plant biology

Plant Biology ISSN 1435-8603

RESEARCH PAPER

Reducing the babel in plant volatile communication: using the forest to see the trees

Y. Ranganathan & R. M. Borges

Centre for Ecological Sciences, Indian Institute of Science, Bangalore, India

Keywords

Chemoinformatics; data mining; *Ficus*; Random Forests; *varSelRF*; volatiles.

Correspondence

R. M. Borges, Centre for Ecological Sciences, Indian Institute of Science, Bangalore 560012, India.

E-mail: renee@ces.iisc.ernet.in

Editor

J. Sparks

Received: 16 July 2009; Accepted: 15 September 2009

doi:10.1111/j.1438-8677.2009.00278.x

ABSTRACT

While plants of a single species emit a diversity of volatile organic compounds (VOCs) to attract or repel interacting organisms, these specific messages may be lost in the midst of the hundreds of VOCs produced by sympatric plants of different species, many of which may have no signal content. Receivers must be able to reduce the babel or noise in these VOCs in order to correctly identify the message. For chemical ecologists faced with vast amounts of data on volatile signatures of plants in different ecological contexts, it is imperative to employ accurate methods of classifying messages, so that suitable bioassays may then be designed to understand message content. We demonstrate the utility of 'Random Forests' (RF), a machine-learning algorithm, for the task of classifying volatile signatures and choosing the minimum set of volatiles for accurate discrimination, using data from sympatric *Ficus* species as a case study. We demonstrate the advantages of RF over conventional classification methods such as principal component analysis (PCA), as well as data-mining algorithms such as support vector machines (SVM), diagonal linear discriminant analysis (DLDA) and *k*-nearest neighbour (KNN) analysis. We show why a tree-building method such as RF, which is increasingly being used by the bioinformatics, food technology and medical community, is particularly advantageous for the study of plant communication using volatiles, dealing, as it must, with abundant noise.

INTRODUCTION

Plants produce a diversity of volatile organic compounds (VOCs) from above- and belowground tissues such as leaves, flowers, fruit and roots (Laothawornkitkul *et al.* 2009). These VOCs are usually lipophilic molecules with high vapour pressures at ambient temperatures, and mainly consist of terpenoids, phenylpropanoids, and fatty acid and amino acid derivatives (Dudareva & Pichersky 2008). More than 1700 VOCs have been found in floral scents (Knudsen *et al.* 1993, 2006; Raguso 2008), and many more are emitted from other tissues (Dudareva *et al.* 2004). These VOCs may serve specific functions of attracting pollinators, fruit dispersers and parasitoids of herbivores, repelling herbivores, and also alerting neighbouring plants or neighbouring parts of the same plant about attacks by herbivores and pathogens (Gershenzon & Dudareva 2007; Felton & Tumlinson 2008; van Dam 2009; Dicke 2009). The function of some of these VOCs in biotic interactions is known, but that of the vast majority still remains to be discovered (Pichersky *et al.* 2006; Lewinsohn & Gijzen 2009). Besides biotic interactions, VOCs such as isoprene and monoterpenes have also been implicated in basic physiological functions within the plant, such as protection against thermal and oxidative damage (Owen & Peñuelas 2005). While plant species and habitats vary in their rates of VOC emission (Guenther 1997; Arneth *et al.* 2008; Lappalainen *et al.* 2009; Steinbrecher *et al.* 2009; Winters *et al.* 2009), the quantity of VOCs released is enormous, *e.g.* the global annual emission of biogenic VOCs is 700–1000 Tg C (Laothawornkitkul *et al.* 2009). Given the quantity and diversity of VOCs produced by plants in any given habitat, and the fact that some of these VOCs are emitted only or in greater amounts in certain contexts, the challenge for the biological entity interacting with the emitting plant is to pick out the signal against the background of VOC noise (van Dam & Poppy 2008). This interacting entity could be a mutualistic or antagonistic plant or animal or a researcher attempting to find a pattern of VOC emission in specific contexts, *e.g.* after herbivory or before pollination. Pattern recognition is therefore essential for successful communication using volatiles, and requires not only that one signal be differentiated from another signal but that noise within VOC blends is also ignored.

While techniques to measure and characterise VOCs are improving, and large amounts of VOC data are being generated (Fernie 2007), these efforts are not yet matched by suitable statistical analysis of data (van Dam & Poppy 2008; Loreto *et al.* 2008). According to van Dam & Poppy (2008), the field of plant volatile analysis needs to adopt methods from bioinformatics, a discipline that also deals with copious amounts of data, *e.g.* as in microarray analysis or genome-wide data mining. This search for new methods to analyse VOC data has led to the recent use of methods other than conventional principal component analysis (PCA), discrimi-

nant analysis (DA), or MANOVA, *e.g.* artificial neural network analysis (Cajka *et al.* 2009). In this paper, we demonstrate the use of Random Forests (RF), a machine-learning algorithm belonging to the class of data-mining techniques, in the analysis of VOC data. This new technique is being increasingly used in data-rich fields such as bioinformatics, chemoinformatics, medical diagnostics, food technology, astronomy and speech analysis, to select the most appropriate candidate variables from the surrounding data babel (Svetnik *et al.* 2003; Cannon *et al.* 2006; Díaz-Uriarte & de Andrés 2006; Granitto *et al.* 2007a,b; Zhang *et al.* 2008; Gao *et al.* 2009; Rong *et al.* 2009). We use data from Borges *et al.* (2008) on VOCs emitted by sympatric *Ficus* syconia in seed dispersal phase as an illustrative example. With this dataset, we also compare the performance of RF to other recent data-mining algorithms and tree-building methods, such as support vector machines (SVM), diagonal linear discriminant analysis (DLDA) and *k*-nearest neighbour analysis (KNN).

