



Published in final edited form as:

J Nat Prod. 2012 March 23; 75(3): 432–443. doi:10.1021/np200878s.

Dereplication, Residual Complexity and Rational Naming – the Case of the *Actaea* Triterpenes [⊥]

Feng Qiu[†], Ayano Imai[†], James B. McAlpine[†], David C. Lankin[†], Ian Burton[†], Tobias Karakach[†], Norman R. Farnsworth^{†,§}, Shao-Nong Chen[†], and Guido F. Pauli^{†,*}

[†]Department of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, University of Illinois at Chicago, Chicago, IL 60612, U.S.A

[‡]Institute for Marine Biosciences, National Research Council, Halifax, Nova Scotia B3H 3Z1, Canada

Abstract

The genus *Actaea* (including *Cimicifuga*) has been the source of ~200 cycloartane triterpenes. While they are major bioactive constituents of complementary and alternative medicines, their structural similarity is a major dereplication problem. Moreover, their trivial names seldom indicate the actual structure. This project develops two new tools for *Actaea* triterpenes that enable rapid dereplication of more than 150 known triterpenes and facilitates elucidation of new compounds. A predictive computational model based on classification binary trees (CBTs) allows *in silico* determination of the aglycone type. This tool utilizes the Me ¹H NMR chemical shifts and has potential to be applicable to other natural products. *Actaea* triterpene dereplication is supported by a new systematic naming scheme. A combination of CBTs, ¹H NMR deconvolution, characteristic ¹H NMR signals, and quantitative ¹H NMR (qHNMR) led to the unambiguous identification of minor constituents in residually complex triterpene samples. Utilizing a 1.7 mm cryo-microprobe at 700 MHz, qHNMR enabled characterization of residual complexity at the 10–20 μg level in a 1–5 mg sample. The identification of five co-occurring minor constituents, belonging to four different triterpene skeleton types, in a repeatedly purified natural product emphasizes the critical need for the evaluation of residual complexity of reference materials, especially when used for biological assessment.

[⊥]Dedicated to Dr. Gordon M. Cragg for his pioneering work on the development of natural product anticancer agents.

*Corresponding Author, Tel: +1 (312) 355-1949. Fax: +1 (312) 355-2693. gfp@uic.edu.

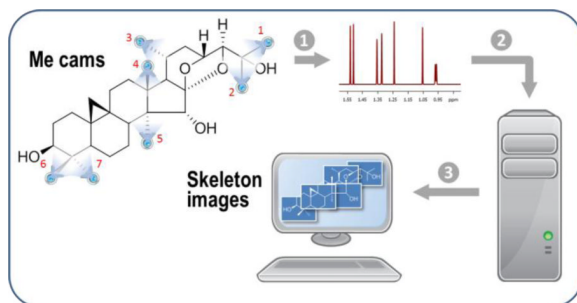
§Deceased on September 10, 2011.

ASSOCIATED CONTENT

Supporting Information

Contains ¹H NMR spectra of the samples E1/E2 and identification of sugar moieties by characteristic signals, screenshots of ActaPredict and ActaMatch, and the complete list of *Actaea* triterpenes in our in-house database. This material is available free of charge via the Internet at <http://pubs.acs.org>.

In addition, the authors are open to performing dereplication runs on an individual basis for interested readers. For instructions on how to submit the required information and ¹H NMR data, visit <http://tiger.uic.edu/~gfp/content/dereplicate.htm> or contact the author of correspondence.



INTRODUCTION

The rapid identification of known natural products, a process known as dereplication, is important for targeting isolation of bioactive compounds of interest from natural resources. The most common dereplication methods, LC-UV and LC-MS, provide limited structural information and require calibration with authentic standards, respectively, in particular for stereochemical assignments and quantification. NMR techniques have been increasingly used in dereplication of natural products.¹ However, full interpretation of ¹H NMR spectra is still challenging, especially when working with components of complex mixtures. In order to overcome these limitations, a few rapid *in silico* dereplication tools based on database searches have been developed.^{2–6} One recent example is the excellent AntiMarin database built by Drs. John W. Blunt and Murray H. G. Munro.⁶ This database features a search function based on molecular weight and exact counts of the number of methyl (Me), methylene, and methine groups. Dereplication of a given compound is performed by searching the most compatible hit(s) limited by these criteria. The success of dereplication is solely dependent of availability of exact or similar compounds in the database. However, this kind of database search lacks an intelligent program which recognizes the query compound and determines/predicts its compound class and partial or even full structures based on a general rule, e.g., the unique pattern of NMR data.

We were particularly interested in triterpenes from various species of *Actaea*. The previously classified genera *Cimicifuga* and *Souliea*, now reclassified as part of the genus *Actaea*,⁷ have been the source of almost 200 triterpenes possessing the cycloartane skeleton.⁸ This creates a major dereplication problem for both new and known members of this compound class from these plants, which are a major source of herbal medicines and have been associated with numerous biological activities.⁸

Moreover, almost all *Actaea* triterpenes have been accorded trivial names, to a large extent derived from the Latin binomial or common names associated with the source plant. These names, at best, provide clues as to the origin of the compound, but seldom have any indication of the actual structure to the non-cognoscenti, and certainly do not help the scientist in the search for novel triterpenes. A non-comprehensive list of these names includes:⁸ acerinol, acerionol, acteol, bugbanoside, cimiaceroside, cimicidanol, cimicidol, cimicifoetiside, cimicifol, cimicifugenol, cimicifugoside, cimifoetiside, cimifoside, cimigenol, cimigol, cimilactone, cimiracemoside, cimiside, dahurinol, foetidinol, heracleiforinol, and shengmanol. The names are also not practical, even for the specialist, because the similarity of names gives no indication of similarity of structure. One example is the cimiracemosides A, M, and P, which have completely different ring systems and differ in the sites of oxygenation at C-12, C-15, C-16, C-21, C-23, and C-26. Another illustration is reflected by the fact that most of the triterpenes from this genus occur as glycosides, usually at the C-3-oxygen, and for the most part these are named with the suffix “-oside”, whereas the aglycones have the suffix “-ol”. However, even this simple convention is not

universally followed as cimicidol, cimicifol, and acteol are all glycosides, whereas acerinol and heracleiforinol, although being alcohols, no longer have that functional group at C-3. Another inconsistency is the original name of hydroshengmanol vs. the subsequent use of hydroxyshengmanol, where both should be considered misnomers of a tautomeric shengmanol.

Here we propose a new rational naming system, and introduce a novel dereplication system based on a binary classification of only the Me signals in the ^1H NMR spectra of these complex molecules. The naming system will simplify the deduction of all known *Actaea* triterpene structures as well as congeners yet to be discovered, given the knowledge of only the cycloartane skeleton. In reclassifying *Cimicifuga* within the genus *Actaea*, botanists have given the chemists an opportunity to systematize this nomenclature, which adds to the aforementioned reasoning for the proposed new naming scheme. The use of *Actaea* as the basis for the new naming system is further justified by the recent discovery of several very closely related triterpenes from *Actaea vaginata* (previously *Souliea vaginata*), a species never classified as a *Cimicifuga*.^{9–11}

The novel dereplication system relies on the fact that most of these compounds have five to seven skeletal Me groups serving as the “surveillance units” (“Me cams”) for their neighboring segments of the molecules. Therefore, their full structures can be mapped by combining all “surveillance images” provided by each of the Me groups as “surveillance units” (Figure 1). In fact, the history of using only Me groups in the determination of structures initially dates back to the late 1950s and into the 1960s for steroids^{12,13} and triterpenes.^{14–17} These studies analyzed the additive intramolecular shielding or deshielding effects of proximate substituents on the chemical shifts of the Me groups, and as a result, the substitution pattern of the substituents could be deduced and the structures of triterpenes could be elucidated using this approach. Two more recent studies used this approach for the structural elucidation of cardiac glycosides^{18,19} and unsaturated C₂₇ sterols.²⁰ We hypothesize that the relationship between the Me shifts and structural characteristics is statistically correlated, and that correlation can be further integrated into a pattern recognition model for structural determination. Starting from this hypothesis, we aimed to establish a novel methodology which uses classification binary trees (CBTs)^{21–23} for a rapid and automatic structural dereplication. In the present study, an in-house database currently containing the Me shifts of more than 170 *Actaea* triterpenes was assembled and a search function based on the Pearson’s Coefficient (r) developed. Using the Me shifts in the database as the training set, the CBTs for the classification of *Actaea* triterpenes were generated by classification and regression tree (CART)²⁴ analysis. All of these triterpenes are listed with structures in Supporting Information S6 and with Me shifts in S7. A detailed explanation of their common stereochemical properties and the depictions used or omitted, both in the present text and in the Supporting Information, is given at the preface to S6.

Furthermore, triterpenes are a good example to demonstrate an important signature of natural products, which is that certain levels of characteristic impurity patterns, referred to as residual complexity, remain visible along the entire (bio-)analytical pathway.²⁵ The term, residual complexity, refers to the easily overlooked impurity profile of isolated natural products, which may exert a significant influence on their accurate biological assessment.^{25–27} Residual complexity can be static or dynamic, referring to impurity patterns that are either constant or fluctuating depending on conditions, respectively. Therefore, it is important to both qualitatively and quantitatively characterize the impurity profile of isolated natural products. In the present study, the classification models and database search were utilized as two *in silico* tools to dereplicate *Actaea* triterpenes in residually complex samples of purified reference materials.

RESULTS AND DISCUSSION

The New Naming System

All of the known cycloartane triterpenes from *Actaea* fall into only a few basic structural skeletons. As far as C-20 to C-27 are concerned, there are acyclic compounds in which these carbons have no connections between themselves other than the basic carbon chain, as shown in the Chart. Then there are other compounds in which some of these carbons are involved in one or two rings, usually formed by ether or acetal oxygens, often back to C-16. The new system then would name the acyclic aglycone compounds as *actanols*, those with a single oxygen bridge forming a further ring as *actamonoxols*, and those with two oxygen-containing rings *actabinoxols*. These names all include the 3β -hydroxy group. Where this group is part of a glycosidic linkage, the suffix would be “-oside”, e.g., actabinoside. All of the substituents and other structural modifications need to be affixed using standard chemical nomenclature, with prefixes arranged in alphabetical order.

The stereochemistry of these triterpenes would be designated via the Cahn-Ingold-Prelog (CIP) system rather than the simpler α/β system commonly used in steroid nomenclature, because the α/β nomenclature fails in bicyclic caged rings that occur in many of the actabinoxols. This problem has been faced by partial use of both systems, however the CIP system works universally, and we are advocating its use in all cases except for glycosidic linkages, where the α/β and *D/L* system for sugars is well accepted and fully understood. The CIP system does have a disadvantage in that the stereodesignation of a specific stereocenter can change without a requisite change in the configuration, but rather by changes in CIP-preferences of nearby substituents, and hence a change in the precedence number of substituents to that carbon (Supporting Information S1).

