J|A|C|S
COMMUNICATIONS

Published on Web 04/17/2004

Genome Mining To Identify New Plant Triterpenoids

Gia C. Fazio, Ran Xu, and Seiichi P. T. Matsuda*

Department of Chemistry and Department of Biochemistry and Cell Biology, Rice University, Houston, Texas 77005

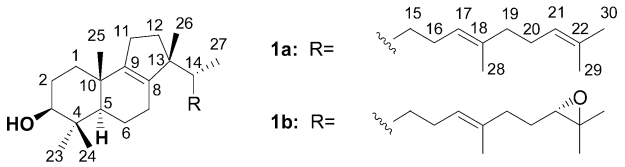
Received December 21, 2003; E-mail: matsuda@rice.edu

Plants are sessile organisms that cannot flee or fight to avoid being consumed. As part of their defense mechanisms, plants biosynthesize tens of thousands of known natural products, many of which have medicinally useful properties. The triterpenoids are a major group of plant natural products that includes sterols, triterpene saponins, and related structures. We describe herein experiments in which exploiting *Arabidopsis thaliana* genomic information uncovered two new triterpenoids. Genome mining can uncover compounds that have eluded classical methodologies.

The cyclic triterpenoid skeletons are produced by oxidosqualene cyclases.[1] The *Arabidopsis* genome contains 13 sequences homologous to known oxidosqualene cyclases.[2] We PCR-amplified from *Arabidopsis* cDNA[3] the 2301 bp putative coding sequence At5g48010, which is predicted to encode a protein 47% identical to the *Arabidopsis* cycloartenol synthase.[4] In preparation for in vitro experiments with heterologously expressed protein, the amplicon was subcloned into the yeast expression vector pRS426GAL.[5] The resultant construct pGCF4.0 was expressed in the yeast squalene epoxidase/lanosterol synthase double mutant RXY6. This strain is ideal for in vitro analysis of oxidosqualene cyclases because the lanosterol synthase mutation removes the native oxidosqualene cyclase and the squalene epoxidase mutation abolishes oxidosqualene formation, preventing the formation of At5g48010 metabolites in vivo.

A crude homogenate derived from a 5-mL culture of galactose-induced RXY6[pGCF4.0] was incubated with synthetic racemic oxidosqualene. TLC analysis revealed activity comparable to similarly expressed cycloartenol synthase. After reaction, the organic-soluble components were derivatized (1:1 bis(trimethylsilyl)trifluoroacetamide/pyridine). Because GC−MS revealed that only one component had the appropriate mass for the TMS ether of a $C_{30}H_{50}O$ triterpene alcohol (*m*/*z* 498), the protein encoded by At5g48010 converted oxidosqualene to a single product. Material for structure determination was generated using a yeast strain that was metabolically engineered to generate this compound in vivo. The yeast strain SMY8 lacks lanosterol synthase and consequently cannot cyclize oxidosqualene, causing this substrate to accumulate in vivo.[6] A 1-L culture of SMY8[pGCF4.0] induced with galactose was grown in synthetic complete medium lacking uracil and supplemented with heme and ergosterol. The cell pellet was saponified with KOH/ethanol, organics were extracted with hexane, and the nonsaponifiable lipid was purified by silica gel column chromatography, yielding ~1 mg of a colorless oil slightly less polar than lanosterol. HMBC, HSQC, COSYDEC, NOE, and 1D NMR experiments established the structure as (3*S*,13*S*,14*R*)-malabarica-8,17,21-trien-3-ol, to which we assigned the trivial name thalianol (**1a**, Chart 1). Thalianol synthase resembles cycloartenol synthase in that it generates a single major product cleanly. In contrast, the previously characterized *Arabidopsis* LUP1[7] and At1g78960[8] are multifunctional enzymes that convert oxidosqualene to mixtures of at least six and seven distinct triterpene alcohols, respectively. The good activity and accurate product formation of heterologously expressed thalianol synthase suggest that the enzyme is folded appropriately, and the enzyme would produce thalianol if expressed in *Arabidopsis* in the presence of substrate.

