Sesquiterpene Lactones-Based Classification of the Family Asteraceae Using Neural Networks and *k*-Nearest Neighbors

Dimitar Hristozov,[†] Fernando B. Da Costa,*[,†,‡] and Johann Gasteiger[†]

Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany, and Laboratório de Farmacognosia, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Av. do Café s/no., 14040-903, Ribeirão Preto, SP, Brazil

Received February 6, 2006

In a recent publication we described the application of an unsupervised learning method using self-organizing maps to the separation of three tribes and seven subtribes of the plant family Asteraceae based on a set of sesquiterpene lactones (STLs) isolated from individual species. In the present work, two different structure representations—atom counts (2D) and radial distribution function (RDF) (3D)—and two supervised classification methods—counterpropagation neural networks and *k*-nearest neighbors (*k*-NN)—were used to predict the tribe in which a given STL occurs. The data set was extended from 144 to 921 STLs, and the Asteraceae tribes were augmented from three to seven. The *k*-NN classifier with $k = 1$ showed the best performance, while the RDF code outperformed the atom counts. The quality of the obtained model was assessed with two test sets, which exemplified two possible applications: (1) finding a plant source for a desired compound and (2) based on a plant species chemical profile (STLs): (a) study the relationship between the current taxonomic classification and plant's chemistry and (b) assign a species to a tribe by majority vote. In addition, the problem of defining the applicability domain of the models was assessed by means of two different approaches—principal component analysis combined with Hotelling T^2 statistic and an a posteriori probability-based rule.

I. INTRODUCTION

Plant chemotaxonomy is generally focused on the investigation of chemical relationships among taxa at different taxonomic hierarchical levels, like families, tribes, genera, or species. For instance, these relationships are investigated taking into account the observation and the analysis of biosynthesis, occurrence, and special trends of accumulation of certain structural classes of secondary metabolites within a given taxon. Among several possible applications, the obtained information can be used for classification of plants into groups. Hence, chemotaxonomy had a strong influence on plant systematics, and new taxonomic classifications were developed taking into account the distribution of such metabolites.[1] Moreover, the impact of this subject may also exert an effect on other areas of research like plant metabolism, biological activities, geographic distribution of secondary metabolites, analysis of cultivars, and quality control of plant drugs.

Among the terpenoids, the sesquiterpene lactones (STLs) comprise the most used structural class of secondary metabolites for chemotaxonomic studies in the family Asteraceae. They are considered taxonomic markers, show many biological activities, and are of commercial as well as of ecological value.[2,3] According to Bremer,[4] the family Asteraceae has ca. 23 000 species arranged into 17 tribes, 82 subtribes, and around 1500 genera where the STLs are widespread. Currently there are more than 4000 known STLs, and around 40 different carbon skeletons comprise the most important group.[5,6] Consequently, this large amount of chemical information is a rich source of data which has great value for the investigation of relationships among taxa at tribal, subtribal, generic, or species levels. Such investigation can be done by statistical methods, but so far only principal component analysis (PCA) has been explored.[6] Hence, machine learning techniques, for example classification methods, may provide useful ways for exploration of these data.

The use of classification methods for predictive purposes raises the question about the applicability domain of the built models. Given the fact that a machine learning technique is always trained on data covering a limited area of the chemical universe, it becomes important that it is able to differentiate between (i) compounds which are very far from the training data and (ii) compounds which are in or very near to the space, spanned by the training data. While in the field of QSAR and chemoinformatics the use of such methods is still in an early stage, as reviewed by Netzeva et al.,[7] different approaches have been proposed in the field of machine learning, termed "novelty detection". They can be separated in two major groups—statistical methods[8] and neural networks.[9] Targeting specifically the *k*-NN classifier, a number of rejection strategies have been proposed,[8,10,11] mostly utilizing the a posteriori probability estimates.

The representation of chemical structures is another important aspect that can change dramatically the quality of the obtained models. There are several ways to describe a chemical structure,[12] e.g. based on the connectivity informa-

* Corresponding author phone: (55)-16-3602-4312; e-mail: febcosta@fcfrp.usp.br.
† Universität Erlangen-Nürnberg.
‡ Universidade de São Paulo.

10.1021/ci060046x CCC: $37.00
Published on Web 12/02/2006

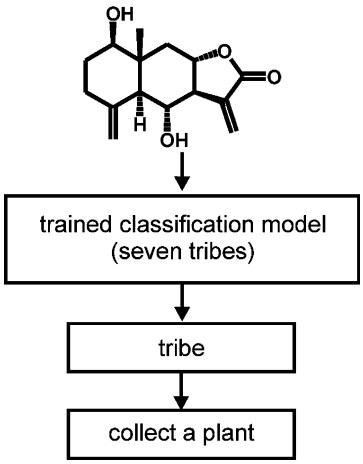

Figure 1. Illustration of the first use case of the proposed classification model: classifying compound(s) of interest with the aim of finding plant source(s) for them.

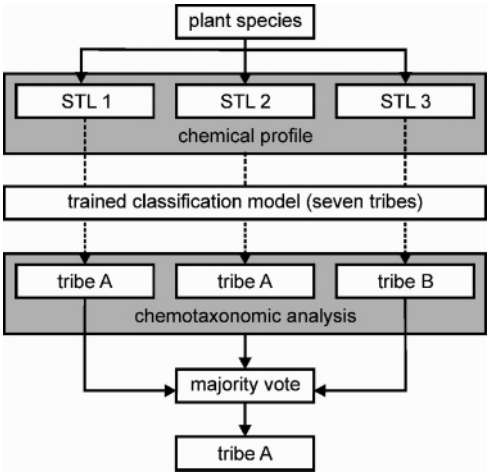

Figure 2. Illustration of the second use case of the proposed classification model: starting with an STL profile of a given species (STL1+STL2+STL3, first gray area), the classification is made for each single STL from this profile, and the results are used to analyze chemical relationships between different tribes—the second gray area; in addition, the plant species can be assigned to a tribe by majority vote.

tion alone (2D) or on the spatial arrangement of the atoms in a molecule (3D).

Recently we described the separation of seven subtribes of the plant family Asteraceae based on sesquiterpene lactones (STLs) using a data set which consisted of ca. 150 STLs.[13] That approach was based on unsupervised learning through the projection of the 3D structures of STLs encoded by their radial distribution functions (RDF) into self-organizing maps (SOMs). The study provided useful insights into the STL-based chemosystematics.

In this work, we present classification models which predict the Asteraceae tribe from which a given STL has been isolated based on two different structure representations—atom counts augmented with stereo information[14] ("2D") and a description of the 3D structure by the RDF code.[15] The applicability of such classification is demonstrated with two use cases: (1) targeted collection of plant material—given a known and/or desired STL, the most probable tribe where it can be found is suggested by the classification, thus substantially limiting the possible plant sources (Figure 1); and (2) once the STL profile of a given plant species is known, the proposed model can be used in the subsequent chemotaxonomic analysis, and the species can be assigned to a tribe based on majority vote (Figure 2). The second use case can lead to new insights into the relationship between the current taxonomic classification and plant's chemistry. Counter-propagation (CPG) neural networks[16] and *k*-nearest neighbors (*k*-NN) were used to build the classification models. A comparison is presented between two different approaches for rejecting patterns which are likely to be misclassified: (i) distance metric based on PCA and Hotelling T^2 statistic[17] and (ii) a probability-based approach that takes into account the distances to the nearest neighbors as well as their class membership.[11]

Thus, this work has the following objectives: (1) development of a classification model, which is capable of classifying STLs from plant species into their corresponding Asteraceae tribes; (2) comparison between different supervised learning techniques and structure representations; and (3) comparison between different approaches for defining the applicability domain of a model. An advantage of this model is that it can deal with literature-new or previously unreported structures rather than only retrieving known STLs from a data set.

