drawn inside a circle, and the edges are not labeled. The labeled atoms C, H, P, and O represent respectively Carbon, Hydrogen, Phosphorus, and Oxygen are the vertices. The bonds between the atoms are the edges, emphasizing that single edges represent single bonds and double edges represent double bonds.

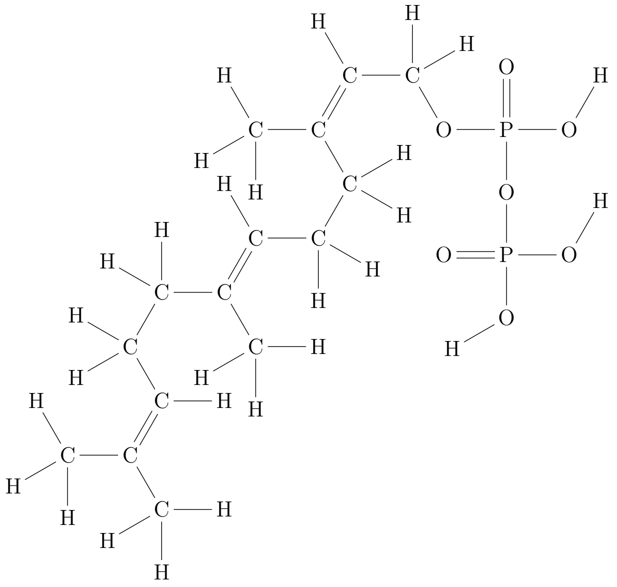


Figure 2.13: Molecule of FPP represented as an undirected graph.

Graph Grammar

Graph grammars, or graph rewriting systems, are proper generalizations of term rewriting systems (Andersen et al., 2013). Graph grammar rules provide a suitable rewriting formalism to express the feasible chemical transformations in molecules abstracted as undirected graphs. Using graph grammars rules, the undirected graphs that describe molecules can be transformed into other distinct undirected graphs describing distinct molecules. Structural changes in molecules during chemical reactions can be modeled as graph rewriting rules while preserving the fundamental chemical principles like mass conservation, atomic types, and cyclic shifts of electron pairs (Andersen et al., 2013).

Double Pushout (DPO) (Löwe, 1993) is a particular rewriting formalism for graph transformations that covers changes of undirected graph molecules in a rather explicit and detailed way (Andersen et al., 2013). The DPO considers transformation rules of form $p = (L \xleftarrow{l} K \xrightarrow{r} R)$ where L, R, and K are called the left graph, right graph, and context graph, respectively. The map l is graph morphism $l : K \to L$ and r is a graph morphism $r : K \to R$.

The rule p transforms G to H, in symbols $G \overset{p,m}{\Longrightarrow} H$, if there is a pushout graph D and a 'matching morphism' $m{:}L \Rightarrow G$ such that Diagram 2.1 is valid:

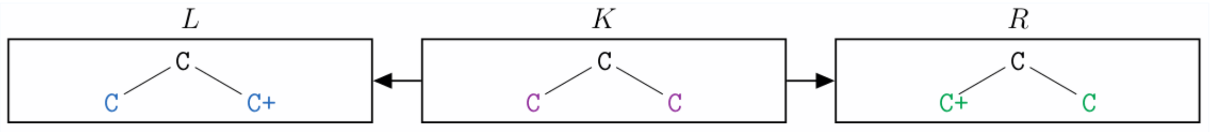
$$\begin{tikzcd}
L \arrow[d, "m"'] & K \arrow[l, "l"'] \arrow[r, "r"] \arrow[d] & R \arrow[d] \\
G & D \arrow[l] \arrow[r] & H
\end{tikzcd} \tag{2.1}$$

There is a ‘gluing condition’ which determines the existence of D whether the rule p is applicable to match in G. The ’gluing condition’ has two constraints:

1. There are not distinct elements x, y of L with $m(x) = m(y)$ and $y \notin l(K)$.
2. No edge e of $G : m(l)$ is incident to a node in $m(L : l(K))$.

The first constraint ensures the atoms’ conservation, while the second one ensures that a bond is not both in the context and in the transformed graph at the same time.

An example of a DPO rule $((L \xleftarrow{l} K \xrightarrow{r} R))$ representing a 1,3 hydrideshift is specified by three graph fragments in Graph Modeling Language (GML) format (Himsolt, 1997) as can be seen in the Figure 2.14:



(a) Graphical representation of the graph grammar rule.

```
# (Sandbeck, 2016) https://doi.org/10.1021/acs.joc.5b02553
rule [
 ruleID "1,3 hydride shift"
 left [
  node [ id  1 label "C+" ]
  node [ id  3 label "C" ]
 ]
 context [
  node [ id  2 label "C" ]
  edge [ source  1 target  2 label "-" ]
  edge [ source  2 target  3 label "-" ]
 ]
 right [
  node [ id  1 label "C" ]
  node [ id  3 label "C+" ]
 ]
]
```

(b) GML code for the graph grammar rule.

Figure 2.14: A graph grammar rule representing an 1,3 hidrid shift.

Metabolic Networks as Hypergraphs

Metabolic networks can be abstracted considering both multiple reactions and multiple chemical mechanisms in a single reaction. Both levels of abstraction comprise many-to-many relationships. One example of this is the KEGG reaction (R08695) for the β-farnesene formation. In the level of reaction, the FPP is converted into β-farnesene and OPP, as can be seen in the Figure 2.15.

However, in the level of mechanism, there are hidden intermediate molecules and chemical transformations. After the cleavage of the OPP, a FPP cation molecule is formed. Then, the FPP cation loses a H^+ giving the final product β-farnesene. This example of β-farnesene formation is shown in Figure 2.15 with their differences for both reaction and mechanism levels highlighted.

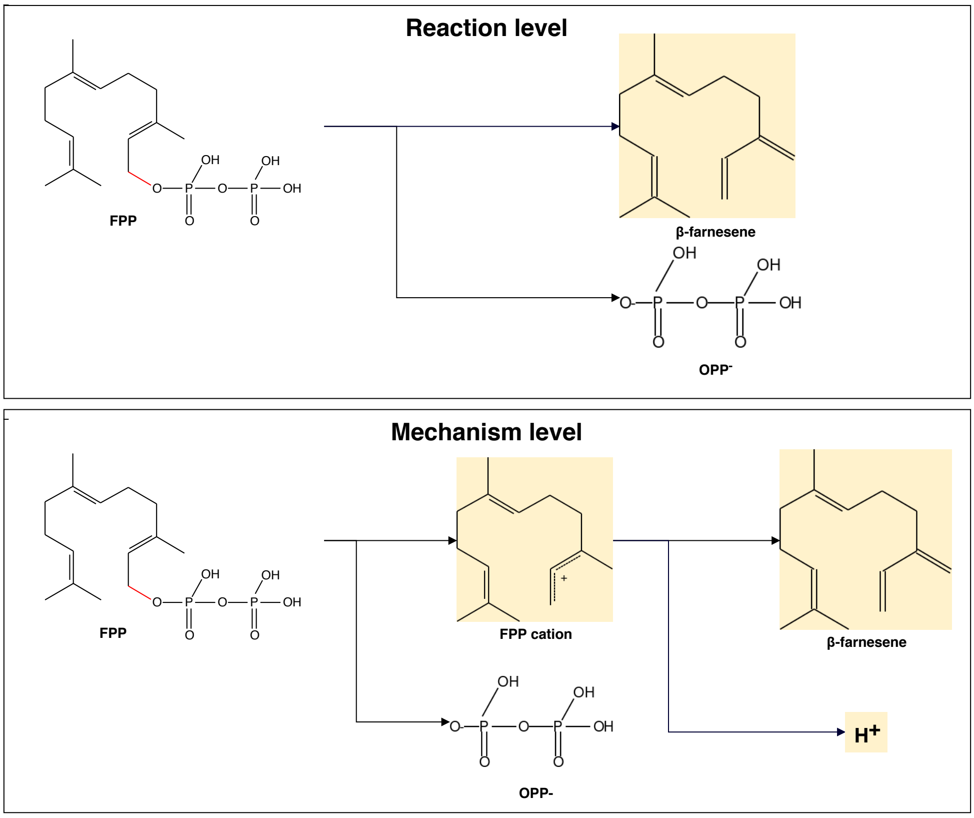


Figure 2.15: Reaction and mechanism levels of abstraction for the β-farnesene formation.

Directed hypergraphs are a suitable topological representation for metabolic networks in both reactions and chemical mechanisms levels. Directed hypergraphs can model both chemical reactions and their chemical mechanisms representing their many-to-many relationships with molecules through hyperedges (Andersen et al., 2017). A directed hypergraph is a hypergraph with directed hyperedges.

A hypergraph $\mathcal{H}$ can be defined as a pair $(\mathcal{V},\ \mathcal{E})$ consisting of (Gallo et al., 1993):

(i) a nonempty finite set of vertices $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$, with $V_i \subseteq V$ for $i = 1, \ldots, n$

(ii) a set of edges $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$, with $E_i \subseteq E$ for $i = 1, \ldots, m$.

The edges $\mathcal{E}$ in a hypergraph are called hyperedges. A directed hyperedge is an ordered pair, $E = (V_1, V_2)$, of disjoint subsets of $\mathcal{V}$; V_1 is the tail of E while V_2 is its head (Gallo et al., 1993). Figure 2.16 shows an example of directed hypergraph and its graphical representation.

In the context of molecules as undirected graphs (G), $\mathcal{V} = \{G_1, G_2, \ldots, G_n\}$. The chemical transformations occurring with the molecules are represented by $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$.

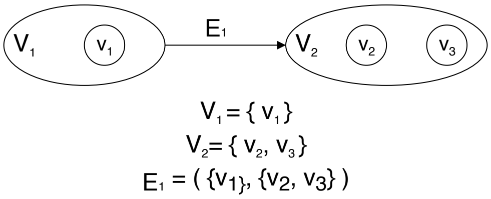
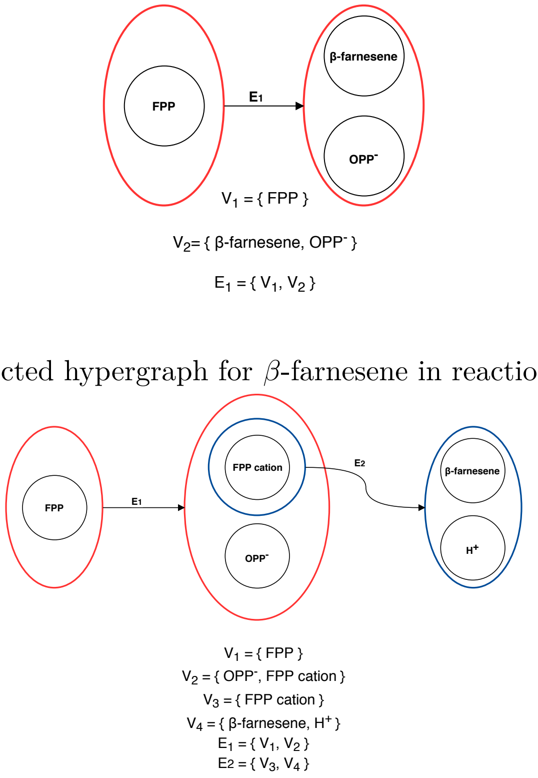


Figure 2.16: Example of a hypergraph.

Taking the example shown in Figure 2.15, both in reaction and mechanism levels, the compounds can be abstracted as vertices (represented as undirected graphs) and the transformations as hyperedges forming a directed hypergraph.

The Figure 2.17 shows the same example of the Figure 2.15 under the directed hypergraph perspective. Figure 2.17a graphically presents the abstraction of the directed hypergraph for β-farnesene in reaction level. Figure 2.17b graphically presents the abstraction of the directed hypergraph for β-farnesene in chemical mechanism level.



(a) Directed hypergraph for β-farnesene in reaction level.

(b) Directed hypergraph for β-farnesene in chemical mechanism level. There is a dependency between the highlighted red and blue sets.

Figure 2.17: β-farnesene formation abstracted as a directed hypergraph for both reaction and mechanism levels.

Representing chemical networks with MedØlDatschgerl

The framework MedØlDatschgerl (Andersen et al., 2016) provides a set of tools to generate and investigate large chemical networks. In MedØlDatschgerl, each molecule is handled as an undirected graph where an atom is a vertex, and a bond is an edge. Then, based on DPO graph grammar rules, the molecule graphs can be rewritten when they match partially, or totally, a rule simulating a feasible chemical mechanism (Andersen et al., 2013) (See Diagram 2.1).

A chemical universe of molecules can be reached from a set of start compounds by iterative application of a finite number of reactions (Andersen et al., 2014). Then, a derivation graph can be generated by successively applying the rules to an initial set of molecules and their successive and cumulative results. The resulting chemical reaction network of a simulation is a directed hypergraph, which preserves the atoms' type, mass, and charge (Andersen et al., 2016).

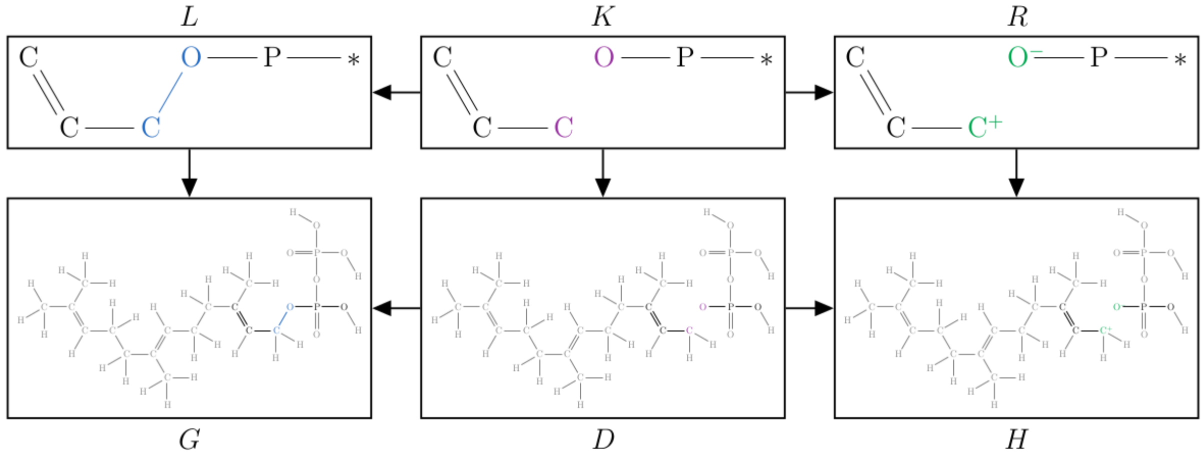
In addition, MedØlDatschgerl can work with Simplified Molecular-Input Line-Entry System (SMILES) format (Minkiewicz et al., 2017), allowing the chemical structures to be represented unambiguously and in a manner that permits automated reasoning (Andersen et al., 2016). The FPP molecule, for instance, can be represented as the SMILES string:“$CC(= CCCC(= CCCC(= CCOP(= O)(O)OP(= O)(O)O)C)C)C$”. Figure 2.18 shows an example of a DPO graph grammar rule for diphosphate cleavage from FPP. Figure 2.18 also shows the match of the rule with the molecule and the subsequent rewriting of the original graph (FPP) to two other molecules: farnesyl cation and a diphosphate anion.

The MedØlDatschgerl can be driven using PyMØD, a Python 3 module for bindings library libMØD. It produces a hypergraph, which can be computationally exploited in-memory and conveniently exported in a pdf report. It is important to note that the resulting hypergraph is not stored after execution of the program.

2.3.2 NoSQL Graph Databases

NoSQL databases have occupied significant space in managing large volumes of data in many areas including omics. This can be explained by the need for solutions for managing High-Throughput Sequencing Technologies (HTS) or Next Generation Sequencing (NGS) data. These are expected to meet simultaneously fast data access requirements through database systems, high-performance analytics tools, and efficient and interactive visualization capabilities for large volumes of data (De Brevern et al., 2015).

Although 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



```
rule [
 left [
  node [ id 16 label "O" ]
  node [ id  3 label "C" ]
  edge [ source 3 target 16 label "-" ]
 ]
 context [
  node [ id  1 label "C" ]
  node [ id  2 label "C" ]
  node [ id 17 label "P" ]
  node [ id 18 label "*" ]
  edge [ source  1 target  2 label "=" ]
  edge [ source  2 target  3 label "-" ]
  edge [ source 16 target 17 label "-" ]
  edge [ source 17 target 18 label "-" ]
 ]
 right [
  node [ id 16 label "O-" ]
  node [ id  3 label "C+" ]
 ]
]
```

Figure 2.18: Graph grammar rule and its application over a graph representing a molecule.

mechanisms for storing and retrieving data, whose modeling is an alternative to traditional relational databases. According to Corbellini (Corbellini et al., 2017), there are different types of NoSQL management systems for structured data storage systems, commonly classified as Key-Value; Wide Column or Column Families; Document-oriented; and Graph-oriented databases, henceforth, graph databases. Despite this classification, the NoSQL databases may be hybrids, that is, they can use more than one storage management model.

Graphs naturally describe a problem domain and graph databases assemble simple abstractions of vertices and relationships in connected structures, thus allowing models to be build and mapped more closely to the problem domain. Graph databases are Database Management Systems (DBMS) with Create, Read, Update, and Delete (CRUD) methods, which can store graphs natively or serializing the graph data into a different database model (Robinson et al., 2013). The schema in graph databases can store data in the vertices, and also in the edges, depending on the database (Silva et al., 2017).

A significant aspect of graph databases is the way they manage relationships between entities. It is similar to storing pointers between two objects in memory. Regarding queries performance, indexes can make the data recovery more efficient.

The lack of schema in NoSQL graph databases, despite offering flexibility, can also remove the interoperability pattern from the data (Lysenko et al., 2016). A graph database schema may positively influence the maintainability of the graph databases. It enables the study of the best graph schema for the data, and their relationships with regard to 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) (Van Erven et al., 2018) offers rich modeling diagrams for this purpose.

The performance and intuitiveness of queries in graph databases seems to be the main reason to use them as discussed in (Fabregat et al., 2018). Graph queries are more concise and intuitive compared with equivalent SQL queries, which are complicated by joins. In addition, the engine of the graph databases is different, which leads to another point for investigation regarding the relationship between an engine and performance.

According to the site db-engines.com, there are currently 29 graph database management systems. The site measures their popularity and updates a rank monthly. Figure 2.19 shows the ranking to June 2018.

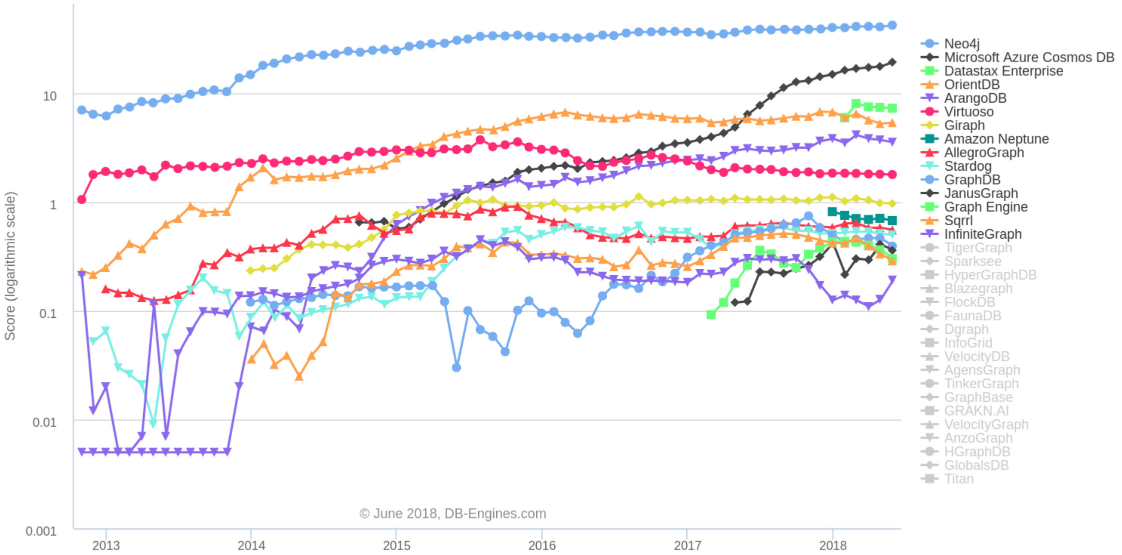


Figure 2.19: DB-Engines Ranking - trend of graph DBMS popularity. (Source: DB-Engines (2018)).

Graph Databases in Molecular Biology

With the advent of NoSQL databases, a fundamental question loomed: would the NoSQL databases be ready for Bioinformatics? Have and Jensen (Have et al., 2013) published a paper answering this question for NoSQL graph databases. In their work, they measured the performance of Neo4J v1.8 and PostgreSQL[3] v9.05 on STRING (Szklarczyk et al., 2017) data executing some operations. They found, for example, that the graph database was able to find the best scoring path between two proteins faster by a factor of almost 1000 times. Also, the graph database was able to find the shortest path, when constraining the maximal path length to two edges, 2441 times faster than the relational database. 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 (Pareja-Tobes et al., 2015) provides 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, by considering the intrinsic and extrinsic semantic features. Corbacho *et al* (Corbacho et al., 2013) used the Bio4J graph database for Gene Ontology (GO) analyzes in *Cucumis melo*. ncRNA-DB (Bonnici et al., 2014) is a database that integrates ncRNAs data interactions from a large number of well established on-line repositories built on top of the OrientDB. It is accessible through a web-based platform, a command-line, and a Cytoscape app called ncINetView.

Henkel *et al.* (Henkel et al., 2015) used the Neo4J to integrate the data from distinct system biology model repositories. This database offers to the community, curated and reusable models describing biological systems through queries in *Cypher Query Language*, the native query language of Neo4J.

Important software for the analysis and visualization of biological networks is Cytoscape[4]. The cyNeo4j plugin (Summer et al., 2015) was designed to link Cytoscape and Neo4j and enables the interactive execution of an algorithm by sending requests to the server.

Lysenko *et al.* (Lysenko et al., 2016) 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 (Balaur et al., 2016) uses Neo4J to store genetic and epigenetic events observed at different stages of colorectal cancer. The graph database enhanced the explo-

[3] https://www.postgresql.org

[4] www.cytoscape.org

ration of different queries related to colorectal tumor progression when compared to the primary source StatEpigen[5].

The Network Library (Summer et al., 2016) used Neo4J to integrate data from several biological databases through a clean and well-defined pipeline. 2Path (Silva et al., 2017) is a metabolic network implemented in Neo4J to manage terpenes biosynthesis data. It uses open data from several sources and was modeled to integrate important biological characteristics like the cellular compartmentalization of the reactions.

Biochem4j (Swainston et al., 2017) is another work that seeks integration of open data from different sources using Neo4J. It goes beyond a database and provides a framework for this integration and exploration of an ever-widening range of biological data sources.

GeNNet (Costa et al., 2017) is an integrated transcriptome analysis platform that uses Neo4J to unify scientific workflows by storing the results of the analysis.

BioKrahn (Messina et al., 2018) is a graph-based deductive and integrated database that takes advantage of the power of knowledge graphs and machine reasoning, to solve problems in the domain of biomedical science, such as interpreting the meaning of data from multiple sources or manipulated by various tools. It contains resources related to genes, proteins, miRNAs, and metabolic pathways.

Arena-Idb is a plataform for the retrieval of comprehensive and non-redundant annotated ncRNA interactions (Bonnici et al., 2018). It uses two different DBMS: a relational MySQL, and the graph database Neo4J, which is applied to handle the visualization of the networks in a web page.

Messaoudi (Messaoudi et al., 2018) evaluated the performance time needed for storing, deleting and querying the 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 reached best performances for queries when more in-depth levels of graph traversal were required.

Reactome (Fabregat et al., 2018) is a well established open-source, open-data, curated and peer-reviewed database of pathways, which recently adopted the graph database as storage strategy due to performance issues associated with queries traversing highly interconnected data. The adoption of graph database improved the queries reducing the average query time by 93%.

Table 2.4 summarizes the contributions of each reported work in this review.

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

Table 2.4: Contributions of graph-oriented databases for Molecular Biology.

Graph-oriented database	Main contribution	Other contributions	Source
Neo4J	Biological networks	Protein-protein interaction	(Have et al., 2013)
Neo4J	Gene annotation	GO analyses	(Corbacho et al., 2013)
Neo4J	Data integration	-	(Henkel et al., 2015)
Neo4J	Data integration	-	(Pareja-Tobes et al., 2015)
Neo4J	Biological networks	Diseases association	(Lysenko et al., 2016)
Neo4J	Cancer	Epigenetic events	(Balaur et al., 2016)
Neo4J	Data integration	-	(Summer et al., 2016)
Neo4J	Biological networks	**Metabolic networks**	(Silva et al., 2017)
Neo4J	Data integration	-	(Swainston et al., 2017)
Neo4J	Transcriptome analyses	-	(Costa et al., 2017)
grakn.ai	Data integration	Biomedical analyses	(Messina et al., 2018)
Neo4J/OrientDB	Biomedical analyses	-	(Messaoudi et al., 2018)
Neo4J	Biological networks	**Metabolic networks**	(Fabregat et al., 2018)
Neo4J	Biological networks	ncRNAs interactions	(Bonnici et al., 2018)

Chapter 3

Material and Methods

This chapter presents the material and the methods for the *in silico* reconstruction of the sesquiterpenes metabolic network of *Copaifera multijuga* Hayne (CmH). Section 3.1 presents the CmH target sesquiterpenes for this *in silico* reconstruction. Section 3.2 presents the graph grammar rules designed to reproduce the chemical mechanisms of the biosynthesis of these sesquiterpenes. It also presents the generation of the chemical network using these graph grammars rules. Section 3.3 presents the graph database schema for storing the generated metabolic network in the Neo4J. It presents the database update with the identification of the predicted compounds based on their chemical structure. Also, it describes the database update with biosynthesis scenarios from the literature. Section 3.4 shows the annotation of sesquiterpene synthases from CmH transcriptome[1] using information from the previous phases.

