MASTERARBEIT | MASTER'S THESIS

Titel | Title

Gene Evolution and the history of plant defenses in conifers.

verfasst von | submitted by

Sebastian Ley BEd

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Education (MEd)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 199 502 520 02
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Lehramt Sek (AB) Unterrichtsfach Biologie und Umweltbildung Unterrichtsfach Mathematik
Betreut von | Supervisor:	Ass.-Prof. Dr. Leyla Jean Seyfullah Privatdoz. BSc MSc

Abstract – English

Terpenes, the largest and most diverse class of plant chemicals, play essential roles in plant defense. Terpenoids (=TPs) are derived from universal building blocks and achieve their diversity through specialized terpene synthases (=TPSs). By focusing on a group of terpenes, the labdenoids, which may fossilize to form amber, it is possible to compare extant species to their evolutionary past. Labdenoids are formed by Class I/II diterpene synthases.
To construct a phylogenetic tree of all conifer TPS sequences, a lot of programs were used. The programs used to acquire, sort, clean and align the TPS sequences, then construct a tree are in order: BLASTp, HMMER, Phyx, MAFFT, IQTREE. The resulting tree is then displayed using Figtree to form the main result of this work, a comprehensive tree detailing the evolutionary history of the terpene synthases. With this tree it is possible to infer a significant duplication and subsequent deletion event in the history of the conifers. The percentage of data available for each family leads to insights into the state of research. By collating phylogenetic data with anatomical data on the conifers, paleontological research is predicted and reconfirmed.
Lastly, the place of Pinaceae as a representative of the conifers is discussed.

Abstract – Deutsch

Terpene, die größte und diverseste Klasse von pflanzlichen Chemikalien, spielen essenzielle Rollen in deren Verteidigung. Terpenoide (=TPs) sind abgeleitet von universellen Bausteinen und erreichen ihre Diversität durch spezialisierte Terpensynthasen (=TPSs). Durch den Fokus auf eine Gruppe von Terpenen, die Labdane, welche fossilisieren, um Bernstein zu formen, ist es möglich heute lebende Spezies mit ihrer evolutionären Vergangenheit zu vergleichen. Labdane werden von Class I/II Diterpensynthasen gebildet.
Um einen phylogenetischen Baum aller Koniferen TPSs zu bilden, werden viele Programme benötigt. Die Programme, um die TPS Sequenzen zu gewinnen, sortieren, säubern, auszurichten und einen Baum zu konstruieren sind in dieser Reihenfolge: BLASTp, HMMER, Phyx, MAFFT, IQTREE. Der resultierende Baum wird dann mithilfe von Figtree dargestellt, um das Hauptresultat dieser Arbeit zu bilden, einen vollständigen Baum, der die evolutionäre Geschichte der TPSs darstellt.
Mit diesem Baum ist es möglich, ein signifikantes Duplikations- und darauffolgendes Eliminationsereignis in der Geschichte der Koniferen abzulesen. Der Prozentsatz von erhältlichen Daten gibt Einsicht in den Stand der Forschung. Durch Kollationieren der phylogenetischen Daten mit Anatomischen, lässt sich paläologische Forschung voraussagen und erneut bestätigen.
Letztlich wird der Platz der Pinaceae als Repräsentant der Koniferen diskutiert.

Table of contents

Introduction

Terpenes represent the largest and most diversified class of chemical compounds produced by plants (Alicandri et al., 2020; Tholl, 2015). The structural diversity and evolutionary success of at least 40 000 compounds largely depends on the flexibility of building molecules of various sizes. While all terpenoids are derived from the same building blocks, they have a variety of functions (Tholl, 2015). Humans use them as fragrances and flavors, as pharmaceutical agents and as insecticides (Tholl, 2006). To plants they have many uses, but the majority are used for specialized chemical interactions. These range from communication to plant defense (Alicandri et al., 2020; Tholl, 2006).
In this work I focus on the conifers, a subset of the gymnosperms. The conifers, also called Pinophyta, are defined as a group of cone-bearing seed plants. The conifers contain the following families: Araucariaceae, Podocarpaceae, Sciadopityaceae, Cupressaceae, Taxaceae and Pinaceae. In conifers, resin is produced from terpenoids as an answer to many kinds of external attacks. In doing so they can offer us a glimpse into the past, as the originally liquid resin can solidify through polymerization, then slowly turning into amber through maturation in sediments. This lets us compare fossil amber deposits to resin producing plants today.
In this work I plan to connect the dots between amber producers of the past and resin producers of today. By taking a close look at the genes responsible for resin production and then constructing a comprehensive phylogenetic tree, I hope to uncover part of the evolutionary history of the conifers.

As previously discussed, terpenoids are an incredibly diverse group of chemicals. As a first step, the focus needs to be narrowed. I will do this by defining what terpenoids are, how they are formed, and which ones are important for our purposes. Following this is a short introduction of the tools used to conduct this research.

Terpene Structures

Terpenoids consist of a number of basic five-carbon (C_5) isoprenoid units. They are usually categorized as hemiterpenoids (C_5), monoterpenoids (C_{10}), sesquiterpenoids (C_{15}), diterpenoids (C_{20}), triterpenoid (C_{30}), tetraterpenoid (C_{40}) or polyterpenoids (C_{5n}), based on the number of C_5units they contain. Many of these compounds are volatile and aromatic (Alicandri et al., 2020; Tholl, 2006; Tholl, 2015).

Terpenoids are biosynthesized by proteins called terpene synthases (=TPS). Their functional diversity is determined by their modular structure. Three common domains, denoted as γ, β, and α, are variably arranged (see Figure 1). Based on the presence of either one or two active sites, and their related catalytic motifs, TPSs are said to be monofunctional or bifunctional, respectively, and are categorized into

class-I, class-II, or class-I/II enzymes. The class-I TPSs host only one active site and include all the monoterpene-synthases (=MTPSs), all the Sesquiterpene-synthases (=STPSs) and part of the diterpene-synthases (=DTPSs). The class-II TPSs, also host just one active site and include only DTPSs. Finally, a limited number of three-domain (γβα) TPSs contain all the three functional active sites. These bifunctional TPSs, all of which are DTPSs, are called class-I/II enzymes (Alicandri et al., 2020; Zhou & Pichersky, 2020).