II. MATERIALS AND METHODS

2.1. Data Sets. The structures of the 921 STLs and the information concerning their taxonomic origin were taken from comprehensive surveys published in the literature.[2,18−40] The taxa (tribes and genera) presented herein are in accordance to the division made by Bremer.[4] This information was used to assign all chemical structures to their corresponding taxa. One hundred and nine (around 11%) STLs were reported in more than one tribe, i.e., they overlap. In this work, each of the overlapping STL was assigned to only one tribe. The criterion used to assign such structures was their predominance within the tribes from where they were reported. It means that if a certain STL occurs in more than one tribe it was placed in the one in which it was reported more times, i.e., in the one with the highest number of occurrences. The number of occurrences of each STL in the corresponding tribes was obtained through a systematic survey in SciFinder.[41] This criterion does not lead to misrepresentation of tribes with respect to their actual chemical profile since only the most representative STLs of each tribe were considered. Thus, such representative STLs cover a broad array of chemodiversity since their skeletons and most common substitutional features are present in the data set. During the data collection, care was taken to select as many different STLs types as possible, i.e., to select structures that belong to a great variety of skeletal classes and subclasses among all tribes. The most common major biogenetic types of compounds, germacranolides, guaianolides, and eudesmanolides, which together comprise the tripartite chemistry[4] of the family Asteraceae, were selected in a major scale. Analogously, the biogenetically advanced but less abundant skeletal types, like pseudoguaianolides, elemanolides, xanthanolides, and seco-derivatives, were also taken into consideration. Ring functionalization and substitution features such as side chain esters and sugar moieties were also taken into account as well as the presence of dimers

Table 1. Distribution of STLs in Their Corresponding Asteraceae Tribes

tribe	abbreviation	number of STLs
Anthemideae	ANT	175
Cardueae	CAR	73
Eupatorieae	EUP	202
Heliantheae	HLT	296
Inuleae	INU	50
Lactuceae	LAC	48
Vernonieae	VER	77
total		921

Table 2. Second Test Set—STLs from Nine Plant Species

species	tribe	abbreviation	number of STLs
Anthemis wiedemanniana[42]	ANT	ANT. es1	3
Achillea depressa[43]	ANT	ANT. es2	7
Centaurea babylonica[44]	CAR	CAR.es1	6
Centaurea acaulis[45]	CAR	CAR.es2	6
Stevia alpina var. *glutinosa*[46]	EUP	EUP.es1	9
Viguiera eriophora[47]	HLT	HLT.es1	11
Carpesium abrotanoides[48]	INU	INU.es1	8
Leontodon palisae[49]	LAC	LAC.es1	5
Vernonanthura lipeoensis[50]	VER	VER.es1	17
total			72

and other minor components of low occurring nature. As nowadays the great majority of all reported structures of STLs have their stereocenters fully assigned, in this work only those with total assigned stereochemistry were selected. Table 1 gives the distribution of the 921 STLs into their corresponding tribes.

The most represented tribe in the data set in terms of number of occurrences and skeletal diversity of STLs is Heliantheae (HLT), from which the largest number and diversity of STLs was reported so far. The number and variety of structures of STLs in the remaining six tribes followed the same criterion, i.e., the selection of STL-rich and representative taxa. Most of these taxa have already been the subject of chemotaxonomic studies based only on their STL profile.[2,3,13]

The data set was then split semirandomly—the corresponding distribution of STLs in the tribes was preserved—in three subsets: a training set with around 70% (644 structures) of the data set, a cross-validation (135 structures), and a test set (142 structures). The test set was not used in the model building but only after all parameters of the underlying classification method were established. It illustrates the first use case, i.e., given an STL that is of particular interest—even literature-new structures—the classification is made in order to find a plant source for it (cf. Figure 1).

To exemplify the second use case (cf. Figure 2) a second test set consisting of 72 STLs from nine plant species from the seven tribes was built. The information about these STLs, i.e., their plant sources and chemical structures, was collected from recent articles as indicated in Table 2. In this case, starting with an STL profile of a given species—including previously unreported structures—, the classification is made for each single STL, and the results are used to analyze relationships among different tribes in terms of their chemical profiles. This second set is summarized in Table 2.

In addition to the chemotaxonomic analysis, the second test set was also used with the final model to predict the tribe to which a species belongs. A prediction was made for each single compound from a species chemical profile, and the majority vote was used to determine the tribe. For example, given a profile, consisting of seven STLs, if five of them were predicted to belong to a certain tribe A and two of them to tribe B, then the species was assigned to tribe A (Figure 2).

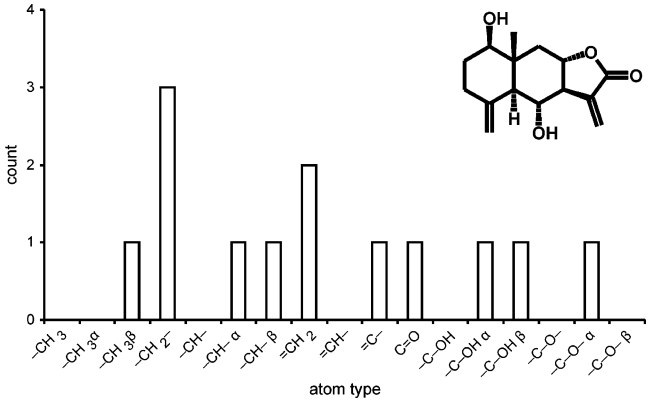

Figure 3. Example of an STL coded with histogram of atom counts.

2.2. Structure Representation. Two different structure representations—atom counts and radial distribution functions—were investigated.

2.2.1. Atom Counts. This representation involves a histogram of atom counts with atom types, which was built for each structure. These atom types were previously defined and originally used for the prediction of ^{13}C NMR spectra using artificial intelligence.[14] In addition, as can be deduced from the 2D representation of the structures, the relative configuration of some centers, i.e., the orientation of some bonds, was taken into account. As already mentioned, the structures in the data set have fully assigned stereochemistry, and in this work their corresponding 2D structures were drawn in such a way which is common in most of the publications regarding STLs. This code was built by considering the same atom types as different if, for instance, a sp^3 carbon atom has an α- or a β-oriented substituent, i.e., a hydroxyl group or an ester moiety below or above of the plane, respectively. The originally reported atom types were chosen with the aim to describe any common organic structure.[14] In our study, the great majority of the structures of STLs have mainly sp^3 and sp^2 carbon as well as oxygen atoms. As a consequence, only the following atom types were considered: $-CH_3$, $-CH_2$, $-CH-$, $-C-$, $=CH_2$, $=CH$, $=C-$, $C=O$, $-C-OH$, $-C-O-$, and $=C-OH$ as well as other minor types. The inclusion of stereo information as already mentioned resulted in 27 different atom types. However, the examination of the histograms which were generated using all the 27 atom types revealed that the number of compounds in the training set, e.g. those that have stereo information for certain atom types, was very low. So, their corresponding counts were removed, and the final representation consists of a histogram with 17 bins (different atom types) for each STL: $-CH_3$, $-CH_3$ α-oriented, $-CH_3$ β-oriented, $-CH_2-$, $-CH-$, $-CH-$ α-oriented, $-CH-$ β-oriented, $=CH_2$, $=CH-$, $=C-$, $C=O$, $-C-OH$, $-C-OH$ α-oriented, $-C-OH$ β-oriented, $-C-O-$, $-C-O-$ α-oriented, and $-C-O-$ β-oriented. Figure 3 gives an example of the histogram, obtained for the shown STL.

