Prediction of Anti-inflammatory Plants and Discovery of Their Biomarkers by Machine Learning Algorithms and Metabolomic Studies

DOI 10.1055/s-0034-1396206
Planta Med 2015; 81: 450–458

Publisher and Copyright:
© 2015 by
Georg Thieme Verlag KG
Rüdigerstraße 14
70469 Stuttgart
ISSN 0032-0943

Reprint with the permission by the publisher only

Thieme

Prediction of Anti-inflammatory Plants and Discovery of Their Biomarkers by Machine Learning Algorithms and Metabolomic Studies

Authors **Daniela Aparecida Chagas-Paula**[1], **Tiago Branquinho Oliveira**[1], **Tong Zhang**[2], **RuAngelie Edrada-Ebel**[2], **Fernando Batista Da Costa**[1]

Affiliations
[1] AsterBioChem Research Team, Laboratory of Pharmacognosy, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, SP, Brazil
[2] Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, Scotland

Key words
- Asteraceae
- anti-inflammatory
- cyclooxygenase
- lipoxygenase
- machine learning algorithms
- metabolomics

received July 22, 2014
revised October 22, 2014
accepted Dec. 14, 2014

Bibliography
DOI http://dx.doi.org/10.1055/s-0034-1396206
Published online January 23, 2015
Planta Med 2015; 81: 450–458

ISSN 0032-0943

Correspondence
Prof. Dr. Fernando B. Da Costa
Laboratory of Pharmacognosy
School of Pharmaceutical Sciences of Ribeirão Preto
University of São Paulo
Av. do Café s/n
14040-903 Ribeirão Preto, SP
Brazil
Phone: + 55 16 36 02 06 61
Fax: + 55 16 36 02 48 79
febcosta@fcfrp.usp.br

Abstract

Nonsteroidal anti-inflammatory drugs are the most used anti-inflammatory medicines in the world. Side effects still occur, however, and some inflammatory pathologies lack efficient treatment. Cyclooxygenase and lipoxygenase pathways are of utmost importance in inflammatory processes; therefore, novel inhibitors are currently needed for both of them. Dual inhibitors of cyclooxygenase-1 and 5-lipoxygenase are anti-inflammatory drugs with high efficacy and low side effects. In this work, 57 leaf extracts (EtOH-H_2O 7 : 3, v/v) from Asteraceae species with *in vitro* dual inhibition of cyclooxygenase-1 and 5-lipoxygenase were analyzed by high-performance liquid chromatography-high-resolution-ORBITRAP-mass spectrometry analysis and subjected to *in silico* studies using machine learning algorithms. The data from all samples were processed by employing differential expression analysis software coupled to the Dictionary of Natural Products for dereplication studies. The 6052 chromatographic peaks (ESI positive and negative modes) of the extracts were selected by a genetic algorithm according to their respective anti-inflammatory properties; after this procedure, 1241 of them remained. A study using a decision tree classifier was carried out, and 11 compounds were determined to be biomarkers due to their anti-inflammatory potential. Finally, a model to predict new biologically active extracts from Asteraceae species using liquid chromatography-mass spectrometry information with no prior knowledge of their biological data was built using a multilayer perceptron (artificial neural networks) with the back-propagation algorithm using the biomarker data. As a result, a new and robust artificial neural network model for predicting the anti-inflammatory activity of natural compounds was obtained, resulting in a high percentage of correct predictions (81%), high precision (100%) for dual inhibition, and low error values (mean absolute error = 0.3), as also shown in the validation test. Thus, the biomarkers of the Asteraceae extracts were statistically correlated with their anti-inflammatory activities and can therefore be useful to predict new anti-inflammatory extracts and their anti-inflammatory compounds using only liquid chromatography-mass spectrometry data.

Abbreviations

AI:	anti-inflammatory
ANN:	artificial neural networks
COX-1:	cyclooxygenase-1
DNP:	Dictionary of Natural Products
ID:	identification number
5-LOX:	5-lipoxygenase
MSA:	multivariate statistical analysis
MAE:	mean absolute error
NSAID(s):	nonsteroidal anti-inflammatory drug(s)
O2PLS:	orthogonal-orthogonal partial least squares
PCA:	principal component analysis
PLS:	partial least squares
RT:	retention time

Supporting information available online at http://www.thieme-connect.de/products

Introduction

The study of low-molecular-weight metabolites of organisms or biological matrices by comprehensive and sensitive analytical techniques with the aim of understanding their system biology is termed metabolomics [1–5]. The analytical techniques that are most often used for this purpose are NMR, HPLC-MS, and GC-MS. These techniques are used alone or in combination with the aim of maximizing the number of metabolites belonging to a studied sample under a determined condition [1–5]. Although natural plant products are still the most successful source of biologically active or lead compounds [6–8], the majority of classical studies only evaluate the extracts' major compounds, usually through time-consuming isolation steps. Extracts with a higher biological activity than their isolated active constituents, such as the extract of *Artemisia annua* L. that showed greater antimalarial properties than the well-known active compound artemisinin [9–11], have already been described. In these cases, the metabolomic tools together with suitable MSA, as a non-reductionist approach aiming to find active compounds in a known target [12–14], could determine which metabolites are correlated with a certain biological property, either isolated or in mixtures [4, 5, 12, 13, 15–18].

The analytical techniques used in metabolomics generate an enormous amount of spectroscopic or/and chromatographic data about the analyzed samples, and the use of MSA is always required for data mining and visual interpretation. Generally, MSA is used to classify the samples into different groups or to determine which compounds are correlated with the investigated property or characteristics of the sample [12, 19–22]. For this purpose, the MSA techniques used most frequently are PCA, PLS, and O2PLS-DA [12, 19, 20]. Additionally, the MSA models must be properly validated so that further properties or characteristic predictions of new (uninvestigated) samples can be made based on the information about their chemical composition (metabolome).

However, data analysis becomes more problematic when a large number of variables are measured and the data sets grow in both size and complexity, or when a nonlinear relationship is expected. Within this context, machine learning methods, and more specifically classification algorithms, are promising alternatives to address these issues [23]. For example, the genetic search algorithm [24–27] has been used to select variables prior to classification [25, 28–30].

The J48 Weka classifier [31] is a simple but powerful supervised MSA approach that uses an algorithm (C4.5) and builds decision trees from a set of training data [27, 32, 33]. More specifically, at each node of a tree, the C4.5 chooses the attributes (variables) of the data that most effectively split its set of samples into subsets enriched in one class [32, 33]. At the conclusion of the process, the attributes selected on the tree nodes can be the properties or characteristics of the class, such as the samples' (bio)markers. Although the J48 decision tree is a well-known and established technique that has been used in several cases of data analysis [34, 35], thus far it has not been used in metabolomic studies to determine biomarkers (i.e., potentially bioactive compounds) in plant extracts. One of the great advantages of the J48 decision tree among other machine learning methods is that the results are very easy to interpret [32], and once the model is built, the software is no longer necessary. With new data from a new sample, it is possible to predict the samples' class without using the software and algorithm, relying instead on the information provided by the original tree on J48. Because it expresses the results as a tree of conditions, it is possible to obtain decisions (the results and the predictions) easily by simply following these conditions, unlike the majority of MSA models (PCA, PLS, O2PLS), which usually require inserting the input data from a new sample to obtain class predictions (output data).