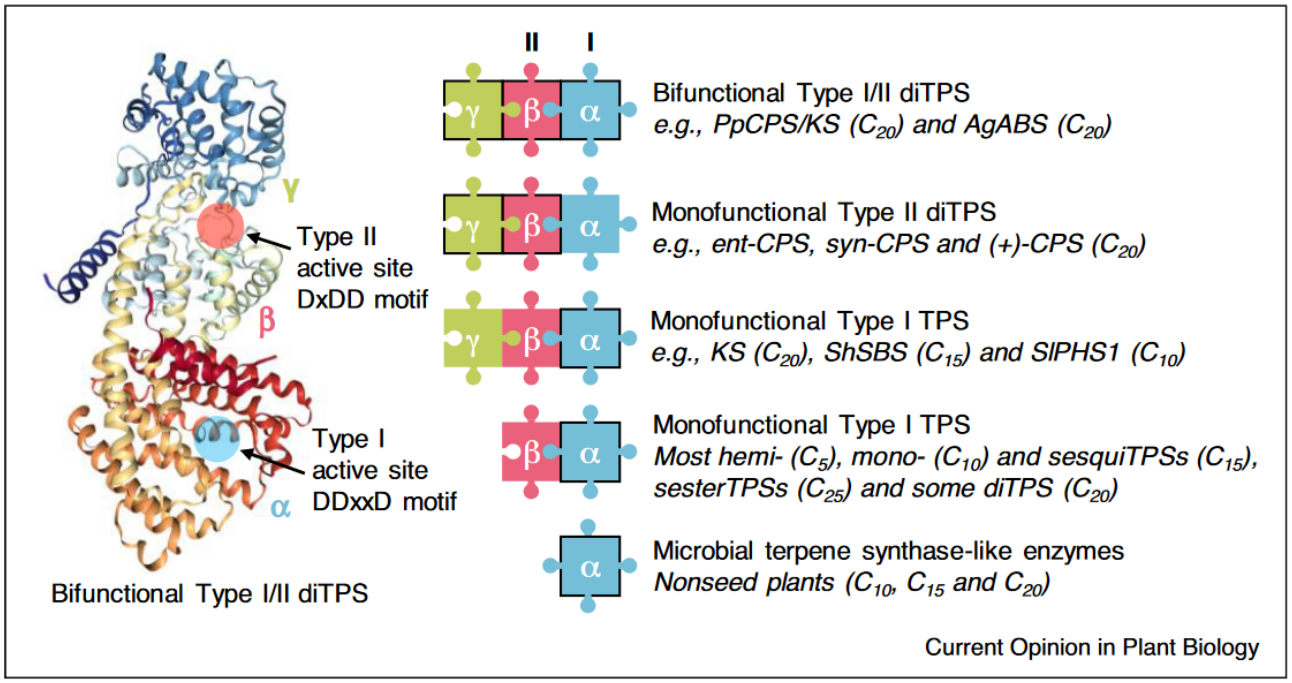

Figure 1 from Zhou & Pichersky, 2020. The representative structure of a bifunctional type I/II diterpene synthase is shown on the left. On the right are structural features of plant TPSs based on the combination of the γ, β, and α domains is. Functional domains are marked with boxes.

Of particular importance for this work is a class of diterpenes called labdanes. They do not volatilize away during fossilization but may polymerize over time to form amber (conversation with Seyfullah, Anderson & Crelling, 1996). Generally speaking, there are two possible biosynthetic pathways (see Figure 2) to form Labdane-related-Diterpenoids (=LRDs). Both pathways start from Geranylgeranyl diphosphate (=GGPP). The first pathway turns GGPP into an intermediate product using a Class II DTPS. These intermediate products may then be further converted using a Class I DTPS into LRDs. The second pathway skips the intermediate step using bifunctional Class I/II DTPSs, directly turning GGPP into LRDs (Gao et al., 2021). The direct pathway is more likely to yield LRDs, therefore we focus on Class I/II DTPSs when trying to figure out possible connections between amber forming species of chemicals.

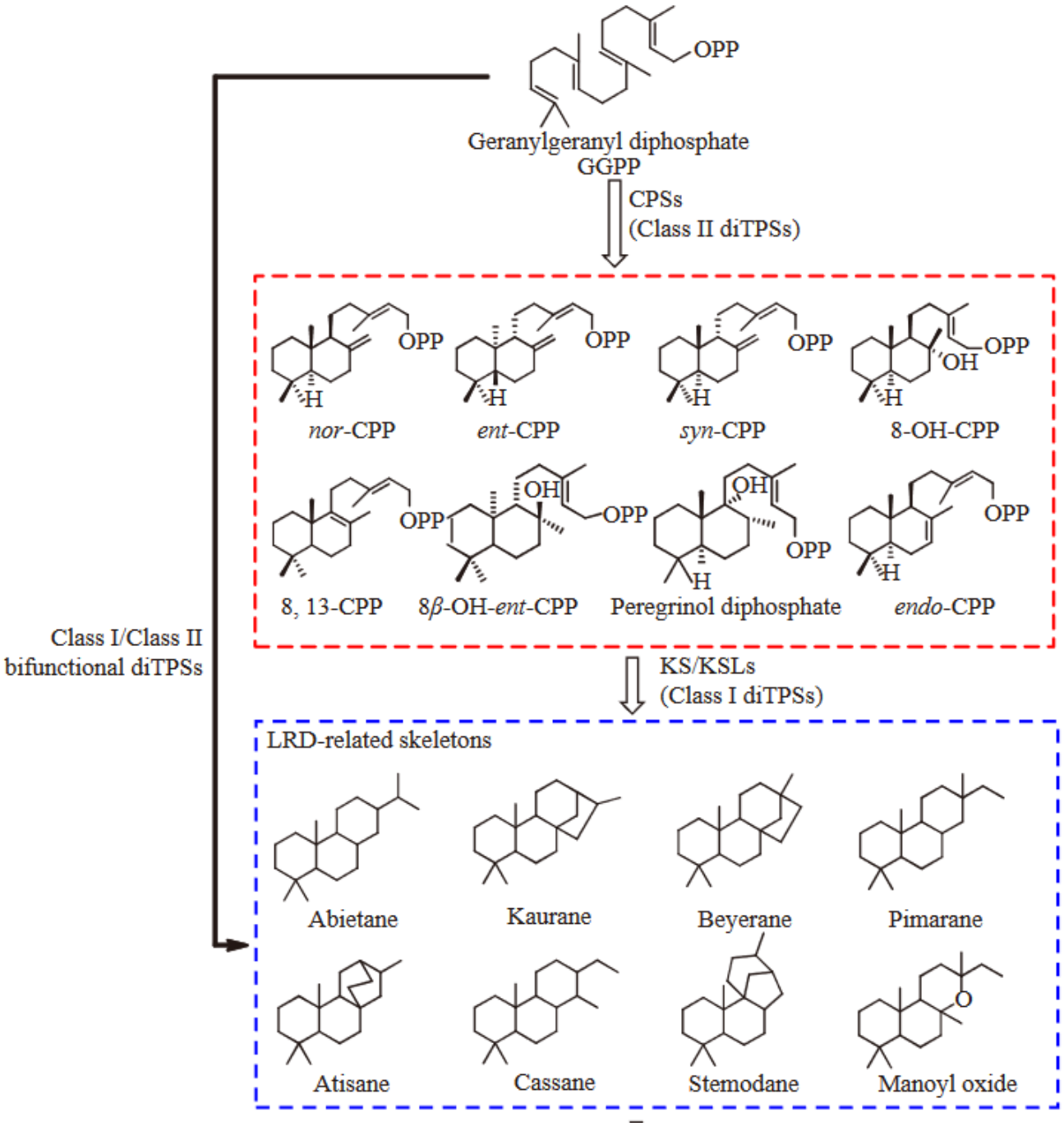

Figure 2 from Gao et al., 2021. General biosynthetic pathway of LRDs. CPS: copalyl pyrophosphate synthase; KS: kaurene synthase; KSL: kaurene synthase-like enzyme