The rationale behind this descriptor is that different plant species from specific tribes have special trends of accumula-

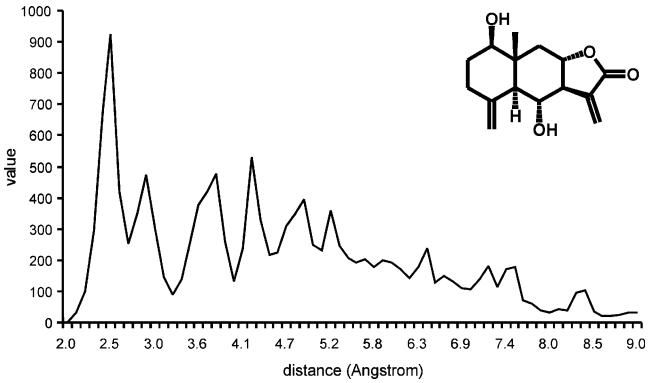

Figure 4. Example of an STL coded with 64 dimensional radial distribution function in the interval 2.0−9.0 Å using atomic number as a property for each atom.

tion of compounds which show certain features at specific positions of their skeletons, e.g. α- or β-oriented hydroxyl groups and their corresponding esters, reduction of double bonds, epoxy groups, etc.

2.2.2. Radial Distribution Function (RDF). The second representation is based on RDF codes[15] calculated using the three-dimensional structures. Single, low energy 3D conformations were generated for the STLs from their 2D constitution using CORINA.[51,52] The RDF codes were calculated with the descriptor calculation package ADRIANA.CODE[53] according to the following equation

$$g(r) = \sum_{i=1}^{N-1} \sum_{j>i}^{N} A_i A_j e^{-B(r-r_{ij})2} \quad (1)$$

where N is the number of atoms in a molecule, A_i and A_j are properties associated with the atoms i and j, respectively, r_{ij} represents the distance between atoms i and j, and B is a smoothing factor. The above formula was applied with the parameter A set to the atomic number of the considered atom, and 32, 64, and 128 dimensional RDF codes were calculated. The function $g(r)$ was defined in the interval 2.0−9.0 Å. RDF of an ensemble of atoms can be interpreted as a probability distribution of the atoms in 3D space. This code is independent of the number of atoms, it is unique regarding the three-dimensional arrangement of the atoms and is invariant against translation and rotation of the entire molecule. The RDF code can distinguish diasteroisomers as well as epimers but not enantiomers. Although the STLs show several chiral centers and may potentially occur as enantiomers pairs, the STL-producing plants do not biosynthesize stereoisomers, and the fact that RDF does not reflect such differences is not relevant in this study. Figure 4 shows the same STL as in Figure 3 and its corresponding RDF code.

2.3. Classification Methods. 2.3.1. Counterpropagation (CPG) Neural Network. A CPG neural network consists of a SOM block (input layer) and an additional output block. The input data are stored in a two-dimensional grid of neurons, each containing as many elements (weights) as there are input variables (Figure 5). All the units of the inputs (object + properties) are linked to the SOM block.[16] In this work, the input data are the structures of STLs as represented by their atom counts or RDF codes. During the training, a winning neuron is chosen by the SOM block for each individual object (coded structure of a STL). Each neuron in the SOM block is in turn connected to a neuron in the

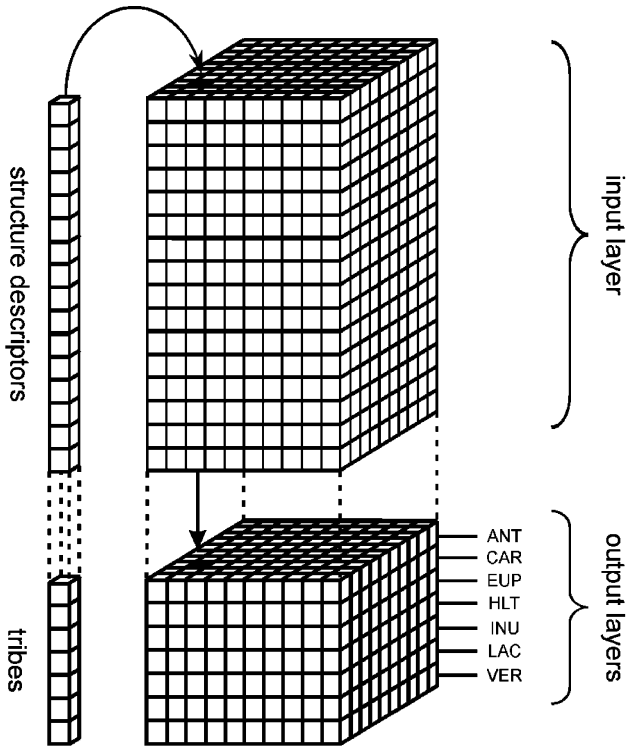

Figure 5. Architecture of the counterpropagation neural network for the classification of STLs into seven Asteraceae tribes with seven output layers—one for each tribe.

output block (Figure 5). In this study, the output block consisted of seven layers, one layer for each of the seven tribes of Asteraceae. The weights of the output layers are adapted in order to become closer to the output value of the presented object. Predictions for new compounds are made by determining the winning neuron, defined as the neuron with the smallest Euclidian distance between its weight vector and the X-variables of the STL. CPG neural networks were calculated using SONNIA.[54]

2.3.2. *k*-Nearest Neighbor (*k*-NN). This method belongs to the family of "lazy" learners. It is memory-based and requires no model to be fit.[55] All training examples are stored in the memory, and the prediction of a new pattern is made by finding its k-nearest neighbors in terms of some predefined distance measure and assigning the pattern to the class to which the majority of the nearest neighbors belong. In this study, the measure was the Euclidian distance, which is the most common used. Although it is a very simple method, k-NN has been successfully utilized in many real world applications. It is able to approximate highly irregular class boundaries which are inevitable when highly overlapping classes have to be classified. The choice of k is crucial and very important because it controls the bias-variance tradeoff of the method. A low number of neighbors provides low bias and high variance, while high values of k tend to reduce the variance but increase the bias.[56]

2.4. Model Validation. The quality of the initial model was assessed by means of the first test set (142 STLs, see section 2.1). The final model, which is based on the entire set of 921 STLs, was evaluated by means of the second test set (72 STLs, see Table 2) as well as with stratified 10-fold cross-validation.[55] The data set was randomly divided into 10 subsets, each of which contained approximately the same distribution of the seven classes as the whole data set. Then

a model was fitted taking nine of these subsets as a training set and the remaining one as a test set. The procedure was repeated 10 times until each subset has been used as a test set once. Due to the random splitting the estimates can vary; therefore, the whole procedure was repeated 10 times, leading to 100 model fitting runs. The confusion matrix of the average cross-validated predictions was computed at the end of the procedure. From it the following per class and overall statistics were derived:

- recall (also known as sensitivity), which is obtained by dividing the number of correctly classified compounds of a given tribe by the total number of compounds in this tribe;
- precision, which is obtained by dividing the number of correctly classified compounds of a given tribe by the total number of compounds which were predicted to belong to this tribe;
- F-measure, which is calculated as $2 \times recall \times precision/recall + precision$ and combines the recall and precision in a single efficiency measure (it is the harmonic mean of precision and recall);
- Cohen's *kappa*, which gives a measure of the classification accuracy after accounting for chance effects. It is independent of the prevalence of a given class[57] and is therefore better suited for data sets in which the different classes are not evenly distributed than the overall correct classification rate. It takes values between 0 and 1, with values of 0.0−0.4 considered to indicate slight to fair model performance, values of 0.4−0.6 moderate, 0.6−0.8 substantial, and 0.8−1.0 almost perfect.

