BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 18 2012, pages 2333–2341
doi:10.1093/bioinformatics/bts437

System biology

Advance Access publication July 18, 2012

Metabolite identification and molecular fingerprint prediction through machine learning

Markus Heinonen[1,2,*], Huibin Shen[1], Nicola Zamboni[3] and Juho Rousu[3,4]

[1]Department of Computer Science, University of Helsinki, Helsinki, 00014, Finland, [2]Helsinki Institute for Information Technology, Finland, [3]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich 8093, Switzerland and [4]Department of Information and Computer Science, Aalto University, Espoo 00076, Finland

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: Metabolite identification from tandem mass spectra is an important problem in metabolomics, underpinning subsequent metabolic modelling and network analysis. Yet, currently this task requires matching the observed spectrum against a database of reference spectra originating from similar equipment and closely matching operating parameters, a condition that is rarely satisfied in public repositories. Furthermore, the computational support for identification of molecules not present in reference databases is lacking. Recent efforts in assembling large public mass spectral databases such as MassBank have opened the door for the development of a new genre of metabolite identification methods.

Results: We introduce a novel framework for prediction of molecular characteristics and identification of metabolites from tandem mass spectra using machine learning with the support vector machine. Our approach is to first predict a large set of molecular properties of the unknown metabolite from salient tandem mass spectral signals, and in the second step to use the predicted properties for matching against large molecule databases, such as PubChem. We demonstrate that several molecular properties can be predicted to high accuracy and that they are useful in *de novo* metabolite identification, where the reference database does not contain any spectra of the same molecule.

Availability: An Matlab/Python package of the FingerID tool is freely available on the web at http://www.sourceforge.net/p/fingerid.

Contact: markus.heinonen@cs.helsinki.fi

Received on June 4, 2012; revised on July 3, 2012; accepted on July 4, 2012

1 INTRODUCTION

In metabolomics, mass spectrometry (MS) provides the key measurement technology for quantifying and qualifying chemical signals to provide biological knowledge of cellular processes (Kell, 2004). An MS measurement of a biological sample results in a set of peaks representing the mass-to-charge ratios and intensities of the different compounds of the sample. Identification of these molecules through mass spectra is a prerequisite for further biological interpretation and is the most time-consuming and laborious step in metabolomics experiments (Werner *et al.*, 2008). The identification process is still mostly not automatized, thus requiring extensive manual analysis and expert knowledge (Neumann and Böcker, 2010).

Some structural information on a given compound can be obtained by tandem mass spectrometry (MS/MS) through ionization or collision-induced dissociation fragmentation experiments. Tandem mass spectrum contains peaks representing various fragmentation products of the unknown compound, including the difficulty to predict rearrangement reaction products. A compound fragments in specific patterns according to its structure, the collision energy and the experimental configuration (McLafferty, 1983). Elucidation of tandem spectra is at the core of molecular identification (Neumann and Böcker, 2010; Heinonen *et al.*, 2008). The standard method of performing compound identification is to measure the tandem mass spectrum of the unknown compound and to query the spectrum against annotated reference libraries of standardized spectra (NIST, MassBank, Wiley Registry) (Werner *et al.*, 2008), followed by extensive domain expert analysis. In practice, the reference database requires a closely matching type of equipment and operational parameters for reliable matching (Horai *et al.*, 2010).

Machine learning approaches for metabolite identification from MS/MS data have not been widely studied. Early related work includes STIRS (Dayringer *et al.*, 1976) that uses a nearest neighbor approach to model the statistical relationships between spectral features and molecular substructures, the neural network work of Curry and Rumelhart (1990), and the decision tree approach by Breiman *et al.* (1984). However, the impact of these methods has remained dormant perhaps due to predating the era of systems biology and limited available data. In contrast, current state-of-the-art methods are based on combinatorial algorithms and database searches. The MetFrag software identifies metabolites by matching candidate metabolites with closest combinatorially simulated spectrum to the observed one (Wolf *et al.*, 2010). Analysis of isotopic patterns can give additional clues on the metabolite's elemental composition (Böcker *et al.*, 2009).

We introduce a novel two-step pattern-recognition approach (see Fig. 1) to the metabolite identification problem. Instead of directly learning a mapping between the spectrum and the metabolite, we first predict a set of characterizing fingerprints of the metabolite from its tandem mass spectrum using a kernel-based approach. We learn our fingerprint prediction model from a large set of tandem mass spectra obtained from public mass spectral database MassBank (Horai *et al.*, 2010). In the next

*To whom correspondence should be addressed.

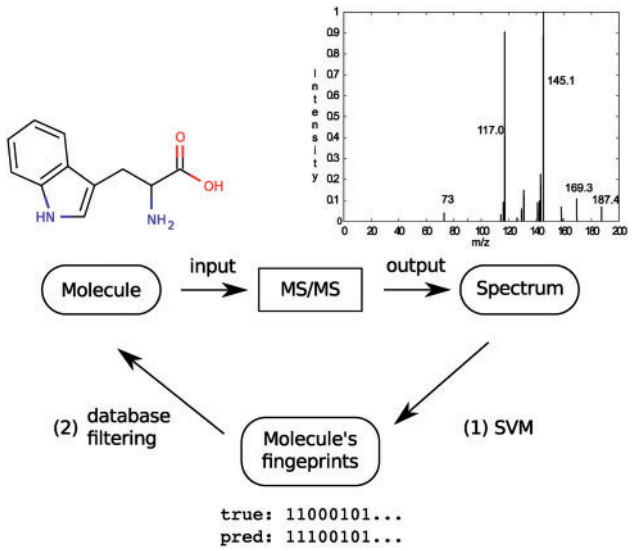

Fig. 1. The overview of the two-step metabolite identification framework. An example molecule Tryptophan (mass 204.2 Da) produces a characterizing MS/MS spectra, which is used to predict the original molecule through fingerprints. The predicted fingerprints, along with neutral mass measurement, are used to filter a molecular repository for candidates

step, we match the predicted fingerprints against a large molecular database to obtain a list of candidate metabolites. The metabolite identification model generalizes to metabolites not present in reference spectral databases. Due to the machine learning approach, data from any type of mass spectrometer are supported.