Chart 1. Tricyclic Triterpenoids Produced by At5g48010



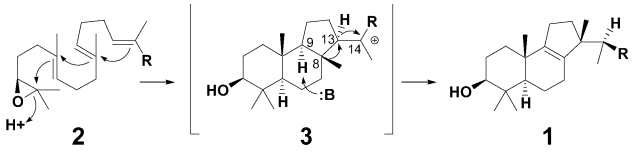
In addition to thalianol, SMY8[pGCF4.0] accumulates a more polar compound that exhibits mass spectra consistent with a $C_{30}H_{50}O_2$ triterpene alcohol. NMR experiments established the structure as a second novel compound, (3*S*,13*S*,14*R*,21*S*)-21(22)-epoxy-malabarica-8,17-dien-3-ol (**1b**). Triterpene epoxide **1b** would arise from thalianol synthase-mediated cyclization of (3*S*,22*S*)-2(3),-22(23)-bis-oxidosqualene, which accumulates in yeast lanosterol synthase mutants.[9] When cyclase activities are limiting, oxidosqualene is cyclized inefficiently and it can re-enter squalene epoxidase and be oxidized on the distal terminus. Epoxide **1b** would be formed in *Arabidopsis* if oxidosqualene levels increase because local epoxidase levels surpass the ability of cyclases to consume oxidosqualene. Lanosterol synthase, cycloartenol synthase, and onocerin synthase are related oxidosqualene cyclases that also cyclize bis-oxidosqualene normally, leaving the distal epoxide unchanged.[9,10]

Both compounds **1a** and **1b** are novel structures in the malabaricane series. A probable biosynthetic mechanism for **1a** is shown in Scheme 1. Protonation of oxidosqualene (**2a**) initiates carbocation formation and subsequent cyclization to **3a**, a 6,6,5, chair−chair, tricyclic carbocationic intermediate with B-ring stereochemistry reminiscent of the dammarenyl cation. A hydride shift from C13 to C14, a β-methyl shift from C8 to C14, and C9 deprototonation yields thalianol (**1a**). The C13 epimer of **3a** is also a conceivable intermediate, but would mandate a syn shift of C26 from C8 to C13. Bis-epoxide **2b** undergoes the analogous reaction to yield thalianol epoxide (**1b**).

Thalianol synthase is the first characterized member of the *PEN* gene family, which are sequences named for their similarity to oxidosqualene cyclases that generate pentacyclic triterpenes.[2] Thalianol synthase is 46−49% identical to plant cycloartenol synthases, 48−58% identical to dicot oxidosqualene cyclases that generate pentacyclic triterpene alcohols,[2,11] and 42% identical to an oat β-amyrin synthase.[12] A phylogenetic tree including characterized plant OSCs and predicted *Arabidopsis* OSCs establishes that thalianol synthase is evolutionarily more similar to enzymes that form pentacyclic nonsteroidal triterpenoid skeletons than to cycloartenol synthases,[2] and thalianol synthase shares conservation patterns with these enzymes at all known catalytically relevant positions. Tyr410 and Ile481 are conserved cycloartenol synthase residues, and mutagenesis experiments with *Arabidopsis* cycloartenol synthase have established that these residues guide

10.1021/ja0318784 CCC: $27.50

Scheme 1. Cyclization of Oxidosqualene to Thalianol



catalytic events in the B ring.[5,13] Thalianol synthase has Phe and Val at corresponding positions, like enzymes that form pentacyclic nonsteroidal triterpene alcohols. This similarity may reflect that common catalytic forces are used to promote formation of a chair conformation in ring B. Like all other known oxidosqualene cyclases, thalianol synthase maintains the active-site acid corresponding to the *Arabidopsis* cycloartenol synthase Asp481 residue.