METHODS

Dataset

We used data published in Borges *et al.* (2008) on the volatiles produced by the syconia of sympatric *Ficus hispida*, *F. exasperata* and *F. tsjahela* at the seed dispersal stage. The headspace samples were collected from Agumbe Reserve Forest in the Indian Western Ghats (details of study site, volatile collection, composition and analysis are available in Borges *et al.* 2008). The dataset consisted of 49 samples and 77 VOCs. *F. hispida* and *F. exasperata* are dioecious species, in which only syconia on female trees produce seeds while those on male trees breed the pollinating mutualistic wasps. *F. tsjahela* is a monoecious species, in which all syconia produce seeds and wasps. Since seed dispersers should be attracted only to seed-bearing female syconia in the dioecious species because the male-bearing syconia contain developing and eclosing wasps, we predicted that the VOC signature of male and female ripened syconia should be distinct and different from each other. However, since in *F. exasperata* male and female ripened syconia are not usually available simultaneously owing to asynchrony between the sexes, we predicted that the VOC signature of male and female ripened syconia in this species could overlap, while they would be distinctive in the synchronous *F. hispida*. We found these predictions to be true (Borges *et al.* 2008). Using this dataset as a test case, we now address the following classification problem. We assume that in a forest where ripened syconia of these three *Ficus* species are available simultaneously, the problem to be solved by a biological entity, such as a seed disperser of *Ficus*, is to pick out the volatile signature of the ripened syconia of *F. tsjahela*, female *F. hispida* or female *F. exasperata* from other signatures in this data universe using the entire set of 77 VOCs recovered from these samples. Therefore, in this case, we are attempting to differentiate signal from noise within VOC blends produced by the five groups (male/female *F. hispida*, male/female *F. exasperata* and *F. tsjahela*), and also differentiate signal from signal (a unique VOC set to identify a group from the background) for each group. Moreover, since *F. hispida* is largely bat-dispersed and the other two species are largely bird-dispersed (Borges *et al.* 2008), we expect that the signatures of the seed-bearing syconia for each group should be distinguishable from the background or 'the rest' in this case in order to attract specific dispersal agents. We examine the efficacy of Random Forests (RF) in solving this classification problem. We emphasise that this classification problem was not attempted in Borges *et al.* (2008), in which clusters of the seed dispersal data were visualised with the help of PCA using only 33 out of the 77 VOCs (*i.e.* those VOCs with >5% occurrence in an individual sample).

VOC data and Random Forests

VOC data usually consist of samples of volatile emissions collected over several plant individuals or from the same plant over several time points. Whatever the type of sample or its method of collection, the key feature of a VOC dataset is that it is analogous to a microarray gene expression dataset in the sense that there are many more variables than samples. A typical headspace sample would include between 50 and 100 VOCs, just as a typical microarray would yield expression levels for hundreds of genes. This nature of a VOC dataset limited our use of the entire VOC dataset in earlier PCA analyses (Borges *et al.* 2008) and also limits the usage of classical multivariate analysis methods such as MANOVA or LDA. Conventional multivariate analysis methods require sample sizes to be proportionately increased for each added variable and also assume normality of the dataset, as well as the absence of auto-correlation between variables, besides having other limitations (Stevens 1992). Random Forests (Breiman 2001) is a classification algorithm with the following features that make it best suited for volatile analyses: (i) it allows for more variables than samples; (ii) it has a good classification efficiency, even with a lot of background noise; (iii) it is capable of arriving at a minimal set of variables, which can be used as predictors of that particular group; (iv) it is robust to interactions and correlations among variables; (v) it gives measures of relative variable importance (this objective and identifying the minimum variable set are more efficiently achieved using the *varSelRF* modifications to RF provided by Díaz-Uriarte & de Andrés 2006); and (vi) it can also be used to analyse time series data that record patterns in volatile emissions over time (achieved by the 'dyn' package available at http://cran.r-project.org). RF builds sets of decision trees using bootstrapping from the set of samples, and also selects a variable set of attributes (different VOCs in this case) at each node of the many decision trees thus generated. In this way, RF is also different from other tree-building methods such as PAUP: phylogenetic analysis using parsimony (Perdiguero-Alonso *et al.* 2008). RF also uses the unselected samples in a given bootstrap iteration to calculate an out-of-bag (OOB) error; *i.e.* the classification error obtained when the OOB (unselected samples) samples are examined; approximately one-third of the samples are unselected in each iteration. A major advantage of RF is that it does not overfit the data (Breiman 2001; Granitto *et al.* 2007a,b) so that even if minor fluctuations in variable strength (VOC concentrations or proportions in this case) lead to the building of thousands of classification trees, these fluctuations are not given undue importance in the final model; thus only the minimum set of important predictor variables is obtained.

Classification using RF

Between group classification

In this case, we retained the identity of the groups (*e.g.* female *F. hispida* or male *F. exasperata*) and attempted to find the prediction error of group membership.