In Section 2, we review the basics of mass spectrometry and the current state-of-the-art of metabolite identification through reference databases. We introduce three classes of mass spectral features and a probability product kernel over the spectral features in Section 3. We also discuss the use of multiple measurements from the same metabolite. We propose a statistical approach to retrieve candidate metabolites using the predicted properties in Section 4. In Section 5, we experiment with the proposed FingerID method in fingerprint prediction and metabolite identification. We conclude this article with discussion in Section 6.

2 METABOLITE IDENTIFICATION THROUGH TANDEM MASS SPECTROMETRY

A tandem mass spectrum is generated by selecting an unknown ion band and its mass-to-charge ratio to undergo fragmentation (see Fig. 1). Under mild fragmentation conditions, the peak corresponding to the molecule ion is still visible in the tandem spectrum at the same mass-to-charge. During fragmentation, an ion is often cleaved into two fragments, one of which retains the charge (Small molecules appear often as singly charged ions. We assume single-charged ions throughout the paper for clarity.) and is visible in the tandem mass spectrum. The complementary fragment is a neutral loss invisible in the spectrum.

The first step in compound identification is to constraint the mass and elemental composition of the compound through peak masses. Computational methods utilize the compound's peak and its isotope peak masses to compute set of possible elemental compositions (Böcker *et al.*, 2009). However, mass measurement accuracy defines the set of compatible compositions through the scope of error in the mass measurements. High- or ultra-high-resolution analyzers, such as Time-of-flight and fourier transform analyzers, can achieve low mass errors in the range of 1–5 ppm (Werner *et al.*, 2008).

We distinguish four increasingly challenging cases for annotating an unknown spectrum measured at specific device through querying mass spectral databases:

(1) The database contains a spectrum of the molecule, measured with the same device and experimental parameters
(2) The database contains a spectrum of the molecule, measured on the same device but with different experimental parameters (e.g. collision energy)
(3) The database contains a spectrum of the molecule, measured on a different device
(4) The database does not contain the spectrum. Annotation requires *de novo* metabolite identification

In the first Case (a), we expect a simple retrieval of the most similar spectrum from the reference database to give reliable identifications, with no need for more complicated prediction schemes. Cases (b) and (c) are expected to be less suited to simple retrieval and leave room for improvement through machine learning approaches. Finally, Case (d) is not possible to tackle with retrieval from a reference database and can be seen as the prime motivation for machine learning method development.

The standard query methods rely on a distance function to compute the best hits of the unknown spectra against a library. The most common approach is to count the number or proportion of shared peaks. Several methods include peak intensities with experimentally defined weights (Horai *et al.*, 2010; Wolf *et al.*, 2010; Dworzanski *et al.*, 2004) or probabilistic measures (Pavlic *et al.*, 2006; Oberacher *et al.*, 2009).

In our experiments we use retrieval from the MassBank database as the reference database method, giving match scores for query spectra. (MassBank also has an advanced metabolite identification service, which however is not available in batch mode.) The score is the Pearson correlation between weighted peak vectors between query $\mathbf{w}^{(q)}$ and target $\mathbf{w}^{(t)}$ spectra with elements

$$w_i = int_i^{\alpha} \cdot mass_i^{\beta},$$

where $\alpha = 0.5$ and $\beta = 2$ (Horai *et al.*, 2010; Stein, 1994). Alignment of the peaks is ensured by allowing some mismatch, default value at 0.3.

3 KERNELS FOR MASS SPECTRA

In this section, we build three classes of features extracted from mass spectra that are relevant to the fingerprint prediction task. The features are used in two families of mass spectral kernels for SVM classification: an integral mass accuracy kernel and a high mass accuracy kernel, where the peaks generate a gaussian mixture model densities. Furthermore, we introduce methods to utilize multiple collision energy spectra through kernel fusion.

We consider mass spectra of molecule as a collection $\chi = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \in \mathcal{X}$ of k two-dimensional peak tuples $\mathbf{x}_i \in \mathbb{R}^2$ (see Fig. 1). (The number of peaks k varies from spectrum to spectrum in the dataset.) A peak $\mathbf{x} = (mass, int)^T$ represents the mass-to-charge value and the intensity of the peak measurement. We normalize all intensities to range [0,1]. Often we have a series of spectra of a molecule measured with increasing collision energies as $(\chi_{10eV}, \ldots, \chi_{50eV})$.

A mass spectral kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semi-definite similarity function between sets. The kernel K induces a mapping $\phi : \mathcal{X} \to \mathcal{H}$ of input to a Hilbert space, such that $K(\chi,\chi') = \langle \phi(\chi), \phi(\chi') \rangle$ where $\chi, \chi' \in \mathcal{X}$. The power of kernels comes from the fact that a simple similarity value can implicitly construct a rich feature representation of non-vectorial objects.

We learn a function $f : \mathcal{X} \to \{-1, 1\}^m$ mapping the tandem mass spectrum χ of an unknown metabolite into the m-dimensional fingerprint target vector $\mathbf{y} = (y_i)_{i=1}^m \in \{-1,1\}^m$ using a dataset of n spectra $\{\chi_1, \ldots, \chi_n\}$. We opt to use n binary SVM's each learning the subtask $f_i : \mathcal{X} \to \{-1, 1\}$. The learning function is $f_i(\chi) = \mathbf{sign}(\mathbf{w}_i^T \phi(\chi)) = \mathbf{sign}(\sum_j \alpha_j K(\chi_j, \chi))$, where $\mathbf{w}_i \in \mathcal{H}$.

The main types of fingerprints are topological, physico-chemical and electrical properties of the molecule (Steffen *et al.*, 2009). In Section 4, we utilize the fingerprints $\mathbf{y}$ to filter candidates from molecular databases.

3.1 Integral mass kernel

The integral kernel consists of a feature mapping $\phi : \mathcal{X} \to \mathbb{R}^d$. The feature representation gathers the peak intensities at discrete intervals, i.e. $\phi_{\text{peaks}}(\chi) = (\phi_{\text{peaks}}(\chi)_i)_{i=1}^d$ with

$$\phi_{\text{peaks}}(\chi)_i = \sum_{(\text{mass},\text{int}) \in \chi} \delta_{i \pm 0.5}(\text{mass}) \cdot \text{int},$$

where $\delta_{i \pm 0.5}(\text{mass})$ is an indicator function giving 1 if $i - 0.5 < \text{mass} < i + 0.5$, and 0 otherwise. The 'peaks' features gather all intensity of peaks that are rounded to its nearest integral mass. The dimensionality d of the feature space is equivalent to the largest feature index with non-zero value, i.e. the mass of the largest peak in the dataset.