This discovery of thalianol biosynthesis encourages further search for natural products by genome mining. Malabaricane derivatives are rare; the Dictionary of Natural Products Database http://www.chemnetbase.com/scripts/dnpweb.exe contains only 66 6,6,5-triterpenoids, almost all of which are oxygenated derivatives of compounds that have cyclized but are not rearranged. Their phylogenetic distribution is unusual; these compounds have been found in relatively few species but in most eukaryotic kingdoms. Tricyclic triterpenoids are most numerous in the animals. Three sponges[14] accumulate 49 different 6,6,5 triterpenoids. Toadstools in the fungal genus *Cortinarius* accumulate six unrearranged malabaricanes with antibacterial and antitumor activities.[15] Ferns generate the greatest diversity of cyclization products, with three structures that are structurally consistent with being direct cyclization products. These include an unrearranged 13α malabaricatriene from *Polypodium polypodioides*, a pair of C13 epimers from *Lemmaphyllum microphyllum*, and two rearranged Δ^7 and Δ^8 structures from *Polypodiodes niponica*.[16] These lack oxidation at C-3 and are presumably squalene rather than oxidosqualene cyclization products. The Δ^8 tricycle is a 3-deoxy analogue of thalianol that has differences in chemical shifts of the C27 protons, suggesting that its C14 stereocenter (which was not determined in the 1989 study) is epimeric to that in thalianol. Although dicotyledonous plants are prolific producers of nonsteroidal triterpenoids, they have yielded only 12 malabaricanes, none of which have rearranged rings. These compounds have been found in only five species representing four dicotyledonous orders (Asterales, Fabales, Sapindales, and Violales).[17] The presence of thalianol synthase in *Arabidopsis* establishes that malabaricane biosynthetic ability also exists in Brassicales. Tricyclic triterpenoid biosynthesis probably did not arise independently in these five orders, but may have evolved in a common ancestor of these orders, and malabaricanes are probably more broadly distributed than isolation patterns indicate. *Arabidopsis* and many other dicots may accumulate only modest amounts of malabaricanes or produce them only in response to specific stimuli.

The earliest efforts at natural product isolation preceded chromatography and were consequently limited to purifying dominant components that could be readily obtained by distillation or crystallization. Dramatic improvements in methods to fractionate extracts and to determine the structures of trace components have provided a wealth of natural products, many of which have important biological activities. However, natural product biosynthesis is heavily influenced by external stimuli such as infection or insect herbivory, and extracts of uninduced plants consequently contain only a subset of the products that these organisms can biosynthesize. The compounds that can be isolated from a native source are consequently not a comprehensive accounting of the organism's biosynthetic capacity, but reflect the state of the tissues upon harvest.

Arabidopsis is generally considered to be a metabolically simple plant; it generates interesting glucosinolates but is not a notable producer of natural products. Although the *Arabidopsis* genome encodes extensive capability to generate triterpene rings systems, the repertoire of known *Arabidopsis* triterpenoids is limited by low-level native production and limitations in classical isolation methodology. Plants apparently have much more extensive natural product capability than can be ascertained by classical isolation methods, but only produce these compounds when and where they are needed. We suggest that heterologous expression of enzymes uncovered by genomic sequencing provides a uniquely systematic and comprehensive approach to identifying the natural products generated by an organism.

Acknowledgment. We are indebted to Dr. William K. Wilson for numerous contributions, including determining the structures of **1a** and **1b**. We thank Ying Liu for subcloning At5g48010. G.C.F is a Robert A. Welch Fellow. We are grateful to the National Science Foundation (MCB-0209769), the Robert A. Welch Foundation (C-1323), and the Herman Frasch Foundation for financial support.