One versus the rest classification

In this case, we determined the characteristic set of variables (VOCs) that define a particular group (*e.g.* female *F. hispida*) from all other samples, and for this purpose we masked the identity of the other groups (*the rest*). We attempted to find the smallest number of variables (VOCs) with which each group can be distinguished from *the rest*. The package *varSelRF* was used with R software version 2.9.0 (R Development Core Team 2009) for this purpose. The variable selection was allowed to run for 100 iterations. Running the algorithm for 1000 iterations did not yield a different result when compared with 100 iterations (data not shown). We therefore opted to use the results obtained from 100 iterations alone, as these are computationally intensive algorithms. For all analyses presented in this paper, only the proportions of the different VOCs present in the samples were used; this was purely for illustrative purposes; RF can use proportions or actual concentrations of each variable. A coefficient of variation (CV) was calculated for the predictor variables in each of the groups to determine whether RF consistently picks VOC predictors with low variability. An average out-of-bag (OOB) probability of membership in the groups (in the 'one' *versus* 'the rest' classification) was also estimated. This is the probability (calculated *à posteriori*) of a sample belonging to that group. We also obtained prediction errors for group membership for the bootstrapping procedures as well as the mean decrease in accuracy (MDA) when particular variables (VOCs) are removed from the model. The MDA provides an importance score for that variable. The '*importance*' function of the package *randomForest* for R was used to calculate MDA.

Comparative efficiency of Random Forests

To compare the relative performance of RF with other classification methods being currently evaluated in the literature for use in data-rich fields, we used the .632+ bootstrap method with 200 iterations for all methods. This bootstrap method is an improvement on the 'leave-one-out' cross-validation method and gives a better estimate of prediction error (Efron & Tibshirani 1997; Díaz-Uriarte & de Andrés 2006). We compared the prediction errors of support vector machines (SVM), diagonal linear discriminant analysis (DLDA), *k*-nearest neighbour (KNN) to RF. SVM is a data-mining algorithm that uses the concept of data kernels and support vectors that maximise the distance between parallel supporting planes between kernels of data (Bennett & Campbell 2000). DLDA is a type of linear discriminant analysis that uses maximum likelihood and diagonal covariance matrices (Dudoit *et al.* 2002). KNN is a classification method that classifies each sample based on minimum distance to *k* nearest neighbours of the sample (Hastie *et al.* 2001). The '*errorest*' function of the package *ipred* for R was used for SVM, the '*geSignatureBoot*' function of the package *geSignatures* for R was employed for DLDA and KNN, and the '*varSelRFBoot*' function of the package *varSelRF* was used for RF. All packages mentioned in this paper are freely obtainable from http://cran.r-project.org.

RESULTS

The OOB membership probability plots showed that *Ficus tsjahela*, as well as male and female *F. hispida*, were clearly separated from *the rest*, while male and female *F. exasperata* extractions were not clearly classifiable from *the rest* (Fig. 1). In the case of female *F. exasperata*, one sample actually grouped with *the rest*, while in male *F. exasperata* many more samples grouped with *the rest*. The variable selection procedure indicated similar trends (Table 1). The model frequency in this Table indicates the percentage of times the same predictor volatiles appeared in the 100 iterations. Female *F. hispida* extractions were clearly classifiable from *the rest* by 2-amyl acetate and iso-amyl acetate. This combination of

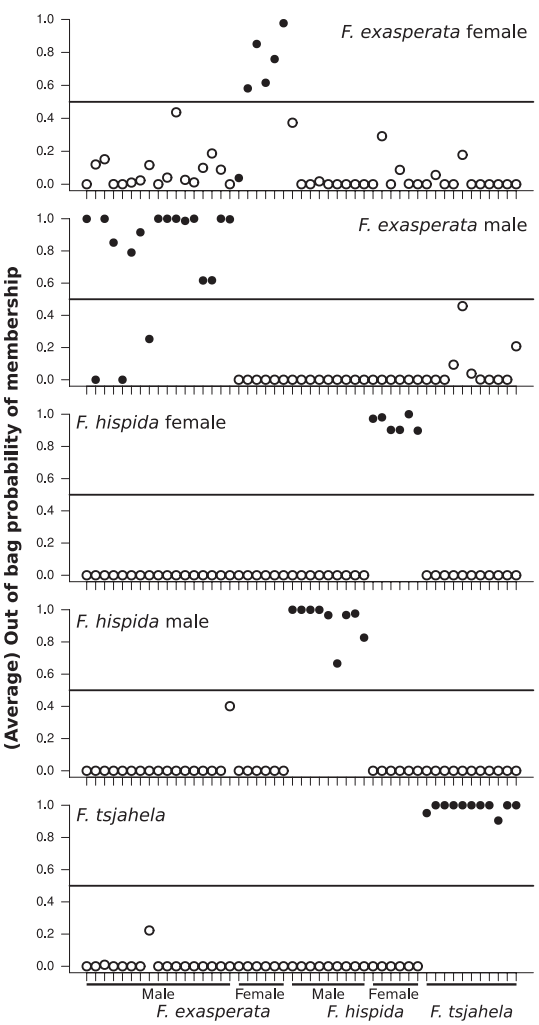

Fig. 1. The average out-of-bag (OOB) probability of membership of samples in the different groups. Samples that have a distinct volatile profile have an OOB probability close to one. Samples in the group of interest are indicated by filled circles; samples constituting *the rest* are indicated by open circles.

predictor compounds appeared in 100% of the iterations. Male *F. hispida* extractions were also uniquely classifiable by indole and α-trans-bergamotene. This combination of predictive compounds also appeared in 100% of the iterations. *F. tsjahela* had α-pinene and camphene that differentiated this group from *the rest* at 100% model frequency. On the other hand, male *F. exasperata* especially was relatively poorly classifiable and there were many misclassified samples with low model frequency (31%) (Table 1). This result for *F. exasperata* was as expected (see Discussion). Based on classifications achieved by RF, a biological entity (such as a seed disperser) could have less difficulty in finding female figs compared to finding male figs. This is an important and biologically relevant result (see Discussion). Additionally, the coefficient of variation (CV) was also lower for those predictor compounds whose contribution to the volatile signature of each sample was high (Table 1). The RF procedure also indicated the mean decrease in accuracy (MDA) in classification when the principal predictor compounds, as well as others, are removed from the model (Table 2). These results demonstrate the relative importance of certain VOCs in defining the volatile signature of the group.

