Article

Exploring plant sesquiterpene diversity by generating chemical networks

Waldeyr M. C. da Silva [1,5,3]*, Jakob L. Andersen [4], Maristela Holanda [2], Maria Emília M. T. Walter [3], Marcelo M. Brigido [5], Peter F. Stadler [2,6–9], and Christoph Flamm [7]

1 Federal Institute of Goiás, Rua 64, esq. c/ Rua 11, s/n, Expansão Parque Lago. CEP: 73813-816. Formosa, GO, Brazil; waldeyr.mendes@ifg.edu.br
2 Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade de Brasília, Brasília-DF, Brazil; mholanda@cic.unb.br,mariaemilia@unb.br
3 Bioinformatics Group, Department of Computer Science; Interdisciplinary Center for Bioinformatics; University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany; studla@bioinf.uni-leipzig.de
4 Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, DK-5230 Odense, Denmark; jlandersen@imada.sdu.dk
5 Departamento de Biologia Celular, Universidade de Brasília, 70910-900. Brasília-DF, Brazil; brigido@unb.br;
6 German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig; Competence Center for Scalable Data Services and Solutions Dresden-Leipzig; and Leipzig Research Center for Civilization Diseases, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
7 Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria; xtof@tbi.univie.ac.at
8 Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
9 Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501
* Correspondence: waldeyr.mendes@ifg.edu.br; Tel.: +55-61-99671-6025

Abstract: Plants produce a diverse portfolio of sesquiterpenes that are important in their response to herbivores and the interaction with other plants. Their biosynthesis from farnesyl diphosphate depends on the sesquiterpene synthases. Here, we investigate to what extent metabolic pathways can be reconstructed just from knowledge of the final product and the reaction mechanisms catalyzed by sesquiterpene synthases. We use the software package `MedØlDatschgerl` (`MØD`) to generate chemical networks and elucidate pathways contained in them. As examples, we successfully consider the reachability of the important plant sesquiterpenes β-caryophyllene, α-humulene, and β-farnesene. We also introduce a graph database to integrate simulation results with experimental biological evidence for selected predicted sesquiterpenes biosynthesis.

Keywords: plant; sesquiterpenes; biosynthesis; graph grammars; database;

1. Introduction

Terpenes form a large and diverse class of natural products appearing particularly in the essential oils of many plants. They have commercial uses in medicine and as fragrances in perfumery. Synthetic derivatives of natural terpenes are also used as aromas and food additives [1]. Ecologically they perform key functions both in direct plant defense and in indirect mechanisms involving herbivores and their natural enemies [2].

Terpenes are produced throughout the tree of life from the C_5 compounds isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP). Leopold Ružička [3] formulated the "biogenetic isoprene rule" according to which terpenes are the result of concatenating isoprene units in a "head-to-tail" fashion to form chains, from which then rings are formed. Prenyltransferases condense IPP and DMAPP to geranyl pyrophosphate (GPP), farnesyl pyrophosphate (FPP), and geranylgeranyl pyrophosphate (GGPP) as the entry points to the world of monoterpenes (C_{10}), sesquiterpenes (C_{15}), and diterpenes (C_{20}) [4]. FPP and GGPP can then be condensed further to form squalene (C_{30}) and even larger molecules [5].

 doi:10.20944/preprints201903.0015.v1

These intermediates are substrates for a very large class of enzymes known collectively as terpene synthases (TPSs) to produce a wide variety of compounds. There are two major classes of TPS, which are distinguished by essential amino acid motifs [6,7]. Class I TPS, which are of interest in this contribution, convert linear, all-trans, isoprenoids, geranyl (C_{10})-, farnesyl (C_{15})-, or geranylgeranyl (C_{20})-diphosphate into numerous monoterpenes, sesquiterpenes, and diterpenes. They bind their substrate by coordination of a divalent metal ion catalytic site 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 [8–10].

The mechanisms of sesquiterpenes synthesis involve the formation of C–C bonds, cationic intermediates, Wagner-Meerwein rearrangements, carbocation capture by water and hydride, as well as methyl- and allyl-shifts caused by conformational changes of intermediate cations [11–13]. A combinatorial reaction cascade inside the TPS combining different cyclization/rearrangement reactions yields a highly diverse set of sesquiterpenes from just a handful of simple acyclic precursors molecules. Nevertheless, there are only four dominating types of cyclization reactions initiating the complex reaction cascade: $C_1 - C_{10}$, $C_1 - C_{11}$, $C_1 - C_6$, and $C_1 - C_7$ [14]. Even though the electrophilic chemical reaction mechanisms [11] primarily determine the diversity of the products, many other factors influence multiproduct sesquiterpene enzymes and their cyclization cascades, such as pH, the metal cofactor [4], evolutionary forces for the functional divergence [15], and plant-plant interactions [16].

The complexity of the portfolio of sesquiterpenes and their synthesis pathways calls for computational models that explain the observed diversity and can be used to predict biosynthetic pathways for particular terpene compounds. Within the field of systems biology, the construction and exploration of metabolic networks is grounded in determining metabolic capabilities from the enzymes encoded in the genome and extensive knowledge on enzymes themselves, collected, e.g. in the `BRENDA` database [17]. Several software tools have been developed to predict pathways from annotated genome sequences and to analyze and explore them. `Pathway Tools` [18], for instance, is a system designed for the `BioCyc` database collection [19]. Complementarily, versatile tools have become available to predict the metabolism of xenobiotics, in particular aiming at drugs. Starting from a known parent molecule, pattern recognition and/or machine learning techniques and knowledge-based rules are used to identify a "site of metabolism" a set of chemical products from biotransformation. Many of these tools are restricted to quite specific metabolic processes, although recently much more generally applicable tools such as `Biotransformer` [20] have become available. `RetroRules` [21] is a database of >400,000 highly specific reaction rules intended for metabolic engineering and in particular the prediction of novel and alternative products from *de novo* reactions of promiscuous enzymes. Possible reactions can in principle also be obtained with methods such as AFIR [22] from Potential Energy Surfaces, however, the computational cost makes it difficult to use them for exploring large chemical spaces.

Here we take a somewhat different approach in that we do not start from an extensive base of detailed chemical and biological knowledge but from a very generic representation of the *reaction mechanisms* underlying sesquiterpene synthesis. Molecules are represented as labeled graphs, abstracting away details of the spatial embedding of the molecules. Reactions are considered at the level of graph transformations and only use local context information to determine whether they are applicable to a given molecule or not. A key advantage of this approach, which is described in more detail in the Methods section, is its computational efficiency. The simple representation of molecules and reactions as graphs and graph transformation enables the efficient expansion of large networks of logically possible reactions. Integer flows in this networks correspond to feasible synthesis pathways, which can be identified efficiently as solutions of Integer Linear Programs (ILP). The software package `MØD` [23,24] makes it possible to prescribe potential synthesis pathways leading up to a user-defined molecule using a prescribed set of reaction mechanism (here the types of rearrangements catalyzed by TPSs) and a prescribed set of starting materials available in the already mapped part of the terpene metabolism. We have shown that such an explorative approach is indeed feasible and yields reasonable

 doi:10.20944/preprints201903.0015.v1

results without the need first to mine an extensive knowledge base. As examples, we consider the biosynthesis common plant sesquiterpenes as β-caryophyllene, α-humulene and β-farnesene.

2. Results

Here we present the results of three simulations and discuss the potential to expand them, how selected portions of specific pathways fit into conditions previously described in the literature, acknowledging their advantages and limitations.