A short introduction of the programs used

BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi and https://db.cngb.org/onekp/)
The word is an acronym for "basic local alignment search tool". BLAST identifies homologous sequences using a heuristic method which initially finds short matches between two sequences, without taking the whole sequence into account. After an initial match, it attempts to start local alignments from these initial matches (Altschul et al., 1990). I used the BLASTp variant, which compares protein queries to protein databases (Altschul et al., 1997).

HMMER (http://hmmer.org/)
"The HMMER webserver is a free-to-use service which provides fast searches against widely used sequence databases and profile hidden Markov model (HMM) libraries using the HMMER software suite" (Potter et al., 2018). In essence, homologies are detected and identified by constructing a profile-HMM ("Hidden Markov Model," 2024) and then comparing to either a target sequence or a database ("HMMER," 2024). In this case the target database was Pfam (Mistry et al., 2021).

Phyx (https://github.com/FePhyFoFum/phyx)
Phyx is a program suite written for linux and Mac. It can perform various phylogenetic analyses and is capable of handling large collections of sequences. I used the program to prune and clean the list of sequences.

MAFFT (https://mafft.cbrc.jp/alignment/server/index.html)
MAFFT is a free online tool (Katoh et al., 2019) used to create **M**ultiple sequence **a**lignments using **F**ast-**F**ourier-**T**ransformation (Katoh et al., 2002). Multiple sequence alignments are used to infer evolutionary relationships and can highlight homologous features between sequences. Alignments highlight mutation events such as deletions ("Multiple Sequence Alignment," 2024).

IQ-TREE (http://www.iqtree.org/)
IQ-TREE is a free online tool. It takes a *multiple sequence alignment* as input and will reconstruct an evolutionary tree that is best explained by the input data (Nguyen et al., 2015). Multiple different algorithms are compared and the most suitable is chosen automatically (Minh et al., 2020).

Material and Methods

Sequence acquisition

Sequences from the genus *Araucaria*, *Agathis*, Podocarpaceae, Cupressaceae, Sciadopityaceae, Zamiaceae, Cycadaceae, Taxaceae, *Cephalotaxus*, Ginkgoaceae, *Welwitschia*, Gnetales, Nymphaeales and Amborellales were obtained via protein BLAST (BLASTp) from the 1kp project page (*BLAST Service of CNGB*, n.d.) and the NCBI project page (*BLAST: Basic Local Alignment Search Tool*, n.d.). We used the sequence from the e-beta farnesene synthase of *Pinus sylvestris* (GenBank ID:ADH29869.1) as the input query sequence. It is a known terpene synthase sequence. This input query sequence allowed me to find similar sequences from all the families listed above. Additional sequences were acquired from the review on the evolution and functional diversity of terpene synthases in the *Pinus* species (Alicandri et al., 2020).

Sequence selection

All hits were downloaded and concatenated to form a complete list in the Fasta format. The sequences of this list were then manually submitted one by one to HMMER to search against the Pfam database (*Biosequence Analysis Using Profile Hidden Markov Models | HMMER*, n.d.). The presence of motif PF03936 and motif PF01397 was used as a criterion to identify Terpene Synthase sequences. If a sequence lacked either of these motifs it was marked for removal using an excel sheet (see Appendix "BIG table"). These marked sequences were then removed from the concatenated list of all sequences using phyx (*Phyx*, n.d.). To use the program, the concatenated list of all sequences had to undergo multiple rounds of manual cleaning, as phyx only allows certain data formats and symbols. This would turn out to be a recurring theme.

Multiple Sequence Alignment

The pruned list of usable sequences was uploaded to an online multiple sequence alignment program called MAFFT (Katoh et al., 2019). MAFFT only accepts certain data formats and occurring symbols, so another round of cleaning was in order. MAFFT then produced a multiple sequence alignment, which was once again downloaded in Fasta format. The programs settings were left on default for this.

Phylogenetic tree

The resulting alignment was then put into IQ-TREE (Minh et al., 2020), an online tool for creating phylogenetic trees based on multiple sequence alignments. IQ-TREE only accepts certain data formats and occurring symbols but is smart enough to clean up most errors by itself. Some errors had to be corrected by hand for the program to function. These errors within the multiple sequence alignment and the list of all usable sequences were corrected. The resulting multiple sequence alignment was uploaded to IQ-TREE once more and resulted in the first completed phylogenetic tree, downloaded in the ".treefile" format. This treefile was then displayed using FIGtree to make the information digestible (see Figure 3).

Further data exploration

First readings of the tree revealed possible findings. To be sure we chose to run a secondary round of analysis. Previously unavailable sequences from the genus *Wollemia* were acquired. The *Wollemia* sequences were similarly pruned and cleaned as described above, then added to the list of sequences. All previous steps, including a lot of manual cleaning, were repeated, and allowed me to create a full phylogenetic tree (see Appendix "Supertree").