Although such a naming scheme will occasionally result in reasonably long names, they will be readily understood for all *Actaea* triterpenes by organic chemists with nothing further than the basic knowledge of the attached Chart. Some representatives of *Actaea* triterpenes are listed in Table 1 and their structures are given in Chart 2. The first five skeletons cover more than 90% of the known triterpenes from *Actaea*. There are, however, a limited number of compounds that do not share these basic structures. Some compounds, which are missing carbons at the end of the chain, are readily accommodated by the “nor” prefix, and there are a couple of types where carbon-carbon bonds are cleaved and use of the “seco” prefix is required.

Characteristics of the Collected Data

The data sets of ^1H NMR spectra, predominantly measured in pyridine-*d*₅, for ~170 initially included *Actaea* triterpenes, representing 75% of all *Actaea* triterpenes found in SciFinder, were collected in our in-house database (Supporting Information S6 and S7). The major types of compounds are (both trivial and new names given; see also Table 2): cimigenols (CG, 50, 33%; acta-16,23;16,24-binoxols), acteols (AT, 16, 11%; acta-16,23;23,26-binoxols), hydroshengmanols (HS, 16, 11%; acta-16,23-monoxols), cimircemosides (CR, 12, 8%; acta-16,23;22,25-binoxols), 23-*O*-acetylshengmanols (AS, 11, 7%; 16-oxo-actanols), cimicidanol (CA, 7, 5%; 16,23-dioxo-actanols), dahurinols (DA, 5, 3%; acta-16,23-monoxols), foetidinol (FO, 5, 3%; acta-16,23-carbamonols), and cimicidols (CO, 4, 3%; 16,23-dioxo-actanols). Some rare sub-types, such as 15,16-secocimicidols (SE), alkaloids, cimicifugenols, cimilactones, and heracleiforinols comprising one or two known compounds, were also included in the database.

Canonical Discriminant Analysis

A canonical discriminant analysis (CDA) was initially performed for all compounds in the database using their Me δ_{H} values (Me1, Me2, etc.) in ascending order. All the Me groups with $\delta_{\text{H}} < 1.90$ were used in this analysis as Me signals with $\delta_{\text{H}} > 1.90$ are either acetyl (OAc) or methoxy groups (OMe) which are not essential base structural components for *Actaea* triterpenes. In order to create a dimensionally homogenous data set for CDA analysis, compounds with less than seven analyzed Me groups (δ_{H} 0.70–1.90) are given extra variable(s) with value 0. The result for all compounds is visualized in Figure 2, representing a 3D CDA plot. The first factor (CDA-1) represents 77.9% variation in the original data, whereas CDA-2 and CDA-3 account for 17.2 and 2.3% variation, respectively. All three factors explain a total of 97.4% of cumulative variance in a highly significant analysis (Wilks' $\lambda = 0.00$, $F_{\text{approx}} = 35.16$, $df_1 = 84$, $df_2 = 718$, $P < 0.0001$). The majority of triterpenes (~80%) which have seven Me groups ($\delta_{\text{H}} < 1.90$) are clustered in a space shown in Figure 2B and their classification results are listed in Table 3. Relying only on the variances of the Me shifts, all the *Actaea* triterpenes in the database can be classified with an overall correct rate of 86.9% by the model derived from CDA analysis. Considering inescapable variations of reported ^1H chemical shift information due to inconsistencies in, e.g., temperature and calibration (TMS vs. residual solvent), the discriminative power of the model could be further improved in the future by using a standardized NMR acquisition protocol.

Development of the Classification Binary Trees

CART is a machine learning technique ideal for large and unbalanced data sets with many descriptors.²⁸ It generates a tree-like graph or model as a binary-decision support tool to identify the origin or class of the samples.²⁴ In order to build a more accurate classification model for *Actaea* triterpenes, the classification binary trees (CBTs) were developed by CART analysis and used to partition the compounds into structurally similar clusters of aglycones. The compounds in the database were initially divided into three subgroups according to the number of Me groups (five, six, or seven; $\delta_{\text{H}} < 1.90$) within the molecules. Figure 3 shows the resulting CBT from CART analysis to classify *Actaea* triterpenes with five (CBT-1) or six Me groups (CBT-2) by using their Me shifts. Both CBTs consist of three terminal nodes (leaves) and two non-terminal nodes. From the top (root) of the tree, the compounds were split into groups according to the Me shifts used as descriptors at each node. Using these two CBTs, all of the triterpenes with five or six Me groups in the database were correctly classified. Similarly, the CBT for the classification of *Actaea* triterpenes with seven Me groups (CBT-3) is depicted in Figure 4, which is characterized by 14 terminal nodes and 13 non-terminal nodes, with an overall success rate of 94.4%. The percentage of correct classification for each structural sub-type is shown in Table 4.

In CART analysis, variable importance is usually determined by looking at every node in which a variable appears and taking into account its suitability as a splitter.²⁸ The importance score of the variables used in the generation of CBT-3 was calculated using the Salford Predictive Miner as follows: Me7 100.00, Me6 98.44, Me4 68.19, Me1 64.78, Me5 62.80, Me3 57.30, Me2 50.71, and Me8 29.20. These scores reflect the contribution of each Me signal to the classification of *Actaea* triterpenes, with the contribution stemming from all the variables' roles as primary splitters and as surrogates to any of the primary splitters. Here, Me7 and Me6 are ranked as the two most important. More than 75% of the Me7 and Me6 protons are assigned to either H-26 or H-27. Both of these Me groups are located in the aglycone side chain, which often cyclizes with C-16, and, thus, are highly indicative of the major structural differences of *Actaea* triterpenes. This explains why Me7 and Me6 are important indicators of the aglycone type. By further looking at the resulting CBT, Me7 is found to be the major classifier for cimiracemosides (node #11 and #13) and cimigenols

(node #10) with an OAc group at C-25. This is highly consistent with their structural characteristics. Under the neighboring effect of a 25-epoxy function, both H-26 and H-27 signals of cimracemosides shift downfield to $\delta_{\text{H}} > 1.75$. The acetylation of the OH group at C-25 has been seen only in cimigenols, which results in H-26 shifting to the range of δ_{H} 1.59–1.75.

The variable Me4 can be used to classify 15,16-secocimcidols: their Me4 protons are either H-18 or H-27 with apparent $\delta_{\text{H}} > 1.50$. Me1 is ranked as the fourth most important classifier, covering ~50% of the investigated compounds which are cimigenols (node #3 and #5), cimicidols (node #12), cimicidanols (node #8), 23-*O*-acetylshengmanols (node #16), and hydroshengmanols (node #6). Consistently, the protons of their Me1s are all H-21, which is also a Me group in the side chain of the aglycone. While any of the first four important variables cannot distinguish dahurinols (node #1) and cimigenols (node #2), Me5 works well to differentiate these two types of compounds. The two Me groups, Me3 and Me2, show much less importance because the underlying protons are the geminal Me groups at C-4 (H-29 and H-30), which are located in the least structurally diverse region of the aglycones. However, the differences in their chemical shifts, regardless of their assignment, are also useful to split the compounds into subgroups which can be further classified using other discriminating Me groups.

Interestingly, by using CBT-3, in the terminal nodes #2, 3, 5, 8, and 10, the 40 known cimigenols are partitioned into five subgroups, and each group has its own structural characteristics. All thirteen cimigenols in node #10 have an OAc group at C-25. Four of the seven cimigenols in node #1 have an OAc at C-25 and an extra OAc within the sugar moiety. Cimigenols in node #5 either have an OAc at C-12 or OMe at C-25. As a matter of fact, it is easy to distinguish OAc and OMe according to the Me chemical shift. The signals for OAc are usually observed at 2.0 ± 0.2 ppm, while OMe groups resonate at 3.2 ± 0.2 ppm. In addition, six cimigenols classified in node #8 have an OH group at C-12. However, all 17 cimigenols in node #3 are free of any OAc or OMe within the aglycone. These results indicate that the presence and position of OH, OAc, and OMe in cimigenols may also be determined solely based on the Me shifts.

In addition to the dereplication capability, the CBT models have potential to predict the aglycone type of the unknown *Actaea* triterpenes yet to be discovered. Owing to limited data available, leave-one-out cross-validation (LOOCV) was used to estimate the accuracy of the predictions. Overall, the predictions are 80.4% correct for the CBT-3. As summarized in Table 5, cimigenols, cimracemosides, and hydroshengmanols, which comprise the majority of compounds in the database, have excellent prediction rate of 80.0, 91.7, and 100%, respectively. For 23-*O*-acetylshengmanols, three of 11 (72.7% correct) are incorrectly predicted as cimicidanols. Both these two types have the same epoxide group at C-24 and C-25, leading to somewhat difficulty in differentiating them by the terminal Me groups in the side chain. The minority of compounds, including cimicidanols, cimicidols, and dahurinols, are 50–60% correctly predicted. Despite their structural similarity with other types of *Actaea* triterpenes, using more descriptors, e.g., multiplicity of the Me groups and chemical shift of cyclopropane methylene (H-19a/b), and/or more data sets if available in the future expectedly improve their accuracy of predictions.

Dereplication of *Actaea* Triterpenes in Residually Complex Mixtures

Traditionally, the complexity of natural product mixtures makes it a challenge to identify the components by full interpretation of NMR spectra. In the present study, the concept of using only Me shifts for the dereplication of multicomponent mixtures, such as residually complex (impure) mixtures of triterpenes, has the particular advantage that Me resonance are usually singlets of relatively high intensity. While the Me groups resonate in the same range of 0.8–

2.0 ppm as several aliphatic methines and methylenes, the signals of the latter are much more complex and their intensities distribute over a much broader range due to J coupling. In approximation, comparing a *ddd* methylene (1H) with a singlet Me (3H) signal, the individual spectral lines of the former are ~25 fold lower in intensity. Accordingly, impurities of more than 4% become visible even in overlapped regions of the spectra. While chemical shift dispersion limits the number of components for which all Me signals can be identified, 2D NMR methods prove useful in the further unraveling of this “hidden” spectroscopic information.