2.1. Protonation-Dependent Diphosphate Cleavage

Natural class I terpene synthetases precisely pre-orient their acyclic diphosphate substrates in their active-site pockets in a reaction ready conformation. The complex reaction cascade is than initiated by cleaving off the diphosphate group from the pre-oriented substrate molecule, a pH-dependent [4] process that requires divalent metal ions as co-factor [25,26]. It has been shown, that this ionization step of the allylic diphosphate is the rate-limiting step in class I terpene synthetase catalysis [27]. The resulting highly reactive carbocation is than guided by the enzyme via rearrangements of the carbon skeleton towards diverse poly-cyclic product molecules. The highly reactive carbocation intermediates can scavenge nucleophiles, present in the enzyme's active-site, or saturate the positive charge by deprotonation, terminating the rearrangement cascade in an early stage. An example where the rearrangement cascade is minimal is the isomerization of FPP to Nerolidyl Diphosphate (NPP) [28], where the diphosphate anion terminates the reaction cascade by reattaching to the the carbocationic intermediate immediately after an allylic rearrangement has occurred (see Fig. 1).

In other words, NPP is reachable from FPP via a defined sequence of reactions from a predefined reaction set. It has been shown by reduction to the *word problem* for (semi)groups [29], that this type of reachability questions for chemical transformation systems is Turing undecidable. Therefore, the best we can do is to try out if in a given reaction network a path between two molecules of interest can be found. This test can easily be performed by combining a systematic reaction network generation, using the graph rewrite framework `MØD`, with an ILP-based pathway search within the generated reaction network. Furthermore, the reachability question for molecules of interest can be investigated under varying conditions by modulating the sets of reactions and / or additional molecules, e.g. nucleophiles, present during reaction networks expansion and pathway search. Terpene cyclization involves highly reactive cationic intermediates vulnerable to nucleophilic attack or elimination reactions. This characteristics allows us to tie the changing mixture of reachable sesquiterpene products to variations in external constraints which can in many cases be mapped to changes of environmental conditions such as seasonal changes, soil composition or periods of drought. A particular set of external conditions will in the following be called a scenario.

Using only the cleavage of the diphosphate group and the addition of a nucleophile to a cation in the reaction set, the reachability of NPP from FPP is trivial to ascertain in a constraint independent simulation, Fig. 1.

2.2. Synthesis of β-caryophyllene, α-humulene, β-farnesene, and their side products

(E)-β-caryophyllene is produced from FPP, and is emitted by different plant tissues, often in response to herbivore attack [30–32]. There is ample evidence that TPS catalysis produces a mixture of sesquiterpenes rather than a single sesquiterpene product [2,33,34].

Figure 2 shows that β-caryophyllene, α-humulene and β-farnesene are closely related in the reaction network. The same mechanism potentially produces several additional compounds. While the $P0,0$ is an intermediate compound, the predicted side compounds $P0,1$ and $P0,2$ could not be identified using chemical public databases such as PubChem [35] and ChemSpider [36].

 doi:10.20944/preprints201903.0015.v1

2.3. Large-scale exploration of terpene space

Simulation 03 is an explorative way to get the diversity of the feasible compounds. Starting with an FPP and a water molecule, during seven iterations, all the set of rules were iteratively applied.

2.4. Database Storage

Detailed mechanistic simulations of complex metabolic pathways typically involve extensive curation and annotation of the results. Thus they are a potentially valuable resource, which can be leveraged more efficiently if the data are provided with a comprehensive and consistent data schema can support the FAIR Guiding Principles for scientific data management [37] and facilitate the exchange and interoperability. Graph databases are a suitable tool for this purpose that has been demonstrated to be an efficient and convenient way to store and explore metabolic networks [38,39]. Here we use a graph database to enriches simulation data with experimental evidence related to the predicted sesquiterpenes.

Graph databases can both store data into nodes and relationships. Using our proposed graph database (`2Path-Sesquiterpenes`), the generated metabolic network was stored in a graph database, whose schema is a graph that minimizes the transition from the generated hypergraph. The labeled nodes represent the predicted compounds, the reactions transforming them, and the biological evidence related to these compounds. The labeled relationships represent the relations between these nodes. Once stored, the generated chemical mechanisms level metabolic network can be traversed and handled through graph databases query languages as Cypher[1] or Gremlin[2], and visualized using tools such as Cytoscape [40].

The `2Path-Sesquiterpenes` database nodes labeled as *Scenarios* provide a kernel of manually curated biological experimental evidence about constraints under which the compound are produced. In the `2Path-Sesquiterpene`, a relationship labeled as *OCCURS* is created between a node *Scenarios* and a node *Compound* when there is a biological scenario supporting the biosynthesis of this predicted compound. The *Scenarios* provide NCBI accession number for the enzymes, PUBMED accession number for the associated publication with the experimental results, experimental conditions, plant tissue[3], compound yield, EC numbers for the reactions, and cross-references to KEGG [41], Rhea [42] and ExploEnz (IUBMB) [43] in a taxonomic range of species. Figure 4 shows an example of query result for a stored simulation into Neo4J[4] graph database.

3. Discussion

Metabolic networks have been abstracted by various data structures, including substrate graphs, bipartite graphs, directed hypergraphs, reaction graphs, stoichiometric matrix and petri-net [44], [45], [46]. Direced hypergraphs can overcome conceptual limitations of graph modeling of biological processes such as multilateral relationships, which are not compatible with graph edges [44]. These hypergraphs properties allow for multilateral relationships between the nodes resulting in a suitable description of biological processes [44]. For example, in a metabolic reaction such as ($Compound_1 + Compound_2 \rightarrow Compound_3 + Compound_4$), a hypergraph allows edges to connect more than two nodes.

The rule-based simulations using graph transformation with MedØlDatschgerl can reach various compounds. These compounds are abstracted as undirected graphs, while the chemical mechanisms are abstracted as hyperedges in a directed hypergraph. Depending on the rule combination, it is possible to provide context-free results regarding sesquiterpenes biosynthesis reactions. By exploring

---

1 https://neo4j.com/developer/cypher-query-language/
2 https://tinkerpop.apache.org/gremlin.html
3 EMBL-EBI Plant Ontology – https://www.ebi.ac.uk/ols/ontologies/po
4 Neo4J graph database - https://neo4j.com

 doi:10.20944/preprints201903.0015.v1

the possible chemical mechanisms of these reactions, it is reasonable to establish connections with experimental results to draw from this universe of possibilities, those that can occur naturally or synthetically under certain circumstances. Constraint-based models (CBM) have enormous potential to enhance the understanding of reconstructed/predicted metabolic networks and predictive computational models by integrating biological evidence [47], [48], [49]. Studies combining genome, transcriptome, and metabolome data can address important questions, as biosynthetic pathways, selection of plants of interest, the environment influence in the gene expression and the metabolome profile, and many further questions.

`2Path-Sesquiterpenes` focusses plant sesquiterpene biosynthesis and aggregating putative biological meaning to the generated network, but there are some related works as AFIR [22]/GRRM, RetroRules [21], and Biotransformer [20]. Isegawa *et al* [50] had made simulations predicting pathways for terpene formation from a humulyl cation and others intermediate molecules using AFIR [22]/GRRM whose chemical results are aligned to ours, which allowed for a cross-confirmation. AFIR [22]/GRRM approach is fundamentally distinct from `2Path-Sesquiterpenes` because it uses both a different data structure to design molecules and applies artificial forces between two or more reacting molecules. Both Biotransformer and RetroRules [21] uses chemical reaction descriptions and rules encoded by SMARTS [51] and SMIRKS [52] and they are at the same time next to this work concerning objectives, but quite distinct regarding method. These three related works can deal with compounds chemical bonds geometry, while using `2Path-Sesquiterpenes`, the geometry of chemical bonds are indistinguishable due to the data structure (undirected graphs) that abstracts the compounds molecules. The Table 1 shows a summary of some features of the `2Path-Sesquiterpenes` and its related works.