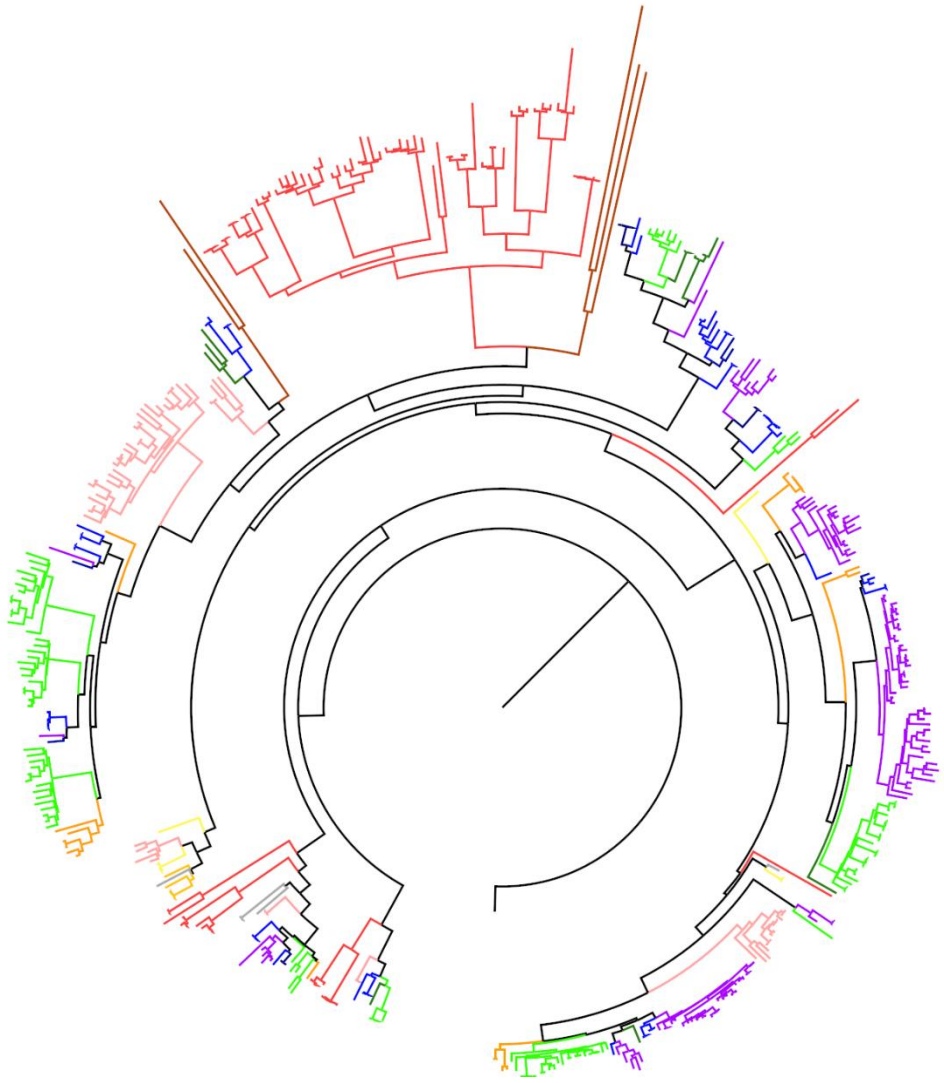

Figure 3: The OGTree. The sequneces were color coded by genus and each end point represents one sequence. The sequence labels were removed for readability. This first tree allowed for some conjectures, which had to be analyzed further. The overall structure remained consistent with the Supertree in Figure 4.

Two sections warranted deeper analysis: [Group 5] and [Group 6] (See Figure 4). The reasons for this will be discussed in the Results and Discussion sections. Groups 5 and 6 were each analyzed isolated from the rest of the tree. Each groups sequences underwent the steps of selection, alignment and tree building as described above once more. Additionally, I paid careful attention if any information on the sequences function was included in the databanks.

Lastly I analyzed the three dimensional structure of one sequence from group 4 and group 5 each. The structure of group 4's sequences is known, while group 5's was previously unknown. We used Q675L5.2 (Pinaceae) as a representative of group 4 and RSCE_scaffold_2003732 (Araucariaceae) as a representative of group 5. I plugged them into Phyre2 using the programs standard options. The resulting .zip files containing the analysis, supporting information and pictures were downloaded.

Results

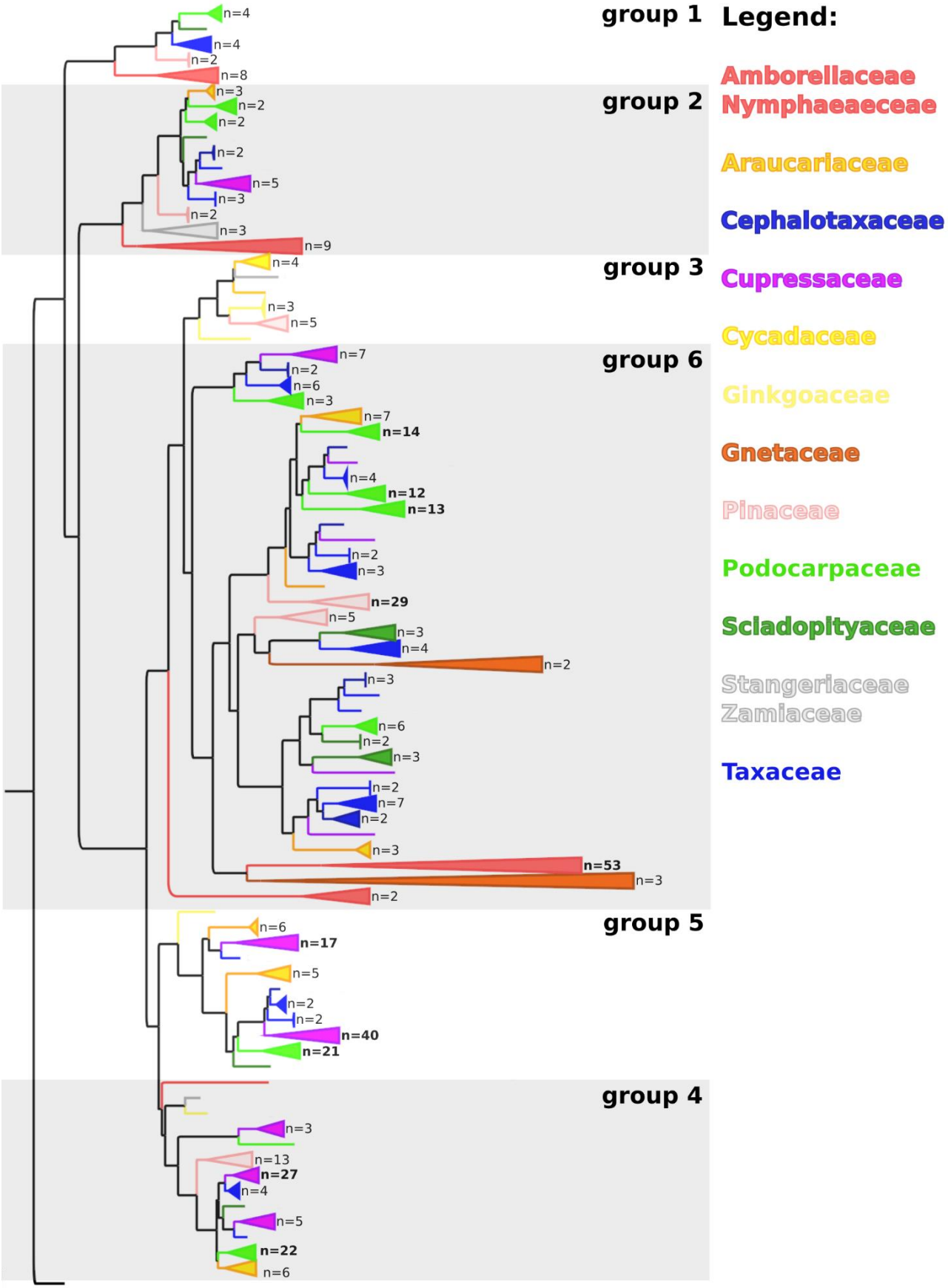

Figure 4 The collapsed Supertree. The main result of this work. All sequences are color coded by genus. Clusters of sequences from the same genus are displayed as triangles. The number of sequences n in a cluster is displayed next to it. Six distinct groups were identified.

