Building a map of plant triterpenoid occurrences using literature and citizen-collected mass spectrometry samples

A. OVERVIEW AND AIMS

Plant cultivation has been a crucial and inseparable part of human history because plants are a major source of food. Crop plant resilience, like that of many other organisms, is greatly influenced by chemistry. Understanding plant chemicals can therefore help meet the important goals of efficient plant growth and high crop yields. These goals in turn will help to (i) reduce the negative environmental impacts of agriculture and (ii) produce food using fewer resources.

One class of chemicals that has a big influence on plant health and growth is the triterpenoids. These compounds have a general molecular formula of $C_{30}H_nO_m$, usually have five 5- or 6-member carbons rings, as well as one or more oxygen-containing functional groups (Fig. 1). They also have a variety of bioactivities. For example, triterpenoids like lupeol and ursolic acid are associated with post-harvest weight and firmness of highbush blueberries (*Vaccinium corymbosum*) (Moggia et al. 2016), the triterpenoids α-amyrin, β-amyrin, and simiarenol seem to help the grain crop *Sorghum bicolor* be exceptionally tolerant to drought conditions (Busta et al. 2021), and the triterpenoids ursolic acid, echinocystic acid, and oleanolic acid help the desert shrub *Rhazya stricta* stay hydrated at elevated temperature (Schuster et al 2016).

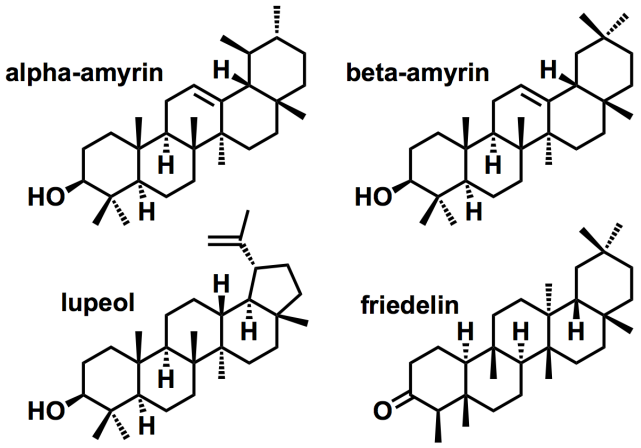

Figure 1. *Some common and structurally simple triterpenoids.*

Triterpenoids can also help defend a plant against other living organisms. For example, birch trees produce the triterpenoid betulin that deters the Colorado Potato beetle (Huang et al., 1995), cork trees produce the triterpenoid friedelane to defend against the parasites *Trypanosoma cruzi* and *Leishmania infantum* (Moiteiro et al., 2006), and citrus trees produce limonoid triterpenoids with anti-malarial (against *Plasmodium falciparum*), anti-microbial (against bacteria *Streptococcus pyogenes*, *Staphylococcus aureus,* and *Botrytis cinerea*) and insecticidal (against fall armyworm) activities (Roy et al., 2006).

Finally, triterpenoids also serve as building blocks for more structurally complex classes of biological defense compounds. One example of such triterpenoid derivatives are the saponins. Saponins are amphipathic compounds made from a combination of fat-soluble triterpenoid rings and a water-soluble oligosaccharide chain (Fig. 2). Experimental evidence suggests that their

amphipathic character allows saponins to permeabilize membranes, particularly the intestinal cells of insects such as pea aphid and cotton leafworm (De Geyter et al., 2011; De Geyter 2012), leading to feeding deterrence, growth inhibition, or outright fatality (Singh et al. 2018). In agriculture, these pest management functions are usually carried out using synthetic pesticides, but synthetic pesticides are often not biodegradable and can be toxic to humans and other non-target organisms, causing damage to the environment and negatively affecting the ecosystem. Replacing synthetic pesticides with natural plant metabolites like saponins (triterpenoid derivatives), therefore, will likely achieve three important objectives: (i) ensure the safety of the crop, (ii) minimize the toxicity of crop products, and (iii) make agriculture more sustainable (Singh et al. 2018).