Table 1. Summary of some features of the `2Path-Sesquiterpenes` and its related works.

	Molecules	**Reactions**	**Focus**	**Storage for results**	**Biological Evidence**
`2Path-Sesquiterpenes`	Undirected graphs	Graph rewrite rules	Plant sesquiterpenes	Graph database	Scenarios
Isegawa *et al* [50]	Internal	AFIR/GRRM	Sesquiterpenes	Internal	-
RetroRules [21]	SMART	SMIRKS	General	-	RetroRules
BioTransformer [20]	SMART	SMIRKS	General	-	MetXBioDB

Another matter is to make the simulation results findable, accessible, inter-operable, and reusable (FAIR). For this purpose, `2Path-Sesquiterpenes` offers the option of storing the predictions in a graph database. Metabolic network databases have been constructed since 1989 [53] through distinct methods, and many of them have been made available over time mostly due both to advances in metabolic network reconstruction methods and the expansion of omic data. Despite their extensive range, such as KEGG [41], Metacyc [54], and Reactome [55], to name a few, most of them do not provide information at the level of chemical mechanisms and intermediate compounds of a reaction, except for MACiE [56], which still offering rare data on biosynthetic sesquiterpene reactions. Also, gaph databases can bring significant query performance improvements for selected problems including metabolic networks [55], confirming their suitability for this purpose.

4. Methods

4.1. Graph Transformation, Hypergraphs, and Integer Hyperflows

Labeled graphs are commonly used in the chemical literature to represent molecules. Vertex labels identify atom types, while edge labels are used to indicate bond types. Chemical reactions thus correspond to transformations of graphs with particular features:

(i) Reactions may change the number of molecules, hence both input (substrate) and output (product) graphs are not necessarily connected.
(ii) All atoms are preserved, i.e., a chemical reaction defines a bijection between the vertex sets of input and output graphs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2019 doi:10.20944/preprints201903.0015.v1

(iii) Electrons are preserved as well, implying restrictive conditions on the way how edges (bonds) can change, corresponding to chemical *reaction mechanisms*.

Graph grammars are formal systems describing rule-based graph transformation that generalize the much more commonly used term-rewriting systems [57]. We favor the so-called double pushout (DPO) formalism [58] as a model of chemistry because it guarantees the structural reversibility of reactions [59] and it conveniently exposes the representation of the chemical transition state as part of the rule. In DPO graph rewriting, a transformation rule is of the form

$$p = (L \xleftarrow{l} K \xrightarrow{r} R) \tag{1}$$

where L, R, and K are the left graph, right graph, and context graph, respectively. These three graphs are connected by graph morphisms $l\colon K \to L$ and $r\colon K \to R$ describing the embedding of the context, into the L and R. The application of the rule p to a graph G requires that L "matches" a part of G. The existence of another graph morphism captures this, the *matching morphism* $m\colon L \to G$. Together the rule p and the matching morphism m uniquely define the transformation $G \overset{p,m}{\Longrightarrow} H$ of the substrate G to the product H by requiring that all morphisms in the following commutative diagram exist:

$$\begin{array}{ccccc} L & \xleftarrow{\text{l}} & K & \xrightarrow{\text{r}} & R \\ {\scriptstyle \text{m}}\downarrow & & \downarrow & & \downarrow \\ G & \longleftarrow & D & \longrightarrow & H \end{array} \tag{2}$$

In the context of modeling chemistry, we consider only injective graph morphisms, and we require that the restrictions of r and l to the vertex sets are bijective, ensuring preservation of atoms.

Since each chemical reaction transforms a set of substrate molecules into a set of product molecules, chemical networks are directed hypergraphs, with molecules as vertices and concrete reactions as hyperedges. Iterated application of reaction rules to a set of starting molecules generates the network (directed hypergraph) of reachable molecules, i.e., the chemical space defined by the given starting molecules and reaction rules.

Chemical reactions preserve mass, atom types, and charges. Chemical reaction pathways therefore from flows in the reaction hypergraph that connect a set of input molecules with a set of output molecules [44,60]. As an immediate consequence, reachability questions in chemical reaction networks translate into the existence of integer flows [24], which is efficiently evaluated by means of integer linear programming (ILP).

4.2. Simulations

Simulations were performed using MedØlDatschgerl (`MØD`) [23], a software package that combines a DPO graph rewriting engine and a ILP solver to generate and analyze large-scale reaction networks. `MØD` provides a confortable Python interface (the Python 3 module `PyMØD` comprising bindings to the underlying library `libMØD`) as well as a system of generic exploration strategies [61] to guide and restrict the generation process. Graph transformation rules are specificied manually in GML format [62], substrate molecules can be provided either in GML or SMILES [63] format.

We have designed DPO graph transformation rules representing chemical mechanisms involved in the production of the plant sesquiterpenes β-caryophyllene, α-humulene and β-farnesene from its precursor FPP, exploring the generation of distinct compound through simulations with varying sets of rules. The presented simulations are available on GitHub. Figure 5 shows the rules employed in the Simulation 01, Figure 6 shows the rules employed in the Simulation 02, and Figure 7 shows the rules that, together with rules of the Figures 5 and 6, were employed in the Simulation 03.

 doi:10.20944/preprints201903.0015.v1

5. Conclusions

Rule-based generative transformation systems, such as the graph grammars used here provide a mathematically sound way to answer reachability questions in combinatorial, potentially infinite, search spaces. Here we have used DPO graph transformations, as implemented by MedØlDatschgerl [23] to model the specific combination of cyclizations reactions catalyzed by sesquiterpene synthases (STPS) using a small number of transformation rules. This is sufficient to explain the diversity of sesquiterpenes, including the most common plant sesquiterpenes:β-caryophyllene, α-humulene and β-farnesene, through combinations of specific cyclization reactions. The generative approach produces a local view of the metabolic or chemical network that is naturally represented as a hypergraph. Reachability then translates to the existence of pathways, which can be decided by integer linear programming.

The networks and pathways are exported to a PDF report and optionally can be stored and traversed in a graph database adhering to FAIR Guiding Principles [37]. This makes it possible to integrate simulation results with *Scenarios* contained in database, in particular experimental evidence on biosynthesis reactions.

Computational *de novo* pathway discovery is of particular interest for reactions catalyzed by multi-product enzymes as in the case of STPS, because the combinatorial complexity of products in multistep synthesis quickly exceeds the limits of manual analysis. It also enables a systematic analysis of synthetic and heterologous biology with potential applications e.g. in sustainable bioeconomy.

The work presented here is intended as proof of concept. In future work, it will be expanded in several directions. Additional graph grammar rules can be included to extend the space of reachable sesquiterpenes and/or to include other classes of terpenes. Functionalization to a broad array of terpenoids can also be modeled by generative approach discussed here. On the other hand, we plan to expand the collection of experimental scenarios and develop a user-friendly interface to facilitate the integration of experimental knowledge and computational predictions.

Author Contributions: Conceptualization, W.M.C.S. and C.F.; methodology C.F., J.L.A., M.H., M.E.M.T.W., and P.F.S.; software, C.F. and J.L.A.; simulations, W.M.C.S.; data curation, J.L.A. and C.F; interpretation of results, C.F., J.L.A., M.M.B., and P.F.S.; writing first draft W.M.C.S.; all authors contributed to writing and editing.

Funding: This research was funded in part by CAPES through a sandwich scholarship to W.M.C.S. It was additionally supported in part by the Independent Research Fund Denmark, Natural Sciences, grant DFF-7014-00041.