2.5. Determination of Prediction Space. 2.5.1. Principal Component Analysis (PCA) and Hotelling T^2. Eriksson et al.[17] gave a detailed description of applying PCA in concert with Hotelling T^2 score as a method for defining the prediction space. The methodology was also used in a recent classification study.[58] In summary, a PCA is performed on the training set, and the obtained loading matrix is used to calculate scores on the external set. Using these scores and a given confidence level, a pattern from the external set is decided to belong to the prediction space if its Hotelling's score satisfies the following criterion

$$T_i^2 < \frac{A(N^2-1)}{N(N-A)}F(p=\alpha) \tag{2}$$

where $F(p=\alpha)$ is a tabulated value for a F distribution using a confidence level α, A is the number of principal components used to build the Hotelling's test, and N is the number of compounds of the training set, and

$$T_i^2 = \sum_{a=1}^{A} \frac{t_{ia}^2}{s_a^2} \tag{3}$$

where s_a^2 is the variance explained by principal component a, and t_{ia} is the score of compound i for principal component a.

A graphical overview of this approach is shown in Figure 6. It shows the PCA-score plot of our test set (142 STLs) described by 64-dimensional RDF codes.

The ellipse defines the 95% confidence region and was determined from the training set−779 STLs. This number of STLs refers to the training set, which comprises 644 structures (around 70% of the data set) and 135 more structures from the cross-validation set (around 15%). All points outside the ellipse are supposed to be outside the prediction space and are subsequently not classified. Its application is not limited to the first two PCs as suggested in Figure 6. The ellipsoid can be defined by any sensible number of PCs. In this study, we used the first 5 PCs, which covered approximately 85% of the variance. This approach is rather global, i.e., it rejects patterns that fall clearly outside of the volume, spanned by the training set. However, if a pattern falls inside this volume it is always accepted, even if the training data density around it is sparse.

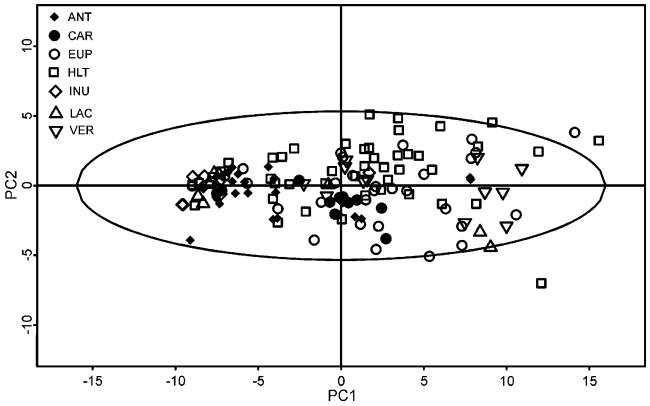

Figure 6. PCA-score plot of the test set (142 STLs) with 64 dimensional RDF codes. The ellipse defines the 95% confidence region, as calculated by using the training set (779 STLs). Each point represents an STL, the tribe from which it was isolated being indicated by the corresponding symbol.

2.5.2. Reject Rule. In a statistical classifier, an estimation $\hat{p}(x|w_i)$ of the probability density function (pdf) of each class in the feature space is computed from a set of training observations. In the voting k-NN classifier the decision rule is equivalent to the Bayes' minimum error rule with the following pdf estimation[11]

$$\hat{p}(x|w_i) = \frac{k_i}{n_i v} \tag{4}$$

where n_i is the number of prototypes of the class w_i, k_i is the number of prototypes from class w_i among the k-nearest neighbors, and v is the volume of a hypersphere comprising all of them. The Bayes' theorem can be applied to this pdf estimation to obtain a posteriori probability estimation $\hat{P}(w_i|x)$. However, for finite size data sets and small values of k it is often preferable to apply a measure which takes into account the distances as well as the class membership of the nearest neighbors.[11] A commonly used estimation and the one which was applied in this study is

$$\hat{P}(w_i|x) = \frac{\sum_{j \in s_i} \frac{1}{d(x,y_j)}}{\sum_{j=1}^{k} \frac{1}{d(x,y_j)}} \tag{5}$$

where d is the distance measure used, and s_i is the set of subindices of the prototypes from class w_i among the k-nearest neighbors retrieved $y_1 \ldots y_k$. A suitable small value should be used when the distance computation gives a result of zero.

Table 3. Classification Statistics of the CPG Neural Network Models Applied on the First Test Set[b]

	atom counts			RDF codes, 64 dimensional		
tribe	recall	precision	F-measure	recall	precision	F-measure
Anthemideae	0.589	0.493	0.535	0.726	0.498	0.59
Cardueae	0.527	0.481	0.499	0.554	0.534	0.536
Eupatorieae	0.573	0.583	0.576	0.57	0.48	0.518
Heliantheae	0.693	0.582	0.631	0.582	0.547	0.562
Inuleae	0	NaN[a]	NaN[a]	0	NaN[a]	NaN[a]
Lactuceae	0.444	0.895	0.589	0.344	0.718	0.458
Vernonieae	0.492	0.724	0.58	0	NaN[a]	NaN[a]
kappa	0.439			0.38		

[a] NaN "not a number"—indicates that no STLs were predicted as occurring in the corresponding tribe. [b] 142 STLs, 19 × 22-dimensional SOM layer, rectangular topology, 50 epochs, average values over ten runs with different seeds.

Table 4. Classification Statistics of the 1-Nearest Neighbor Models Applied on the First Test Set

	atom counts			RDF codes, 64 dimensional		
tribe	recall	precision	F-measure	recall	precision	F-measure
Anthemideae	0.759	0.824	0.790	0.741	0.741	0.741
Cardueae	0.636	0.621	0.628	0.818	0.6	0.692
Eupatorieae	0.777	0.747	0.762	0.8	0.857	0.828
Heliantheae	0.878	0.778	0.824	0.911	0.788	0.845
Inuleae	0.325	0.299	0.31	0.375	1	0.545
Lactuceae	0.455	1	0.625	0.667	0.857	0.75
Vernonieae	0.867	0.948	0.904	0.667	0.8	0.727
kappa	0.691			0.723		

This measure can be interpreted as the confidence of the classifier on its decision. An ambiguity threshold $T_a \in [0,1]$ can be applied to reject the patterns that are not clearly classified in one class, i.e., the value of the estimator is smaller or equal to T_a.

III. RESULTS

3.1. CPG Neural Network Models. To select the best size and topology of the SOM block, different size/topology combinations were examined with both structure representations. The initial size of the SOM block was set to $5 \times \sqrt{N}$, where N is the number of compounds in the training set following an empirical rule[59] and was increased until $0.8 \times N$. Each size was tested with toroidal and rectangular topology of the SOM block, with both structure representations and with training time of 50 epochs. The quality of the obtained models was judged using the validation set. After the best one has been selected—19 × 22-dimensional SOM block with rectangular topology for both structure representations—the validation set was merged with the training set, and a new model was built using the already determined best parameters. Table 3 gives the classification statistics of this model when applied to the first test set with atom counts and 64 dimensional RDF codes. Since the training of a SOM is a stochastic approximation process, the given values are averaged over 10 runs with different seeds.