RF was extremely good at classifying samples with the VOC data. For instance, out of the 77 seed dispersal VOCs found in the five groups examined, only 11 compounds (14% of the complete repertoire of compounds) (Table 1) were sufficient to classify all of the 49 samples with 5.84% error (Table 3). This translates to approximately only three samples being misclassified. RF clearly scored over the other tree-building and classification methods examined (Table 3). For example, the prediction error for RF was 5.8% compared to 34.6% for SVM, 22.1% for DLDA and 14.9% for KNN in the *between group* comparison (Table 3). The same lower prediction error for RF was generally found in the comparison of *one* versus *the rest* for all the other groups (Table 3). It must be noted that all classification methods evaluated in this paper performed better than when no information was provided other than representation of the different groups in the total number of samples (no information category in Table 3).

DISCUSSION

While there are several methods available for multivariate analysis and clustering, RF is ideally suited for the analysis of VOC data, as this case study of dispersal stage volatiles in sympatric *Ficus* has illustrated. RF enables the selection of candidate variables from a large dataset consisting of many more variables (77 VOCs) than samples (49 belonging to five groups) and also provides model statistics for these variables. Such variables can then form the basis of biological assays under controlled conditions. Information such as the mean decrease in accuracy when particular VOCs are removed from the model can also be used to determine candidate VOCs that are important in the particular VOC message that is intended for communication. It must be emphasised that we have merely used our dataset as an illustrative example of RF. Moreover, we attempted this classification exercise in the absence of knowledge of the overall background volatile landscape (*i.e.* volatiles produced by other plant species at the same site). As we discuss later, this is an important limitation of the dataset and not of the classification method.

Since plants and animals are faced with a complex volatile landscape, the characterisation of the statistical structure of this environment is an important first step towards understanding how volatile signals are encoded (Wright & Thomson 2005). Evaluating the statistics of visual scenes has correspondingly played an important role in understanding visual systems and how they function (Mackay 1986; Field 1987; Wright & Thomson 2005). In the case of VOCs too, understanding and characterising natural olfactory landscapes has been recently advocated, since many studies are now showing the effect of background compounds such as isoprene or various mixtures on olfactory perception and corresponding behavioural or physiological responses (Dicke *et al.* 2003; Mumm & Hilker 2005; Laothawornkitkul *et al.* 2008; Loivamäki *et al.* 2008). Conducting experiments on volatile communication in natural conditions using natural contexts is therefore being increasingly encouraged (Hunter 2002; Hale *et al.* 2009). However, moving from the controlled olfactory environments of the laboratory to complex natural olfactory landscapes will involve presenting and detecting signals where the signal-to-noise ratio is likely to be very low. Under this scenario, RF can be a powerful exploratory classification tool since it has been demonstrated to work extremely well with noise in a variety of noisy environments ranging from proteomics to astronomy (Gunther *et al.* 2003; Svetnik *et al.* 2003; Cannon *et al.* 2006; Díaz-Uriarte & de Andrés 2006; Granitto *et al.* 2007a,b; Kwak *et al.* 2008; Zhang *et al.* 2008; Gao *et al.*

Table 1. Predictor volatiles, the frequency of their occurrence in Random Forest models and their proportions in sample headspace.

species	sex	model frequency	predictor volatiles	percentage in headspace[a]	CV[a,b]
Ficus hispida	female	100	2-amyl acetate	63.3	0.3
			iso-amyl acetate	1.3	1.5
	male	100	indole	32.1	0.6
			α-trans-bergamotene	20.9	0.5
Ficus tsjahela	monoecious	100	α-pinene	31.5	0.2
			camphene	3.1	0.3
Ficus exasperata	female	98	γ-terpinene	21.7	0.5
			p-cymene	5.4	0.3
			β-caryophyllene	0.2	1.5
	male	31	daucene	2.9	1.0
			β-copaene	0.9	0.9

[a]Data from Borges *et al.* (2008).
[b]Coefficient of variation.

Table 2. Mean decrease in accuracy (MDA) when particular volatiles are removed from the model for each 'group' *versus* 'the rest'. Only non-zero values of MDA are reported.