The peak ions are complemented by undetected neutral loss peaks which indicate the masses of cleaved fragments from the precursor (mother) ion. The 'neutral loss' features $\phi_{\text{nloss}}(\chi) = (\phi_{\text{nloss}}(\chi)_i)_{i=1}^d$ are

$$\phi_{\text{nloss}}(\chi)_i = \sum_{(\text{mass},\text{int}) \in \chi} \delta_{i \pm 0.5}(\text{prec}(\chi) - \text{mass}) \cdot \text{int},$$

where we add the intensities of all peaks whose mass difference to the precursor prec(χ) rounds to i. The neutral losses are invariant of the mother ion mass and allow cleavage detection of e.g. phosphate group PO_4^- which has a standard atomic weight of 94.97.

In tandem mass spectrometry, the fragment ions themselves can undergo further fragmentation. We can generalize the notion of neutral losses to include secondary fragmentation reactions between non-mother ion peaks by multiplying the intensities of any two pairs of peaks with a fixed integral mass difference. The peak 'difference' feature representation $\phi_{\text{diff}}(\chi) = (\phi_{\text{diff}}(\chi)_i)_{i=1}^d$ consists of

$$\phi_{\text{diff}}(\chi)_i = \sum_{\substack{(\text{mass},\text{int}) \in \chi \\ (\text{mass}',\text{int}') \in \chi}} \delta_{i \pm 0.5}(|\text{mass} - \text{mass}'|) \cdot \text{int} \cdot \text{int}'.$$

The kernels utilizing these three classes of features are

$$K_{\text{peaks}}(\chi, \chi') = \langle \phi_{\text{peaks}}(\chi), \phi_{\text{peaks}}(\chi') \rangle$$
$$K_{\text{nloss}}(\chi, \chi') = \langle \phi_{\text{nloss}}(\chi), \phi_{\text{nloss}}(\chi') \rangle$$
$$K_{\text{diff}}(\chi, \chi') = \langle \phi_{\text{diff}}(\chi), \phi_{\text{diff}}(\chi') \rangle.$$

The full integral kernel $K_{\text{full}} = K_{\text{peaks}} + K_{\text{nloss}} + K_{\text{diff}}$ corresponds to a concatenation of the feature sets $\phi_{\text{full}} = (\phi_{\text{peaks}}; \phi_{\text{nloss}}; \phi_{\text{diff}})$.

Polynomial combinations of the features are easily computed by applying a polynomial kernel. We experiment with linear and quadratic kernels in Section 5. The integral kernel is directly positive semi-definite (Shawe-Taylor and Christianini, 2004).

3.2 High-resolution mass kernel

The explicit feature representation of the discrete kernel allows direct inspection of the feature weights. However, it rounds all peaks to a nearest integer value. The unique peaks with same rounded mass are erroneously binned together. We can expand the feature space by narrowing the bin widths; however, this introduces an alignment problem in the dot product.

Instead, we generalize the feature mapping with a kernel between distributions by applying the probability product kernel of Kondor and Jebara (Kondor and Jebara, 2003). We fit probabilistic models $p(\chi)$ and $p'(\chi')$ over spectra χ and χ' and define the kernel between spectra as a kernel between the corresponding probability distributions, $K(\chi,\chi') \equiv K(p,p')$. We use a probability product kernel

$$K(p, p') = \int_{\mathbb{R}^2} p(\mathbf{x}) p'(\mathbf{x}) d\mathbf{x},$$

which has an interpretation as an expectation of one distribution under the other

$$\int_{\mathbb{R}^2} p(\mathbf{x}) p'(\mathbf{x}) d\mathbf{x} = \mathbb{E}_p[p'(\mathbf{x})] = \mathbb{E}_{p'}[p(\mathbf{x})],$$

called the expected likelihood kernel as in Jebara *et al.* (2004).

For countable and finite quantities, the probability product corresponds to a dot product $\langle p,p' \rangle$ in ℓ_2 with a feature representation $\phi : \chi \to p(\chi)$ that collects the probabilities into a vector. Note that in general $K(\chi,\chi) \neq 1$ as the probability product is not a proper distribution. We normalize all kernels to [0, 1].

We opt to model the probabilities $p(\chi) = \frac{1}{k} \sum_{i=1}^k p_i(\mathbf{x})$ as gaussian mixture models estimated using maximum likelihood density estimation separately for the three feature classes. Each feature (e.g. peaks) $\mathbf{x}_i = (\text{mass}_i, \text{int}_i)$ corresponds to a non-isotropic two-variate normal distribution $p_i(\mathbf{x}_i) \sim (\mathcal{N})(\mathbf{x}_\rangle, \pm)$ over the mass and intensity of peak i. We assume zero covariance between mass and intensity by setting $\Sigma = [\sigma_{\text{mass}}, 0; 0, \sigma_{\text{int}}]$.

The probability product kernel can be computed in closed form as

$$K(\chi, \chi') = K(p, p') = \int_{\mathbb{R}^2} p(\mathbf{x})p'(\mathbf{x})d\mathbf{x} = \frac{1}{k}\frac{1}{k'}\sum_{i,j}^{k,k'} \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}|\Sigma'|^{\frac{1}{2}}|\Sigma_\dagger|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}\left(\mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i + \mathbf{x}_{j'}{}^T\Sigma'^{-1}\mathbf{x}_{j'} + \mathbf{x}_\dagger^T\Sigma_\dagger^{-1}\mathbf{x}_\dagger\right)\right),$$

where $\Sigma_\dagger = \Sigma^{-1} + \Sigma'^{-1}$, $\mathbf{x}_\dagger = \Sigma^{-1}\mathbf{x}_i + \Sigma'^{-1}\mathbf{x}'_j$, and $\mathbf{x}_i \in \chi$ and $\mathbf{x}'_j \in \chi'$ (cf. (Kondor and Jebara, 2003)).