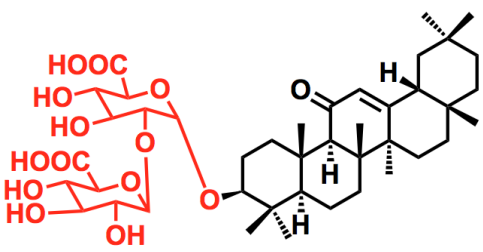

Figure 2: *An example of the amphipathic saponin group compounds: glycyrrhizin. This compound includes a nonpolar & fat soluble triterpenoid group (black) and a polar & water soluble sugar group (red).*

Based on the studies summarized above, triterpenoids and their derivatives have the potential to make a big impact on our agricultural system, but only if we can understand how they function. However, we have a ***problem*** in that we know of only a limited number of biological systems in which to study triterpenoid function in detail. Therefore, there is a ***critical need*** to understand which triterpenoid compounds can be found in which plant species so we can identify new study systems for future triterpenoid research. This is a critical need to which an analytical chemist can make a major contribution. Accordingly, the ***objective in this proposal*** is to build two "maps" (example in Fig. 3) that show the presence of different triterpenoid compounds along with their relative abundance using existing and new data.

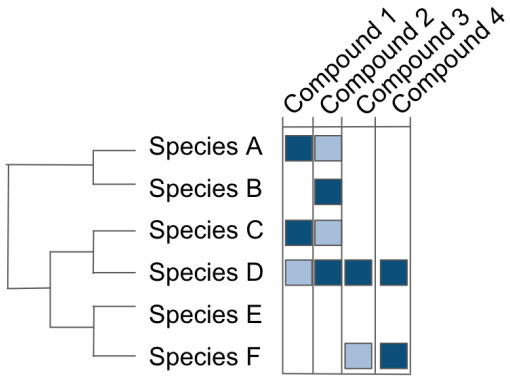

Figure 3: *An example of a "map" of compound occurrence. a phylogenetic tree of all the plant species on the left; a table indicating which triterpenoid compound is absent (white), minor (light blue) or major (dark blue) within wax tissues in each respective species.*

In order to meet the objective of this proposal, I will carry out the following two aims:

Aim 1: Document known occurrences of plant triterpenoids based on published studies.
Aim 2: Document new occurrences of plant triterpenoids using at least 100 plant samples.

Triterpenoids have been documented before in certain individual plant species. In contrast, this study will focus on (i) systematically gathering and organizing information on existing sources of plant triterpenoids into a literature review and also (ii) acquiring new triterpenoid occurrence data by analyzing samples coming from many different plant species sent by citizen scientists. The combined dataset generated during this project will be a collection of triterpenoid information on a large, plant kingdom-spanning scale, which has never been done on this scale before. The maps that I create will give us a greater number of biological systems in which to study triterpenoids and triterpenoid derivatives like saponins. This will greatly contribute to future experiments that directly test triterpenoid function. This in turn will help to make agricultural production better in three different ways: (i) keep the plants safe, (ii) reduce pesticides residues on plant products, and (iii) make agriculture more environmentally friendly.

B. APPROACH

In plants, triterpenoid compounds can be found within all sorts of tissues: roots, stems, leaves, bark, among others. However, in order to finish this project on time, triterpenoids occuring on plant surfaces will be the only target considered in this project (studying triterpenoids in all possible tissues would take more than 2 years). As a result, plant surface waxes, which contain a lot of triterpenoids and are easy to access, will be used as a tool for studying triterpenoids in this project.

Aim 1

The goal of aim one is to construct a map of triterpenoid occurrence based on published information. In order to build this map, there are three major steps: (i) find and select scientific papers reporting the occurrence of specific triterpenoids from waxes of specific plant species, (ii) analyze those papers and use them to build a spreadsheet of triterpenoid occurrence, and (iii) analyze this spreadsheet and construct an occurrence map.