ANNs comprise a supervised MSA using a classification algorithm. The most common ANN is the multilayer perceptron and the most well-known and widely used learning algorithm is the back-propagation [32, 36]. A great advantage of the ANN is that it is a nonlinear MSA and therefore works well when a complex or nonlinear relationship is expected. Furthermore, the nonlinear MSA is usually superior in classifying samples than the linear MSA, and it is therefore useful for making predictions concerning the properties of complex samples [19, 20, 32]. A disadvantage of nonlinear MSA is that the determination of potential (bio) markers is complicated [19]. In this context, it can be summarized that supervised methods such as decision trees are very useful and suitable for determining (bio)markers, while supervised methods such as ANNs are better for predictions provided that the validation procedure is carried out properly.

Several plant species from the Asteraceae family have been systematically investigated in our laboratory using LC-MS, and the information concerning the direct inhibition of COX-1 and 5-LOX using their extracts was obtained (**Table 1**). The COX and LOX enzymes are very important in inflammatory processes and, therefore, dual inhibitors on COX-1 and 5-LOX should perform as AI medicines with a higher efficacy and lower side effects than the currently available NSAIDs [37–42].

The extracts from the Asteraceae species that were systematically evaluated in this study comprise 57 species from a diverse evolution degree, i.e., species from different genera and tribes within the family (**Table 1**). Because these extracts cover substantial metabolome diversity, prediction models created with this set of data would be comprehensive, at least for the Asteraceae species, one of the largest and most widespread families in plant kingdom [43]. In fact, metabolomic studies and MSA have been applied to guide the discovery of the biomarkers of some organisms (fungi, bacteria, or plants); however, in most of these studies, only one species (different extracts) has been evaluated [44–50]. So far, only one classification study has used ANNs to predict antioxidant activity in essential oils from food plants [51], while the majority of classification studies seek to find diagnostic predictions of human diseases [19, 34, 52]. Thus, studies involving general models to predict new active extracts using LC-MS-based metabolomics and MSA as reported in this work are new.

Therefore, we report herein the use of machine learning methods to determine the biomarkers with AI potential (dual inhibition of COX and LOX) in extracts of 57 Asteraceae species. Moreover, models to detect biomarkers in new extracts based solely on their metabolomic data, and with no prior knowledge of biological data, were also established. This combined strategy that uses metabolomics, *in vitro* bioactivity, decision trees, and ANNs has never before been explored; therefore, the time- and money-consuming steps usually required for compound isolation, identification, and AI evaluation are avoided.

Table 1 Extracts and substances from 57 Asteraceae species tested in the COX-1 and 5-LOX bioassay. IC_{50} values < 40 µg/mL were considered active (+).

Species	Sample code	Tribe	5-LOX (µg/mL)	COX-1 (µg/mL)
Achillea millefolium Ledeb. [yarrow]	1	Anthemideae Cass.	(−)	(−)
Achyrocline satureioides (Lam.) DC. [macela]	2	Gnaphalieae (Cass.) Lecoq & Juill.	(−)	(−)
Acmella oleracea (L.) R. K. Jansen [toothache plant]	3	Heliantheae Cass.	(−)	(−)
Ageratum conyzoides L. [billygoat-weed]	4	Eupatorieae Cass.	(−)	(−)
Anteremanthus hatschbachii H. Rob.	5	Vernonieae Cass.	(−)	(−)
Arctium lappa L. [greater burdock]	6	Cynareae Less.	(+)	(−)
Arnica montana L. [arnica]	8	Heliantheae Cass.	(+)	(−)
Artemisia absinthium L. [wormwood]	9	Anthemideae Cass.	(+)	(−)
Artemisia annua L. [sweet wormwood]	10	Anthemideae Cass.	(−)	(−)
Baccharis dracunculifolia D. C. [alecrim do campo]	12	Astereae Cass.	(−)	(−)
Baccharis trimera (Less.) DC. [carqueja]	13	Astereae Cass.	(−)	(−)
Bidens pilosa L. [beggar-ticks]	14	Coreopsideae Lindl.	(−)	(−)
Calea cuneifolia DC.	15	Neurolaeneae Rydb.	(−)	(−)
Calendula officinalis L. [marigold]	16	Calenduleae Cass.	(−)	(+)
Chronopappus bifrons (DC. ex Pers.) DC.	18	Vernonieae Cass.	(−)	(+)
Cichorium intybus L. [chicory]	19	Cichorieae Lam. & DC.	(+)	(+)
Cynara scolymus L. [artichoke]	20	Cardueae Cass.	(−)	(+)
Dasyphyllum brasiliense var. *latifolium* (*D. Don*) *Cabrera* [espinho agulha]	21	BarnadesieaeD.Don	(−)	(+)
Echinacea purpurea (L.) Moench [purple coneflower]	22	Heliantheae Cass.	(+)	(−)
Emilia sonchifolia L. DC [lilac tassel flower]	23	Senecioneae Cass.	(−)	(−)
Eremanthus polycephalus (DC.) MacLeish	24	Vernonieae Cass.	(−)	(+)
Helianthus annuus L. [sunflower]	26	Heliantheae Cass.	(−)	(−)
Heterocoma gracilis Loeuille, J. N. Nakaj. & Semir	27	Vernonieae Cass.	(−)	(+)
Lactuca sativa L. [common lettuce]	28	Cichorieae Lam. & DC.	(+)	(−)
Lychnophora diamantinana Coile & S. B. Jones	29	Vernonieae Cass.	(+)	(−)
Lychnophora ericoides Mart. [arnica da serra]	33	Vernonieae Cass.	(+)	(−)
Lychnophora tomentosa (Mart. ex DC.) Sch. Bip.	34	Vernonieae Cass.	(+)	(−)
Matricaria chamomilla L. [chamomile]	35	Anthemideae Cass.	(+)	(−)
Mikania glomerata Sprengl. [guaco]	37	Eupatorieae Cass.	(+)	(−)
Mikania hirsutissima DC. [cipó cabeludo]	38	Eupatorieae Cass.	(+)	(−)
Mikania laevigata Schultz Bip. ex Baker [guaco]	39	Eupatorieae Cass.	(+)	(−)
Minasia scapigera H. Rob.	40	Vernonieae Cass.	(+)	(+)
Piptolepis monticola Loeuille	41	Vernonieae Cass.	(+)	(+)
Prestelia eriopus Sch. Bip.	42	Vernonieae Cass.	(+)	(+)
Pluchea quitoc D. C.	43	Inuleae Cass.	(−)	(−)
Smallanthus sonchifolius (Poepp. & Endl.) H. Robinson [yacon]	45	Heliantheae Cass.	(−)	(+)
Solidago microglossa DC. [arnica do campo]	46	Astereae Cass.	(+)	(+)
Sonchus oleraceus L. [sowthistle]	48	Cichorieae Lam. & DC.	(+)	(−)
Sphagneticola trilobata (L.) Pruskei	49	Heliantheae Cass.	(+)	(+)
Stevia rebaudiana (Bertoni) Bertoni [sweetleaf]	50	Eupatorieae Cass.	(+)	(−)
Tridax procumbens L. [tridax daisy]	51	Heliantheae Cass.	(+)	(−)
Tanacetum parthenium L. (feverfew)	53	Anthemideae Cass.	(+)	(−)
Tanacetum vulgaris L. [tansy]	54	Anthemideae Cass.	(−)	(−)
Taraxacum officinale Weber ex FH Wigg. [dandelion]	55	Cichorieae Lam. & DC.	(−)	(+)
Tithonia diversifolia (Hemsl.) A. Gray [tree marigold]	56	Heliantheae Cass.	(+)	(+)
Vernonia condensata Baker [boldo baiano]	25	Vernonieae Cass.	(−)	(−)
Vernonia herbacea (Vell.) Rusby	57	Vernonieae Cass.	(+)	(+)
Vernonia platensis (Spreng.) Less.	58	Vernonieae Cass.	(+)	(+)
Vernonia polyanthes Less. [assa peixe]	59	Vernonieae Cass.	(+)	(+)
Vernonia rubriramea Mart. Ex DC.	60	Vernonieae Cass.	(+)	(+)
Viguiera arenaria Baker	61	Heliantheae Cass.	(+)	(−)
Viguiera bracteata Gardner	62	Heliantheae Cass.	(+)	(−)
Viguiera discolor Baker	63	Heliantheae Cass.	(+)	(−)
Viguiera filifolia Sch. Bip. Ex Baker	64	Heliantheae Cass.	(+)	(−)
Viguiera linearifolia Chodat & Hassl.	65	Heliantheae Cass.	(−)	(−)
Viguiera robusta Gardner	66	Heliantheae Cass.	(+)	(+)
Viguiera trichophylla Dusén	67	Heliantheae Cass.	(+)	(+)
Reference inhibitors (RI)				
Indomethacin		RI of COX-1	(−)	(+)
Nordihydroguaiaretic acid		RI of 5-LOX	(+)	(−)