Acknowledgments: W.M.C.S. thanks to CAPES for the scholarship and gratefully acknowledges the hospitality of Leipzig and Vienna Universities.

Conflicts of Interest: "The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results".

Abbreviations

The following abbreviations are used in this manuscript:

AAM	Atom to Atom Mapping
AFIR	Artificial Force-Induced Reaction
CBM	Constraint-based models
DMAPP	Dimethylallyl Pyrophosphate
DPO	Double pushout graph rewriting
FAIR	Findability, Accessibility, Interoperability, and Reuse of digital assets
FBA	Flux Balanced Analysis
FPP	Farnesyl Diphosphate
GGPP	Geranylgeranyl Diphosphate
GML	Graph Modeling Language
GRRM	Global Reaction Route Mapping
GPP	Geranyl Diphosphate
ILP	Integer Linear Programming
IPP	Isopentenyl Pyrophosphate
IUBMB	International Union of Biochemistry and Molecular Biology
MEP	Methylerythritol phosphate pathway
MVA	Mevalonate pathway
NPP	Nerolidyl Diphosphate
OPP	Diphosphate
PDF	Portable Document Format
PGDB	Pathway/Genome database
STPS	Sesquiterpene Synthases
TPS	Terpene Synthase

References

1. Breitmaier, E. *Terpenes: Flavors, Fragrances, Pharmaca, Pheromones*; Wiley-VCH, 2006. doi:10.1002/9783527609949.
2. Cheng, A.X.; Xiang, C.Y.; Li, J.X.; Yang, C.Q.; Hu, W.L.; Wang, L.J.; Lou, Y.G.; Chen, X.Y. The rice (E)-β-caryophyllene synthase (OsTPS3) accounts for the major inducible volatile sesquiterpenes. *Phytochemistry* **2007**, *68*, 1632–1641.
3. Ružička, L. The isoprene rule and the Biogenesis of terpenic compounds. *Cell. Mol. Life Sci.* **1953**, *9*, 357–367. doi:10.1007/BF02167631.
4. Vattekkatte, A.; Garms, S.; Brandt, W.; Boland, W. Enhanced structural diversity in terpenoid biosynthesis: enzymes, substrates and cofactors. *Organic & Biomolecular Chemistry* **2018**, *16*, 348–362.
5. Wink, M. *Biochemistry of Plant Secondary Metabolism*; Vol. 40, John Willey & Sons Inc., 2010.
6. Chen, F.; others. 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* **2011**, *66*, 212–229.
7. Liu, W.; others. Structure, function and inhibition of ent-kaurene synthase from Bradyrhizobium japonicum. *Scientific reports* **2014**, *4*.
8. Lesburg, C.A. Crystal Structure of Pentalenene Synthase: Mechanistic Insights on Terpenoid Cyclization Reactions in Biology. *Science* **1997**, *277*, 1820–1824.
9. Oldfield, E.; Lin, F.Y. Terpene biosynthesis: Modularity rules. *Angewandte Chemie - International Edition* **2012**, *51*, 1124–1137.
10. Kempinski, C.; Jiang, Z.; Bell, S.; Chappell, J. Metabolic engineering of higher plants and algae for isoprenoid production. *Advances in Biochemical Engineering/Biotechnology* **2015**.
11. Degenhardt, J.; Köllner, T.G.; Gershenzon, J. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry* **2009**, *70*, 1621–1637.
12. Schifrin, A.; others. A single terpene synthase is responsible for a wide variety of sesquiterpenes in Sorangium cellulosum Soce56. *Organic & biomolecular chemistry* **2016**, *14*, 3385–3393.
13. Tholl, D. Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Current opinion in plant biology* **2006**, *9*, 297–304.
14. Christianson, D.W. Structural and Chemical Biology of Terpenoid Cyclases. *Chemical Reviews* **2017**, *117*, 11570–11648.
15. Chen, H.; others. Positive Darwinian selection is a driving force for the diversification of terpenoid biosynthesis in the genus Oryza. *BMC plant biology* **2014**, *14*, 239.
16. Kigathi, R.N.; Weisser, W.W.; Reichelt, M.; Gershenzon, J.; Unsicker, S.B. Plant volatile emission depends on the species composition of the neighboring plant community. *BMC Plant Biology* **2019**, *19*, 58. doi:10.1186/s12870-018-1541-9.

17. Jeske, L.; Placzek, S.; Schomburg, I.; Chang, A.; Schomburg, D. BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.*, *47*, D542–D549. doi:10.1093/nar/gky1048.
18. Karp, P.D.; Latendresse, M.; Caspi, R. The pathway tools pathway prediction algorithm. *Standards in genomic sciences* **2011**, *5*, 424.
19. Karp, P.D.; Billington, R.; Caspi, R.; Fulcher, C.A.; Latendresse, M.; Kothari, A.; Keseler, I.M.; Krummenacker, M.; Midford, P.E.; Ong, Q.; Ong, W.K.; Paley, S.M.; Subhraveti, P. The BioCyc collection of microbial genomes and metabolic pathways. *Briefings Bioinf.* **2017**. doi:10.1093/bib/bbx085.
20. Djoumbou-Feunang, Y.; Fiamoncini, J.; Gil-de-la Fuente, A.; Greiner, R.; Manach, C.; Wishart, D.S. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *Journal of Cheminformatics* **2019**, *11*, 2.
21. Duigou, T.; du Lac, M.; Carbonell, P.; Faulon, J.L. RetroRules: a database of reaction rules for engineering biology. *Nucleic acids research* **2018**, *47*, D1229–D1235.
22. Maeda, S.; others. Artificial Force Induced Reaction (AFIR) Method for Exploring Quantum Chemical Potential Energy Surfaces. *Chemical Record* **2016**, *16*, 2232–2248.
23. Andersen, J.L.; Flamm, C.; Merkle, D.; Stadler, P.F. A software package for chemically inspired graph transformation. International Conference on Graph Transformation. Springer, 2016, pp. 73–88.
24. Andersen, J.L.; Flamm, C.; Merkle, D.; Stadler, P.F. Chemical Transformation Motifs — Modelling Pathways as Integer Hyperflows. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, Vol. 5963, pp. 1–14.
25. Picaud, S.; Olsson, M.E.; Brodelius, M.; Brodelius, P.E. Cloning, expression, purification and characterization of recombinant (+)-germacrene D synthase from Zingiber officinale. *Archives of Biochemistry and Biophysics* **2006**, *452*, 17–28.
26. Farzadfar, S.; Zarinkamar, F.; Behmanesh, M.; Hojati, M. Magnesium and manganese interactively modulate parthenolide accumulation and the antioxidant defense system in the leaves of Tanacetum parthenium. *Journal of Plant Physiology* **2016**, *202*, 10–20.
27. Zhang, F.; others. Protonation-dependent diphosphate cleavage in FPP cyclases and synthases. *ACS Catalysis* **2016**, *6*, 6918–6929.
28. Cane, D.E.; Iyengar, R. The enzymic conversion of farnesyl to nerolidyl pyrophosphate: Role of the pyrophosphate moiety. *Journal of the American Chemical Society* **1979**, *101*, 3385–3388.
29. Smith, W.D. Computational complexity of synthetic chemistry — Basic facts. Technical Report availabe at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.9276`, 1997.
30. Kollner, T.G.; Held, M.; Lenk, C.; Hiltpold, I.; Turlings, T.C.; Gershenzon, J.; Degenhardt, J. A Maize (E)-beta-Caryophyllene Synthase Implicated in Indirect Defense Responses against Herbivores Is Not Expressed in Most American Maize Varieties. *the Plant Cell Online* **2008**, *20*, 482–494.
31. Tholl, D.; Chen, F.; Petri, J.; Gershenzon, J.; Pichersky, E. Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from Arabidopsis flowers. *Plant Journal* **2005**, *42*, 757–771.
32. Irmisch, S.; Krause, S.T.; Kunert, G.; Gershenzon, J.; Degenhardt, J.; Köllner, T.G. The organ-specific expression of terpene synthase genes contributes to the terpene hydrocarbon composition of chamomile essential oils. *BMC Plant Biology* **2012**, *12*.
33. Chen, F. Biosynthesis and Emission of Terpenoid Volatiles from Arabidopsis Flowers. *the Plant Cell Online* **2003**, *15*, 481–494.
34. Yu, F.; Okamto, S.; Nakasone, K.; Adachi, K.; Matsuda, S.; Harada, H.; Misawa, N.; Utsumi, R. Molecular cloning and functional characterization of α-humulene synthase, a possible key enzyme of zerumbone biosynthesis in shampoo ginger (Zingiber zerumbet Smith). *Planta* **2008**, *227*, 1291–1299.
35. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; others. PubChem 2019 update: improved access to chemical data. *Nucleic acids research* **2018**, *47*, D1102–D1109.
36. Pence, H.E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education* **2010**, *87*, 1123–1124.
37. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; others. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **2016**, *3*.