Step 1: Literature Search [complete]. The first step I performed was an extensive search for articles that mentioned specific triterpenoids occurring on the surfaces of specific plant species. Some examples of queries used in this search are: “triterpenoid”, “wax” and “plant”. I picked only articles dated from the 1970s to as recently as 2020 and from these, it was necessary to filter again and select articles with quantitative tables and with descriptions of instruments and extraction methods used, as well as the specific plant tissue studied. This resulted in 51 different articles, which I organized into a spreadsheet with their authors, date, and title. I then put the 51 articles into five categories: (i) articles describing triterpenoid occurrence (37 articles), articles describing triterpenoid function (4 articles), articles describing triterpenoid evolution (4 articles), articles describing triterpenoid biosynthesis (2 articles), and titles to which full text access was not available (4 articles). Then, as the aim is building a qualitative library of triterpenoid occurrence, only the 37 papers that focused on triterpenoid occurrence were retained for more review.

Step 2: Extract Data from Relevant Papers [complete]. With the list of 37 articles in hand, it was time to start building the triterpenoid library. I went through each of the 37 papers carefully and documented which triterpenoids were in each plant species. To do this, I recorded eight things: (i) the plant genus, (ii) the plant species, (iii) the tissue location where it was being extracted, (iv) the extract solvent being used for sample analysis, (v) the instrumental system being used for analysis, (vi) the compound system/common name, (vii) their relative abundance in the sample (major or minor) and finally, (viii) their respective reference paper. In the end, I documented 705 triterpenoid occurrences (129 unique triterpenoids) across 79 plant species based on 37 scientific articles.

Step 3: Analyze Extracted Literature Data [in progress]. I am currently using the literature data to build a map of triterpenoid occurrences. The final map will contain two main components: a phylogenetic tree of all the plant species and a table indicating which triterpenoids compound is absent, minor or major in each respective species (see an example in Fig. 3). The 705 triterpenoid occurrences are found in a total of 79 plant species; and I plan to utilize the software RStudio to help with the construction of this tree. I have used RStudio to analyze and make plots on CSV data before, so this is a perfectly possible task for my second year of graduate school. As the re-organized data table can be easily built from the triterpenoid occurrence excel template, the remaining work is spending the coming two months to analyze the data and build a phylogenetic tree; and with that the first aim of my thesis will be achieved. In addition, if time permits after data analysis; I plan to do more interpretation on the map about what information it would tell us about plant triterpenoids.

Potential Pitfalls and Alternative Approaches - A potential pitfall which can occur: contradicting information from different sources about 1 type of triterpenoid can cause serious problems in building the final map. In this event, my plan is to read the conflicting articles thoroughly, and make a decision to rely on whichever has a more reliable and detailed method section. This can help to know if the research was done with good attention and proper conduct.

Aim 2

The goal of the second aim is to build a map of triterpenoid occurrence based on at least 100 plant samples that I analyze with gas chromatography-mass spectrometry (GC-MS). This map is intended to be more quantitative than the first one, and will be constructed via three major steps: (i) sample preparation, (ii) GC-MS analysis, and (iii) data analysis and map construction using RStudio.

Step 1: Sample documentation & preparation [in progress]. Many samples of different plant species across the US are sent to Dr. Busta's lab by citizen scientist collaborators. These samples are collected on cotton swabs, and it is necessary to document all of them first. A spreadsheet was created for this purpose and each sample that arrives is given an ID number. The ID number of each sample is recorded along with the collector's name, the state of origin, and most importantly, the plant species' scientific and common name. At this point, after the first documenting step is finished, the samples will be extracted and prepared for GC-MS analysis. Each GC vial is labeled according to the ID of its sample and 1 mL of chloroform is added to each vial. Each swab is removed from its pouch and is immersed in the respective vial containing chloroform for 1 minute. After that, the swabs can be discarded and all the chloroform can evaporate. When chloroform is confirmed to be completely evaporated, 125 uL of pyridine and 125 uL of N,O-Bis(trimethylsilyl)trifluoroacetamide (BSTFA) can be added to the vials to convert triterpenoid alcohols into trimethylsilyl derivatives. Doing this helps to increase volatility and MS fragmentation for triterpenoid molecules. Then they are capped, vortexed, and finally incubated at 70°C for 45 minutes; and after that they will be ready for GC-MS analysis.