The full phylogenetic tree, as seen in the appendix “Supertree”, is the main result of this work. It is impossible to fit on a page due to its size. The tree seen in Figure 4 is a heavily condensed and annotated version of the full tree. To create the collapsed tree, agglomerations of sequences originating from the same genus were collapsed together into clusters. This preserves the structure and evolutionary relationships, while enabling the tree to be viewed in this work. The tree was then split up into 6 distinct groups for ease of analysis. These groups were chosen through the trees structure. Four of the groups could be identified directly (see Table 1), as they featured some of the known Pinaceae sequences acquired from “On the Evolution and Functional Diversity of Terpene Synthases in the Pinus Species: A Review” (Alicandri et al., 2020). Group 5 does not contain any of the known Pinaceae sequences and remained unknown at first. Via its close relation to Group 4, Group 5 was presumed to also be Class I/II DTPS. Group 6 is by far the largest group and is quite diverse, containing at least Class I MTPS, Class I bi-domain STPS and many more unknowns. So, while Group 5 was treated as putative Class I/II DTPS, Group 6 was treated as a complete unknown.

Table 1: initial TPS classifications by group.

	TPS - family	**Identified directly**
Group 1	Class II DTPS	Yes
Group 2	Class I DTPS	Yes
Group 3	Class I tri-domain STPS	Yes
Group 4	Class I/II DTPS	Yes
Group 5	Class I/II DTPS	No
Group 6	diverse	No

Group 4 and 5 will be the focus of the rest of this analysis, as they are the only ones that can form amber. As discussed previously, only Class I/II DTPS can produce the labdenoids necessary to survive the fossilization and polymerization processes needed to form amber. The genera contained in these groups are: Amborellaceae, Araucariaceae, Cephalotaxaceae, Cupressaceae, Ginkgoaceae, Pinaceae, Podocarpaceae, Sciadopityaceae, Taxaceae and Zamiaceae. The respective numbers of clusters and sequences for these genera can be seen in Figure 4 but are summarized in Table 2.

Table 2: distribution of Class I/II DTPS sequences in groups 4 and 5. The sequences were arranged by genus, then sorted by the total number of clusters and total number of sequences. Additionally the number of clusters and sequences per group are displayed.

Genus	Group 4 clusters	Group 4 sequences	Group 5 clusters	Group 5 sequences	Total clusters	Total sequences
Cupressaceae	3	35	2	57	5	92
Podocarpaceae	2	23	1	21	3	44
Araucariaceae	1	6	2	11	3	17
Taxaceae	2	5	3	5	5 (4)	10
Cephalotaxaceae	1	1	1	1	2	2
Ginkgoaceae	1	1	1	1	2	2
Sciadopityaceae	1	1	1	1	2	2
Pinaceae	1	13	-	-	1	13
Amborellaceae	1	1	-	-	1	1
Zamiaceae	1	1	-	-	1	1

Most of the genera can be found in both groups. Cupressaceae has the highest number of sequences at a whopping 92. Comparing this to the total number of sequences, 184, this means that Cupressaceae is responsible for ***exactly*** half of all sequences featured in group 4 and 5. Podocarpaceae has 44 sequences, split into 3 clusters. Two of these clusters are split between groups 4 and 5 and are almost identical in size at 22 and 21 sequences respectively. Araucariaceae features three clusters of similar size: 5,6 and 6 sequences respectively. Taxaceae features 10 total sequences split up into 5 clusters. Two clusters found in group 5 are closely related (see Figure 4), only broken up by a singular Cephalotaxaceae sequence. Arguably these clusters could have been merged, bringing the total number of clusters down to four. Some genera are only featured in group 4. They are Amborellaceae, Zamiaceae and importantly Pinaceae. Pinaceae features a medium number of sequences at 13, all of which are found in one singular cluster in group 4.

Now that the description of the results is complete, lets dive into their significance and the broader implications derived from them.

Discussion

What the results tell us

Groups 4 and 5 are both Class I/II DTPS. As Class I/II DTPS are responsible for production of LRDs I can confidently state that the genera included in groups 4 and 5 produce terpenes required for resin production. The genera included in groups 4 and 5 happen to be the genera that make up the conifers. Surprisingly Amborellaceae and Zamiaceae are also present near the root of group 4, but they are not relevant to this discussion.
The number of sequence clusters of a reasonable size is a good indication for the number of Class I/II DTPS copies a given family has. And indeed, one can find clusters of (almost) all conifer families present in both groups. As *Ginkgo biloba* is also present in both groups, this suggests a duplication event in the distant past of the conifers. Part of the DNA responsible for DTPS was duplicated. The sole exception is Pinaceae, which only has one cluster of sequences in group 4. At some point in their evolutionary history the second copy must have undergone a deletion event. This evidence has not been described before to date. There is evidence of many more duplications and subsequent deletions, but it is extremely difficult to pin down when they might have happened, as they are not as immediately obvious as the Pinaceae. All I can state with confidence is that Cupressaceae, Taxaceae and Araucariaceae have at least three copies each, while Podocarpaceae and Sciadopityaceae are likely to only have two copies of Class I/II DTPS each.

Data coverage

In this thesis, I can only work with a subset of the gymnosperms. This stems from the fact that data is limited. Even if a species is known to exist, it does not necessarily have genetic information assigned to it. Even if genetic information exists, it is sometimes misattributed, misnamed, falsely categorized or inaccessible. A lot of information we have today is therefore incomplete or outdated. This issue is exacerbated while working with data from multiple sources. As there is no one unified source, different databanks use different formats, layouts and names for the same species. Thankfully, some conventions have established themselves and are used across the board, namely the Fasta data format. Even then, the contents within are often formatted in interesting ways. These are the reasons for the many hours spent manually cleaning, converting and translating data, described in the Material and Methods section.
The question remains if the subset used in this work is representative of the gymnosperms and by extension, the conifers. In Table 3 I compare the data acquired to the gymnosperms at large. For this comparison I use the database of gymnosperm taxonomy found under www.conifers.org (Earle, n.d.). It provides descriptions for

each order, family and genus of the gymnosperms known today. In addition, I add the number of genera of each family that appear in groups 4 and 5.

Table 3 Comparing known genera to what is available in the databanks used and the two important groups. Percentages are always in comparison to the absolute number of genera in each family. Relevant genera were color coded in a way to remain consistent with Figure 4. The colored genera also make up the conifers + Ginkgo.