RDF codes with different dimensionalities, i.e., 32 and 128, were examined as well. The obtained *kappa* values—0.342 and 0.298—were somehow lower, and we concluded that 64 is the best RDF dimensionality for our case.

3.2. *k*-NN Models. The selection of the optimal number of neighbors for the *k*-NN classifier was made by varying *k* from 1 to 9. Analogously to the CPG neural network approach, the validation set (135 STLs) was used at this point, and once the optimal value for *k* had been determined this set was merged with the training data. The best results were obtained with $k = 1$. Table 4 gives the classification statistics of the 1-nearest neighbor models applied to the first test set (142 STLs) with atom counts and 64 dimensional RDF codes. If more than one nearest neighbor was found, i.e., the tested pattern had the same distance to more than one pattern in the training set, the result was determined by majority vote with possible ties broken at random.

As occurred with the CPG neural network, the 64 dimensional RDF code gave the best performance compared to 32 and 128 dimensional ones, which gave *kappa* values of 0.584 and 0.636, respectively.

3.3. Comparison between the Obtained Models. It is generally agreed that the more data the model has been trained on, the more likely it will be reliable.[55] Therefore our final models were built using the entire data set, i.e., 921 STLs, with the settings described above. The CPG neural network classifier was dropped at this stage since Tables 3 and 4 show that the 1-nearest neighbor performed significantly better on this type of data. The classification statistics of the 1-nearest neighbor models with both structure representations based on 10 times 10-fold stratified CV are given in Table 5.

As can be seen in Table 5, the use of RDF codes gave slightly better results. However, there is substantial variance in each single CV run. Although the presented results are averaged over 10 CV runs, which in turn reduces the variance, the use of a statistical test rather than comparing the values directly has been suggested.[55] To test if the difference between using atom counts and RDF codes with 1-nearest neighbor classifier is statistically significant, we used a two-tailed paired *t*-test, using the average *kappa* values, produced in each single 10-fold CV run. The same splits of the data set were used with each structural descriptor, thus allowing the application of this test.[55] The test produced a *p*-value of 0.0066; therefore, the null hypothesis—that the

Table 5. Classification Statistics of the Final 1-Nearest Neighbor Models Based on Ten Times 10-Fold Cross-validation

	atom counts			RDF codes, 64 dimensional		
tribe	recall	precision	F-measure	recall	precision	F-measure
Anthemideae	0.731	0.727	0.729	0.726	0.774	0.749
Cardueae	0.726	0.768	0.746	0.863	0.700	0.773
Eupatorieae	0.752	0.749	0.751	0.807	0.795	0.801
Heliantheae	0.780	0.729	0.754	0.780	0.740	0.760
Inuleae	0.460	0.548	0.500	0.500	0.641	0.562
Lactuceae	0.625	0.682	0.652	0.646	0.756	0.697
Vernonieae	0.792	0.871	0.830	0.649	0.714	0.680
kappa	0.665			0.682		

Table 6. Assignment of the Individual Plant Species into Their Corresponding Tribes of Asteraceae Using the Majority Vote[a]

species	number of STLs	correctly classified STLs	species classified as
ANT. es1	3	2	**ANT**
ANT. es2	7	4	**ANT**
CAR.es1	6	4	**CAR**
CAR.es2	6	1	ANT
EUP.es1	9	8	**EUP**
HLT.es1	9	8	**HLT**
INU.es1	8	2	EUP
LAC.es1	5	3	**LAC**
VER.es1	17	5	**VER**

[a] The correctly assigned species are given in bold.

two means of the *kappa* values follow the same distribution—can be rejected with high confidence. Due to the fact that the individual CV estimates were obtained on the same data set they are not actually independent, and the above test can confirm only that these estimates are different and not the "true" classifier performance across different training sets. However, it has been shown that repeated 10-fold CV is one of the most accurate estimators available; thus, we conclude that the RDF code indeed outperforms the atom counts.

3.4. Predictions on the Second Test Set. In addition to provide another measure for the quality of the built model, the purpose of the second test set (72 STLs) was to illustrate the applicability of the proposed model to chemotaxonomic analysis (cf. Figure 2 and Discussion) as well as the ability to correctly classify individual plant species rather than single STLs. The entire data set (921 STLs) was used as training data, and the STL profiles of the plant species (Table 2) were classified using 1-nearest neighbor with 64 dimensional RDF codes. The classification based on single STLs produced a *kappa* value of 0.447. In addition, each species was assigned to a tribe using a majority vote, and the results are summarized in Table 6.

3.4. Defining the Prediction Space. The methods for defining the prediction space were tested initially on the first test set (142 STLs) using the remaining (779 STLs) of the entire data set for defining the space and consequently on the second test set (72 STLs) using the full data set (921 STLs) for defining the prediction space. 1-nearest neighbor with 64 dimensional RDF codes was used as a classification method.

3.4.1. PCA and Hotelling T^2. The rejection rate was varied by using different confidence levels α, cf. eq 2. All STLs, which were identified outside the prediction space at a given level, were left out of the test set, i.e., were rejected, and no prediction was made for them. Figure 7 shows the overall classification quality, which was obtained on the first and second test set given by the *kappa* statistic, at different rejection rates. All STLs were described by 64 dimensional RDF codes.

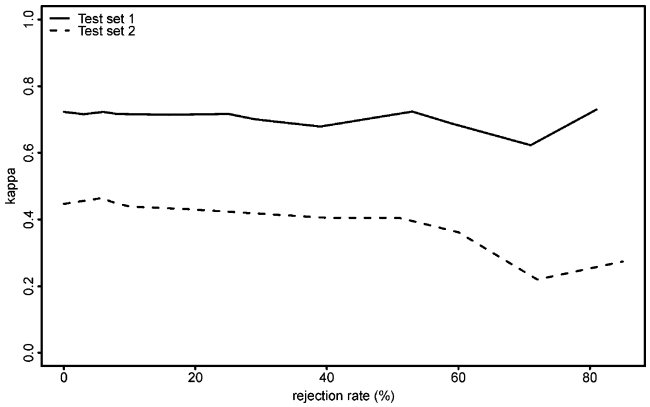

Figure 7. Overall classification quality against the rejection rate with PCA based approach.

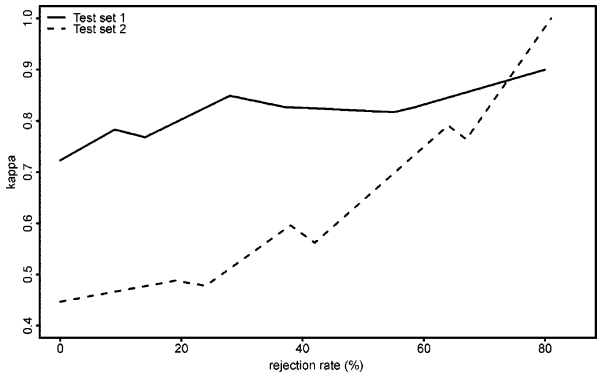

Figure 8. Overall classification quality against the rejection rate with the reject rule.

Even at high rejection rates no significant improvement in the classification quality can be observed. For the second test set, the quality even decreases with the rejection rate, thus rather than excluding the misclassified patterns some of the correctly classified ones have been deemed new.

