BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 6 2009, pages 787–794
doi:10.1093/bioinformatics/btp056

Systems biology

Automated procedure for candidate compound selection in GC-MS metabolomics based on prediction of Kovats retention index

V. V. Mihaleva[1], H. A. Verhoeven[2], R. C. H. de Vos[2], R. D. Hall[2] and R. C. H. J. van Ham[1,3,*]

[1]Applied Bioinformatics, Plant Research International, [2]Centre for BioSystems Genomics (CBSG), Droevendaalsesteeg 1 and [3]Laboratory of Bioinformatics, Wageningen University, Dreijenlaan 3, Wageningen, The Netherlands

Received on July 21, 2008; revised on December 11, 2008; accepted on January 24, 2009

Advance Access publication January 28, 2009

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Matching both the retention index (RI) and the mass spectrum of an unknown compound against a mass spectral reference library provides strong evidence for a correct identification of that compound. Data on retention indices are, however, available for only a small fraction of the compounds in such libraries. We propose a quantitative structure-RI model that enables the ranking and filtering of putative identifications of compounds for which the predicted RI falls outside a predefined window.

Results: We constructed multiple linear regression and support vector regression (SVR) models using a set of descriptors obtained with a genetic algorithm as variable selection method. The SVR model is a significant improvement over previous models built for structurally diverse compounds as it covers a large range (360–4100) of RI values and gives better prediction of isomer compounds. The hit list reduction varied from 41% to 60% and depended on the size of the original hit list. Large hit lists were reduced to a greater extend compared with small hit lists.

Availability: http://appliedbioinformatics.wur.nl/GC-MS

Contact: roeland.vanham@wur.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Metabolomics is a rapidly evolving field that aims to provide an unbiased, qualitative and quantitative characterization of the metabolites present in a biological sample. A typical biological sample contains hundreds of metabolites present in a wide range of concentrations. Gas chromatography coupled to mass spectrometry (GC-MS) is currently the method of choice for analysis of volatile organic compounds (VOCs) due to its high sensitivity and speed of detection. The method provides both mass spectra and retention time as characteristics of the measured compounds. Identification is usually done by comparing the measured spectrum with the spectra in a reference library. The number of possible identifications obtained in such a comparison depends on the search criteria used and is determined either by specifying the number of hits that should be returned or by requesting all hits with a matching factor (MF) above a specified value. The MF between the experimental and the reference mass spectrum depends on the quality of the machine output and the way the raw data are processed. In metabolomics approaches such as VOC-profiling, it is mostly impossible to completely separate the compounds contained in a complex biological sample. Overlaps in chromatographic separation may hamper the deconvolution of experimental sample spectra. As a result, the extracted individual spectra may be incomplete (missing m/z values) or contain fragments of other components (additional m/z values). Therefore, relaxed search criteria must often be used to ensure that compounds are not missed, even though this may come at the expense of long lists of hits that need to be evaluated.

The list of candidate compound identifications using GC-MS can be reduced by taking the retention time into consideration (Adams, 2001; Eckel and Kind, 2003). The chromatographic retention time of a compound is dependent on experimental conditions such as column type, temperature program and gradient. The retention index (RI) proposed by Kovats (1958) is therefore used as a standardized parameter in reporting GC data. Although the system of RI allows for direct comparison of GC data from different runs, instruments and laboratories, the currently available libraries on RI have a limited coverage and are much smaller compared with mass spectral (MS) libraries. For example, within the widely used NIST05 library (Ausloos *et al*., 1999), experimental data on RI is available for only 9% of the compounds. Because standards are not always available or difficult to chemically synthesize, it would be extremely useful to have a reliable and broadly applicable method for prediction of the RI for those compounds without an experimentally determined RI.

Quantitative structure–retention relationship (QSRR) models provide an estimation of RI based on descriptors derived from the chemical structure of the compounds. The QSRR models developed in the last decade have been summarized in the recent review by Heberger (2007). For model generation, multivariate methods such as multiple linear regression (MLR) (Farkas *et al*., 2004; Hemmateenejad *et al*., 2007; Hu *et al*., 2005;

*To whom correspondence should be addressed.

Jalali-Heravi and Kyani, 2004; Rayne and Ikonomou, 2003; Safa and Hadjmohammadi, 2005), artificial neural networks (Hemmateenejad *et al*., 2007; Jalali-Heravi and Kyani, 2004) and support vector machines (SVM) (Luan *et al*., 2005) have most frequently been used. The main drawback of most of these models is that they have been built for particular chemical classes (Farkas *et al*., 2004; Hemmateenejad *et al*., 2007; Hu *et al*., 2005; Jalali-Heravi and Kyani, 2004; Luan *et al*., 2005; Rayne and Ikonomou, 2003; Safa and Hadjmohammadi, 2005) and cover relatively small numbers of compounds. This greatly limits their use in untargeted analysis of crude biological samples, which may contain hundreds of metabolites from a broad range of compound classes (Tikunov *et al*., 2005). To our knowledge, only two methods have been published so far that can in principle deal with structurally diverse compounds. Garkani-Nejad *et al.* (2004) modeled RI on a relatively small set of compounds (846) relevant to toxicology. A limitation of their model is that it does not cover molecules with RI smaller than 1100 units. More recently, Stein *et al.* (2007) developed a method based on estimating the contribution of a large set of functional groups to the RI. A drawback of this approach is that the use of functional groups alone does not allow for discrimination between isomers, which is important because isomers commonly have different biological activities (Constantinou *et al*., 2008; Fitzgerald *et al*., 2005; Kashfi *et al*., 2005; Preuss *et al*., 2006; Umemura *et al*., 1996).