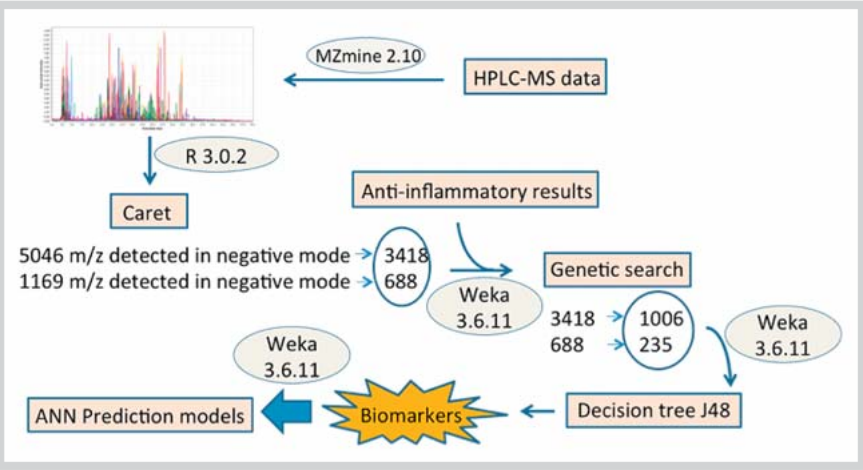

Fig. 1 Combined strategy for the treatment of LC-MS-based metabolomic data and machine learning studies to determine biomarkers and build prediction models using ANNs. (Color figure available online only.)

Results and Discussion

The LC-MS data for 57 samples were aligned and deconvoluted, obtaining a total of 6052 peaks (1169 in the positive and 5046 in the negative mode of detection). The missing or irrelevant values that were considered to be on the noise threshold that could disturb the MSA [53] were cleaned using a proper algorithm [54], thus reducing the data set to 4106 peaks (3418 in the negative and 688 in the positive mode). A further selection of relevant variables according to their respective AI properties was carried out by a genetic algorithm; as a result, only 1241 peaks (1006 in the negative and 235 in the positive mode) were considered to be relevant data (**Fig. 1**).

This reduced data set of the extracts was analyzed by the linear supervised MSA J48 decision tree. This decision tree algorithm chooses the variables of the data that most effectively split its set of samples into subsets enriched in one class [32,33], which in our case correspond to AI properties. Because our aim was to determine the biomarkers with AI properties using supervised MSA, both the AI results and the selected metabolomic data for all extracts were used as input data. After the decision tree models were built (**Fig. 1**), it was possible to determine 11 biomarkers with AI properties (the nodes of the decisions tree, **Fig. 2**) in the extracts (*b* = dual inhibitors of COX-1 and 5-LOX enzymes, *c* = inhibitors of COX-1 only, and *n* = non-inhibitors of both enzymes, **Fig. 2**). The statistical values of the decision trees showed that the models are relevant, with R^2 (goodness of fit) > 0.9 and Q^2 (goodness of prediction) > 0.6. According to previously published data on the use of MSA in metabolomics, values of R^2 and Q^2 > 0.5 are considered significant for analysis of complex data [12,22,55,56], meaning that the J48 decision tree was suitable to classify the 57 Asteraceae extracts according to their AI properties, as well as to point to relevant biomarkers.

Decision trees comprise a simple technique that is rich in information, showing results completely and clearly and in a format that is easy to understand (**Fig. 2**). For example, examining the decision tree in **Fig. 2**, it is possible to observe that four of the bioactive extracts coded in parentheses as (b) (dual inhibitors) have the peak ID = 2610 (*m/z* 447.1306 on RT: 17.5 min; **Table 2**) with a peak area > 4080.65, and the peak ID = 3144 (*m/z* 513.2714 on RT: 19.0 min; **Table 2**) with a peak area > 85.07, while six COX-1 inhibitors have the same peak ID = 2610 with a peak area > 4080.65, but a peak area < 85.07 for the peak ID = 3144. Thus, when the peak ID = 2610 occurs in the extracts with a peak area > 4080.65 concomitantly with the peak ID = 3144 with a peak area > 85.07, it means that this extract is probably a dual inhibitor, as these two biomarkers are common in the extract, while non-dual inhibitor extracts do not meet these criteria.

Following the same argument, the peak IDs 3144 (*m/z* 513.2714 on RT: 19.0 min with a peak area > 85.07), 815 (*m/z* 509.2246 on RT: 16.9 min with a peak area > 3 833 191.55), and 829 (*m/z* 411.1797on RT: 25.3 min with a peak area < 128.33) on the metabolomic data detected in the negative mode, and the peak IDs 1637 (*m/z* 349.1643 on RT: 28.2 min with a peak area > 310.21), 1116 (*m/z* 600.2654 on RT: 16.8 min with a peak area > 9230.05), and 692 (*m/z* 449.1076 on RT: 19.0 min with a peak area > 2 856 584.62) on the metabolomic data detected in the positive mode are the biomarkers of dual inhibitor extracts (**Fig. 2**, **Table 2**).

Therefore, according to this procedure, without having performed any time- or money-consuming COX and LOX enzymatic assays, and using only LC-MS metabolomic data, it is possible to predict whether a nonbiologically evaluated extract displays the AI property of dual inhibition of the COX-1 and 5-LOX. One necessary condition to fulfill these criteria is that the metabolomic data must be obtained under the same chromatographic and MS conditions. Thus, the only work is to check the peak areas of the biomarkers selected, which is displayed by the decision trees (**Fig. 2**).