38. Brandizi, M.; Singh, A.; Rawlings, C.; Hassani-Pak, K. Towards FAIRer Biological Knowledge Networks Using a Hybrid Linked Data and Graph Database Approach. *Journal of integrative bioinformatics* **2018**, *15*.
39. da Silva, W.M.; Wercelens, P.; Walter, M.E.M.; Holanda, M.; Brígido, M. Graph Databases in Molecular Biology. Brazilian Symposium on Bioinformatics. Springer, 2018, pp. 50–57.
40. Smoot, M.E.; Ono, K.; Ruscheinski, J.; Wang, P.L.; Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **2010**, *27*, 431–432.
41. Kanehisa, M.; Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **2000**, *28*, 27–30.
42. Alcántara, R.; Axelsen, K.B.; Morgat, A.; Belda, E.; Coudert, E.; Bridge, A.; Cao, H.; De Matos, P.; Ennis, M.; Turner, S.; others. Rhea—a manually curated resource of biochemical reactions. *Nucleic acids research* **2011**, *40*, D754–D760.
43. McDonald, A.G.; Boyce, S.; Tipton, K.F. ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic acids research* **2008**, *37*, D593–D597.
44. Klamt, S.; Haus, U.U.; Theis, F. Hypergraphs and cellular networks. *PLoS Comput Biol* **2009**, *5*, e1000385. doi:10.1371/journal.pcbi.1000385.
45. Wang, L.; Dash, S.; Ng, C.Y.; Maranas, C.D. A review of computational tools for design and reconstruction of metabolic pathways. *Synthetic and Systems Biotechnology* **2017**, *2*, 243–252.
46. Cherdal, S.; Mouline, S. Modelling and Simulation of Biochemical Processes Using Petri Nets. *Processes* **2018**, *6*, 97.
47. Blazier, A.S.; Papin, J.A. Integration of expression data in genome-scale metabolic network reconstructions. *Frontiers in Physiology* **2012**, *3 AUG*, 299.
48. Øyås, O.; Stelling, J. Genome-scale metabolic networks in time and space. *Current Opinion in Systems Biology* **2018**, *8*, 51–58.
49. Fang, C.; Fernie, A.R.; Luo, J. Exploring the Diversity of Plant Metabolism. *Trends in Plant Science* **2018**, *24*, 83–98.
50. Isegawa, M.; Maeda, S.; Tantillo, D.J.; Morokuma, K. Predicting pathways for terpene formation from first principles–routes to known and new sesquiterpenes. *Chemical Science* **2014**, *5*, 1555–1560.
51. Systems, D.C.I. SMARTS – a language for describing molecular patterns. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, 2008. [Online; accessed 30-Jan-2019].
52. Systems, D.C.I. A reaction transform language. http://daylight.com/dayhtml/doc/theory/theory.smirks.html. [Online; accessed 30-Jan-2019].
53. Selkov, E.E.; others. Factographic data bank on enzymes and metabolic pathways. *Studia Biophysica* **1989**, *129*, 155–164.
54. Caspi, R.; others. The MetaCyc Database of metabolic pathways. *Nucleic Acids Research* **2014**, *42*, 471–480.
55. Fabregat, A.; others. Reactome graph database: Efficient access to complex pathway data. *PLoS Computational Biology* **2018**, *14*, 1–13.
56. Holliday, G.L.; others. MACiE: exploring the diversity of biochemical reactions. *Nucleic acids research* **2012**, *40*, D783–D789.
57. Ehrig, H.; Ehrig, K.; Prange, U.; Taenthzer, G. *Fundamentals of Algebraic Graph Transformation*; Springer-Verlag: Berlin, D, 2006. doi:10.1007/3-540-31188-2.
58. Löwe, M. Algebraic approach to single-pushout graph transformation. *Theoretical Computer Science* **1993**, *109*, 181 – 224.
59. Andersen, J.L.; Flamm, C.; Merkle, D.; Stadler, P.F. Inferring chemical reaction patterns using rule composition in graph grammars. *Journal of Systems Chemistry* **2013**, *4*, 4.
60. Zeigarnik, A.V. On Hypercycles and Hypercircuits in Hypergraphs. In *Discrete Mathematical Chemistry*; Hansen, P.; Fowler, P.W.; Zheng, M., Eds.; American Mathematical Society: Providence, RI, 2000; Vol. 51, *DIMACS series in discrete mathematics and theoretical computer science*, pp. 377–383.
61. Andersen, J.L.; Flamm, C.; Merkle, D.; Stadler, P.F. Generic Strategies for Chemical Space Exploration. *Int. J. Comp. Biol. Drug Design* **2014**, *7*, 225–258.
62. Himsolt, M. GML: A portable graph file format, 1997.
63. Minkiewicz, P.; Iwaniak, A.; Darewicz, M. Annotation of peptide structures using SMILES and other chemical codes-practical solutions, 2017.

 doi:10.20944/preprints201903.0015.v1

Sample Availability: A tutorial to setup the environment and the Python code for the simulations using MedØlDatschgerl framework are available in the GitHub.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2019 doi:10.20944/preprints201903.0015.v1

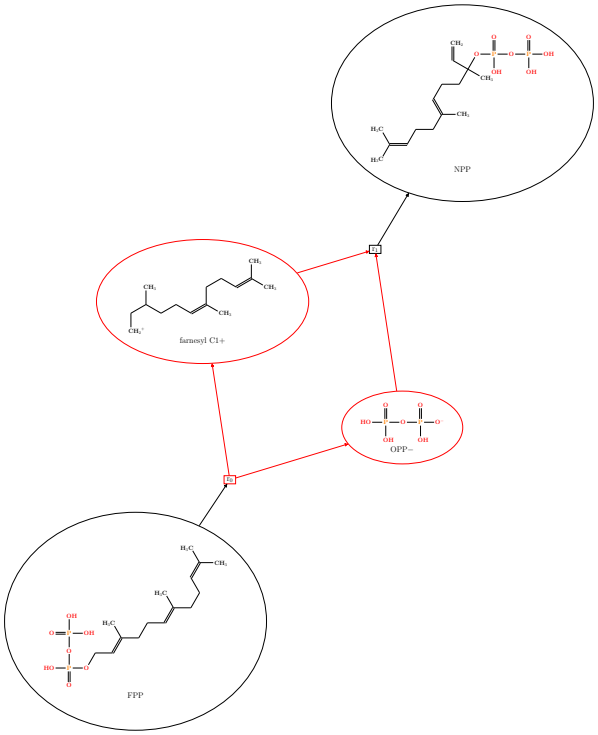


Figure 1. Plotted result of the Simulation 01 for OPP cleavage from FPP.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2019 doi:10.20944/preprints201903.0015.v1