A chain of software executions is called workflow (Leipzig, 2016), and this method is described as a workflow where each step adds new information to the *in silico* metabolic network. Each information layer comes from a combination of human and computational interactions, making the workflow semi-automatic. The Figure 3.1 summarizes the a workflow.

3.1 Sesquiterpenes of *Copaifera multijuga Hayne*

The sesquiterpenes produced by CmH and the chemical mechanisms of their enzymatic biosynthesis reactions were identified from the literature. This first literature review provided a list of sesquiterpenes identified from the oil-resin of CmH. This complete list of sesquiterpenes and their respective abundance was presented in Table 2.2. In this work, a set of 27 compounds were taken among them for the *in silico* metabolic reconstruction.

[1] Unpublished data.

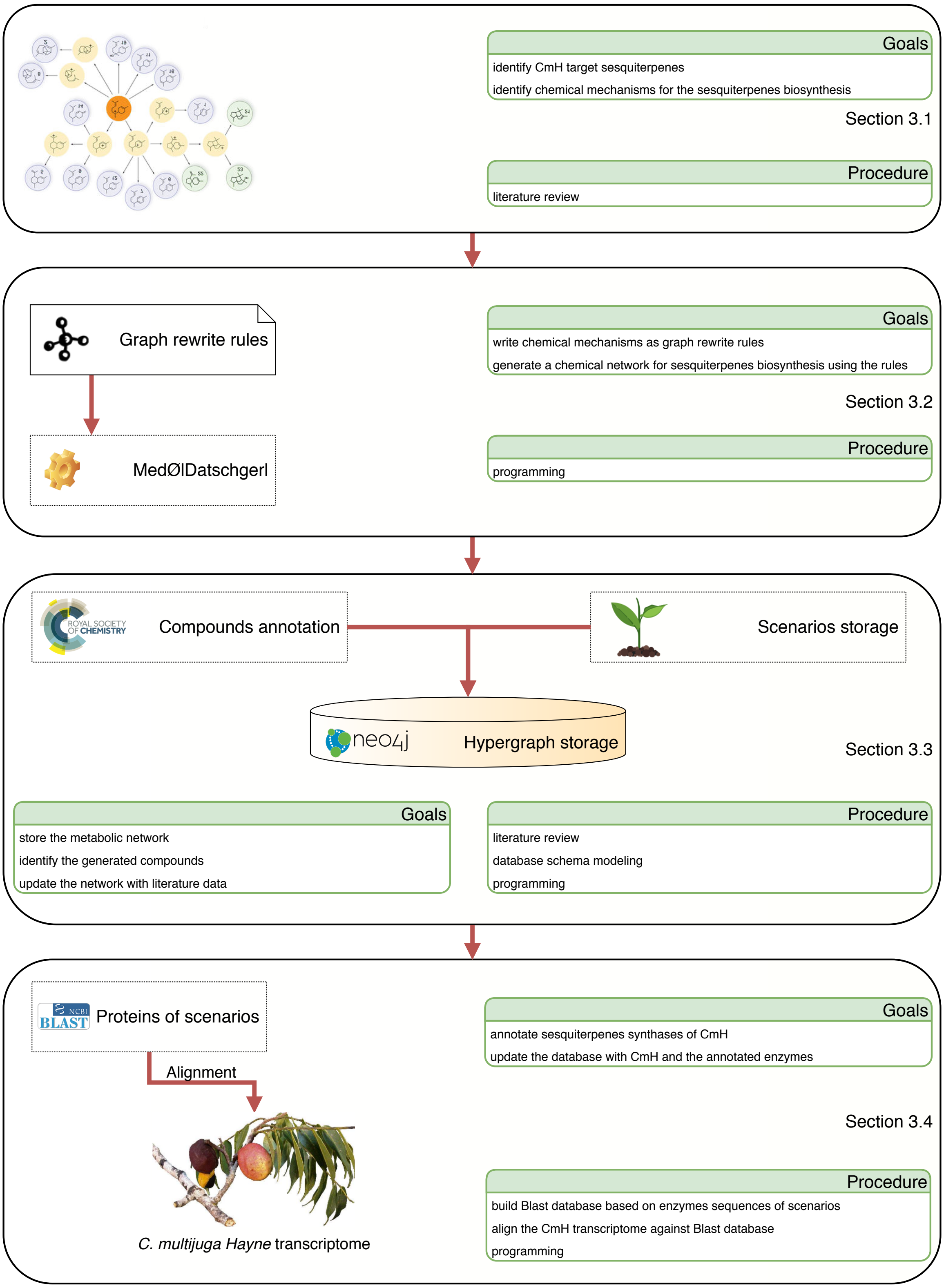

Goals
identify CmH target sesquiterpenes
identify chemical mechanisms for the sesquiterpenes biosynthesis
Section 3.1
Procedure
literature review
Graph rewrite rules
MedØlDatschgerl
Goals
write chemical mechanisms as graph rewrite rules
generate a chemical network for sesquiterpenes biosynthesis using the rules
Section 3.2
Procedure
programming
Compounds annotation
Scenarios storage
neo4j
Hypergraph storage
Section 3.3
Goals
store the metabolic network
identify the generated compounds
update the network with literature data
Procedure
literature review
database schema modeling
programming
Proteins of scenarios
Alignment
C. multijuga Hayne transcriptome
Goals
annotate sesquiterpenes synthases of CmH
update the database with CmH and the annotated enzymes
Section 3.4
Procedure
build Blast database based on enzymes sequences of scenarios
align the CmH transcriptome against Blast database
programming

Figure 3.1: Overview of the workflow for *in silico* reconstrucion of the sesquiterpenes metabolic network of *Copaifera multijuga* Hayne (CmH).

This set sum approximately 83% of the CmH oil-resin and it is highlighted by the red border in the Figure 3.2.

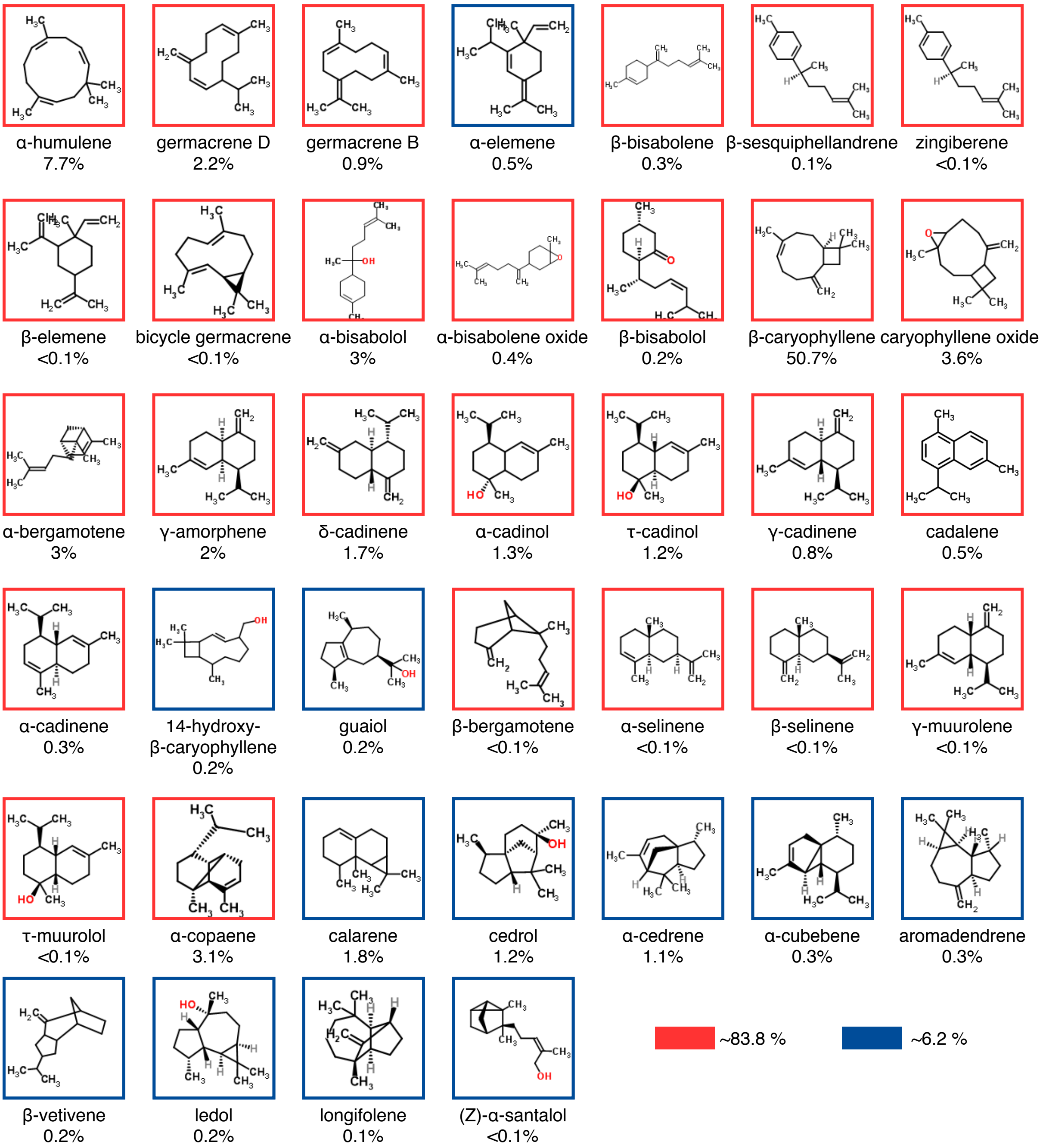


Figure 3.2: Sesquiterpenes found in oil-resin of *CmH* supplemented by their average percentages. The red border highlights the target compounds of this work.

A second literature review provided the chemical bases for building a set of graph grammar rules representing the chemical mechanisms. These chemical mechanisms are limited to the essential set of metabolic reactions for achieving the *Copaifera multijuga* Hayne (CmH) target sesquiterpenes.

3.2 Generation of the Sesquiterpenes Metabolic Network

The FPP is the pivotal substrate for the enzymatic reactions that catalyze the production of sesquiterpenes. Taking the FPP or its isomer NPP (See Figure 2.11) as a precursor, graph grammar rules were written to represent the chemical mechanisms responsible for the production of the target sesquiterpenes shown in Figure 3.2.

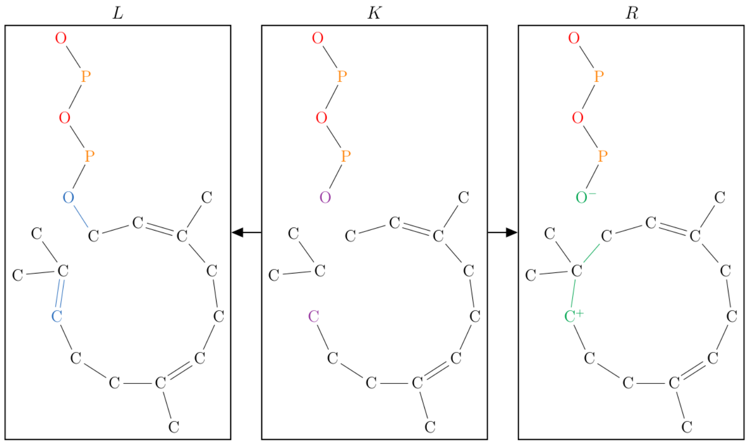
Each rule has its particular bibliographical reference and was written in an individual file using the Graph Modeling Language (GML) format (Himsolt, 1997), which can be read by MedØlDatschgerl. The set of 18 designed rules express together a set capable of achieving, during the simulation, each of the 27 target sesquiterpenes of CmH and a vast number of predicted compounds.

The designed rules may consider molecules (undirected graphs) partially or entirely. It keeps the computational simulation as generic as possible. For example, in the rule shown in Figure 3.3, the undirected graph representing the diphosphate OPP is not fully expressed. A sub-graph is sufficient for a match with the diphosphate molecule since there is not a different molecule with the same sub-graph during the simulation. Consequently, there is a fine-tuning between how generic or specific the rules are to work together.

The initial cyclizations from FPP after the OPP loss are 1-11 leading to (E)-humulyl cation, and 1-10 leading to (E,E)-germacradienyl cation. Alternatively, the four initial cyclizations of NPP are 1-6 leading to the bisabolyl cation, 1-7 leading to the cycloheptanyl cation, 1-10 leading to (Z,E)-germacradienyl cation or 1-11 leading to (Z)-humulyl cation. The initial cyclization 1-7 was not identified in plants. The list of graph grammar rules used to represent the undirected graph transformations are presented in Table 3.1:

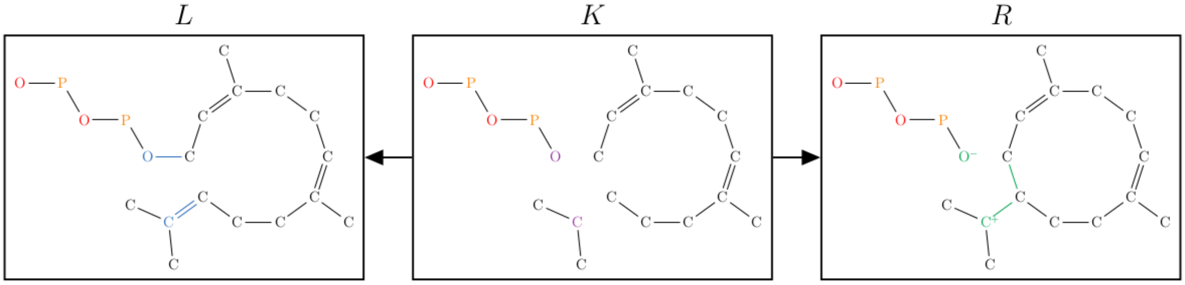
Table 3.1: List of graph grammar rules used to represent the undirected graph transformations.

Graph grammar rule	Figure	Reference
OPP loss from FPP and subsequent 1-11 ring closure	3.3	(Christianson, 2017; Degenhardt et al., 2009)
OPP loss from FPP and subsequent 1-10 ring closure	3.4	(Christianson, 2017; Degenhardt et al., 2009)
OPP loss from NPP and subsequent 1-11 ring closure	3.5	(Christianson, 2017; Degenhardt et al., 2009)
OPP loss from NPP and subsequent 1-10 ring closure	3.6	(Christianson, 2017; Degenhardt et al., 2009)
OPP loss from NPP and subsequent 1-6 ring closure	3.7	(Christianson, 2017; Degenhardt et al., 2009)
Formation of farnesysl cation $C3^+$	3.8	(Christianson, 2017; Degenhardt et al., 2009)
1,2 hydride shift	3.9	(Sandbeck et al., 2016)
1,3 hydride shift	3.10	(Sandbeck et al., 2016)
allyl shift	3.11	(Rinkel et al., 2016)
Capture of a water molecule	3.12	(Vattekkatte et al., 2018)
2-7 ring closure	3.13	(Steele et al., 1998)
2-6 ring closure	3.14	(Dickschat, 2016)
1-11 ring closure	3.15	(de Kraker et al., 1998; Garms et al., 2010)
2-10 ring closure	3.16	(Degenhardt et al., 2009)
Cope rearrangement	3.17	(Colby et al., 1998; Takeda, 1974)
Oxyreduction of δ-cadinene	3.18	(Townsend, 2005)
Reprotonation of $C7$ in germacrenoids	3.19	(Christianson, 2017; Steele et al., 1998)
Deprotonation (H^+ loss)	3.20	(Vattekkatte et al., 2018)



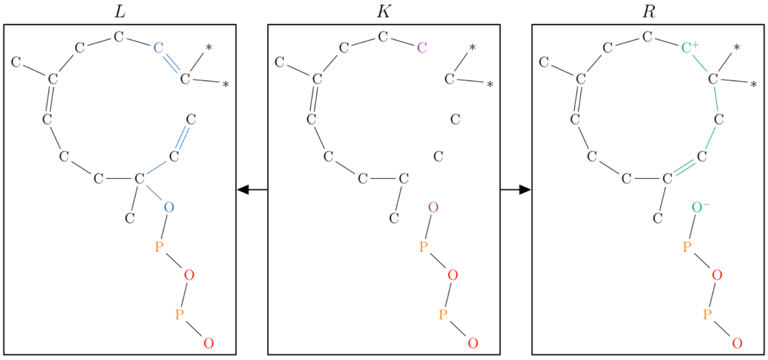
```
# (Christianson,2017) https://doi.org/10.1021/acs.chemrev.7b00287
# (Degenhardt, 2009) https://doi.org/10.1016/j.phytochem.2009.07.030
rule [
    ruleID "FPP OPP loss and 1-11 ring closure"
    left [
        node [ id 10 label "C" ] edge [ source 10 target 11 label "=" ]
        node [ id 16 label "O" ] edge [ source 1 target 16 label "-" ]
    ]
    context [
        node [ id  1 label "C" ]    node [ id  2 label "C" ]
        node [ id  3 label "C" ]    node [ id  4 label "C" ]
        node [ id  5 label "C" ]    node [ id  6 label "C" ]
        node [ id  7 label "C" ]    node [ id  8 label "C" ]
        node [ id  9 label "C" ]    node [ id 11 label "C" ]
        node [ id 12 label "C" ]    node [ id 13 label "C" ]
        node [ id 14 label "C" ]    node [ id 15 label "C" ]
        node [ id 17 label "P" ]    node [ id 18 label "O" ]
        node [ id 19 label "P" ]    node [ id 20 label "O" ]
        edge [ source  1 target  2 label "-" ]  edge [ source  2 target  3 label "=" ]
        edge [ source  3 target  4 label "-" ]  edge [ source  3 target 15 label "-" ]
        edge [ source  4 target  5 label "-" ]  edge [ source  5 target  6 label "-" ]
        edge [ source  6 target  7 label "=" ]  edge [ source  7 target  8 label "-" ]
        edge [ source  7 target 14 label "-" ]  edge [ source  8 target  9 label "-" ]
        edge [ source  9 target 10 label "-" ]  edge [ source 11 target 12 label "-" ]
        edge [ source 11 target 13 label "-" ]  edge [ source 16 target 17 label "-" ]
        edge [ source 17 target 18 label "-" ]  edge [ source 18 target 19 label "-" ]
        edge [ source 19 target 20 label "-" ]
    ]
    right [
        node [ id 10 label "C+" ] edge [ source 10 target 11 label "-" ]
        node [ id 16 label "O-" ] edge [ source  1 target 11 label "-" ]
    ]
]
```

Figure 3.3: Graph grammar rule representing the OPP cleavage from the FPP molecule and the subsequent $C1$,$C11$ ring closure.



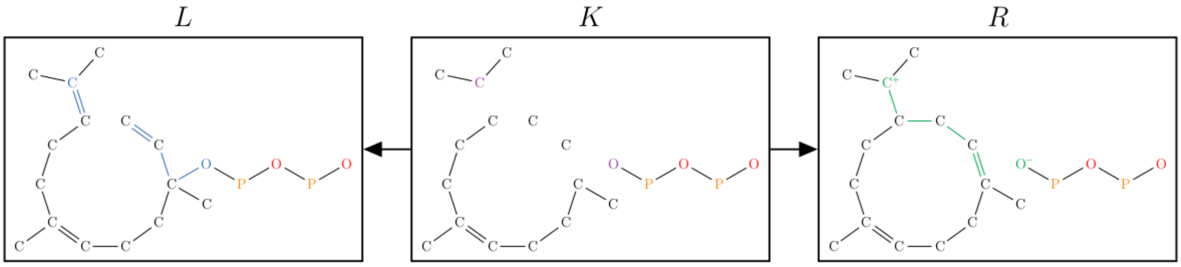
```
# (Christianson,2017) https://doi.org/10.1021/acs.chemrev.7b00287
# (Degenhardt, 2009) https://doi.org/10.1016/j.phytochem.2009.07.030
rule [
    ruleID "FPP OPP loss and 1-10 ring closure"
    left [
        node [ id 11 label "C" ] edge [ source 10 target 11 label "=" ]
        node [ id 16 label "O" ] edge [ source 1 target 16 label "-" ]
     ]
    context [
        node [ id  1 label "C" ]    node [ id  2 label "C" ]
        node [ id  3 label "C" ]    node [ id  4 label "C" ]
        node [ id  5 label "C" ]    node [ id  6 label "C" ]
        node [ id  7 label "C" ]    node [ id  8 label "C" ]
        node [ id  9 label "C" ]    node [ id 10 label "C" ]
        node [ id 12 label "C" ]    node [ id 13 label "C" ]
        node [ id 14 label "C" ]    node [ id 15 label "C" ]
        node [ id 17 label "P" ]    node [ id 18 label "O" ]
        node [ id 19 label "P" ]    node [ id 20 label "O" ]
        edge [ source  1 target  2 label "-" ]  edge [ source  2 target  3 label "=" ]
        edge [ source  3 target  4 label "-" ]  edge [ source  3 target 15 label "-" ]
        edge [ source  4 target  5 label "-" ]  edge [ source  5 target  6 label "-" ]
        edge [ source  6 target  7 label "=" ]  edge [ source  7 target  8 label "-" ]
        edge [ source  7 target 14 label "-" ]  edge [ source  8 target  9 label "-" ]
        edge [ source  9 target 10 label "-" ]  edge [ source 11 target 12 label "-" ]
        edge [ source 11 target 13 label "-" ]  edge [ source 16 target 17 label "-" ]
        edge [ source 17 target 18 label "-" ]  edge [ source 18 target 19 label "-" ]
        edge [ source 19 target 20 label "-" ]
    ]
    right [
        node [ id 11 label "C+" ] edge [ source  1 target 10 label "-" ]
        node [ id 16 label "O-" ] edge [ source 10 target 11 label "-" ]
    ]
]
```

Figure 3.4: Graph grammar rule representing the OPP cleavage from the FPP molecule and the subsequent $C1$,$C10$ ring closure.



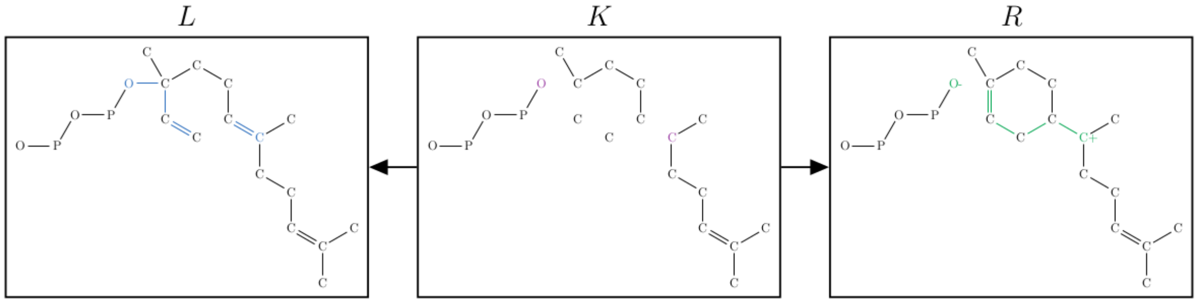
```
# (Christianson, 2017) https://doi.org/10.1021/acs.chemrev.7b00287
# (Degenhardt, 2009) https://doi.org/10.1016/j.phytochem.2009.07.030
rule [
    ruleID "NPP OPP loss and 1-11 ring closure"
    left [
        node [ id 10 label "C" ]    node [ id 16 label "O" ]
        edge [ source  1 target  2 label "=" ]  edge [ source  2 target  3 label "-" ]
        edge [ source  3 target 16 label "-" ]  edge [ source 10 target 11 label "=" ]
    ]
    context [
        node [ id  1 label "C" ]    node [ id  2 label "C" ]
        node [ id  3 label "C" ]    node [ id  4 label "C" ]
        node [ id  5 label "C" ]    node [ id  6 label "C" ]
        node [ id  7 label "C" ]    node [ id  8 label "C" ]
        node [ id  9 label "C" ]    node [ id 11 label "C" ]
        node [ id 12 label "*" ]    node [ id 13 label "*" ]
        node [ id 14 label "C" ]    node [ id 15 label "C" ]
        node [ id 17 label "P" ]    node [ id 18 label "O" ]
        node [ id 19 label "P" ]    node [ id 20 label "O" ]
        edge [ source  3 target  4 label "-" ]  edge [ source  3 target 15 label "-" ]
        edge [ source  4 target  5 label "-" ]  edge [ source  5 target  6 label "-" ]
        edge [ source  6 target  7 label "=" ]  edge [ source  7 target  8 label "-" ]
        edge [ source  7 target 14 label "-" ]  edge [ source  8 target  9 label "-" ]
        edge [ source  9 target 10 label "-" ]  edge [ source 11 target 12 label "-" ]
        edge [ source 11 target 13 label "-" ]  edge [ source 16 target 17 label "-" ]
        edge [ source 17 target 18 label "-" ]  edge [ source 18 target 19 label "-" ]
        edge [ source 19 target 20 label "-" ]
    ]
    right [
        node [ id 10 label "C+" ]   node [ id 16 label "O-" ]
        edge [ source  1 target 11 label "-" ]  edge [ source  1 target  2 label "-" ]
        edge [ source  2 target  3 label "=" ]  edge [ source 10 target 11 label "-" ]
    ]
]
```

Figure 3.5: Graph grammar rule representing the OPP cleavage from the NPP molecule and the subsequent $C1$,$C11$ ring closure.