It should also be noted in **Fig. 2** that only one non-dual inhibitor extract was incorrectly classified as a dual inhibitor in the decision tree created using metabolomic data detected in the positive mode (R^2 = 0.9), while all of the data were classified correctly in the decision tree made with the metabolomic data detected in the negative mode (R^2 = 1.0). This result explains the differences in the corresponding R^2 for the data obtained in the positive (perfect fit of the MSA model) and negative modes. Regardless, errors were rare and both models were well fitted.

In fact, the information provided by the decision trees can also be useful to make predictions when a model is well fitted and also contains good statistical values for the cross-validation, as shown in our study (**Fig. 2**). However, carrying out an external validation is always recommended [56,57]. Therefore, because the objective of the model using decision trees reported herein was solely to find the biomarkers that show dual inhibition for COX-1 and 5-LOX using the most comprehensive data possible, no data were left out to be used as a group for external validation. Be-

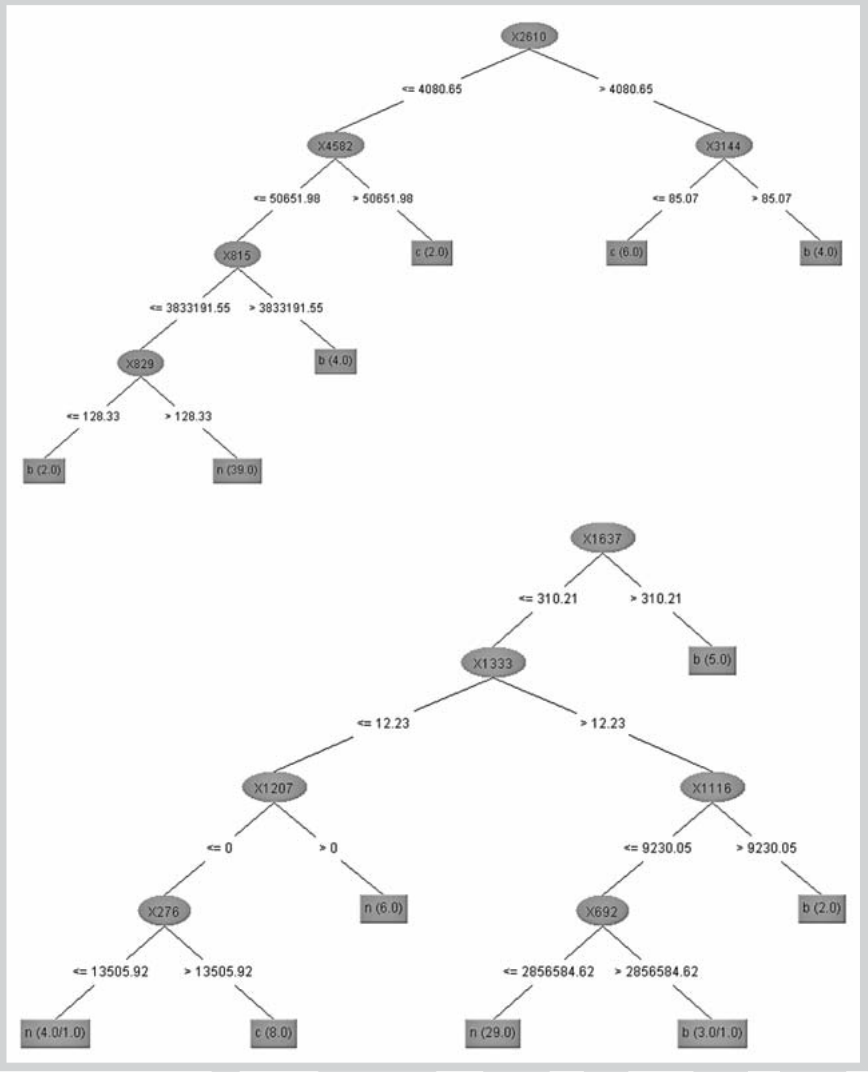

Fig. 2 J48 decision tree using the anti-inflammatory results and the metabolomic data of detection in the negative mode (figure above, $R^2 = 1.0$ and $Q^2 = 0.7$) or the positive mode (figure below, $R^2 = 0.9$ and $Q^2 = 0.6$) as input data. The numbers inside the circles (the nodes of the decision tree) after the X are the IDs of the peaks that have their *m/z*, peak area, and RT values associated. These IDs correspond to the biomarkers with anti-inflammatory properties; (**b**) are extracts with dual inhibition of the COX-1 and 5-LOX enzymes, (**c**) are extracts that inhibit only the COX-1 enzyme, and (**n**) are the extracts unable to inhibit both enzymes. The number between the parentheses next to the anti-inflammatory properties codes (**b, c**, and **n**) provides the information on how many extracts were classified as anti-inflammatory correctly or incorrectly. The values between the nodes and the outputs correspond to the peak area of the biomarkers.

cause predictions are also very important in this type of approach, we decided to build predictive models using ANNs (with proper external validation), which are suitable to handle and predict complex data [19, 20, 32].

As ANNs are prone to overfitting when the input data include several variables [19, 20, 36], not all metabolomic data were used in this step. For the input data to build the models (training), only the peak areas of the biomarkers previously determined by the decisions trees, as well as the AI results of selected extracts, were used. As a result, the Q^2 values obtained for the ANN models (**Fig. 1S**, Supporting Information; **Table 3**) were even higher than those obtained for the decision trees (**Fig. 2**), which means that in this case, the ANNs were better at predicting new AI extracts based on their metabolomic data than the decisions trees, as expected. Additionally, the values of external validation (**Table 3**) showed that the ANN models are robust to make predictions and are not overfitted. They showed a high percentage of correct predictions (> 81%), a high precision for dual inhibition (100%), and low error values (MAE < 0.3).

None of the non-dual inhibitor extracts were predicted as dual inhibitors, even those that were able to inhibit only the COX-1 or the 5-LOX enzymes (external validation group 3, **Table 3**). Extracts that were able to inhibit only one of the two enzymes were included in the third validation group and were always predicted as non-inhibitors, thus showing that the model does not provide false positive outputs. Thus, the ANN models are suitable to predict new dual inhibitor extracts, which are highly likely to actually be dual inhibitors, although some false negative results may always occur. False negative results explain 19% of incorrect predictions.

The good results obtained by the ANN models corroborate the biomarkers determined by the decision trees as good biomarkers for AI properties, and therefore can be useful to predict new AI extracts as well as their AI compounds. However, it should be kept in mind that the ANN models were validated using new extracts only from the Asteraceae family. Nevertheless, we believe that they are also able to deal with extracts from other plant families provided such extracts contain (or not) the biomarker(s) indicated by the decision trees that have been used to train the ANN models. A limitation of these models is that they do not have the ability to deal with previously unknown bioactive compounds. In summary, the ANN models determine the presence or absence of biomarkers in the extracts taking into account the chromatographic peak areas of the compounds and do not make predictions like the conventional models based on chemical structures and their calculated descriptors.

The identification of the chemical structures of the biomarkers present in the extracts was not specific enough using known der-

Table 2 Biomarkers with anti-inflammatory properties selected by decision trees (J48) in the metabolomic data obtained in the negative and positive modes.