This study establishes an *in silico* dereplication approach to identify *Actaea* triterpenes in both pure forms and residually complex mixtures by using a combination of the CBTs and database search, using three steps (Figure 5). *Step 1 [ActaPredict]*: The Me shifts are analyzed by the CBTs, and the triterpene skeleton is determined; the substituents on the skeleton as well as the sugar moieties are identified by the presence of characteristic ^1H NMR signals. *Step 2 [ActaMatch]*: The ^1H NMR data of Me groups are also used to search the hits with $r > r_0$, where r_0 is the threshold of similarity defined by the user; according to our experience, exact hits have r values > 0.998 . *Step 3*: The results from steps 1 and 2 are compared for consistency. Both steps 1 and 2 are programmed and automated within a spreadsheet and incorporated into an application suite named “ActaFinder”.

Two examples (E1, E2) of residually complex triterpene reference materials were chosen to illustrate this approach. Both materials resulted from a multistep fractionation of EtOAc partition of *Actaea racemosa* (black cohosh) crude extracts, using VLC and MPLC. The sample E1 (2.4 mg) was initially subjected to ^1H NMR analysis using the conditions stated in the Experimental section. Each *Actaea* triterpene gives rise to a pair of doublets in the range of δ_{H} 0.2–1.0, due to its cyclopropane methylene protons, H-19a/b. Based on these characteristic signals, it is known that this sample contained two major triterpenes denoted by **10** and **11**, respectively (Figure 6). The Me signals of each triterpene were readily recognized based on their integral values relative to the individual H-19 signals. The overlapped Me signals (Me2 of **10** and Me1 of **11** at 1.08 ppm) were deconvoluted by using the Line Fitting function in the MestReNova software, and the individual spectra of the two triterpenes extracted from the ^1H NMR spectrum of the mixture.

Compounds **10** and **11** have seven and six Me groups with $\delta_{\text{H}} < 1.90$, respectively. By using the CBT-3 partitioning in ActaPredict, **10** was dereplicated as an acta-16,23;22,25-binoxol, formerly often designated as cimracemoside. An additional Me signal at 2.14 ppm indicates that an OAc may be present at C-12, a position which is commonly acetylated in actabinoxols. Its H-19a signal was observed at 0.98 ppm, indicating the presence of a $\Delta^{7,8}$ -double bond. Close inspection of the region for the sugar moieties (3.7–5.0 ppm) identified a characteristic *dd* signal (11.9, 1.4 Hz, H-5' b) of arabinopyranose (*arap*) at 3.790 ppm, bearing the same integral as H-19b of **10** (Supporting Information S3). Compound **11** was dereplicated as the xylopyranoside of a 21-hydroxylated acta-16,23;16,24-binoxol, formerly classified as 21-hydroxycimigenol, by CBT-2 partitioning. This was substantiated by the lack of a doublet among the Me signals due to hydroxylation of the Me at C-21. A characteristic *dd* signal (11.2, 9.8 Hz, H-5' b) of xylopyranose (*xylp*) was observed at 3.755 ppm, exhibiting the same integral as H-19a of **11**. Summarizing all the dereplication results and further observations, the structures of **10** and **11** were identified as shown in Figure 6. In addition, because the ^1H NMR spectra were acquired under quantitative conditions (qHNMR), their molar ratio was determined to be 70:30 from the integrals of their H-19a/b signals.

Similarly, the residually complex sample E2 (4.1 mg) was used for *in silico* dereplication. The H-19a/b signals in the ^1H NMR spectrum indicated that its composition is more

complicated, with at least six minor triterpenes being present along with the main component, **12**. Initial identification targeted the major constituents **12**, **13**, and **14**, which had content of more than 10 mol% and allowed full Me deconvolution: Based on the integral of their H-19b signals, individual Me signals were identified and extracted from the ^1H NMR spectrum of E2 by deconvolution of the overlapped peaks. Compound **12** was dereplicated as an acta-16,23;16,24-binoxol (formerly: cimigenol) in node #5 of the CBT-3. A Me signal at 2.14 ppm further indicated that **12** is acetylated at C-12. Compound **13** was dereplicated as a 24,25-epoxy derivative of an acta-16,23;23,26-binoxol (formerly: acteol) with an OAc (2.15 ppm) at C-12. Compound **14** was dereplicated as a 23-acetate of a 16-oxo-actanol (formerly: 23-*O*-acetylshengmanol). The fact that none of the H-19a signals is shifted downfield to ~ 1.00 ppm indicates that all three triterpenes are saturated at C-7 and C-8. A characteristic *dd* signal (11.9, 1.4 Hz, H-5' b) of arabinopyranose (*arap*) was observed at 3.832 ppm, exhibiting the same integral as H-19a of **13** (Supporting Information S4). Two overlapped *dd* signals were observed at 3.730 and 3.744 ppm, and both are characteristic for H-5' b of *xylp*. Their integrals were identical with those of H-19b of **12** and **14**, respectively. Therefore, the structures of **12**, **13**, and **14** were identified as shown in Figure 7. Their molar ratio was measured by qHNMR as 62:22:16 based on the integral of their H-19b signals. While the aforementioned general considerations put the threshold of identifiable Me signals (vs. overlapping CH_2/CH protons) around the 5% level, we were still able to tentatively identify *R*- and *S*-actein [(12*R*)-12-acetoxy-(24*R*,25*S*)-24,25-epoxy-(26*R*&*S*)-26-hydroxy-3-*O*- β -*D*-xylopyranosylacta-(16*S*,23*R*)-16,23;23,26-binoxoside in the new naming scheme] as two minor constituents of E2 (~ 3 and $\sim 5\%$ impurities, respectively). Evidence for this assignment came from the CBT analysis and characteristic $^2/3\text{JHMBC}$ cross peaks between the small Me-28 signals at 0.87 and 0.80 ppm and bridgehead carbons C-8/13/14, which all resonate in the narrow range ~ 44 – 46 ppm. This demonstrates the power of cryo-microprobe NMR analysis of residually complex natural products.

In order to verify the dereplication results by the CBTs, the ^1H NMR data of the Me groups of individual triterpenes identified in the mixture samples E1 and E2 were searched by ActaMatch. The results are shown in Table 6 and indicate that all the investigated compounds are highly correlated with their best hits ($r > 0.998$). It is noteworthy that the triterpenes with the same aglycone but different sugar moieties may exhibit a high correlation with $r > 0.999$ in terms of the ^1H NMR properties of Me groups. For example, adding to cimracemoside G, two hits, both of which are cimracemoside F data from two different sources, matched to compound **10** with a high r value of 0.9994 and 0.9997, respectively (Supporting Information S5). For cimracemosides F and G, the different sugar moieties *xylp* vs. *arap* have only a negligible chemical shift effect on the H-29 and H-30 Me groups, which results in a minor difference in the r value. However, inconsistencies in the NMR experimental conditions of reported data may also contribute to this minor difference. As a result, rather than identifying matches solely on the basis of correlation ranking, glycosides often require verification on the basis of characteristic sugar signals which are readily available. Compound **14** is another good example to illustrate this concept: whereas the best match to **14** was an arabinopyranoside with $r = 0.9997$, the sugar was identified as xylopyranose based on the characteristic ^1H NMR signals.

CONCLUSION

This study introduces two new tools for the efficient study of triterpenes present in *Actaea* plants, a genus extensively used in complementary and alternative medicine and a major source of these natural products. A new semi-systematic naming scheme links compound name to the actual chemical structure, and a rapid dereplication tool utilizes the readily available information of ^1H NMR chemical shifts of Me groups as well as an in-house database. A rationale naming scheme plays an important role in dereplication, because

unambiguous compound names provide crucial links between the literature and the actual structures. *Actaea* triterpenes served as examples to demonstrate how these tools were developed and utilized in practice. By using the Me shifts as indicators of structural characteristics, two classification models based on CDA and CBTs were generated for *in silico* classification of *Actaea* triterpenes according to their aglycone type. This concept has potential to be adopted for any other class of natural products with characteristic and readily accessible chemical shift information, such as Me groups.

Both CDA and CBTs exhibit high accuracy when classifying the *Actaea* triterpenes in our in-house database using only Me chemical shifts. Comparing these two methods in practical use, CBTs are more straightforward, simple to understand, and interpret. The CBT model can be implemented in procedural computer algorithm such as VBA code. Therefore, CBTs are not only efficient in the dereplication of triterpenes, as shown, but are also a promising dereplication tool for other natural products, such as steroids, other terpenoids, and peptides.

Looking forward, using a combination of characteristic ^1H Me shifts (Me-28) and $^{2/3}J_{\text{C,H}}$ HMBC coupling patterns, we were able to tentatively assign *R*- and *S*-actein as two minor constituents in E2 (Figure 7), present only at the 3–5% level. While further results will be reported in due course, it is safe to conclude that the presented approach, combined with contemporary (q)NMR methodology using 700 MHz 1.7 mm cryo-microprobe equipment, has future potential for the standard characterization of residual complexity of natural products reference materials, allowing analysis of several minor constituents down to the 10–20 μg level in a 1–5 mg sample. Recently, the power of HSQC in the analysis of complex mixtures has been shown.^{29,30} Future studies will also adopt HSQC-DEPT which is not only more sensitive than other 2D ^1H – ^{13}C experiments, but also provides an extra dimension by tying the ^{13}C chemical shifts to the appropriate ^1H chemical shifts for the methyl groups. HSQC-DEPT is particularly useful in determining and differentiating the methyl groups of individual triterpenes in complex mixtures and, thus, improves the dereplication process.

In our experience, even repeatedly purified reference materials of biosynthetically diverse natural products such as triterpenes often exhibit surprisingly high degrees of residual complexity. In this regard, the case of E2 is particularly noteworthy, because it shows that constitutionally and spatially distinct natural products can exhibit similar chromatographic behavior, even in multi-step purification procedures: this study identified E2 as a mixture of more than five compounds which belong to at least four different skeleton types: one acta-16,23;16,24-binoxol, one actanol, and three acta-16,23;24,26-binoxols belonging to two different sub-types. While the different abundance levels are important parameters of residual complexity, the observed co-occurrence of considerably different chemical species is of broader importance regarding bioactivity. First, the evaluation of the degree and pattern of residual complexity of purified natural products should be considered a prerequisite for their biological assessment. Second, assumptions about 3D structural similarities can potentially be misleading and have to be verified for each particular sample used in a bioassay. The dereplication tools introduced here, in combination with qualitative and quantitative ^1H NMR analysis, might inspire future applications for a wider range of natural products.