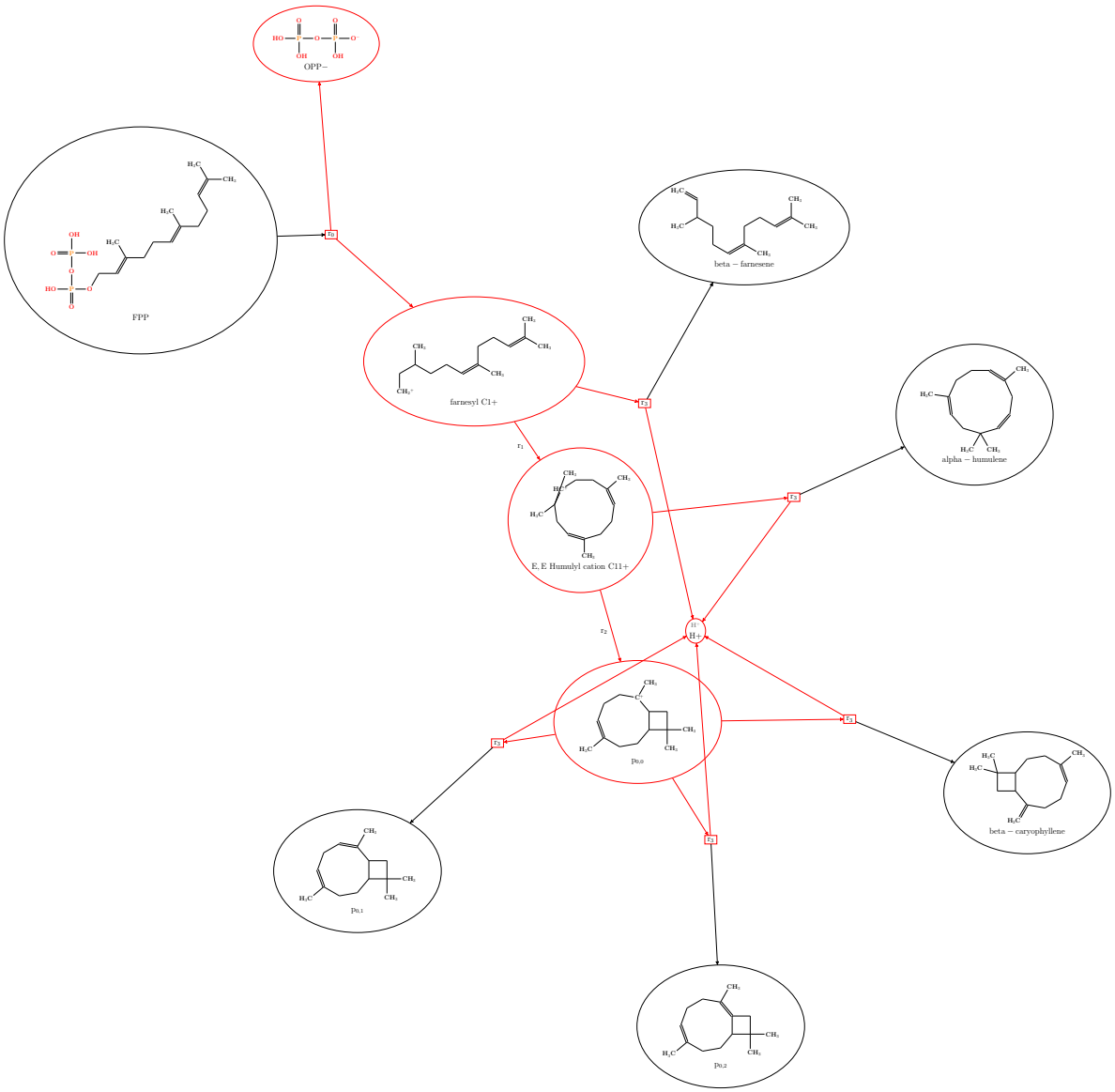


Figure 2. Plotted result of the Simulation 02 for β-caryophyllene, α-humulene, β-farnesene, and side predicted compounds from FPP.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2019 doi:10.20944/preprints201903.0015.v1

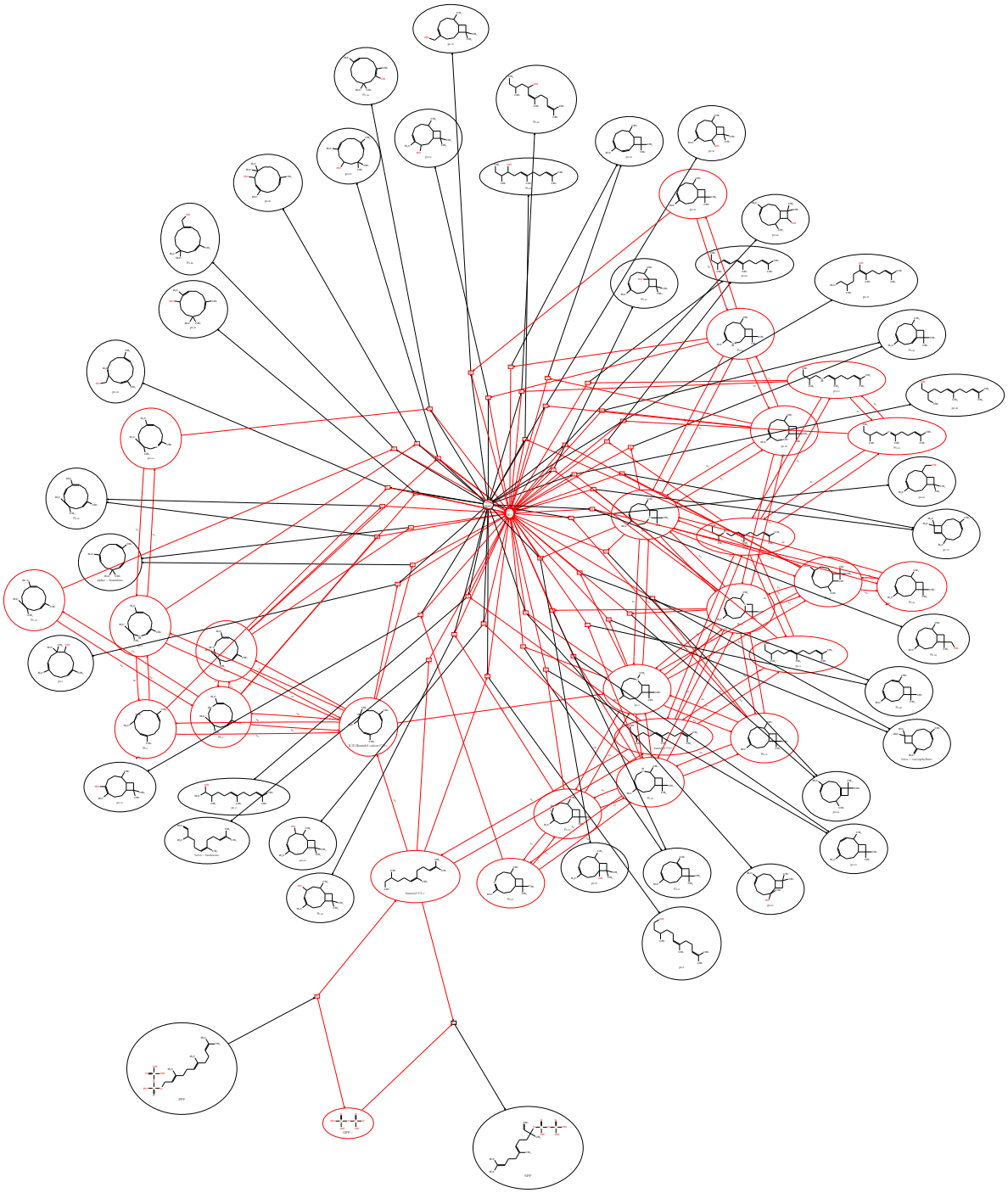


Figure 3. The exploratory strategy yield 55 predicted molecules. Stable, neutral molecules are outlined in black, while cations and anions are outlined in red.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2019 doi:10.20944/preprints201903.0015.v1