In this study, we present MLR and support vector regression (SVR) models for the prediction of the RI based on a large set of compounds (22 690) from a wide range of chemical classes. The RI predicted by the SVR model was used to rank and filter out potentially false positive annotations obtained from searching mass spectra against the NIST05 MS library. The ranking of the hits was determined by the relative error obtained by comparing the experimental RI with the predicted RI. The proposed model was then tested using a sample consisting of a mixture of standard compounds and a biological sample of tomato fruit volatiles for which a set of putative identifications had previously been established (Tikunov *et al*., 2005). We show that our method is able to detect likely false identifications with performance better than the method proposed by Stein *et al.* (2007), especially with regard to the RI prediction of isomers. The hit list reduction ranges from 41% to 60% when 3 and 10 hits per experimental mass spectrum were retrieved, respectively. The last hit list contained all identified compounds and only one was filtered out. The procedure allows for the retrieval of more hits per experimental spectrum for low abundant metabolites as, in general, extraction of a pure experimental spectrum of these metabolites is difficult. The ranking and filtering algorithm has been implemented as a Python 2.5 procedure.

2 METHODS

2.1 Compound selection

Compounds with experimentally determined Kovats retention indices were obtained from the NIST05 library in a structure distribution format (SDF). The file contained 120 757 records and provided the name, active phase, temperature program used and 2D structure. Multiple records were present for many compounds, reporting the RIs for different stationary phases and/or temperature programs. Data were available for non-polar or slightly polar (up to 5% phenyl groups) stationary phases. It has been shown by Stein *et al.* (2007) that, on average, RI determined at different temperature programs and/or column types (capillary or packed) varied within 12 RI units. Therefore, the median RI was used for compounds with multiple records. Unique compounds were defined by means of the InChI (Stein *et al*., 2003) string generated from the SDF file using Openbabel (http://openbabel.sourceforge.net). This resulted in an initial, non-redundant set of 24 509 compounds.

2.2 MLR model and modeling sets

The relation between RI and the compound structure was described as a linear function of a set of predefined descriptors:

$$\mathrm{RI}_{\mathrm{mlr}} = c_0 + \sum_{i=1}^{n} c_i D_i \quad (1)$$

where c_0 is the intercept, c_i is the regression coefficient of descriptor D_i and n is the number of descriptors.

Many different types of descriptors have been proposed in the literature to encode physico-chemical and electronic properties and topology, geometry, size and shape. As the structures used here were available as 2D coordinates only, geometry optimization would be required to obtain reliable 3D structures. Because this step is computationally costly and difficult to automate, we only used the 2D structures to generate several groups of descriptors. These were calculated using Dragon (Todeschini *et al*., 2003) and include constitutional, topological, walk and count paths, connectivity indices, topological charge indices, functional groups count, atom-centered fragments and molecular properties. Previous study has shown that there is a strong correlation between RI and the boiling point (BP) of a compound (Eckel and Kind, 2003). Experimental data on BP were available for only 2909 compounds. The BP of the remaining compounds was estimated using the MPBPVP program (Stein and Brown, 1994), part of the EPI suite (http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm). In total, a set of 586 descriptors was obtained. After removing descriptors with zero values or nearly constant values and descriptors with a correlation coefficient with the other descriptors greater than 0.90, 159 descriptors remained. The values for different descriptors varied several orders of magnitude and were therefore scaled to unit variance before modeling to ensure an equal weight of the descriptors in the regression model.

Compounds with extreme descriptor values might influence the MLR model. The leverage of each compound was estimated and used to detect outliers in the chemical space defined by the descriptors. Leverages were calculated as the diagonal elements of the hat matrix, $X(X'X)^{-1}X'$, where X is the descriptor matrix. A cutoff value of $2p/n$was used, where p is the number of descriptors and n is the number of compounds. In total, 1819 compounds had leverage greater than this cutoff. The majority of these compounds were metal complexes, compounds containing boron atoms, highly halogenated compounds (especially fluorinated), cyclic siloxanes, polyglycols and small molecules of up to five atoms including a herero-atom. Since most of these compounds are of non-biological origin, they were excluded from the analysis.

The final set used for modeling comprised 22 690 compounds and covered a RI range from 360 to 4120 units. The RI data were sorted and binned into groups of 30 compounds each. Compounds were then randomly selected from these bins and equally partitioned over training, monitoring and test sets.

2.3 Descriptor selection

Descriptor selection in QSRR modeling is essential to ensure that the models are robust and attain optimal predictive power. Moreover, a smaller number of predictive descriptors allows for an easier interpretation of the model. Genetic algorithms (GA) are global optimization methods that have been successfully applied in QSAR studies for descriptor selection (Gao *et al*., 2002; Hasegawa *et al*., 1997; Rogers and Hopfinger, 1994). At each step of the GA optimization procedure, many models are evaluated and the information of the fittest models is propagated to the next step. Descriptors in a GA procedure are combined in a linear string to form chromosomes