EXPERIMENTAL SECTION

Construction of the In-house Database

In order to obtain sufficient data to develop effective classification models, an extensive literature search was carried out through SciFinder (American Chemical Society, Washington D.C.) to locate reports of *Actaea* triterpenes with spectroscopic data. The ^1H

NMR data of cyclopropane methylene (H-19a/b) and all Me groups including chemical shift (δ_{H} , ppm), multiplicity and assignment for each reported triterpene were collected and entered into a spreadsheet database using Microsoft Excel 2010. All chemical shift values were recorded with two decimal places for a homogenous dataset. The Me groups of each compound were given a series of names as Me1, Me2, Me3, etc. in ascending order of chemical shift. While pyridine- d_5 was used for most triterpene glycosides, the less polar solvent CDCl_3 was used in a few cases, especially for triterpene aglycones. Due to the effect of various solvents on the chemical shifts, the NMR solvent was also noted for each compound. Mining of other NMR acquisition parameters including temperature was omitted due to the frequent lack of reporting in the literature.

After collection of these data sets and construction of the spreadsheet database, an *in silico* “search-and-match” function named “ActaMatch” (Supporting Information S5) was developed using Microsoft Visual Basic for Applications 7.0 (VBA). Its search function is based on the Pearson’s Coefficient (r) as a measure of the similarity of the pattern of Me shifts between the investigated compound and the compounds in the database (Eq. 1):

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (\text{Eq. 1})$$

where x and y denote the Me shifts of the investigated compound and any compound in the database, respectively. By entering the Me shifts of the investigated compound and an appropriate r value (r_0), the VBA-coded program automatically determines which compound(s) in the database fulfill $r > r_0$ and lists the hit(s) on the output page.

Development of Classification Models

The canonical discriminant analysis (CDA) was performed in Microsoft Excel 2010 with the XLSTAT-Pro 7.5 add-on (Addinsoft, Paris, France), using the triterpene type as dependent variable and the Me shifts as explanatory variables. The classification binary trees (CBTs) were generated by classification and regression tree (CART) analysis within the Salford Predictive Miner v6.6 (Salford Systems, San Diego, CA), by using the Me shifts as the descriptors. The resulting CBTs were further implemented in a computer algorithm by VBA in Microsoft Excel 2010, leading to an *in silico* tool named “ActaPredict” (Supporting Information S4).

^1H NMR Analysis of *Actaea* triterpenes

Investigated samples of purified but residually complex reference materials of *Actaea* triterpenes were initially analyzed as follows. Each sample was dissolved in 35 μL of pyridine- d_5 (99.96% D, Aldrich-Sigma, St. Louis, MO) and transferred to a 1.7 mm NMR tube. The ^1H NMR spectra were recorded on a Bruker Avance-III 700 MHz NMR spectrometer equipped with a 1.7 mm cryo-microprobe (Bruker BioSpin, Karlsruhe, Germany) using the following acquisition parameters: Pulse program zg30, 128 scans, 32 K complex points, 14423 Hz spectral width, acquisition time 2.3 s, and receiver gain 57. FIDs were processed using MestReNova v7.0.2-8636 software (Mestrelab Research, Santiago de Compostela, Spain). Line resolution was improved by Lorentzian-Gaussian (LG) window functions (LB -2.0 , GB 0.10) and three times of zero-filling, prior to Fourier transformation of the FID data. For mixtures of *Actaea* triterpenes, individual Me groups were distinguished according to their matching integral using the highly disperse H-19 signals as reference. Overlapped Me signals were deconvoluted by using the advanced functionality of Global Spectra Deconvolution (GSD) within MestReNova software.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported in parts by grant P50 AT000155 from the National Center for Complementary and Alternative Medicine (NCCAM) and the Office of Dietary Supplements (ODS), both of the National Institutes of Health (NIH), as well as grant RC2 AT005899 from NCCAM/NIH. We thank Dr. Tanja Gödecke at UIC for providing *Actaea racemosa* extract, and Dr. Shi-Hui Dong for his generous help during manuscript preparation. We are also grateful to Drs. Chen Peng and Carlos Cobas at MestReLab Research for support with the MestReNova software package. The authors kindly acknowledge helpful discussions with Dr. John Walter, IMB/NRC, Halifax (Canada).

REFERENCES

1. Li JW-H, Vederas JC. *Science*. 2009; 325:161–165. [PubMed: 19589993]
2. Steinbeck C. *Nat. Prod. Rep.* 2004; 21:512–518. [PubMed: 15282633]
3. Kalchhauser H, Robien W. *J. Chem. Inform. Comput. Sci.* 1985; 25:103–108.
4. Corley DG, Durley RC. *J. Nat. Prod.* 1994; 57:1484–1490.
5. Bradshaw J, Butina D, Dunn AJ, Green RH, Hajek M, Jones MM, Lindon JC, Sidebottom PJ. *J. Nat. Prod.* 2001; 64:1541–1544. [PubMed: 11754607]
6. Lang G, Mayhudin NA, Mitova MI, Sun L, van der Sar S, Blunt JW, Cole ALJ, Ellis G, Laatsch H, Munro MHG. *J. Nat. Prod.* 2008; 71:1595–1599. [PubMed: 18710284]
7. Compton JA, Culham A, Jury SL. *Taxon*. 1998; 47:593–634.
8. Li J-X, Yu Z-Y. *Curr. Med. Chem.* 2006; 13:2927–2951. [PubMed: 17073639]
9. Zhou L, Yang J-S, Zou J-H, Tu G-Z. *Chem. Pharm. Bull.* 2004; 52:622–624. [PubMed: 15133220]
10. Zhou L, Yang J-S, Wu X, Zou J-H, Xu X-D, Tu G-Z. *Heterocycles*. 2005; 64:1409–1414.
11. Zhou L, Yang J-S, Tu G-Z, Zu J-H. *Chem. Pharm. Bull.* 2006; 54:823–826. [PubMed: 16755051]
12. Shoolery JN, Rogers MT. *J. Am. Chem. Soc.* 1958; 80:5121–5135.
13. Bhacca, NS.; Williams, DH. *Applications of NMR Spectroscopy in Organic Chemistry; Illustrations From the Steroid Field*. San Francisco, CA: Holden-Day; 1964.
14. Lavie D, Benjaminov BS, Shvo Y. *Tetrahedron*. 1964; 20:2585–2592.
15. Tursch B, Savoie R, Ottinger R, Chiurdoglu G. *Tetrahedron Lett.* 1967; 8:539–543.
16. Cheung HT, Wong C-S, Yan TC. *Tetrahedron Lett.* 1969; 10:5077–5080.
17. Cheung HT, Williamson DG. *Tetrahedron*. 1969; 25:119–128.
18. Pauli, GF. Ph.D. Dissertation. Düsseldorf, Germany: Heinrich-Heine-University; 1993. Cardenolide aus *Adonis aleppica* Boiss. - Isolierung und Strukturaufklärung.
19. Pauli, GF. Attacking Cardenolides in the 1ppm Range; Annual Meeting of the Gesellschaft für Arzneipflanzenforschung (GA); Halle, Germany. 1995.
20. Wilson WK, Sumpter RM, Warren JJ, Rogers PS, Ruan B, Schroepfer GJ Jr. *J. Lipid Res.* 1996; 37:1529–1555. [PubMed: 8827525]
21. Kokkinofa R, Petrakis PV, Mavromoustakos T, Theocharis CR. *J. Agr. Food Chem.* 2003; 51:6233–6239. [PubMed: 14518949]
22. Petrakis PV, Touris I, Liouni M, Zervou M, Kyrikou I, Kokkinofa R, Theocharis CR, Mavromoustakos TM. *J. Agr. Food Chem.* 2005; 53:5293–5303. [PubMed: 15969510]
23. Petrakis PV, Agiomyrgianaki A, Christophoridou S, Spyros A, Dais P. *J. Agr. Food Chem.* 2008; 56:3200–3207. [PubMed: 18422335]
24. Brown, SD.; Tauler, R.; Walczak, B. *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*. Vol. Vol. 3. Amsterdam: Elsevier; 2009. p. 542-567.
25. Chen S-N, Lankin DC, Chadwick LR, Jaki BU, Pauli GF. *Planta Med.* 2009; 75:757–762. [PubMed: 19145555]

26. Schinkovitz A, Pro SM, Main M, Chen S-N, Jaki BU, Lankin DC, Pauli GF. *J. Nat. Prod.* 2008; 71:1604–1611. [PubMed: 18781813]
27. Jaki BU, Franzblau SG, Chadwick LR, Lankin DC, Zhang F, Wang Y, Pauli GF. *J. Nat. Prod.* 2008; 71:1742–1748. [PubMed: 18798682]
28. Steinberg, D.; Golovnya, M. *CART 6.0 User's Guide*. San Diego, CA: Salford Systems; 2006. p. 61-63.
29. Lewis IA, Schommer SC, Hodis B, Robb KA, Tonelli M, Westler WM, Sussman MR, Markley JL. *Anal. Chem.* 2007; 79:9385–9390. [PubMed: 17985927]
30. Xi Y, de Ropp JS, Viant MR, Woodruff DL, Yu P. *Anal. Chim. Acta.* 2008; 614:127–133. [PubMed: 18420042]

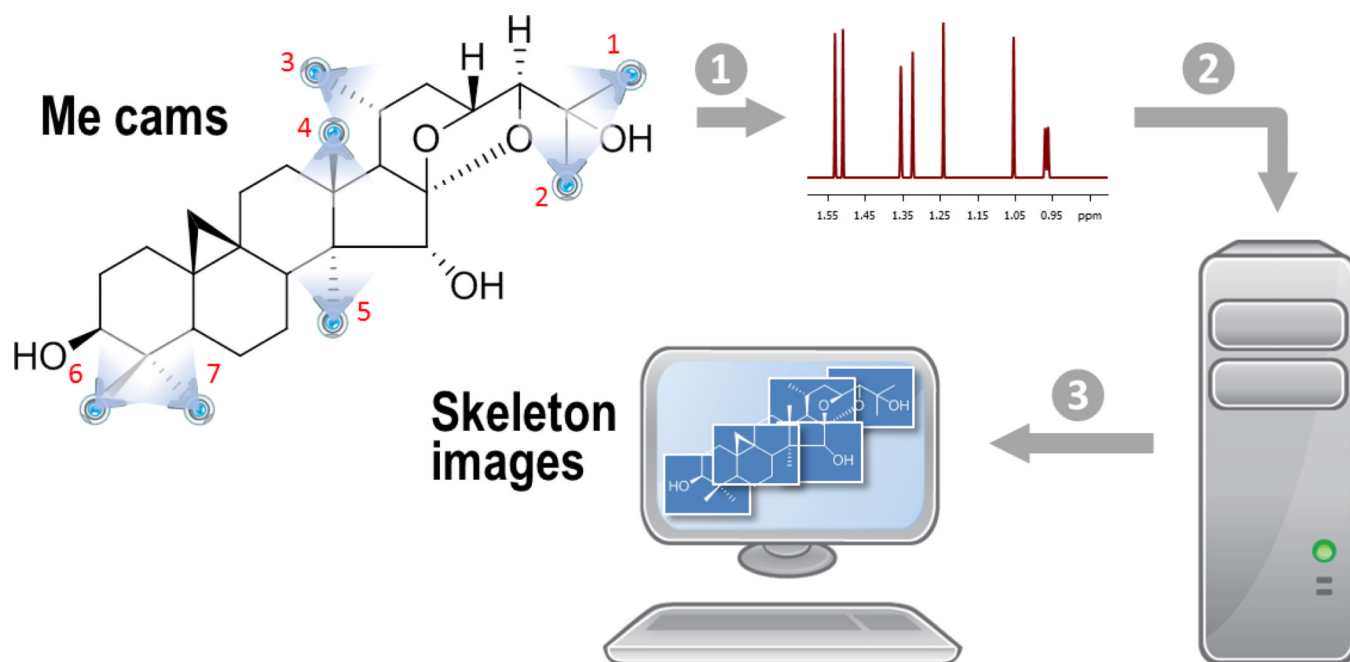


Figure 1. The concept of using methyl (Me) groups of triterpenes as partial structural indicators (“Me cams”) to map the full skeleton of the molecules

Imitating the mechanism of a biometric system, the raw data (e.g., Me shifts and multiplicities) collected by each of the “Me cams” ① are processed *in silico* for pattern recognition ② and converted to visible images, representing partial structures from which the full structure can be assembled ③.