ID	*m/z*	RT	Adducts	Error (ppm)	Molecular formula	Hits on dictionary of natural products
Negative mode						
815	509.2246000	16.9	$[M-H]^-$	1.248	$C_{22}H_{38}O_{13}$	1*: β-D-glucopyranose,1-[8-(β-D-glucopyranosyloxy)-2,6-dimethyl-2-octenoate],(6*R*-2*E*)
829	411.1797000	25.3	$[M+Cl]^-$	1.413	$C_{18}H_{34}O_8$	1: D-glucopyranose,2,3-dihexanoate
2610	447.1305870	17.5	$[M-H]^-$	2.051	$C_{22}H_{24}O_{10}$	59: phenolic compounds, flavonoids, and isocoumarins; ex.: the phenolic compound allo-inositol,1,4-bis(4-hydroxybenzeneacetate) occurs in Asteraceae (e.g., *Taraxacum linearisquameum* Soest) and 6′-*O*-trans-caffeoylsalicin occurs in other families
3144	513.2714390	19.0	$[M-H]^-$	1.789	$C_{26}H_{42}O_{10}$	15*: diterpenes
4582	470.9872900	12.4	$[M-H]^-$	1.023	$C_{23}H_6O_{11}N$	0
Positive mode						
276	415.2112519	34.5	$[M+H]^+$	0.634	$C_{24}H_{30}O_6$	43**: sesquiterpene lactones and diterpenes
692	449.1075653	19.0	$[M+H]^+$	0.607	$C_{21}H_{20}O_{11}$	101: flavonoids; ex.: quercetrin*** occurs in Asteraceae (in the dual inhibitor extract of *Solidago microglossa* #46 [61]); luteolin,7-β-D-glucopyranoside is widespread in plants and occurs in Asteraceae
1116	600.2654479	16.8	$[M+NH_3]^+$	0.635	$C_{28}H_{38}O_{13}$	0
1333	270.1696844	29.1	$[M+NH_3]^+$	1.112	$C_{14}H_{20}O_4$	1: butanoic acid,3-methyl-,4-(1*R*,2*S*-dihydroxypropyl)phenyl ester
1207	509.1290488	20.6	$[M+H]^+$	0.160	$C_{23}H_{24}O_{13}$	39: pinacetin 3-*O*-β-D-glucopyranoside occurs in Asteraceae (*Artemisia absinthium* #9)
1637	349.1642519	28.2	$[M+H]^+$	0.897	$C_{19}H_{24}O_6$	162: sesquiterpene lactones; ex.: many of them have been reported in Asteraceae, and two of them were described for a species whose extract showed dual inhibition (*Tithonia diversifolia* #56): tagitinin C*** and tagitinin F

The dual inhibitor biomarkers are in bold; *none of them were previously described for species from the Asteraceae family; **none of them were previously described for species from family the Asteraceae family; however, sesquiterpene lactones and diterpenes are widespread within the family [62]; ***confirmed by coinjection with authentic standards

Table 3 Statistical values of the prediction model using ANNs and its validations.

	Metabolomic data	R^2	Q^2	MAE	External validation 1	External validation 2	External validation 3
ANN	Negative mode	0.81	0.72	0.30	0.77	0.57	0.54
	Positive mode	0.90	0.81	0.20	0.66	0.85	0.90

Training group: non-inhibitors (n) = 1, 10, 12, 14, 23, 43, 55; dual inhibitors (b): 40, 41, 56, 57. Testing group 1: n = 13, 15, 26, 25, 53, 52; b: 66, 67, 49, 46. Testing group 2: n = 13, 25; b: 58, 60, 19, 42, 46. Testing group 3: n = 13, 15, 25; LOX inhibitors: 9, 29, 22; COX inhibitor: 24; b: 19, 42, 46

eplication strategies [3, 16, 58] along with a comprehensive database (**Table 2**). Some biomarkers had more than one hit according to the database; in addition, two biomarkers yielded zero hits, suggesting that they may be new compounds (**Table 2**). Most of the standards are unavailable to validate the results according to RT, and MS/MS experiments were not performed, as the MS^1 levels for these metabolites were not intensive enough to generate reliable MS^2 data. **Fig. 2** shows that such biomarkers correspond to minor compounds with low peak areas in the extracts, which could explain why these biomarkers have not yet been isolated; however, their isolation is currently in progress because these compounds have been shown to be strongly correlated with their biological properties. In this light, their prediction, structure identification, and the confirmation of their AI properties seem to be very important. Nevertheless, two compounds that have been reported earlier to show AI activity [59], the flavonoid glycoside quercetrin (ID 692, **Table 2**) and the sesquiterpene lactone tagitinin C [60] (ID 1637, **Table 2**), were chromatographed under the same analytical conditions (LC and MS) and further detected by coinjection of *Solidago microglossa* (code 46,

Table 1) and *Tithonia diversifolia* (code 56, **Table 1**), two bioactive extracts that showed dual inhibition. Thus, this data corroborates the results obtained by the decision trees.

Based on the results, the combined strategy proposed herein, which comprises suitable treatment of LC-MS-based metabolomic data from 57 Asteraceae extracts as well as the use of machine learning algorithms combined with obtained biological data, allowed us to determine 11 biomarkers with AI potential, i.e., dual inhibition of COX-1 and 5-LOX enzymes. Decision trees were used for the first time to identify biomarkers in plant extracts. Additionally, it was possible to build robust models using ANNs that display good validation values and a high percentage of correct predictions and precision. The decision trees and ANN models reported herein can therefore be used to predict and screen for AI potential of unknown extracts based on their LC-MS metabolomic profile, thus eliminating the time- and money-consuming steps of AI assays and compound isolation and/or purification, at least for Asteraceae species. The identification of the biomarkers' chemical structures using comprehensive databases (e.g., DNP that was used in this study) may not be specific enough; however, independent of the identification of the biomarkers, using decision trees and ANN models, it will be possible to predict the biologically active extracts knowing only peak areas on the LC-HRMS profile data, provided those data are obtained under the same experimental conditions. Thus, this strategy can be useful in holistic studies that seek to find biomarkers with a certain determined property or characteristic of complex samples (such as the extracts) and also to reveal active secondary metabolites.