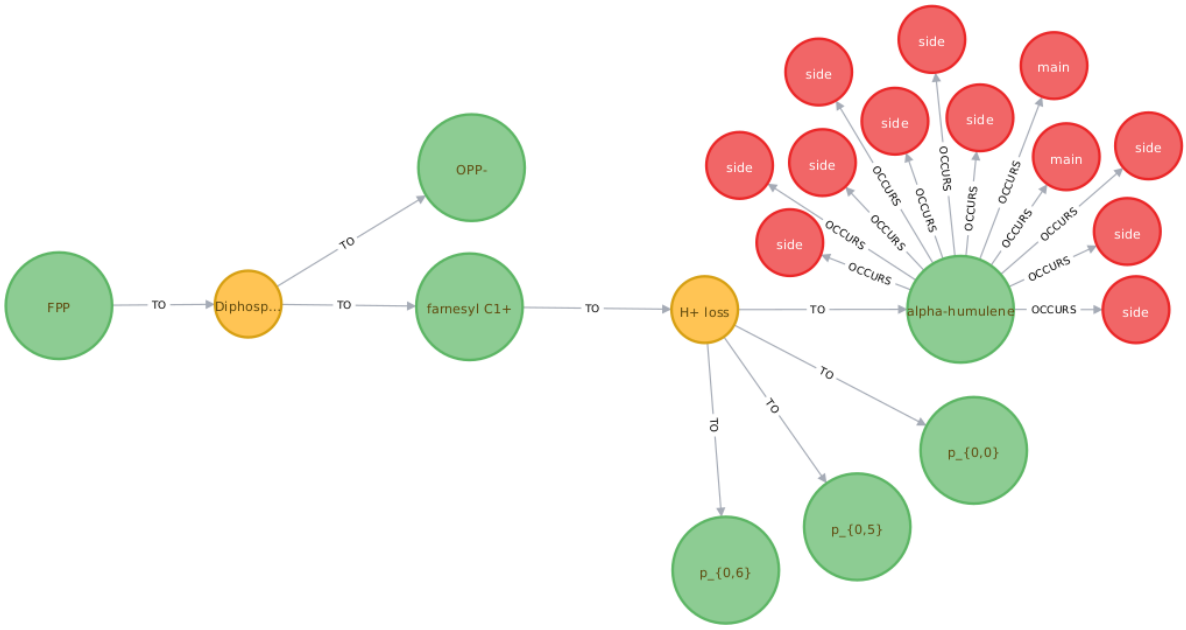
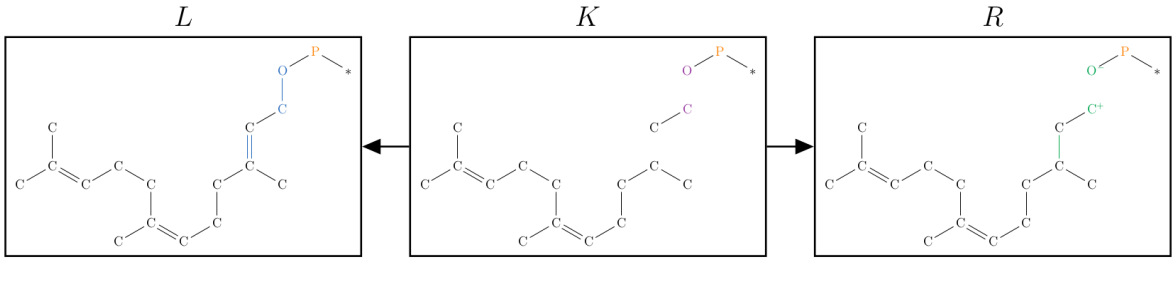
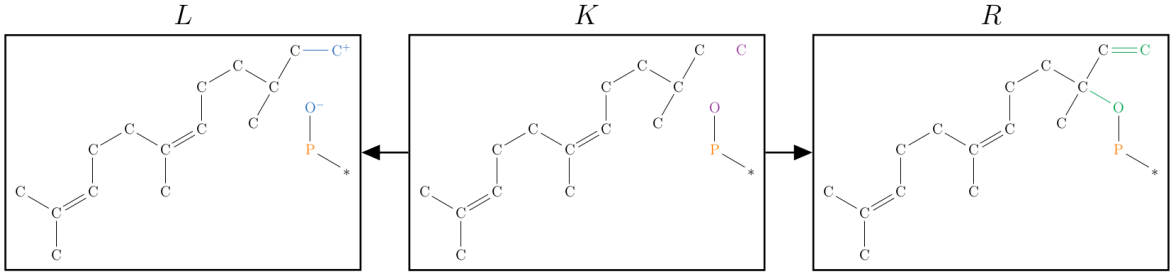


Figure 4. Example of a reaction stored in the Neo4J graph database. Yellow nodes (label Rules) denote the rules diphosphate loss and H^+ loss. Green nodes (label Compound) denote compounds from FPP to α-humulene and other generated compounds. Red nodes (label Scenario) denote the experimental scenarios with compound yield for the α-humulene. The details inside each Scenario node, as the experimental conditions, the plant tissue, EC numbers for the reactions, and cross-references, can be reached using the web interface, which is native for Neo4J graph database.



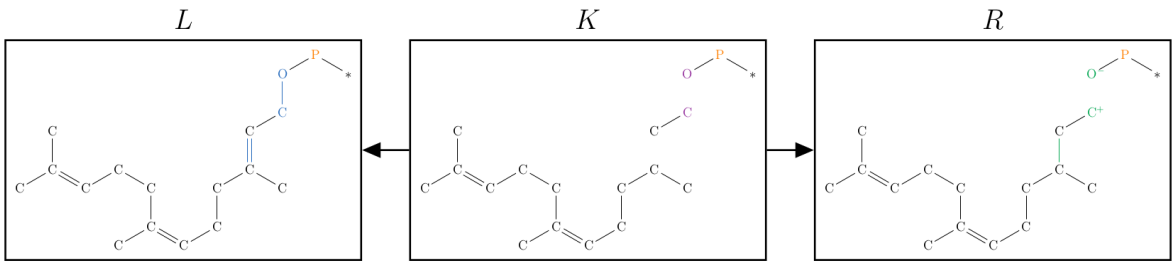
(a) OPP loss from FPP.



