METABOLITE IN SILICO IDENTIFICATION SOFTWARE (METISIS): A MACHINE LEARNING APPROACH TO TANDEM MASS SPECTRAL IDENTIFICATION OF METABOLITES

by

LARS J. KANGAS

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
Department of Electrical Engineering and Computer Science

AUGUST 2012

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of LARS J. KANGAS find it satisfactory and recommend that it be accepted.

John H. Miller, Ph.D., Chair

Robert R. Lewis, Ph.D.

Li Tan, Ph.D.

ACKNOWLEDGEMENTS

This thesis is the result of team work with great help from many, including my advisors Drs. John Miller, Bob Lewis, and Li Tan who allowed me to see this work to an end.

Dr. John Miller was initially skeptical about the algorithm proposed for this thesis. This is understandable as my proposed solution was drastically different from other attempts to solve the same problem. Dr. Miller’s most appreciated contributions to this work are that his skepticism challenged me to ensure that details had solid foundations in science and engineering and maybe more importantly, he made our project meetings a lot of fun with many laughs. Thank You Dr. Miller.

Dr. Bob Lewis spent many hours with me to ensure the correctness of all equations and graphs modeling a mass spectrometer. I would not have succeeded in completing an accurate thesis without his help. Thank You Dr. Lewis.

Dr. Tom Metz, Pacific Northwest National Laboratory (PNNL), provided expertise about lipid metabolites and mass spectrometry. The idea to solve this problem started two years prior to this thesis work when Dr. Metz proposed the need for an algorithm that could accurately identify metabolites from mass spectrometry data. Thank You Dr. Metz.

Dr. Giorgis Isaac, PNNL, ran mass spectrometer experiments to analyze lipids and provided the data that were used in our algorithm development and testing. Dr. Isaac’s expertise in lipid metabolites provided much of what I know of these species. Thank You Dr. Isaac.

I wish to thank others at WSU and PNNL who provided much help: Bojana Ginovska-Panovska, our chemical expert, educated us on charges in molecules and defined chemical reactions for us to incorporate into our algorithm. Brian Schrom implemented the charge prediction algorithm. Luning Wang ran performance tests and implemented chemical reaction

templates. Dr. Paul Keller helped with proof readings and formatting of journal papers. Thank You all for the support of this thesis. It has been a lot of fun because I got to work with you.

Thank You Pamela for enduring yet another degree. Our daughter was born when I worked on a fourth college degree. Then there was one more degree before this last one. This time our daughter and I were in college at the same time—she finished her first degree before me, but she is short of my six degrees.

I want to acknowledge that funding for this research was obtained from the Laboratory Directed Research and Development program at PNNL, a national laboratory operated by Battelle for the U.S. Department of Energy under Contract DEAC06-76RL0-1830. Experimental work was performed at the Environmental Molecular Sciences Laboratory, a national scientific user facility located at PNNL and sponsored by the U. S. Department of Energy Office of Biological and Environmental Research.

METABOLITE IN SILICO IDENTIFICATION SOFTWARE (METISIS): A MACHINE LEARNING APPROACH TO TANDEM MASS SPECTRAL IDENTIFICATION OF METABOLITES

Abstract

by Lars J. Kangas, Ph.D.
Washington State University
AUGUST 2012

Chair: John H. Miller

Liquid chromatography-mass spectrometry-based metabolomics has gained importance in the life sciences, yet it is not supported by software tools for high throughput identification of metabolites based on their fragmentation spectra. An algorithm (MetISIS: metabolite *in silico* identification software) and its implementation are presented and show great promise in generating *in silico* spectra of metabolites for the purpose of structural identification. Instead of using chemical reaction rate equations or rules-based fragmentation libraries, the algorithm uses machine learning to find accurate bond cleavage rates in a mass spectrometer employing collision-induced dissociation tandem mass spectrometry.

A preliminary test of the algorithm and a comparison to another algorithm with 45 lipids shows both high sensitivity and specificity.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The field of proteomics has enjoyed considerable success in part due to software tools like SEQUEST (Eng, 1994) and Mascot (Perkins, 1999), which enable high throughput identifications of detected peptides (and their corresponding proteins) based on their fragmentation spectra as generated by collision-induced dissociation (CID). These tools have benefited from the close link between DNA and protein sequences and the fact that the polymeric structure of amino acid residues in proteolytic peptides provides a convenient basis for interpreting peptide tandem mass spectra. However, small molecules other than peptides have to be considered as two- or three-dimensional structures of atoms or functional groups of atoms, and these differences require novel algorithms. Indeed, while liquid chromatography-mass spectrometry (LC-MS)-based metabolomics has gained importance in the life sciences, it is not supported by software tools for high throughput identification of metabolites based on their fragmentation spectra in a manner analogous to the identification of peptides. The majority of recent published metabolomics informatics efforts have focused on the development of descriptive databases of metabolites (Kumar, 2012; Sakurai, 2011; Goto, 2002; Psychogios, 2011) or their metabolic pathways or software for analysis and visualization of metabolomics data (Karnovsky, 2011; Kastenmüller, 2011), often with complementary transcriptomics or proteomics datasets (Redestig, 2011; García-Alcalde, 2011). The success of these informatics efforts largely depends on the degree to which metabolite features (characterized by observed masses and elution times in LC-MS-based metabolomics studies) can be assigned chemical structures.

While several databases containing fragmentation spectra of metabolites have been developed for chemical standards analyzed using both LC-MS (Smith, 2005; Brown, 2009;

Psychogios, 2011 and gas chromatography-mass spectrometry (GC-MS) (Kind, 2009; Kopka, 2005), these databases are limited by the commercial availability, or the effort required for synthesis, of authentic standards. Further, these databases do not exploit the wealth of information contained therein for the development of tools for predicting chemical structures of previously uncharacterized metabolites based on their fragmentation spectra.

To date, the main approaches for predicting in silico tandem mass spectra of non-peptide small molecules are based on either chemical reaction equations, libraries of fragmentation spectra/pathways, or bond cleavage probabilities using bond strengths. None of these approaches (briefly discussed below) have shown sufficient accuracy in generating in silico spectra to enable automated and correct identifications of non-peptide small molecules.

Chemical reactions involving unimolecular dissociation are commonly studied using the Rice–Ramsperger–Kassel–Marcus (RRKM) (Marcus, 1952) and quasiequilibrium theories (QET) (Rosenstock, 1952). Chemical reaction time evolutions are described in systems of differential equations as in a master equation approach. These theories are invaluable to understanding chemical reaction mechanisms and energies required for state transitions. Yet, while quantum calculations like RRKM and QET explain the dissociation of activated ions, they are insufficient in explaining the activations of ions in inelastic collisions for molecules larger than a few atoms (Sleno and Volmer, 2004). Predicting or identifying fragment ions in tandem mass spectra is difficult for large molecules; indeed, little is known regarding the rates at which ions gain internal energy in activation methods and at what energies bonds dissociate. A few small peptides such as leucine enkephalin and bradykinin have been empirically studied as to their fragmentation behaviors; however, these results cannot be translated to the fragmentation of non-peptide small molecules (Drahos and Vékey, 1999; Gabelica, 2003; Vékey, 1996). Hence, it remains difficult to explain many ions and their intensities in fragmentation spectra or to generate accurate *in silico* spectra knowing only the molecular properties, such as atomic compositions and bonding patterns.

Tools such as Mass Frontier from Thermo Scientific (Highchem, http://www.highchem.com/) and ACD/MS Fragmenter, (ACDLabs, http://www.acdlabs.com/products/adh/ms/ms_frag/) generate fragments using a large library of rules describing fragmentation pathways. This can become unmanageable in that rules are not necessarily exclusive—one rule can affect another rule. Sometimes the correct rules are not available or are not available with sufficient specificity. Mass Frontier generates "bar code" spectra where all ions have the same intensity because bond cleavage rates are not considered. Bar code spectra are not sufficient when many molecules generate the same fragment ions. In these instances, only the relative ion intensities will aid the correct identification.

Hill et al. (2005) and Wolf et al. (2010) chose a bond disconnection approach to generate fragments from molecules. Hill et al. have user-defined criteria for bond cleavages, while Wolf et al. generate all possible topological fragments in their tool MetFrag and then score these by measures such as bond dissociation energies (BDEs). Unfortunately, BDEs vary significantly as molecules increase in size beyond only a few atoms and where atoms beyond the nearest atoms influence the outcome. For example, Bach et al. (1996) showed that the O—O BDEs were predicted at 22.73 kcal/mol for CH3C(CH2)O—OH and at 48.32 kcal/mol for CH3C(O)O—OH. It may be surmised that many BDEs are either not available or sufficiently accurate.

Thus, defining an accurate algorithm that identifies non-peptide small molecules from collision activated tandem mass spectra is still an open problem. The developed algorithm MetISIS: metabolite *in silico* identification software shows great promise in generating *in silico* spectra of metabolites for the purpose of structural identification. That is, the algorithm uses no chemical reaction equations (in which parameters have to be estimated), no fragmentation rules from observed pathways, and no bond cleavage rates provided by bond strengths. Instead, the algorithm simulates the fragmentation process in a mass spectrometer model using a machine learning approach to overcome the difficulties that result from unknown quantities and simplifying assumptions.

An evaluation, described in Chapter 10, with a small subset of metabolites showed MetISIS to have significantly higher accuracy, sensitivity, and speed in identifying metabolites than MetFrag (Wolf, 2010; Hildebrandt, 2011) whose developers in turn showed improved performance above that of the Mass Frontier software from Thermo Scientific (Highchem, http://www.highchem.com/).

The ultimate goal of MetISIS development is to provide the first means to analyze metabolites like the practitioners in proteomics have enjoyed for the identifications of peptides. MetISIS has the potential to become a standard for metabolite identification in LC-MS, enabling highly accurate high-throughput metabolomics studies. Similar to the way in which SEQUEST (Eng, 1994) and Mascot (Perkins, 1999) have allowed the proteomes of a large number of species to be mapped, MetISIS will allow metabolomes to be mapped for a vastly improved understanding of systems biology. Metabolites have not been utilized as biomarkers as much as peptides due to the lack of good identification tools.

MetISIS, thus shows promise to be a tool in the health sciences and in the metabolomics community where metabolites have to be identified to diagnose diseases and to understand the biology behind diseases.

2 ION FRAGMENTATION IN MASS SPECTROMETRY

Mass spectrometry is an analytical technique to measure masses of ions or more accurately, the mass-to-charge ratios (*m/z*)—the correct mass can be determined if the charge is known. An ion is a molecule that has an excess of one or more positive or negative charges.

2.1 Ion Fragmentation

Mass spectrometry experiments can be performed to measure either only a molecule's mass or also masses of fragments of that molecule. The objective of a fragmentation step is to cleave a few covalent bonds in the ion so that the detected fragment ions provide more information about the identity of the molecule being analyzed. Identification of these species is typically called elucidating their structures like when identifying metabolites from mass spectrometry.

In all methods, a precursor ion is a charged species of the molecule initially injected into the instrument and product ions are the result of fragmentations. The old terms parent ions and daughter ions are still frequently used for precursor ions and product ions, respectively.

When the mass spectrometry experiment uses one fragmentation step, this is often called tandem mass spectrometry or MS/MS, or alternatively MS^2 to reflect that MS^n is used to show that more than one MS step is used. MS^3 equals MS/MS/MS and says that the experiment was run such that fragment ions from the first fragmentation step were fragmented in an additional step.

The covalent bonds cleaving in MS/MS for a specific species tend to follow certain patterns that can be exploited for structure elucidations (observe that the work presented here used only MS/MS configured mass spectrometers). During the fragmentation steps, the weakest bonds have higher probabilities to cleave. If the fragmentation is performed on ions of which there are, for example, 50k copies, 40k of these may cleave at one specific bond, 5k may cleave at another bond, and the last 5k may not cleave at all. The resulting spectrum from the experiment is a histogram—a count of how many ions was detected as having cleaved at specific *m/z*s.