Materials and Methods

Plant material and extracts

Selected species for this study are summarized in **Table 1**. The purchased powdered species have the corresponding documents to attest to their authenticities and are as follows: #2 – Sítio da Mata, Cajuru-SP; #3 – Market, Belém-PA; #4 – Sítio Irmas Maries, Jardinópolis-SP; #8 – Santos Flora, São Paulo-SP; #19 – Market, Ribeirão Preto-SP; #50 – Sítio da Mata, Cajuru-SP. The donated samples have been deposited in the Herbarium of the Institute of Biosciences (SPF), University of São Paulo, SP, Brazil and are as follows (extract # according to **Table 1**): #1 – A.M.S. Pereira 1428; #5 – Loeuille et al. 537; #6 – Sítio Irmas Maries, Jardinópolis-SP; #9 – Sítio Irmas Maries, Jardinópolis-SP; #10 – Sítio Irmas Maries, Jardinópolis-SP; #12 – Sítio Irmas Maries, Jardinópolis-SP; #13 – A.M.S. Pereira 1426; #14 – A.M.S. Pereira 1424; #15 – M. Nogueira & L.E. Gregório 35; #16 – A.M.S. Pereira 1430; #18 – Lusa et al. 63; #20 – Sítio Irmas Maries, Jardinópolis-SP; #21 – D.A. Chagas-Paula 10; #22 – A.M.S. Pereira 1425; #23 – D.A. Chagas-Paula 06; #24 – Loeuille et al. 531; #25 – A.M.S. Pereira 1418; #26 – D.A. Chagas-Paula 09; #27 – Lusa et al. 61; #28 – D.A. Chagas-Paula 08; #29 – Loeuille et al. 530; #33 – Sítio Irmas Maries, Jardinópolis-SP; #34 – Loeuille et al. 528; #35 – Sítio Irmas Maries, Jardinópolis-SP; #37 – Sítio Irmas Maries, Jardinópolis-SP; #38 – A.M.S. Pereira 1400; #39 – A.M.S. Pereira 1419; #40 – Loeuille et al. 529; # 41 – Lusa et al. 62; #42 – Loeuille et al. 524; #43 – A.M.S. Pereira 1427; #45 – A.M.S. Pereira 1422; #46 – A.M.S. Pereira 1421; #48 – A.M.S. Pereira 1429; #49 – D.A. Chagas-Paula 05; #51 – D.A. Chagas-Paula 07; #53 – Sítio Irmas Maries, Jardinópolis-SP; #54 – A.M.S. Pereira 1420; #55 – Sítio Irmas Maries, Jardinópolis-SP; #56 – A.M.S. Pereira 1423; #57 – E.S.A. 94148; #58 – E.S.A. 94146; #59 – Sítio Irmas Maries, Jardinópolis-SP; #60 – Sítio Irmas Maries, Jardinópolis-SP; #61 – M. Magenta et al. 275; #62 – M. Magenta et al. 440; #63 – M. Magenta et al. 307; #64 – A.B. Bombo et al. 56; #65 – A.B. Bombo et al. 62; #66 – M. Magenta et al. 454; #67 – A.B. Bombo et al. 51.

This project has authorization from the Genetic Heritage Management Council (CNPq, process #010055/2012-6).

Dried plant materials were ground, and 20 mg of each powder was extracted with 2 mL of EtOH-H2O (7 : 3, v/v) in an orbital shaker (110 rpm and 30 °C) for 24 h, partitioned with n-hexane, and filtered through a 0.2-µm PTFE membrane (Millipore©). The solvent was eliminated in a rotary evaporator, and the material was split into two aliquots and used for analysis of both the LC-MS and AI assays.

Anti-inflammatory assays

The extracts were systematically evaluated using COX-1 (catalog #560101 Cayman Chemical's ACE™) and 5-LOX (catalog #760700 and #60401) screening kits. Reference inhibitors for COX-1 and 5-LOX were indomethacin (≥ 99% purity, Sigma-Aldrich®) and nordihydroguaiaretic acid (≥ 97% purity, Sigma-Aldrich®), respectively.

High-performance liquid chromatography-mass spectrometry-based metabolomic study

HPLC-HRMS data of the biomarker compounds for 57 plant extracts were obtained as shown in **Tables 1** and **2**. The MS data were acquired using Thermo Fisher Scientific Exactive™ equipment powered by Orbitrap™ technology coupled to an HPLC system (Accela, Thermo Fisher Scientific). The parameters of the analysis can be found in the Supporting Information. The raw data (the total ion current chromatogram along with the respective mass spectral data and RT) were sliced into individual positive and negative data sets using the RecalOffline tool from Xcalibur 2.1 (Thermo Fisher Scientific) before importing them (in. mzML format) into MZMine 2.10 software (MZmine Development Team) for further data treatment. Quercetrin and tagitinin C used in the coinjections were previously isolated in our laboratory, and their spectral data are available upon request.

High-performance liquid chromatography-mass spectrometry data treatment

MZmine 2.10 is an open-source software used to perform *peak detection* (mass detection = exact mass: noise level = 10000 or 1.0E4; filtering = FTMS shoulder peaks filter; mass resolution = 50000), *chromatogram building* [min time span = 0.3 min (18 s); min height = 5.0E4 (50000); *m/z* tolerance = 0.001], *chromatogram deconvolution* [peak recognition = local minimum search; chromatographic threshold = 5%; search minimum in RT range = 0.2 min (12 s); minimum relative height = 15%; minimum absolute height = 50000 (5.0E4); min ration of peak top/edge = 5; peak duration range = 0.3–10.0 min], *deisotoping* [*m/z* tolerance = 0.001; RT tolerance = 0.1 absolute (6 s); maximum charge = 2; representative isotope = most intense)], *alignment* [join aligner: *m/z* tolerance = 1 ppm; RT tolerance type = absolute; absolute RT tolerance = 0.5 min (30 s); weight for *m/z* = 15 and for RT= 10], and *gap filling* (intensity, *m/z* and RT tolerance = 1%; 0.001 *m/z* and 30 s, respectively). After removing a few non-peak-shape features through a visual survey, both data sets were exported from MZMine 2.10 as a. csv file for further statistical analysis (**Fig. 1**).

This file contains ID, *m/z*, RT, and area of each peak. The data from this file were additionally cleaned using the standard parameters of the package Caret in the free software R 3.0.2 (R Development Core Team). The function nearZeroVar was used to eliminate the near zero variance predictors in the data set and variables that were unique or with a skewed frequency of distribution [53]. After the missing or irrelevant values had been cleaned from the data set, the selection of the attributes important to the classes was made by a genetic search. The genetic search is a method that can be used to find the best set of variables for ANN analysis [24]. The parameters for making the attribute selection by genetic algorithm in the Weka software 3.6.11 (Machine Learning Group at the University of Waikato) were: Attribute Evaluator = CfsSubsetEval, Search Method = GeneticSearch, PopulationSize = Z = 50, MaxGenerations = G = 100, CrossoverProb = C = 0.6, MutationProb = M = 0.033, ReportFrequency = R = 20, and Seed = S = 1 (**Fig. 1**).

Machine learning

After data treatment, the J48 decision trees were built in Weka using the following parameters: Classifier = Trees – J48, BinarySplits = False, ConfidenceFactor = C = 0.25, Debug = False, MinNumObj = 2, NumFolds = 3, ReducedErrorPruning = False, SaveInstanceData = False, Seed = 1, SubtreeRaising = True, Unpruned = False, and UseLaplace = False (**Fig. 1**).

The same procedures were used for the LC-MS data detected in the positive and negative modes. In the data analysis using ANNs, however, these parameters were different for each type of data.

The prediction models using ANNs were also built in Weka using the data set containing only the variables revealed by the decision trees: IDs 1637, 1333, 1207, 1116, 276, and 692 in the data detected in the positive mode, and IDs 2610, 4582, 3144, 815, and 829 in the data detected in the negative mode.