Step 2: Acquiring the GC data [in progress]. Prepared GC samples are run on the HCAMS 109 GC-MS system using a GC method developed for this project. Since I do not know the concentration of analytes in the sample (they were collected by citizen scientists), I use an initial injection volume of 1 ul. After a sample is run, I inspect the resulting total ion chromatogram (TIC) and use that to categorize the sample as "good", "bad", or "ok" (Fig. 4). The "good" ones are TICs with very clear and much larger target peaks; and the "ok" are TICs with some peaks of contaminants and target peaks are often not higher than the end of run baseline. The "bad" ones are samples with a TIC where peaks are essentially absent and only noise can be observed so chances for improvement are unlikely. In general, all of the "ok" and some of the "good" ones are the most likely to be put on a rerun for improvement, with adjustment of the injection volume. If the peaks stand out from the baseline but are not visible enough (lots of noise along the baseline) or stood lower than the "ramp baseline" (can be seen within 45-50 minutes retention time parts of the TIC in Fig. 4), the injection volume can be increased. In addition, peaks are expected to fall within the intensity range of around 10^6, so injection can also be increased or decreased if the peaks go out of this range. At the same time, it should be noted that the injection volume is kept to not exceed 3 uL since this can overload the instruments' injection port liner. The purpose of these reruns is to maximize the number of GC samples with "good" data quality.