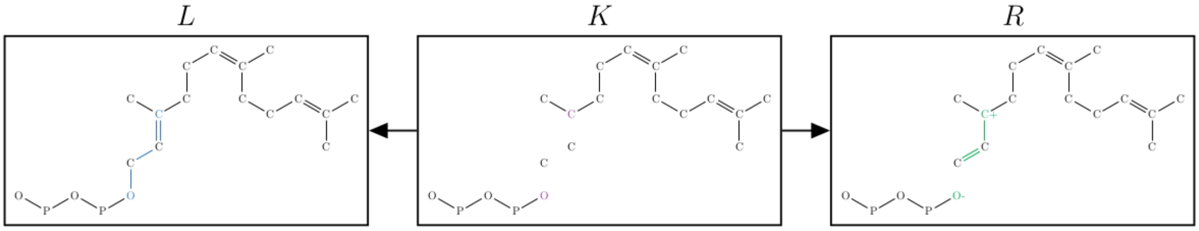
```
# (Christianson, 2017) https://doi.org/10.1021/acs.chemrev.7b00287
# (Degenhardt, 2009) https://doi.org/10.1016/j.phytochem.2009.07.030
rule [
    ruleID "NPP OPP loss and 1-10 ring closure"
    left [
        node [ id 11 label "C" ]     node [ id 16 label "O" ]
        edge [ source 10 target 11 label "=" ]   edge [ source 1 target  2 label "=" ]
        edge [ source  2 target  3 label "-" ]   edge [ source 3 target 16 label "-" ]
     ]
    context [
        node [ id  1 label "C" ]     node [ id  2 label "C" ]
        node [ id  3 label "C" ]     node [ id  4 label "C" ]
        node [ id  5 label "C" ]     node [ id  6 label "C" ]
        node [ id  7 label "C" ]     node [ id  8 label "C" ]
        node [ id  9 label "C" ]     node [ id 10 label "C" ]
        node [ id 12 label "C" ]     node [ id 13 label "C" ]
        node [ id 14 label "C" ]     node [ id 15 label "C" ]
        node [ id 17 label "P" ]     node [ id 18 label "O" ]
        node [ id 19 label "P" ]     node [ id 20 label "O" ]
        edge [ source  3 target  4 label "-" ]   edge [ source  3 target 15 label "-" ]
        edge [ source  4 target  5 label "-" ]   edge [ source  5 target  6 label "-" ]
        edge [ source  6 target  7 label "=" ]   edge [ source  7 target  8 label "-" ]
        edge [ source  7 target 14 label "-" ]   edge [ source  8 target  9 label "-" ]
        edge [ source  9 target 10 label "-" ]   edge [ source 11 target 12 label "-" ]
        edge [ source 11 target 13 label "-" ]   edge [ source 16 target 17 label "-" ]
        edge [ source 17 target 18 label "-" ]   edge [ source 18 target 19 label "-" ]
        edge [ source  9 target 20 label "-" ]
    ]
    right [
        node [ id 11 label "C+" ]    node [ id 16 label "O-" ]
        edge [ source  1 target  2 label "-" ]   edge [ source  2 target  3 label "=" ]
        edge [ source 10 target 11 label "-" ]   edge [ source  1 target 10 label "-" ]

    ]
]
```

Figure 3.6: Graph grammar rule representing the OPP cleavage from the NPP molecule and the subsequent $C1$,$C10$ ring closure.



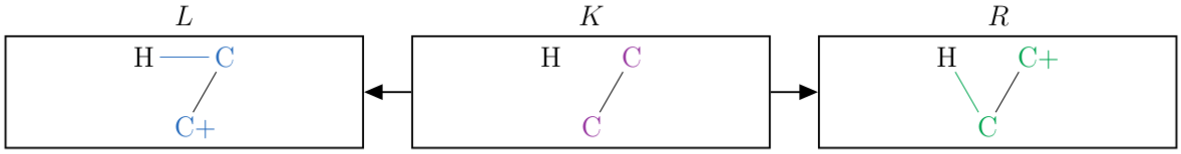
```
# (Christianson, 2017) https://doi.org/10.1021/acs.chemrev.7b00287
# (Degenhardt, 2009) https://doi.org/10.1016/j.phytochem.2009.07.030
rule [
    ruleID "NPP OPP loss and 1-6 ring closure"
    left [
        node [ id 7 label "C" ] node [ id 16 label "O" ]
        edge [ source  1 target  2 label "=" ]  edge [ source 2 target  3 label "-" ]
        edge [ source  6 target  7 label "=" ]  edge [ source 3 target 16 label "-" ]
    ]
    context [
        node [ id  1 label "C" ]    node [ id  2 label "C" ]
        node [ id  3 label "C" ]    node [ id  4 label "C" ]
        node [ id  5 label "C" ]    node [ id  6 label "C" ]
        node [ id  8 label "C" ]    node [ id  9 label "C" ]
        node [ id 10 label "C" ]    node [ id 11 label "C" ]
        node [ id 12 label "C" ]    node [ id 13 label "C" ]
        node [ id 14 label "C" ]    node [ id 15 label "C" ]
        node [ id 17 label "P" ]    node [ id 18 label "O" ]
        node [ id 19 label "P" ]    node [ id 20 label "O" ]
        edge [ source  3 target  4 label "-" ]  edge [ source  3 target 15 label "-" ]
        edge [ source  4 target  5 label "-" ]  edge [ source  5 target  6 label "-" ]
        edge [ source  7 target  8 label "-" ]  edge [ source  7 target 14 label "-" ]
        edge [ source  8 target  9 label "-" ]  edge [ source  9 target 10 label "-" ]
        edge [ source 10 target 11 label "=" ]  edge [ source 11 target 12 label "-" ]
        edge [ source 11 target 13 label "-" ]  edge [ source 16 target 17 label "-" ]
        edge [ source 17 target 18 label "-" ]  edge [ source 18 target 19 label "-" ]
        edge [ source 19 target 20 label "-" ]
    ]
    right [
        node [ id 7 label "C+" ]    node [ id 16 label "O-" ]
        edge [ source  1 target  2 label "-" ]  edge [ source  2 target  3 label "=" ]
        edge [ source  1 target  6 label "-" ]  edge [ source  6 target  7 label "-" ]
    ]
]
```

Figure 3.7: Graph grammar rule representing the OPP cleavage from the NPP molecule and the subsequent $C1$,$C6$ ring closure.



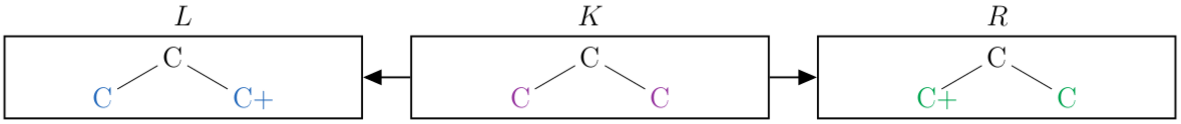
```
# (Christianson, 2017) https://doi.org/10.1021/acs.chemrev.7b00287
# (Degenhardt, 2009) https://doi.org/10.1016/j.phytochem.2009.07.030
rule [
    ruleID "FPP OPP loss and farnesyl cation C3+ formation"
    left [
        node [ id 16 label "O" ]    node [ id  3 label "C" ]
        edge [ source 1 target 16 label "-" ]
        edge [ source  1 target  2 label "-" ]
        edge [ source  2 target  3 label "=" ]
     ]
    context [
        node [ id  1 label "C" ]    node [ id  2 label "C" ]
        node [ id  4 label "C" ]    node [ id  5 label "C" ]
        node [ id  6 label "C" ]    node [ id  7 label "C" ]
        node [ id  8 label "C" ]    node [ id  9 label "C" ]
        node [ id 10 label "C" ]    node [ id 11 label "C" ]
        node [ id 12 label "C" ]    node [ id 13 label "C" ]
        node [ id 14 label "C" ]    node [ id 15 label "C" ]
        node [ id 17 label "P" ]    node [ id 18 label "O" ]
        node [ id 19 label "P" ]    node [ id 20 label "O" ]
        edge [ source  3 target  4 label "-" ]  edge [ source  3 target 15 label "-" ]
        edge [ source  4 target  5 label "-" ]  edge [ source  5 target  6 label "-" ]
        edge [ source  6 target  7 label "=" ]  edge [ source  7 target  8 label "-" ]
        edge [ source  7 target 14 label "-" ]  edge [ source  8 target  9 label "-" ]
        edge [ source  9 target 10 label "-" ]  edge [ source 10 target 11 label "=" ]
        edge [ source 11 target 12 label "-" ]  edge [ source 11 target 13 label "-" ]
        edge [ source 16 target 17 label "-" ]  edge [ source 17 target 18 label "-" ]
        edge [ source 18 target 19 label "-" ]  edge [ source 19 target 20 label "-" ]
    ]
    right [
        node [ id  3 label "C+" ]   node [ id 16 label "O-" ]
        edge [ source  1 target  2 label "=" ]
        edge [ source  2 target  3 label "-" ]
    ]
]
```

Figure 3.8: Graph grammar rule representing the formation of the farnesyl cation $C3^{+}$.



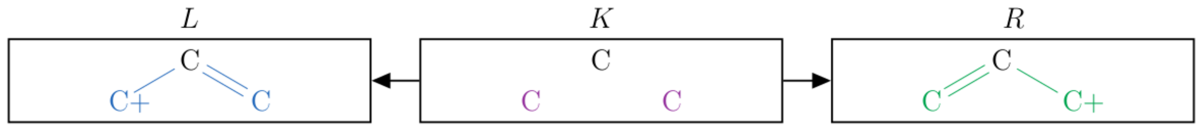
```
# (Sandbeck, 2016) https://doi.org/10.1021/acs.joc.5b02553
rule [
 ruleID "1,2 hydrid shift"
 left [
  node [ id 1 label "C" ]
  node [ id 2 label "C+" ]
  edge [ source 1 target 3 label "-" ]
 ]
 context [
  node [ id 3 label "H" ]
  edge [ source 1 target 2 label "-" ]
 ]
 right [
  node [ id 1 label "C+" ]
  node [ id 2 label "C" ]
  edge [ source 2 target 3 label "-" ]
 ]
]
```

Figure 3.9: Graph grammar rule representing an 1,2 hydride shift.



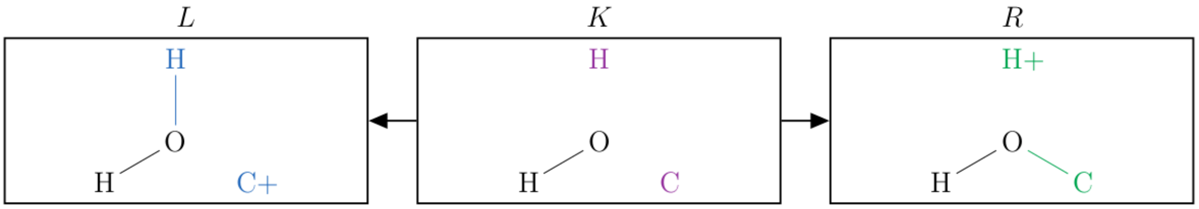
```
# (Sandbeck, 2016) https://doi.org/10.1021/acs.joc.5b02553
rule [
 ruleID "1,3 hydride shift"
 left [
  node [ id  1 label "C+" ]
  node [ id  3 label "C" ]
 ]
 context [
  node [ id  2 label "C" ]
  edge [ source  1 target  2 label "-" ]
  edge [ source  2 target  3 label "-" ]
 ]
 right [
  node [ id  1 label "C" ]
  node [ id  3 label "C+" ]
 ]
]
```

Figure 3.10: Graph grammar rule representing an 1,3 hydride shift.



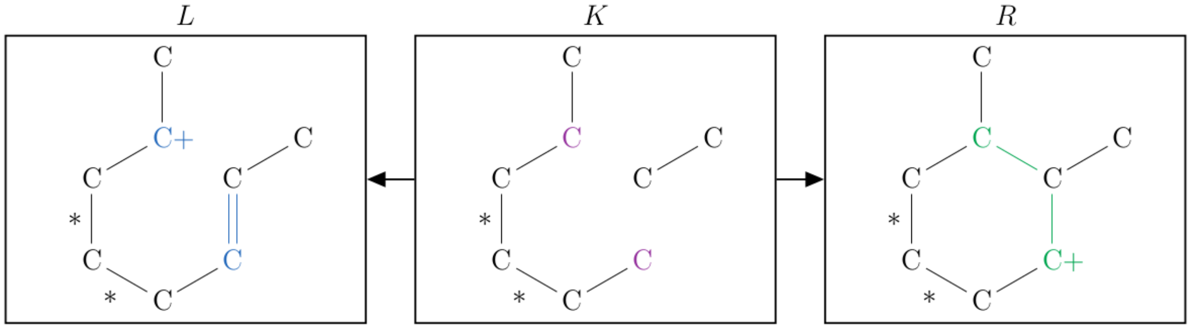
```
# (Rinkel, 2016) https://doi.org/10.1002/anie.201608042
rule [
 ruleID "allylic charge shift"
 left [
  node [ id 1 label "C" ]        node [ id 3 label "C+" ]
  edge [ source 1 target 2 label "=" ]  edge [ source 2 target 3 label "-" ]
 ]
 context [
  node [ id 2 label "C" ]
 ]
 right [
  node [ id 1 label "C+" ]       node [ id 3 label "C" ]
  edge [ source 1 target 2 label "-" ]  edge [ source 2 target 3 label "=" ]
 ]
]
```

Figure 3.11: Graph grammar rule representing an allyl shift.



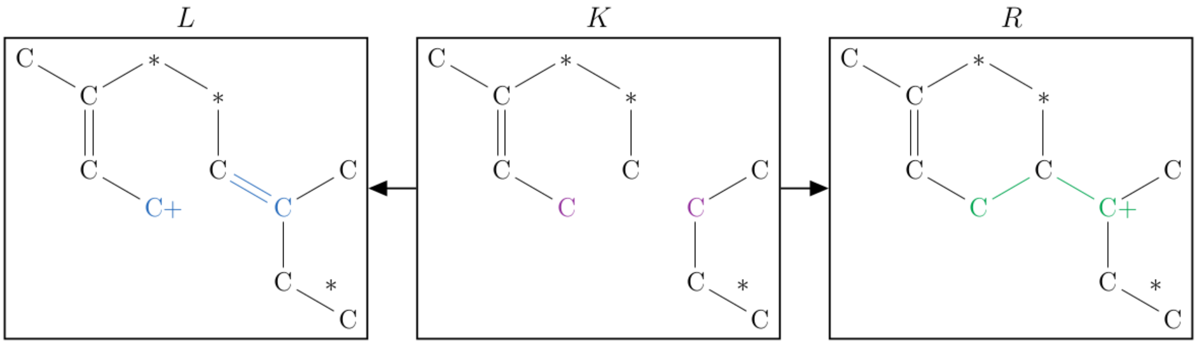
```
# (Vattekkatte, 2018) https://doi.org/10.1039/C7OB02040F
rule [
 ruleID "Capture of H2O"
 left [
  node [ id 1 label "C+" ]
  node [ id 2 label "H" ]
  edge [ source 2 target 3 label "-" ]
 ]
 context [
  node [ id 3 label "O" ]
  node [ id 4 label "H" ]
  edge [ source 3 target 4 label "-" ]
 ]
 right [
  node [ id 1 label "C" ]
  node [ id 2 label "H+" ]
  edge [ source 1 target 3 label "-" ]
 ]
]
```

Figure 3.12: Graph grammar rule representing the capture of water by a cation molecule.



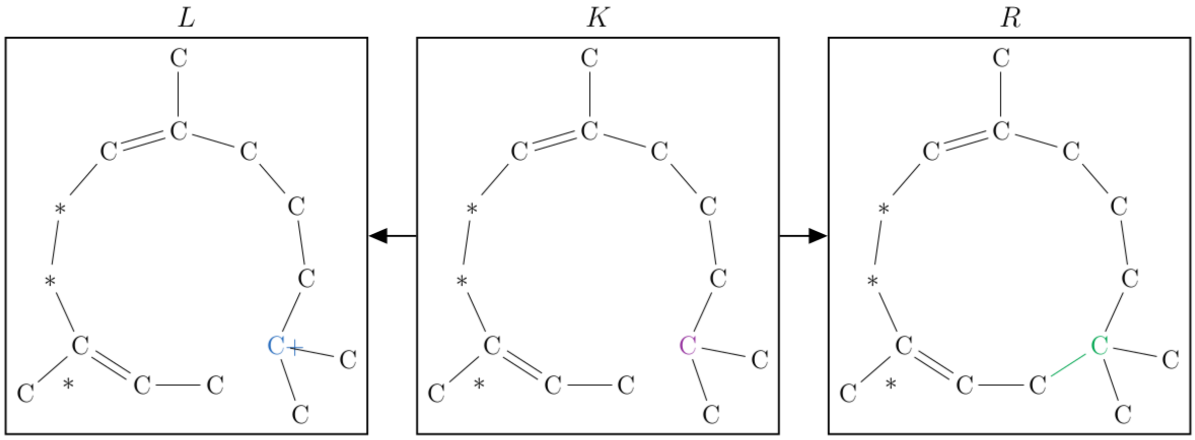
```
# (Steele, 1998) https://doi.org/10.1074/jbc.273.4.2078
rule [
 ruleID "2-7 ring closure"
 left [
  node [ id  7 label "C+"]
  node [ id  3 label "C" ]
  edge [ source  2 target  3 label "=" ]
 ]
 context [
  node [ id  1 label "C" ]
  node [ id  2 label "C" ]
  node [ id  4 label "C" ]
  node [ id  5 label "C" ]
  node [ id  6 label "C" ]
  node [ id  8 label "C" ]
  edge [ source  1  target  2 label "-" ]
  edge [ source  3  target  4 label "-" ]
  edge [ source  4  target  5 label "*" ]
  edge [ source  5  target  6 label "*" ]
  edge [ source  6  target  7 label "-" ]
  edge [ source  7  target  8 label "-" ]
 ]
 right [
  node [ id  7 label "C" ]
  node [ id  3 label "C+" ]
  edge [ source  2 target 3 label "-" ]
  edge [ source  2 target 7 label "-" ]
 ]
]
```

Figure 3.13: Graph grammar rule representing the $C2$, $C2$ ring closure.



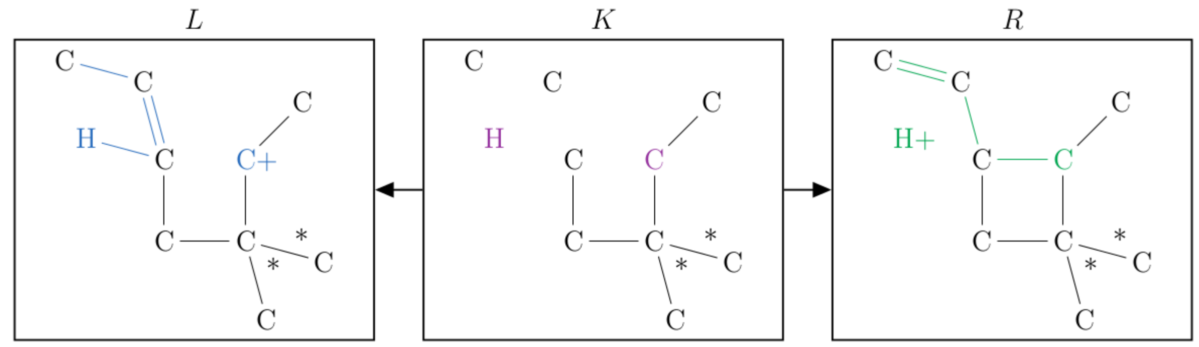
```
# (Dickschat, 2016) https://doi.org/10.1039/C5NP00102A
rule [
 ruleID "1-6 ring closure"
 left [
  node [ id  1 label "C+" ] node [ id  7 label "C"  ]
  edge [ source  6 target  7 label "=" ]
 ]
 context [

  node [ id  2 label "C" ]  node [ id  3 label "C" ]
  node [ id  4 label "*" ]  node [ id  5 label "*" ]
  node [ id  6 label "C" ]  node [ id  8 label "C" ]
  node [ id  9 label "C" ]  node [ id 14 label "C" ]
  node [ id 15 label "C" ]
  edge [ source  1 target  2 label "-" ]  edge [ source  2 target  3 label "=" ]
  edge [ source  3 target  4 label "-" ]  edge [ source  3 target 15 label "-" ]
  edge [ source  4 target  5 label "-" ]  edge [ source  5 target  6 label "-" ]
  edge [ source  7 target  8 label "-" ]  edge [ source  7 target 14 label "-" ]
  edge [ source  8 target  9 label "*" ]
 ]
 right [
  node [ id  1 label "C" ]  node [ id  7 label "C+" ]
  edge [ source  6 target  7 label "-" ]  edge [ source  1 target  6 label "-" ]
 ]
]
```

Figure 3.14: Graph grammar rule representing the $C1$, $C6$ ring closure.



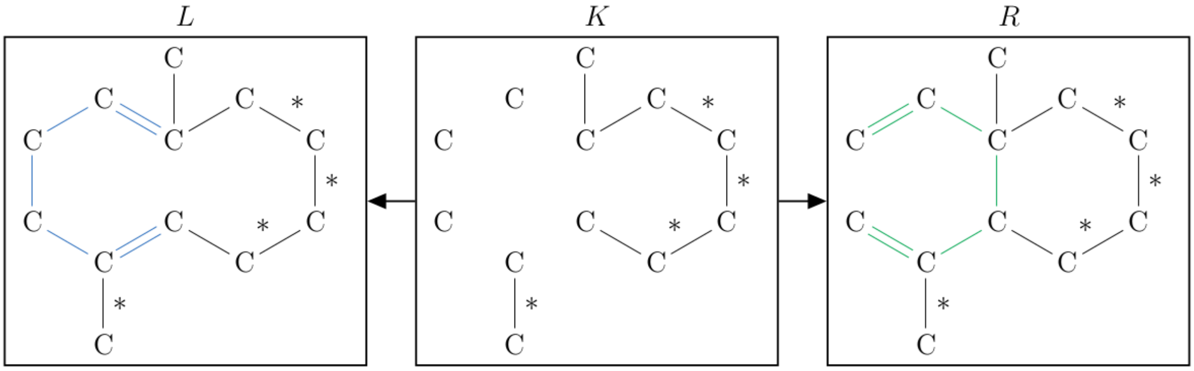
```
# (de Kraker, 1998) https://doi.org/10.1104/pp.117.4.1381
# (Garms, 2010) https://doi.org/10.1021/jo100917c
rule [
 ruleID "1-11 ring closure"
 left [
  node [ id  11 label "C+" ]
 ]
 context [
  node [ id  1 label "C" ]    node [ id  2 label "C" ]
  node [ id  3 label "C" ]    node [ id  4 label "*" ]
  node [ id  5 label "*" ]    node [ id  6 label "C" ]
  node [ id  7 label "C" ]    node [ id  8 label "C" ]
  node [ id  9 label "C" ]    node [ id 10 label "C" ]
  node [ id 12 label "C" ]    node [ id 13 label "C" ]
  node [ id 14 label "C" ]    node [ id 15 label "C" ]
  edge [ source  1 target  2 label "-" ]
  edge [ source  2 target  3 label "=" ]
  edge [ source  3 target  4 label "-" ]
  edge [ source  4 target  5 label "-" ]
  edge [ source  5 target  6 label "-" ]
  edge [ source  6 target  7 label "=" ]
  edge [ source  7 target  8 label "-" ]
  edge [ source  8 target  9 label "-" ]
  edge [ source  9 target 10 label "-" ]
  edge [ source  3 target 15 label "*" ]
  edge [ source  7 target 14 label "*" ]
  edge [ source 10 target 11 label "-" ]
  edge [ source 11 target 12 label "-" ]
  edge [ source 11 target 13 label "-" ]
  edge [ source  7 target 14 label "-" ]
 ]
 right [
  node [ id 11 label "C" ]
  edge [ source  1 target  11 label "-" ]
 ]
]
```

Figure 3.15: Graph grammar rule representing the $C1$, $C11$ ring closure.



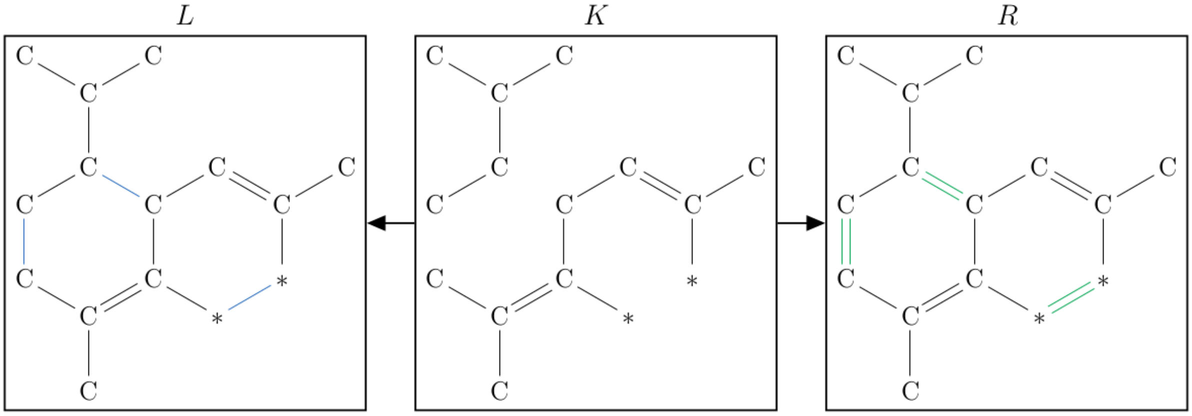
```
# 2-10 closure rule
# Last review: March, 11, 2018 by Waldeyr Mendes Cordeiro da Silva
# from Kempinski_2015.pdf, Biosynthesis and Biological Functions of Terpenoids in Plants, Figure 3b

rule [
 ruleID "2-10 closure"
 left [
  node [ id 10 label "C+"]
  edge [ source  2 target  3 label "=" ]
  node [ id  3 label "C" ]
 ]
 context  [
  node [ id  1 label "C" ]
  node [ id  2 label "C" ]

  node [ id  4 label "C" ]
  node [ id  5 label "C" ]
  node [ id  6 label "C" ]
  node [ id  7 label "C" ]
  node [ id  8 label "C" ]
  node [ id  9 label "C" ]
  node [ id 11 label "C" ]
  node [ id 12 label "C" ]
  node [ id 13 label "C" ]
  edge [ source  1 target  2 label "-" ]
  edge [ source  1 target 11 label "-" ]
  edge [ source  3 target  4 label "-" ]
  edge [ source  4 target  5 label "*" ]
  edge [ source  5 target  6 label "*" ]
  edge [ source  6 target  7 label "=" ]
  edge [ source  7 target  8 label "-" ]
  edge [ source  8 target  9 label "*" ]
  edge [ source  9 target 10 label "-" ]
  edge [ source 10 target 11 label "-" ]
  edge [ source 11 target 12 label "*" ]
  edge [ source 11 target 13 label "*" ]
 ]
 right [
  node [ id 10 label "C"]
  edge [ source  2 target  3 label "-" ]
  node [ id  3 label "C+" ]
  edge [ source  2 target 10 label "-" ]
 ]
]
```

Figure 3.16: Graph grammar rule representing the $C2$, $C10$ ring closure.