The parameters for the ANN models using data from the positive mode were: Classifier = MultilayerPerceptron, Gui = False or True, AutoBuild = True, Debug = False, Decay = False, HiddenLayers = H = 4, LearningRate = L = 0.3, Momentum = M = 0.2, NominalToBinaryFilter = True, NormalizeAttributes = True, NormalizeNumericClass = True, Reset = False, Seed = S = 0, TrainingTime = N = 500, ValidationSetSize = V = 0, and ValidationThreshold = E = 20. The parameters for the ANN models using data from the negative mode were the same as earlier described, with the exception of HiddenLayers = H = 3, Learning Rate = L = 0.25, and Momentum = M = 0.25 (**Fig. 1** and **Fig. 1S**).

Other parameters were also evaluated but those described here provided the models with the best statistical values.

Compound identification

The following parameters were used in MZmine 2.10 for peak identification according to their *m/z* in high resolution: 1) adduct search for the data from mass detection in the HR ESI positive mode $[M + Na]^+$, $[M + K]^+$, $[M + NH_3]^+$, and $[M + MeCN + H]^+$, and $[M + Cl]^-$ in the HR ESI negative mode; RT tolerance = 0.1 min (6 s), *m/z* tolerance = 0.001, max adduct peak height = 50%; 2) complex search: ionization method = +H for data detected in the positive mode and -H for those detected in the negative mode; RT tolerance = 0.1 min (6 s), *m/z* tolerance = 0.001, max complex peak height = 50%.

The molecular formulae of the biomarkers were determined using Xcalibur according their *m/z* in high resolution considering an error below 3 ppm. The identification procedure was carried out by searching for the molecular formulae in DNP and filtering those compounds that had been previously described for the Asteraceae species (i.e., a chemotaxonomic filter), when it was possible.

Supporting information

HPLC-HRMS parameters and ANN models built with data detected in the negative and positive modes are available as Supporting Information.

Acknowledgements

The authors thank FAPESP (grants #2010/10940–2 and #2010/51454–3) for financial support as well as CNPq and CAPES for funding and grants. We also thank Prof. J.K. Bastos (FCFRP-USP) for providing us with pure quercetrin and Prof. A.M.S. Pereira (UNAERP) for several samples of plant material.

Conflict of Interest

The authors declare no conflict of interest.

References

1 *Villas-Bôas SG, Mas S, Akesson M, Smedsgaard J, Nielsen J.* Mass spectrometry in metabolome analysis. Mass Spectrom Rev 2005; 24: 613–646
2 *Weckwerth W, Morgenthal K.* Metabolomics: from pattern recognition to biological interpretation. Drug Discov Today 2005; 10: 1551–1558
3 *Wolfender J, Marti G, Ferreira Queiroz E.* Advances in techniques for profiling crude extracts and for the rapid identification of natural products: dereplication, quality control and metabolomics. Curr Org Chem 2010; 14: 1808–1832
4 *Sumner LW, Mendes P, Dixon RA.* Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry 2003; 62: 817–836
5 *Bedair M, Sumner LW.* Current and emerging mass-spectrometry technologies for metabolomics. Trends Anal Chem 2008; 27: 238–250
6 *Newman DJ.* Natural products as leads to potential drugs: an old process or the new hope for drug discovery? J Med Chem 2008; 51: 2589–2599
7 *Newman DJ, Cragg GM.* Natural products as sources of new drugs over the last 30 years from 1981 to 2010. J Nat Prod 2012; 75: 311–335
8 *Harvey AL.* Natural products in drug discovery. Drug Discov Today 2008; 13: 894–901
9 *De Donno A, Grassi T, Idolo A, Guido M, Papadia P, Caccioppola A, Villanova L, Merendino A, Bagordo F, Fanizzi FP.* First-time comparison of the *in vitro* antimalarial activity of *Artemisia annua* herbal tea and artemisinin. Trans R Soc Trop Med Hyg 2012; 106: 696–700
10 *Phillipson JD.* Phytochemistry and medicinal plants. Phytochemistry 2001; 56: 237–243
11 *Wagner H.* Synergy research: approaching a new generation of phytopharmaceuticals. Fitoterapia 2011; 82: 34–37
12 *Yuliana ND, Khatib A, Choi YH, Verpoorte R.* Metabolomics for bioactivity assessment of natural products. Phyther Res 2011; 25: 157–169
13 *Verpoorte R, Choi YH, Kim HK.* Ethnopharmacology and systems biology: a perfect holistic match. J Ethnopharmacol 2005; 100: 53–56
14 *Abdelmohsen UR, Cheng C, Viegelmann C, Zhang T, Grkovic T, Ahmed S, Quinn RJ, Hentschel U, Edrada-Ebel R.* Dereplication strategies for targeted isolation of new antitrypanosomal actinosporins A and B from a marine sponge associated-*Actinokineospora* sp. EG49. Mar Drugs 2014; 12: 1220–1244
15 *Shyur LF, Yang N.* Metabolomics for phytomedicine research and drug development. Curr Opin Chem Biol 2008; 12: 66–71
16 *Lang G, Mayhudin NA, Mitova MI, Sun L, van der Sar S, Blunt JW, Cole ALJ, Ellis G, Laatsch H, Munro MHG.* Evolving trends in the dereplication of natural product extracts: new methodology for rapid, small-scale investigation of natural product extracts. J Nat Prod 2008; 71: 1595–1599
17 *Nicholson JK, Lindon JC.* Metabonomics. Nature 2008; 455: 1054–1056