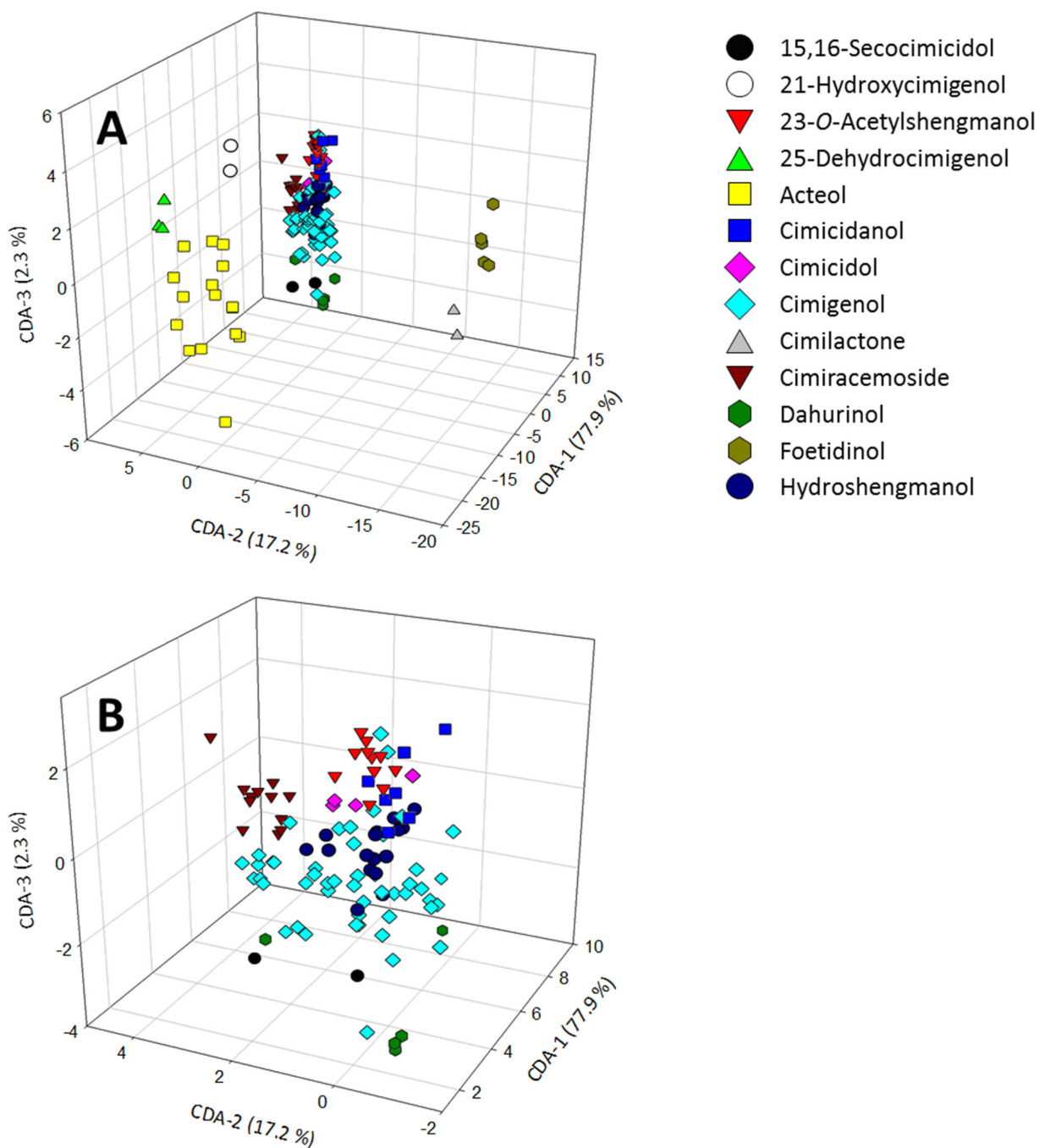


Figure 2.

Classification of all *Actaea* triterpenes contained in the in-house database based on CDA analysis. The 3D plot (panel A; axes CDA-1 = 77.9%, CDA-2 = 17.2%, CDA-3 = 2.3%) shows that the first 3 factors account for 97.4% of the total variance in the Me shifts of the compounds. Panel B shows the sub-cluster of all triterpenes with seven Me groups ($\delta_{\text{H}} < 1.90$) having CDA-1 scores between 0 and 10, which form further sub-clusters depending on the specific skeleton types.

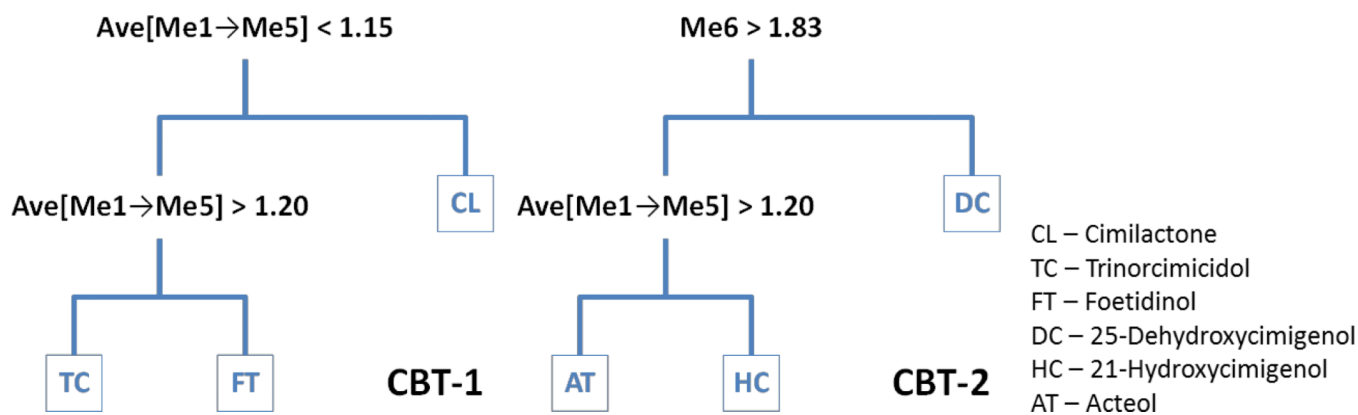


Figure 3. The CBTs developed for the classification of *Actaea* triterpenes with five (CBT-1) and six (CBT-2) Me groups ($\delta_H < 1.90$). Ave[Me1→Me5] denotes the average of all five Me shifts (Me1 to Me5). In case the answer to a given descriptor/splitter is yes, it branches to the right child node.

CBT-3

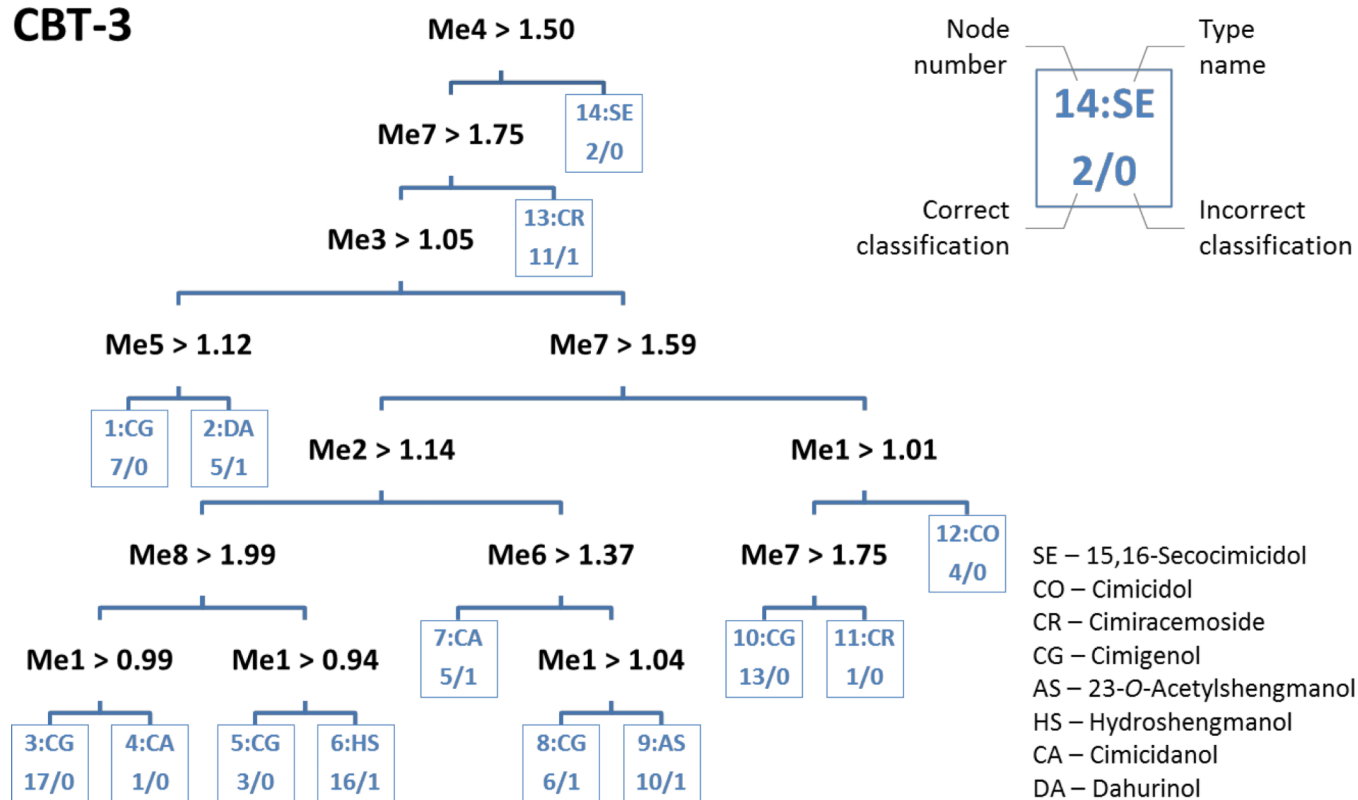


Figure 4. The CBTs developed for classification of *Actaea* triterpenes with seven Me groups ($\delta_H < 1.90$).