Supporting Information Available: Details of cDNA cloning, expression, product isolation, NMR signal assignments, and GC−MS and NMR spectra of **1a** and **1b** (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Xu, R.; Fazio, G. C.; Matsuda, S. P. T. *Phytochemistry* **2004**, *65*, 261−290.
(2) Husselstein-Muller, T.; Schaller, H.; Benveniste, P. *Plant Mol. Biol.* **2001**, *45*, 75−92.
(3) Minet, M.; Dufour, M.-E.; Lacroute, F. *Plant J.* **1992**, *2*, 417−422.
(4) Corey, E. J.; Matsuda, S. P. T.; Bartel, B. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 11628−11632.
(5) Hart, E. A.; Hua, L.; Darr, L. B.; Wilson, W. K.; Pang, J.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **1999**, *121*, 9887−9888.
(6) Corey, E. J.; Matsuda, S. P. T.; Baker, C. H.; Ting, A. Y.; Cheng, H. *Biochem. Biophys. Res. Commun.* **1996**, *219*, 327−331.
(7) (a) Herrera, J. B. R.; Bartel, B.; Wilson, W. K.; Matsuda, S. P. T. *Phytochemistry* **1998**, *49*, 1905−1911. (b) Segura, M. J. R.; Meyer, M. M.; Matsuda, S. P. T. *Org. Lett.* **2000**, *2*, 2257−2259.
(8) Kushiro, T.; Shibuya, M.; Masuda, K.; Ebizuka, Y. *Tetrahedron Lett.* **2000**, *41*, 7705−7710.
(9) Field, R. B.; Holmlund, C. E. *Arch. Biochem. Biophys.* **1977**, *180*, 465−471.
(10) (a) Corey, E. J.; Gross, S. K. *J. Am. Chem. Soc.* **1967**, *89*, 4561−4562. (b) Heintz, R.; Schaefer, P. C.; Benveniste, P. *J. Chem. Soc., Chem. Commun.* **1970**, 946−947. (c) Rowan, M. G.; Dean, P. D. G.; Goodwin, T. W. *FEBS Lett.* **1971**, *12*, 229−232.
(11) Ebizuka, Y.; Katsube, Y.; Tsutsumi, T.; Kushiro, T.; Shibuya, M. *Pure Appl. Chem.* **2003**, *75*, 369−374.
(12) Haralampidis, K.; Bryan, G.; Qi, X.; Papadopoulou, K.; Bakht, S.; Melton, R.; Osbourn, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 13431−13436.
(13) Herrera, J. B. R.; Wilson, W. K.; Matsuda, S. P. T. *J. Am. Chem. Soc.* **2000**, *122*, 6765−6766.
(14) (a) Ravi, B. N. W.; Robert, J.; Croft, K. D. *J. Org. Chem.* **1981**, *46*, 1998−2001. (b) Ravi, B. N.; Wells, R. J. *Aust. J. Chem.* **1982**, *35*, 39−50. (c) Zhang, W.-H.; Che, C.-T. *J. Nat. Prod.* **2001**, *64*, 1489−1492.
(15) Sontag, B.; Froede, R.; Bross, M.; Steglich, W. *Eur. J. Org. Chem.* **1999**, *255*, 5−260.
(16) (a) Ageta, H.; Arai, Y. *J. Nat. Prod.* **1990**, *53*, 325−332. (b) Masuda, K.; Shiojima, K.; Ageta, H. *Chem. Pharm. Bull.* **1983**, *31*, 2530−2533. (c) Arai, Y.; Hirohara, M.; Ageta, H. *Tetrahedron Lett.* **1989**, *30*, 7209−7212.
(17) (a) Jakupovic, J.; Eid, F.; Bohlmann, F.; El-Dahmy, S. *Phytochemistry* **1987**, *26*, 1536−1538. (b) Faini, F.; Castillo, M.; Torres, R.; Delle Monache, G.; Gacs-Baitz, E. *Phytochemistry* **1995**, *40*, 885−890. (c) Chawla, A.; Dev, S. *Tetrahedron Lett.* **1967**, 4837−4843. (d) Marner, F. J.; Freyer, A.; Lex, J. *Phytochemistry* **1991**, *30*, 3709−3712. (e) Ziegler, H. L.; Strk, D.; Christensen, J.; Olsen, C. E.; Sittie, A. A.; Jaroszewski, J. W. *J. Nat. Prod.* **2002**, *65*, 1764−1768.

JA0318784