3 BOND CLEAVAGES

Collision induced dissociation (CID) is an MS mechanism to fragment ions by cleaving covalent bonds. CID generates and moves kinetic energy from atom/molecule collisions to internal energy for the molecule. In a population of molecules, bond cleavages occur at rates that are functions of the distribution of internal energies. As internal energy is added through collisions, initially the cleavage rate is low for a specific bond but then increases to a maximum value, which cannot be maintained because the ion trap gets depleted of candidate molecules that still have that specific bond. The possible candidates are also depleted by molecules possibly being fragmented through other bond cleavages. In a linear ion trap MS instrument, the driving force for collisions is a function of the precursor ions' *m/z*s. Thus, in any bond cleavage that produces fragment ions, those fragment ions will have too low *m/z*s to continue being heated. The fragment ions will slowly cool and eventually drop below the internal energy levels where they experience bond cleavages.

We define bond-cleavage energy as the internal energy level of the ion where half of the ions in the population would have cleaved that bond when the ions are slowly heated. This is illustrated by a break-down diagram like Figure 3.1 specific to each bond. Figure 3.1 shows a hypothetical example (Vékey, 1996) where 50% of the precursor ions are expected to have dissociated at a specific bond at approximately 2.2 eV internal energy levels. Most molecules will have multiple fragmentation pathways that compete and the percentage breakdown thus refer to each individual pathway.

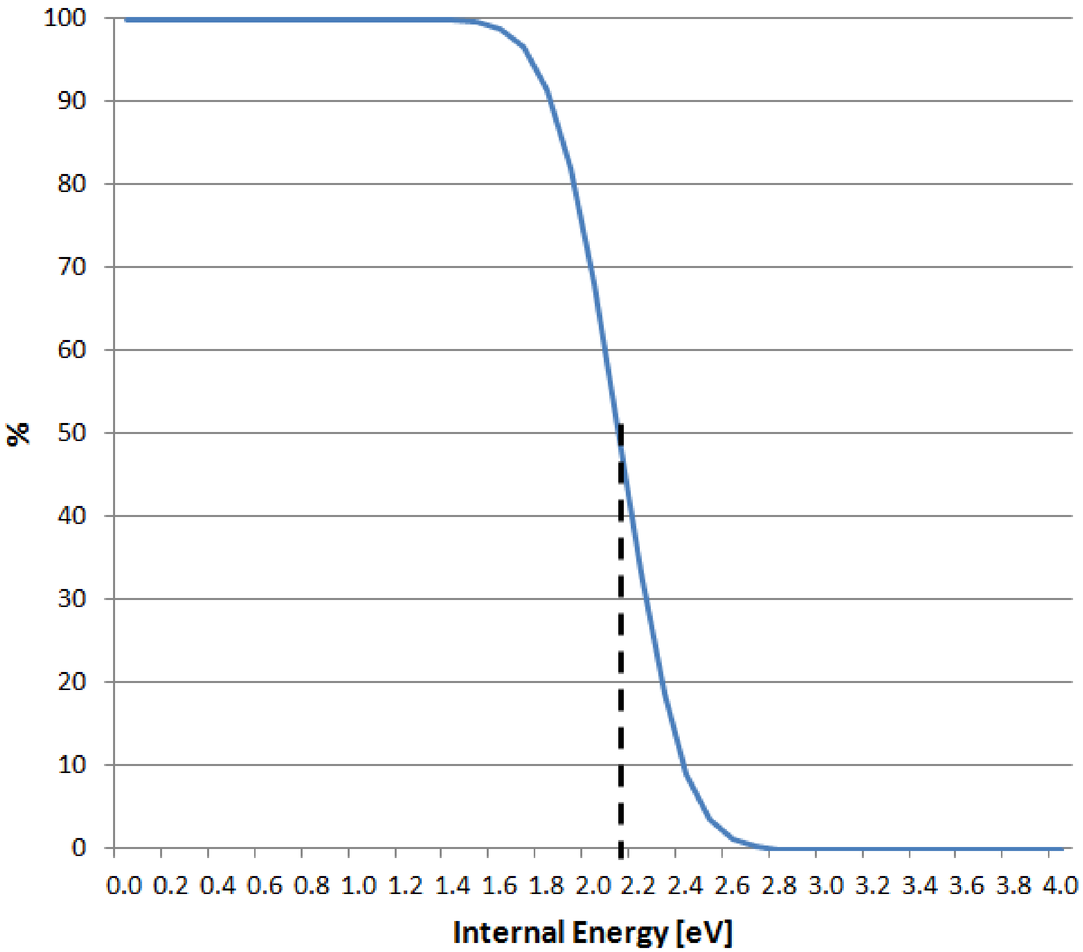


Figure 3.1. Breakdown Diagram. The percentage of precursor ions that has a specific bond still intact at increasing internal energy levels.

4 ALGORITHM DESCRIPTION

The MetISIS algorithm which is based on Monte Carlo simulation is conceptually simple. The algorithm has two phases: first it learns to predict bond cleavage energies from which cleavage rates can be calculated.

In the second phase, the algorithm generates *in silico* tandem mass spectra from molecular structures and uses these spectra in identifications of lipids. Figure 4.1 shows the flowchart of the algorithm. The three components above the dashed line are associated with the machine-learning phase, and the dotted rectangle around Molecules and Experimental Spectra signifies that these are given in pairs to the algorithm during the learning phase. The algorithm is, in essence, learning the mapping function from molecular structures to spectra. The components below the dashed line are those that are involved in generating *in silico* spectra both in the training phase and from a trained algorithm.

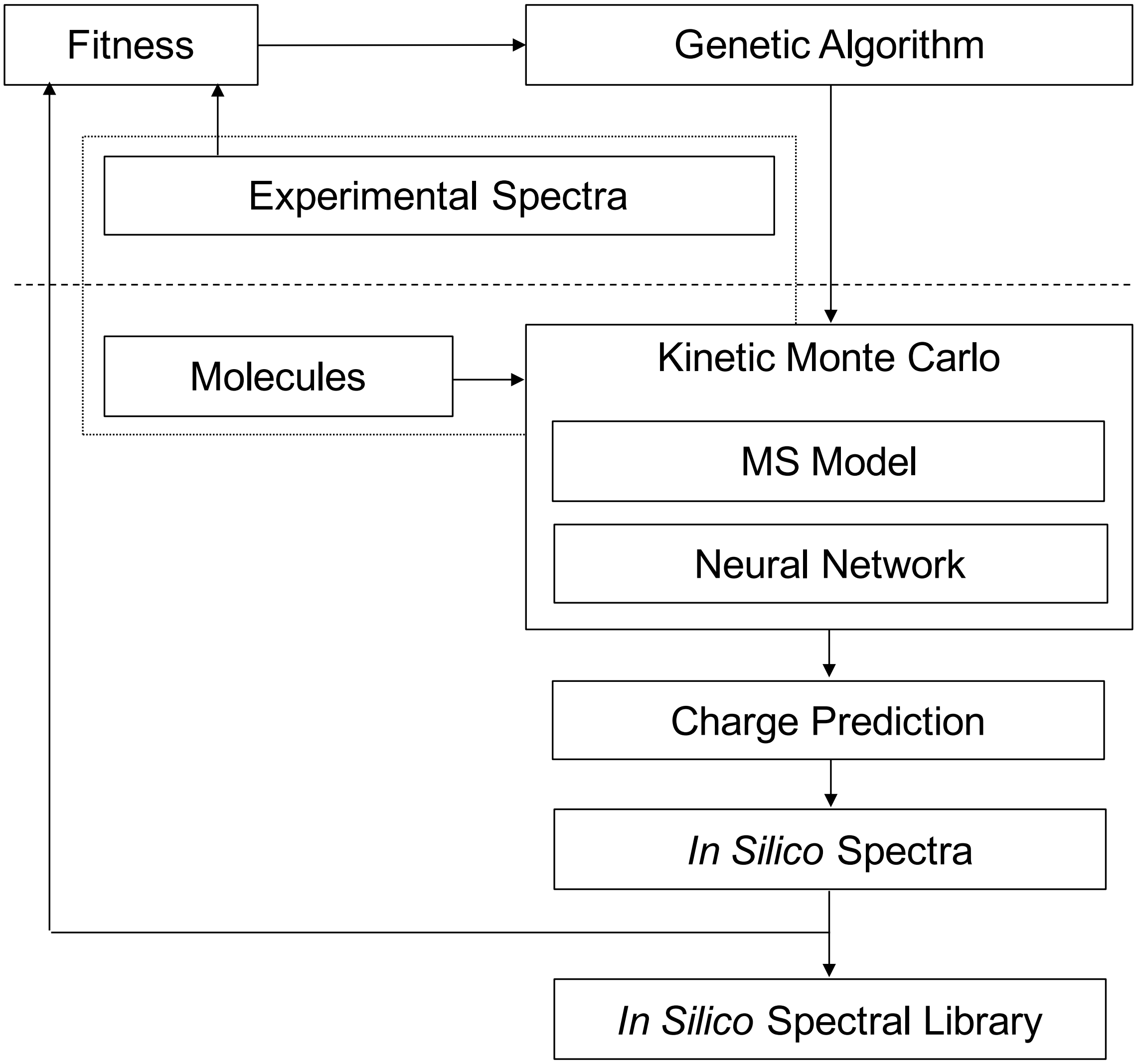


Figure 4.1. Flowchart of the MetISIS algorithm.

The machine-learning phase where bond-cleavage energies are learned proceeds as follows: based on molecule/experimental spectrum pairs, the algorithm uses a model of the linear ion trap and the artificial neural network (ANN) in kinetic Monte Carlo (KMC) simulations to incrementally learn bond cleavage energies in CID. For each molecule in a training set, a KMC simulation uses the molecule's structural information to generate an *in silico* tandem mass spectrometer spectrum that is compared to an experimental spectrum from the same molecule. The squared Pearson product-moment correlation coefficient (R-square) measured between these two spectra is added to a cumulative fitness score. That fitness score represents the goodness of one hypothetical solution to a genetic algorithm discussed in Section 6.1.1 that iteratively continues to optimize a best hypothetical solution in a set of hypothetical solutions. The variables that are optimized in each solution are the weights for the ANN discussed in Chapter 6 that predicts bond cleavage energies. Improved predictions increase the fidelity of the *in silico* spectra.

In the application phase, the best solution determined in the learning phase from the training lipids is used to generate *in silico* tandem mass spectra for novel lipids, i.e. a library is populated with *in silico* spectra based on a large database of lipid molecular structures. Finally, experimental tandem mass spectra of unknown lipids are searched against the library of *in silico* spectra for matches that define a ranked list of candidate identifications.

4.1 LTQ Linear Ion Trap Mass Spectrometer Model

The Thermo Scientific LTQ linear ion trap MS is a tandem in time spectrometer coupled to an electrospray ionization (ESI) source (Fenn, et al., 1989). The instrument ionizes samples in either positive of negative mode, i.e., protonates or deprotonates molecules to give them a charge.

In the first step the instrument selects precursor ions, and then proceeds to fragment these one by one to identify product ions in an MS/MS step. An MS step alone gives the correct mass-to-charge ratio (m/z) measurements of ions, and if the ions have charge one which is typical for most metabolites, then the m/z corresponds to the mass of the molecule as an adduct (i.e., the neutral metabolite plus a charged residue added by electrospray ionization) in Daltons (Da). An adduct is the single chemical reaction product of two or more distinct molecules, for example the protonated [M+H]+ or [M+NH4]+ adducts where *M* is the mass of the molecule analyzed in an MS experiment.

The fragmentation of precursor ions in the MS/MS step cleaves bonds to generate product ions, fragments of the original molecules. The fragment ions provides a richer set of information to identify the molecules; two different molecules that have the same mass are indistinguishable by MS alone, but if they have different fragmentation pathways, then the fragment ions from MS/MS will distinguish the two molecules.

The ions in an MS instrument are increased in temperature to cause the dissociation of molecular bonds. The molecules to be measured in the MS instrument are injected at room temperature. The molecules gain some internal energy in the ionization in ESI, but lose most of this from collisional cooling before the next step, the excitation in an ion trap using collision induced dissociation (CID).