and are encoded by '1' if the descriptor is included in a next generation of the model and by '0' if it is excluded from the model in that generation. Here, the length of the chromosomes was set to 159—the total number of descriptors. At the start, a population of 100 individuals was created by randomly selecting the descriptors. The coefficients c_i of Equation (1) were determined using the training set and the root mean square error (RMSE) of prediction of the monitoring set was obtained for each model. A linear ranking was used to assess the fitness of each individual based on RMSE and fitness-proportional selection of the chromosomes was applied to generate subsequent generations. The uniform crossover probability was set to 0.7 and the probability for mutation was set to 0.05. These values were within the range of the commonly used values in GA procedures (Broadhurst *et al*., 1997; Lucasius and Kateman, 1993). To obtain a small set of descriptors, chances for the direction of a mutation were set to 90% for flipping 1 to 0 and 10% for flipping 0 to 1. Three GA optimization runs were performed, starting from different initial populations and using 1000 generations per run. Ten percent of the best chromosomes were selected for further analysis. The data analysis and modeling were performed using the GA toolbox (http://www.shef.ac.uk/acse/research/ecrg/gat.html) and in-house written Matlab 7.0 scripts.

2.4 SVR model

In SVR, the descriptor matrix is first mapped into a higher dimensional feature space by the use of a kernel function, and then a linear model is constructed in this feature space. In-depth theoretical background on SVR can be found in the introductions by Cristianini and Shauwe-Taylor (2000) and Vapnik (1995). A radial basis function (RBF) was used as a kernel function. The optimal values of the SVR parameters, the regularization parameter C, ε of the ε-insensitive loss function and the width (γ) of RBF, were found by a grid search ($C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$, $\varepsilon = 2^{-8}, 2^{-7}, \ldots, 2^{3}$ and $\gamma = 2^{3}, 2^{1}, \ldots, 2^{-11}$). LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm) was used as SVR implementation.

2.5 Calibration setup

Fifty analytical grade chemicals (Sigma) with known RI were measured for the calibration of the experimental setup. The compounds were pooled in 10 groups of 5 compounds with well-separated RI's within each group. Per vial, 100 μg of each compound was added. The headspace volatiles present in the vials were automatically extracted by exposing a 65 mm polydimethylsiloxane-divinylbenzene SPME fiber (Supelco) to the vial head space for 20 min under continuous agitation and heating at 50°C, and injected into the GC-MS via a Combi PAL autosampler (CTC Analytics AG). The fiber was inserted into a GC 8000 (Fisons Instruments) injection port and volatiles were desorbed for 1 min at 250°C. Chromatography was performed on an HP-5 column (50 m length, 0.32 mm diameter and 1.05 μm film thickness) using helium as carrier gas (37 kPa). The GC interface and MS source temperatures were 260°C and 250°C, respectively. The GC temperature program began at 45°C (2 min), was then raised to 250°C at a rate of 5°C/min, and finally held at 250°C for 5 min. The total run time, including oven cooling, was 65 min. Mass spectra in the 35–400 m/z range were recorded by an MD800 electron impact MS (Fisons Instruments) at a scanning speed of 2.8 scans/s and an ionization energy of 70 eV. The collected data are presented in Supplementary Material (Table S1).

2.6 Filtering of library hits

All raw GC-MS data were processed using AMDIS (Stein, 1999) and the extracted mass spectra were searched against the NIST05 library. The set of 50 compounds with known RI was measured and the system was calibrated to obtain the experimental RI, denoted as RI_{exp}. The final, GA-selected set of descriptors was calculated for all records in the NIST05 library and their RI was predicted using the SVR model. The predicted RI, denoted as RI_{SVR}, was saved as a text file together with the Chemical Abstracts Service (CAS) number and the name belonging to a particular record. For each hit, the predicted RI_{SVR} was retrieved by searching the CAS number or the name when the CAS number was not available. The relative RI error was determined as:

$$\%\ \text{rel.error} = 100 \times \left(RI_{SVR} - RI_{exp}\right)/RI_{exp}. \quad (2)$$

The distribution of the relative error was used to obtain thresholds to define the rank of the NIST05 library hits. A normal distribution fit of the relative error resulted in a mean value (μ) of 0.94 and a standard deviation (SD, σ) of 5.28. Hence, threshold values of 1σ, 2σ and larger than 2σ were used to define ranks one, two and three, respectively, which in turn correspond to 'good', 'moderate' and 'poor' agreement between the experimental and predicted RI. In the final list, only hits ranked one and two were retained. The filtering procedure was fully automated as a Python 2.5 script.

2.7 Test samples

2.7.1 Mixture of standard compounds A mixture of 22 analytical grade chemicals was analyzed by GC-MS. The sample was prepared by adding 10 μl of each liquid standard and 5 mg of each solid standard in a 4 ml screw cap vial. Then, 10 μl of this mix was transferred into a 10 ml screw cap vial and diluted with 5 ml ethyl acetate and 1 μl of this diluted standard mixture was injected into the GC-MS. The same instrument and experimental conditions were used as described for the calibration.

2.7.2 Biological samples The raw GC-MS data of a single tomato fruit volatiles sample was selected from a data set of 188 samples (Tikunov *et al*., 2005). In this particular sample, the majority of the previously identified metabolites were present and it was therefore most suitable for testing the model. In addition to the tomato fruit sample, melon and rice samples were used to estimate the fraction of library hits that were filtered out by our procedure. One gram of ground rice (*Oryza sativa* L., cultivar Perurutong) was weighed and introduced into a 10 ml glass vial and capped. The sample was rotated for 24 h to achieve saturation of the headspace with all volatiles. The headspace was sampled with an SPME device, using the blue polydimethylsiloxane/divinylbenzene (PDMS/DVB) fiber during 30 min at room temperature. Samples of melon fruits of different ripening stages of the varieties *Galia* and *Charantais* were mixed together. Sample preparation and incubation was done following the same protocol as for the tomato fruit sample (Tikunov *et al*., 2005). Injection, desorption and GC-MS analysis of all biological samples were performed with the same parameters and settings as described for the calibration.