We compute the feature class specific kernels K^ϕ_{peaks}, K^ϕ_{nloss} and K^ϕ_{diff} by iterating over peaks $\{\mathbf{x}_i\}$, neutral losses $\{[\text{prec}(\chi) - \text{mass}_i; \text{int}_i]^T\}$ or peak differences $\{\mathbf{x}_i - \mathbf{x}_j; i<j\}$, respectively.

The final kernel is $K^\phi_{\text{full}} = K^\phi_{\text{peaks}} + K^\phi_{\text{nloss}} + K^\phi_{\text{diff}}$. Polynomials of the kernel measure the tuple products of features analogous to the integral case. The kernel is a valid positive semi-definite function (Jebara *et al.*, 2004).

3.3 Multiple collision energies

Databases such as MassBank contain a series of spectra measured using various collision energies (CE). Complementary structural information can be attained from spectra of different collision energies. A higher collision energy places more energy into the fragmentation process, usually producing more small secondary and tertiary fragments. Gradually ramping up the collision energy might reveal the gradients of the peaks, which are related to the energy landscape of the fragments. In practice, spectra are measured with fixed collision energies, often from 10 to 50 eV. A simple strategy to facilitate learning from various collision energies is to denote all spectra as independent data points and use majority voting over all collision energies.

3.3.1 Sum kernels Spectra at different collision energies provide complementary information on the molecule. Thus, we can model the process by having CE-specific variables. Let ϕ_e be the feature space of measurements at CE-e. Then, the full feature space is a concatenation of CE, specific spaces; $\phi = (\phi_{e_1}; \ldots; \phi_{e_n})$, where e_i is the ith collision energy. The kernel is computed as the sum of CE-specific kernels, $K = K_{e_1} + \cdots + K_{e_n}$. In this model, we take all similarity information into account with equal weights. The method supports natural data where molecules have partial spectra series with some measurements missing.

3.3.2 Merged spectra It is common practice to merge the differing collision energy spectra together by summing the intensities into a single spectrum. This corresponds to summing the feature values across spectra, i.e. the feature space is $\phi = \sum_{e_i} \phi_{e_i}$. This can be computed with kernel matrices by $K = \sum_{e,e'} K_{e,e'}$, where $K_{e,e'}(i,j)$ is kernel value of $\phi_e(i)$ and $\phi_{e'}(j)$. This formulation corresponds to a sum kernel model with additional similarities between features across different CE added.

4 METABOLITE IDENTIFICATION THROUGH FINGERPRINTS

The predicted fingerprints can be directly used to characterize the class and properties of the measured metabolite. However, our primary interest is to support metabolite identification, which calls for converting a set of predicted molecular fingerprints into a score function that can be used to retrieve candidate molecules from large molecular databases such as PubChem (Wang *et al.*, 2009).

As different fingerprints can be predicted with varying accuracy from the mass spectra, the reliability of the fingerprints in identification of the molecule varies correspondingly. We develop a probabilistic model to tackle this uncertainty, in such a manner that the more reliably predicted fingerprints receive more weight when matching to a large set of candidate molecules.

Let $\mathbf{p} = (p_i)_{i=1}^m \in [0.5,1]^m$ denote the vector of prediction accuracies (reliability of fingerprint prediction) of the m fingerprints $\mathbf{y} = (y_i)_{i=1}^m$, obtained, e.g. from a cross-validation experiment. Taking the individual fingerprints as independent, we obtain a probability for a fingerprint vector $\mathbf{y}$ to be the true fingerprint vector given a predicted vector $\mathbf{y}'$ from the Poisson-binomial distribution

$$P(\mathbf{y}|\mathbf{p}; \mathbf{y}') = \prod_{i=1}^m p_i^{[\![y_i = y'_i]\!]}(1 - p_i)^{[\![y_i \neq y'_i]\!]}. \quad (1)$$

Now, given a fingerprint vector $\mathbf{y}'$ predicted from a mass spectrum, and a set $\mathcal{M}$ of candidate molecules M with the true fingerprint vectors $\mathbf{y}(M)$ pre-computed, we can use Equation (1) to score the candidate molecules by

$$\text{score}(M) = P(\mathbf{y}(M) \mid \mathbf{p}, \mathbf{y}').$$

A ranking is established by checking the number of molecules M' in the dataset $\mathcal{M}$ that have as high score or higher than the given molecule M:

$$\text{rank}(M) = |\{M' \in \mathcal{M} \mid \text{score}(M') \geq \text{score}(M)\}|.$$

Thus, all molecules with the same score receive the same rank.

We, note that scoring and ranking a set of candidate molecules is efficient. The time complexity of evaluating the probabilities of database's fingerprints given a prediction is $\mathcal{O}(Nm)$ operations, where N is the size of the database and m the number of fingerprints.

In our experiments, the success of metabolite identification query is simply determined by rank (M^*) of the true molecule M^* among the candidates.

In our experiments, we use in addition to MassBank, the Kegg database (Kanehisa *et al.*, 2006), which represents over 14 000 common organic molecules, and the PubChem database, which is the largest open repository of known molecular universe with over 30 million unique structures. Utilizing special purpose databases such as Kegg act as prior knowledge that our target molecules are most likely metabolites.

5 EXPERIMENTS

We experiment with our method, called FingerID, in predicting fingerprints and identifying metabolites with three datasets: *QqQ, Ltq* and *Lipids*. The 'QqQ' dataset is of nominal mass accuracy and contains positive-mode Quadrupole measurements of 514 metabolites. The metabolites are also measured with five different collision energies from 10 to 50 eV, however 11.5 % of the measurements are missing. The 'Ltq' dataset is an ultra-high accuracy positive-mode Orbitrap dataset of 293 miscellaneous

metabolites. Finally, 'Lipids' is an ultra-high-accuracy negative-mode Orbitrap dataset of 403 internally homogeneous phophatidylethanolamines. All data are obtained from the MassBank database (Horai *et al.*, 2010). The average mass offsets and standard deviations are computed using MassBank annotations and are listed in Table 1.

We initially examined a set of 528 unique structural fingerprints from OpenBabel (FP3, FP4 and MACCS). However, many of the fingerprints appear in either all or none of the molecules in the dataset and are excluded from the set of potentially useful fingerprints (Table 1). For each dataset, we retain only non-uniform fingerprints.