The initial temperatures of molecules entering the ion trap and CID are estimated to have a Gaussian distribution of temperatures with a mean of 298K (Drahos and Vékey, 1999). The temperature of molecules translates to molecular vibrations, periodic motions in the form of translational and rotational motions. Higher temperatures or thermal energy are associated with larger vibrations in the molecules normal modes of vibration. The mean internal energy (E_{therm}) per oscillator (s is the number of oscillators defined in Equation (4.3)) is,

$$E_{\text{therm}}/s \; = C_{\text{peptide}} k_{\text{B}} T, \tag{4.1}$$

where k_{B} is the Boltzmann constant 8.617343 × 10^{-5}eV K^{-1}, and C_{peptide} is a temperature dependent factor ~0.2 for peptides with temperature T in Kelvin (Drahos and Vékey, 1999).

$$C_{\text{peptide}} = 5.61 \times 10^{-4}T - 1.24 \times 10^{-7}T^2, \tag{4.2}$$

where T is temperature in Kelvins.

Equation (4.1) and Equation (4.2) use C_{peptide}, a function specific to peptides. Drahos and Vékey suggest that most organic compounds have similar functions, and that the C_{peptide} function can be used for most organic compounds.

Assuming that only the vibrational oscillation is dominant, a molecule then has degrees of freedom (s),

$$s = 3n - 6, \tag{4.3}$$

where n is the number of atoms in the ion and n > 2. A diatom (n = 2) has one degree of freedom.

The width W of the distribution is

$$W = 1.8 \times 10^{-4}T\sqrt{s} \tag{4.4}$$

determined by regression analysis (Drahos and Vékey, 1999).

The internal energy distribution of the ion is Gaussian like (Drahos and Vékey, 1999), and is defined at energy E by

$$P(E) = \frac{1}{W\sqrt{\pi/2}} exp\left[-2\frac{(E-E_{\text{therm}})^2}{W^2}\right] \tag{4.5}$$

4.2 Ion Trap and Collision Induced Dissociation

Collision induced dissociation (CID), a slow-heating method to fragment molecules in an ion trap, transforms kinetic energy to internal energy (primarily vibrational) by collisions with an inert gas (helium in this configuration).

The (precursor) ions in this dipole-excited ion trap oscillate at high velocities in an oscillating electric field with a frequency of a few hundred kHz. Energy transfer from hundreds to a few thousands of collisions slowly increases internal ion energies until bonds cleave. The fragments of a precursor ion immediately stop oscillating because their m/zs no longer match the resonant frequency of the precursor ions. Note that the precursor ion with charge +1 or -1 cleaves into an ion with the same charge and one neutral fragment, which is not detected.

The fragment ion that no longer is heated by collisions starts to lose energy to the environment. A cooling schedule for this internal energy decrease was modeled by (Zhang, 2004). T_{eff}, effective internal temperature decrease exponentially to the temperature of the buffer gas T_0,

$$T_{\mathrm{eff}} = \left(T_{\mathrm{precursor}} - T_0\right)\exp\left(-r_{\mathrm{c}}t\right) + T_0, \tag{4.6}$$

where t is the elapsed time after the precursor has been fragmented, and for an ion with mass M, r_{c} is

$$r_{\mathrm{c}} = r_c^0(M/1000)^c, \tag{4.7}$$

where r_c^0 is a cooling rate of an ion of mass 1000 Dalton and c is a constant. r_c^0 = 104.6s^{-1} and c = 0.74 after an optimization in the paper.

4.3 Activation Time in the CID

The ions continue CID collisional excitations in the ion trap for a preset time, 30 msec in this configuration. Too short time will not fragment ions or not fragment sufficiently many ions to generate good fragmentation counts for ion spectra. Increasing the activation time beyond where all precursors have experienced at least one cleavage increase the cycle time of the CID without changing the composition of fragment ions in the ion trap, since all cleavages cause the fragments to become unexcited and they will thus not experience further collisions. The cooling schedule will also reduce the occurrences of spontaneous cleavages in the absence of collisions due to residual internal energy. Thus, consistent with the ion trap instrument, the Kinetic Monte Carlo (KMC) simulation in MetISIS does not excite fragment molecules to further collisions with the inert gas, but the fragment molecules continue to be tested for spontaneous fragmentations.

The Kinetic Monte Carlo (KMC) simulation requires a rate of collisions for which the collision ions are selected. Time is moved forward to keep track of the cumulative time that is limited to the excitation time for the simulation.

The KMC simulation increments time by $-\ln(\zeta)/R$, where R is the total rate for all possible collision events in the system, and ζ is a uniformly sampled random number in (0, 1] (Voter, 2005). KMC thus uses two random numbers, one to select the ion for a collision event and one to move time forward.

For a collision between an ion and a target atom, a larger ion moving with larger velocity in a higher number density n of the collision gas will have a higher probability of collisions.

The collisional cross-section σ for an ion and target atom is

$$\sigma = \pi(r_1 + r_2)^2, \tag{4.8}$$

where r_1 is the van der Waal's radius for the target gas and the radius of the ion, r_2, is

$$r_2 = \sqrt[3]{\textstyle\sum_i R_i^3}, \tag{4.9}$$

where R_i is the radius of each atom i in the ion.

The sampled mean free time, τ , can be defined from the gas number density, collisional cross-section, and velocity as

$$\tau = -\frac{1}{\rho\sigma v}\ln(\zeta'), \tag{4.10}$$

where ζ' is a (new) random number in (0-1], σ is the cross-section given by Equation (4.8), v is the relative velocity between an ion and a target atom, and ρ is the number density of the gas given by the ideal gas equation

$$\rho = \frac{\mathrm{N_A}P}{\mathrm{R}T}, \tag{4.11}$$

where $\mathrm{N_A}$ is Avogadro's number, P is the pressure, T is the temperature, and R is the gas constant.

Observe that the term "$-\ln(\zeta)$" in Equation (4.10) and in the calculation for the time increment for the KMC simulation has a mean of unity. The random numbers are added to increase the heterogeneity of the population of KMC simulated ions.

4.4 Ion Spectrum

At the completion of the activation time for the CID method, the ions in the ion trap are moved to a detector that accumulates the ions onto a spectrum according to their m/z. The spectrum is thus a histogram showing counts of ions at each m/z. The resolution of the specific MS instrument determines the ability to distinguish two ions with almost the same m/z values.

The LTQ linear ion trap in CID activations has a relatively high cutoff m/z below which the instrument will not detect ions. For this ion trap, that cutoff is often called the one-third-rule to say that ions with m/zs below approximately one third of the precursor ion m/zs are not detectable. This cutoff is formally named the low mass cutoff (LMCO), and for the LTQ linear ion trap it is defined from q, the activation value, by

$$\mathrm{LMCO} = \frac{m/z_{precursor}q}{0.908}, \tag{4.12}$$

where 0.908 is called *qz,* a dimensionless parameter that determines the stability of ions' trajectories in the ion trap and depends on ions' *m*/*z*s, the size of the trap, and the amplitude and frequency of the fundamental radio frequency in the trap.

The activation value q, also called the Mathieu stability parameter, is a space charge parameter set on a linear ion trap mass spectrometer to stabilize ions in the trap, i.e., to ensure that the ions are not prematurely ejected from the trap. q was set at 0.18 in the MS experiments for this work. Default value for q is 0.25, but is often lowered to detect smaller fragments which is a compromise because a higher q tends to produce higher fragmentation efficiencies (Schwartz, 2002). As the q values inserted into Equation (4.12) suggest, the one-third-rule is in the name only, and does not describe the exact LMCO. (Observe that "*qz*" is not a product of q and charge z.)

4.5 Kinetic Monte Carlo Simulation

A kinetic Monte Carlo (KMC) algorithm simulates the dynamical evolution of a system by moving time forward relative to probabilities of stochastically sampled events [Bortz *et al.*, 1991; Gillespie, 1976; Meng and Weinberg, 1994; Young and Elcock, 1966]. A system simulated with a KMC usually has events occurring at different time scales. The algorithm has to span multiple time scales where the less frequent events still need to occur within the (limited) iterations. Dynamically moving the time as a function of the frequency of events thus makes it possible to simulate systems over vastly longer periods of time by "skipping" over time when no events occur [Voter, 2005].

The KMC algorithm simulates the slow heating of ions in a linear ion trap (Sleno and Volmer, 2004). Initially, the model linear ion trap is loaded with a number of replicates of the same ion with internal energies stochastically sampled from an electrospray ionization model (all distributions in the algorithm are assumed to be Gaussian instead of “Gaussian like” or Boltzmann (Drahos *et al.*, 1999; Drahos and Vékey, 1999; Gabelica and De Pauw, 2005; Naban-Maillet *et al.*, 2005; Pak *et al.*, 2008). The iterations of the KMC proceed by selecting an event—a specific ion to experience one collision with an inert gas atom in the linear ion trap. After the collision, the ion is tested to see if one of its bonds will cleave at the achieved internal energy of the ion. Typically, the ions experience a large number of collisions before reaching energy levels sufficient for fragmentation. The simulation stops when the KMC has accumulated incremental time steps equivalent to the excitation time set for the linear ion trap (30 ms). Next, all simulated intact ions and fragment ions are added to an *in silico* spectrum.

4.6 The Major KMC Steps

Step 1. Selecting a collision event

In a slow heating environment like a linear ion trap, both larger and faster ions moving in an environment of inert (e.g., helium) atoms have higher rates of collisions compared to smaller and slower ions. The rate is not only an increasing function of each ion’s collision cross-section and velocity but also of the number density of the collision gas. As shown above in Equation (4.10), the inverse of the product of collision cross-section, molecular velocity, and the number density of the collision gas is the mean free time, i.e., the time interval between collisions.

Suppose we have N distinct ions, each with a collision rate r_i, where $i \in [1 \ldots N]$. (The mean time between collisions for an ion i is $1 / r_i$.) We define R_k, the cumulative sum of r_i, as

$$R_k \equiv \begin{cases} 0, \ k = 0 \\ \sum_{i=1}^{k} r_i \,, \ 1 \ \leq k \ \leq N \end{cases} . \tag{4.13}$$

The total collision rate, then, is R_N.

Assuming a Poisson process, given a continuous random number ζ_0 ($0 \leq \zeta_0 < 1$) we could generate a time to the next collision with any ion as

$$\Delta t = -\frac{ln\,\zeta_0}{R_N} \tag{4.14}$$

and use this to advance the KMC simulation "clock".

The collision rate r_i determines the relative collision probability for species i, so we could use another continuous random number ζ_1 to determine the colliding ion by solving

$$\frac{R_{k-1}}{R_N} \leq \zeta_1 < \frac{R_k}{R_N} . \tag{4.15}$$

An efficient way to do this is to precompute bounds and do a binary search on ζ_1.

Step 2. Performing a collision

Each collision between an ion and an inert atom in the linear ion trap provides at the most a few hundredths of one eV, while the critical energy needed for dissociation can be several eV (Sleno and Volmer, 2004) or approximately 0.007eV per degree of freedom for a molecule (Vékey, 1996).

The ions in a dipole-excited ion trap oscillate at high velocities in an oscillating electric field of a few hundred kHz. The kinetic energy of the ion as a function of its mass and velocity is, in the collision with an inert atom, calculated into a center of mass frame kinetic energy. This is the maximum ion collision energy that can be converted from kinetic energy to internal energy for the ion.

The KMC simulation assumes that the ion velocities in the ion trap oscillate like a cosine function. Using a random number, the velocity of an ion is sampled from the cosine curve to compute collision energy.

The normalized collision energy schedule, E_{schedule} (Volt peak-to-peak) in the ion trap is an *m/z* adjusted collision energy following an estimated linear regression line (Gabelica *et al.,* 2003; Lopez *et al.*,1999; Thermo Product Support Bulletin, PBS 104),

$$E_{\text{schedule}} = 0.002\, m/z \;+\; 0.4 \tag{4.16}$$

for a 30% collision energy (maximum 5 Volt peak-to-peak in an oscillating electric field). Adjusted for a user specified collision energy percent (*Coll*) it is

$$E = Coll/30\; E_{\text{schedule}}\,. \tag{4.17}$$

The relative velocity *v* between an ion and a target atom is

$$v = \sqrt{\frac{2E}{m}}\,, \tag{4.18}$$

where *m* is the mass. (The target is assumed stationary.)