male *Ficus exasperata* *versus* the rest	MDA	female *Ficus* *exasperata versus* the rest	MDA	male *Ficus hispida* *versus* the rest	MDA	female *Ficus hispida* *versus* the rest	MDA	*Ficus tsjahela* *versus* the rest	MDA
daucene	0.056	γ-terpinene	0.023	indole	0.044	iso-amyl acetate	0.022	camphene	0.053
β-copaene	0.032	p-cymene	0.010	α-trans-bergamotene	0.040	2-amyl acetate	0.021	α-pinene	0.048
allo-aromadendrene	0.029	β-caryophyllene	0.010	isolepidozene	0.019	2-nonyl acetate	0.016	α-terpinene	0.028
undecane	0.026	terpinolene	0.007	2-heptyl acetate	0.015	n-amyl acetate	0.015	β-pinene	0.022
γ-terpinene	0.021	γ-muurolene	0.006	(E)-ocimene	0.014	2-nonanone	0.012	α-muurolene	0.019
γ-muurolene	0.016	tridecane	0.006	longifolene	0.013	3-octenyl acetate	0.012	α-selinene	0.015
(E)-β-farnesene	0.011	α-trans-bergamotene	0.003	2-heptanone	0.010	γ-terpinene	0.009	2-phenyl ethanol	0.013
(Z)-ocimene	0.010	(Z)-3-hexenyl acetate	0.003	allo-aromadendrene	0.008	2-heptyl acetate	0.008	germacrene D	0.012
pentadecane	0.009	limonene	0.003	aromadendrene	0.007	α-pinene	0.007	δ-cadinene	0.007
β-elemene	0.006	tetradecane	0.003	δ-cadinene	0.007	(E)-ocimene	0.006	anisole	0.006
α-cis-bergamotene	0.006	isolepidozene	0.002	α-pinene	0.007	hexyl acetate	0.004	γ-terpinene	0.005
δ-cadinene	0.005	benzyl acetate	0.002	p-cymene	0.005	isolepidozene	0.003	allo-aromadendrene	0.005
sabinene	0.004	α-copaene	0.002	terpinolene	0.004	p-cymene	0.003	terpinolene	0.004
(E)-ocimene	0.004	(Z)-ocimene	0.002	isoledene	0.003	γ-muurolene	0.002	α-gurjunene	0.004
α-copaene	0.004	daucene	0.002	n-amyl acetate	0.003	α-cis-bergamotene	0.001	4-ethyl anisole	0.004
NI 2	0.004	1,8-cineole	0.002	γ-terpinene	0.003	α-trans-bergamotene	0.001	α-thujene	0.004
isoledene	0.004	α-thujene	0.002	β-pinene	0.003	α-copaene	0.001	α-copaene	0.003
p-cymene	0.004	allo-aromadendrene	0.002	iso-amyl acetate	0.002	limonene	0.001	undecane	0.003
β-pinene	0.003	(E)-ocimene	0.002	daucene	0.002	myrcene	0.001	isolepidozene	0.002
myrcene	0.003	α-pinene	0.001	2-amyl acetate	0.002	2-heptanone	0.001	1,8-cineole	0.002
hexadecane	0.003	2-heptyl acetate	0.001	methyl anthranilate	0.002	β-pinene	0.001	β-cedrene	0.002
indole	0.003	methyl salicylate	0.001	α-thujene	0.002	allo-aromadendrene	0.001	isoledene	0.002
β-santalene	0.003	indole	0.001	perillene	0.002	(E)-2-hexenyl acetate	0.001	myrcene	0.001
β-caryophyllene	0.003	isoledene	0.001	germacrene D	0.002	bicyclogermacrene	0.001	(E)-ocimene	0.001
α-pinene	0.003	α-humulene	0.001	α-copaene	0.001	germacrene D	0.001	sabinene	0.001
(Z)-3-hexenyl acetate	0.002	pentadecane	0.001	undecane	0.001	methyl salicylate	0.001	p-cymene	0.001
perillene	0.002	α-gurjunene	0.001	hexyl acetate	0.001	sabinene	0.001	NI 1	0.001
camphene	0.002	hexyl acetate	0.001	2 octyl acetate	0.001	(Z)-3-hexenyl acetate	0.001	(Z)-ocimene	0.001
tetradecane	0.002	(E)-β-farnesene	0.001	2-nonanone	0.001			α-trans-bergamotene	0.001
α-terpinene	0.001	sabinene	0.001	α-cis-bergamotene	0.001				

NI=not identified.

	no information	SVM	DLDA	KNN	random forest
between groups	0.6530	0.3462	0.2208	0.1496	0.0584
female *Ficus hispida versus* the rest	0.1224	0.0590	0.0011	0.0018	0.0177
male *Ficus hispida versus* the rest	0.1837	0.1267	NA	0.0161	0.0054
Ficus tsjahela versus the rest	0.2245	0.0984	0.0011	0.0010	0.0107
female *Ficus exasperata versus* the rest	0.1224	0.0878	NA	0.1341	0.0589
male *Ficus exasperata versus* the rest	0.3469	0.2256	0.1339	0.1255	0.0855

NA = not available.

Table 3. Error rates estimated for the different classification methods using the .632+ bootstrap method with 200 bootstrap samples. The 'no information' column denotes the error rate at random when information from the groups is not used (Díaz-Uriarte & de Andrés 2006).

2009; Rong *et al.* 2009). In the context of noise, it is also significant that the compounds picked as predictors by RF in our case study were mostly those that had very low CVs (Table 1). This is biologically relevant since a VOC that is to be used to constitute a signal with reliable information content should not vary greatly in its representation in the volatile signature. Low CVs of biologically relevant VOCs have been found recently in other systems, *e.g.* in orchid pollination (Salzmann *et al.* 2007). Therefore the variability in scent components is a critical factor in their inclusion into a signal, and identifying such VOCs as predictor variables can be important in generating testable hypotheses about communication using volatiles. The mean decrease in classification accuracy when individual volatiles are deleted from the model (Table 2) can also help in identifying a set of candidate VOCs suitable for biological assays when details of sensitivity of the receiver to these compounds are known from electroantennography or other types of investigation. Table 2 also indicates those VOCs whose inclusion into the classification model is equivalent, in the sense that the decline in model prediction accuracy on their addition or removal is the same. Such information is immensely useful, especially when samples of the natural background olfactory landscapes are also included into the classification problem, since these very compounds may constitute 'generic' VOC noise generated by basal plant physiology. Most researchers in this field also collect samples of ambient air when acquiring headspace samples of specific interest. These samples of ambient air usually remain unquantified. The inclusion of such samples into the classification problem would increase our understanding of the complexity of the communication problem under natural settings. The identification of predictor variables by RF can also be used to make predictions about samples outside the current dataset, a situation analogous to using a set of predictor genes or their protein products in bioinformatics to predict or diagnose a disease. Thus RF does answer the call made by van Dam & Poppy (2008) to induct methods employed in bioinformatics into the analysis of plant volatiles.