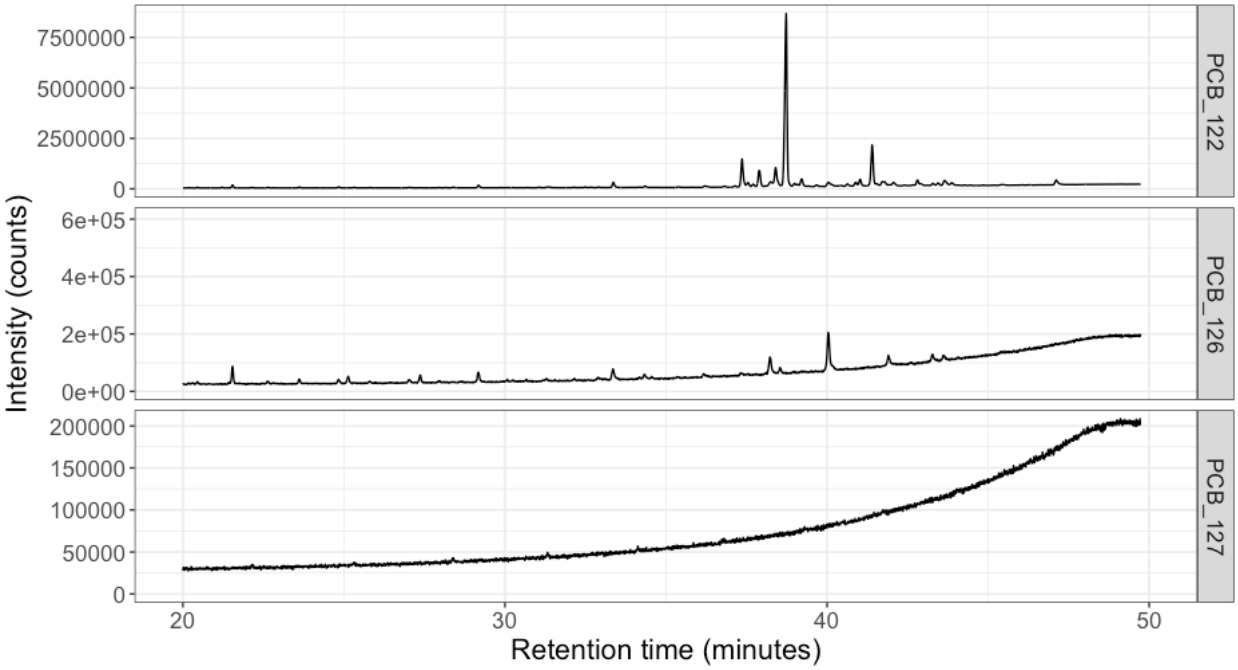

Figure 4: *An example of how GC sample TICs are being categorized: a "good" TIC (PCB_122 on top), an "OK" TIC (PCB_126 in the middle), and a "bad" TIC (PCB_127 at the bottom).*

Step 3: Analyzing the GC data [in progress]. During the first eight months of my work as a graduate student, I have been working to run and rerun all the samples to collect "good" data from as many samples as possible. At the moment, I am working on selecting all the principal and crucial peaks from the TIC of the 54 currently "good" samples and exporting their mass spectra into CSV files. The plan is that these CSV files will then be analyzed using a library of mass spectra of common organic compound groups in plants provided by Dr. Busta and the analytical software RStudio. This will help to identify the main components (alkanes, alcohols, aldehyde, fatty acids, triterpenoids …) along with their percentages in each species sample and these can be documented in the last added column "Composition" of the Google Excel. Although data of all the stated organic components is expected to be obtained; triterpenoid data will be prioritized and with this in mind, a map similar to the one made in the first aim can be created. On the other hand, it is important to know that this map will be expected to be a lot more quantitative: the compounds are not defined simply as "major" or "minor" compounds but they will come with actual number (composition percentages) in their samples. In addition to that, this map will hopefully also be able to demonstrate a wide range of chemical compounds for each plant species and not only the triterpenoid groups. With this map from at least 100 samples being created (54 samples already collected and run are considered good enough to be part of the maps), my second aim can be realized.

Potential Pitfalls and Alternative Approaches - This second map is built using data I personally acquire has a number of potential pitfalls. One possible pitfall is that generating all the data including percentage numbers for all main compounds becomes a very time-consuming task. In the event that this happens, my solution would be to focus exclusively on quantifying triterpenoids in each sample (which is also the main topic of this thesis) and delivering quality information for each of them. In addition to that, another potential pitfall is that there can be a large amount of unidentified compounds. Even though this is not very likely to be the case (we have a good MS library), the best way to solve this problem is using an online database like mz.cloud.org for access to more compound spectra. And in the worst scenario if the compound can not be identified with the online database, I will take detailed notes of those spectra and mark them for future research. A final pitfall is that the GC-MS instrument being used is more efficient in detecting some triterpenoids over the others. A possible solution for this is using another instrumental system like gas chromatography-thermal conductivity detector (GC-TCD) to do a correction factor and compare that with the GC-MS.

C. TIMELINE AND CONCLUSION

At the end of December 2021, I am hoping to finalize construction of the two plant chemistry maps: the literature-based triterpenoid map (Aim 1) and a *de novo* GC analysis-based triterpenoid map (Aim 2). The literature-based triterpenoid map will provide a picture of a qualitative abundance (minor or major) of 705 triterpenoids in 79 plant species, and the *de novo* triterpenoid map will provide a quantitative number of relative percent composition of each triterpenoid present in at least 100 samples from different plant species. With the maps in hand, I can start writing my thesis and prepare a presentation for thesis defense in spring 2022. I aim to finish all tasks by the end of April 2022 (Fig. 5).