5.1 Experimental settings

We used the libSVM implementation with 5-fold cross-validation over the fingerprints in each dataset. We choose the optimal C parameter from $\{10^0,10^1,10^2,10^3,10^4\}$; however, in general it had a small effect. Each fingerprint was predicted independently as a binary classification task. The baseline classifier (default) always blindly votes for the most common fingerprint assignment according to the dataset.

Note that due to the cross-validation the method is assessed with a test set that contains only spectra not seen in learning the model. This is analogous to the Case (d), where the metabolites are novel and not previously annotated.

We experimented with various values for the width parameter σ_{mass} of the gaussian kernel (figure not shown). The value of two times the empirical standard deviation yielded consistently high results. We set the σ_{int} such that the similarity in density between maximal (1.0) and minimal (0.0) intensity peak is half of the mode, resulting in $\sigma_{\text{int}} = 0.849$.

5.2 Fingerprint prediction performance

The aggregate mean results of the fingerprint prediction experiments are summarized in Figure 2. The $F_1 = 2PR/(P+R)$, where R is the recall and P is the precision, measures the balance of the model in predicting both positive and negative and improvement over the default classifier. The high-resolution kernel on all features comes out as the best on average. However, 'peaks' and 'neutral loss' kernel is almost as good. Quadratic kernel helps prediction on average only in the case of integral kernel on 'peaks' or 'peaks+neutral loss' features (open markers).

The dataset-specific results are summarized in Tables 2 and 3 denoting the average prediction accuracy and F_1, respectively, over the kernel and dataset. In general, the high-resolution mass kernels (lower part of the table) achieve better results than the discrete kernels (upper part). The quadratic high-resolution kernel over all three feature types achieves best results in 5 cases out of 9.

The difference in three feature classes is consistent in all datasets. In the 'Ltq' dataset, the 'peaks' features give 86.7%, 'neutral losses' 88.8% and 'peak differences' 83.9% accuracies. Combining 'peaks' and 'neutral losses' improves the result to 91.1% while combining 'peak differences' does not improve the results. In general, the 'peak difference' features are useful only in the 'QqQ' dataset with a single CE dataset.

In the 'QqQ' dataset, metabolites are measured with five collision energies. The 40 eV CE is alone the most informative data source; however, utilizing all spectra of different collision energies increases the results notably. In the high-resolution kernel, best results are achieved by merging the spectra (91.1%), while in the discrete case summing the kernels directly gives best results (90.7%).

Individual fingerprint prediction performance is depicted in Figure 3. The figure shows the predictive accuracy for the 150 least accurately predicted fingerprints using the worst single CE as the sorting criterion. The merged spectrum always surpasses the predicting made by individual CE's.

The 'Lipids' dataset achieves extremely high-prediction accuracies of over 97% with almost all kernels. This is in contrast to the 'QqQ' dataset, where utilizing any individual CE measurement gives only small improvements over the baseline predictor. This is partly explained by the integral measurement accuracy of the 'QqQ' dataset, while the other two datasets come from ultra-high-resolution analyzers.

Table 1. The dataset statistics

Spectral dataset	Device	Size	Mode	Peak offset mean	Peak error SD	Effective fingerprints
QqQ	misc	514	Pos			286
	API3000	445	Pos	0.128	0.164	
	QuattroPremier XE	49	Pos	−0.092	0.073	
	TSQ 7000	14	Pos	−0.124	0.036	
	TSQ Quantum AM	3	Pos			
	Q-TRAP	3	Pos			
Ltq	LTQ Orbitrap XL	293	Pos	0.0	0.049	128
Lipids	LTQ Orbitrap	403	Neg	−0.135	0.090	20

Only a subset of fingerprints are exhibited in each dataset's molecules.

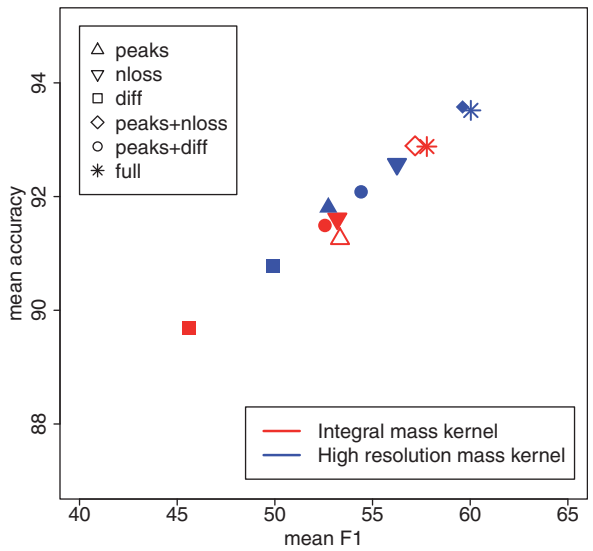

Fig. 2. Scatter plot of the aggregate average accuracy/F_1 across the three datasets with different kernel features. The open markers represent higher accuracy/F_1 ratio in a quadratic kernel

Table 2. The classification accuracies (%). The kernel with best accuracy is highlighted with bold in each dataset column