3 RESULTS AND DISCUSSION

3.1 Regression models performance

3.1.1 MLR model The relationship between the structure of a compound and its RI was modeled as a linear function of a set of descriptors. The GA descriptor selection procedure on an initial set of 159 descriptors was repeated three times by randomly selecting different initial populations. The frequency of selection of the descriptors in the best 10 000 chromosomes (10% of the total number of chromosomes) in the three repetitions is shown in Supplementary Material (Fig. S1). These frequencies were used to determine the order in which the descriptors were selected to enter the model. A model of 19 descriptors was selected as the model with the smallest number of descriptors and small RMSE, μ and σ values. The descriptors and their coefficients in the regression function are listed in Table 1. The GA-selected set of descriptors can be divided into two main groups, namely descriptors related to properties of the molecule as a whole (12 descriptors) and descriptors accounting for specific atom types or functional groups (7 descriptors). As the descriptors were scaled to unit variance, the

Table 1. Regression coefficients of the MLR-GA selected descriptors

Number	Label	Description (type)	Coefficient ± SD
68	X1v	Valence connectivity index chi-1 (4)	252.41 ± 6.33
75	X3sol	Solvation connectivity index chi-3 (4)	221.88 ± 6.49
159	BP	BP (8)	155.26 ± 3.88
156	TPSA	Fragment-based polar surface area calculated using N,O polar coefficients (8)	153.47 ± 3.80
12	nHM	Number of heavy atoms (nHA) (1)	−142.09± 2.86
101	nCb-	Number of substituted benzene carbon atom (Csp2) (6)	111.67 ± 3.78
22	w	Detour index (2)	100.69 ± 3.61
30	DELS	Molecular electrotopological variation (2)	−81.59± 4.17
10	nO	Number of oxygen atoms (1)	−60.53± 3.12
81	GGI6	Topological charge index of order 6 (5)	−50.68± 3.04
42	ICR	Radial centric information index (2)	48.72 ± 2.13
158	ALOGP2	Squared Ghose–Crippen octanol–water partition coefficient (8)	−44.23 ± 2.65
122	C-001	CH3R/CH4 (7)	−36.72± 1.64
35	PW3	Path/walk 3—Randic shape index(2)	−33.70± 2.86
63	PCD	Difference between multiple path count and path count (3)	33.24 ± 3.17
102	nCconj	Number of non-aromatic conjugated carbon atom (Csp2) (6)	30.84 ± 1.30
13	nX	Number of halogen atoms (1)	17.53 ± 3.75
37	PW5	Path/walk 5—Randic shape index (2)	−15.45± 2.37
109	nRCONR2	Number of tertiary aliphatic amides (6)	9.02 ± 1.27

The descriptor number in the final set of 159 descriptors. The type is constitutional (1), topological (2), walk and path counts (3), connectivity indices (4), topological charge indices (5), functional groups count (6), atom-centred fragments (7), molecular properties (8).

magnitude of the coefficients is related to the importance of the descriptors for determining RI. A high-positive coefficient was found for the first group including TPSA (NO), the valence (X1v) and solvation (X3sol) connectivity indices and the BP. These descriptors are related to size, shape and the degree of branching of the molecule. The X1v descriptor takes into account the presence and position of the hetero atoms in the molecule. The X3sol descriptor includes the dispersion interaction with the stationary phase. The Detour index, w, is a descriptor that can discriminate between acyclic and cyclic compounds (Trinajstic *et al*., 1997). The Randic shape indices PW3 and PW5 are topological descriptors obtained as the quotient of

Table 2. Regression model evaluation

	R^2		RMSE		Number of compounds
	MLR	SVR	MLR	SVR	
Training	0.9756	0.9818	104	90	7573
Monitoring	0.9799	0.9818	98	93	7756
Test	0.9685	0.9720	121	114	7361

Correlation coefficient (R^2) and RMSE for the MLR and the SVR models. The values for the training set are based on 10-fold cross-validation. The monitoring set was used for the GA variable selection and the SVR parameters selection.

paths and walks of length 3 and 5, respectively (Randic, 2001). These descriptors account for the isomeric variation of molecules of the same size. Disperse interactions between the compounds and the stationary phase govern the retention behavior on non-polar columns and this is reflected by the set of descriptors with a positive effect on RI. Descriptors that showed a strong negative correlation with RI included the nHM and high polarity (DELS and GGI6). All compounds, for which nHM had a non-zero value, were trimethylsilane derivatized compounds. The addition of the bulky $Si(CH_3)_3$ group not only increases the size of the molecule but also the degree of branching and the compactness of the molecule. Hence, this represents the negative influence of the derivatization on RI. The charge transfer in the molecule is represented by DELS. With increased polarity of the molecule, the interaction with the non-polar stationary phase is weakened. As a result, a large value of the DELS descriptors (high polarity) will result in lower RI values.