3.4.2. Rejection Rule. The rejection rate was varied using increasing values for the ambiguity threshold T_a. All STLs were first classified, and the probability given by eq 5 was calculated by using the five nearest neighbors. If the calculated probability for a given structure was smaller or equal to the predefined threshold, the structure was removed from the test set. After all such STLs have been removed, the classification statistics were calculated from the remaining and classified with high confidence patterns. Figure 8 shows the overall classification quality, which were obtained on the first and second test set given by the *kappa* statistic at

different rejection rates. All STLs were described by 64 dimensional RDF codes.

The classification quality generally increases with increase in the rejection rate with small fluctuations. Even at small rejection rates—less than 20%—there is an improvement in the overall classification quality.

IV. DISCUSSION

The main goal of this study was to develop a methodology which allows the assignment of different STLs, isolated from taxa of the family Asteraceae, into their corresponding tribes. This methodology is a valuable tool in various cases—collection of new plant material with the aim of finding compounds with special structural features (cf. Figure 1), classification of novel compounds, and to help in studies about the relationship between taxonomy and chemistry (cf. Figure 2). The proposed method allows the classification of novel compounds (unreported or literature-new). This is important since currently one is limited to search in different databases, which can deal only with already known structures. To achieve that, we used two supervised learning algorithms and two different structure representations.

Classification Methods. With regards to the classification algorithms, comparing the *kappa* values given in Tables 3 and 4 clearly show that the 1-nearest neighbor classifier outperforms the CPG neural network by approximately a factor of 2. This observation together with the fact that 1-nearest neighbor outperforms all other number of neighbors tested suggests that the decision boundaries between the different tribes are highly irregular and can be approximated only locally.[56] This is not unexpected since it is known that often a given compound of high structural complexity occurs simultaneously in two or more taxa in the plant kingdom and this is certainly valid within Asteraceae. This natural overlap, or co-occurrence of natural compounds, makes the decision boundaries hard to capture, and thus the 1-nearest neighbor classifier gave the best performance. In addition, the overall characteristic of structures of secondary metabolites may cause a dilemma in the assignment of a given structure into its corresponding class. In the present study, if a compound is reported to appear in more than one tribe it was assigned exclusively to the one in which it is more frequently found. Although this is the most common approach when dealing with such multilabeled data,[60] it additionally blurs the boundaries between the classes.

The most misclassified tribe was Inuleae (INU), regardless of the used classification method or structure representation. This can be attributed to various reasons: first, the number of STLs belonging to this tribe in the data set (50) might not be enough (although the STLs from this tribe are quite similar among each other and certainly an increase in the data set would not cause dramatic changes in the results); second, around 50% of the 50 present STLs have also been isolated from taxa belonging to some of the other six tribes, i.e., INU has the highest percentage of STLs which co-occur within the other Asteraceae tribes, thus indicating that it does not present a distinguishable pattern of STLs. The most common co-occurrence is between INU and HTL, and subsequently almost all STLs marked to belong to INU are misclassified as belonging to HLT. This can be seen as an indication of chemical similarity between these two tribes with regard to certain skeletal types of STLs, like the guaianolides, and other closely related biogenetic derivatives, which is in agreement with previous chemotaxonomic studies.[2,3]

Structural Descriptors. At first glance, both structural descriptors gave a similar performance but as shown by the paired *t*-test the RDF code produce statistically significant improvement. Shifting to the individual tribes rather than comparing the overall performance, one can notice that by using the RDF code the recall for CAR and LAC is increasing, while it is decreasing for VER. This can be attributed to the fact that in VER nearly all STLs have an 8α-oxygen function,[3] which is better reflected by the atom counts. Moreover, the skeletal type variability of the STLs from VER is quite low when compared to HLT or EUP. The precision remains pretty much the same with regard to the different representations. However, using the RDF code no STLs from other tribes are predicted as INU—precision of one (cf. Table 4)—, while in the case of atom counts some, mainly HLT, were misclassified as belonging to INU. Again it can be explained by the fact that some STLs from HLT (the major and more diversified group) show overlap with some compounds from INU. Despite the improvement brought by using RDF codes, the fact that the atom counts have approximately four times smaller dimensionality could still advocate for their usage. However, it should be stressed that the atom counts used in this study have been augmented with stereo information, i.e., consideration of the stereo bond types—α or β—was performed. This fact prevents this descriptor from being purely 2D, i.e., completely deducible from the connectivity alone and puts the requirement of some prior knowledge about the 3D configuration of the stereo centers. We tried to build a classification model by omitting the stereo augmentation, but the quality of the results decreased drastically with a resulting *kappa* value of 0.231, compared to 0.665 (cf. Table 5). This is understandable having in mind that, for example, the 8α- or 8β-oxygen orientation is very important for some tribes such as VER, EUP, or HLT from the chemotaxonomic point of view.[2,3] Therefore, the RDF codes are preferable since they bring statistically significant improvement and are also a more general descriptor, while the selected atom types for the histogram might be (slightly) influenced by the particularity of a certain subset of data.

Applicability. The proposed model allows the classification of a single STL into a corresponding Asteraceae tribe to be done with high confidence. As illustrated in Figure 1, such an assignment can be utilized for the selection of a plant source for a given (or desired) natural compound, even if it is a novel structure. Although a tribe is somewhat high in the taxonomic hierarchy within the family, selecting the most probable tribe already limits the possible plant sources significantly since there is a large number of species within Asteraceae. Thus, the proposed methodology allows for the "targeted" collection of plant material based on chemotaxonomic relationships. When applied to real life situations, this feature can drastically narrow the search for a plant as well as to save time and money. This procedure is carried out in several plant screening projects and is certainly a valuable strategy for the discovery of biologically active natural products.[61,62] In the same direction, another possible usage of this model is its usefulness in structure elucidation, since

spectroscopic data of compounds from plant sources with similar STL profile can be easily assessed and used for further comparison.

From the other point of view, it is often the case in a phytochemistry laboratory when a plant material has been collected and its corresponding compounds were isolated and identified. In such a case the proposed model can help in the consequent chemotaxonomic analysis, especially when one is not familiar with the chemistry and other special features regarding STLs from this huge plant family. This is illustrated by our second test set in which the whole profile of a given Asteraceae species was investigated. This allowed us to assign the species as a whole to a given tribe by using the majority vote on the predictions for each STL as well as to draw conclusions about the similarity in the secondary metabolite chemistry of the species across different Asteraceae tribes.