Many studies have demonstrated the superiority of RF over other classification methods, including newer data-mining methods such as SVM (Díaz-Uriarte & de Andrés 2006; Granitto *et al.* 2007a; Perdiguero-Alonso *et al.* 2008; Fusaro *et al.* 2009; Rong *et al.* 2009), as also shown with our case study. In our particular case study, we also demonstrate (Table 3) that some classification methods (*e.g.* DLDA or KNN) may perform better than RF when only a binary classification is required (*e.g.* female *Ficus hispida versus* 'the rest'); however, even RF can do better in some such cases (*e.g.* female *F. exasperata versus* 'the rest'). Furthermore, in our own case study, DLDA was unable to perform a binary classification in two cases (marked as NA) when no results could be obtained, while RF and KNN were able to perform classifications in such cases (Table 3). Methods such as DLDA and KNN also do much worse than RF when many groups need to be distinguished (Table 3). Furthermore, using PCA with this same dataset, Borges *et al.* (2008) were able to visualize well separated clusters of seed dispersal volatile groups only when compounds whose concentration was <5% in each individual sample were excluded, while RF found predictor variables with high model frequency even when these compounds occurred at <5% concentration [*e.g.* iso-amyl acetate in female *F. hispida* (1.3%), β-caryophyllene in female *F. exasperata* (0.2%), β-copaene in male *F. exasperata* (0.9%) and camphene in *F. tsjahela* (3.1%); Table 1]. It must also be noted that our dataset also contained many instances when particular VOCs were absent; *i.e.* the dataset had many zero values. With our dataset, RF had clear advantages over PCA, which in any case suffers from limitations of statistical interpretation. Overall, therefore, RF scores significantly over the other available methods, including PCA.

More importantly, RF also provided information on the probability of membership of each sample in the investigated group (Fig. 1); such information can be used to validate expected patterns, to examine reasons for the membership of particular samples in certain classifications or to detect outliers due to experimental error. For example, several male *F. exasperata* samples are misclassified with 'the rest'. This has a biological explanation. In *F. hispida*, dispersal stage figs are available simultaneously on trees of both sexes in the population (synchrony between sexes), while in *F. exasperata* there is low overlap between syconia production in the sexes (asynchrony between sexes) (see Borges *et al.* 2008). Therefore, while there should be selection on female figs to have a VOC profile different from male figs or 'the rest', such that only seed figs are consumed so that seeds are dispersed (female figs are well classified in both species; Fig. 1 and Table 1), selection for this difference is expected to be greater in dioecious species that produce male and female figs simultaneously in the same season, *i.e. F. hispida* in this case. Thus, there should be less selection pressure on *F. exasperata* than on *F. hispida* in the dispersal phase to make the male fig volatile signature different from that of seed figs or 'the rest', and this why several male *F. exasperata* figs are misclassified with 'the rest'. Furthermore, the ability of RF to separate female signatures from 'the rest' is biologically relevant since seed-bearing female figs, rather than wasp-containing male figs, should be consumed.

Another advantage of RF in chemical ecology is that it can deal not only with categorical variables (within a classification framework), such as volatile type, chirality, or any other fea-

ture of stereochemistry that one may want to include, but also with continuous variables, such as concentrations of volatiles or their ratios (in a regression framework) or a combination of both types of variables. Furthermore, as in all good model building, RF allows portions of the data to be used as training sets so that the tree-building algorithm can be refined (Breiman 2001). Such flexibility and feature diversity facilitate comprehensive exploration of volatile landscapes where complex algorithms may be required to determine how olfactory systems find their targets (Bruce *et al.* 2005; Pareja *et al.* 2009). Such algorithms can also deal with complex datasets that show great intraspecific variability in VOC emissions (*e.g.* Degen *et al.* 2004). From a biologically relevant perspective, using RF to find a limited set of predictor variables from a universe of 77 VOCs in our test dataset can provide a practicable set of compounds to be employed in biological assays with model seed dispersal agents, such as bats or birds, that are the natural dispersers of these *Ficus* species.

Despite the fact that RF has come into recent use in ecology, including forestry, parasitology and migratory movements (Prasad *et al.* 2006; Cutler *et al.* 2007; Iverson *et al.* 2008; Perdiguero-Alonso *et al.* 2008; Oppel *et al.* 2009), it has not yet been employed in understanding the chemical ecology of communication using volatiles in plants. Its versatility can be a great boon in this field and needs further examination. RF is itself in the process of being constantly evaluated (*e.g.* Amaratunga *et al.* 2008). We suggest that RF could be a valuable addition to the chemical ecologist's tool kit, a view that should, however, always be tempered by Wolpert's 'no free lunch theorem'; *i.e.* there is no one algorithm that can be universally suitable for all classification problems (Svetnik *et al.* 2003).

ACKNOWLEDGEMENTS

This research was funded by the Ministry of Environment and Forests, Government of India. We thank Mahua Ghara and Anusha Krishnan for useful discussions and the reviewers for useful comments.

REFERENCES

Amaratunga D., Cabrera J., Lee Y.-S. (2008) Enriched random forests. *Bioinformatics*, **24**, 2010–2014.