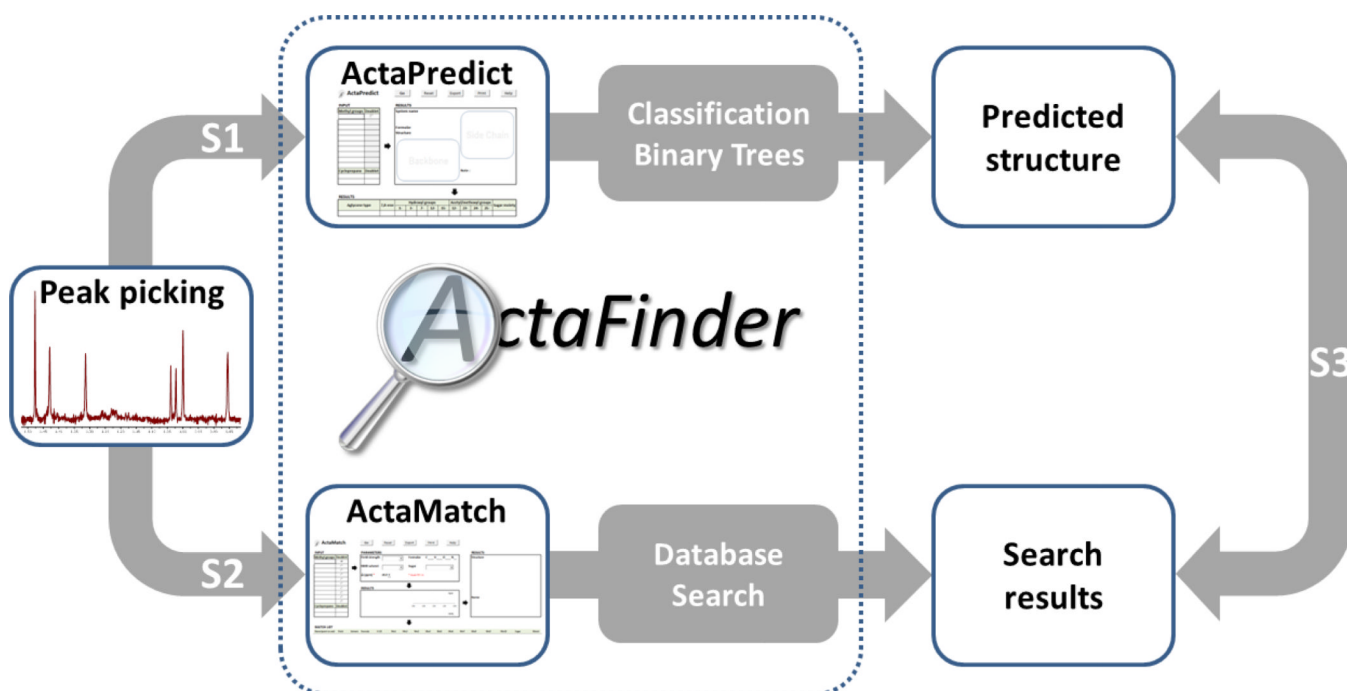


Figure 5. The new *in silico* tool ActaFinder comprised of two modules ActaPredict (step S1) and ActaMatch (step S2) was used for the automatic dereplication of *Actaea* triterpenes. In a third step (S3), the results of S1 and S2 are compared for consistency. This approach can potentially be adopted for other natural products using characteristic and readily accessible ^1H chemical shift information, such as of Me groups.

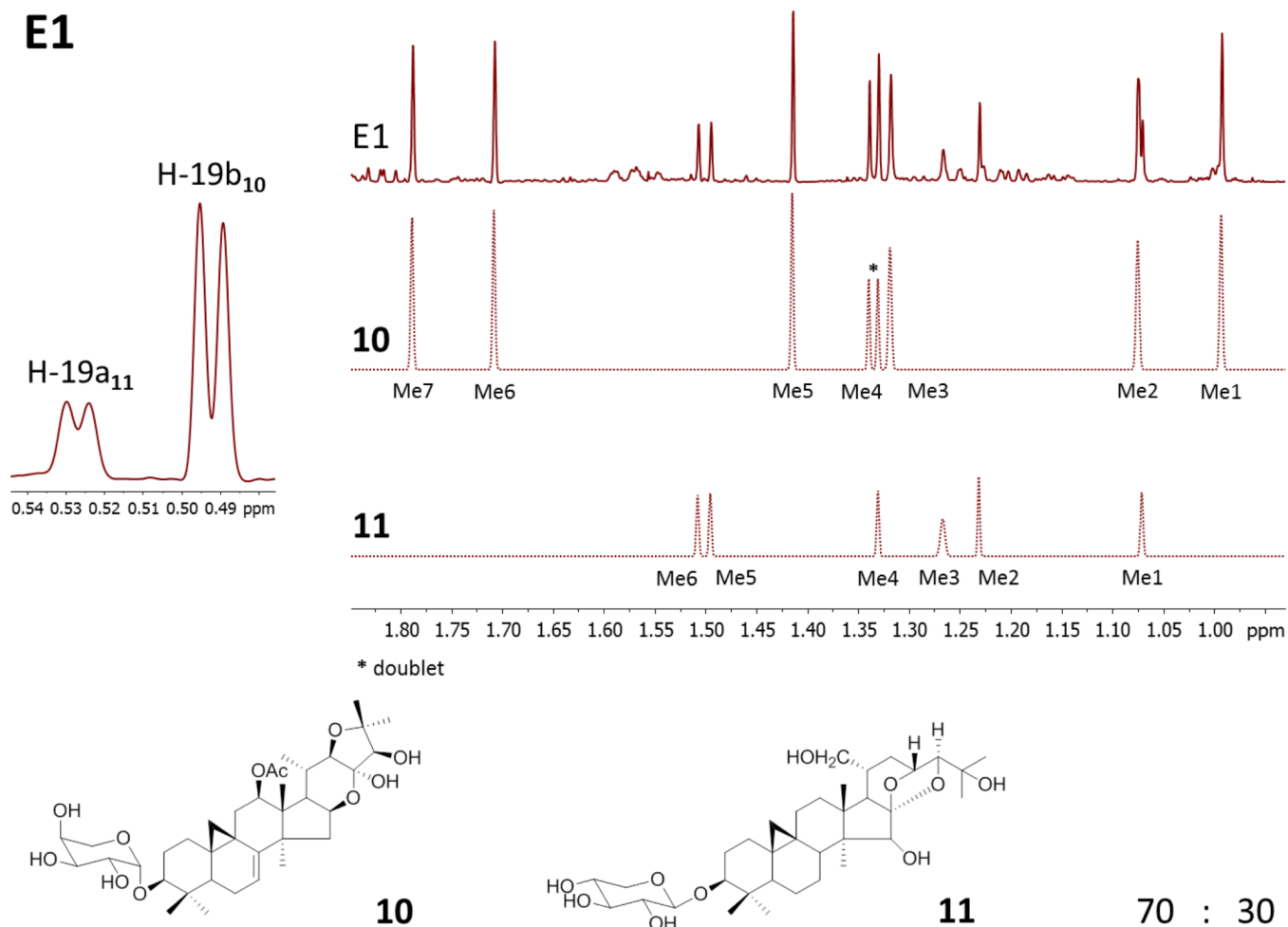


Figure 6. The exact composition of a residually complex sample of a purified *Actaea* triterpene, E1, was analyzed by a combination of ^1H NMR spectral deconvolution, CBT partitioning of Me ^1H chemical shifts, and characteristic ^1H NMR sugar signals. Sample E1 exhibited moderate residual complexity, which is frequently found with *Actaea* triterpene reference materials, and can be considered a “clean” 70:30 mixture of the two triterpenes **10** and **11**.