The results of predicting the three different sets are listed in Table 2. The correlation coefficient R^2 and RMSE were similar for all sets. The values of RMSE were in the interval 98–121. A somewhat smaller value, 79 RI units, was found by Garkani-Nejad *et al.* (2004) for a set of 846 structurally diverse compounds and an MLR model. In their model, the connectivity indices were also the descriptors with the highest positive coefficients. BP in combination with 3D descriptors was used by Farkas *et al.* (2004) to predict RI for a set of oxygen, nitrogen and sulfur containing saturated heterocyclic compounds. In our model, BP of the majority of the compounds is estimated from the structure (Stein and Brown, 1994) and for some compounds the error of the BP estimation might be significant. This may have contributed to the final error of the RI estimation.

3.1.2 SVR model Non-linear transformation of the chemical space defined by the GA-selected descriptors was done using a RBF kernel. A grid search was used to find the best values for the SVR parameters. For each combination of the parameters, the model was constructed using the training set and evaluated by the RMSE of the monitoring set. The best model was obtained with $C = 2$, $\gamma = 2^{-5}$ and $\varepsilon = 2^{-8}$. Compared with the MLR model, the SVR model resulted in higher R^2 and smaller RMSE values for all sets. The distribution of the relative error presented in Figure 1 shows an increase of the fraction of compounds with low relative error in the SVR model. As a result, a smaller SD of 5.26% was obtained form the normal distribution fit of the relative error expressed in percentage compared with 6.12% for the MLR model. The two distributions were centered at +0.42% (MLR) and +0.94% (SVR). This corresponds to a slight overestimation of RI by both models.

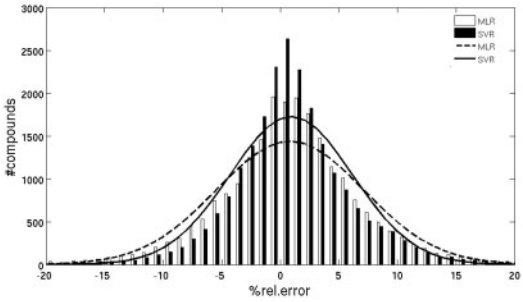

Fig. 1. Distribution of the relative prediction error in percentage (bars) and the fitted normal probability distribution (lines) of the proposed MLR and SVM models, respectively.

3.2 SVR versus NIST05 prediction model

Next, the RI predictions based on our SVR model were compared with those obtained with the NIST05 (Stein *et al*., 2007) model. The models are based on different modeling techniques and different sets of descriptors. The NIST05 model is based on 84 different groups of atoms and reflects the specific elemental composition and arrangement of the molecule. Our SVR model, on the other hand, includes about four times less descriptors with the majority of the descriptors related to properties derived from the 2D molecular structure. As for the atom specific descriptors, these are mainly accounting for the presence of a particular element. For example, in the SVR model the number of oxygen atoms is included. In the NIST05 model, eight different oxygen containing groups are needed. The NIST05 model can therefore be considered as a specific model and our SVR model as a generic model.

Three different sets were selected for the comparison. Set 1 contained all 11 229 compounds predicted by the SVR and the NIST05 models. Set 2 comprised a subset of 4138 isomers. This set was selected by using only the functional groups because isomers are characterized by the same set of functional groups. Set 3 contained isomers for which the experimental BP values were available, which corresponded to ∼50% of the total number of compounds with known experimental BP value. Compared with the NIST05 model, the SVR model gave the highest percentages of rank 1 predictions and the lowest percentages of rank 3 predictions for each of the sets (see Table 3). These rankings correspond to 'good' and 'poor' agreement between the experimental and predicted RI, respectively. The largest difference in performance of the two models was observed for the compounds in Set 3, for which the NIST05 model predicted rank 1 for 73% of the compounds compared with 88% for the SVR model. Only 1% of the compounds in this set had poor prediction (rank 3) in the SVR model. This analysis shows that out SVR model outperforms the NIST05 model.

3.3 Filtering procedure evaluation

3.3.1 Mixture of standard compounds The RI prediction model and the ranking and filtering of the NIST05 hits was tested on a mixture of 22 standards measured using fluid injection. The compounds were selected to cover a relatively large range of RIs, from 800 to 1800. Included were unsaturated cyclic hydrocarbons, alcohols, aldehydes, ketones, thiols and heterocyclic compounds. The raw data were processed using AMDIS and extracted mass spectra were searched against the NIST05 MS library. For each mass spectrum, 10 MS hits were retrieved from the library and ranked. In Figure 2, the predicted RI values of the MS hits are plotted against the experimental RI values for each component. Hits with very different RI were retrieved for some of the components whereas for other components all hits had similar RI values. For a single component there were no hits with rankings 1 or 2. This component was extracted at the end of the chromatogram and contained only a few m/z values.

The match between a spectral library and an experimental spectrum (MF), as reported by AMDIS, ranges between 0 (no match) and 100 (perfect match). The distribution of the rank over the MF is given in Figure 3a. Except for pantolactone, all standards were retrieved with an MF higher than 90 (Supplementary Material, Table S2). The chromatographic peaks of pantolactone and D-limonene were partially overlapping with pantolactone having a low abundance, which resulted in a lower quality of the extracted mass spectrum. All 22 compounds passed the RI filter with six compounds having a moderate (rank 2) agreement between experimental and predicted RI values. As shown in Figure 3a, the fraction of hits with rank 1 decreased with the decrease of the value of MF. However, even an MF value as

Table 3. Ranking of different sets predicted by the SVR and the NIST05 models

	Rank 1		Rank 2		Rank 3	
	NIST05 (%)	SVR (%)	NIST05 (%)	SVR (%)	NIST05 (%)	SVR (%)
Set 1	72	79	20	17	8	4
Set 2	76	85	18	13	6	2
Set 3	73	88	23	11	4	1

Set 1 contains 11 229 compounds, Set 2 is a subset of 4138 isomers and Set 3 is a subset of 1055 isomers with known experimental BP.