Arneth A., Schurgers G., Hickler T., Miller P.A. (2008) Effects of species composition, land surface cover, CO_2 concentration and climate on isoprene emissions from European forests. *Plant Biology*, **10**, 150–162.

Bennett K.P., Campbell C. (2000) Support vector machines: hype or hallelujah? *SIGKDD Explorations*, **2**, 1–13.

Borges R.M., Bessière J.-M., Hossaert-McKey M. (2008) The chemical ecology of seed dispersal in monoecious and dioecious figs. *Functional Ecology*, **22**, 484–493.

Breiman L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Bruce T.J.A., Wadhams L.J., Woodcock C.M. (2005) Insect host location: a volatile situation. *Trends in Plant Science*, **10**, 269–274.

Cajka T., Hajslova J., Pudil F., Riddellova K. (2009) Traceability of honey origin based on volatile pattern processing by artificial neural networks. *Journal of Chromatography A*, **1216**, 1458–1462.

Cannon E.O., Bender A., Palmer D.S., Mitchell J.B.O. (2006) Chemoinformatics-based classification of prohibited substances employed for doping in sport. *Journal of Chemical Information and Modeling*, **46**, 2369–2380.

Cutler D.R., Edwards T.C. Jr, Beard K.H., Cutler A., Hess K.T., Gibson J., Lawler J.J. (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

van Dam N.M. (2009) How plants cope with biotic interactions. *Plant Biology*, **11**, 1–5.

van Dam N.M., Poppy G.M. (2008) Why plant volatile analysis needs bioinformatics – detecting signal from noise in increasingly complex profiles. *Plant Biology*, **10**, 29–37.

Degen T., Dillmann C., Marion-Poll F., Turlings T.C.J. (2004) High genetic variability of herbivore-induced volatile emission within a broad range of maize inbred lines. *Plant Physiology*, **135**, 1928–1938.

Díaz-Uriarte R., de Andrés S.A. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Dicke M. (2009) Behavioural and community ecology of plants that cry for help. *Plant, Cell and Environment*, **32**, 654–665.

Dicke M., de Boer J.G., Höfte M., Rocha-Granados M.C. (2003) Mixed blends of herbivore-induced plant volatiles and foraging success of carnivorous arthropods. *Oikos*, **101**, 38–48.

Dudareva N., Pichersky E. (2008) Metabolic engineering of plant volatiles. *Current Opinion in Biotechnology*, **19**, 1–9.

Dudareva N., Pichersky E., Gershenzon J. (2004) Biochemistry of plant volatiles. *Plant Physiology*, **135**, 1893–1902.

Dudoit S., Fridlyand J., Speed T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77–87.

Efron B., Tibshirani R.J. (1997) Improvements on cross-validation: the .632+ method. *Journal of the American Statistical Association*, **92**, 548–560.

Felton G.W., Tumlinson J.H. (2008) Plant–insect dialogs: complex interactions at the plant–insect interface. *Current Opinion in Plant Biology*, **11**, 457–463.

Fernie A.R. (2007) The future of metabolic phytochemistry: larger numbers of metabolites, higher resolution, greater understanding. *Phytochemistry*, **68**, 2861–2880.

Field D.J. (1987) Relations between the statistics of natural images and the response profiles of cortical cells. *Journal of the Optical Society of America A*, **4**, 2379–2394.

Fusaro V.A., Mani D.R., Mesirov J.P., Carr S.A. (2009) Predication of high-responding peptides for targeted protein assays by mass spectrometry. *Nature Biotechnology*, **27**, 190–198.

Gao D., Zhang Y.-X., Zhao Y.-H. (2009) Random forest algorithm for classification of multiwavelength data. *Research in Astronomy and Astrophysics*, **9**, 220–226.

Gershenzon J., Dudareva N. (2007) The function of terpene natural products in the natural world. *Nature Chemical Biology*, **3**, 408–414.

Granitto P.M., Gasperi F., Biasioli F., Trainotti E., Furlanello C. (2007a) Modern data mining tools in descriptive sensory analysis: a case study with a random forest approach. *Food Quality and Preference*, **18**, 681–689.

Granitto P.M., Biasioli F., Aprea E., Mott D., Furlanello C., Märk T.D., Gasperi F. (2007b) Rapid and non-destructive identification of strawberry cultivars by direct PTR-MS headspace analysis and data mining techniques. *Sensors and Actuators B, Chemical*, **121**, 379–385.

Guenther A. (1997) Seasonal and spatial variations in natural volatile organic compound emissions. *Ecological Applications*, **7**, 34–45.

Gunther E.C., Stone D.J., Gerwien R.W., Bento P., Heyes M.P. (2003) Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles *in vitro*. *Proceedings of the National Academy of Sciences USA*, **100**, 9608–9613.

Hale R., Swearer S.E., Downes B.J. (2009) Separating natural responses from experimental artefacts: habitat selection by a diadromous fish species using odours from conspecifics and natural stream water. *Oecologia*, **159**, 679–687.

Hastie T., Tibshirani R., Friedman J. (2001) *The elements of statistical learning*. Springer, New York, USA.

Hunter M.D. (2002) A breath of fresh air: beyond laboratory studies of plant volatile–natural enemy interactions. *Agricultural and Forest Entomology*, **4**, 81–86.

Iverson L.R., Prasad A.M., Matthews S.N., Peters M. (2008) Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *Forest Ecology and Management*, **254**, 390–406.

Knudsen J.T., Tollsten L., Bergstrom L.G. (1993) Floral scents – a checklist of volatile compounds isolated by headspace techniques. *Phytochemistry*, **33**, 253–280.