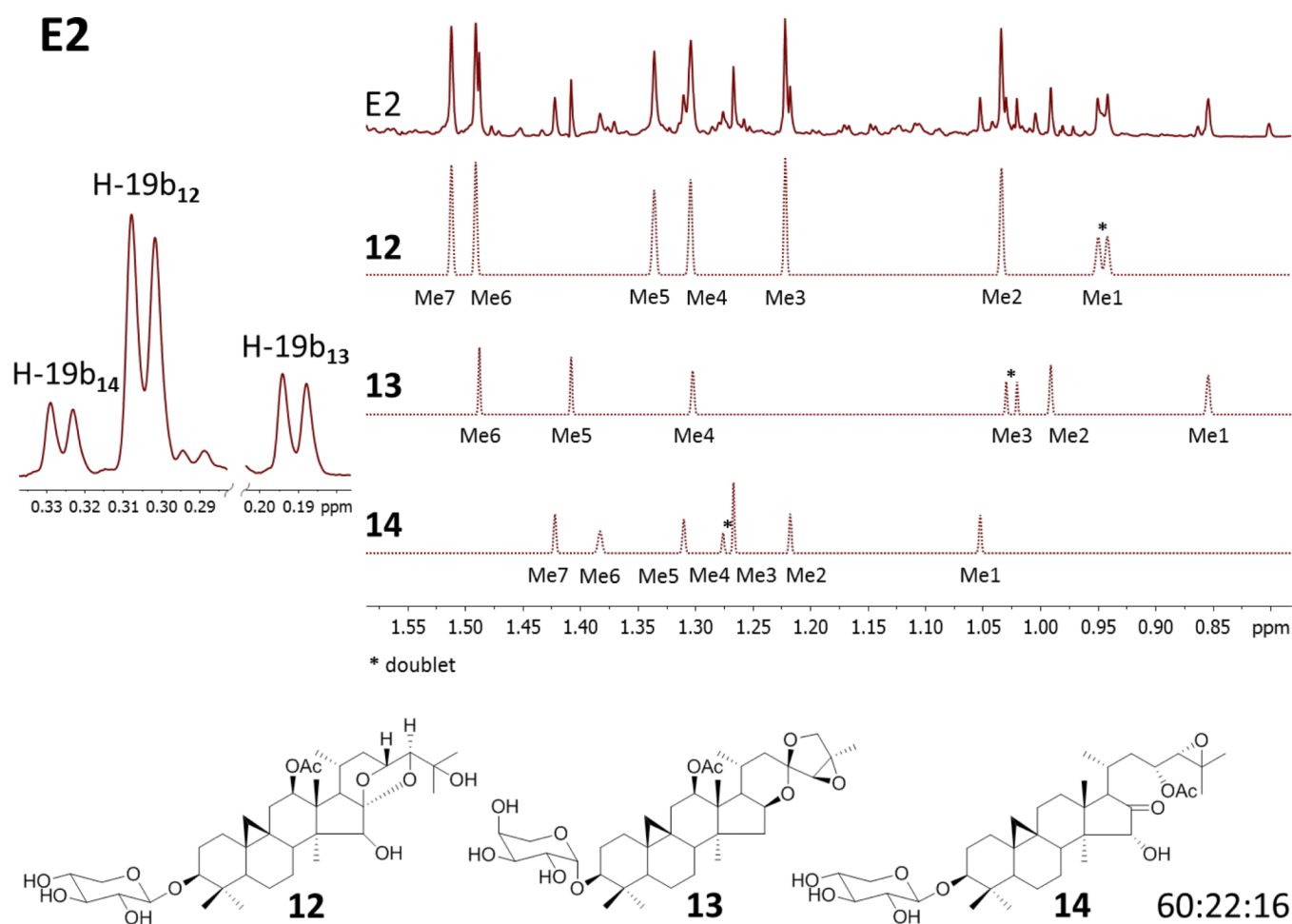


Figure 7.

Using the analogous approach as for sample E1 (Figure 6), analysis of E2 led to the identification of three major triterpenes, **12**, **13** and **14**, in this residually rather complex mixture. Because these compounds were present in a 62:22:16 ratio, their Me signals were readily distinguished and amenable to CBT dereplication. Interestingly, the minor impurities giving rise to Me singlets at 0.87 and 0.80 ppm could be assigned to *R*- and *S*-actein, at ~3 and ~5 mol% abundance, respectively, using their known Me-28 chemical shifts and characteristic HMBC coupling patterns (see main text).

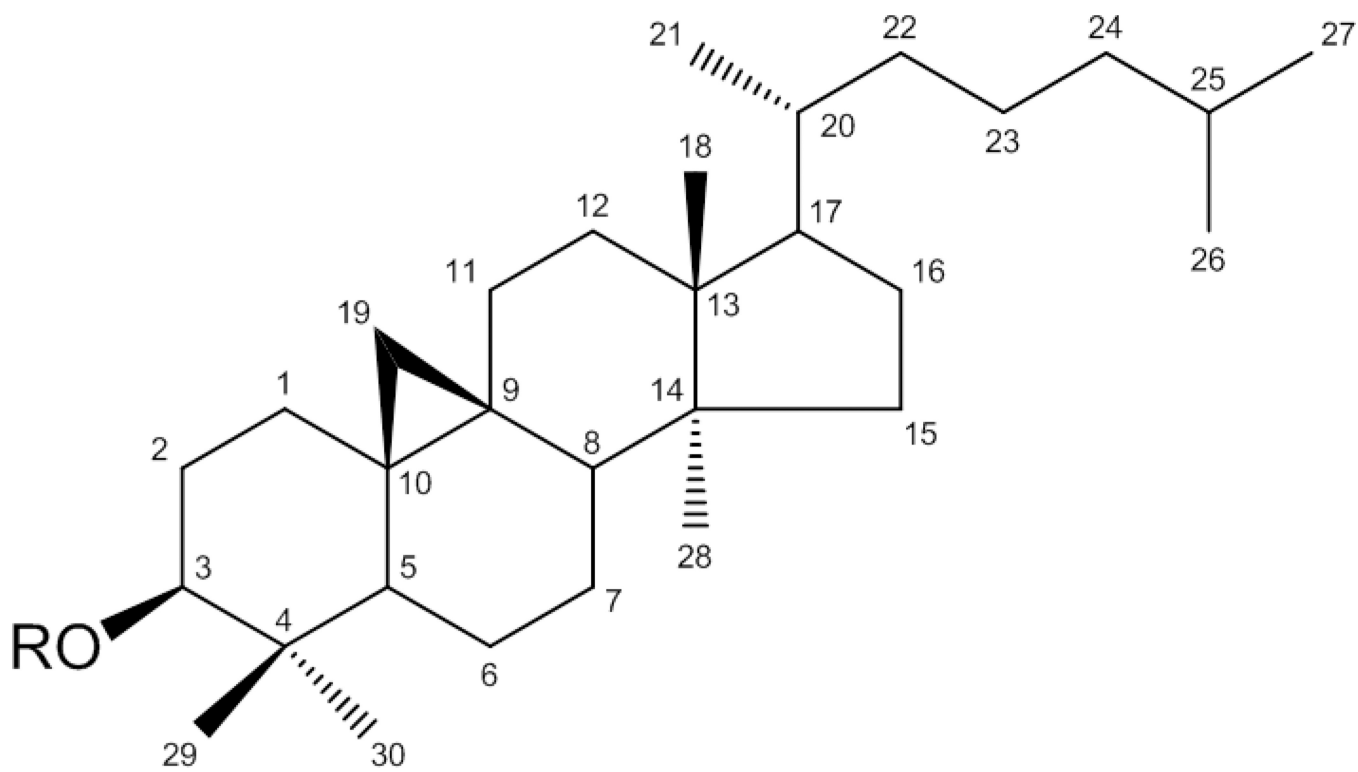


Chart 1.

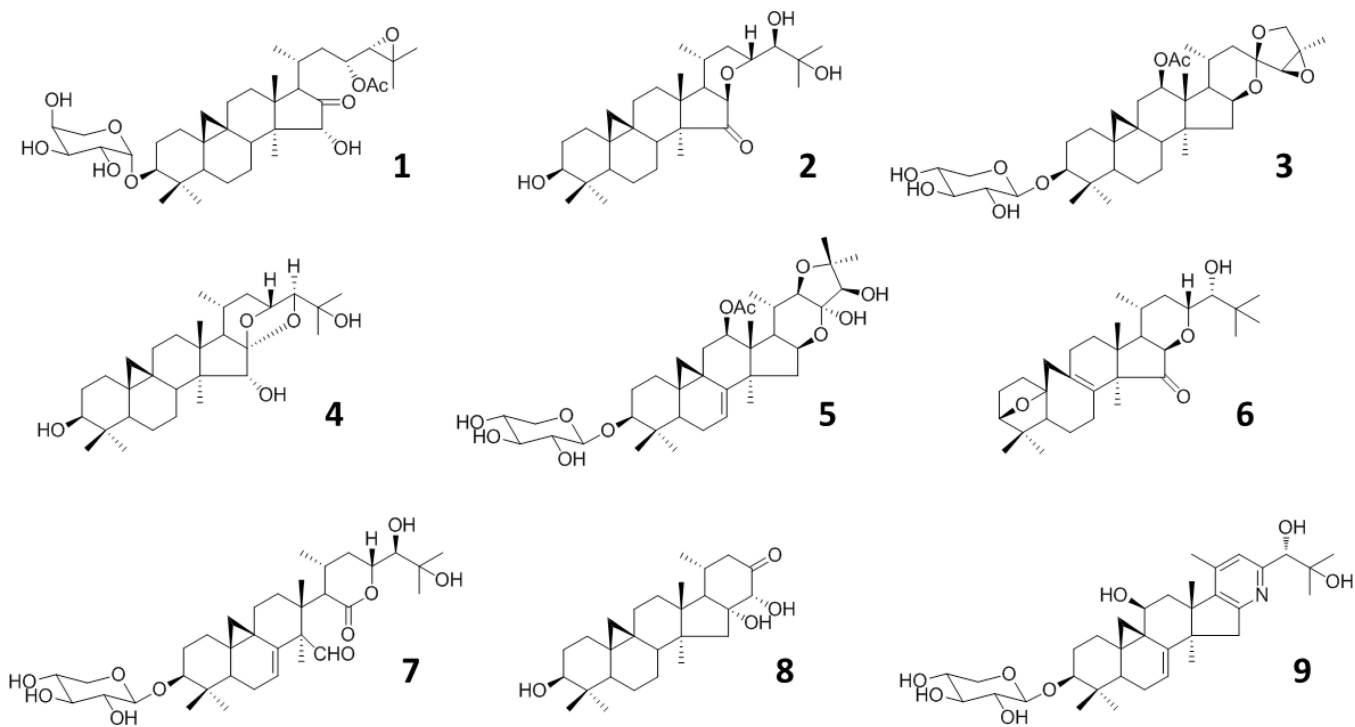


Chart 2.

Table 1Representatives of the *Actaea* Triterpenes. Structural Types 1–5 Cover More Than 90% of Known Structures.

| Common name | New systematic name |
|--|--|
| 23- <i>O</i> -acetylshengmanol arabinoside (1) | (23 <i>R</i>)-23-acetoxy-(24 <i>S</i>)-24,25-epoxy-(15 <i>R</i>)-15-hydroxy-16-oxo-3- <i>O</i> - β -D-arabinopyranosylactanoside |
| dahurinol (2) | (24 <i>R</i>)-24,25-dihydroxy-15-oxoacta-(16 <i>R</i> ,23 <i>R</i>)-16,23-monoxol |
| 23- <i>epi</i> -26-deoxyactein (3) | (12 <i>R</i>)-12-acetoxy-(24 <i>R</i> ,25 <i>R</i>)-24,25-epoxy-3- <i>O</i> - β -D-xylopyranosylacta-(16 <i>S</i> ,23 <i>R</i>)-16,23;23,26-binoxoside |
| cimiracemoside F(4) | (12 <i>R</i>)-12-acetoxy-7,8-didehydro-(23 <i>R</i> ,24 <i>S</i>)-23,24-dihydroxy-3- <i>O</i> - α -L-xylopyranosylacta-(16 <i>S</i> ,22 <i>R</i>)-16,23;22,25-binoxoside |
| cimigenol (5) | (15 <i>R</i>)-15,25-dihydroxyacta-(16 <i>S</i> ,23 <i>R</i> ,24 <i>S</i>)-16,23;16,24-binoxol |
| acerionol (6) | 3-deoxy-8,9-didehydro-(24 <i>S</i>)-24,25-dihydroxy-(3 <i>S</i> ,10 <i>S</i>)-3,10-epoxy-15-oxo-9,10-secoacta-(16 <i>R</i> ,23 <i>R</i>)-16,23-monoxol |
| compound 7 ^a | 7,8-didehydro-(24 <i>R</i>)-24,25-dihydroxy-15-formyl-16-oxo-15,16-seco-3- <i>O</i> - β -D-xylopyranosylacta-(23 <i>R</i>)-16,23-monoxoside |
| foetidinol (8) | (16 <i>R</i> ,24 <i>R</i>)-16,24-dihydroxy-23-oxo-25,26,27-trinoracta-16,23-carbamonol |
| cimicifugadine (9) [an alkaloid] | (11 <i>S</i> ,24 <i>S</i>)-11,24,25-trihydroxy-7,8,16,17,20,22,23, <i>N</i> -octadehydro-3- <i>O</i> - β -D-xylopyranosylacta-16,23-monozoside |

^aNo common name has been assigned

Table 2

The Major Types of *Actaea* Triterpenes Included in the In-house Database, Along with Their Common (⊗) and New Systematic Names (⊙). The Methyl Groups Used for the Dereplication Models are Indicated in Red.