In the dipole excited oscillating electric field in the ion trap, the instantaneous velocity magnitude v_{inst} is sinusoidal:

$$v_{\text{inst}} = v\;|cos\,\zeta\pi|, \tag{4.19}$$

where ζ is a uniformly sampled random number in [0, 1].

The instantaneous laboratory frame kinetic energy of the ion E_{lab} is then

$$E_{\text{lab}} = \frac{1}{2} m v_{\text{inst}}^{2}\,. \tag{4.20}$$

The energy of interest is the kinetic energy in the center-of-mass reference frame, E_{com} that is the maximum collision energy potentially available as internal energy to the ion (Shukla and Futrell, 2000; Sleno and Volmer, 2004),

$$E_{\text{com}} = \frac{m_{\text{target}}}{m_{\text{target}} + m_{\text{ion}}} E_{\text{lab}} \ . \tag{4.21}$$

Not all the kinetic energy available is transformed into internal energy for a molecule because, for example, many collisions between a target atom and a molecule are glancing only. Different collision efficiencies are suggested in the literature, but while there is a consensus on there being less than maximal energy transfer, the exact numbers for different molecules are only estimated (Schneider, *et al*., 2001; Wells & McLuckey, 2005); Cunningham & Glisch, 2006). As a first approximation, we selected a few suggested collision efficiencies for known peptide molecules and found a regression line (Equation (4.22)) to apply in the KMC simulation.

$$E_{\text{actual}} = E_{\text{com}}\,(0.0006 m_{\text{ion}} + 0.2195). \tag{4.22}$$

Equation (4.22) implies a linear increase in the efficiency of energy transfer with increasing ion mass and gives an efficiency range from about 0.4 to 0.9 for lipids from 250 to 1060 Dalton. E_{actual} is the amount of incremental energy from each collision added to the molecule's internal energy.

After an energy transfer, KMC simulation continues with a dissociation test to see if the ion has reached sufficient internal energy to cleave a bond. Mean internal energy is referred to as an internal temperature in the discussion that follows. The relationship between these two quantities is described in equations (4.1) and (4.2) from (Drahos and Vékey, 1999).

Step 3. Calculating the cleavage probability after a collision

The internal energy of an ion is thermal-like from both the heating in the ESI and the CID (McLuckey and Goeringer, 1997; Naban-Maillet *et al.*, 2005; Pak *et al.*, 2008). While at low energies the distribution is Poisson, after additional heating in the CID, the distribution tends towards Gaussian with variance a function of energy and degrees of freedom as proposed by Drahos and Vékey (1999).

The ion selected in Step 2 is tested for bond dissociation by an ANN first assigning cleavage energies to bonds. Next, probabilities of internal energies are assigned to each ion. Figure 4.2 shows the integration of the cumulative probability densities $Q_k(T)$ for four hypothetical bonds $k \in [1, 2, 3, 4]$. We further define $Q_k(T)$ the fractions

$$Q_k(T) = \int_{E_{0,k}}^{\infty} P(E,T)dE \qquad (4.23)$$

where $E_{0,k}$ is the specific energy level at which bond k cleaves from the probability density function $P(E,T)$ of the ion internal energies E at temperature T.

Note that each bond cleavage is assumed an independent event and the bond dissociation test allows either no bond cleavage or exactly one bond to cleave in the calculations below.

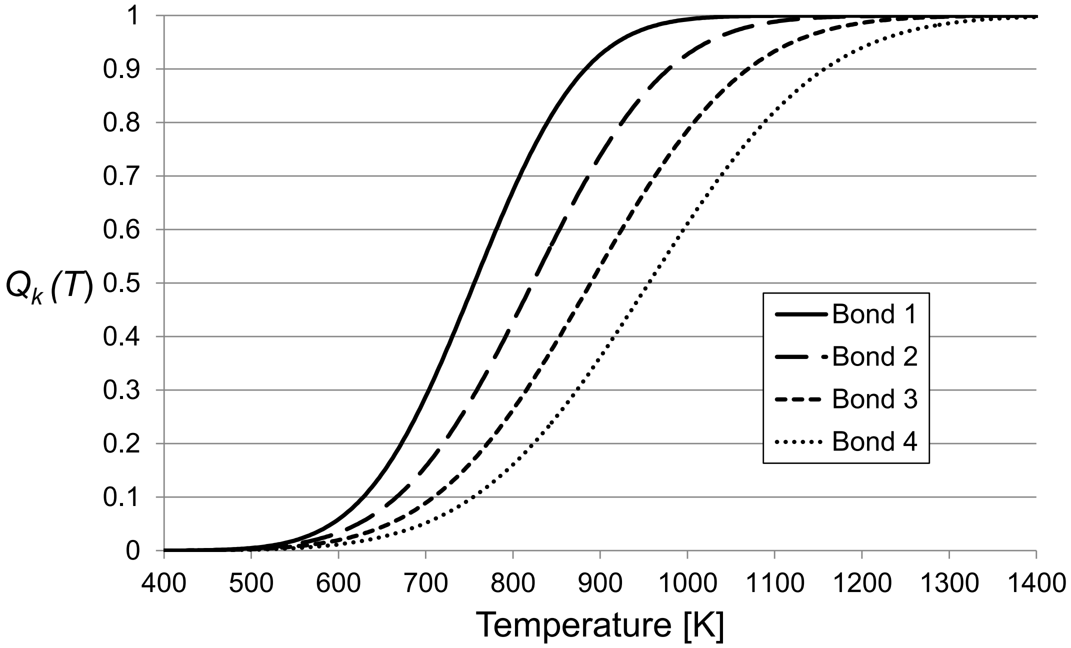


Figure 4.2. Fractions of Bonds Cleaving at Increasing Internal Energies.

The cumulative probability Q_C of one and only one of B bonds cleaving at temperature T is

$$Q_C(T) = \sum_{k=1}^{B} S_k\,(T), \tag{4.24}$$

where $S_k\,(T)$ is the contingent probability of bond k breaking:

$$S_k(T) = \; Q_k(T)\; \prod_{j\neq k}(1 - Q_j(T)). \tag{4.25}$$

Allowing only one bond to break or no bond to break, the relative cumulative probability $\tilde{Q}_C$ of one and only one of B bonds cleaving is

$$\tilde{Q}_C(T) = \frac{Q_C(T)}{Q_C(T) + S_{NC}(T)}\,, \tag{4.26}$$

where the probability of no cleavage is

$$S_{NC}(T) = \; \prod_{j=1}^{B}(1 - Q_j(T)) \tag{4.27}$$

and the relative contingent probability $\tilde{S}_k(T)$ of bond k breaking is

$$\tilde{S}_k(T) = \frac{S_k(T)}{Q_C(T) + S_{NC}(T)}\,. \tag{4.28}$$

Observe that we continue to use tildes with relative probabilities when allowing only one bond break or no bond break. Figure 4.3 shows the cumulative probability for a hypothetical lipid with four bonds with their relative contingent probabilities $\tilde{S}_k(T)$, where k range from 1 to 4.

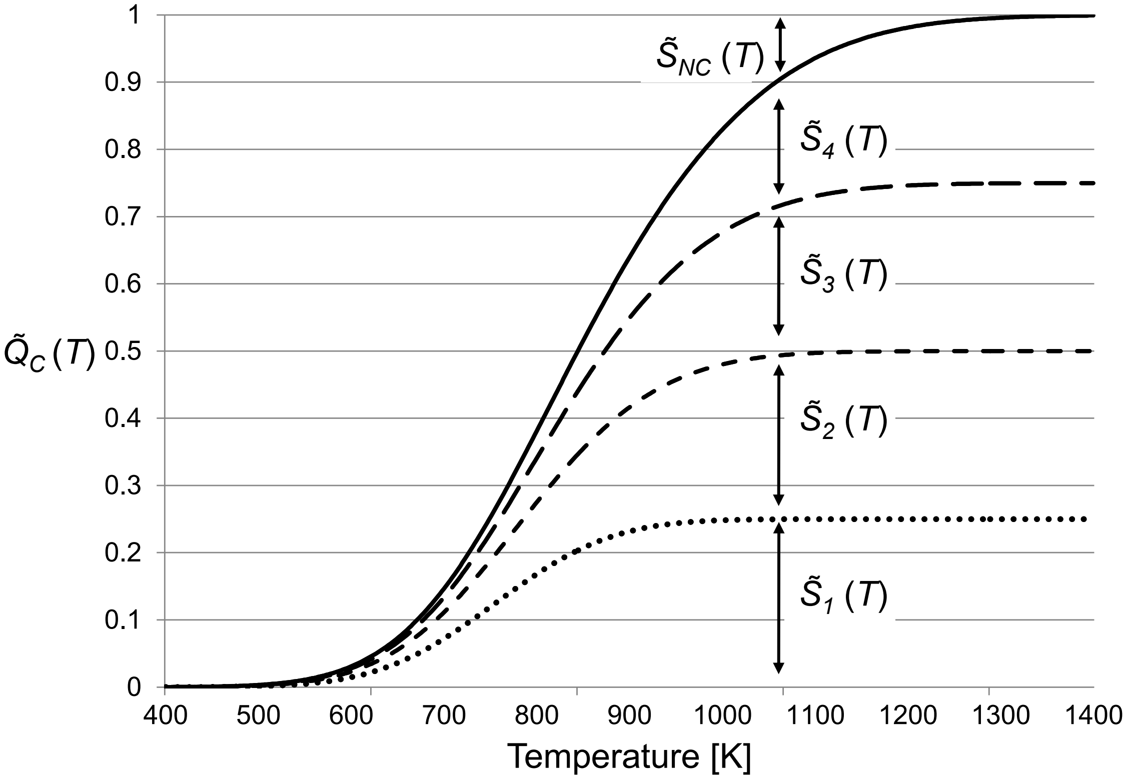


Figure 4.3. Hypothetical Cumulative Probabilities for a Lipid with Four Bonds. $\tilde{S}_{NC}(T)$ is the contingent probability of no bond cleaving.

Assuming multiple cleavages do not occur, the contingent probability of no cleavage, $\tilde{S}_{NC}(T)$ occurring at temperature T is

$$\tilde{S}_{NC}(T) = 1 - \tilde{Q}_C(T). \tag{4.19}$$

This contingent probability $\tilde{S}_{NC}(T)$ is shown in the figure as the vertical extent above the bonds. An event, a specific bond cleaving or a "no cleavage" is selected by a random number in the range (0-1).

The ion selected for a collision has at this time reached the end of this KMC iteration. If a bond cleaves and produces two separate fragments, a singly-charged ion and a neutral molecule, both fragments would replace the molecule that was fragmented in the ion trap. A charge prediction model labels each fragment as either an ion or a neutral.

The fragment ions lose their velocities as their *m/z*s no longer resonate with the dipole excitation frequency in the trap. This removes them from the collision selection, excluding them from additional collisions in subsequent KMC iterations. Further, a cooling schedule is applied to fragments in the ion trap (Zhang, 2004), meaning that the internal energies of fragments decrease due to a lack of collisional heating.

Even if fragments lose their abilities to gain further energies from collisions, the potential spontaneous cleavages of fragments are allowed in the simulations. The internal energy of a precursor ion before fragmentation is proportionally distributed to the fragments according to their degrees of freedom. The algorithm also predicts new bond cleavage energies for the bonds in the fragment(s); hence, there is no assumption that a bond in the precursor ion before fragmentation has the same cleavage energy as that “same” bond in one of the fragments.

5 MOLECULE VECTOR ENCODING

Machine learning algorithms, such as ANNs, support vector machines, and K-means clustering, take element vectors as input representations. An ANN was developed for MetISIS to predict rates at which different molecular bonds cleave in mass spectrometer collision induced dissociation (CID) mechanisms. A vector representation for molecular bonds was thus required, and further, this representation had to include information about that bond in the context of the atoms and other bonds in the molecule.

A molecule encoding scheme was developed that is based on the covalent bonds and atoms of molecules. It is assumed that the atom and covalent bond information is sufficient as input for a prediction model because two molecules that have the same atom/bond configurations in small neighborhoods will likely have the same relative 3-D coordinates for the atoms in those neighborhoods. A prediction model that learns from atom and covalent bond information from one set of molecules will be able to predict the bond cleave rates for other molecules.