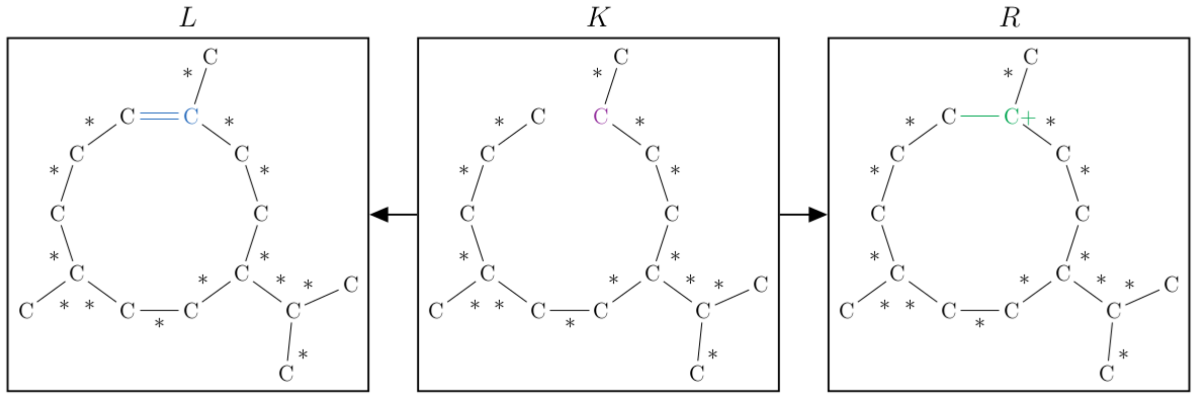
```
# (Takeda, 1974) https://doi.org/10.1016/S0040-4020(01)90674-X
# (Colby, 1998) https://doi.org/10.1073/pnas.95.5.2216
rule [
 ruleID "Cope Rearrangment (HEAT)"
 left [
  edge [ source  2 target  3 label "=" ]  edge [ source  3 target  4 label "-" ]
  edge [ source  4 target  5 label "-" ]  edge [ source  5 target  6 label "-" ]
  edge [ source  6 target  7 label "=" ]
 ]
 context [
  node [ id  1 label "C" ]  node [ id  2 label "C" ]
  node [ id  3 label "C" ]  node [ id  4 label "C" ]
  node [ id  5 label "C" ]  node [ id  6 label "C" ]
  node [ id  7 label "C" ]  node [ id  8 label "C" ]
  node [ id  9 label "C" ]  node [ id 10 label "C" ]
  node [ id 14 label "C" ]  node [ id 15 label "C" ]
  edge [ source  1 target  2 label "-" ]
  edge [ source  1 target 10 label "*" ]
  edge [ source  9 target 10 label "*" ]
  edge [ source  8 target  9 label "*" ]
  edge [ source  7 target  8 label "-" ]
  edge [ source  7 target 14 label "-" ]
  edge [ source  3 target 15 label "*" ]
 ]
 right [
  edge [ source  2 target  3 label "-" ]
  edge [ source  2 target  7 label "-" ]
  edge [ source  3 target  4 label "=" ]
  edge [ source  5 target  6 label "=" ]
  edge [ source  6 target  7 label "-" ]
 ]
]
```

Figure 3.17: Graph grammar rule representing cope rearrangement.



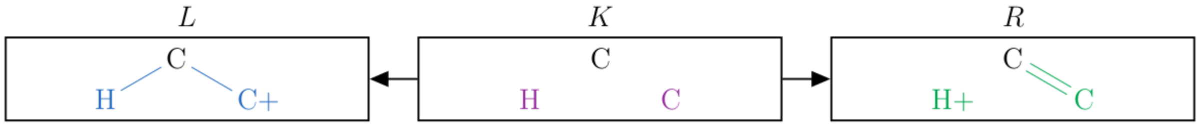
```
# (Townsend,2005) https://doi.org/10.1104/pp.104.056010
rule [
 ruleID "delta-cadinene oxyreduction"
 left [
   edge [ source  4 target  5 label "-" ]
   edge [ source  8 target  9 label "-" ]
   edge [ source  1 target 10 label "-" ]
 ]
 context [
  node [ id  1 label "C" ]  node [ id  2 label "C" ]
  node [ id  3 label "C" ]  node [ id  4 label "*" ]
  node [ id  5 label "*" ]  node [ id  6 label "C" ]
  node [ id  7 label "C" ]  node [ id  8 label "C" ]
  node [ id  9 label "C" ]  node [ id 10 label "C" ]
  node [ id 11 label "C" ]  node [ id 12 label "C" ]
  node [ id 13 label "C" ]  node [ id 14 label "C" ]
  node [ id 15 label "C" ]
  edge [ source  1 target  2 label "-" ]
  edge [ source  2 target  3 label "=" ]
  edge [ source  3 target  4 label "-" ]
  edge [ source  5 target  6 label "-" ]
  edge [ source  6 target  7 label "=" ]
  edge [ source  7 target  8 label "-" ]
  edge [ source  9 target 10 label "-" ]
  edge [ source 10 target 11 label "-" ]
  edge [ source 11 target 12 label "-" ]
  edge [ source 11 target 13 label "-" ]
  edge [ source  3 target 15 label "-" ]
  edge [ source  7 target 14 label "-" ]
  edge [ source  1 target  6 label "-" ]
 ]
 right [
   edge [ source  4 target  5 label "=" ]
   edge [ source  8 target  9 label "=" ]
   edge [ source  1 target 10 label "=" ]
 ]
]
```

Figure 3.18: Graph grammar rule representing an oxyreduction in an δ-cadinene molecule.



```
# (Christianson,2017) https://doi.org/10.1021/acs.chemrev.7b00287
# (Steele, 1998) https://doi.org/10.1074/jbc.273.4.2078
rule [
    ruleID "Germacrenes Reprotonation (C7)"
    left [
        node [ id  7 label "C" ]
        edge [ source  6 target  7 label "=" ]
    ]
    context [
        node [ id  1 label "C" ]    node [ id  2 label "C" ]
        node [ id  3 label "C" ]    node [ id  4 label "C" ]
        node [ id  5 label "C" ]    node [ id  6 label "C" ]
        node [ id  8 label "C" ]    node [ id  9 label "C" ]
        node [ id 10 label "C" ]    node [ id 11 label "C" ]
        node [ id 12 label "C" ]    node [ id 13 label "C" ]
        node [ id 14 label "C" ]    node [ id 15 label "C" ]
        edge [ source  1 target  2 label "*" ]
        edge [ source  2 target  3 label "*" ]
        edge [ source  3 target  4 label "*" ]
        edge [ source  3 target 15 label "*" ]
        edge [ source  4 target  5 label "*" ]
        edge [ source  5 target  6 label "*" ]
        edge [ source  7 target  8 label "*" ]
        edge [ source  7 target 14 label "*" ]
        edge [ source  8 target  9 label "*" ]
        edge [ source  9 target 10 label "*" ]
        edge [ source 10 target 11 label "*" ]
        edge [ source 10 target  1 label "*" ]
        edge [ source 11 target 12 label "*" ]
        edge [ source 11 target 13 label "*" ]
    ]
    right [
        node [ id  7 label "C+" ]
        edge [ source  6 target  7 label "-" ]
    ]
]
```

Figure 3.19: Graph grammar rule representing the reprotonation of $C7$ in germacrenoid molecules.



```
# (Christianson, 2017) https://doi.org/10.1021/acs.chemrev.7b00287
rule [
    ruleID "Deprotonation (H+)"
    left [
        node [ id 1 label "C+" ] node [ id 3 label "H" ]
        edge [ source 1 target 2 label "-" ] edge [ source 2 target 3 label "-" ]
    ]
    context [
        node [ id 2 label "C" ]
    ]
    right [
        node [ id 1 label "C" ] node [ id 3 label "H+" ]
        edge [ source 1 target 2 label "=" ]
    ]
]
```

Figure 3.20: Graph grammar rule representing a deprotonation (H^+ loss).

Simulating the Sesquiterpenes Biosynthesis

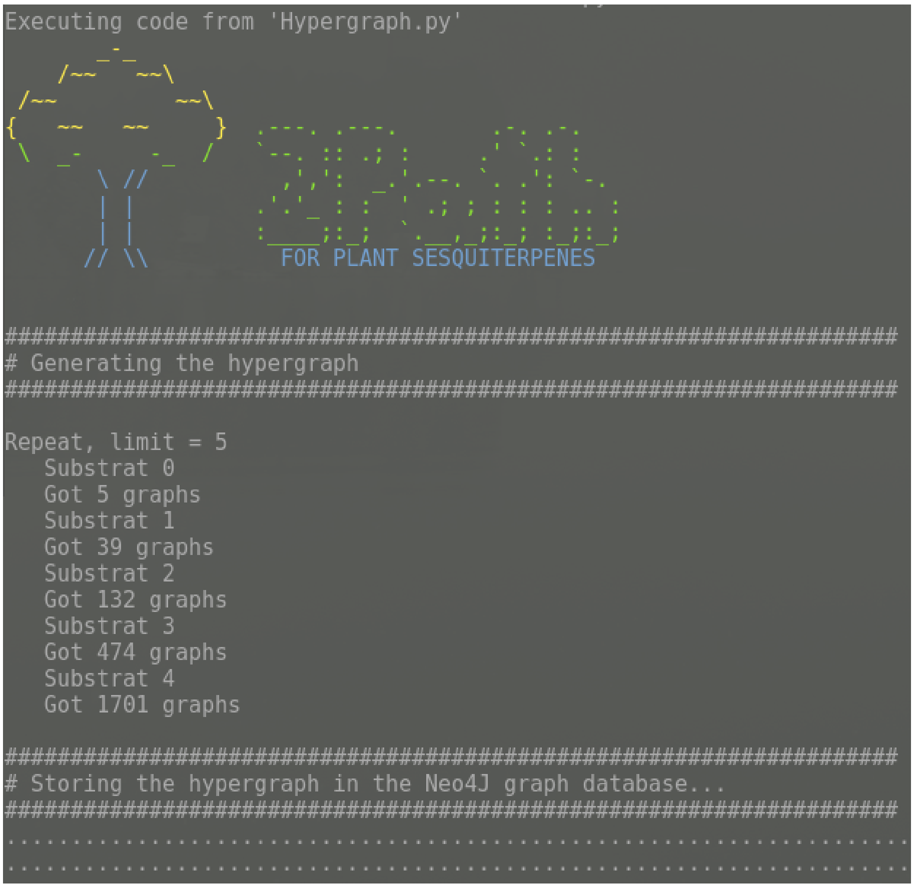
The set of start compounds was defined comprising FPP, NPP, and H_2O molecules, represented as undirected graphs generated from their SMILES. Also, was defined the set of target compounds comprising those 27 compounds showed in Figure 3.2, also represented as undirected graphs. Thus, with the starting and finishing points defined, and driven by the Python script *Hypergraph.py*, MedØlDatschgerl calculated the derivation graph using the designed graph grammar rules to reach the target compounds from the start compounds. All the target compounds were generated after five iterations using a general breadth-first expansion strategy (Andersen et al., 2014).

Defining the target compounds as constraints, their presence in the generated chemical network was checked using integer programming. The IBM CPLEX® Optimization Studio 127 was used for this task. Also, using PETRI-NET (Peterson, 1981) prove it was possible to get the temporal ordering of reaction sequence from the integer programming results.

The derivation graph is a directed hypergraph (See Section 2.3.1) representing the metabolic network of reactions and chemical mechanisms that transforms start compounds into target compounds preserving the atoms type, mass, and charge. The resulting range of compounds is not limited to the target compounds, and a massive number of predicted compounds is also generated including all intermediates, as cations.

3.3 Storing the Sesquiterpenes Network of CmH

After the generation of the metabolic network as a hypergraph, it was stored in the Neo4J graph database. Figure 3.21 shows the running of the derivation graph calculation and its storage in the Neo4J.



```
Executing code from 'Hypergraph.py'
       -
      _ _
   /~~   ~~\
 /~~         ~~\
{   ~~   ~~     }
 \   _-     -_  /
      \ //
       | |
       | |
      // \\          FOR PLANT SESQUITERPENES


####################################################################
# Generating the hypergraph
####################################################################

Repeat, limit = 5
   Substrat 0
   Got 5 graphs
   Substrat 1
   Got 39 graphs
   Substrat 2
   Got 132 graphs
   Substrat 3
   Got 474 graphs
   Substrat 4
   Got 1701 graphs

####################################################################
# Storing the hypergraph in the Neo4J graph database...
####################################################################
....................................................................
....................................................................
```

Figure 3.21: Generating the chemical network as a hypergraph and storing it in the Neo4J database.

Since the chemical network is a hypergraph, and the Neo4J graph database does not support a hypergraph directly, a schema was modeled to accommodate the hypergraph as a graph. Graph Description Diagram for Graph Databases (GRAPHED) (Van Erven et al., 2018) was used to model this graph database schema.

By using GRAPHED, the vertex is represented by a rounded rectangle, with a mandatory label inside, and optional elements such as 'Type' and 'Attribute' between brackets. The optional 'Type' information is used to indicate the domain of the identifier. This format is similar to tables in the relational model with the attributes' name on the left, followed by its type on the right.

The vertices *Organism*, *Sequence*, and *Scenario* were modeled using this notation. The hyperedges representing the many-to-many relation of *Compound* vertices, was effectively implemented by creating an additional vertex *Rule* that centralizes the *n:n* relationships between the vertices. Also, in this schema there are attributes both in the vertices and in the edges, since the Neo4J graph database has support for this feature. Figure 3.22 presents the effective way the graph database was implemented.

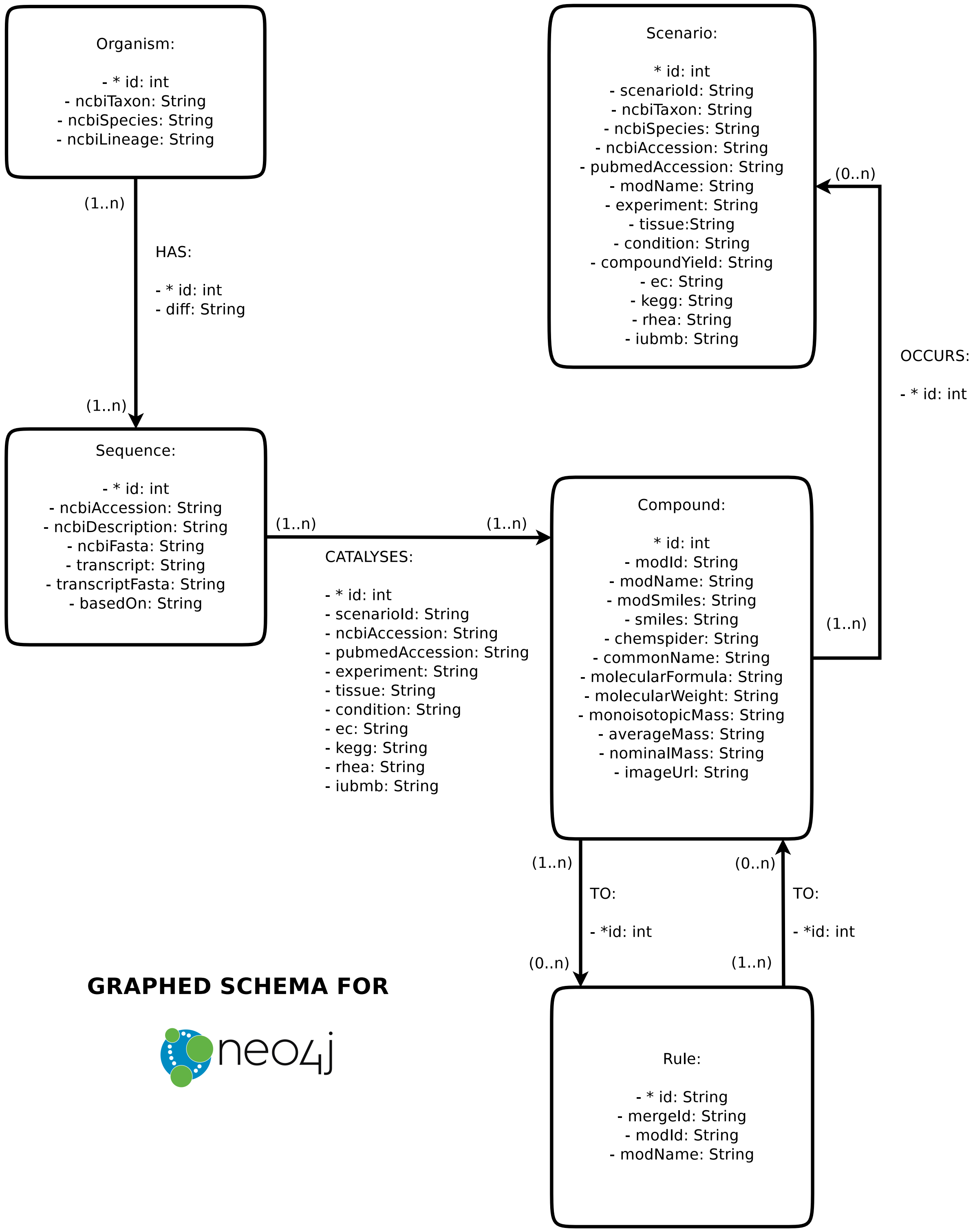
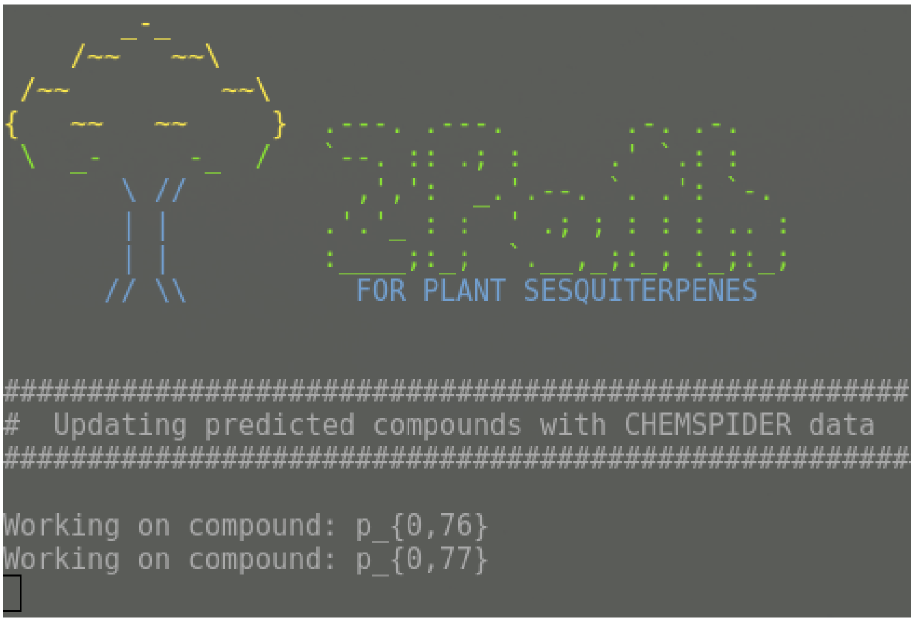


Figure 3.22: Implemented graph database schema for Neo4J graph database.

One of the most important practices for database documentation is the data dictionary (Batini et al., 1992), which provides information about the database including the definitions of all schema objects in the database. The Appendix II presents the data dictionary for the modeled schema.

Annotation of Predicted Compounds

The resulting hypergraph has a considerable number of predicted compounds generated together with the target compounds using the same set of graph grammar rules. The Royal Society of Chemistry provides the ChemSpider (Pence and Williams, 2010) Web Service which was used to identify these predicted compounds by comparing their chemical structure. The Python program *Compounds.py* performed the annotation of the predicted compounds and their storage in the Neo4J as shown in Figure 3.23.



```
                                  FOR PLANT SESQUITERPENES


######################################################
#  Updating predicted compounds with CHEMSPIDER data
######################################################

Working on compound: p_{0,76}
Working on compound: p_{0,77}
```

Figure 3.23: Updating the predicted compounds with annotations from Chemspider datsbase.

Scenarios for Sesquiterpenes Biosynthesis

The literature provided a collection of scenarios for biosynthesis of a variety of sesquiterpenes. The scenarios are identified by a ‘scenarioId’ and contain information such as experiment, tissue, condition, yield, EC of the enzymes for the particular enzymatic activity. The Planteome ontology[2] controls the vocabularies used for tissue and experiment

[2] http://www.planteome.org

conditions. For example, 'PO:0009046:flower' represents the ontology for the characteristic reproductive structure of angiosperms. These scenarios were imported to the Neo4J database and then associated to the annotated compounds by the Python program *Scenarios.py*. Figure 3.24 shows the storage of the scenarios in the Neo4J, and Appendix III.1 shows these identified scenarios.

Scenarios in the database embody the identified experimental results of enzymes from various plants catalyzing the synthesis of the same sesquiterpenes found in the CmH oil-resin.

```
FOR PLANT SESQUITERPENES

#################################################
Storing on scenarios...

WARNING: It can take some time.
#################################################
```

Figure 3.24: Updating the database with the scenarios for sesquiterpenes biosynthesis.

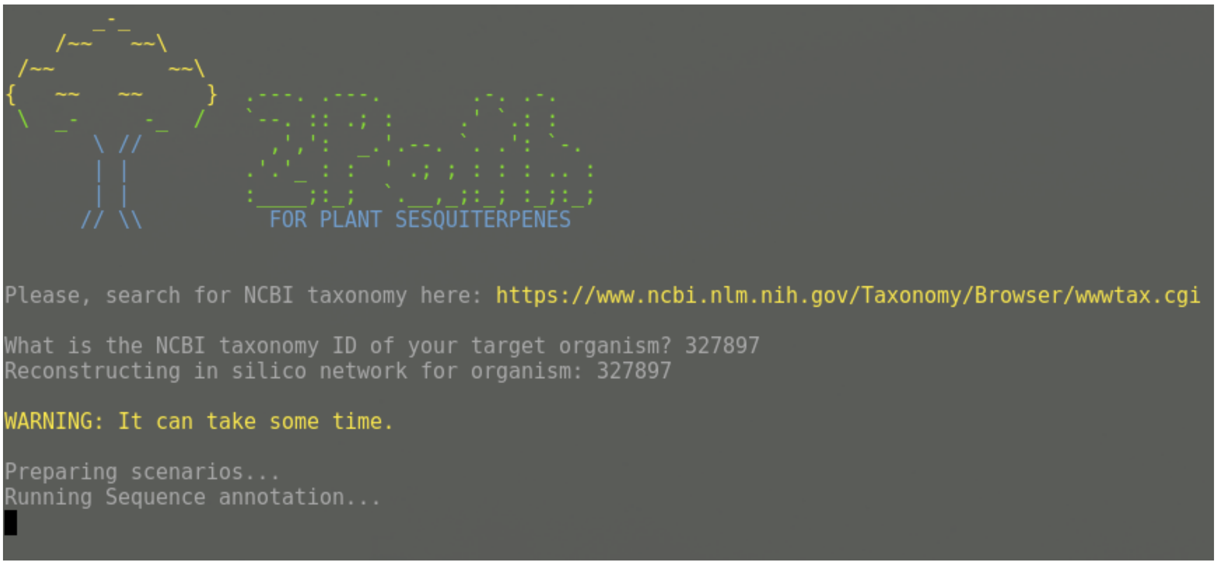
3.4 Annotation of CmH Enzymes

In 2011, a CmH transcriptome was sequenced from a cDNA library extracted from leaves of young and healthy plants of CmH collected from the greenhouse of Medicinal Plants of the University of Amazon with about 1 and 1.5 m tall, and 0.5 cm and 3.0 cm thick stem (Bastos, 2011). The reads[3], sequenced using Roche 454®, were filtered and assembled. The filtering phase aimed to prevent misassembled contigs in the later stages by detecting and correcting read errors. The assembly phase aimed to reconstruct extended sequences (contigs) from smaller sequence (reads) through sequence alignments using Trinity (Haas et al., 2013).

[3] Fragments of DNA released by the sequencers in text format.

A Blast (Altschul et al., 1997) database was built using the amino acids residues from enzymes retrieved from the scenarios in the database. Then, the CmH transcriptome was aligned with 'blastx' (nucleotides as query and proteins as database) against this Blast database to identify and annotate the sequences of enzymes catalyzing the synthesis of the target sesquiterpenes. After that, the results were parsed and stored in the database.

This is the final stage for the reconstruction of the metabolic network of sesquiterpenes of CmH. It was performed with an interactive prompt by the Python program *Reconstruction.py*. Figure 3.25 shows this running.



```
FOR PLANT SESQUITERPENES

Please, search for NCBI taxonomy here: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi

What is the NCBI taxonomy ID of your target organism? 327897
Reconstructing in silico network for organism: 327897

WARNING: It can take some time.

Preparing scenarios...
Running Sequence annotation...
```

Figure 3.25: Updating the database with the annotated transcripts of *Copaifera multijuga* Hayne (CmH).