Chemotaxonomic Analysis and Majority Vote. As can be seen from Table 6, seven out of nine species were assigned correctly to their actual tribe. However, while for some of them almost all STLs were classified correctly—ANT.es1, EUP.es1, HLT.es1—this was not the case for the rest of the species. For both plant species which belong to ANT, all STLs from the second test set possess skeletons and substitution features which are typical for this specific tribe. Subsequently, they were classified correctly as expected. However, the decision for ANT.es2 was very close—with 4 STLs correctly classified as ANT and 3 assigned to HLT. This result suggests a high degree of similarity between the secondary metabolite chemistry of the species *Achillea depressa* and species from the tribe HLT. The first CAR species was correctly assigned with four correct classifications out of six, while the second species was completely misclassified—our model predicted correctly only one of the six reported STLs. From the remaining five structures, three were predicted as belonging to ANT and two as HLT. The species was incorrectly classified as belonging to ANT although the confidence in this decision is rather low. Having in mind that the reported STLs for this species are very common across whole Asteraceae, including for example the compound costunolide, it is not surprising that the classification was incorrect and scattered over several tribes. One must have in mind that costunolide is regarded as the biogenetic precursor of all STLs and is widespread within Asteraceae, thus not being a typical compound of any tribe. The species from EUP and HLT were very confidently assigned to their correct tribe, being an expected result. It is because these tribes are the two most represented in the training set (Table 1) and also because at a certain extent their STL profiles are characteristic. The INU.es1 was misclassified, with the majority of its compounds predicted as HLT. The STLs from INU were found generally hard to separate from the rest (Tables 4 and 5), and these results confirm that based on the current taxonomic classification it is difficult to handle this tribe correctly based solely on the STL profile. However, as already mentioned, INU and HLT are not very well separated with regards to their STL profiles, thus our results confirm the earlier observations.[2,3] The single LAC species was classified correctly. The two misclassified STLs were assigned to ANT and CAR. The more complete profile from all species listed in Table 2 belongs to VER.es1. Some of the 17 reported STLs are frequently found in other tribes—mainly EUP and ANT. The rest are typical for VER, although somewhat unusual due to the position/type of substituents, low occurring skeletons and/or oxidative levels. Applying our model on this rich profile of VER.es1, we were able to classify the species correctly as belonging to VER based on five correct predictions (Table 6). In concert with the above observations, four STLs were misclassified as ANT, three as EUP, and three as HLT. Although we use the term "misclassified", it is not unlikely that some of this compounds have been isolated from species from these tribes as well. Building a classification which is capable of dealing correctly with such multilabeled data is a possible extension of the method proposed herein, and this will be the subject of further studies.

Prediction Space. Another possible reason for misclassification with any model might be that the new instances, which are to be tested, lie far away from the prediction space. Although the models presented herein are aimed at a specific class of secondary metabolites, i.e., STLs, the problem is still present since several novel STLs are reported each year,[18−39] some of them with rather unusual structures, including for example dimers.[63] This was what we examined in turn. As can be seen from Figure 7, the Hotelling test proved unsuitable for the task at hand. Even at high rejection rates the quality of the classification does not increase or even decrease. This means that the test rejected the actually correctly classified STLs rather than rejecting the misclassified ones, which was our expectation. This can be attributed to the fact that, although this approach can identify global outliers, i.e., patterns which do lie relatively far from the space spanned by the whole data, it does not take into account neither the class membership nor can it deal with local spots of low data density. Due to a phenomenon known as "the curse of dimensionality"[56] such spots are inevitable when high dimensional patterns are used. In contrast, the second approach—reject rule based on a measure which combines the a posteriori probabilities with the actual distances cf. eq 5—is rather local and does take the class membership into account. It is incapable of defining outliers, since even at large distances the a posteriori probability estimate can be rather high if all or most of the neighbors are from the same class. Nevertheless, as shown in Figure 8, it brings an almost steady improvement to the classification quality at increased reject rates. Figure 8 shows the improvement in the overall classification quality, but it does not contain any information about the species. We examined if the classification of the species listed in Table 2 (second test set) improves at a rejection rate of about 28%, i.e., 27 out of 72 STLs were rejected from the second test set. Initially, seven out of the nine species were classified correctly. After rejecting around one-third of the data, the assignment of the species into their corresponding tribe by majority vote remains the same. However, the proportion of true votes, i.e., the confidence of the classification, increased. Interestingly almost all—13 out of 17—of the STLs in the VER.es1 were rejected. Thus, their unusual substitution patterns, skeletons as well as oxidative levels, have been captured by the method.

V. CONCLUSION

A classification model capable of classifying a special type of secondary metabolites—sesquiterpene lactones (STLs)—

into seven Asteraceae tribes was developed. The applicability of the presented model for (1) identifying plant sources for a given STL and (2) for studying the relationship in the secondary metabolism across different tribes of individual plant species was shown. The *k*-NN classifier with $k = 1$ gave the best results, regardless of the used structural descriptor, thus suggesting highly irregular class boundaries. Two chemical structure descriptors—histogram of atom counts, augmented with stereo information and RDF codes—were investigated. The RDF code gave better results, and the difference in the performance was statistically significant. Therefore, it is proposed in concert with the 1-nearest neighbor classifier, which clearly outperformed the CPG neural network. Two approaches of identifying patterns which are likely to be misclassified were studied: (1) distance metric based on principal component analysis and Hotelling T^2 statistic[17] and (2) rejection rule, based on the a posteriori probabilities estimates and the distances to the nearest neighbors.[11] While the former did not bring any significant improvement, the latter provided a useful way to reject patterns, in which classification the method was not confident enough. This approach is an example of how statistical methods and machine learning tools can be combined to a real life problem in order to study the intricate mechanism of naturally occurring compounds from plants.

ACKNOWLEDGMENT

FBC is grateful to the *Alexander von Humboldt Foundation* (Germany) for a Research Fellowship at the Computer-Chemie-Centrum. Simon Spycher is thanked for his valuable comments.

REFERENCES AND NOTES

(1) Wink, M. Evolution of Secondary Metabolites from an Ecological and Molecular Phylogenetic Perspective. *Phytochemistry* **2003**, *64*, 3−19.

(2) Seaman, F. C. Sesquiterpene Lactones as Taxonomic Characters in the Asteraceae. *Bot. Rev.* **1982**, *48*, 123−551.

(3) Zdero, C.; Bohlmann, F. Systematics and Evolution Within the Compositae, Seen With the Eyes of a Chemist. *Plant Syst. Evol.* **1990**, *171*, 1−14.

(4) Bremer, K. Classification. In *Asteraceae: Cladistics and Classification*; Bremer, K., Ed.; Timber Press: Portland, Oregon, 1994; pp 13−23

(5) Emerenciano, V. P.; Ferreira, M. J. P.; Branco, M. D.; Dubois, J. E. The Application of Bayes' Theorem in Natural Products as a Guide for Skeletons Identification. *Chemom. Intell. Lab. Syst.* **1998**, *40*, 83−92.

(6) Alvarenga, S. A. V.; Ferreira, M. J. P.; Emerenciano, V. P.; Cabrol-Bass, D. Chemosystematic Studies of Natural Compounds Isolated from Asteraceae: Characterization of Tribes by Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 27−37.

(7) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; van de Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure−Activity Relationships. *ATLA, Alt. Lab. Anim.* **2005**, *33*, 155−174.

(8) Markou, M.; Singh, S. Novelty Detection: a Review − Part 1: Statistical Approaches. *Signal Process.* **2003**, *83*, 2481−2497.

(9) Markou, M.; Singh, S. Novelty Detection: a Review − Part 2: Neural Network Based Approaches. *Signal Process.* **2003**, *83*, 2499−2521.

(10) Fumera, G.; Roli, F.; Giacinto, G. Reject Option With Multiple Thresholds. *Pattern Recognit.* **2000**, *33*, 2099−2101.

(11) Arlandis, J.; Perez-Cortes, J. C.; Cano, J. Rejection Strategies and Confidence Measures for a k-NN Classifier in an OCR Task. In *16th IEEE International Conference on Pattern Recognition (ICPR'02)*; Proceedings of the 16th IEEE International Conference on Pattern Recognition, Quebec, Canada, August 11−15, 2002; IEEE Computer Society, Conference Publications: Los Alamitos, California, 2002; Vol. 1, pp 10576−10579.