Order	**Family**	**Genera**	**Data coverage**		**Group 4+5 coverage**	
			Total	%	Total	%
Cycadales	Zamiaceae	10	3	30%	1	10%
	Cycadaceae	1	1	100%	0	0%
Ginkgoales	Ginkgoaceae	1	1	100%	1	100%
Gnetales	Gnetaceae	1	1	100%	0	0%
	Welwitschiaceae	1	0	0%	0	0%
	Ephedraceae	1	0	0%	0	0%
Pinales	Pinaceae	11	4	36%	3	27%
Araucariales	Araucariaceae	3	3	100%	3	100%
	Podocarpaceae	20	16	80%	13	65%
Cupressales	Sciadopityaceae	1	1	100%	1	100%
	Cupressaceae (+Taxodiaceae)	28	24	86%	24	86%
	Taxaceae (+Cephalotaxaceae)	6	6	100%	5	83%

From this one can infer some trends. First of all, and perhaps most obvious:

Commercial viability drives research. The genera not represented in the databases whatsoever, as is the case with Welwitschiaceae and Ephedraceae, have little to no economic value. These two generally occupy dry arid habitats and are unpalatable to livestock. Similarly, Zamiaceae has little to no economic value and has a low rate of data coverage. This trend of economic viability continues with the Pinaceae. There are only four of them in the dataset. They happen to be the four genera of commercial value: *Abies*, *Picea*, *Pinus* and *Pseudotsuga*.

High accessibility makes a good target for research. If a genus is easily found and widely distributed it is more likely to be researched. This is the case for Podocarpaceae, Cupressaceae and Taxaceae, which all show high rates of data coverage. Cupressaceae is extremely prevalent, it can be found on all continents except antarctica. It is the family with the most genera out of all the gymnosperms. The data coverage of 86% is remarkable. Podocarpaceae are mostly confined to the

southern hemisphere and are native to tropical or subtropical forests. Still, they can be found on most continents across the globe. Lastly, Taxaceae can be found in most of Europe, North America and Asia.

Unique species are interesting. Ginkgoaceae and Sciadopityaceae both have exactly one species each (monotypic). *Ginkgo biloba* is often regarded as a living fossil, has a unique appearance and a pungent smell. *Sciadopytis verticillate* is only native to Japan, has commercial value for its wood and as an ornamental piece. As it grows slowly it is often found near the country's oldest temples.
Lastly, regarding Cycadaceae and Gnetaceae. Both appear to have extremely high rates of data coverage. This is not the case. This stems from the way the data is presented. In the table I focused on the number of genera included in the dataset. For most families, the number of genera is a good enough indicator of the number of species. Cycadaceae only has one genus, *Cycas*, but includes 120 different species. *Gnetaceae* also has one genus, *Gnetum*, with 40 species. The dataset includes two species of *Cycas* (*Cycas micholitzii* and *Cycas taitungensis*) and one species of *Gnetum* (*Gnetum montanum*).

When comparing the coverage in groups 4 and 5 to the total number of genera it is noticeable that only the conifers + Ginkgo are well represented. The odd ones out are Podocarpaceae and Pinaceae. For Podocarpaceae the comparatively low data coverage may be explained by the lack of research. The family is often thought of as lacking when it comes to detailed analysis of its chemistry and ecology (Langenheim, 2003; Anderson & Crelling, 1996). In that regard, 65% data coverage might indicate more attention in recent years. Pinaceae is an odd case and will be discussed in detail in its own chapter.It makes sense, that the conifers are rich in Class I/II DTPS DNA among the gymnosperms. For the exact reasons one only needs to look at the anatomy of conifers today.

Anatomy and its potential link to amber

The anatomy of conifers today informs us of their past. From modern anatomy one can extrapolate likely ancestral anatomy. Most conifers today produce resin as a response to physical injury. The resin acts as an insect repellant, anti-bacterial agent and to close off the wound (Langenheim, 2003). But only four genera produce resin in specialized secretory structures: Pinaceae, Araucariaceae, Podocarpaceae and Cupressaceae. Out of these four only Pinaceae and Araucariaceae are known to produce large quantities of resin today.
The secretory structures of Pinaceae have been studied most intensively. Differences between individual species can even be compared. Pinaceae develop specialized resin ducts along the stem, trunk and leaves of the tree. In case of injury, the resin is pumped towards the wound to close it. This means that little to no resin is created in-situ as a response. Nonetheless, Pinaceae can excrete large amounts of resin. The secretory structures of the remaining three have been studied less extensively. They produce resin in small, localized pockets or cysts. In case of injury the nearby pockets excrete the resin within, and additional resin is created in-situ. Taxaceae produce TPS similar to Pinaceae. They produce large amounts of leaf resin but have no secretory structures in the stems. (Langenheim, 2003; Anderson & Crelling, 1996; Seyfullah et al., 2018).

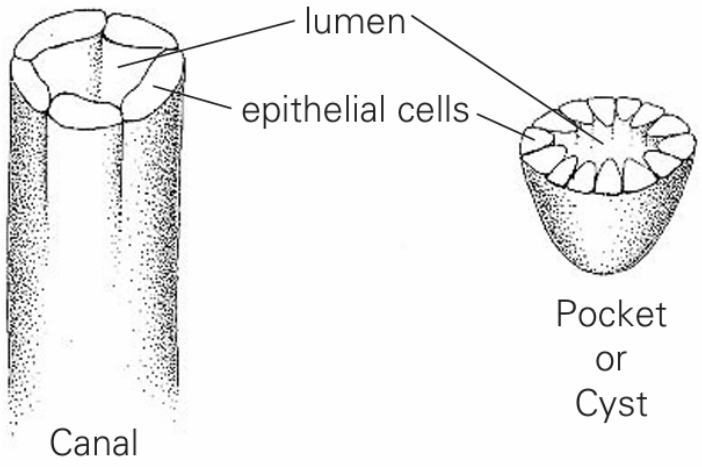

Figure 5 form (Langenheim, 2003). Comparison of canals (ducts) or pockets. Resin is produced in the epithelial cells and stored in the lumen.

To summarize: My phylogenetic data suggests, that Cupressaceae, Podocarpaceae, Araucariaceae, Pinaceae and Taxaceae (incl. Cephalotaxaceae) have the Class I/II DTPS necessary to produce Labdenoids. Anatomical information suggests that, excluding Taxaceae, these genera also have specialized secretory structures for resin production. Podocarpaceae is not well studied in comparison to the others, while Pinaceae is studied extensively, this study is limited to a few species. Current opinion in plant biology is that the Araucariales are likely the oldest extant order, so they

would have had the most time to produce resin (Leslie et al., 2018). Araucariaceae and Pinaceae produce the most resin today.
This suggests the following: Araucariaceae should have the highest amount of amber attributed. Following should be Cupressaceae and Pinaceae. Lastly Podocarpaceae should have the least amber attributed to it.