Chapter 4

Results and Discussion

Although the metabolic network of *Copaifera multijuga* Hayne (CmH) is the most evident result, it is not the only one. The defined workflow can be used to reconstruct the metabolic networks of any other plant from its transcriptome. Also, it is possible to generate different networks for the same organism, changing the set of graph grammar rules and executing new computational simulations. The workflow is modular, and each step can be performed individually. Another implicit result is the consolidation of the use of graphs databases for storage of metabolic networks as a viable and efficient alternative (Silva et al., 2017). Using the graph database, it is possible to plan biological experiments by combining information from reactions, compounds, scenarios, and sequences of sesquiterpene synthases through queries.

In this chapter, the results are discussed highlighting these outcomes, limitations, perspectives and compared with related works, where possible.

4.1 *In silico* Sesquiterpenes Metabolic Network of *Copaifera multijuga* Hayne (CmH)

The *in silico* sesquiterpenes metabolic network of *Copaifera multijuga* Hayne (CmH) has been reconstructed and stored in the 2Path Database for Sesquiterpenes (See Section 4.2). It covers a range of enzymatic metabolic reactions forming sesquiterpenes, including predicted compounds and chemical mechanisms for these reactions, which were generated based on graph grammar rules applied to only the set of initial precursor molecules: FPP, NPP and H_2O. Figure 4.1 shows an overview of the plant terpene metabolism and the place occupied by the generated chemical network in this context.

The predicted compounds in this chemical network were annotated using data from ChemSpider (Pence and Williams, 2010), comprising the common name, molecular for-

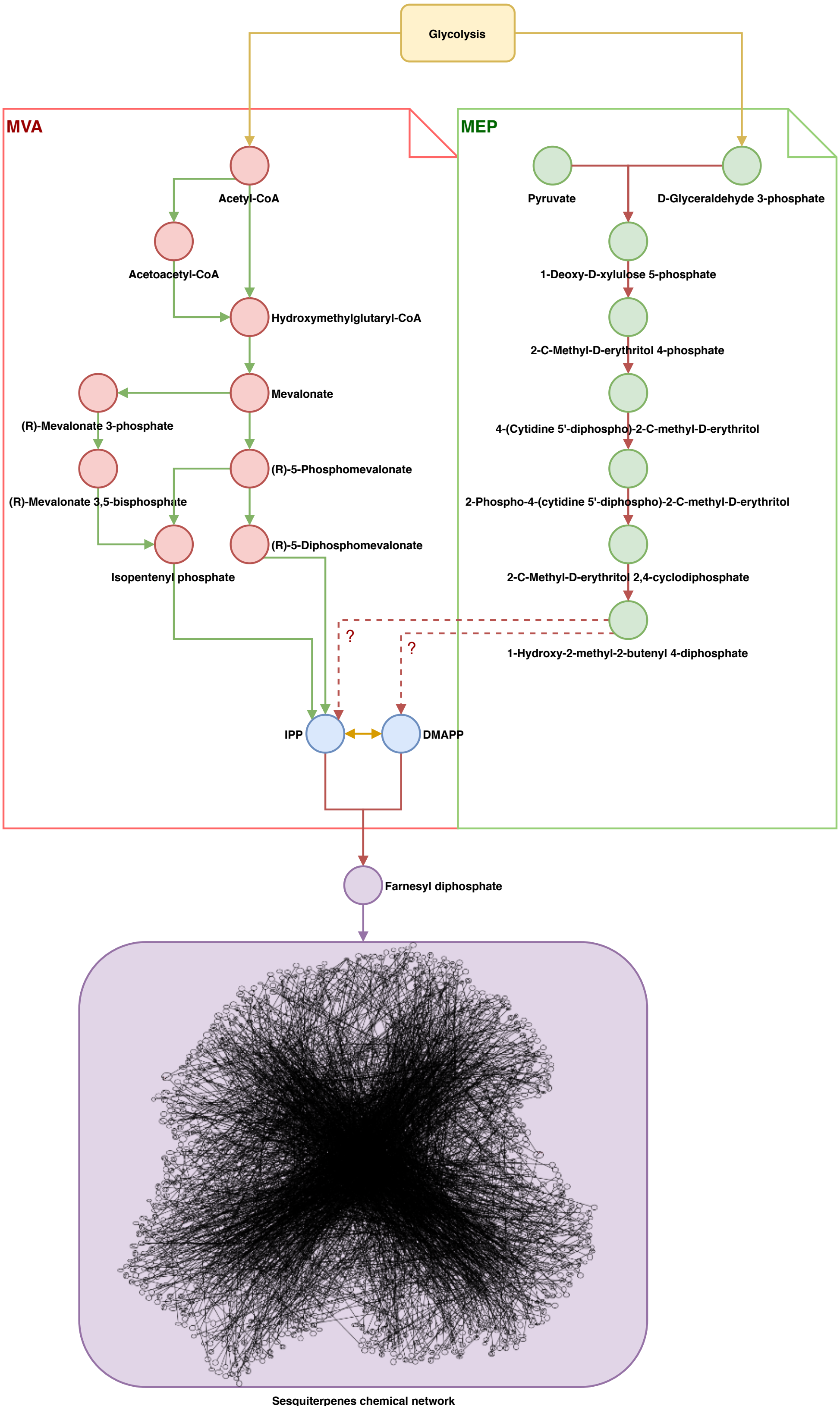

Glycolysis
MVA
MEP
Acetyl-CoA
Acetoacetyl-CoA
Hydroxymethylglutaryl-CoA
Mevalonate
(R)-Mevalonate 3-phosphate
(R)-Mevalonate 3,5-bisphosphate
(R)-5-Phosphomevalonate
(R)-5-Diphosphomevalonate
Isopentenyl phosphate
Pyruvate
D-Glyceraldehyde 3-phosphate
1-Deoxy-D-xylulose 5-phosphate
2-C-Methyl-D-erythritol 4-phosphate
4-(Cytidine 5'-diphospho)-2-C-methyl-D-erythritol
2-Phospho-4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol
2-C-Methyl-D-erythritol 2,4-cyclodiphosphate
1-Hydroxy-2-methyl-2-butenyl 4-diphosphate
?
?
IPP
DMAPP
Farnesyl diphosphate
Sesquiterpenes chemical network

Figure 4.1: Generated chemical network in the context of plant terpene metabolism. The purple box shows all generated feasible reactions from FPP and NPP.

mula, molecular weight, SMILES, a two-dimensional image of their chemical structures, and mono-isotopic, average and nominal mass. The universe of compounds achieved using the set of graph grammar rules exceeds the target compounds opening a way to investigate possible alternative sesquiterpene biosynthesis. Figure 4.2 shows the predicted chemical network with cation molecules in red, and the electrically stable molecules in blue. Figure 4.3 shows the CmHsesquiterpenes reached using the set of graph grammar rules (See Table 3.1). A document with all predicted compounds can be downloaded from `http://www.biomol.unb.br/2path/docs/2path15_predicted_compounds.pdf`.

In nature, enzymes regularly catalyze reactions, enhancing their magnitude. Despite this, there are cases, including the metabolism of terpenes, where the reaction control and specificity take priority over rate enhancement (Vattekkatte et al., 2018). Plant mono- and sesquiterpene synthases generate substantial amounts of different acyclic and cyclic products making them multiproduct enzymes (Vattekkatte et al., 2018). According to Degenhardt *et al.* (Degenhardt et al., 2009), the electrophilic chemical mechanisms controlled by these enzymes influence the diversity of products. But many other factors influence multiproduct sesquiterpene enzymes and their cyclization cascades. Some of them are unknown, while others have been studied, such as pH, the metal cofactor (Vattekkatte et al., 2018), and evolutionary forces for the functional divergence (Chen et al., 2014).

Sesquiterpene synthases belong to Class I terpene synthases. Their catalytic process starts with the cleavage of Diphosphate (OPP) mediated by a metal ion, which also neutralizes the negative charge on the OPP during the reaction, preventing a premature quenching of the cation. Sesquiterpene synthases have a preference for Mg 2+ as cofactor *in vitro*, but they also accept Mn^{2+} in low concentrations (Vattekkatte et al., 2018). Although less expressive, there are examples of plant sesquiterpene synthases that show catalytic activity in the presence of other metal ions (Köllner et al., 2008).

pH influences both concentration and the specificity of terpenes. In *Medicago trucatula*, for example, the enzymatic activity of $MtTPS5$ occurred in a limited pH range between 5 and 11. This enzyme decreased the production of cadalene and increased the production of germacrene while the pH became basic.

Particularly for plants, the multiproduct ability of sesquiterpene synthases gives ecological advantages by producing a range of direct and indirect defensive compounds against herbivores. This range is also influenced by the tissue and environmental conditions. For instance, Köllner *et al.* (Kollner et al., 2008) showed that *Zea mays* $TPS23$ gene provides two distinct expression patterns in different tissues. Aboveground, the $TPS23$ expression after herbivory by lepdopteran larvae on leaves produced a blend of volatile terpenes attracting parasitic wasps. Underground, in the roots, the $TPS23$ produced only β-caryophyllene after damage by *Diabrotica virgifera* attracting pathogenic nematodes.

The evolution of enzymes over time is related to their cellular processes (Vattekkatte et al., 2018). In terpene synthases, the multiproduct activity reflects the tendency of nature to form mechanisms that maximize the range of products with a minimal number of steps (Vattekkatte et al., 2018). Studies have pointed out that TPSs of plants are often more related to their own TPSs with similar function than to other plant species (Kollner et al., 2008). Chen *et al.* (Chen et al., 2014) found that the volatile sesquiterpene blend, containing germacrenes, produced by the $TPS1$ gene of *Oryza* species may be adaptive. They suggested that a positive Darwinian selection drives this diversification of terpenoid biosynthesis in the genus *Oryza*.

As the blend of the produced sesquiterpenes is broad and influenced by several assay conditions, the curated annotation of a TPS sequence becomes a difficult task. The annotation of an enzyme is dependent on the scenario in which the synthesis occurs because a single enzyme catalyzes the synthesis of several sesquiterpenes.

Enzymes catalyzing the biosynthesis of sesquiterpenes in many scenarios were collected from the literature and stored in the 2Path database for sesquiterpenes (See Section 4.2). These scenarios provided a framework of evidence beyond the similarity of the sequences to support the annotation of the CmH enzymes. They included the NCBI accession number for the enzymes, the PUBMED accession number for the associated publication with the experimental results, the experimental conditions, the plant tissue, the compound yield, EC numbers for the reactions, and cross-references to KEGG, Rhea (Morgat et al., 2016) and International Union of Biochemistry and Molecular Biology (IUBMB) in a taxonomic range of species.

The CmH transcriptome was assembled using the software Trinity (Haas et al., 2013), and the report for the assembly is shown in Table 4.2. A total of 28 sequences of CmH were annotated as sesquiterpene synthases using these scenarios, and their sequence identifiers are shown in Table 4.1.

An *in silico* metabolic network of sesquiterpene biosynthesis that considers chemical mechanisms and a scenario is valuable for a biological investigation of the target organism. The importance of an *in silico* metabolic network lies both in the range of its explicit information, and in the knowledge that can be deduced from it, allowing several biological questions to be answered using this information.

4.2 2Path Database for Sesquiterpenes - 2Path15

Before choosing the graph database to store the metabolic network of CmH, we developed and published a proof of concept for terpenoids, materialized in the 2Path

Table 4.1: Transcripts of CmH annotated as sesquiterpenes synthases.

Nr.	Transcript	Nr.	Transcript
1	TRINITY_DN17151_c0_g1_i1	15	TRINITY_DN10226_c0_g1_i1
2	TRINITY_DN7739_c0_g1_i1	16	TRINITY_DN971_c0_g1_i1
3	TRINITY_DN28883_c0_g1_i1	17	TRINITY_DN5133_c0_g1_i1
4	TRINITY_DN9683_c0_g1_i1	18	TRINITY_DN29555_c0_g1_i1
5	TRINITY_DN8241_c0_g1_i1	19	TRINITY_DN24690_c0_g1_i1
6	TRINITY_DN21516_c0_g1_i1	20	TRINITY_DN23966_c0_g1_i1
7	TRINITY_DN11019_c0_g1_i1	21	TRINITY_DN813_c0_g2_i1
8	TRINITY_DN813_c0_g1_i1	22	TRINITY_DN3207_c0_g1_i1
9	TRINITY_DN22210_c0_g1_i1	23	TRINITY_DN4599_c7_g35_i1
10	TRINITY_DN4599_c7_g34_i1	24	TRINITY_DN4599_c7_g33_i1
11	TRINITY_DN10862_c0_g1_i1	25	TRINITY_DN26667_c0_g1_i1
12	TRINITY_DN2357_c0_g2_i1	26	TRINITY_DN2357_c0_g1_i1
13	TRINITY_DN8904_c0_g1_i1	27	TRINITY_DN27904_c0_g1_i1
14	TRINITY_DN6486_c0_g1_i1	28	TRINITY_DN21250_c1_g1_i1

Table 4.2: CmH transcriptome assembly stats.

Stats	
Total assembled bases	19384458
Average contig	461.11
Median contig lenght	411
N10	1007
N20	637
N30	516
N40	462
N50	433
Percent GC	32.44

database (Silva et al., 2017). 2Path database integrates data from several repositories of plant metabolism, filtering for terpenoid data. The next natural step was to produce new data to expand the 2Path database with information on the CmH.

This expansion is following a bottom-up flow. Detailed information on sesquiterpenes biosynthesis was generated by this work using data from the literature and computational simulations. This new information should fit as members of the terpenoid metabolic network existing in 2Path database. As this integration is not yet implemented, two databases are kept: 2Path for terpenoid metabolism and *2Path15* for sesquiterpene metabolism in an allusion to the 15 carbons skeleton of sesquiterpenes.

A suitable database model and schema makes the search for both explicit and implicit information more efficient. The 2Path15 graph database schema shown in Figure 3.22, presents the organisms connected to the enzymes, which are connected to the two potential initial precursor molecules: FPP and NPP. In this way, it is possible to connect an organism and its sequences to the generated chemical network and the scenarios. Table 4.3 summarizes some important numbers from the complete *in silico* network in the database.

Table 4.3: Database of CmH sesquiterpene metabolic network in numbers.

Object	**Amount**	**Cypher query**
Relationships	7888	MATCH ()–>() RETURN count(*)
Vertices	5507	MATCH (n) RETURN count(n)
Compound	2354	MATCH (n:Compound) RETURN count(n)
Organism	17	MATCH (n:Organism) RETURN count(n)
Rule	3043	MATCH (n:Rule) RETURN count(n)
Sequence	50	MATCH (n:Sequence) RETURN count(n)
Scenario	43	MATCH (n:Scenario) RETURN count(n)
Total store size	22.13 MB	-

Figure 4.4 shows the complete sesquiterpene network. Particularly for the CmH, the pathway for any of the target sesquiterpenes can be demanded through a simple Cypher query. For example, Figure 4.5 shows the pathways for CmH where the β-caryophyllene is the main compound produced. Both in the Figures 4.4 and 4.5, purple vertices are organisms, blue vertices are enzymes, green vertices are compounds, gray vertices are chemical mechanisms (or rules), and the red vertices are scenarios.

Also, each of the resulting compounds, target or not, can be queried from the 2Path15 database using Cypher language, which enables several biological questions to be answered. For example, which enzymes catalyze β-caryophyllene as the main compound in leaves of adult plants? Alternatively, what are the species whose enzymes were expressed in recombinant experiments? The answer to these questions and the correspondent Cypher queries are shown in Tables 4.4 and 4.5.

Table 4.4: Example 01 of biological question/answer using cypher query.

Which enzymes catalyze beta-caryophyllene as the main compound in leaves of adult plants?

```
MATCH   (c:Compound{modName:"beta-caryophyllene"})-->(n:Scenario{yield:"main"})
WHERE   n.tissue=~"(?i).*leaf.*" AND n.experiment = "adult"
RETURN  n.condition , n.ncbiSpecies , n.ec , n.ncbiAccession
```

condition	ncbi species	ec	ncbi accession
methyl jasmonate exposure	Medicago truncatula	4.2.3.57	AAV36464
Spodoptera spp. Exposure	Medicago truncatula	4.2.3.57	AAV36464

Since the data structure of the generated metabolic network is a particular graph, using graph database is a very intuitive way both to store and query the data. Cypher

Table 4.5: Example 02 of biological question/answer using cypher query.

What are the species whose enzymes were expressed in recombinant experiments?

```
MATCH (c:Compound)-->(n:Scenario)
WHERE n.experiment=~"recombinant" AND c.modName = "beta-caryophyllene" AND n.yield="main"
RETURN n.ncbiSpecies, n.ec, n.ncbiAccession, n.rhea, n.kegg, c.modName, c.averageMass
```

ncbi species	ec	ncbi accession	rhea	kegg	modName	average mass
Artemisia annua	4.2.3.57	AAL79181	28297	R08541	beta-caryophyllene	204.3511
Helianthus annuus	4.2.3.57	AAY41422	28297	R08541	beta-caryophyllene	204.3511
Zea diploperennis	4.2.3.57	ABY79209	28297	R08541	beta-caryophyllene	204.3511
Zea mays subsp. huehuetenangensis	4.2.3.57	ABY79210	28297	R08541	beta-caryophyllene	204.3511
Zea luxurians	4.2.3.57	ABY79211	28297	R08541	beta-caryophyllene	204.3511
Zea mays subsp. mexicana	4.2.3.57	ABY79212	28297	R08541	beta-caryophyllene	204.3511
Zea mays subsp. parviglumis	4.2.3.57	ABY79213	28297	R08541	beta-caryophyllene	204.3511
Zea perennis	4.2.3.57	ABY79214	28297	R08541	beta-caryophyllene	204.3511
Matricaria chamomilla var. Recutita	4.2.3.57	AFM43734	28297	R08541	beta-caryophyllene	204.3511
Selaginella moellendorffii	4.2.3.57	AFR34007	28297	R08541	beta-caryophyllene	204.3511
Cucumis sativus	4.2.3.57	AAU05952	28297	R08541	beta-caryophyllene	204.3511
Medicago truncatula	4.2.3.57	AAV36464	28297	R08541	beta-caryophyllene	204.3511

queries corroborate the intuitiveness by their convenient semantic structure. The result adequately stored in a graph database becomes Findable, Accessible, Interoperable, and Reusable. These are precisely the assumptions of science data management and knowledge discovery proposed by Wilkinson *et al.* (Wilkinson et al., 2016).

4.3 2Path15 Workflow

The provided a workflow for the reconstruction of *in silico* metabolic networks covers essential aspects of the biosynthesis of sesquiterpenes in plants. Based on this workflow, a series of partial results are combined to produce the *in silico* metabolic network.

Firstly, target sesquiterpenes from an organism can be collected from the literature and graph grammar rules designed to represent the chemical mechanisms of the enzymatic reactions that produce them. Alternatively, a set of arbitrarily defined rules can be used without necessarily defining target compounds.

The hypergraph resulting from the application of these rules is generated and stored in the Neo4J graph database using a database schema modeled to accommodate the hypergraph as a graph. The chemical structures of the predicted compounds is used to annotate them using data from the ChemSpider (Pence and Williams, 2010) Web Service. The core of the metabolic network can be augmented with scenarios, which are based on experimental results from the literature. Finally, the complete *in silico* metabolic network of sesquiterpenes is built from transcriptome data.

Different metabolic networks can be generated in two ways:

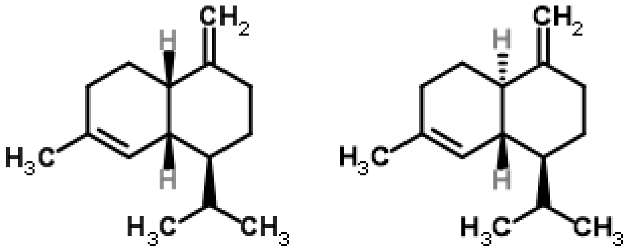
1. For different organisms by giving transcriptome data as input

2. For a target organism by changing the set of graph grammar rules or scenarios

The database supports the reconstruction of metabolic networks for multiple organisms, which makes it possible to compare them. Because the workflow is semi-automatic, and each phase can be performed separately, it simplifies the creation of new networks and the updating of the database with new scenarios.

4.4 Limitations

The main limitation of the workflow lies in the impossibility of abstracting chemical geometry in the molecules represented by undirected graphs. For example, the molecules γ-muurolene and γ-cadinene are distinguished by the geometry of the hydrogen bonded to the carbon 6 as can be seen in Figure 4.6. Such difference cannot be expressed using an undirected graph. Thus, during the generation of the metabolic network, compounds whose difference is only in the geometry of chemical bonds are indistinguishable.



(a) γ-muurolene. (b) γ-cadinene.

Figure 4.6: Abstraction for representing molecules as undirected graph.

This limitation also was the reason for bypassing the isomerization of the FPP to NPP (See Figure 2.11). In the simulations it was assumed that both are initial molecules.

4.5 Related works

As shown in Section 2.3, the reconstruction of metabolic networks is a diversified process with many methods. Some of them comprise chemical mechanisms through computational simulations as Artificial Force-Induced Reaction (AFIR) (Maeda et al., 2016) and 'Modelling Pathways as Integer Hyperflows' (Andersen et al., 2017). Isegawa *et al* (Isegawa et al., 2014) used AFIR to predict pathways applying artificial forces to molecules inducing them to approach each other. This work brought energetically viable predictions of cyclization/rearrangement pathways for carbocation precursors to sesquiterpenes.

The *in silico* metabolic network of CmH sesquiterpenes used the ‘Modelling Pathways as Integer Hyperflows’ to generate the chemical mechanisms. Compared to the work of Isegawa *et al* (Isegawa et al., 2014), this approach limited by the chemical geometry. Both methods have similar chemical goals but use different computational approaches.

The *in silico* metabolic network of CmH sesquiterpenes was later stored in the 2Path15 database supporting additional analyses. The closest work available with this feature is the Reactome (Fabregat et al., 2018), which is also a metabolic network database, recently updated to use the same strategy of storing. Although storing a rich universe of data on plant metabolism, Reactome[1], currently (Release 57 - May 2018) has a small set of terpene reactions (2), proteins (6), and no chemical mechanism details.