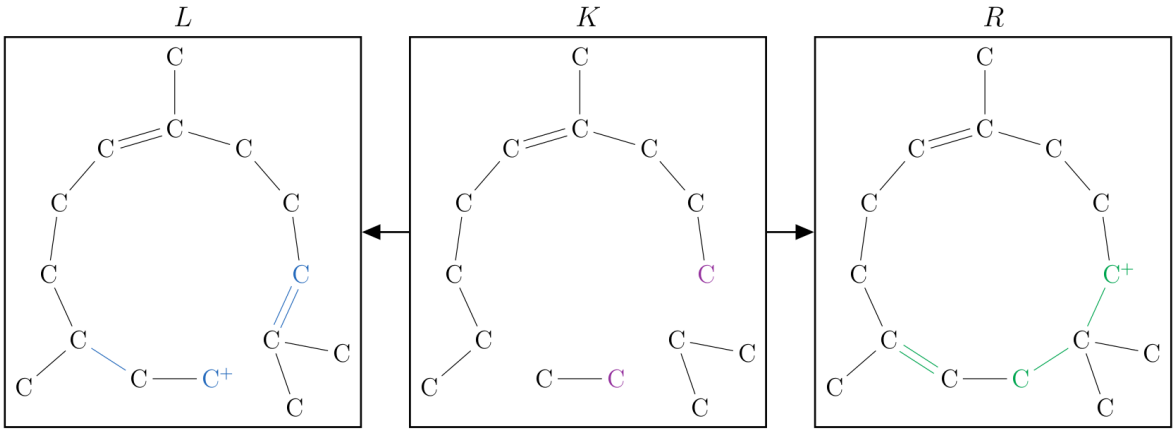
(b) OPP gain from FPP.

Figure 5. Graphical representation of the set of graph grammar rules used in the Simulation 01.

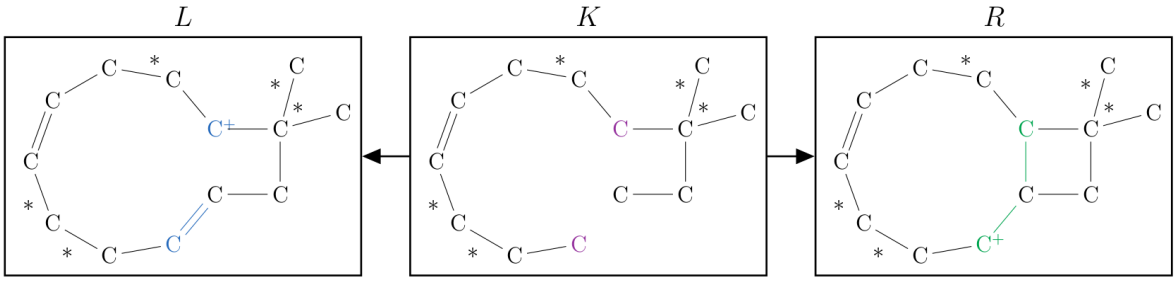
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2019 doi:10.20944/preprints201903.0015.v1



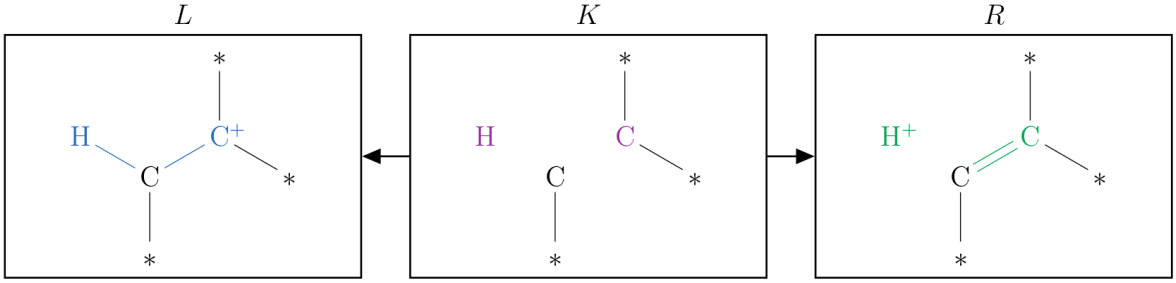
(a) OPP loss from FPP.



(b) C_1 to C_{11} cyclization.



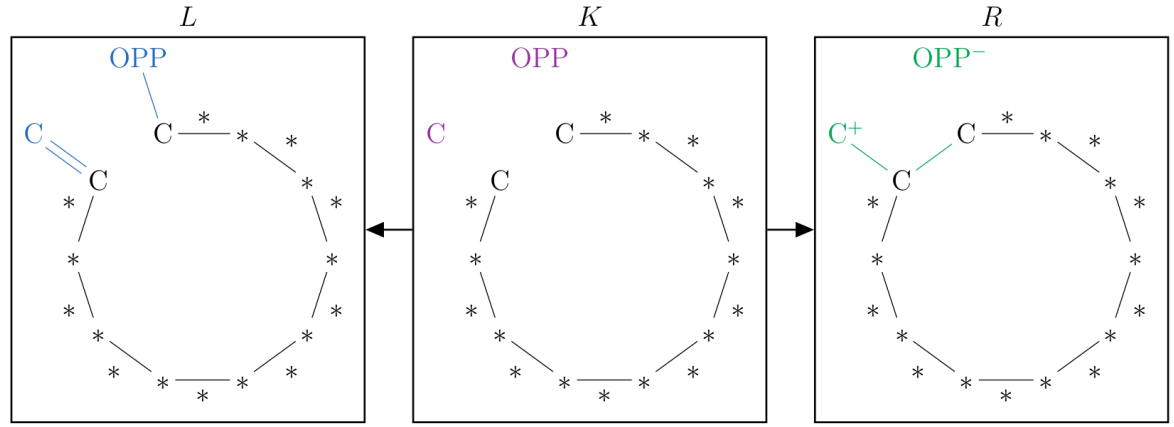
(c) C_2 to C_{10} cyclization.



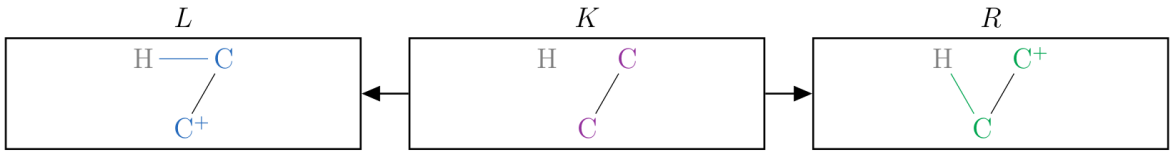
(d) H^+ loss.

Figure 6. Graphical representation of the set of graph grammar rules used in the Simulation 02.

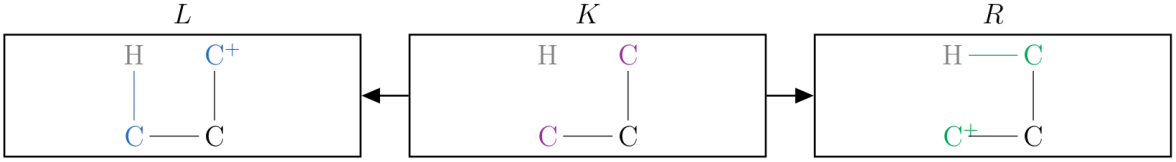
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2019 doi:10.20944/preprints201903.0015.v1



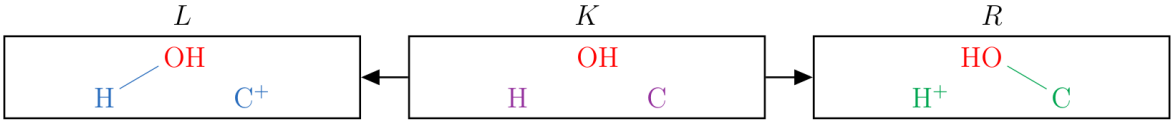
(a) C_1 to C_{10} cyclization.



(b) C_1 to C_2 hidrid shift.



(c) C_1 to C_3 hidrid shift.



(d) H_2O gain.

Figure 7. Graphical representation of the set of graph grammar rules that, together with Rules of the Figures 5 and 6, where used in the Simulation 03.