Now to compare this hypothesis to actual data of modern plant families suggested as sources of amber (see Figure 6, Langenheim, 2003). The longer a genus has produced amber, the more of it can be found. While sudden bursts of amber production are theoretically possible, this has never been proven. Araucariaceae is the source of the most coniferous amber. Cupressaceae (including Taxodiaceae) is the second largest source. Pinaceae is third and Podocarpaceae is last (Langenheim, 2003; Seyfullah et al., 2018). This remains consistent with my earlier hypothesis.

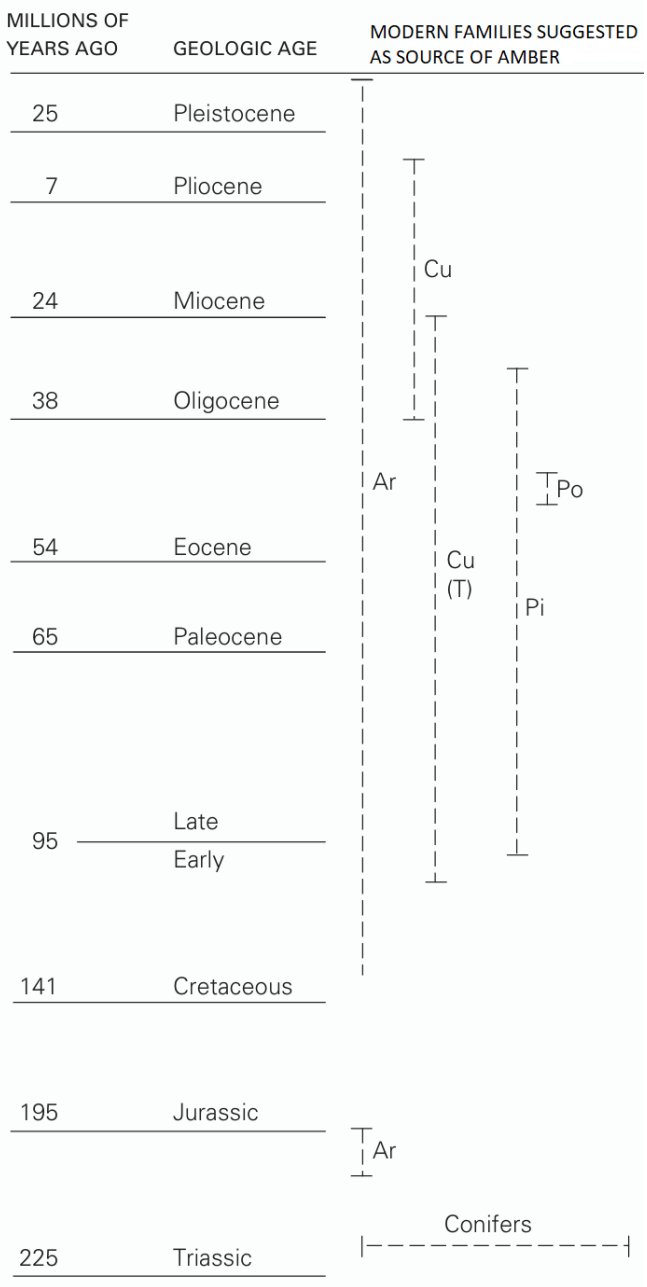

Figure 6 from (Langenheim, 2003) Modern plant families reported in literature as sources of amber through geologic time:

Ar, Araucariaceae; Cu, Cupressaceae; Cu (T), Cupressaceae including Taxodiaceae; Pi, Pinaceae; Po, Podocarpaceae.

Vertical line segments approximate the general range of time through which the family has been recorded

A place to put Pinaceae

For the class Pinophyta (= literally the plants that are like Pinaceae), Pinaceae is a remarkably bad representative. Not only does it feature anatomical structures unique among the conifers, but its resin production also differs. As I have shown, it is the only genus among the conifers with only one copy of Class I/II DTPS.

Pinaceae features the highest number of species among all gymnosperms. The family is found across the globe and commercially viable. It comes as no surprise that Pinaceae is extensively studied. As an example, Alicandri et al., 2020; Mao et al., 2019; Hall et al., 2013 all focus on different Pinaceae species to infer information about the conifers. But perhaps Pinaceae should not be used as a sole representative. As discussed above, Pinaceae is the only family that develops specialized resin ducts. And while Pinaceae is highly resinous today, the resin does not preserve well. This makes it rare in the fossil record (Seyfullah et al., 2018). Pinaceae only has one Class I/II DTPS, which may explain the resins instability, although this requires further research.

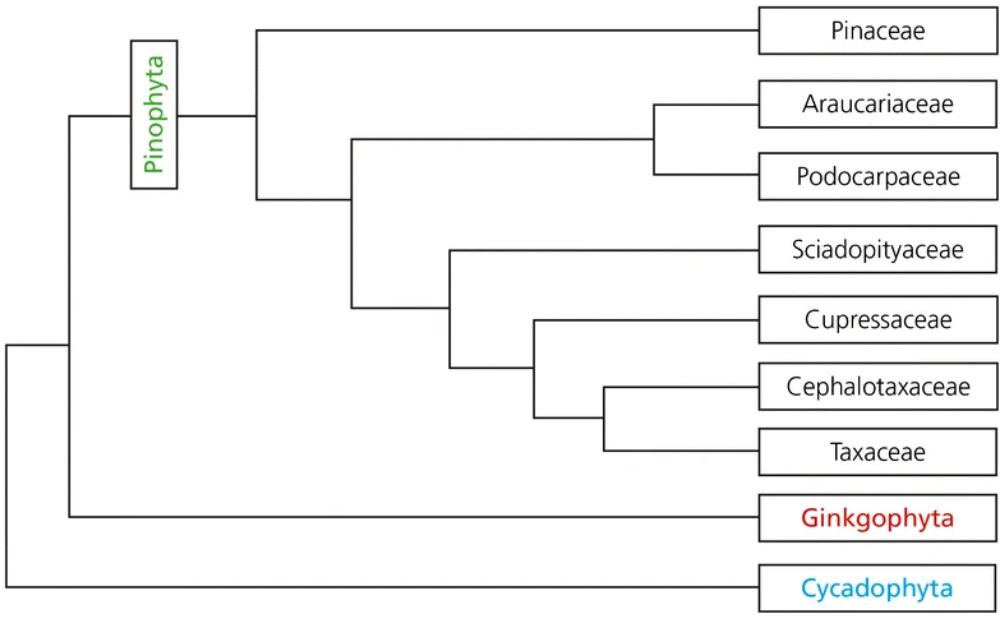

Figure 7 taken from "The Kew Report", Farjon, 2018. A commonly presented phylogeny of conifers.