Kernel	QqQ							Ltq	Lipids
	Single spectra (CE eV)					Multiple spectra			
	10	20	30	40	50	$\sum_e K_e$	merge		
K_p, linear	87.8	88.2	88.8	89.3	89.5	89.5	89.2	85.5	98.4
quadr.	87.9	88.3	88.8	89.4	89.6	89.9	89.8	84.4	98.1
K_{nl}	88.4	88.8	88.8	88.7	89.2	89.4	89.0	86.3	98.8
	88.4	88.9	88.8	88.9	89.2	89.6	89.3	86.1	98.7
K_{df}	87.8	88.0	87.7	87.8	88.2	88.0	87.9	82.6	97.1
	87.8	88.0	87.8	87.9	88.3	87.9	87.9	82.9	96.9
K_{p+nl}	88.5	89.5	89.9	90.1	90.3	**90.7**	90.3	88.3	**99.5**
	88.4	89.4	90.0	90.0	90.3	90.5	90.6	88.1	99.3
K_{p+df}	88.2	88.6	89.0	89.4	89.6	89.4	89.2	85.6	98.7
	88.1	88.7	89.2	89.6	89.8	89.3	89.7	84.8	98.4
$K_{p+nl+df}$	88.5	89.5	90.1	90.1	90.3	90.5	90.3	88.3	99.5
	88.6	89.8	90.3	90.3	**90.5**	90.3	90.7	87.6	99.3
K_p^{φ}	88.0	88.6	89.1	89.1	89.4	89.3	89.4	86.7	98.6
	88.2	89.1	89.5	89.7	89.9	89.3	90.0	85.5	97.3
K_{nl}^{φ}	88.8	89.5	89.3	89.2	89.2	89.8	89.6	88.8	99.1
	89.0	89.8	89.7	89.5	89.6	90.0	90.0	88.1	98.0
K_{df}^{φ}	88.5	88.9	88.6	88.4	88.4	89.2	89.3	83.7	97.8
	88.6	89.0	88.9	88.6	88.6	89.2	89.5	83.9	97.1
K_{p+nl}^{φ}	89.0	89.9	90.1	90.1	90.2	90.5	90.5	**91.1**	99.3
	89.2	**90.1**	90.3	90.3	90.4	90.1	90.8	89.6	97.9
K_{p+df}^{φ}	88.8	89.4	89.5	89.5	89.5	90.0	90.0	86.5	98.8
	88.9	89.5	89.7	89.8	89.8	89.8	90.4	84.9	97.5
$K_{p+nl+df}^{\varphi}$	89.1	90.0	90.3	90.2	90.2	90.6	90.7	90.5	99.3
	89.2	**90.1**	**90.4**	**90.5**	90.4	90.2	**91.1**	88.6	98.0
default	87.3	87.2	87.2	87.2	87.7		87.3	78.7	88.3

Abbreviations: *p* is peaks, *nl* is neutral loss and *df* is difference kernel.

5.3 Direct spectral matching

We characterize the reference database retrieval performance in our three datasets (Table 4). In the reference retrieval, we assume the correct molecule is contained in the reference database (Cases a–c).

On average (the upper portion of the Table 4), the 'QqQ' and 'Lipids' datasets achieve almost 100% retrieval rates. We define the retrieval rate as the proportion of metabolites that are identified correctly within the first 10 candidates. This is due to the existence of multiple measurements of the same metabolite in MassBank: in 'QqQ' almost all metabolites have measurements at five collision energies (Case b), and in 'Lipids' almost all metabolites have a duplicate measurement (case a). By limiting ourselves to spectra measured at the same device but different CE, the retrieval rate becomes 94% in 'QqQ'.

In 'Ltq' utilizing the duplicate measurements does not increase the retrieval rates. However, when spectral matches are restricted to measurements made with different devices, the retrieval drops from 63.2 to 28.2%. This is an expected drop, as the mass spectrometer defines the fragmentation process and thus has a major impact on the spectral signals.

5.4 Metabolite identification through fingerprints

Herein, we examine the performance of our method in *de novo* identification (Case d), where the test set contains only spectra of molecules not seen in the training set. This is expected to be a challenging task as the machine learning method needs to generalize from the spectral signals. Notably, the direct spectral matching method cannot work in this case.

Table 5 and Figure 4 indicate the *de novo* metabolite identification performance utilizing the predicted fingerprints in the three datasets. We first search for candidate metabolites matching the measured neutral mass using a mass range of ± 0.5 ('Avg. hits w/ ± 0.5 mass' in Table 5) from either Kegg or PubChem, followed by fingerprint prediction and metabolite identification.

When querying molecular identification from Kegg, the average ranks of the correct metabolites are 5.0 and 3.2 for 'QqQ' and 'Ltq', respectively. The retriaval rates are 85 and 91.8%, respectively. None of the molecules in the 'Lipids' dataset was found from Kegg. Against PubChem several tens of thousands of candidates match the mass range on average. The top 10 rank retrieval rates for PubChem are 29.3, 50.8 and 54.4%, for the three datasets, even though the average ranks are relatively high.

The P_{50} value is the normalized median rank, i.e. it indicates the maximum rank, normalized by the database size, which 50% of the query molecules attain. For instance, in 'Ltq' dataset against PubChem, the P_{50} is 3.2×10^{-4}, which indicates that the better half of the predictions are able to exclude 99.96% of the neutral mass matching molecules in PubChem.

Figure 4 indicates the cumulative portions of the datasets at a given maximum rank. The P_{50} values correspond to the *x*-axis

Table 3. The F_1 (%). The kernel with best F_1 is highlighted with bold in each dataset column

Kernel	QqQ							Ltq	Lipids
	Single spectra (CE eV)					Multiple spectra			
	10	20	30	40	50	$\sum_e K_e$	merge		
K_p, linear	12.4	15.0	20.2	22.9	21.4	23.1	22.0	56.9	89.1
quadr.	13.2	16.4	21.1	23.2	22.2	23.2	25.4	52.6	89.5
K_{nl}	16.4	19.5	18.5	17.3	16.6	21.4	19.7	60.1	90.9
	17.9	20.1	19.7	18.1	17.2	21.8	20.8	58.6	90.4
K_{df}	14.5	15.5	14.2	12.9	12.6	15.1	15.3	47.4	82.9
	13.2	15.7	15.4	13.2	13.2	13.7	15.8	47.1	81.6
K_{p+nl}	15.9	20.3	23.9	22.0	22.7	**27.0**	25.0	64.6	**92.7**
	16.4	20.4	23.0	21.0	22.0	21.7	25.5	63.9	**92.7**
K_{p+df}	15.0	17.0	21.5	21.6	20.4	22.9	22.2	54.2	90.1
	15.0	19.0	22.2	23.9	**24.0**	20.0	25.9	51.0	87.9
$K_{p+nl+df}$	16.7	20.7	25.8	22.9	23.1	**27.0**	26.4	64.8	**92.7**
	17.3	22.2	24.8	23.1	23.3	21.1	26.2	60.7	92.1
K_p^{φ}	13.8	18.0	21.4	19.8	19.9	21.3	22.4	58.0	87.6
	15.6	18.8	21.0	22.7	22.2	18.4	25.1	54.9	82.6
K_{nl}^{φ}	18.3	21.4	22.4	18.8	17.1	23.9	22.8	65.9	91.9
	18.4	20.7	22.5	19.0	18.1	21.6	24.1	61.0	86.1
K_{df}^{φ}	15.9	19.9	20.7	15.6	14.6	20.8	22.8	49.0	85.1
	15.8	20.0	20.7	17.7	15.7	19.4	22.8	48.3	83.7
K_{p+nl}^{φ}	18.9	**23.9**	25.8	24.1	22.2	26.8	27.4	**71.7**	91.9
	19.6	21.0	22.3	21.7	20.9	19.9	25.6	65.1	85.3
K_{p+df}^{φ}	17.4	21.6	23.6	21.8	20.8	24.0	25.7	56.8	89.3
	17.2	21.8	23.3	23.9	22.5	19.6	27.2	51.8	83.3
$K_{p+nl+df}^{\varphi}$	19.1	23.8	**26.7**	**25.0**	22.8	27.0	**29.1**	70.6	91.8
	19.5	22.8	24.8	24.1	22.6	20.7	28.5	62.5	85.2

Abbreviations: *p* is peaks, *nl* is neutral loss and *df* is difference kernel.