[1] http://plantreactome.gramene.org

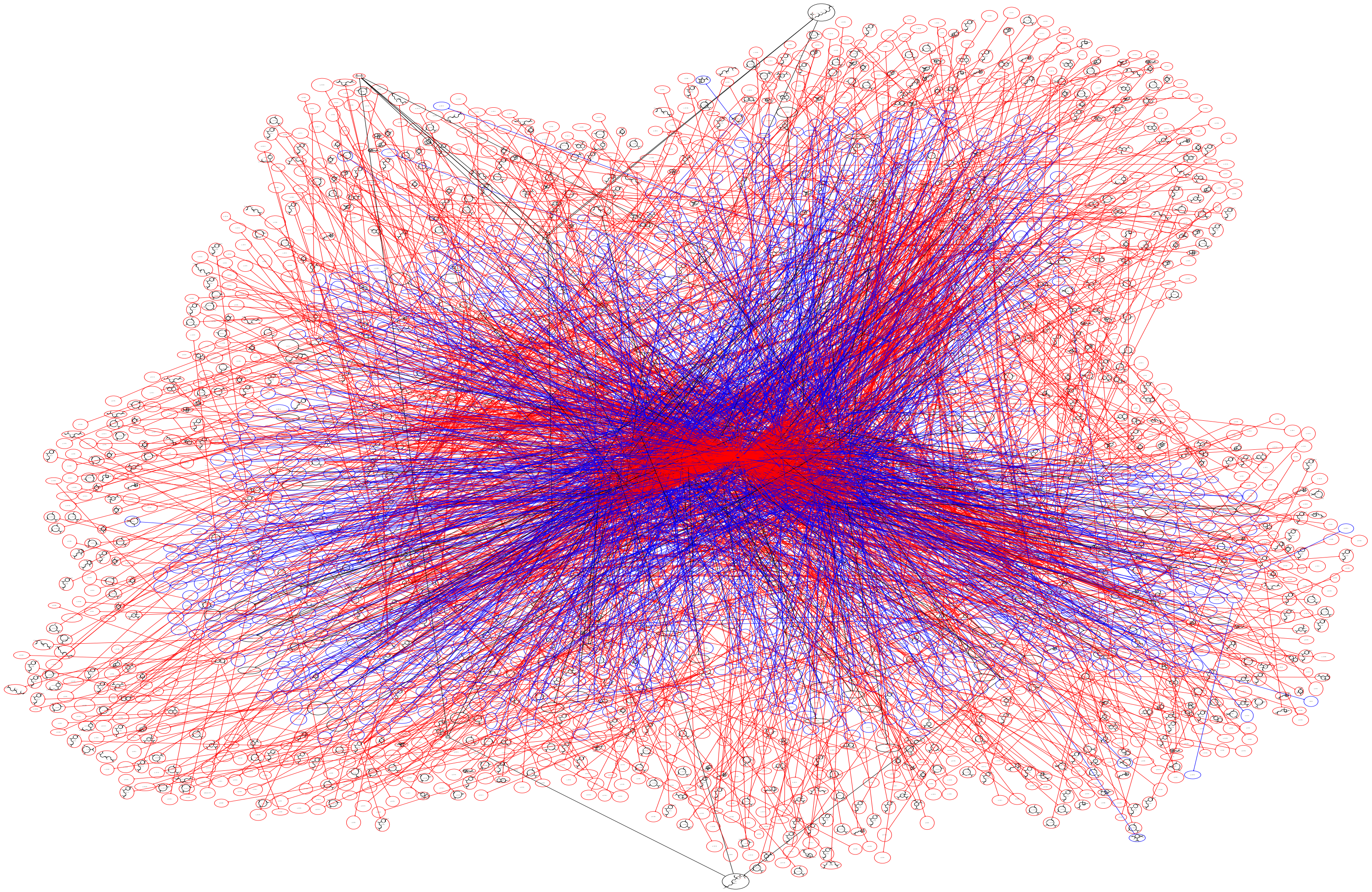
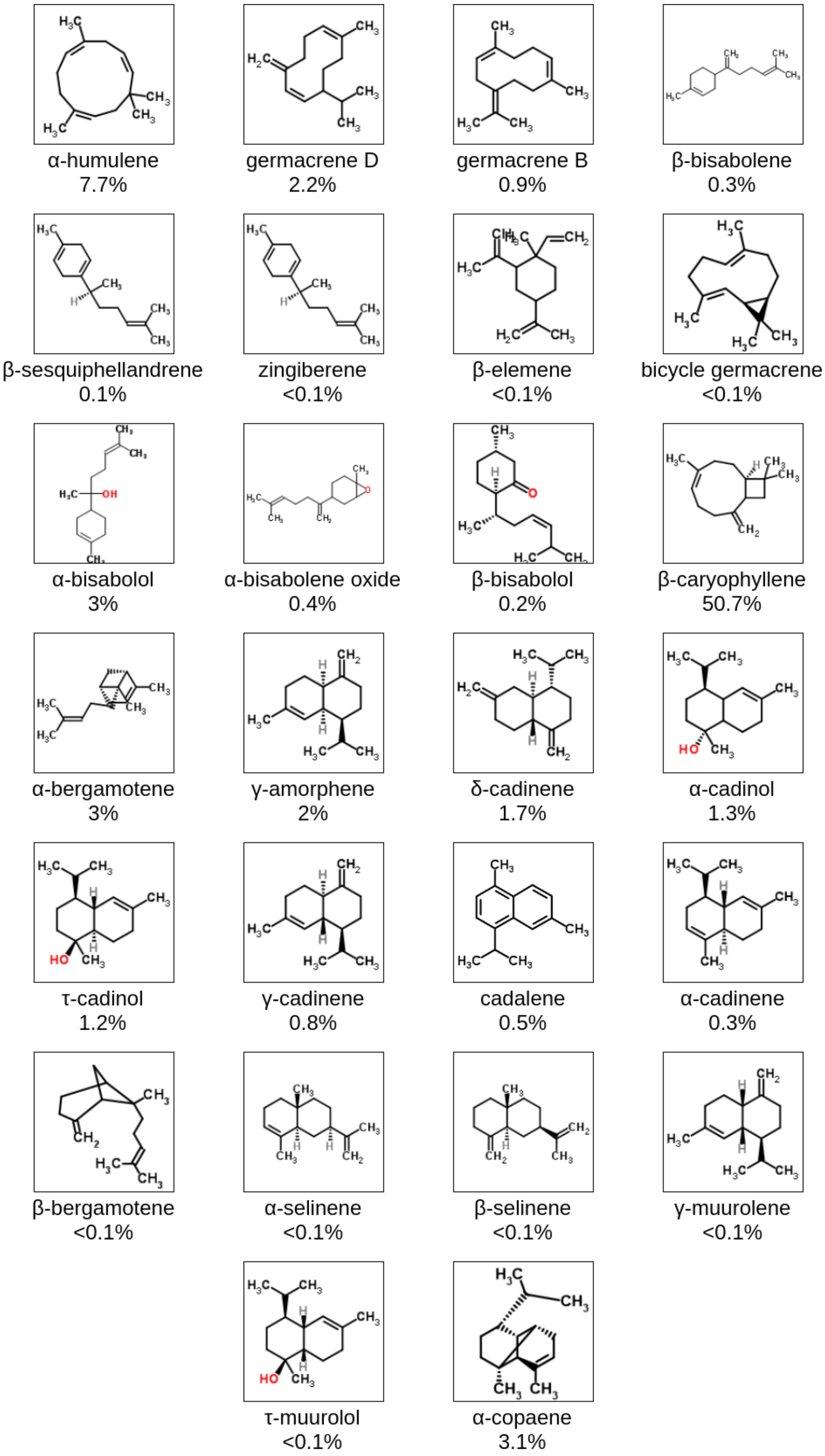
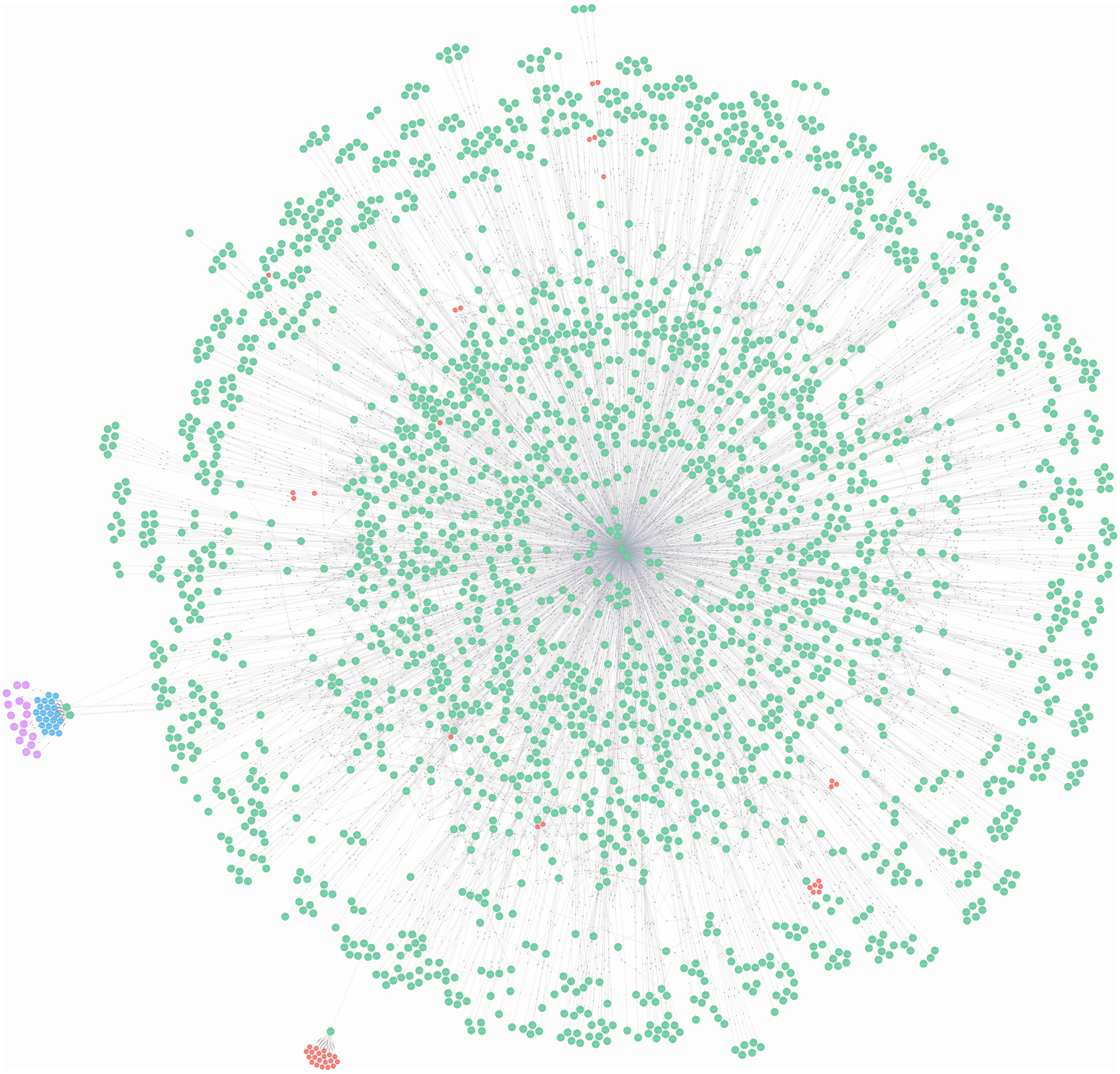


Figure 4.2: Generated chemical network using the workflow step presented in Section 3.2. Blue compounds are electrically stable, while the red vertices represent cations.



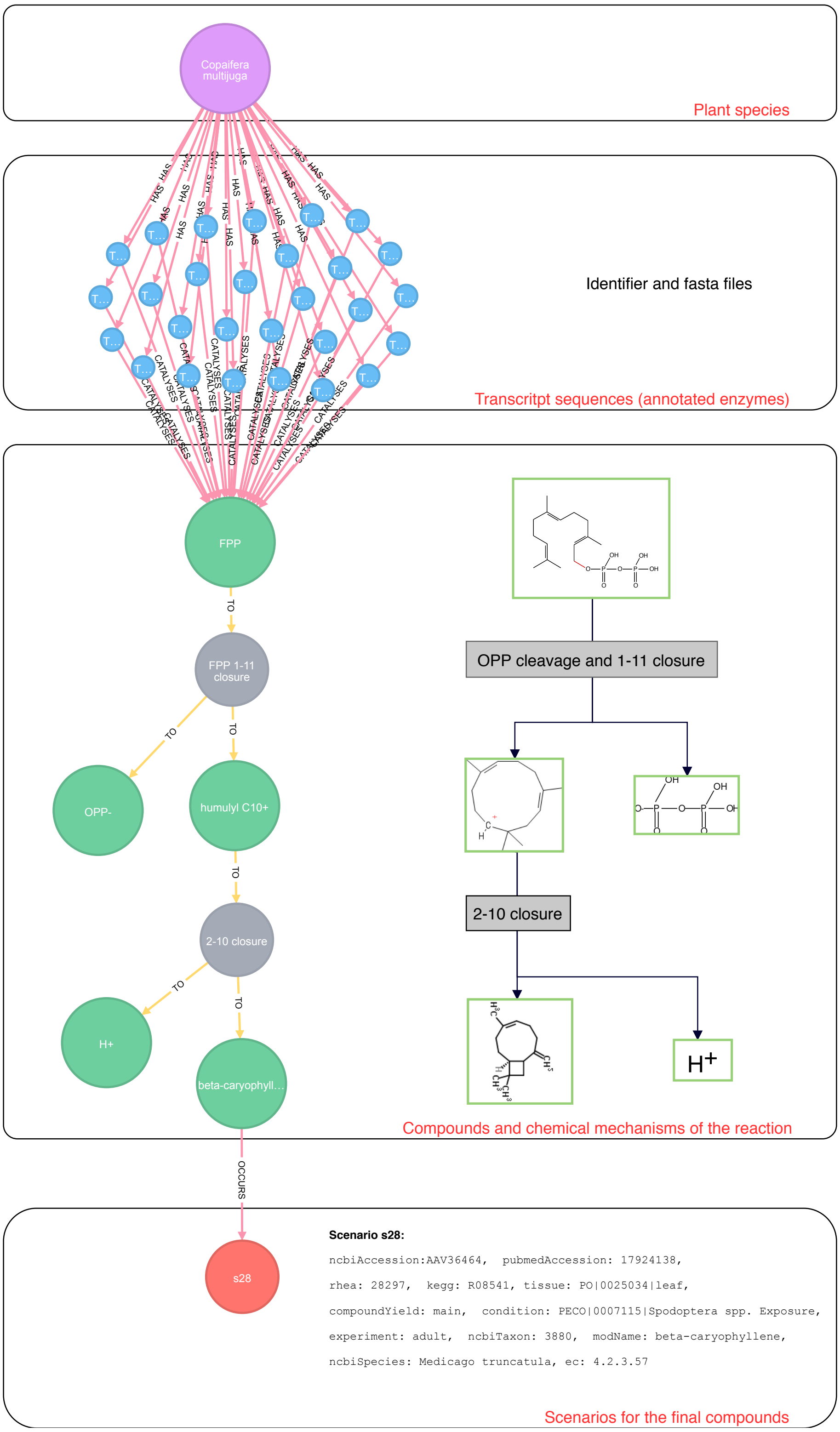
α-humulene 7.7%
germacrene D 2.2%
germacrene B 0.9%
β-bisabolene 0.3%
β-sesquiphellandrene 0.1%
zingiberene <0.1%
β-elemene <0.1%
bicycle germacrene <0.1%
α-bisabolol 3%
α-bisabolene oxide 0.4%
β-bisabolol 0.2%
β-caryophyllene 50.7%
α-bergamotene 3%
γ-amorphene 2%
δ-cadinene 1.7%
α-cadinol 1.3%
τ-cadinol 1.2%
γ-cadinene 0.8%
cadalene 0.5%
α-cadinene 0.3%
β-bergamotene <0.1%
α-selinene <0.1%
β-selinene <0.1%
γ-muurolene <0.1%
τ-muurolol <0.1%
α-copaene 3.1%

Figure 4.3: CmH sesquiterpenes reached in the generated hypergraph using the set of rules presented in 3.1.



```
MATCH (n)-[r]-(m) return n,m,r;
```

Figure 4.4: Complete 2Path15 sesquiterpenes metabolic network. Purple vertices represent the organisms, blue vertices represent enzymes linked to the compounds FPP or NPP (green vertices) indicating their enzymatic activity. The red vertices represent the experimental scenarios for the biosynthesis of final compounds.


Copaifera multijuga
Plant species
HAS
T...
Identifier and fasta files
CATALYSES
Transcritpt sequences (annotated enzymes)
FPP
TO
FPP 1-11 closure
OPP-
humulyl C10+
2-10 closure
H+
beta-caryophyll...
OPP cleavage and 1-11 closure
2-10 closure
H^+
Compounds and chemical mechanisms of the reaction
OCCURS
s28
Scenarios for the final compounds

Scenario s28:

ncbiAccession:AAV36464, pubmedAccession: 17924138,

rhea: 28297, kegg: R08541, tissue: PO|0025034|leaf,

compoundYield: main, condition: PECO|0007115|Spodoptera spp. Exposure,

experiment: adult, ncbiTaxon: 3880, modName: beta-caryophyllene,

ncbiSpecies: Medicago truncatula, ec: 4.2.3.57

```
OPTIONAL MATCH
p1=(target:Compound)-[:OCCURS]->(n:Scenario),
p2=SHORTESTPATH((source:Compound)-[r:TO*..12]->(target:Compound)),
p3=(o:Organism{ncbiTaxon:"327897"})-[*]->(source:Compound)
WHERE source.modName="FPP" AND target.modName IN ["beta-caryophyllene","H+", "OPP-"]
AND n.compoundYield="main" AND n.condition=~"(?i).*Spodoptera.*"
RETURN p1,p2,p3
```

Figure 4.5: CmH pathway for the sesquiterpene β-caryophyllene as main compound in response to *Spodoptera spp.* Exposure.

Chapter 5

Conclusions

This work presents an *in silico* sesquiterpenes metabolic network of *Copaifera multijuga* Hayne (CmH). This network covers the chemical mechanisms of the enzymatic reactions, automatic annotation of predicted compounds and scenarios supporting their biosynthesis. These chemical mechanisms, are essentially a chemical network abstracted as a hypergraph built using graph grammar rules to simulate the transformation of molecules during a reaction. This chemical network was completed and combined with other information through a well-defined workflow to form the *in silico* sesquiterpenes metabolic network of *Copaifera multijuga* Hayne (CmH).

The workflow for the *in silico* reconstruction is a chain of program executions for prediction and annotation of compounds, which fill a graph database together with metadata from biosynthesis scenarios and a subsequent insertion of the organism of interest in this context. The resulting graph database was called 2Path15 in an allusion to the 15 carbons skeleton of sesquiterpenes. The metabolic network was stored under a schema particularly developed for this purpose. Thus, while using the same schema, the database can be updated independently of the generated network. The storing of the metabolic network in a database enables efficient data management and knowledge discovery.

The results obtained from the *in silico* reconstruction of sesquiterpenes metabolic network of CmH, as well as the potential outcomes that can be achieved using the proposed workflow, have significance for the generation of biological knowledge, since the application of these results brings an inherent potential for basic science, biotechnological applications, and sustainable exploitation of the CmH. In addition, designing putative metabolic pathways is of great interest in synthetic biology (Hadadi and Hatzimanikatis, 2015).

5.1 Contributions

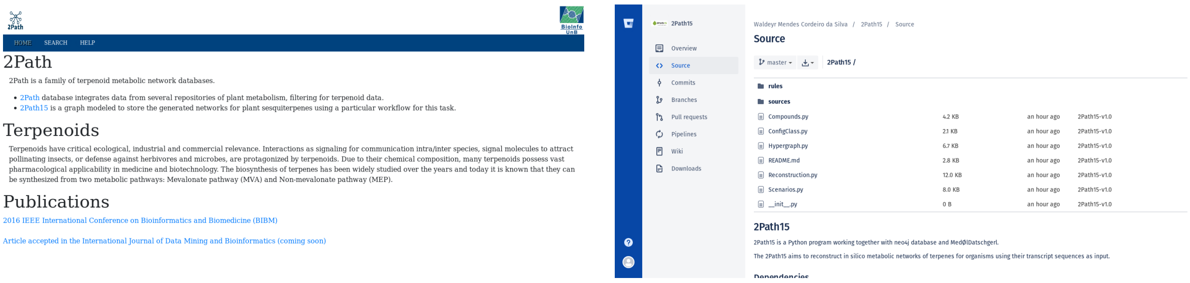
The contributions are related, but not limited, to the achievement of the specific objectives.

- A set of graph grammar rules for the generation of a chemical network representing the chemical mechanisms of the CmH sesquiterpenes biosynthesis reactions
- A workflow for the *in silico* reconstruction of metabolic networks based on the generated chemical network
- Design and implementation of a graph database schema to store the reconstructed network
- The workflow as a public and available computational tool
- A set of predicted sesquiterpene synthases of *Copaifera multijuga* Hayne (CmH) (approximately 80% of the sesquiterpenes identified diversity)
- Scenarios for experiment design on the predicted sesquiterpene synthases

5.2 Availability

2Path general information, 2Path15 code, list of predicted compounds and more is available at:

`http://www.2path.org`



2Path

2Path is a family of terpenoid metabolic network databases.

- 2Path database integrates data from several repositories of plant metabolism, filtering for terpenoid data.
- 2Path15 is a graph modeled to store the generated networks for plant sesquiterpenes using a particular workflow for this task.

Terpenoids

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).

Publications

2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

Article accepted in the International Journal of Data Mining and Bioinformatics (coming soon)

(a) 2Path main page.

Waldeyr Mendes Cordeiro da Silva / 2Path15 / Source

Source

rules			
sources			
Compounds.py	4.2 KB	an hour ago	2Path15-v1.0
ConfigClass.py	2.1 KB	an hour ago	2Path15-v1.0
Hypergraph.py	6.7 KB	an hour ago	2Path15-v1.0
README.md	2.8 KB	an hour ago	2Path15-v1.0
Reconstruction.py	12.0 KB	an hour ago	2Path15-v1.0
Scenarios.py	8.0 KB	an hour ago	2Path15-v1.0
__init__.py	0 B	an hour ago	2Path15-v1.0

2Path15

2Path15 is a Python program working together with neo4j database and MedØlDatschgerl.

The 2Path15 aims to reconstruct in silico metabolic networks of terpenes for organisms using their transcript sequences as input.

Dependencies

(b) 2Path15 git page.

Figure 5.1: 2Path versions available on-line.

5.3 Future work

The database schema supports the expansion of pathways to sets of different reactions of terpene synthases. For example, by expanding the set of graph grammar rules, it is possible to simulate the production of mono-, and sesquiterpenes which use different cyclizations from those proposed for the CmH. It is also possible to simulate the entire pathway showed in Figure 4.1 for terpenoid biosynthesis from MVA and MEP pathways by changing the sets of graph grammar rules and initial compounds.

This is precisely what the 2Path database (Silva et al., 2017) is intended to be - a database that at the same time integrates information from other data sources and stores new terpenoid biosynthesis information covering chemical mechanisms and scenarios.

In addition, and already being developed and tested (Esteves et al., 2018), there is a Web interface to facilitate the work of database users who do not have Cypher query language skills. The complete Web system will cover *in silico* reconstruction from transcriptome experiments with multiple conditions and genomic DNA. In this interface, the user will be able to generate the *in silico* metabolic networks by uploading the genomic data and choosing a set of rules.

References

S. F. Altschul et al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997. 54, 86

J. L. Andersen, C. Flamm, D. Merkle, and P. F. Stadler. Inferring chemical reaction patterns using rule composition in graph grammars. *Journal of Systems Chemistry*, 4 (1):4, 2013. 3, 19, 23

J. L. Andersen, C. Flamm, D. Merkle, and P. F. Stadler. Generic strategies for chemical space exploration. *International journal of computational biology and drug design*, 7 (2-3):225–258, 2014. 23, 49

J. L. Andersen, C. Flamm, D. Merkle, and P. F. Stadler. A software package for chemically inspired graph transformation. In *International Conference on Graph Transformation*, pages 73–88. Springer, 2016. 23, 86

J. L. Andersen, C. Flamm, D. Merkle, and P. F. Stadler. Chemical Transformation Motifs — Modelling Pathways as Integer Hyperflows. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, volume 5963, pages 1–14, 2017. 3, 21, 62, 84, 86

G. I. Arimura et al. Herbivore-induced terpenoid emission in Medicago truncatula: Concerted action of jasmonate, ethylene and calcium signaling. *Planta*, 227(2):453–464, 2008. 12

A. Baher et al. Integrated database to support research on escherichia coli. *Argonne Technical Report*, ANL92/1, 1992. 17

I. Balaur et al. Epigenet: a graph database of interdependencies between genetic and epigenetic events in colorectal cancer. *Journal of Computational Biology*, 2016. 26, 28

A. P. M. R. Bastos. *Análise cromatográfica, morfológica e molecular da síntese do oleoresina em plantas jovens de Copaifera multijuga Hayne (FABACEAE – CAESALPINIOIDEAE).* PhD thesis, UFAM, 2011. 53, 86

Carlo Batini, Stefano Ceri, Shamkant B Navathe, et al. *Conceptual database design: an Entity-relationship approach*, volume 116. Benjamin/Cummings Redwood City, CA, 1992. 52

S. Bazzani. Promise and reality in the expanding field of network interaction analysis: Metabolic networks. *Bioinformatics and Biology Insights*, 8(phenotype I):83–91, 2014. 2, 3, 16, 83, 84

J. Bohlmann, G. Meyer-Gauen, and R. Croteau. Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proceedings of the National Academy of Sciences*, 95 (8):4126–4133, 1998. 13

J. Adrian Bondy, U. S. R Murty, et al. *Graph theory with applications*, volume 290. Citeseer, 1976. 18

V. Bonnici et al. Comprehensive Reconstruction and Visualization of Non-Coding Regulatory Networks in Human. *Frontiers in Bioengineering and Biotechnology*, 2(December): 69, 2014. ISSN 2296-4185. 26

V. Bonnici et al. Arena-Idb : a platform to build human non-coding RNA interaction networks, 2018. 27, 28

F. Boyer and A. Viari. Ab initio reconstruction of metabolic pathways. *Bioinformatics*, 19, 2003. 16, 17

M. V. H. Brito et al. Copaiba oil effect on urea and creatinine serum levels in rats submitted to kidney ischemia and reperfusion syndrome. *Acta cirúrgica brasileira / Sociedade Brasileira para Desenvolvimento Pesquisa em Cirurgia*, 20(3):243–6, 2005. 1, 8, 82

H. D. Brum et al. Copaíba-roxa, Copaifera mutlijuga Hayne. Technical report, INPA, Manaus-AM, 2009. 6, 7

R. Caspi et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 38:D473–D479, 2009. 2, 15, 83

Ron Caspi et al. The metacyc database of metabolic pathways. *Nucleic Acids Research*, 42(D1):471–480, 2014. 3, 16, 84

F. Chen et al. Biosynthesis and emission of terpenoid volatiles from arabidopsis flowers. *The Plant Cell*, 15(2):481–494, 2003. 12

F. Chen et al. Characterization of a root-specific arabidopsis terpene synthase responsible for the formation of the volatile monoterpene 1, 8-cineole. *Plant Physiology*, 135(4): 1956–1966, 2004. 12

F. Chen et al. The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant Journal*, 66(1):212–229, 2011. 2, 13, 14, 83

H. Chen et al. Positive darwinian selection is a driving force for the diversification of terpenoid biosynthesis in the genus oryza. *BMC plant biology*, 14(1):239, 2014. 14, 57, 58

D. W. Christianson. Structural biology and chemistry of the terpenoid cyclases. *Chemical Reviews*, 106(8):3412–3442, 2006. 12

D. W. Christianson. Structural and Chemical Biology of Terpenoid Cyclases. *Chemical Reviews*, 117(17):11570–11648, 2017. 2, 3, 15, 33, 83, 84

S. M. Colby et al. Germacrene C synthase from Lycopersicon esculentum cv. VFNT cherry tomato: cDNA isolation, characterization, and bacterial expression of the multiple product sesquiterpene cyclase. *Proceedings of the National Academy of Sciences of the United States of America*, 95(5):2216–2221, 1998. 33

J. Corbacho 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. 26, 28

A. Corbellini et al. Persisting big-data: The nosql landscape. *Information Systems*, 63: 1–23, 2017. 24

J. A. S. Costa. *Copaifera multijuga Hayne.* jardim botânico do rio de janeiro. REFLORA - Jardim Botânico do Rio de Janeiro, 2018. URL http://reflora.jbrj.gov.br/reflora/floradobrasil/FB82967. 5

R. L. Costa et al. Gennet: an integrated platform for unifying scientific workflows and graph databases for transcriptome data analysis. *PeerJ*, 5:e3509, 2017. 27, 28

DB-Engines. Db-engines ranking - trend of graph dbms popularity, 2018. URL https://db-engines.com/en/ranking/graph+dbms. 25

A. G. De Brevern et al. Trends in IT innovation to build a next generation bioinformatics solution to manage and analyse biological big data produced by NGS Technologies. *BioMed Research International*, 2015, 2015. 23