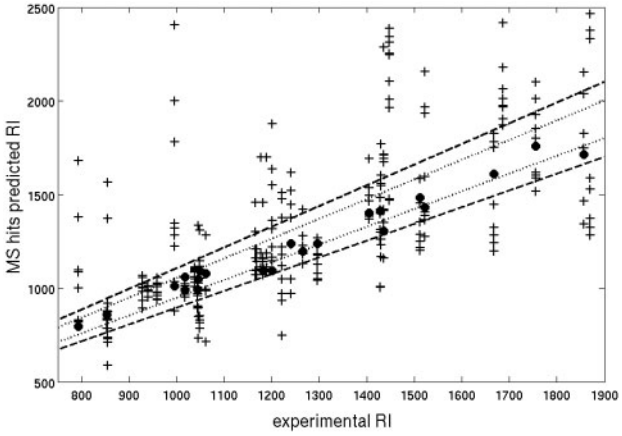

Fig. 2. Experimental RI of the extracted components and the predicted retention indices (RI) for MS hits for the mixture of standards. The MS hits are represented by + and 22 standard compounds by black circles. The dotted and the dashed lines represent the (±)1σ and (±)2σ relative difference interval, respectively.

high as 90 might return a wrong hit. An example is given by the hit list of methylparaben. The first hit was indeed methylparaben and the second hit was 3-hydroxybenzylhydrazide. In 3-hydroxybenzylhydrazide, the methoxy group (-OCH_3) is substituted by a hydrazide group (-NH-NH_2). The mass spectra of these two compounds are very similar but their RIs differ greatly. The difference between the experimental RI of methylparaben and the predicted RI of 3-hydroxybenzhydrazide was 254 RI units. Hence, 3-hydroxybenzylhydrazide was filtered out.

Despite the relatively large error of prediction of RI, the combination of RI and MF provides a powerful means to increase the confidence of the identification. With the proposed procedure of filtering out hits with absolute relative RI error larger than 10.52% ($\pm 2\sigma$), none of the 22 compounds would have been missed with only six compounds found to have a moderate agreement between the predicted and experimental RI values. The ranking of the hits can thus be used to guide the selection of compounds to be tested with authentic standards for final confirmation of the putative identity of unknown metabolites.

3.3.2 Tomato volatile compounds In a biological sample, such as tomato fruit, hundreds of volatile compounds can be present and it is not always possible to achieve good chromatographic separation for all of these. In such cases, incomplete or contaminated mass spectra will be extracted, resulting in low MF values being reported by AMDIS. Based on spectral hits only, Tikunov *et al.* (2005) putatively identified 68 compounds present in different tomato genotypes. Of these 68, 43 compounds have been confirmed by using authentic standards. The abundance of these compounds was used as a criterion for selecting an appropriate test sample for our model. In total, 237 components were retrieved for this sample by AMDIS. The distribution of the rank over the MF values of the hits is shown in Figure 3b. Compared with the standard mixture, only a small fraction of the hits had MF values higher than 90. From these, 82% passed the RI filter. The fraction of rejected hits increased to 44% for hits with MFs between 80 and 90.

The hit list was searched for the previously characterized metabolites and 50 out of 68 could be retrieved (Supplementary Material, Table S3). The remaining 18 compounds had too low abundance in the selected sample to give reliable mass spectra. In Table S3 (Supplementary Material), also the RI values found in the literature (RI_{lit}) are given. The MF values for these compounds were in a broad range, from 63 to 95. Three of the compounds did not pass the RI filter. An overestimation of RI by the model resulted in giving rank 3 to β-damascenone. However, the relative error of 10.83% was very close to the filtering threshold. The other two hits, 5-methyl-3-methylene-5-hexen-2-one and 2-phenyl-3-buten-1-ol, were obtained for low abundant components and the extracted spectra had low MF values. In order to get a better estimation of the experimental MS spectra, the original hit list as proposed by Tikunov *et al.* (2005) for these two components was evaluated. This list was obtained by manual extraction of the MS spectra using a characteristic m/z value for the deconvolution. The resulting hit list for these two components is presented in Table S4 (Supplementary Material). All hits for the component at RI_{exp} of 1016 were unsaturated ketones. The open chain keton (first hit) had a lower predicted RI value than the cyclic compounds (second and third hit). The difference in the predicted RI was also confirmed by the experimental data. Most probably, this component corresponds to an ethanone group connected to a cyclohexene or cyclopentene ring. Structurally very different compounds were retrieved for the component at RI_{exp} of 1036. The predicted RIs for 2-phenyl-3-buten-1-ol and 3,5-diphenyl-1-pentane were much higher than the experimentally observed values. Unfortunately, no literature data are available for these compounds. The other three hits, (Z)-β-methylstyrene, cyclopropylbenzene and benzocyclopentane, had too similar MF and RI values to be able to discriminate between them.