This view of a molecule as atoms and covalent bonds has long been seen as a connected undirected graph with bonds and atoms as edges and vertices, respectively. Leonhard Euler published the first paper on graph theory in 1736, describing a problem related to crossing the seven bridges of Königsberg [Euler, 1736].

The term graph was first used by James Joseph Sylvester in linking molecules to algebra [Sylvester, 1878], but preceding him, Scottish chemists Archibald Scott Couper in1858 and Crum Brown in 1864 had presented graphs and trees showing molecules as atoms and valences [Couper, 1858; Brown, 1864]. The term *chemical structure* was put forward in these same years by the Russian chemist Butlerov [Butlerov, 1861]. Thus, graphs have from the beginning been closely related to representations of molecules.

In the graphical representation of molecules used in this thesis, the undirected graph G(V,E) has a set V of organic atom types from {C,H,O,N,S,P}, where the letters are standard identifications for carbon, hydrogen, oxygen, nitrogen, sulfur, and phosphorus, respectively, and a set E of pairs of atoms in V defining bonds. The bonds are further labeled as single, double, or triple bonds (bond orders). A graph is represented by an adjacency matrix A with entries A(*i*,*j*) where *i* and *j* are atoms in V and the entries A(*i*,*j*) are bond orders in {0,1,2,3}; A(*i*,*j*) = 0 says the atom pair has no covalent bond.

To make an ANN input vector from a molecule, its graph is first transformed into two rooted trees using queues for breadth-first traversals over the graph. Cycles are not allowed in trees—these are broken by keeping a list of visited vertices and allowing only one visit per vertex.

Two atoms defining a bond that may be cleaved in the CID process are selected as root vertices. The vector used as input for machine learning to predict bond cleavage propensities is thus made from two vectors, one from each rooted tree, which embody the information about the neighborhood of a specific bond. It is the cleavage rate for that bond that is to be predicted. Another path may exist between the two atoms that define the bond. To avoid putting the same atom into both trees, the two trees are generated together in an alternating deepening traversal of the trees and the use of a common list of visited atoms. Thus, to make an ANN input vector, two trees are always generated, i.e., an input vector to the ANN consists of two vectors, one from each tree. Where the discussion of encoding scheme that follows refers to one rooted tree and its vector, this also applies to the other rooted tree and its vector for the other atom defining a bond.

The path through the tree from the root to a given atom, considering every atom and bond in the path, is calculated to a vector index, i.e., an offset to an element in a vector. The vector element at that index is incremented by one (all elements are initially 0). One vector element is incremented for each atom in the neighborhood, and if multiple atoms are indistinguishable in the neighborhood, they increment the same vector element.

Assuming that the root atom is in the set {C,H,O,N,S,P}, then without restrictions, any of these atom types could be bonded to other atoms in {C,H,O,N,S,P} by a bond order in {1,2,3}. Thus, the unrestricted index possibilities are the cardinality of {C,H,O,N,S,P} times the cardinality of {1,2,3} times the cardinality of {C,H,O,N,S,P} or 6 x 3 x 6 = 108. This is the number of unique indices in a neighborhood of two—a neighborhood *n* includes all atoms reachable by traversing *n*-1 bonds from a root atom. For example, a neighborhood of one includes the root atom only and a neighborhood of two has the root atom and other atoms covalently bonded to the root atom.

A neighborhood of size *n* has $6 \text{ x } (3 \text{ x } 6)^{n-1}$ unique indices—a number that grows too large for machine learning with growing *n* if every unique index is an element in a vector. Fortunately, different atom types restrict the number of bonds they establish. Further, nature is kind in that the majority of combinations of atom types and bond orders do not occur. This results in a sparse space that can be compressed without loss of information. This sparseness is discussed below where Table 5.1 shows that with the implemented neighborhood of eight, only a small fraction (0.0000002) of all atom/bond combinations occur in data used to train the ANN.

Figure 5.1 shows partial encoding trees for a neighborhood of size three. The trees have all combinations of atoms and bond orders as indices without regard for what is possible in nature. For example, carbon C_0 is shown to have a triple bond to hydrogen H_{19} despite the fact that bonds to hydrogen are always first order. The trees are thus heavily pruned when applied.

There are six root atoms in Figure 5.1 with indices 0 to 5; the encoding scheme involves six separate trees of which only one is selected to match the root atom on one side of the bond for which the cleavage rate will be predicted. Only the carbon (C) root atom is shown fully expanded to the next atoms through single, double, and triple bonds. From these atoms, only the single bonded carbon is expanded to the third level atoms. The figure also shows the phosphors (P) triple bonds expanded. These phosphors have the highest indices at each depth of the trees; a neighborhood of one has six indices (0 to 5), a neighborhood of two has 108 (0 to 107), and a neighborhood of three has 1944 (0 to 1943) as seen in the figure (and in Table 5.1). The indices are thus sequenced across the six trees. (Alternatively, the six trees could be considered as one tree if a start root was added as a parent to the six first atoms at the left in the figure.)

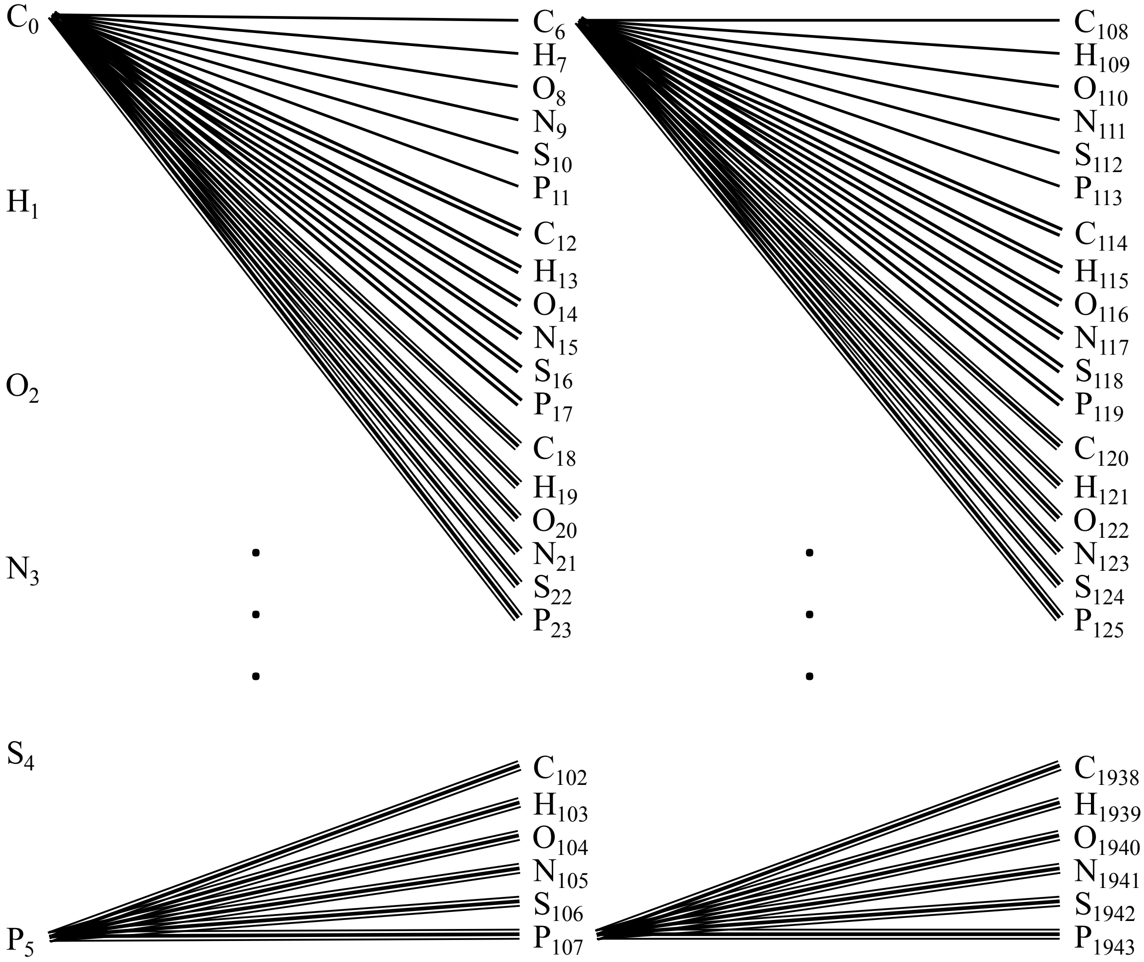


Figure 5.1. Molecule Encoding Trees. The subscripts show the indices.

The offsets into the vectors generated from the encoding trees are the indices associated with the atoms and the elements at these offsets are the number of occurrences of these atoms. The vectors start with all elements zero. As the encoding trees are traversed, the elements in the vectors are incremented at the right offsets. (The vectors are later packed where the offsets no longer correspond to indices.)

Figure 5.2 shows a molecule with atom indices as subscripts for the atoms shown in Figure 5.1. Two atoms have no indices: the hydrogen H bound to the oxygen would have an index but was left out only because it was not shown in Figure 5.1, the hydrogen bound to carbon C_{114} is outside of a neighborhood of three that is shown in Figure 5.1 and would not be encoded in the vector.

The cleavage rate for a bond between a C and the O is to be predicted; hence vectors are made with C and O as roots. For the vector with root C, the vector elements at 0, 6, 7, 109, and 114 are incremented by one for each occurrence of that index: this vector's element at index 6 will have a value of 2, and the element at index 109 will have a value of 3, and elements at 0, 7, and 114 will have ones.

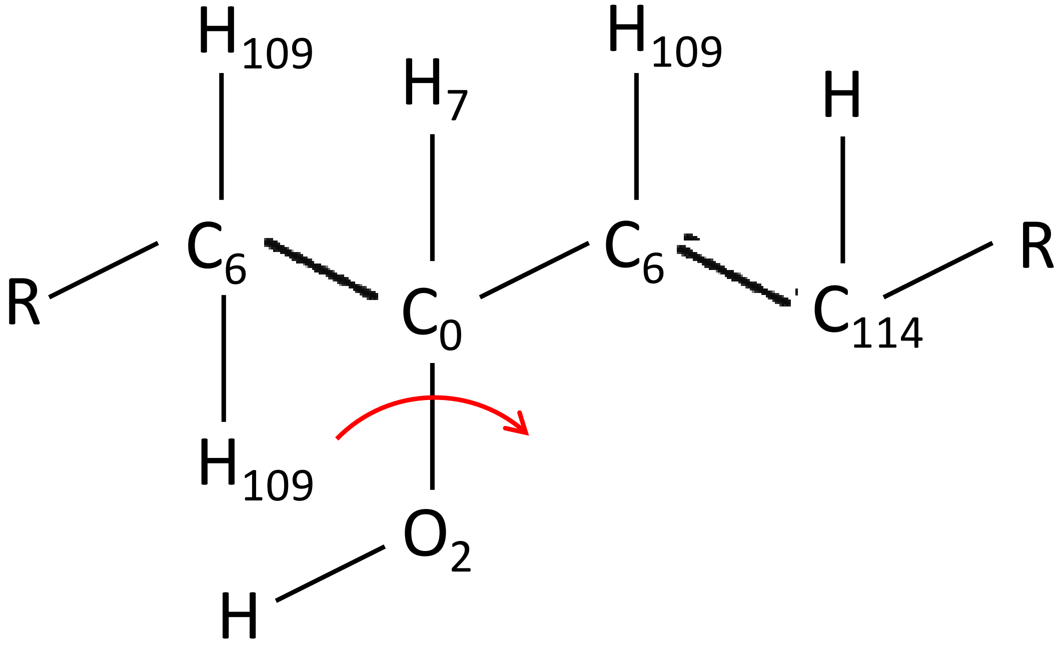


Figure 5.2. Molecule with Atom Indices.