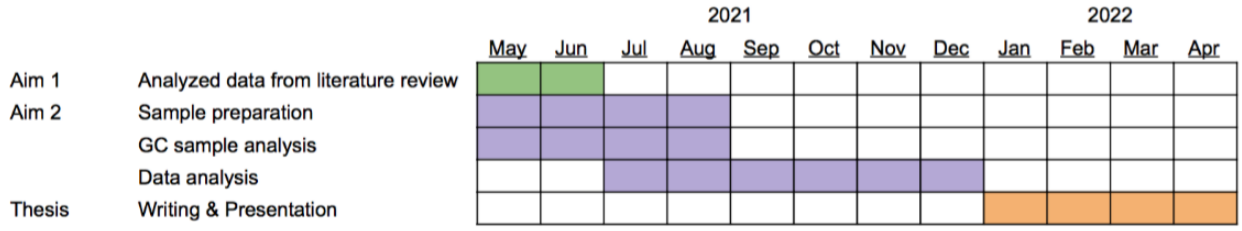

Figure 5: *Project timeline. This is my plan that shows how much time and when I plan to finish all the tasks to achieve my aim 1 (green), aim 2 (purple) and thesis (orange).*

The major impact of this proposal will be to add more available knowledge about triterpenoids that can be used to make a great contribution in improving the agricultural sector in three various ways: (i) keep the crops safe, (ii) provide food products free of synthetic pesticide and (iii) minimize environmental damages during agricultural production.

BIBLIOGRAPHY

Busta, L., Schmitz, E., Kosma, D., Schnable, J., Cahoon, E., **2021.** A co-opted steroid synthesis gene, maintained in sorghum but not maize, is associated with a divergence in leaf wax chemistry. *Proceeding of the National Academy of Sciences,* 118(12).
https://www.pnas.org/content/118/12/e2022982118.short

De Geyter, E., Smagghe, G., Rahbé, Y. and Geelen, D., **2011.** Triterpene saponins of Quillaja saponaria show strong aphicidal and deterrent activity against the pea aphid Acyrthosiphon pisum. *Pest management science*, *68*(2), pp.164-169.
https://onlinelibrary.wiley.com/doi/full/10.1002/ps.2235

De Geyter, E., **2012.** *Toxicity and mode of action of steroid and terpenoid secondary plant metabolites against economically important pest insects in agriculture* (Doctoral dissertation, Ghent University).
https://biblio.ugent.be/publication/2154465

Moggia, C., Graell, J., Lara, I., Schmeda-Hirschmann, G., Thomas-Valdés, S. and Lobos, G.A., **2016**. Fruit characteristics and cuticle triterpenes as related to postharvest quality of highbush blueberries. *Scientia Horticulturae*, *211*, pp.449-457.
https://www.sciencedirect.com/science/article/pii/S0304423816304599

Huang, F.Y., Chung, B.Y., Bentley, M.D. and Alford, A.R., **1995.** Colorado potato beetle antifeedants by simple modification of the birch bark triterpene betulin. *Journal of Agricultural and Food Chemistry*, *43*(9), pp.2513-2516.
https://pubs.acs.org/doi/pdf/10.1021/jf00057a035

Roy, A. and Saraf, S., **2006.** Limonoids: overview of significant bioactive triterpenes distributed in plants kingdom. *Biological and Pharmaceutical Bulletin*, *29*(2), pp.191-201.
https://www.jstage.jst.go.jp/article/bpb/29/2/29_2_191/_article/-char/en

Moiteiro, C., Marcelo Curto, M.J., Mohamed, N., Bailén, M., Martínez-Díaz, R. and González-Coloma, A., **2006.** Biovalorization of friedelane triterpenes derived from cork processing industry byproducts. *Journal of agricultural and food chemistry*, *54*(10), pp.3566-3571.
https://pubs.acs.org/doi/abs/10.1021/jf0531151

Schuster, A., Burghardt, M., Alfarhan, A., Bueno, A., Hedrich, R., Leide, J., Thomas, J., Riederer, M., **2016.** Effectiveness of cuticular transpiration barriers in a desert plant at controlling water loss at high temperatures. *AoB Plants,* 8(10).
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4925923/

Singh, B. and Kaur, A., **2018.** Control of insect pests in crop plants and stored food grains using plant saponins: a review. *LWT*, *87*, pp.93-101.
https://www.sciencedirect.com/science/article/pii/S0023643817306552