18 *Viegelmann C, Margassery LM, Kennedy J, Zhang T, O'Brien C, O'Gara F, Morrissey JP, Dobson ADW, Edrada-Ebel R.* Metabolomic profiling and genomic study of a marine sponge-associated *Streptomyces sp.* Mar Drugs 2014; 12: 3323–3351
19 *Madsen R, Lundstedt T, Trygg J.* Chemometrics in metabolomics – a review in human disease diagnosis. Anal Chim Acta 2010; 659: 23–33
20 *Trygg J, Holmes E, Lundstedt T.* Chemometrics in metabonomics. J Proteome Res 2007; 6: 469–479
21 *Trygg J, Wold S.* Orthogonal projections to latent structures (O-PLS). J Chemom 2002; 16: 119–128
22 *Bylesjö M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J.* OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J Chemom 2006; 20: 341–351
23 *Boccard J, Kalousis A, Hilario M, Lantéri P, Hanafi M, Mazerolles G, Wolfender JL, Carrupt PA, Rudaz S.* Standard machine learning algorithms applied to UPLC-TOF/MS metabolic fingerprinting for the discovery of wound biomarkers in *Arabidopsis thaliana*. Chemom Intell Lab Syst 2010; 104: 20–27
24 *Back B, Laitinen T, Sere K.* Neural networks and genetic algorithms for bankruptcy predictions. Expert Syst Appl 1996; 11: 407–413
25 *Anbarasi M, Anupriya E.* Enhanced prediction of heart disease with feature subset selection using genetic algorithm. Int J Eng Sci Technol 2010; 2: 5370–5376
26 *Leardi R.* Genetic algorithms in chemometrics and chemistry: a review. J Chemom 2001; 15: 559–569
27 *Witten IH, Frank E, Hall MA.* Data mining: practical machine learning tools and techniques, 3rd edition. Burlington: Morgan Kaufmann Publishers; 2011: 490–493
28 *Tong DL, Mintram R.* Genetic Algorithm-Neural Network (GANN): a study of neural network activation functions and depth of genetic algorithm search applied to feature selection. Int J Mach Learn Cybern 2010; 1: 75–87
29 *D'heygere T, Goethals PLM, Pauw ND.* Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecol Model 2003; 160: 291–300
30 *Karegowda AG, Manjunath AS, Jayaram MA.* Comparative study of attribute selection using gain ration and correlation based feature selection. Int J Inf Technol Knowl Manag 2010; 2: 271–227
31 *Machine Learning Group.* Class J48. Available at http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html. Accessed October 10, 2014
32 *Kotsiantis SB.* Supervised machine learning: a review of classification techniques. Informatica 2007; 31: 249–268
33 *Quinlan JR.* Improved use of continuous attributes in C4.5. J Artif Intell Res 2006; 4: 77–90
34 *Endo A, Shibata T, Tanaka H.* Comparison of seven algorithms to predict breast cancer survival. Biomed Soft Comput Hum Sci 2008; 13: 11–16
35 *Zarkami R.* Application of classification trees-J48 to model the presence of roach (*Rutilus rutilus*) in rivers. Casp J Environ Sci 2011; 9: 189–198
36 *Hecht-Nielsen R.* Theory of the backpropagation neural network. International Joint Conference on Neural Networks, Vol. 1. Washington: IEEE; 1989: 593–605
37 *Parente L.* Pros and cons of selective inhibition of cyclooxygenase-2 versus dual lipoxygenase/cyclooxygenase inhibition: is two better than one? J Rheumatol 2001; 28: 2375–2382
38 *Fiorucci S, Meli R, Bucci M, Cirino G.* Dual inhibitors of cyclooxygenase and 5-lipoxygenase. A new avenue in anti-inflammatory therapy? Biochem Pharmacol 2001; 62: 1433–1438
39 *Gaddi A, Cicero AFG, Pedro EJ.* Clinical perspectives of anti-inflammatory therapy in the elderly: the lipoxigenase (LOX)/cycloxigenase (COX) inhibition concept. Arch Gerontol Geriatr 2004; 38: 201–212
40 *Schneider I, Bucar F.* Lipoxygenase inhibitors from natural plant sources. Part 1: Medicinal plants with inhibitory activity on arachidonate 5-lipoxygenase and 5-lipoxygenase[sol]cyclooxygenase. Phyther Res 2005; 19: 81–102
41 *Rao PNP, Knaus EE.* Evolution of nonsteroidal anti-inflammatory drugs (NSAIDs): cyclooxygenase (COX) inhibition and beyond. J Pharm Pharm Sci 2008; 11: 81–110
42 *Celotti F, Durand T.* The metabolic effects of inhibitors of 5-lipoxygenase and of cyclooxygenase 1 and 2 are an advancement in the efficacy and safety of anti-inflammatory therapy. Prostag Oth Lipid M 2003; 71: 147–162
43 *Schmidt TJ.* Toxic activities of sesquiterpene lactones: structural and biochemical aspects. Curr Org Chem 1999; 3: 577–608
44 *Frédérich M, Jansen C, de Tullio P, Tits M, Demoulin V, Angenot L.* Metabolomic analysis of *Echinacea spp.* by ^{1}H nuclear magnetic resonance spectrometry and multivariate data analysis technique. Phytochem Anal 2010; 21: 61–65
45 *Li Z, Zhi H, Zhang F, Sun H, Zhang L, Jia J, Xing J, Qin X.* Metabolomic profiling of the antitussive and expectorant plant *Tussilago farfara* L. by nuclear magnetic resonance spectroscopy and multivariate data analysis. J Pharmacol Biomed Anal 2013; 75: 158–164
46 *Modarai M, Yang M, Suter A, Kortenkamp A, Heinrich M.* Metabolomic profiling of liquid *Echinacea* medicinal products with *in vitro* inhibitory effects on cytochrome P450 3A4 (CYP3A4). Planta Med 2010; 76: 378–385
47 *Bórquez J, Kennelly EJ, Simirgiotis MJ.* Activity guided isolation of isoflavones and hyphenated HPLC-PDA-ESI-ToF-MS metabolome profiling of *Azorella madreporica* Clos. from northern Chile. Food Res Int 2013; 52: 288–297
48 *Inui T, Wang Y, Pro SM, Franzblau SG, Pauli GF.* Unbiased evaluation of bioactive secondary metabolites in complex matrices. Fitoterapia 2012; 83: 1218–1225
49 *Ching J, Soh WL, Tan CH, Lee J, Tan JC, Yang J, Yap C, Koh H.* Identification of active compounds from medicinal plant extracts using gas chromatography-mass spectrometry and multivariate data analysis. J Sep Sci 2012; 35: 53–59
50 *Yuliana ND, Khatib A, Verpoorte R, Choi YH.* Comprehensive extraction method integrated with NMR metabolomics: a new bioactivity screening method for plants, adenosine A1 receptor binding compounds in *Orthosiphon stamineus* Benth. Anal Chem 2011; 83: 6902–6906
51 *Cabrera AC, Prieto JM.* Application of artificial neural networks to the prediction of the antioxidant activity of essential oils in two experimental *in vitro* models. Food Chem 2010; 118: 141–146
52 *Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001; 7: 673–679
53 *Kuhn M.* Building predictive models in R using the caret package. J Stat Softw 2008; 28: 1–26
54 *Macintyre L, Zhang T, Viegelmann C, Martinez IJ, Cheng C, Dowdells C, Abdelmohsen UR, Gernert C, Hentschel U, Edrada-Ebel R.* Metabolomic tools for secondary metabolite discovery from marine microbial symbionts. Mar Drugs 2014; 12: 3416–3448
55 *Eriksson L, Trygg J, Wold S.* CV-ANOVA for significance testing of PLS and OPLS® models. J Chemom 2008; 22: 594–600
56 *Tropsha A, Golbraikh A.* Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr Pharm Des 2007; 13: 3494–3504
57 *Golbraikh A, Tropsha A.* Beware of q2! J Mol Graph Model 2002; 20: 269–276
58 *Corley DG, Durley RC.* Strategies for database dereplication of natural products. J Nat Prod 1994; 57: 1484–1490
59 *García-Mediavilla V, Crespo I, Collado PS, Esteller A, Sánchez-Campos S, Tuñón MJ, González-Gallego J.* The anti-inflammatory flavones quercetin and kaempferol cause inhibition of inducible nitric oxide synthase, cyclooxygenase-2 and reactive C-protein, and down-regulation of the nuclear factor kappaB pathway in Chang Liver cells. Eur J Pharmacol 2007; 557: 221–229
60 *Chagas-Paula DA, Oliveira RB, Rocha BA, Da Costa FB.* Ethnobotany, chemistry, and biological activities of the genus *Tithonia* (Asteraceae). Chem Biodivers 2012; 9: 210–235
61 *Sabir SM, Ahmad SD, Hamid A, Khan MQ, Athayde ML, Santos DB, Boligon AA, Rocha JBT.* Antioxidant and hepatoprotective activity of ethanolic extract of leaves of *Solidago microglossa* containing polyphenolic compounds. Food Chem 2012; 131: 741–747
62 *Heinrich M, Robles M, West JE, Ortiz de Montellano BR, Rodriguez E.* Ethnopharmacology of Mexican Asteraceae (Compositae). Annu Rev Pharmacol Toxicol 1998; 38: 539–565