Table 5.1 shows that, as the radii of neighborhoods increase to the implemented radius of eight, the maximum possible indices grow exponentially, but because only a small fraction of the indices are observed, a neighborhood vector only needs these indices. These are defined here as packed indices, as in packed arrays, where indices are limited to only those of interest. The packed indices are defined from a training set of 22 lipids (Table 9.1), where every atom in a lipid in turn was rooted, and the indices were calculated for other atoms in the neighborhoods.

Table 5.1. Radius effect on vector lengths.

Radius	Max Indices	Packed Indices
1	6	5
2	108	17
3	1,944	42
4	34,992	81
5	629,856	145
6	11,337,408	250
7	204,073,344	407
8	3,673,320,192	627

Figure 5.3 shows that as more entries from the LIPID MAPS database are encoded at neighborhood sizes one to eight, the number of observed indices increases. These increases will all approach asymptotes where no additional indices are observed when new lipids are added (the steps in the curves are the result of adding “new” atom/bond configurations for the first time). We say asymptotes because with any finite neighborhood size, only a finite number of atom/bond configurations are possible when restricted by the number of covalent bonds that each atom type will form. The asymptote for a neighborhood size is the maximum number of possible atom/bond configurations.

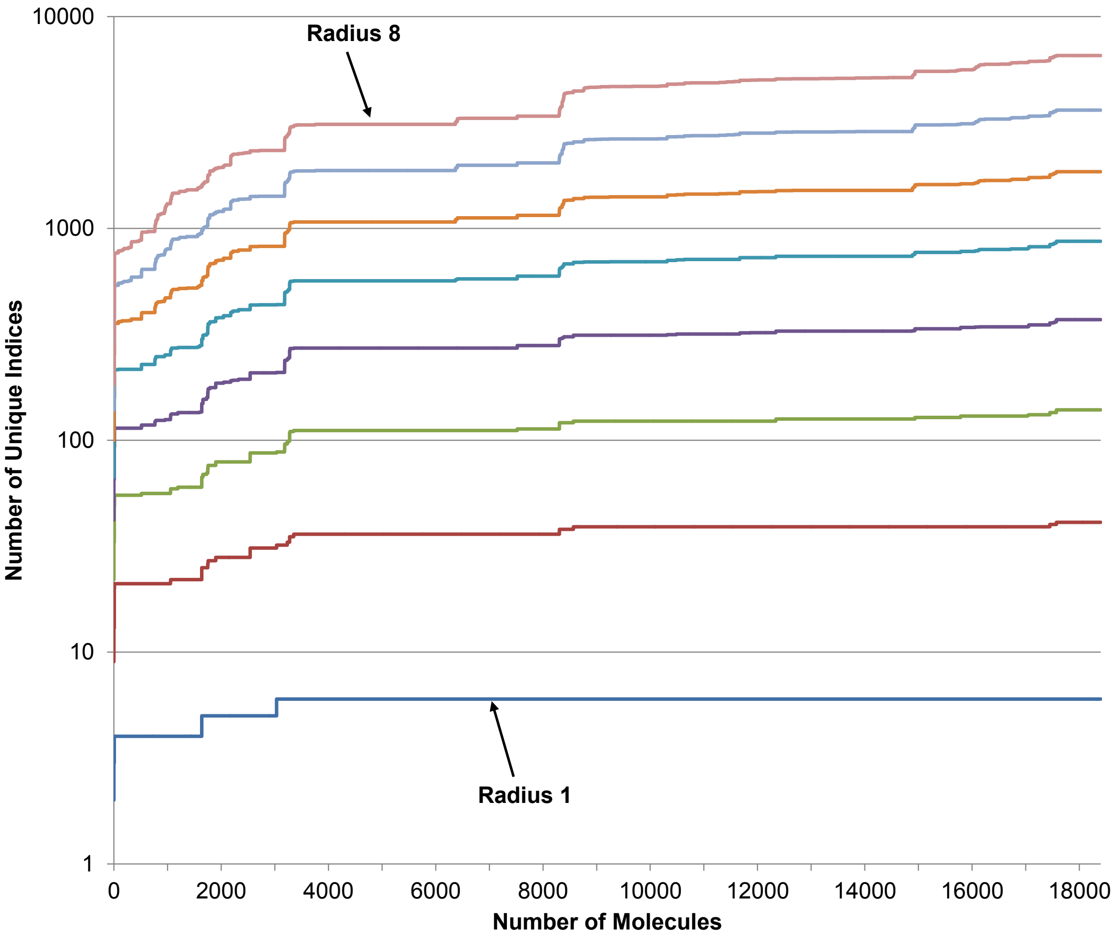


Figure 5.3. Number of Observed Vector Indices. The vertical axis in log scale shows the number of unique atom indices from the encoding algorithm for a growing number of molecules.

The radius 8 neighborhood in Figure 5.3 has 6550 observed indices from a filtered set of 18,399 lipids from the LIPID MAPS database (filter explained below). In the initial training set with 22 lipids, only 627 indices were observed (showed below in Table 6.1).

In a test of the algorithm that involved generation of *in silico* spectra for all lipids in the database, only the 627-element vectors were used. Despite this obvious shortcoming illustrated by Figure 5.3, the rank tests for lipid identifications were remarkably good; possibly because the 627 indices include the most important atoms and bonds in the near neighborhoods needed for predicting bond cleavage temperatures. As more metabolites are added to the training set, the discrepancy of the number of indices used and those in a database should decrease. It is possible that the neighborhood should be increased with more complex metabolites.

It can be shown for the lipids in Figure 5.4 that a neighborhood can be too small to capture important properties in molecules as shown by examples below. The reader is reminded that input vectors are used in the ANN to encode the neighborhoods around bonds in calculations for predicting bond cleavage rates that in turn affect ion intensities in spectra. A bond is defined by its two atoms, each rooted to make two separate vectors as if the bond of interest did not exist. Both vectors are input to the ANN.

Figure 5.4 shows two lipids from the lysophosphocholine (LPC) class. The first is an ester LPC, and the second is an ether LPC. The charged fragment after cleavage of the indicated C-O bond has the N^+ terminal group. Observe that the double bonded oxygen in the neutral fragment is six atoms removed from the bond cleavage that produces a 184 Da ion.

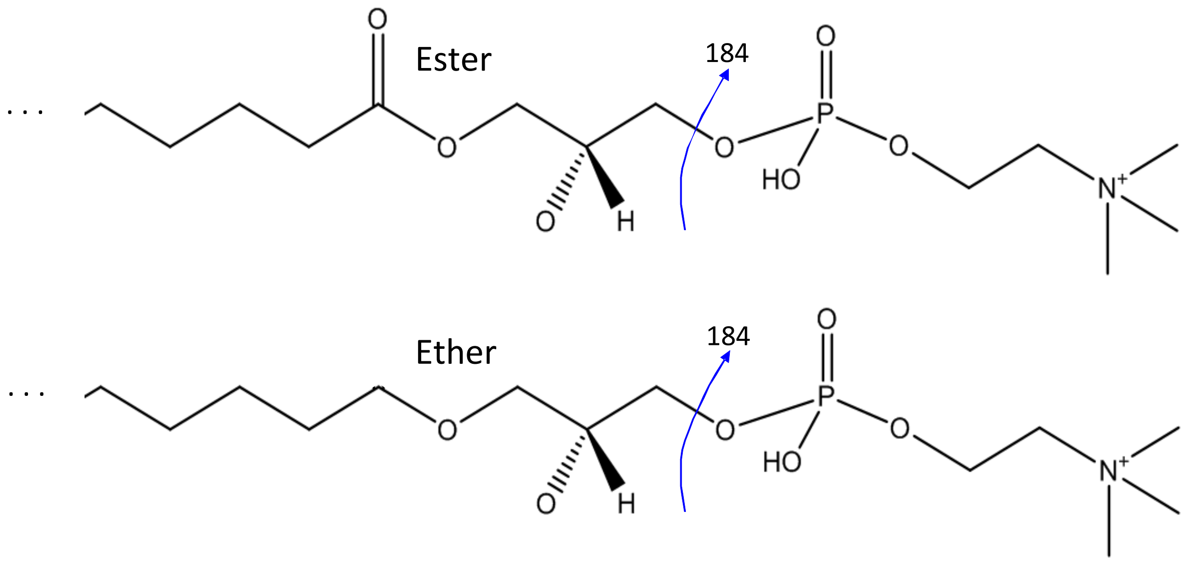
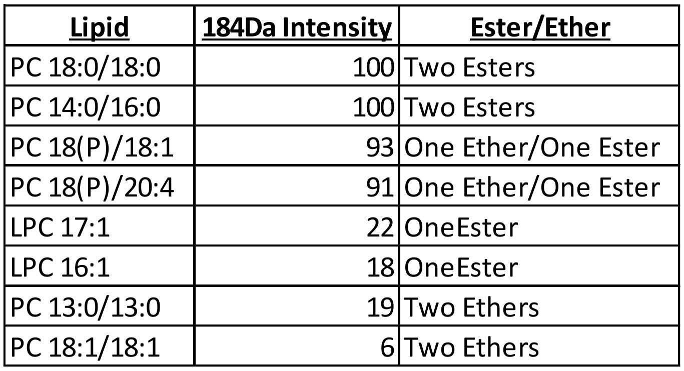


Figure 5.4. Ester and ether LPC lipids.

Table 5.2 shows how the experimental 184 Da ion intensities vary for different ester/ether LPC and PC lipids. As the number of ethers increase and esters decrease, the relative intensity of the 184 Da ion decreases. The double bonded oxygen that distinguishes esters from ethers would not be "visible" to the ANN with a neighborhood radius less than six, and the ANN could not predict the 184 Da intensity difference to separate these two lipid subclasses.

Table 5.2. Relative 184 Da ion intensities in ester and ether lipids.



Lipid	184Da Intensity	Ester/Ether
PC 18:0/18:0	100	Two Esters
PC 14:0/16:0	100	Two Esters
PC 18(P)/18:1	93	One Ether/One Ester
PC 18(P)/20:4	91	One Ether/One Ester
LPC 17:1	22	OneEster
LPC 16:1	18	OneEster
PC 13:0/13:0	19	Two Ethers
PC 18:1/18:1	6	Two Ethers

5.1 Vector Encoding Pseudo Code

Finally, to generate a vector, an element value is 0 if an atom does not have the index in the tree and is 1 if one atom has the index. An element can be greater than 1 if the index occurs more than once. For example, if a carbon as the root is bonded to three other carbons, each by a bond order of one, then these three carbons are indistinguishable and thus have the same index; the vector will have a value 3 at this index.

Equations and pseudo code are provided below, describing the algorithm encoding atoms in a molecule using a breadth-first traversal of a tree structure. The equations show explicitly the calculations of the first for atom positions from a cleaved bond.

$\text{Index}_1 = \text{Atom}_1$

$\text{Index}_2 = \sum_{i=2}^{2} ATypes^{i-1} \times BTypes^{i-2}$

$+ \text{Atom}_1$

$\times ATypes \times BTypes + (\text{Bond}_{1,2} - 1) \times ATypes + \text{Atom}_2$

$\text{Index}_3 = \sum_{i=2}^{3} ATypes^{i-1} \times BTypes^{i-2}$

$+ (\text{Atom}_1$

$\times ATypes \times BTypes + (\text{Bond}_{1,2} - 1) \times ATypes + \text{Atom}_2)$

$\times ATypes \times BTypes + (\text{Bond}_{2,3} - 1) \times ATypes + \text{Atom}_3$

$\text{Index}_4 = \sum_{i=2}^{4} ATypes^{i-1} \times BTypes^{i-2}$

$+ ((\text{Atom}_1$

$\times ATypes \times BTypes + (\text{Bond}_{1,2} - 1) \times ATypes + \text{Atom}_2$

$\times ATypes \times BTypes + (\text{Bond}_{2,3} - 1) \times ATypes + \text{Atom}_3)$

$\times ATypes \times BTypes + (\text{Bond}_{3,4} - 1) \times ATypes + \text{Atom}_4$

. . .,

where *ATypes* is the number of atom types and *BTypes* the number of bond types. $\text{Bond}_{n,n+1}$ is the bond order from atom at tree depth *n* to atom at tree depth *n*+1, and Atom_n is the enumeration of the atom type at tree depth *n*, for example C = 0, H = 1, O = 2, etc.