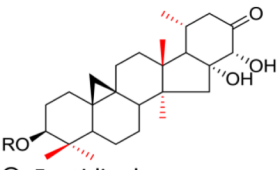
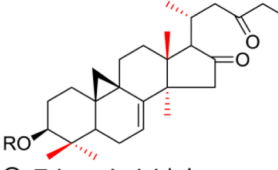
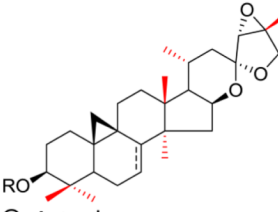
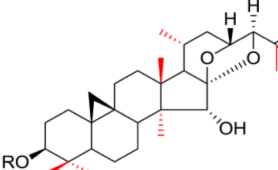
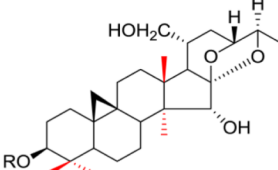
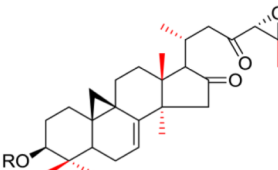
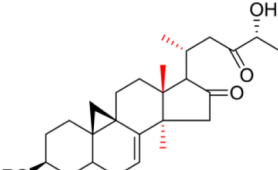
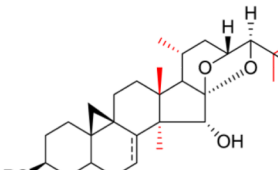
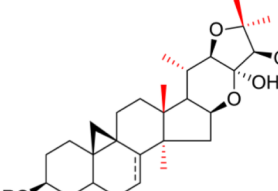
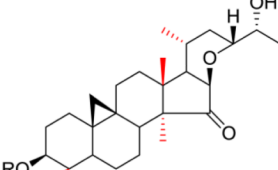
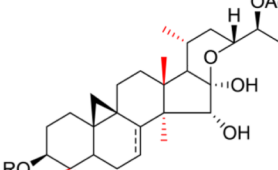
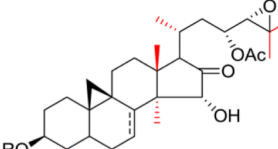
| | | |
|---|--|---|
|  <p>⊗ Foetidinol ⊙ (16<i>R</i>,24<i>R</i>)-16,24-dihydroxy-23-oxo-25,26,27-trinoracta-16,23-carbamonol</p> |  <p>⊗ Trinorcimicidol ⊙ 7,8-didehydro-24-hydroxy-16,23-dioxo-25,26,27-trinoractanol</p> |  <p>⊗ Acteol ⊙ (24<i>R</i>,25<i>R</i>)-24,25-epoxyacta-(16<i>S</i>,23<i>R</i>)-16,23;23,26-binoxol</p> |
|  <p>⊗ 25-Dehydrocimigenol ⊙ 25,26-didehydro-(15<i>R</i>)-15-hydroxyacta-(16<i>S</i>,23<i>R</i>,24<i>S</i>)-16,23;16,24-binoxol</p> |  <p>⊗ 21-Hydroxycimigenol ⊙ (15<i>R</i>)-15,21,25-trihydroxyacta-(16<i>S</i>,23<i>R</i>,24<i>S</i>)-16,23;16,24-binoxol</p> |  <p>⊗ Cimicidanol ⊙ 7,8-didehydro-(24<i>S</i>)-24,25-epoxy-16,23-dioxoactanol</p> |
|  <p>⊗ Cimicidol ⊙ 7,8-didehydro-(24<i>R</i>)-24,25-dihydroxy-16,23-dioxoactanol</p> |  <p>⊗ Cimigenol ⊙ (15<i>R</i>)-15,25-dihydroxyacta-(16<i>S</i>,23<i>R</i>,24<i>S</i>)-16,23;16,24-binoxol</p> |  <p>⊗ Cimiracemoside ⊙ (23<i>R</i>,24<i>S</i>)-23,24-dihydroxyacta-(16<i>S</i>,22<i>R</i>)-16,23;22,25-binoxol</p> |
|  <p>⊗ Dahurinol ⊙ (24<i>R</i>)-24,25-dihydroxy-15-oxoacta-(16<i>R</i>,23<i>R</i>)-16,23-monoxol</p> |  <p>⊗ Hydroschengmanol ⊙ (24<i>S</i>)-24-acetoxy-(15<i>R</i>,16<i>R</i>)-15,16,25-trihydroxyacta-(23<i>R</i>)-16,23-monoxol</p> |  <p>⊗ 23-<i>O</i>-Acetylshengmanol ⊙ (23<i>R</i>)-23-acetoxy-(24<i>S</i>)-24,25-epoxy-(15<i>R</i>)-15-hydroxy-16-oxoactanol</p> |

Table 3

Classification Results for *Actaea* Triterpenes Containing Seven Methyl Groups with $\delta_H < 1.90$ Using the CDA Analysis.

| Type | Total | Correct (%) | Classified Type | | | | | | | | |
|------|-------|-------------|-----------------|----|----|----|----|----|----|----|---|
| | | | SE | CR | CO | CG | AS | CA | HS | DA | |
| SE | 2 | 100 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CR | 12 | 91.7 | 0 | 11 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| CO | 4 | 75.0 | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 0 | 0 |
| CG | 50 | 84.0 | 0 | 1 | 0 | 42 | 2 | 0 | 4 | 1 | 0 |
| AS | 11 | 100 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 0 |
| CA | 7 | 28.6 | 0 | 0 | 0 | 0 | 4 | 2 | 1 | 0 | 0 |
| HS | 16 | 81.2 | 0 | 0 | 0 | 3 | 0 | 0 | 13 | 0 | 0 |
| DA | 5 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |

Table 4

Classification Results for *Actaea* Triterpenes Containing Seven Methyl Groups with $\delta_H < 1.90$ Using the CBT-3 Partitioning.

| Type | Total | Correct (%) | Classified Type | | | | | | | | |
|------|-------|-------------|-----------------|----|----|----|----|----|----|----|---|
| | | | SE | CR | CO | CG | AS | CA | HS | DA | |
| SE | 2 | 100 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CR | 12 | 100 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CO | 4 | 100 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| CG | 50 | 92.0 | 0 | 1 | 0 | 46 | 1 | 0 | 1 | 1 | 1 |
| AS | 11 | 90.9 | 0 | 0 | 0 | 0 | 10 | 1 | 0 | 0 | 0 |
| CA | 7 | 85.7 | 0 | 0 | 0 | 1 | 0 | 6 | 0 | 0 | 0 |
| HS | 16 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 0 |
| DA | 5 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |

Table 5
The Accuracy of the Prediction Performance of the CBT-3 Estimated by Leave-one-out Cross-validation.

| Type | Total | Correct (%) | Predicted Type | | | | | | | | |
|------|-------|-------------|----------------|----|----|----|----|----|----|----|---|
| | | | SE | CR | CO | CG | AS | CA | HS | DA | |
| SE | 2 | 100 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CR | 12 | 91.7 | 0 | 11 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| CO | 4 | 50.0 | 0 | 0 | 2 | 0 | 1 | 1 | 1 | 0 | 0 |
| CG | 50 | 80.0 | 0 | 1 | 1 | 40 | 2 | 4 | 1 | 1 | 1 |
| AS | 11 | 72.7 | 0 | 0 | 0 | 0 | 8 | 3 | 0 | 0 | 0 |
| CA | 7 | 57.1 | 0 | 0 | 0 | 1 | 2 | 4 | 0 | 0 | 0 |
| HS | 16 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 0 |
| DA | 5 | 60.0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 3 |

Table 6
 Dereplication Results of the Triterpenes Contained in the Residually Complex Samples E1 and E2.

| Compound | Me1 | Me2 | Me3 | Me4 | Me5 | Me6 | Me7 | Me8 | r |
|--|-------------------------|-------------|-------------------------|-------------------------|-------------|-------------|-------------|-------------|--------|
| E1-10 ^a | 0.99 | 1.08 | 1.32 | 1.33^c | 1.41 | 1.71 | 1.79 | 2.14 | 0.9998 |
| cimiracemoside G ^b (ara) | 0.97 | 1.05 | 1.28 | 1.31 ^c | 1.39 | 1.68 | 1.76 | 2.11 | |
| E1-11 | 1.08 | 1.23 | 1.27 | 1.34 | 1.50 | 1.51 | - | - | 0.9975 |
| 21-dehydrocimigenol xy ^b | 1.04 | 1.20 | 1.24 | 1.28 | 1.46 | 1.48 | - | - | |
| E2-12 ^a | 0.94^c | 1.03 | 1.22 | 1.30 | 1.34 | 1.49 | 1.51 | 2.14 | 0.9992 |
| 12-O-acetylcimigenol xy ^b | 0.92 ^c | 0.98 | 1.21 | 1.25 | 1.31 | 1.47 | 1.49 | 2.10 | |
| E2-13 | 0.85 | 0.99 | 1.03^c | 1.30 | 1.41 | 1.49 | 2.15 | - | 0.9994 |
| cimiracemoside N ^b (ara) | 0.85 | 0.96 | 1.02 ^c | 1.27 | 1.42 | 1.48 | 2.14 | - | |
| E2-14 | 1.05 | 1.22 | 1.27 | 1.28^c | 1.31 | 1.38 | 1.42 | 2.07 | 0.9997 |
| 23-O-acetylshengmanol ara ^b | 1.05 | 1.21 | 1.25 | 1.26 ^c | 1.30 | 1.37 | 1.40 | 2.06 | |

^aMain component.

^bTriterpene with the best match (*r*).

^cIndicates Me-21 doublets.