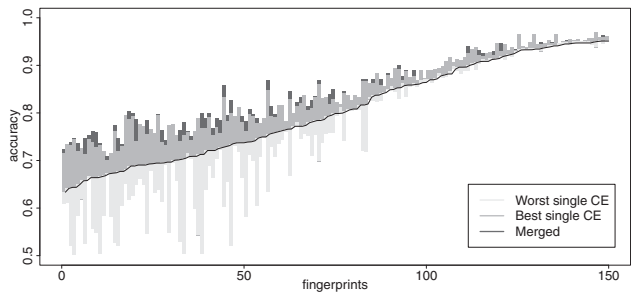

Fig. 3. The fingerprint specific accuracies of the 'QqQ' dataset with the high resolution quadratic full kernel. The bars indicate the accuracy of the fingerprints when using least informative and most informative collision energy spectra, and the merged spectra. Only the 150 least accurate fingerprints are shown. The default classifier is indicated by the bottom of the bars

values at *y*-axis value of 0.5. Due to the large size of PubChem, the ranks are in general several orders of magnitude larger. However, most of the molecules are still within the absolute rank of 1000.

We experimented with thresholding the set of fingerprints used based on the bias and accuracy of the fingerprints. By leaving out some portion of the least accurately predicted and most biased fingerprints the actual molecule ranks decreased (data not shown).

In the next experiment, we test the performance of the FingerID method against the approach of retrieving the closest matching spectrum from MassBank, under the assumption that the spectrum of the metabolite is MassBank but measured with a different collision energy, representing Case (b).

We trained our fingerprint prediction model in a stratified cross-validation setting where spectra of specific collision energy (10–50 eV) were chosen to the test fold and the other collision energies were used in training the model (Table 6). As the baseline, direct spectral matching from Massbank finds a match with correct metabolite in 94% of the cases. The performance of the FingerID method depends on both the molecular database used for retrieval (Kegg or PubChem) and the CE used for training and testing. The performance of FingerID coupled with Kegg is better than direct spectral matching when retrieving the 'middle' collision energy spectra (20–40 eV) but is weaker in predicting the two extremes, especially 10 eV spectra. When identifying against PubChem, the retriaval rate differences become larger from a minimum of 13.1% at 10 eV to a maximum of 70% at 30 eV.

5.5 Comparison to MetFrag

MetFrag is a state-of-the-art metabolite identification method, which assigns a score to candidate metabolite based on the similarity of a simulated spectrum to the observed one (Wolf *et al.*, 2010). The simulated spectrum is produced by combinatorially removing bonds from the parent ion and recording the resulting

Table 4. Statistics of the three metabolite identification cases with the three datasets

		MassBank query (%)		
Case		QqQ	Ltq	Lipids
(a)–(c)	Match found	100%	92.8%	97.0%
	Avg. rank	2.0 ± 0.1	220 ± 623	1.9 ± 0.9
	rank ≤ 10	100	63.2	99.5
(a)	rank ≤ 10	61.7	63.9	99.5
(b)	rank ≤ 10	94.0	–	–
(c)	rank ≤ 10	94.0	28.2	–
(d)	rank ≤ 10	0	0	0

The upper part of the table shows performance of reference queries in general, while the lower part indicates the performance when utilizing spectra only from the four different cases (Section 2). By definition, the (d) case gives no identifications.

Table 5. Statistics of the *de novo* metabolite identification (Case d) with FingerID, retrieving candidate metabolites from Kegg and PubChem, respectively

		Spectral dataset		
Molecular database		QqQ	Ltq	Lipids
Kegg	Avg. hits w/ ± 0.5 mass	24.9	26.6	
	Avg. rank	5.0 ± 7.3	3.2 ± 4.8	N/A
	rank ≤ 10	85.0%	91.8%	
	P_{50}	8.0×10^{-2}	3.8×10^{-2}	
PubChem	Avg. hits w/ ± 0.5 mass	28 648	27 862	12 928
	Avg. rank	$5196 \pm 13{,}168$	$1981 \pm 8{,}652$	11 ± 9
	rank ≤ 10	29.3%	50.8%	54.4%
	P_{50}	5.8×10^{-3}	3.2×10^{-4}	7.7×10^{-4}

fragments as possible explanations for the observed peaks. This parallels the idea that a good candidate should be able to produce all peaks by mostly bond cleavages.

We randomly selected a subset of 20 spectra from both QqQ and Lipids datasets, respectively. The QqQ represents nominal mass spectra with an absolute mass error set to 0.5, while Lipids is a high-resolution dataset with absolute mass error of 0.05. The appropriate mass errors were used in MetFrag. We queried both datasets against both Kegg and PubChem in MetFrag. Analysing the total of 40 spectra took approximately 1 day of manual work, as MetFrag does not support batch processing.

We measure the rank of the correct metabolite with both MetFrag and our method. The results are highlighted in Table 7. FingerID obtains favourable results to MetFrag in most cases, with significantly more retrieval results with top 10 rank, and higher overall recall rate. MetFrag found the correct metabolite for approximately half of the spectra from Kegg and for only a couple of spectra from PubChem. This is due to the default limit of 100 candidate structures, which allows for an analysis of spectral datasets in appropriate time frames.