The encoding algorithm can recursively be defined as

Base case for *n* = 1:

$\text{Base}\,[1] = \text{Atom}_1$

Recursive case *n* > 1:

$\text{Base}\,[n] = \text{Base}[n\text{-}1] \times ATypes \times BTypes + (\text{Bond}_{n-1,n} - 1) \times ATypes + \text{Atom}_n$

$\text{Index}_n = \sum_{i=2}^{n} ATypes^{i-1} \times BTypes^{i-2} + \text{Base}[n]$.

The recursive definition is more attractive for coding because the indices do not need to be recomputed for each atom from the root atom as the encoding progresses if the base values are remembered.

6 FEEDFORWARD ARTIFICIAL NEURAL NETWORK

An artificial neural network (ANN) was configured to use the encoding vectors. ANNs form a class of algorithms that process signals with interconnected neurons similar to the operation of the nervous systems found in living intelligent systems [Priddy, 2005]. ANNs, like many other machine learning algorithms, are divided into supervised and unsupervised learning.

The feedforward ANN is in the supervised category and is probably the best known ANN. It became popular with the invention of the backpropagation algorithm often used in its learning [Werbos, 1974; Parker, 1982; LeCun, 1985; Rumelhart, 1986; Werbos, 1994]. These ANNs are known to be universal classifiers in that in theory, they are capable of classifying any problem to any degree of accuracy [Hornik, 1989]. In reality, the availability of data and the search space configuration determine the accuracy of the models. Yet, these ANNs excel at discerning subtle patterns in large multivariate data sets without preconceived assumptions about the data structure that in turn could be incompletely understood and possibly have complex multivariate relationships.

An example of a feedforward ANN structure is given in Figure 6.1. The data in the form of a vector is fed into the ANN from the left and propagated through the layers of processing neurons until a result vector is output on the right. The neurons in one layer are fully connected to the neurons in the next layer through weighted connections. Each neuron is shown as two separate squares to emphasize the two functions performed in a neuron. The first step is to sum the inputs as products of weights and signals for either the inputs shown on the left to the ANN or the outputs from the neurons in the previous layer. The input *I* to neuron *i* is computed as given in Equation (6.1), where the bias is treated as an additional link weight with a signal value of 1.

$$I_i = bias + \sum_j w_j o_j \qquad (6.1)$$

The second step in a neuron is to compute the output O from neuron i by using a transfer function (f in Figure 6.2) like the hyperbolic tangent function or the more commonly used logistic sigmoid function that is shown in Equation (6.2).

$$O_i = f(I_i) = \frac{1}{1+e^{-I_i}} \tag{6.2}$$

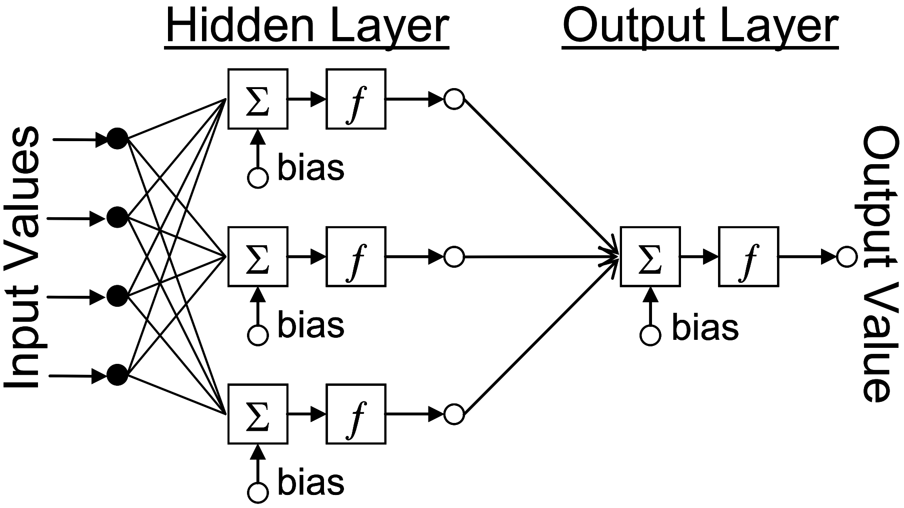


Figure 6.1. An Example of a Feed-forward ANN with One Hidden Layer. The ANN used in this thesis research had one hidden layer with eight hidden nodes.

The Artificial neural networks (ANNs) are thus mathematical models that emulate some of the observed properties of biological nervous systems and as such, draw on the analogies of adaptive biological learning [Priddy, 2005]. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. In Equation (6.1), these are the weights w_j. Learning in feedforward ANNs often occurs by iteratively adjusting the connection weights so that the ANN makes the correct association between input and output of labeled exemplars in a training set.

ANNs belong to the tabula rasa techniques for non-parametric models [Vigneron, et al., 1996]. This essentially means that ANNs when configured have no knowledge—they are born with a "clean slate," and the structure of the data is not specified *a priori* to the ANN model.

Thus, the capability for an ANN to do its task is completely learned from data—here vector encoding of molecules and properties of their bonds.

6.1 Bond Cleavage Energy Prediction

A feedforward ANN was configured to take pairs of the encoding vectors described in Chapter 5 and predict one bond cleavage energy at a time. The ANN was iterated over all bonds in a molecule to find their relative bond cleavage energies.

Table 6.1 shows the structure of input vectors. Two 627-element vectors encode the left and right trees from the two atoms defining a bond. These vectors consist of the packed indices from the above encoding algorithm. "Packed" indicates that all vector elements that do not have indices occurring in any of the lipids in the training set are removed (without packing, the vectors would have ~3.7×10^9 elements).

Six additional metrics were input to the ANN as shown in Table 6.1. The cycle length is that of the shortest path around; for example, a ring, if the bond in focus is in a cycle. The cycle length is 0 if breaking the bond in focus results in two separate fragments.

Table 6.1. Artificial neural network input elements.

Input	No of Inputs
Left tree encoding	627
Right tree encoding	627
Bond order	1
Cycle length	1
Left tree mass	1
Right tree mass	1
Left tree degrees of freedom	1
Right tree degrees of freedom	1
	Total: 1260

6.1.1 Genetic Algorithm

Typically, feedforward ANN weights are learned in a supervised mode using an algorithm like the back-propagation, but as was shown in Figure 4.1, a genetic algorithm (GA) (Goldberg, 1989; Holland, 1975), together with the KMC simulations, trained the ANN because supervised training was not an option. Supervised training requires that the training data has correct examples for the parameter(s) that the ANN is trained to predict. Here the ANN is trained to predict bond cleavage energies, and the true bond cleavage energies are unknown. MetISIS uses KMC simulation and a GA to connect the ANN input vectors to CID spectra that are a function of relative bond energies.

A GA is inspired by natural evolution where individuals in a population evolve to better fit in an environment. The GA is initially configured with a number of individuals initialized with random values or some prior knowledge. Each individual is a candidate solution to a problem. In each generation (iteration) of the GA, the fitness of each individual is assessed by testing it as a solution for the problem. The individuals are sorted on the fitness. Only a number of the most fit individuals go on to the next generation where these individuals produce new individuals through cross-over and mutation operators. The individuals are often either binary- or real-valued sequences. Cross-over is the generation of a new individual by taking some of one parent's, for example, real-valued sequence elements and the other sequence elements from another parent. Mutations are applied with a small probability as small perturbations to the sequence elements of the new individuals from the cross-over operations. The "offspring" receive the parents genes but with an occasional random change in the genes.

Initially, the GA trained the ANN based on experimental CID spectra from 22 lipids. The ANN weights were optimized to better predict relative bond cleavage energies that produced ions and their corresponding intensities in the *in silico* spectra correlated to the experimental data. The GA was configured with 10 individuals, each a set of ANN weights, i.e., each set of weights is a candidate ANN solution. The GA optimized the individuals iteratively with the objective to have the *in silico* spectral ions match those in the experimental spectra using a Pearson R-square correlation. The training algorithm running on a 3 GHz PC was stopped after four months with an R-square of 0.97 against the 22 lipids in the training set.

6.2 Bond Cleavage Prediction Example.

Figure 6.2 shows predicted relative bond cleavage energies for lysophosphatidylcholine 18:0, $[M+H]^+$ 524.5 *m/z*. The figure shows that the hydroxyl bond for a water loss has the lowest energy, 0.182 (3.64 eV) and the head group loss has the second lowest energy, 0.241 (4.82 eV). Cleavage of these bonds generates the two significant peaks in the experimental CID spectrum at 506 and 184 *m/z*, respectively. Observe that the water loss occurs by cleaving off the indicated hydroxyl group and a hydrogen atom from an unspecified location; hence, the predicted mass of resulting fragment ion could be in error by one Da. This type of water-loss reaction, called E1/E2, is further discussed in section 7.3.

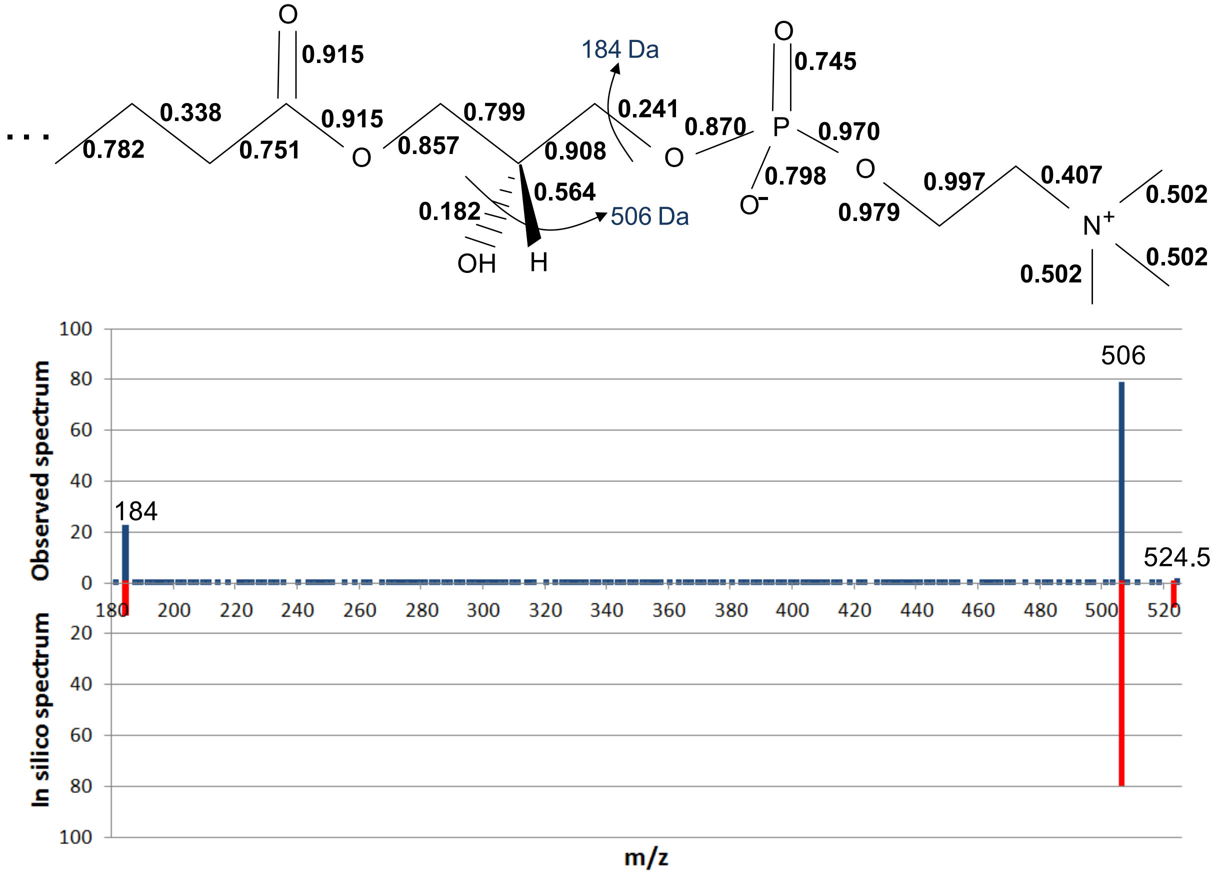


Figure 6.2. Predicted Bond Cleavage Eenergies for a Lipid (lysophosphatidylcholine 18:0, $[M+H]^+$ 524.5 *m/z*) and the Resulting in Silico Spectrum Compared to the Observed Spectrum. The two ions, 184 Da and 506 Da, in the molecule are shown as both observed ions and *in silico* generated ions in the spectrum. The spectra are normalized to 100% total peak intensities.