J. W. de Kraker et al. (+)-Germacrene A biosynthesis . The committed step in the biosynthesis of bitter sesquiterpene lactones in chicory. *Plant physiology*, 117(4):1381–92, 1998. 33

N. de Matos Gomes et al. Characterization of the antinociceptive and anti-inflammatory activities of fractions obtained from copaifera multijuga hayne. *Journal of ethnopharmacology*, 128(1):177–183, 2010. 1, 8, 82

J. Degenhardt, T. G. Köllner, and J. Gershenzon. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry*, 70(15-16): 1621–1637, 2009. 2, 15, 33, 57, 83

E. Demir et al. The biopax community standard for pathway data sharing. *Nature biotechnology*, 28(9):935, 2010. 3, 84

R. J. A. Deus, C. N. Alves, and M. S. P. Arruda. Avaliação do efeito antifúngico do óleo resina e do óleo essencial de copaíba (Copaifera multijuga Hayne). *Revista Brasileira de Plantas Medicinais*, 13(1):1–7, 2011. 1, 8, 82

R. J. A. Deus et al. In vitro fungitoxic effect of the oil resin and the essential oil of copaiba (Copaifera multijuga Hayne). *Revista Brasileira de Plantas Medicinais*, 11(3):347–353, 2009. 1, 8, 82

P. M. Dewick et al. The biosynthesis of C5–C25 terpenoid compounds. *Natural Product Reports*, 19(2):181–222, 2002. 1, 82

J. S. Dickschat. Bacterial terpene cyclases. *Nat. Prod. Rep.*, 33(1):87–110, 2016. 3, 33, 84

A. O. Dos Santos et al. Antimicrobial activity of Brazilian copaiba oils obtained from different species of the Copaifera genus. *Memorias do Instituto Oswaldo Cruz*, 103(3): 277–281, 2008. 1, 8, 82

J. M. Dreyfuss et al. Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus neurospora crassa using farm. *PLoS computational biology*, 9(7):e1003126, 2013. 17

N. Dudareva et al. (e)-β-ocimene and myrcene synthase genes of floral scent biosynthesis in snapdragon: function and expression of three terpene synthase genes of a new terpene synthase subfamily. *The Plant Cell*, 15(5):1227–1241, 2003. 13

G. Esteves, W. M. C. Silva, et al. Human-computer interaction communicability evaluation method applied to bioinformatics. In *World Conference on Information Systems and Technologies*, pages 1001–1008. Springer, 2018. 70, 87

A. Fabregat et al. Reactome graph database: Efficient access to complex pathway data. *PLoS Computational Biology*, 14(1):1–13, 2018. 16, 17, 25, 27, 28, 63

J. D Fischer, G. L Holliday, and J. M. Thornton. The cofactor database: organic cofactors in enzyme catalysis. *Bioinformatics*, 26(19):2496–2497, 2010. 11

J Förster, Iman Famili, and Patrick Fu. Genome-scale reconstruction of the saccharomyces cerevisiae metabolic network. *Genome Research*, pages 244–253, 2003. 17

T. Gaasterland and E. Selkov. Reconstruction of metabolic networks using incomplete information. *International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 3:127–135, 1995. 17

G. Gallo, G. Longo, S. Pallottino, and S. Nguyen. Directed hypergraphs and applications. *Discrete Applied Mathematics*, 42(2-3):177–201, 1993. 21

Y. Gao, R. B. Honzatko, and R. J. Peters. Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Natural Product Reports*, 29(10):1153, 2012. 13

S. Garms, T. G. Köllner, and W. Boland. A multiproduct terpene synthase from medicago truncatula generates cadalane sesquiterpenes via two different mechanisms. *Journal of Organic Chemistry*, 75(16):5590–5600, 2010. 33

A. Goesmann et al. Pathfinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics*, 18(1):124–129, 2002. 16, 17

N. de M. Gomes et al. Antineoplasic activity of Copaifera multijuga oil and fractions against ascitic and solid Ehrlich tumor. *Journal of Ethnopharmacology*, 119(1):179–184, 2008. 1, 8, 82

N. M. Gomes et al. Antinociceptive activity of Amazonian Copaiba oils. *Journal of Ethnopharmacology*, 109(3):486–492, 2007. 1, 8, 82

S. Grossetête, B. Labedan, and O. Lespinet. Fungipath: a tool to assess fungal metabolic pathways predicted by orthology. *BMC genomics*, 11:81, 2010. 16, 17

B. J. Haas et al. De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nature protocols*, 8(8):1494, 2013. 53, 58, 86

N. Hadadi and V. Hatzimanikatis. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways, 2015. 68

E. D. Harris. *Biochemical Facts behind the definition and Properties of Metabolites.* Texas A&M University, 2013. 11

C. T. Have, L. J. Jensen, and J. Wren. Are graph databases ready for bioinformatics? *Bioinformatics*, 29(24):3107–3108, 2013. 26, 28

R. Henkel, O. Wolkenhauer, and D. Waltemath. Combining computational models, semantic annotations and simulation experiments in a graph database. *Database*, 2015, 2015. 26, 28

M. J. Herrgård et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology*, 26(10):1155–1160, 2008. 17

M. Himsolt. Gml: A portable graph file format, 1997. 20, 32

G. L. Holliday et al. Macie: exploring the diversity of biochemical reactions. *Nucleic acids research*, 40(D1):D783–D789, 2012. 3, 17, 84

Y. J. Hong and D. J. Tantillo. Consequences of conformational preorganization in sesquiterpene biosynthesis: theoretical studies on the formation of the bisabolene, curcumene, acoradiene, zizaene, cedrene, duprezianene, and sesquithuriferol sesquiterpenes. *Journal of the American Chemical Society*, 131(23):7999–8015, 2009. 12, 14

M. Hucka et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003. 3, 84

Michael Hucka, Andrew Finney, Herbert Sauro, Hamid Bolouri, John Doyle, and Hiroaki Kitano. The ERATO Systems Biology Workbench: An Integrated Environment for Multiscale and Multitheoretic Simulations in Systems Biology. In *Foundations of Systems Biology*, chapter 6, pages 125–143. The MIT Press, Cambridge, MA, 2001. 17

IBGE. Produção da extração vegetal e da silvicultura (pevs), 2018. URL https://sidra.ibge.gov.br/pesquisa/pevs/quadros/brasil/2016. 7, 8

M. Isegawa et al. Predicting pathways for terpene formation from first principles – routes to known and new sesquiterpenes. *Chemical Science*, 5(4):1555, 2014. 62, 63

Matt Jacobson. The jacobson laboratory at ucsf. https://jacobsonlab.wordpress.com, 2017. URL https://jacobsonlab.wordpress.com. [Online; accessed 20-December-2017]. 3, 84

K. Jørgensen et al. Metabolon formation and metabolic channeling in the biosynthesis of plant natural products. *Current Opinion in Plant Biology*, 8(3 SPEC. ISS.):280–291, 2005. 11

V. F. V. Junior and A. C. Pinto. O gênero copaifera l. *Química Nova*, 25(2):273–286, 2002. 1, 5, 6, 82

V. F. V. Junior et al. Chemical composition and anti-inflammatory activity of copaiba oils from *Copaifera cearensis Huber ex Ducke*, *Copaifera reticulata Ducke* and *Copaifera multijuga Hayne* - a comparative study. *Journal of Ethnopharmacology*, 112(2):248–254, 2007. 8, 10

M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, January 2000. ISSN 0305-1048. 3, 17, 84

P. D. Karp and S. M. Paley. Representations of metabolic knowledge: pathways. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 2:203–211, 1994. 17

C. I. Keeling et al. Functional plasticity of paralogous diterpene synthases involved in conifer defense. *Proceedings of the National Academy of Sciences*, 105(3):1085–1090, 2008. 14

D. B Kell et al. Metabolic footprinting and systems biology: The medium is the message, 2005. 2, 11, 16, 84

N. P. Keller, G. Turner, and J. W. Bennett. Fungal secondary metabolism - from biochemistry to genomics. *Nature Reviews Microbiology*, 3(12):937–947, 2005. 11

C. Kempinski, Z. Jiang, S. Bell, and J. Chappell. Metabolic engineering of higher plants and algae for isoprenoid production. *Advances in Biochemical Engineering/Biotechnology*, 2015. 13

T. G. Kollner, M. Held, C. Lenk, I. Hiltpold, T. C.J. Turlings, J. Gershenzon, and J. Degenhardt. A Maize (E)- -Caryophyllene Synthase Implicated in Indirect Defense Responses against Herbivores Is Not Expressed in Most American Maize Varieties. *the Plant Cell Online*, 20(2):482–494, 2008. URL http://www.plantcell.org/cgi/doi/10.1105/tpc.107.051672. 57, 58

T. G. Köllner et al. Protonation of a neutral (S)-β-bisabolene intermediate is involved in (S)-β-macrocarpene formation by the maize sesquiterpene synthases TPS6 and TPS11. *Journal of Biological Chemistry*, 283(30):20779–20788, 2008. 57

T. Kuzuyama. Mevalonate and nonmevalonate pathways for the biosynthesis of isoprene units. *Bioscience, biotechnology, and biochemistry*, 66(8):1619–1627, 2002. 12

B. M. Lange, T. Rujan, W. Martin, and R. Croteau. Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 97(24):13172–13177, 2000. 12

N. Le Novere. Quantitative and logic modelling of molecular and gene networks. *Nature Reviews Genetics*, 16(3):146, 2015. 2, 15, 83

L. M. Leandro et al. Chemistry and biological activities of terpenoids from copaiba (copaifera spp.) oleoresins. *Molecules*, 17(4):3866–3889, 2012. 1, 8, 82

J. Leipzig. A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics*, page bbw020, 2016. 29

C. A. Lesburg. Crystal Structure of Pentalenene Synthase: Mechanistic Insights on Terpenoid Cyclization Reactions in Biology. *Science*, 277(5333):1820–1824, 1997. 13

S. R. M. Lima, V. F. Veiga, et al. In vivo and in vitro Studies on the Anticancer Activity of Copaifera multijuga Hayne and its Fractions. *Phytotherapy Research*, 17(9):1048–1053, 2003a. 8, 10

S. R. M. Lima et al. In vivo and in vitro studies on the anticancer activity of copaifera multijuga hayne and its fractions. *Phytotherapy Research*, 17(9):1048–1053, 2003b. 1, 8, 82

W. Liu et al. Structure, function and inhibition of ent-kaurene synthase from bradyrhizobium japonicum. *Scientific reports*, 4, 2014. 2, 13, 14, 83

M. Lohr, J. Schwender, and J. E. W. Polle. Isoprenoid biosynthesis in eukaryotic phototrophs: a spotlight on algae. *Plant Science*, 185:9–22, 2012. 12

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*, 28(1):87–99, 2011. 11, 12

M. Löwe. Algebraic approach to single-pushout graph transformation. *Theoretical Computer Science*, 109(1):181 – 224, 1993. ISSN 0304-3975. 19

A. Lysenko et al. Representing and querying disease networks using graph databases. *BioData Mining*, 2016. 25, 26, 28

S. Maeda et al. Artificial Force Induced Reaction (AFIR) Method for Exploring Quantum Chemical Potential Energy Surfaces. *Chemical Record*, 16(5):2232–2248, 2016. 3, 62, 84

D. M. Martin, J. Fäldt, and J. Bohlmann. Functional characterization of nine norway spruce tps genes and evolution of gymnosperm terpene synthases of the tps-d subfamily. *Plant Physiology*, 135(4):1908–1927, 2004. 14

Regina Célia Viana Martins-da Silva. Taxonomia das espécies de copaifera l.(leguminosae-caesalpinioideae) ocorrentes na amazônia brasileira. http://floradobrasil.jbrj.gov.br/jabot/floradobrasil/FB82967, 2006. [Online; accessed 15-January-2017]. 6

D. E. Mendonça and S B. Onofre. Atividade antimicrobiana do óleo-resina produzido pela copaiba - Copaifera multijuga Hayne (Leguminosae). *Brazilian Journal of Pharmacognosy*, 19(2 B):577–581, 2009a. 1, 82

D. E. Mendonça and S. B. Onofre. Atividade antimicrobiana do óleo-resina produzido pela copaiba - Copaifera multijuga Hayne (Leguminosae). *Brazilian Journal of Pharmacognosy*, 19(2 B):577–581, 2009b. 8

P. Mercke et al. Combined transcript and metabolite analysis reveals genes involved in spider mite induced volatile formation in cucumber plants. *Plant Physiology*, 135(4): 2012–2024, 2004. 12

C. Messaoudi, M. A. Mhand, and R. Fissoune. A Performance Study of NoSQL Stores for Biomedical Data NoSQL databases : An Overview, 2018. 27, 28

A. Messina et al. BioGrakn: A knowledge graph-based semantic database for biomedical sciences. In *Advances in Intelligent Systems and Computing*, volume 611, pages 299–309. Springer, 2018. 27, 28

G. Michal and Dietman S. *Biochemical pathways.* John Willey & Sons Inc., 2012. 11

P. Minkiewicz, A. Iwaniak, and M. Darewicz. Annotation of peptide structures using SMILES and other chemical codes-practical solutions, 2017. 23

A. Morgat et al. Updates in rhea—an expert curated resource of biochemical reactions. *Nucleic acids research*, 2016. 58

D. L. Nelson, A. L. Lehninger, and M. M. Cox. *Lehninger principles of biochemistry.* Macmillan, 2008. 11

E. Oldfield and F. Y. Lin. Terpene biosynthesis: Modularity rules. *Angewandte Chemie - International Edition*, 51(5):1124–1137, 2012. 13, 14

S. G. Oliver et al. Systematic functional analysis of the yeast genome. *Trends in biotechnology*, 16(9):373–378, 1998. 11, 16

J. D. Orth, I. Thiele, and B. Ø. Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245, 2010. 3, 16, 84

O. Øyås and J. Stelling. Genome-scale metabolic networks in time and space. *Current Opinion in Systems Biology*, pages 1–8, 2017. 16

T. R. C. Pacheco, L. E. S. Barata, and M. C. T. Duarte. Antimicrobial activity of copaiba (Copaifera spp) balsams. *Revista Brasileira de Plantas Medicinais*, 8:123–124, 2006. 1, 8, 82

P. Pareja-Tobes et al. Bio4j: a high-performance cloud-enabled graph-based data platform. *bioRxiv*, 2015. 26, 28

H. E. Pence and A. Williams. Chemspider: an online chemical information resource, 2010. 52, 55, 61, 86

J. L. Peterson. *Petri Net Theory and the Modeling of Systems*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1981. ISBN 0136619835. 49

P. Pharkya, A. P. Burgard, and C. D. Maranas. Optstrain: a computational framework for redesign of microbial production systems. *Genome Research*, pages 2367–2376, 2004. 16, 17

J. W. Pinney, D. R. Westhead, and G. A. McConkey. Petri net representations in systems biology. *Biochemical Society transactions*, 31(iv):1513–1515, 2003. 17

W3C RDF. Rdf schema 1.1. on-line, 2014. URL `https://www.w3.org/TR/2014/REC-rdf-schema-20140225/`. Resource Description Framework. 3, 84

V. N. Reddy, M. L. Mavrovouniotis, and M. N. Liebman. Petri net representations in metabolic pathways. *International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 1(August):328–336, 1993. 17

O. C. Rigamonte-Azevedo, P. G. S. Wadt, and L. H. de O. Wadt. Copaíba: ecologia e produção de óleo-resina. *Embrapa Acre-Documentos (INFOTECA-E)*, 2004. 6

J. Rinkel et al. Lessons from 1,3-Hydride Shifts in Sesquiterpene Cyclizations. *Angewandte Chemie - International Edition*, 55(43):13593–13596, 2016. 33

D. Ro et al. Microarray expression profiling and functional characterization of attps genes: duplicated arabidopsis thaliana sesquiterpene synthase genes at4g13280 and at4g13300 encode root-specific and wound-inducible (z)-γ-bisabolene synthases. *Archives of Biochemistry and Biophysics*, 448(1):104–116, 2006. 12

I. Robinson, J. Webber, and E. Eifrem. *Graph Databases*. O'Reilly Media, Inc., Sebastopol, CA, USA, 2013. 24

D. J. S. Sandbeck, D. J. Markewich, and A. L. L. East. The Carbocation Rearrangement Mechanism, Clarified. *Journal of Organic Chemistry*, 81(4):1410–1415, 2016. 33

A. O. Santos et al. Effect of brazilian copaiba oils on leishmania amazonensis. *Journal of ethnopharmacology*, 120(2):204–208, 2008. 1, 8, 82

A. Schifrin et al. A single terpene synthase is responsible for a wide variety of sesquiterpenes in sorangium cellulosum soce56. *Organic & biomolecular chemistry*, 14(13):3385–3393, 2016. 2, 12, 15, 83

C. Schnee et al. The products of a single maize sesquiterpene synthase form a volatile defense signal that attracts natural enemies of maize herbivores. *Proceedings of the National Academy of Sciences of the United States of America*, 103(4):1129–1134, 2006. 12

E. E. Selkov et al. Factographic data bank on enzymes and metabolic pathways. *Studia Biophysica*, 129(2-3):155–164, 1989a. 17

E. E. Selkov et al. Factographic data bank on enzymes and metabolic pathways. *Studia Biophysica*, 129(2-3):155–164, 1989b. 16

W M. C. da Silva et al. A terpenoid metabolic network modelled as graph database. *International Journal of Data Mining and Bioinformatics*, 18(1):74–90, 2017. 16, 17, 25, 27, 28, 55, 59, 70, 87

B. Singh and R. A. Sharma. Plant terpenes: defense responses, phylogenetic analysis, regulation and clinical applications. *3 Biotech*, 5(2):129–151, 2015. 12, 14

C. L. Steele, J. Crock, J. Bohlmann, and R. Croteau. Sesquiterpene synthases from grand fir (Abies grandis): Comparison of constitutive and wound-induced activities, and cDNA isolation, characterization, and bacterial expression of δ-selinene synthase and γ- humulene synthase. *Journal of Biological Chemistry*, 273(4):2078–2089, 1998. 33

G. Summer et al. cyneo4j: connecting neo4j and cytoscape. *Bioinformatics*, 31(23): 3868–3869, 2015. 26

G. Summer et al. The Network Library: a framework to rapidly integrate network biology resources. *Bioinformatics*, 32(17):i473–i478, sep 2016. 27, 28

N. Swainston et al. biochem4j: Integrated and extensible biochemical knowledge through graph databases. *PloS one*, 12(7):e0179130, 2017. 27, 28

D. Szklarczyk et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, 45(D1):D362–D368, 2017. 26

K. Takeda. Stereospecific cope rearrangement of the germacrene-type sesquiterpenes. *Tetrahedron*, 30(12):1525–1534, 1974. 33

D. Tholl. Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Current opinion in plant biology*, 9(3):297–304, 2006. 2, 12, 15, 83

D. Tholl et al. Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from arabidopsis flowers. *The Plant Journal*, 42(5):757–771, 2005. 2, 12, 83

B. J. Townsend. Antisense Suppression of a (+)- -Cadinene Synthase Gene in Cotton Prevents the Induction of This Defense Response Gene during Bacterial Blight Infection But Not Its Constitutive Expression. *Plant Physiology*, 138(1):516–528, 2005. 33

G. Van Erven, W. m. C. da Silva, , R. Carvalho, and M. Holanda. Graphed: A graph description diagram for graph databases. In *Trends and Advances in Information Systems and Technologies*, pages 1141–1151. Springer International Publishing, 2018. 25, 50, 86

A Vattekkatte, S. Garms, W. Brandt, and W. Boland. Enhanced structural diversity in terpenoid biosynthesis: enzymes, substrates and cofactors. *Organic & Biomolecular Chemistry*, 16(3):348–362, 2018. 1, 15, 33, 57, 58, 82, 83

A. Vattekkatte et al. Substrate geometry controls the cyclization cascade in multiproduct terpene synthases from <i>Zea mays</i>. *Org. Biomol. Chem.*, 13(21):6021–6030, 2015. 12

V. F. Veiga and A. C. Pinto. O Gênero Copaifera L., 2002. 5

V. F. Veiga et al. The inhibition of paw oedema formation caused by the oil of *Copaifera multijuga Hayne* and its fractions. *Journal of Pharmacy and Pharmacology*, 58(10): 1405–1410, 2006a. 8, 10

V. F. Veiga et al. The inhibition of paw oedema formation caused by the oil of copaifera multijuga hayne and its fractions. *Journal of Pharmacy and Pharmacology*, 58(10): 1405–1410, 2006b. 1, 8, 82

V. F. Veiga et al. Chemical composition and anti-inflammatory activity of copaiba oils from copaifera cearensis huber ex ducke, copaifera reticulata ducke and copaifera multijuga hayne - a comparative study. *Journal of Ethnopharmacology*, 112(2):248–254, 2007. 1, 8, 82

L. Wang et al. A review of computational tools for design and reconstruction of metabolic pathways. *Synthetic and Systems Biotechnology*, 2(4):243–252, 2017. 2, 83, 84

F. L. Westphal et al. Evaluation of the pleuropulmonary alterations after injection of copaiba oil, aqueous extract of crajiru and iodine pvp in the pleural space of mice. *Revista do Colégio Brasileiro de Cirurgiões*, 34(3):170–176, 2007. 1, 82

M. D. Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016. 3, 61, 84

M. Wink. *Biochemistry of Plant Secondary Metabolism*, volume 40. John Willey & Sons Inc., 2010. 1, 2, 11, 12, 83

J. S. Yuan et al. Molecular and genomic basis of volatile-mediated indirect defense against insects in rice. *The Plant Journal*, 55(3):491–503, 2008. 12

F. Zhang et al. Protonation-dependent diphosphate cleavage in fpp cyclases and synthases. *ACS Catalysis*, 6(10):6918–6929, 2016. 2, 3, 14, 83, 84

X. Zhuang et al. Dynamic evolution of herbivore-induced sesquiterpene biosynthesis in sorghum and related grass crops. *The Plant Journal*, 69(1):70–80, 2012. 14

Appendix I

Resumo Estendido

I.1 Introdução

Plantas do gênero *Copaifera* (*Leguminosae-Caesalpinoideae*), comumente chamadas de "Copaíba", crescem abundantemente no Brasil e em vários outros países da América do Sul. A *Copaifera multijuga* Hayne (CmH) é nativa da Amazônia, porém não endêmica ao Brasil, embora ocorra em toda sua região norte. Extraído do tronco das árvores, o óleo-resina das espécies de *Copaifera spp.* é amplamente utilizado na medicina popular e por indígenas da região amazônica para cura (Junior and Pinto, 2002).

Além disso, o óleo-resina de *Copaifera spp.* tem alto potencial biotecnológico associado e tem sido estudado para aplicações antimicrobianas (Dos Santos et al., 2008; Mendonça and Onofre, 2009a; Pacheco et al., 2006), antifúngicas (Deus et al., 2011, 2009), anti-inflamatórias (Brito et al., 2005; de Matos Gomes et al., 2010; Veiga et al., 2006b, 2007), antitumorais (Gomes et al., 2008; Lima et al., 2003b), antinociceptivas (de Matos Gomes et al., 2010; Gomes et al., 2007), antileishmanial (Santos et al., 2008) e cicatrizante (Westphal et al., 2007). O óleo-resina de *Copaifera spp.*, inclusa a CmH, é composto de ácidos resinosos e compostos voláteis, principalmente sesqui e diterpenos (Leandro et al., 2012).

Os terpenos são um grupo grande e variado de produtos naturais que desempenham importantes papéis ecológicos, como defesa e comunicação, além de várias aplicações na indústria e na medicina. Eles são produzidos por uma variedade de organismos como plantas, fungos e bactérias, através de reações metabólicas catalisadas por *Terpene synthases (TPSs)* (Dewick et al., 2002).

Unidades de isopreno com 5 carbonos ($C5$), *Isopentenyl Pyrophosphate (IPP)* e *Dimethylallyl Pyrophosphate (DMAPP)*, são os principais substratos para toda a diversidade de terpenos (Vattekkatte et al., 2018). O alongamento da cadeia de carbonos pela adição de unidades de $C5$ isoprene dá origem a *Geranyl Diphosphate (GPP)* ($C10$), *Farnesyl*

Diphosphate (FPP) ($C15$) e *Geranylgeranyl Diphosphate (GGPP)* ($C20$) (Vattekkatte et al., 2018).

GPP é o substrato para monoterpenos, FPP para sesquiterpenos e $GGPP$ para diterpenos. Entretanto, FPP e $GGPP$ também podem ser dimerizados para formar os precursores de $C30$ e $C40$ terpenes (Wink, 2010). Dependendo da quantidade de unidades de isoprenos, os terpenos são nomeados como monoterpenos ($C10$), sesquiterpenos ($C15$), diterpenos ($C20$), sesterterpenos ($C25$), triterpenos ($C30$), tetraterpenes ($C40$) e politerpenos ($\geq C40$) (Wink, 2010).

As TPSs exibem uma ampla atividade catalítica e podem originar diversos produtos partindo de um mesmo substrato (Schifrin et al., 2016; Tholl et al., 2005). Há duas classes de TPSs: Classe I e Classe II, definidas por aminoácidos que formam os seus sítios catalíticos (Chen et al., 2011) (Liu et al., 2014). FPP é o precursor pivô dos sesquiterpenos através da ação de TPSs Classe I (Zhang et al., 2016).

Independentemente do produto, a biossíntese dos sesquiterpenos a partir do FPP inicia-se pela clivagem do *Diphosphate (OPP)*, a qual é predominantemente dependente de Mg^{2+} (Zhang et al., 2016). O cátion de FPP resultante da clivagem, pode levar diretamente à produção de sesquiterpenos, ou pode sofrer uma rotação seguida da reincorporação do OPP formando um cisoide ou transoide *Nerolidyl Diphosphate (NPP)* (Tholl, 2006). O NPP, também pode ter o OPP clivado e o cátion de NPP resultante, também pode levar à produção de sesquiterpenos.