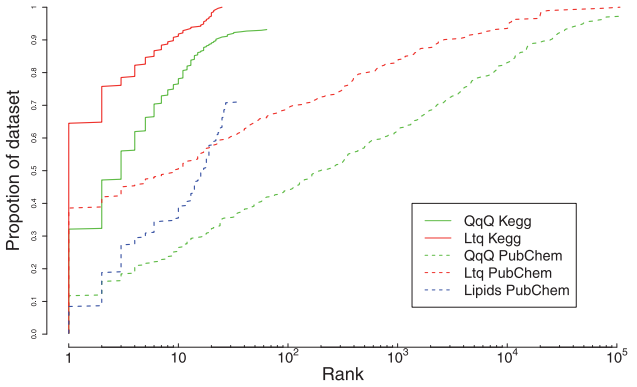

Fig. 4. The cumulative (log) rank distribution of the three datasets against Kegg and PubChem. The vertical axis indicates the ratio of molecules with a maximum rank indicated in the horizontal axis

Table 6. Comparison of metabolite identification against MassBank querying when the measured metabolite is present in a reference database, measured with different collision energy (Case b)

		FingerID					MassBank query
		10 eV	20 eV	30 eV	40 eV	50 eV	
rank ≤ 10	Kegg	76.0%	93.6%	97.6%	95.6%	91.8%	94.0%
	PubChem	13.1%	48.2%	70.0%	67.9%	38.8%	94.0%

Table 7. Comparison of metabolite identification against MetFrag on a subset of 20 spectra from both 'QqQ' and 'Ltq', respectively

Molecular database	Spectral dataset	FingerID			MetFrag		
		match	Avg. rank	rank ≤ 10	match	Avg. rank	rank ≤ 10
Kegg	QqQ	17	3.2	16/17	16	5.1	9/11
	Ltq	20	3.8	18/20	12	5.6	11/12
PubChem	QqQ	11	905	8/16	2	68	0/2
	Ltq	20	58	9/20	1	20	0/1

6 DISCUSSION

We presented a novel approach for *de novo* metabolite identification through intermediate fingerprint prediction based on tandem mass spectra. Our results indicate that it is possible to learn the statistical dependencies between tandem mass spectral signals and molecular properties, which can be used to score and rank metabolites, with good identification performance. Moreover, a sufficiently large set of fingerprint predictions can give useful clues to the actual metabolite identity to the human expert, even if exact automatic identification remains elusive.

The machine learning approach widens the utility of mass spectral databases such as MassBank. Due to the nature of the

method, any type of tandem mass spectral data is applicable. This also allows us to handle, for example, the rearrangement reactions, which result in exceptional and difficult-to-predict fragment products (Heinonen *et al.*, 2008). Utilizing multiple measurements increases the prediction accuracies through kernel-based data fusion. Our method can be easily complemented by inferring additional information about the metabolites, such as sum formulas from isotopic patterns (Böcker *et al.*, 2009) or by performing fragmentation analysis (Wolf *et al.*, 2010) in addition to fingerprint prediction.

Further research is obviously necessary to bring the machine learning-based approach towards a practical tool for metabolomics. In *de novo* metabolite identification, the identification performance depends, on one hand, on the uniqueness of the fingerprints to particular sets of metabolites, and on the other hand, the ability to predict these fingerprints from tandem mass spectra. It is an interesting future research direction to develop computational methods to identify sets of fingerprints that strike a good balance between these two qualities. An interesting machine learning approach for representing the molecular properties would be to use structured prediction (Bakir, 2007) to model the statistical dependencies between the fingerprints and the input spectrum.

Funding: The Academy of Finland grant 118653 and in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

Conflict of Interest: none declared.

REFERENCES

Bakir,G. (2007) *Predicting Structured Data*. MIT Press.

Böcker,S. *et al.* (2009) Sirius: decomposing isotope patterns for metabolite identification. *Bioinformatics*, **25**, 1–9.

Breiman,L. *et al.* (1984) *Classification and Regression Trees*. Chapman and Hall/CRC.

Curry,B. and Rumelhart,D. (1990) Msnet: a neural network that classifies mass spectra. *Tetrahedron Com. Methodol.*, **3**, 213–237.

Dayringer,H. *et al.* (1976) Computer-aided interpretation of mass spectra. Information on substructural probabilities form stirs. *Organic Mass Spectrometry*, **11**, 529–542.

Dworzanski,J.P. *et al.* (2004) Identification of bacteria using tandem mass spectrometry combined with a proteome database and statistical scoring. *Anal. Chem.*, **76**, 2355–2366.

Heinonen,M. *et al.* (2008) Fid: a software for *ab initio* structural identification of product ions from tandem mass spectrometric data. *Rapid Comm. Mass. Spectrom.*, **22**, 3043–3052.

Horai,H. *et al.* (2010) Massbank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom*, **45**, 703–714.

Jebara,T. *et al.* (2004) Probability product kernels. *J. Machine Learn. Res.*, **5**, 819–844.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in kegg. *Nuclead Acids Res.*, **34**, 354–357.

Kell,D. (2004) Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol*, **7**, 296–307.

Kondor,R. and Jebara,T. (2003) A kernel between sets of vectors. *ICML*, Proceedings of the Twentieth International Conference on Machine learning (ICML-2003), Washington DC.

McLafferty,F. (1983) *Tandem Mass Spectrometry*. Wiley, New York.

Neumann,S. and Böcker,S. (2010) Computational mass spectrometry for metabolomics: Identification of metabolites and small molecules. *Anal. Bioanal. Chem.*, **398**, 2779–88.

Oberacher,H. *et al.* (2009) On the instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. optimization and characterization of the search algorithm. *J. Mass Spectrom.*, **44**, 494–502.

Pavlic,M. *et al.* (2006) Combined use of esi-qqtof-ms and esi-qqtof-ms/ms with mass-spectral library search for qualittative analysis of drugs. *Anal. Bioanal. Chem.*, **386**, 69–82.

Shawe-Taylor,J. and Christianini,N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, United Kingdom.

Steffen,A. *et al.* (2009) Comparison of molecular fingerprint methods on the basis of biological profile data. *J. Chem. Inf. Model.*, **49**, 338–347.

Stein,S.E. (1994) Estimating probabilities of correct identification from results of mass spectral library searches. *J. Am. Soc. Mass. Spectrom.*, **5**, 316–323.

Wang,Y. *et al.* (2009) Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.

Werner,E. *et al.* (2008) Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends. *J. Chromatogr. B*, **871**, 143–164.

Wolf,S. *et al.* (2010) *In silico* fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, **11**, 148.