Knudsen J.T., Eriksson R., Gershenzhon J., Ståhl B. (2006) Diversity and distribution of floral scent. *The Botanical Review*, **72**, 1–120.

Kwak J., Willse A., Matsumura K., Opiekum M.C., Yi W., Preti G., Yamazaki K., Beauchamp G.K. (2008) Genetically-based olfactory signatures persist despite dietary variation. *PLoS ONE*, **3**, e3591.

Laothawornkitkul J., Paul N.D., Vickers C.E., Possell M., Taylor J.E., Mullineaux P.M., Hewitt C.N. (2008) Isoprene emissions influence herbivore feeding decisions. *Plant, Cell and Environment*, **31**, 1410–1415.

Laothawornkitkul J., Taylor J.E., Paul N.D., Hewitt C.N. (2009) Biogenic volatile organic compounds in the Earth system. *New Phytologist*, **183**, 27–51.

Lappalainen H.K., Sevanto S., Bäck J., Ruuskanen T.M., Kolari P., Taipale R., Kulmala M., Hari P. (2009) Day-time concentrations of biogenic volatile organic compounds in a boreal forest canopy and their relation to environmental and biological factors. *Atmospheric Chemistry and Physics Discussions*, **9**, 6247–6281.

Lewinsohn E., Gijzen M. (2009) Phytochemical diversity: the sounds of silent metabolism. *Plant Science*, **176**, 161–169.

Loivamäki M., Mumm R., Dicke M., Schnitzler J.-P. (2008) Isoprene interferes with the attraction of bodyguards by herbaceous plants. *Proceedings of the National Academy of Sciences USA*, **105**, 17430–17435.

Loreto F., Kesselmeier J., Schnitzler J.-P. (2008) Volatile organic compounds in the biosphere–atmosphere system: a preface. *Plant Biology*, **10**, 2–7.

Mackay D.M. (1986) Vision – the capture of optical variation. In: Pettigrew J.D., Sandison K.T., Levick W.R. (Eds), *Visual neuroscience*. Cambridge University Press, Cambridge, UK: pp. 365–373.

Mumm R., Hilker M. (2005) The significance of background odour for an egg parasitoid to detect plants with host eggs. *Chemical Senses*, **30**, 337–343.

Oppel S., Powell A.N., Dickson D.L. (2009) Using an algorithmic model to reveal individually variable movement decisions in a wintering sea duck. *Journal of Animal Ecology*, **78**, 524–531.

Owen S.M., Peñuelas J. (2005) Opportunistic emissions of volatile isoprenoids. *Trends in Plant Science*, **10**, 420–426.

Pareja M., Mohib A., Birkett M.A., Dufour S., Glinwood R.T. (2009) Multivariate statistics coupled to generalized linear models reveal complex use of chemical cues by a parasitoid. *Animal Behaviour*, **77**, 901–909.

Perdiguero-Alonso D., Montero F.E., Kostadinova A., Raga J.A., Barrett J. (2008) Random forests, a novel approach for discrimination of fish populations using parasites as biological tags. *International Journal for Parasitology*, **38**, 1425–1434.

Pichersky E., Sharkey T.D., Gershenzon J. (2006) Plant volatiles: a lack of function or a lack of knowledge? *Trends in Plant Science*, **11**, 421.

Prasad A.M., Iverson L.R., Liaw A. (2006) Newer classification and regression tree techniques: Bagging and Random Forests for ecological prediction. *Ecosystems*, **9**, 181–199.

R Development Core Team (2009) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Raguso R.A. (2008) Wake up and smell the roses: the ecology and evolution of floral scent. *Annual Review of Ecology, Evolution, and Systematics*, **39**, 549–569.

Rong J., Li G., Chen Y.-P.P. (2009) Acoustic feature selection for automatic emotion recognition from speech. *Information Processing and Management*, **45**, 315–328.

Salzmann C.C., Nardella A.M., Cozzolino S., Schiestl F. (2007) Variability in floral scent in rewarding and deceptive orchids: the signature of pollinator-imposed selection? *Annals of Botany*, **100**, 757–765.

Steinbrecher R., Smiatek G., Köble R., Seufert G., Theloke J., Hauff K., Ciccioli P., Vautard R., Curci G. (2009) Intra- and interannual variability of VOC emissions from natural and semi-natural vegetation in Europe and neighbouring countries. *Atmospheric Environment*, **43**, 1380–1391.

Stevens J. (1992) *Applied multivariate statistics for the social sciences*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Svetnik V., Liaw A., Tong C., Culberson J.C., Sheridan R.P., Feuston B.P. (2003) Random Forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, **43**, 1947–1958.

Winters A.J., Adams M.A., Bleby T.M., Rennenberg H., Steigner D., Steinbrecher R., Kreuzwieser J. (2009) Emissions of isoprene, monoterpene and short-chained carbonyl compounds from *Eucalyptus* spp. in southern Australia. *Atmospheric Environment*, **43**, 3035–3043.

Wright G.A., Thomson M.G.A. (2005) Odor perception and variability in natural odor scenes. In: Romeo J. (Ed.), *Recent advances in phytochemistry*. Elsevier, Amsterdam, pp 191–226.

Zhang J., Sokal I., Peskind E.R., Quinn J.F., Jankovic J., Kenney C., Chung K.A., Millard S.P., Nutt J.G., Montine T.J. (2008) CSF multianalyte profile distinguishes Alzheimer and Parkinson disease. *American Journal of Clinical Pathology*, **129**, 526–529.