Os mecanismos químicos de biossíntese de sesquiterpenos envolvem a formação de ligação $C-C$, cátions intermiários, rearranjos de Wagner-Meerwein, captura de carbocátions por moléculas de água e *shifts* de hidrogênio e grupos metil e alil, causados por mudanças conformacionais nos cátions (Degenhardt et al., 2009; Schifrin et al., 2016; Tholl, 2006). Apesar da enorme quantidade de combinações de ciclizações possíveis, a variedade de compostos resultantes inicia-se com quatro grupos iniciais de ciclização: $C1 - C10$, $C1 - C11$, $C1 - C6$, and $C1 - C7$ (Christianson, 2017).

Redes metabólicas *in silico* constituem o núcleo da Biologia de Sistemas. Elas são modelos computacionais representativos das reações de biossíntese realizadas pelas células, incluindo interações entre compostos, enzimas, cofatores e outras moléculas em um organismo (Bazzani, 2014). Técnicas e ferramentas computacionais, dados ômicos[1] e da literatura são empregados para a reconstrução de redes metabólicas (Caspi et al., 2009; Wang et al., 2017).

A reconstrução *in silico* de redes metabólicas é dependente da quantidade e qualidade dos dados ômicos (Le Novere, 2015). Em geral, os métodos para reconstrução de redes metabólicas inferem um metaboloma a partir de um genoma e de informação pré-existente

[1] genômica, transcritômica, metabolômica, entre outras

sobre reações metabólicas disponível em bancos de dados Wang et al. (2017). Em organismos não-modelo, para os quais os dados ômicos podem não ser abundantes, o perfil metabólico é uma alternativa para a reconstrução (Kell et al., 2005).

Outra abordagem para reconstrução de redes metabólicas é a predição de compostos e reações a partir de simulações computacionais. Entre as abordagens para geração de redes químicas através de simulações computacionais destacam-se *Artificial Force-Induced Reaction (AFIR)* (Maeda et al., 2016) e *Modelling Pathways as Integer Hyperflows* (Andersen et al., 2017).

Em relação ao foco e nível de detalhes, uma rede metabólica pode ser reconstruída e explorada com objetivos qualitativos, quantitativos ou ambos. Simulações quantitativas em uma rede metabólica podem estimar quantidades do metaboloma, caso em que o método *Flux Balanced Analysis (FBA)* é amplamente utilizado (Orth et al., 2010). Os resultados qualitativos esperados em uma rede metabólica incluem identificação de enzimas, reações, condições que influenciam a formação do metaboloma como localização celular ou tecido, interações com outras biomoléculas e outros mais (Bazzani, 2014). O nível de detalhe dos componentes de uma rede metabólica pode variar, incluindo camadas de conhecimento especializado em cada componente da rede como no caso das redes químicas, onde as reações têm seus mecanismos químicos, compostos iniciais, intermediários e finais representados explicitamente.

Redes metabólicas podem ser armazenadas em arquivos de vários formatos como BioPax (Demir et al., 2010), RDF (RDF, 2014), e SBML (Hucka et al., 2003). Apesar da versatilidade funcionalidade dos arquivos estruturados, bancos de dados permitem o gerenciamento de coleções maiores e mais complexas. Existem diversos bancos de dados de metabolismo como o KEGG (Kanehisa and Goto, 2000) ou o MetaCyc (Caspi et al., 2014). Tais bancos de dados armazenam reações metabólicas enzimáticas compostas de etapas protagonizadas por mecanismos químicos nem sempre explícitos. Mesmo existindo um considerável volume de conhecimento sobre estes mecanismos na literatura, como em (Christianson, 2017), (Dickschat, 2016), e (Zhang et al., 2016), há poucos repositórios especializados disponíveis. Dois exemplos expressivos são o *Jacob Blog* (Jacobson, 2017), que contém um catálogo de esquemas de ciclizações e o banco de dados MACiE (Holliday et al., 2012), cuja cobertura de liases[2] é de aproximadamente 6%.

Neste cenário, um modelo de dados consistente e abrangente para armazenar redes metabólicas corrobora com os princípios FAIR para gerenciamento de dados científicos (Wilkinson et al., 2016). O escopo deste trabalho foi definido com o propósito de criar um *workflow* para predição e gerenciamento de conhecimento sobre a biossíntese de sesquiter-

[2] Liases: classe enzimática a qual pertencem a maioria das *TPSs*.

penos que compõem o óleo-resina da CmH no contexto da reconstrução e armazenamento de redes metabólicas *in silico*.

I.2 Problema

Indisponibilidade de uma rede metabólica *in silico* para a biossíntese de sesquiterpenos da *Copaifera multijuga* Hayne (CmH) com seus mecanismos químicos, compostos iniciais, intermediários e finais representados explicitamente.

I.3 Objetivo

O objetivo desta tese é reconstruir, armazenar e disponibilizar uma rede metabólica *in silico* para a biossíntese de sesquiterpenos presentes no óleo-resin da *Copaifera multijuga* Hayne (CmH) incluindo seus mecanismos químicos.

I.3.1 Objetivos Específicos

- Gerar uma rede química de reações de biossíntese de sesquiterpenos e seus mecanismos químicos, compostos iniciais, intermediários e finais
- Definir e construir um *workflow* para a reconstrução *in silico* de redes metabólicas baseado em uma rede química gerada
- Definir e implementar um banco de dados em grafos para armazenar a rede metabólica reconstruída
- Implementar o *workflow* como uma ferramenta publicamente disponível para a comunidade acadêmica

I.4 Método

O método consiste na acumulação de informação em camadas que completam-se mutuamente para atingir um resultado final que é a própria rede metabólica *in silico* da CmH. A informação em cada camada pode ser proveniente de interação humana ou computacional, tornando o método semi-automático. A sequência das interações está definida como um *workflow* cujas etapas são descritas a seguir.

A primeira parte consistiu em um levantamento biliográfico sobre os sesquiterpenos presentes no óleo-resina da CmH. Um segundo levantamento bibliográfico buscou identificar os mecanismos químicos das reações de biossíntese desses sesquiterpenos.

Representando as moléculas como grafos não direcionados, onde os vétices são átomos e as arestas são ligações químicas, regras de gramática de grafos foram escritas para representar as transformações que ocorrem nas moléculas durante as reações. As transformações foram simuladas computacionalmente utilizando a abordagem *Modelling Pathways as Integer Hyperflows* (Andersen et al., 2017) através do *framework* MedØlDatschgerl (Andersen et al., 2016). Partindo de um conjunto inicial de compostos precursores, FPP, NPP e H_2O, e um conjunto de regras de gramática de grafos, compostos foram preditos, incluindo aqueles identificados no primeiro levantamento bibliográfico.

Esta rede química foi armazenada no banco de dados em grafos Neo4J com um esquema particularmente desenhado para esta finalidade Van Erven et al. (2018). Partindo do banco de dados, compostos preditos foram identificados e atualizados utilizando o *Web Service* ChemSpider (Pence and Williams, 2010). Em seguida, foram levantados da literatura, cenários para a biossíntese dos sesquiterpenos preditos e anotados. Estes cenários, também inseridos no banco de dados, compreendem dados experimentais, incluindo sequências de resíduos de aminoácidos de sesquiterpeno sintases, que amparam as predições e anotações.

O próximo passo foi montar o transcritoma da CmH utilizando como fonte *reads*[3] sequenciadas usando Roche 454®. As *reads* provenieram de uma biblioteca de cDNA extraída de folhas de plantas jovens e saudáveis de CmH, coletadas no viveiro de plantas medicinais da Universidade do Amazonas (Bastos, 2011). As plantas tinham entre 1 e 1,5 m de altura, 0,5 e 3 cm de largura do caule. O software Trinity (Haas et al., 2013) foi usado para a montagem.

Os transcritos montados foram alinhados contra as sequências dos cenários armazenadas no banco de dados usando Blast (Altschul et al., 1997). As sequências do transcritoma anotadas a partir deste alinhamento foram vinculadas às vias metabólicas no banco de dados.

I.5 Resultados

A rede metabólica *in silico* de sesquiterpenos da *Copaifera multijuga* Hayne (CmH), incluindo os mecanismos químicos das reações foi reconstruída e uma visão geral de seu conteúdo é mostrada na Tabela I.1.

Esta rede metabólica *in silico* da *Copaifera multijuga* Hayne (CmH) permite explorar as vias metabólicas biossíntese dos sesquiterpenos presentes no óleo-resina, incluindo seus mecanismos e compostos químicos. Embora preditos computacionalmente, as regras de gramática de grafos usadas para predizer os compostos e reações foram escritas com base

[3] reads: fragmentos de DNA produzidos pelos sequenciadores em formato de texto.

Tabela I.1: Rede metabólica de biossíntese de sesquiterpenos da CmH em números.

Objeto	**Quantidade**
Relacionamentos	7888
Vértices	5507
Compostos	2354
Organismos	17
Aplicações de regras de gramática de grafo	3043
Sesquiterpeno sintases	50
Cenários	43
Espaço em disco	22.13 MB

na literatura especializada, assim como os cenários com resultados experimentais em que as vias metabólicas ocorrem. A maior parte dos sesquiterpenos presentes no óleo-resina foram preditos pela a simulação, ao lado de outros compostos que podem ser produzidos a partir dos mesmos mecanismos químicos. Um conjunto de 28 sequências de transcritos da CmH foram anotados como hipotéticas sesquiterpeno sintases.

Embora a rede metabólica *in silico* da *Copaifera multijuga* Hayne (CmH) seja o resultado mais evidente, não é o único. O *workflow* definido pode ser utilizado para reconstruir redes metabólicas de qualquer outra planta a partir de um arquivo fasta com seu transcriptoma. É possível ainda gerar diferentes redes para um mesmo organismo, alterando o conjunto de regras de gramática de grafos e executando novas simulações computacionais. O *workflow* é modular e cada etapa pode ser executada independentemente.

Outro resultado implícito é a consolidação do uso de bancos de dados em grafos para armazenamento de redes metabólicas como uma alternativa viável e eficiente (Silva et al., 2017). Usando o banco de dados, é possível através de buscas, combinar informações de reações, compostos, cenários, sequências de sesquiterpeno sintases para planejar experimentos biológicos.

Trabalhos futuros incluem a introdução de novas regras de gramática de grafos para expandir o *workflow* de forma a possibilitar resultados que excedam aqueles possíveis de obter usando apenas as regras para os sesquiterpenos da CmH. Paralelamente, uma interface Web está sendo desenvolvida e testada para tornar o uso do *workflow* e a exploração dos resultados menos dependente de habilidades técnicas em computação (Esteves et al., 2018).

Apêndice II

Data Dictionary

Object	**Type**	**Description**
Organism	Node	Label of nodes storing organism data
id	int	id of an organism
ncbiTaxon	String	NCBI unique taxonomy code
ncbiSpecies	String	NCBI name of the species
ncbiLineage	String	NCBI lineage of the species
Sequence	Node	Label of nodes storing sequences data
id	int	id of an sequence
ncbiAccession	String	NCBI unique code for the sequence
ncbiDescription	String	NCBI annotation for the sequence
ncbiFasta	String	NCBI sequence in fasta format
transcript	String	annotation of the sequence
transcriptFasta	String	sequence of the submitted organism in fasta format
basedOn	String	NCBI unique code for the sequence used to annotate the submitted sequence
Compound	Node	Label of nodes storing compounds data
id	int	id of an compound
modId	String	id of an compound during the hypergraph generation
modName	String	annotation of a predicted compound
modSmiles	String	smiles of a predicted compound
chemspider	String	CHEMSPIDER id for a predicted compound
commonName	String	CHEMSPIDER name for a predicted compound
molecularFormula	String	CHEMSPIDER molecular formula for a predicted compound
molecularWeight	String	CHEMSPIDER molecular weight for a predicted compound
monoisotopicMass	String	CHEMSPIDER monoisotopic mass for a predicted compound
averageMass	String	CHEMSPIDER average mass for a predicted compound
nominalMass	String	CHEMSPIDER nominal mass for a predicted compound
imageUrl	String	url of the image for a predicted compound
Scenario	Node	Label of nodes storing scenarios data
id	int	id of a Scenario

scenarioId	String	id of a Scenario in its source
ncbiTaxon	String	NCBI unique taxonomy code
ncbiSpecies	String	NCBI name of the species
ncbiAccession	String	NCBI unique code for the sequence
pubmedAccession	String	NCBI unique code for the associated publication
modName	String	annotation of a predicted compound
experiment	String	type of experiment with controlled vocabulary (examples: recombinant, adult, seedling)
tissue	String	organ or tissue of the plant with controlled vocabulary (examples: PO|0009046|flower, PO|0025034|leaf)
condition	String	condition in which the experiment was conducted with controlled vocabulary (example: PECO|0007115|Spodoptera spp. Exposure)
compoundYield	String	yield of the compound in the experiment (main or side)
ec	String	Enzyme Commission number for the reaction leading to the compound
kegg	String	KEGG database cross reference for the reaction
rhea	String	RHEA database cross reference for the reaction
iubmb	String	IUBMB cross reference for the reaction
Rule	Node	Label of nodes storing rules data
id	int	id of a rule
mergeId	String	generated id for unify the rule applied for the same chemical mechanism
modId	String	id of the rule during the hypergraph generation
modName	String	name of the rule
CATALYSES	Relationship	Label for the relationship between a Sequence and a Compound
id	int	id of a catalyse relationship
scenarioId	String	id of the scenario (the source id) where the catalysis occurs
ncbiAccession	String	NCBI unique code for the sequence
pubmedAccession	String	NCBI unique code for the associated publication
experiment	String	type of experiment with controlled vocabulary (examples: recombinant, adult, seedling)
tissue	String	organ or tissue of the plant with controlled vocabulary (examples: PO|0009046|flower, PO|0025034|leaf)
condition	String	condition in which the experiment was conducted with controlled vocabulary (example: PECO|0007115|Spodoptera spp. Exposure)

ec	String	Enzyme Commission number for the reaction leading to the compound
kegg	String	KEGG database cross reference for the reaction
rhea	String	RHEA database cross reference for the reaction
iubmb	String	IUBMB cross reference for the reaction
HAS	Relationship	Label for the relationship between an organism and a sequence
id	int	id of a organism/sequence rlationship
diff	String	the condition in which the transcript was differentially expressed
OCCURS	Relationship	Label of the relationship between a compound and scenario
id	int	id of a compound/scenario relationship
TO	Relationship	Label of relationships between a compound and a rule or between a rule and a compound
id	int	id of a compound/rule or rule/compound relationship
modName	index	ON :Compound(modName)
modName	index	ON :Rule(modName)

Apêndice III

Scenarios

Tabela III.1: Collection of scenarios for the sesquiterpenes biosynthesis stored in the graph database.

scenarioId	ncbiTaxon	ncbiSpecies	ncbiAccession	pubmedAccession	modName	experiment	tissue	condition	rendiment	ec	kegg	rhea
s1	3702	Arabidopsis thaliana	ABO09887	15918888	alpha-barbatene	adult	PO|0009046|flower	none	side	none	none	none
s2	542674	Phyla dulcis	AFR23371	22867794	alpha-bergamotene	young	PO|0025034|leaf	none	main	4.2.3.54	none	30471
s3	542674	Phyla dulcis	AFR23372	22867794	alpha-bisabolol	young	PO|0025034|leaf	none	main	4.2.3.-	none	none
s4	3702	Arabidopsis thaliana	ABO09887	15918888	alpha-chamigrene	adult	PO|0009046|flower	none	side	none	none	none
s5	542674	Phyla dulcis	AFR23368	22867794	alpha-copaene	young	PO|0025034|leaf	none	main	4.2.3.133	none	33991
s6	3702	Arabidopsis thaliana	AAO85539	15918888	alpha-copaene	adult	PO|0009046|flower	none	side	4.2.3.133	none	33991
s7	3702	Arabidopsis thaliana	ABO09887	15918888	alpha-cuprenene	adult	PO|0009046|flower	none	side	none	none	none
s8	4530	Oryza sativa	ABJ16553	17524436	alpha-curcumene	seedling	PO|0025034|leaf	PECO|0007407|methyl jasmonate exposure	side	none	none	none
s9	3659	Cucumis sativus	AAU05951	15310834	alpha-farnesene	adult	PO|0025034|leaf	PECO|0007407|methyl jasmonate exposure	main	none	none	none
s10	3880	Medicago truncatula	AAV36464	17924138	alpha-humulene	adult	PO|0025034|leaf	PECO|0007115|Spodoptera spp. Exposure	side	4.2.3.104	R08373	31898
s11	127986	Matricaria chamomilla var. Recutita	AFM43734	22682202	alpha-humulene	recombinant	none	none	side	4.2.3.104	R08373	31898
s12	4580	Zea perennis	ABY79214	18296628	alpha-humulene	recombinant	none	none	side	4.2.3.104	R08373	31898
s13	76912	Zea mays subsp. parviglumis	ABY79213	18296628	alpha-humulene	recombinant	none	none	side	4.2.3.104	R08373	31898
s14	4579	Zea mays subsp. mexicana	ABY79212	18296628	alpha-humulene	recombinant	none	none	side	4.2.3.104	R08373	31898
s15	15945	Zea luxurians	ABY79211	18296628	alpha-humulene	recombinant	none	none	side	4.2.3.104	R08373	31898
s16	112001	Zea mays subsp. huehuetenangensis	ABY79210	18296628	alpha-humulene	recombinant	none	none	side	4.2.3.104	R08373	31898
s17	4576	Zea diploperennis	ABY79209	18296628	alpha-humulene	recombinant	none	none	side	4.2.3.104	R08373	31898
s18	311405	Zingiber zerumbet	BAG12020	18273640	alpha-humulene	recombinant	none	none	main	4.2.3.104	R08373	31898
s19	4232	Helianthus annuus	AAY41422	19580670	alpha-humulene	recombinant	none	none	main	4.2.3.104	R08373	31898
s20	3702	Arabidopsis thaliana	AAO85539	15918888	alpha-humulene	adult	PO|0009046|flower	none	side	4.2.3.104	R08373	31898
s21	35608	Artemisia annua	AAL79181	12409018	alpha-humulene	recombinant	none	none	side	4.2.3.104	R08373	31898
s22	4232	Helianthus annuus	AAY41422	19580670	alpha-muurolene	recombinant	none	none	main	none	none	none
s23	4530	Oryza sativa	ABJ16553	17524436	alpha-selinene	recombinant	none	none	side	4.2.3.86	R09886	30383
s24	3702	Arabidopsis thaliana	ABO09887	15918888	beta-acoradiene	adult	PO|0009046|flower	none	side	none	none	none
s25	3702	Arabidopsis thaliana	ABO09887	15918888	beta-barbetene	adult	PO|0009046|flower	none	side	none	none	none
s26	3702	Arabidopsis thaliana	ABO09887	15918888	beta-bisabolene	adult	PO|0009046|flower	none	side	4.2.3.55	R09623	28266
s27	4530	Oryza sativa	ABJ16553	17524436	beta-bisabolene	seedling	PO|0025034|leaf	PECO|0007407|methyl jasmonate exposure	side	4.2.3.55	R09623	28266
s28	3880	Medicago truncatula	AAV36464	17924138	beta-caryophyllene	adult	PO|0025034|leaf	PECO|0007115|Spodoptera spp. Exposure	main	4.2.3.57	R08541	28297
s29	3880	Medicago truncatula	AAV36464	17924138	beta-caryophyllene	adult	PO|0025034|leaf	PECO|0007407|methyl jasmonate exposure	main	4.2.3.57	R08541	28297
s30	3880	Medicago truncatula	AAV36464	17924138	beta-caryophyllene	adult	PO|0025034|leaf	PECO|0007407|methyl jasmonate exposure	main	4.2.3.57	R08541	28297
s31	3880	Medicago truncatula	AAV36464	17924138	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s32	3659	Cucumis sativus	AAU05952	15310834	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s33	88036	Selaginella moellendorffii	AFR34007	22908266	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s34	542674	Phyla dulcis	AFR23370	22867794	beta-caryophyllene	young	PO|0025034|leaf	none	main	4.2.3.57	R08541	28297
s35	127986	Matricaria chamomilla var. Recutita	AFM43734	22682202	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s36	4530	Oryza sativa	ACF05531	18433439	beta-caryophyllene	adult	PO|0025034|leaf	PECO|0007115|Spodoptera spp. Exposure	side	4.2.3.57	R08541	28297
s37	4580	Zea perennis	ABY79214	18296628	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s38	76912	Zea mays subsp. parviglumis	ABY79213	18296628	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s39	4579	Zea mays subsp. mexicana	ABY79212	18296628	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s40	15945	Zea luxurians	ABY79211	18296628	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s41	112001	Zea mays subsp. huehuetenangensis	ABY79210	18296628	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s42	4576	Zea diploperennis	ABY79209	18296628	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s43	4530	Oryza sativa	ABJ16553	17524436	beta-caryophyllene	seedling	PO|0025034|leaf	PECO|0007407|methyl jasmonate exposure	main	4.2.3.57	R08541	28297
s44	311405	Zingiber zerumbet	BAG12020	18273640	beta-caryophyllene	recombinant	none	none	side	4.2.3.57	R08541	28297
s45	4232	Helianthus annuus	AAY41422	19580670	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s46	3702	Arabidopsis thaliana	AAO85539	15918888	beta-caryophyllene	adult	PO|0009046|flower	none	main	4.2.3.57	R08541	28297
s47	35608	Artemisia annua	AAL79181	12409018	beta-caryophyllene	recombinant	none	none	main	4.2.3.57	R08541	28297
s48	3702	Arabidopsis thaliana	ABO09887	15918888	beta-chamigrene	adult	PO|0009046|flower	none	side	none	none	none
s49	4530	Oryza sativa	ABJ16553	17524436	beta-elemene	seedling	PO|0025034|leaf	PECO|0007407|methyl jasmonate exposure	side	none	none	none
s50	4530	Oryza sativa	ABJ16553	17524436	beta-elemene	recombinant	none	none	side	none	none	none
s51	3702	Arabidopsis thaliana	ABO09887	15918888	beta-farnesene	adult	PO|0009046|flower	none	side	none	none	none
s52	4530	Oryza sativa	ABJ16553	17524436	beta-farnesene	seedling	PO|0025034|leaf	PECO|0007407|methyl jasmonate exposure	side	none	none	none
s53	4530	Oryza sativa	ACF05531	18433439	beta-sesquiphellandrene	adult	PO|0025034|leaf	PECO|0007115|Spodoptera spp. Exposure	main	4.2.3.123	none	32699
s54	3702	Arabidopsis thaliana	ABO09887	15918888	beta-sesquiphellandrene	adult	PO|0009046|flower	none	side	4.2.3.123	none	32699
s55	4530	Oryza sativa	ABJ16553	17524436	beta-sesquiphellandrene	seedling	PO|0025034|leaf	PECO|0007407|methyl jasmonate exposure	side	4.2.3.123	none	32699
s56	542674	Phyla dulcis	AFR23369	22867794	bicyclogermacrene	young	PO|0025034|leaf	none	main	4.2.3.100	none	31999
s57	3702	Arabidopsis thaliana	ABO09887	15918888	cuparene	adult	PO|0009046|flower	none	side	none	none	none
s58	542674	Phyla dulcis	AFR23368	22867794	delta-cadinene	young	PO|0025034|leaf	none	main	none	none	none
s59	4232	Helianthus annuus	AAY41422	19580670	delta-cadinene	recombinant	none	none	main	none	none	none
s60	3702	Arabidopsis thaliana	ABO09887	15918888	delta-cuprenene	adult	PO|0009046|flower	none	side	none	none	none
s61	4580	Zea perennis	ABY79214	18296628	delta-elemene	recombinant	none	none	side	none	none	none
s62	76912	Zea mays subsp. parviglumis	ABY79213	18296628	delta-elemene	recombinant	none	none	side	none	none	none
s63	4579	Zea mays subsp. mexicana	ABY79212	18296628	delta-elemene	recombinant	none	none	side	none	none	none
s64	15945	Zea luxurians	ABY79211	18296628	delta-elemene	recombinant	none	none	side	none	none	none
s65	112001	Zea mays subsp. huehuetenangensis	ABY79210	18296628	delta-elemene	recombinant	none	none	side	none	none	none
s66	4576	Zea diploperennis	ABY79209	18296628	delta-elemene	recombinant	none	none	side	none	none	none
s67	4530	Oryza sativa	ABJ16553	17524436	Eudesma-4(14),11-dine	recombinant	none	none	side	none	none	none
s68	4530	Oryza sativa	ACF05531	18433439	gamma-bisabolene	adult	PO|0025034|leaf	PECO|0007115|Spodoptera spp. Exposure	main	none	none	none
s69	3702	Arabidopsis thaliana	ABO09887	15918888	isobazzanene	adult	PO|0009046|flower	none	side	none	none	none
s70	3702	Arabidopsis thaliana	ABO09887	15918888	thujopsene	adult	PO|0009046|flower	none	main	none	none	none
s71	4530	Oryza sativa	ACF05531	18433439	zingiberene	adult	PO|0025034|leaf	PECO|0007115|Spodoptera spp. Exposure	main	4.2.3.65	none	28643
s72	3702	Arabidopsis thaliana	ABO09887	15918888	zingiberene	adult	PO|0009046|flower	none	side	4.2.3.65	none	28643
s73	4530	Oryza sativa	ABJ16553	17524436	zingiberene	seedling	PO|0025034|leaf	PECO|0007407|methyl jasmonate exposure	side	4.2.3.65	none	28643

Apêndice IV

Publications

A terpenoid metabolic network modelled as graph database

DOI: 10.1504/IJDMB.2017.10007186

2Path: A terpenoid metabolic network modeled as graph database

DOI: 10.1109/BIBM.2016.7822709

GRAPHED: A Graph Description Diagram for Graph Databases

DOI: 10.1007/978-3-319-77703-0_111

Human-Computer Interaction Communicability Evaluation Method Applied to Bioinformatics

DOI: 10.1007/978-3-319-77712-2_95