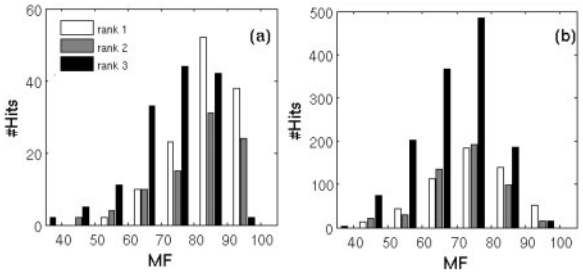

Fig. 3. Distribution of SVR rank and MF of the NIST05 library hits for the mixture of standards (**a**) and the tomato sample (**b**).

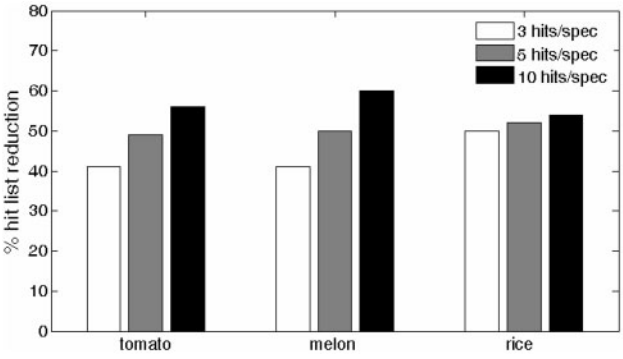

Fig. 4. Percentage hit list reduction of the tomato, melon and rice samples. The 3, 5 and 10 hit per spectrum were retrieved for each sample.

The two examples, pure standards and a crude plant sample, demonstrate that RI estimation can be applied as an automatic post-processing method for candidate compound identifications obtained by searching MS libraries. False identifications can be easily recognized and filtered out based on their predicted RI. In this way, much smaller sets of candidate compounds need to be verified by authentic standards. The proposed SVR model for RI prediction is generic and requires the calculation of a well defined set of descriptors. The model can also be used for predicting the RI of other MS libraries without RI information provided that the exact structure of the compounds is known.

3.3.3 Hit list reduction In practice, the number of hits filtered out by our procedure will depend on the chemical composition of the sample. To investigate this, in addition to the tomato fruit sample, melon and rice samples were analyzed. Sets of 3, 5 and 10 hits per experimental spectrum were processed for each sample. In general, the fraction of hits filtered out increased with increasing the number of hits to be returned. A reduction of 41% was obtained for the melon sample with three hits per spectrum and was increased to 60% with 10 hits per spectrum (Figure 4). The rice sample showed a somewhat

different pattern. More hits (50%) were filtered out using the smallest hit list whereas for the largest hit list the level of reduction (54%) was similar to that of tomato fruit and melon.

In general, low abundant compounds are difficult to detect due to poor quality of the extracted experimental spectrum. As a result, the right compound may appear further down in the hit list. Only one compound is missed when the hit list of the tomato sample is filtered. When three hits per spectrum are retrieved, (Z-) 3-hexenal, benzyl alcohol, 6-methyl-3,5-heptadien-2-one and geranyl acetone, will be missed as they are not within the first three hits.

4 CONCLUSIONS

This study has shown that RI estimation can be successfully applied to enhance compound annotation of unknown volatile compounds in GC-MS by taking into account the chromatographic behavior of the hits obtained by searching MS libraries. The RI prediction SVR model was built using a large set of structurally diverse compounds. Together with the BP, descriptors related to size, shape and the degree of branching were most predictive in the model. The model is an important improvement over previous models built for structurally diverse compounds because it covers a much larger range (360–4100) of RI values and uniquely enables the discrimination of positional isomers. Using the predicted RI, an automated procedure was developed for ranking and filtering the hits obtained after MS library searching. When the procedure was used to process a sample of tomato fruit volatiles, only one out of 50 previously identified compounds was missed. This demonstrates that it is safe to reject hits with a relative RI prediction error $>10.52\%$. The degree of hit list reduction varied from 41% to 60% and depended on the size of the original hit list. Large hit list were reduced to a greater extend compared with small hit lists. The estimated RI can also assist in the selection of compounds to be purchased or chemically synthesized to confirm the identity of unknown metabolites.

ACKNOWLEDGEMENTS

The Netherlands Bioinformatics Centre and Centre for BioSystems Genomics are initiatives under the auspices of The Netherlands Genomics Initiative (NWO/NGI). Prof. Cajo ter Braak is kindly acknowledged for the discussions on the SVR modeling. We thank Laura Pezzolesi for her assistance on the measurements of the standards and the melon sample, and the discussion on the ranking. The tomato and the rice samples were kindly made available by Dr Yury Tikunov and Dr Melissa Fitzgerald, respectively.

Funding: The Netherlands Bioinformatics Centre (to V.V.M. and R.C.H.J.vH.); EU 6th FP project EU-SOL (FOOD-CT-2006-016214 to V.V.M. and R.C.H.J.vH.); Centre for BioSystems Genomics (to R.D.H., R.C.H.dV. and H.A.V.); EU 6th FP project META-PHOR (FOOD-CT-2006-036220) (to R.D.H., R.C.H.dV. and H.A.V.).

Conflict of Interest: none declared.

REFERENCES

Adams,R.P. (2001) *Identification of Essential Oil Components by Gas Chromatography/Quadrupole Mass Spectrometry*. Allured Publishing, Carol Stream, IL.

Ausloos,P. *et al*. (1999) The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.*, **10**, 287–299.

Broadhurst,D. *et al*. (1997) Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Anal. Chim. Acta*, **348**, 71–86.