Finding a place for Pinaceae in the phylogeny of the conifers is also a contentious issue. Commonly the family is placed as illustrated in Figure 7, suggesting that Pinaceae split off early in the evolution of the conifers. The family then split off into its many species along the way. My data suggests, the Pinaceae split from the rest of the conifers, did not happen in isolation. It was much closer to the time when the Araucariales and the Cupressales split off. This is consistent with more modern interpretations, where Pinaceae split into many genera originally. From these original genera, a handful survived and only recently formed the species diversity Pinaceae is known for today (Leslie et al., 2018).

Conclusion and looking to the future

In conclusion, I have explored part of the evolutionary history of the conifers through modern bioinformatics. I highlighted the processes needed to create phylogenetic trees of significant size and the potential pitfalls therein. By conducting a comprehensive phylogenetic analysis of the TPSs of gymnosperms, this study has identified distinct groups sharing similar TPSs. By focusing on Class I/II DTPSs, responsible for LRD production, this study has unveiled significant events in the evolutionary history of the conifers.
This thesis contributes to the fields of evolutionary biology, bioinformatics, and paleobotany by integrating digital phylogenetics with anatomical studies and fossil amber research. I hope that the data acquired in this work can aid future research in these fields and perhaps hitherto unforeseen new branches of biology.

The insights gleamed in this work would have been close to impossible a decade ago. As technology advances and new mathematical models are integrated with ever advancing computational power, evolutionary history becomes increasingly easy to explore. Previously underexplored genera such as Podocarpaceae already show signs of this. This makes me hopeful that one day we will have a complete understanding of this absurd universe of ours. We can never know everything, all the more reason to try!

Bibliography

Alicandri, E., Paolacci, A. R., Osadolor, S., Sorgonà, A., Badiani, M., & Ciaffi, M. (2020). On the Evolution and Functional Diversity of Terpene Synthases in the Pinus Species: A Review. *Journal of Molecular Evolution*, *88*(3), 253–283. https://doi.org/10.1007/s00239-020-09930-8

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Anderson, K. B., & Crelling, J. C. (Eds.). (1996). *Amber, Resinite, and Fossil Resins* (Vol. 617). American Chemical Society. https://doi.org/10.1021/bk-1995-0617

Earle, C. J. E. (n.d.). *The Gymnosperm Database: Home Page*. Retrieved July 4, 2024, from https://conifers.org/

Farjon, A. (2018). The Kew Review: Conifers of the World. *Kew Bulletin*, *73*(1), 8. https://doi.org/10.1007/s12225-018-9738-5

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(Web Server issue), W29-37. https://doi.org/10.1093/nar/gkr367

Gao, K., Zha, W.-L., Zhu, J.-X., Zheng, C., & Zi, J.-C. (2021). A review: Biosynthesis of plant-derived labdane-related diterpenoids. *Chinese Journal of Natural Medicines*, *19*(9), 666–674. https://doi.org/10.1016/S1875-5364(21)60100-0

Gao, Y., Honzatko, R. B., & Peters, R. J. (2012). Terpenoid synthase structures: A so far incomplete view of complex catalysis. *Natural Product Reports*, *29*(10), 1153–1175. https://doi.org/10.1039/c2np20059g

Hall, D. E., Zerbe, P., Jancsik, S., Quesada, A. L., Dullat, H., Madilao, L. L., Yuen, M., & Bohlmann, J. (2013). Evolution of Conifer Diterpene Synthases: Diterpene Resin Acid Biosynthesis in Lodgepole Pine and Jack Pine Involves Monofunctional and Bifunctional Diterpene Synthases1[W][OA]. *Plant Physiology*, *161*(2), 600–616. https://doi.org/10.1104/pp.112.208546

Hidden Markov model. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Hidden_Markov_model&oldid=1223340736

HMMER. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=HMMER&oldid=1193990816

Jia, Q., Brown, R., Köllner, T. G., Fu, J., Chen, X., Wong, G. K.-S., Gershenzon, J., Peters, R. J., & Chen, F. (2022). Origin and early evolution of the plant terpene synthase family. *Proceedings of the National Academy of Sciences*, *119*(15), e2100361119. https://doi.org/10.1073/pnas.2100361119

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166. https://doi.org/10.1093/bib/bbx108

Langenheim, J. H. (2003). *Plant resins: Chemistry, evolution, ecology, and ethnobotany*. Timber Press.

Leslie, A. B., Beaulieu, J., Holman, G., Campbell, C. S., Mei, W., Raubeson, L. R., & Mathews, S. (2018). An overview of extant conifer evolution from the perspective of the fossil record. *American Journal of Botany*, *105*(9), 1531–1544. https://doi.org/10.1002/ajb2.1143

Mao, J., He, Z., Hao, J., Liu, T., Chen, J., & Huang, S. (2019). Identification, expression, and phylogenetic analyses of terpenoid biosynthesis-related genes in secondary xylem of loblolly pine (Pinus taeda L.) based on transcriptome analyses. *PeerJ*, *7*, e6124. https://doi.org/10.7717/peerj.6124

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Multiple sequence alignment. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Multiple_sequence_alignment&oldid=1218424895

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research*, *46*(W1), W200–W204. https://doi.org/10.1093/nar/gky448

Seyfullah, L. J., Beimforde, C., Dal Corso, J., Perrichot, V., Rikkinen, J., & Schmidt, A. R. (2018). Production and preservation of resins—Past and present. *Biological Reviews of the Cambridge Philosophical Society*, *93*(3), 1684–1714. https://doi.org/10.1111/brv.12414

Singh, B., & Sharma, R. A. (2015). Plant terpenes: Defense responses, phylogenetic analysis, regulation and clinical applications. *3 Biotech*, *5*(2), 129–151. https://doi.org/10.1007/s13205-014-0220-2

Sofi, M. Y., Shafi, A., & Masoodi, K. Z. (2022). Chapter 10—NCBI BLAST. In M. Y. Sofi, A. Shafi, & K. Z. Masoodi (Eds.), *Bioinformatics for Everyone* (pp. 95–102). Academic Press. https://doi.org/10.1016/B978-0-323-91128-3.00021-5

Tholl, D. (2006). Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Current Opinion in Plant Biology*, *9*(3), 297–304. https://doi.org/10.1016/j.pbi.2006.03.014

Tholl, D. (2015). Biosynthesis and Biological Functions of Terpenoids in Plants. In J. Schrader & J. Bohlmann (Eds.), *Biotechnology of Isoprenoids* (pp. 63–106). Springer International Publishing. https://doi.org/10.1007/10_2014_295

Zhou, F., & Pichersky, E. (2020). More is better: The diversity of terpene metabolism in plants. *Current Opinion in Plant Biology*, *55*, 1–10. https://doi.org/10.1016/j.pbi.2020.01.005