(12) Gasteiger, J. A Hierarchy of Structure Representations. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; pp 1034−1061.

(13) Da Costa, F. B.; Terfloth, L.; Gasteiger, J. Sesquiterpene Lactone-Based Classification of Three Asteraceae Tribes: a Study Based on Self-Organizing Neural Networks Applied to Chemosystematics. *Phytochemistry* **2005**, *66*, 34−353.

(14) Gastmans, J. P.; Zurita, J. C.; Sahao, J.; Emerenciano, V. P. Prévision Des Spectres De Résonance Magnétique Nucléaire De ^{13}C Par Intelligence Artificielle: Le Problème De La Codification: Prediction of ^{13}C-Nuclear Magnetic Resonance Spectra by Artificial Intelligence: the Problem of Coding Structures. *Anal. Chim. Acta* **1989**, *217*, 85−100.

(15) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D Structure of Organic Molecules from their Infrared Spectra. *Vib. Spectrosc.* **1999**, *19*, 151−164.

(16) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999.

(17) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361−1375.

(18) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1985**, *2*, 147−161.

(19) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1986**, *3*, 273−296.

(20) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1987**, *4*, 473−498.

(21) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1988**, *5*, 497−521.

(22) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1990**, *7*, 61−84.

(23) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1990**, *7*, 515−537.

(24) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1992**, *9*, 217−241.

(25) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1992**, *9*, 557−580.

(26) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1993**, *10*, 397−419.

(27) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1994**, *11*, 533−554.

(28) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1995**, *12*, 303−320.

(29) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1996**, *13*, 307−326.

(30) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1997**, *14*, 145−162.

(31) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1998**, *15*, 73−92.

(32) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1999**, *16*, 21−38.

(33) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **1999**, *16*, 711−730.

(34) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **2000**, *17*, 483−504.

(35) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **2001**, *18*, 650−673.

(36) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **2002**, *19*, 650−672.

(37) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **2003**, *20*, 392−413.

(38) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **2004**, *21*, 669−693.

(39) Fraga, B. M. Natural Sesquiterpenoids. *Nat. Prod. Rep.* **2005**, *22*, 465−486.

(40) Herz, W. Chemistry of the Eupatoriinae. *Biochem. Syst. Ecol.* **2001**, *29*, 1115−1137.

(41) *SciFinder Scholar*, *version 2004*; American Chemical Society: Columbus, Ohio, 2003.

(42) Celik, S.; Rosselli, S.; Maggio, A. M.; Raccuglia, R. A.; Uysal, I.; Kisiel, W.; Bruno, M. Sesquiterpene Lactones from *Anthemis wiedemanniana*. *Biochem. Syst. Ecol.* **2005**, *33*, 952−956.

(43) Trifunovic, S.; Aljancic, I.; Vajs, V.; Macura, S.; Milosavljevic, S. Sesquiterpene Lactones and Flavonoids of *Achillea depressa*. *Biochem. Syst. Ecol.* **2005**, *33*, 317−322.

(44) Bruno, M.; Rosselli, S.; Maggio, A.; Raccuglia, R. A.; Arnold, N. A. Guaianolides from *Centaurea babylonica*. *Biochem. Syst. Ecol.* **2005**, *33*, 817−825.
(45) Bentamene, A.; Benayache, S.; Creche, J.; Petit, G.; Bermejo-Barrera, J.; Leon, F.; Benayache, F. A New Guaianolide and Other Sesquiterpene Lactones from *Centaurea acaulis* L. (Asteraceae). *Biochem. Syst. Ecol.* **2005**, *33*, 1061−1065.
(46) Hernandez, Z. N. J.; Catalan, C. A.; Hernandez, L. R.; Guerra-Ramirez, D.; Joseph-Nathan, P. Sesquiterpene Lactones from *Stevia alpina* var. *glutinosa*. *Phytochemistry* **1999**, *51*, 79−82.
(47) Spring, O.; Zipper, R.; Klaiber, I.; Reeb, S.; Vogler, B. Sesquiterpene Lactones in *Viguiera eriophora* and *Viguiera puruana* (Heliantheae; Asteraceae). *Phytochemistry* **2000**, *55*, 255−261.
(48) Lee, J. S.; Min, B. S.; Lee, S. M.; Na, M. K.; Kwon, B. M.; Lee, C. O.; Kim, Y. H.; Bae, K. H. Cytotoxic Sesquiterpene Lactones from *Carpesium abrotanoides*. *Planta Med.* **2002**, 745−747.
(49) Zidorn, C.; Ellmerer, E. P.; Konwalinka, G.; Schwaiger, N.; Stuppner, H. 13-Chloro-3-O-[Beta]-Glucopyranosylsolstitialin from *Leontodon palisae*: the First Genuine Chlorinated Sesquiterpene Lactone Glucoside. *Tetrahedron Lett.* **2004**, *45*, 3433−3436.
(50) Pollora, G. C.; Bardon, A.; Catalan, C. A. N.; Griffin, C. L.; Herz, W. Elephantopus-Type Sesquiterpene Lactones from a Second *Vernonanthura* Species, *Vernonanthura lipeoensis*. *Biochem. Syst. Ecol.* **2004**, *32*, 619−625.
(51) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(52) *CORINA*, *version 3.0*; Molecular Networks GmbH: Erlangen, Germany, 2003. http://www.mol-net.de (accessed Jan 2006).
(53) *ADRIANA.CODE*, *version 1.0*; Molecular Networks GmbH: Erlangen, Germany, 2006. http://www.mol-net.de (accessed Jan 2006).
(54) *SONNIA − Self-Organizing Neural Network for Information Analysis*, *version 4.1*; Molecular Networks GmbH: Erlangen, Germany, 2002. http://www.mol-net.de (accessed Jan 2006).
(55) Witten, I. H.; Eibe, F. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: San Francisco, 2000.
(56) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, 2001.
(57) Forbes, A. D. Classification-Algorithm Evaluation: Five Performance Measures Based on Confusion Matrices. *J. Clin. Monit.* **1995**, *11*, 189−206.
(58) Spycher, S.; Pellegrini, E.; Gasteiger, J. Use of Structure Descriptors to Discriminate Between Modes of Toxic Action of Phenols. *J. Chem. Inf. Model.* **2005**, *45*, 200−208.
(59) Vesanto, J.; Himberg, J.; Alhoniemi, E.; Parhankangas, J. *SOM Toolbox for Matlab 5*; Technical Report A57; Helsinki University of Technology, Neural Networks Research Centre: Espoo, Finland, 2000.
(60) Boutell, M. R.; Luo, J.; Shen, X.; Brown, C. M. C. Learning Multi-Label Scene Classification. *Pattern Recognit.* **2004**, *37*, 1757−1771.
(61) Hostettmann, K.; Wolfender, J. L. The Search for Biologically Active Secondary Metabolites. *Pestic. Sci.* **1997**, *51*, 471−482.
(62) Cordell, G. A.; Shin, Y. G. Finding the Needle in the Haystack. The Dereplication of Natural Products Extracts. *Pure Appl. Chem.* **1999**, *71*, 1089−1094.
(63) Staneva, J.; Trendafilova-Savkova, A.; Todorova, M. N.; Evstatieva, L.; Vitkova, A.; Staneva, J.; Trendafilova-Savkova, A.; Todorova, M. N.; Evstatieva, L.; Vitkova, A. Terpenoids from *Anthemis austriaca* Jacq. *Z. Naturforsch., C: J. Biosci.* **2004**, *59*, 161−165.

CI060046X