Constantinou,C. *et al*. (2008) Vitamin E and cancer: an insight into the anticancer activities of vitamin E isomers and analogs. *Int. J. Cancer*, **123**, 739–752.

Cristianini,N. and Shauwe-Taylor,J. (2000) *An Introcuction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, Cambridge.

Eckel,W.P. and Kind,T. (2003) Use of boiling point-Lee retention index correlation for rapid review of gas chromatography-mass spectrometry data. *Anal. Chim. Acta*, **494**, 235–243.

Farkas,O. *et al*. (2004) Quantitative structure-retention relationships XIV - Prediction of gas chromatographic retention indices for saturated O-, N-, and S-heterocyclic compounds. *Chemom. Intell. Lab. Syst.*, **72**, 173–184.

Fitzgerald,D.J. *et al*. (2005) Structure-function analysis of the vanillin molecule and its antifungal properties. *J. Agric. Food Chem.*, **53**, 1769–1775.

Gao,H. *et al*. (2002) Enhancement of binary QSAR analysis by a GA-based variable selection method. *J. Mol. Graphics Modell.*, **20**, 259–268.

Garkani-Nejad,Z. *et al*. (2004) Prediction of gas chromatographic retention indices of a diverse set of toxicologically relevant compounds. *J. Chromatogr. A*, **1028**, 287–295.

Heberger,K. (2007) Quantitative structure-(chromatographic) retention relationships. *J. Chromatogr. A*, **1158**, 273–305.

Hasegawa,K. *et al*. (1997) GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *J. Chem. Inf. Comput. Sci.*, **37**, 306–310.

Hemmateenejad,B. *et al*. (2007) Quantitative structure-retention relationship for the Kovats retention indices of a large set of terpenes: a combined data splitting-feature selection strategy. *Anal. Chim. Acta*, **592**, 72–81.

Hu,R.J. *et al*. (2005) QSPR prediction of GC retention indices for nitrogen-containing polycyclic aromatic compounds from heuristically computed molecular descriptors. *Talanta*, **68**, 31–39.

Jalali-Heravi,M. and Kyani,A. (2004) Use of computer-assisted methods for the modeling of the retention time of a variety of volatile organic compounds: a PCA-MLR-ANN approach. *J. Chem. Inf. Comput. Sci.*, **44**, 1328–1335.

Kashfi,K. *et al*. (2005) Positional isomerism markedly affects the growth inhibition of colon cancer cells by nitric oxide-donating aspirin in vitro and in vivo. *J. Pharmacol. Exp. Ther.*, **312**, 978–988.

Kovats,E. (1958) Gas-Chromatographische Charakterisierung Organischer Verbindungen .1. Retentionsindices Aliphatischer Halogenide, Alkohole, Aldehyde Und Ketone. *Helv. Chim. Acta*, **41**, 1915–1932.

Luan,F. *et al*. (2005) Prediction of retention time of a variety of volatile organic compounds based on the heuristic method and support vector machine. *Anal. Chim. Acta*, **537**, 101–110.

Lucasius,C.B. and Kateman,G. (1993) Understanding and using genetic algorithms .1. Concepts, properties and context. *Chemom. Intell. Lab. Syst.*, **19**, 1–33.

Preuss,T.G. *et al*. (2006) Nonylphenol isomers differ in estrogenic activity. *Environ. Sci. Technol.*, **40**, 5147–5153.

Randic,M. (2001) Novel shape descriptors for molecular graphs. *J. Chem. Inf. Comput. Sci.*, **41**, 607–613.

Rayne,S. and Ikonomou,M.G. (2003) Predicting gas chromatographic retention times for the 209 polybrominated diphenyl ether congeners. *J. Chromatogr. A*, **1016**, 235–248.

Rogers,D. and Hopfinger,A.J. (1994) Application of genetic function approximation to quantitative structure-activity-relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.*, **34**, 854–866.

Safa,F. and Hadjmohammadi,M.R. (2005) Use of topological indices of organic sulfur compounds in quantitative structure-retention relationship study. *QSAR Comb. Sci.*, **24**, 1026–1032.

Stein,S.E. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.*, **10**, 770–781.

Stein,S.E. and Brown,R.L. (1994) Estimation of normal boiling points from group contributions. *J. Chem. Inf. Comput. Sci.*, **34**, 581–587.

Stein,S.E. *et al*. (2003) Open standards for chemical information - the IUPAC chemical identifier and data dictionary projects. *Abstr. Pap. Am. Chem. Soc.*, **226**, U304–U304.

Stein,S.E. *et al*. (2007) Estimation of Kovats retention indices using group contributions. *J. Chem. Inf. Model.*, **47**, 975–980.

Tikunov,Y. *et al*. (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol.*, **139**, 1125–1137.

Todeschini,R. *et al*. (2003). DragonX 1.2. Available at http://www.talete.mi.it/products/dragon_description.htm. (last accessed date 10 February 2009).

Trinajstic,N. *et al*. (1997) The detour matrix in chemistry. *J. Chem. Inf. Comput. Sci.*, **37**, 631–638.

Umemura,T. *et al*. (1996) Isomer-specific acute toxicity and cell proliferation in livers of B6G3F1 mice exposed to dichlorobenzene. *Toxicol. Appl. Pharmacol.*, **137**, 268–274.

Vapnik,V. (1995) *The Nature of Statistical Learninr Theory*. Springer-Verlag, New York.