Although the predicted energies give the correct ions as shown in the figure, inconsistencies can be observed in the labeled bonds; for example, one C-C bond in the fatty acyl has a predicted value of 0.338, which is known to be incorrect both from basic chemistry and the absence of an associated peak in experiment CID spectra. With the addition of more training exemplars and more training of the algorithm, the values should approach correct cleavage energies.

6.3 Overfitting the Artificial Neural Network

Large weight sets in ANNs relative to the number of available training vectors are associated with overfitting, which results in a tendency for ANNs to learn the training data well, but then not to generalize this knowledge to novel data. Overfitting is usually observed by ANNs exhibiting perfect or near perfect tasks in predictions/classifications for each training pair, inputs and outputs in the training set, but when new inputs are applied to the ANN, the outputs can deviate, sometimes extremely, from the expected outputs.

When training ANNs using, for example, the backpropagation algorithm, this overfitting can be measured using a cross-validation set, which is tested in parallel to the training set during learning to recognize if the ANN over-memorizes the training data. Over-memorization or loss of generality in the ANN is observed when the error decreases against the training set while the error against the cross-validation set increases. After the proper amount of learning is completed, accuracy is determined by predictions for a test set of labeled exemplars never used in the training process. In addition of stopping learning early, two methods commonly used to mitigate overfitting in a feedforward ANN are (1) to reduce the size of the hidden layer when possible and (2) adding more training data or reconfiguring input/output vectors.

A cross-validation data set was not used in training the ANN in MetISIS for two reasons: the important input values describing atoms and bonds in molecules are discrete and finite, making the ANN essentially act as a very large table look up—it doesn't have the complexity of continuous-valued functions that are more difficult to define. Second, the most important atoms and bonds related to predicting the energy required to cleave a bond are those in close proximity to the bond in focus. Already a small set of molecules, like used here, have a complete set of atom/bond configurations close to the bond predicted.

The ANN in MetISIS does not appear to overfit the training data as observed in testing. We conclude this from that the error against the testing set is about the same as the error against the training set. An overfit ANN typically produces small errors against a training set and a significantly larger error against a testing set. The outputs predicted by the ANN are well-behaved for all test molecules.

Unexpected outputs suggesting overfitting can occur when vectors input to an ANN in testing or in usage significantly deviate from the input vectors used in training. The scheme developed for encoding atoms and bonds in molecules to input vectors calculates specific vector indices. These vectors are packed in MetISIS to only indices that are observed in the training set. When MetISIS is tested or put to use, novel molecules are only encoded to the packed vector indices. Thus, the ANN is not given vector elements that were not in the packed training vectors. If the novel molecules result in calculated atom/bond indices that are not in the packed vectors, then these indices are discarded.

Finally, we recognize that overfitting need to be revisited in a future investigation when more data is available. This may involve using a cross-validation data set during training. Such data set would be beneficial in finding the optimal number of hidden nodes in the ANN—an exercise not yet performed as this preliminary research used a too small data set to give us meaningful results.

7 CHEMICAL REACTIONS

Computer modeling of chemical reactions was implemented as rules in the late 1960s with DENDRAL that incorporated artificial intelligence and machine learning to elucidate chemical structures from mass spectrometry (Lederberg, 1987). The 1960s and 1970s saw several algorithm developments to model chemical reactions (Armitage, 1967; Willett, 1970; Bersohn, 1976). Bersohn and Esack (Berssohn, 1976) programmed more than 200 reactions based on patterns of chemical reactions in functional groups. The method implemented in MetISIS was first suggested by Vleduts in 1963 in a seminal article in chemical reaction modeling (Vleduts, 1963). Vleduts argued that chemical reactions in computer models should be based at an atom/bond level. Recently Pennerath and coworkers developed an algorithm that ranks formability of bonds in chemical reactions (Pennerath, et al., 2010). The algorithm uses the graph-mining algorithm GemsBond to mine chemical reaction databases for information, and then applies machine learning to rank order pairs of bonds as to which bond is more likely to form.

7.1 Chemical Reaction Prediction

The algorithm in MetISIS predicts what covalent bond to cleave, but that cleavage in turn will cause other covalent bonds to change bond order through chemical reactions that return molecules to stable states. These reactions are usually grouped into additions, eliminations, substitutions, and rearrangements. Incorporating these reaction mechanisms into MetISIS ensures that the simulated fragments from a bond cleavage are consistent with basic chemistry. Two advantages result from correctly accounting for the chemical reactions after bond cleavages:

(1) exact m/z values for the fragments, and (2) exact determination of fragment charges from chemical bonding.

A relatively simple reaction, the E1/E2 elimination occurs frequently as a water loss from lipids in CID. Figure 7.1 (a) shows the three steps for loss of a water molecule. The C-O bond at (1) cleaves, but this would only make the O-H the leaving group. To make water loss, another bond must break that release an H atom to combine with the O-H leaving group from the first bond cleavage. Processed (2) and (3) in Figure 7.1 illustrate one possibility for the secondary fragmentation reaction. To make the valence electrons balance, the bond at (3) folds to make a double bond between the two carbons. Figure 7.1 (b) shows the two separated molecules.

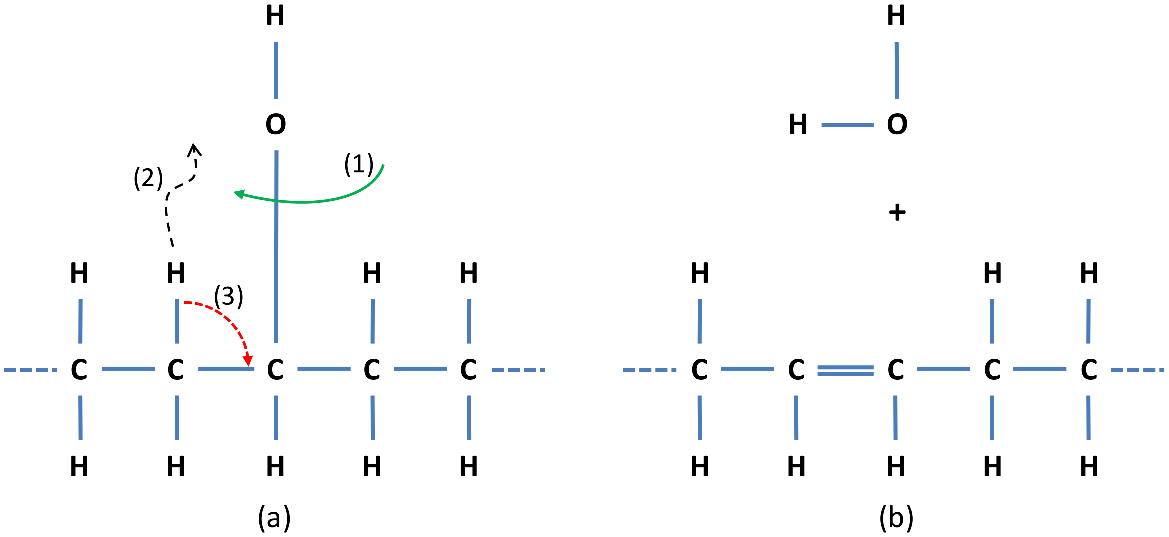


Figure 7.1. Water Loss by E1/E2 Chemical Elimination Reaction.

If only the cleavage at (1) in Figure 7.1(a) occurred without steps (2) and (3) the leaving group would have a mass of 17 Da, while the true mass leaving is 18 Da of a H_2O molecule. A mass error of 1Da occurs for each of the two fragments after the water loss so that the mass balance conservation law holds for both the correct water loss scheme and the one implemented in MetISIS.

Chemical reactions are implemented in MetISIS as templates consisting of pairs of graph adjacency (connectivity) matrices: one matrix to match to the molecules in an atom/bond neighborhood around the bond that will be cleaved and one matrix that shows the changes in the bonds that need to be applied to the molecule to simulate the complete chemical reaction. The mass balance conservation law states that the atoms between the two matrices do not change—only bonds are reconfigured.

Figure 7.2 shows the molecular structures for the water loss template (the same reaction as described in Figure 7.1). Only atoms that are essential to the reaction (template) and essential to distinguishing one reaction (template) from other reaction templates are included.

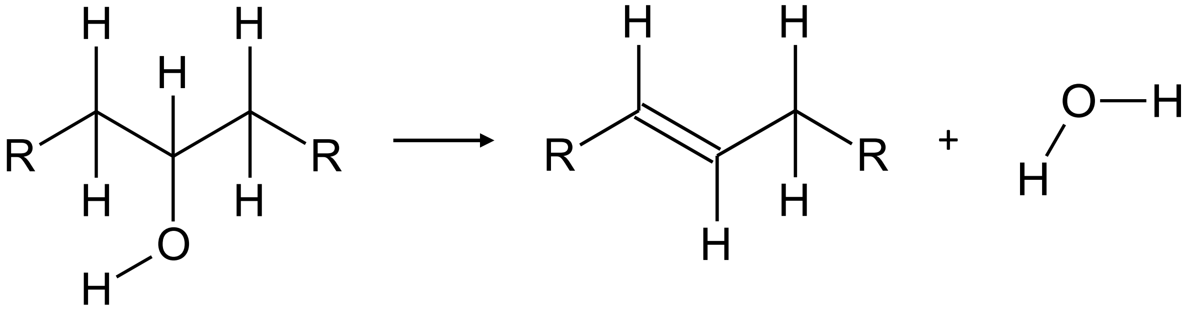


Figure 7.2. E1/E2 Water Loss Reaction Structures

Each reaction template includes a probability of it being selected from all templates that are applicable to a specific bond cleavage. In the Figure 7.2 water loss example, the reaction took a second hydrogen atom from the left side and needed a reaction template to describe this rearrangement. Another template describes how the water loss could have taken the hydrogen from the right side instead since these hydrogen atoms are indistinguishable. Each of these two templates has a probability of 0.5 of being selected. A random number is generated to proportionally to the probabilities select one applicable reaction template.

The molecule encoding scheme described in Chapter 5 is applied to both templates and the molecules to which they are matched, thereby making the matching invariant of atom numbering.

An initial set of 49 reaction templates has been developed to cover the known fragmentation pathways for 97 lipid metabolites. As the capacity of MetISIS is increased to identify other metabolites, new templates will be developed and added to MetISIS. The initial effort in defining these templates has been to cover the most significant ions observed in the 97 metabolites. It is expected that relatively few templates—possibly a few hundred—will cover the most prevalent chemical reactions for a large set of metabolites.

8 CHARGE PREDICTION

In mass spectrometry, only ions are detectable. Most metabolites have charge + or -1 because of their relatively small sizes; larger molecules tend to have more possible atoms that can easily take charges.

Changing a neutral molecule into an ion is the first step in an ion trap mass spectrometer modeled in this work. The lipid data described here was generated with an electrospray ionization (ESI) mechanism configured to give molecules positive charges. The lipids were observed as either protonated ions [M+H]+ or adduct ions [M+NH$_4$]+.

The importance of the charges in molecules can be described with two similar lipids: Figure 8.1 (a) shows lysophosphocholine (PC) lipid labeled with the 184Da phosphocholine ion that is significant in identifying this lipid class. Replacing the ammonium in the head group with an amine changes this lipid to a lysoglycerophosphoethanolamines (PE) as in Figure 8.1 (b) that has neutral head group fragment, the 141Da phosphoethanolamine, in positive MS. The figure shows that cleaving the carbon oxygen bond next to the phosphate group produces an 184.08Da ion to the right for the PC lipid while in the PE lipid the same cleavage produces a 341.34Da ion to the left.

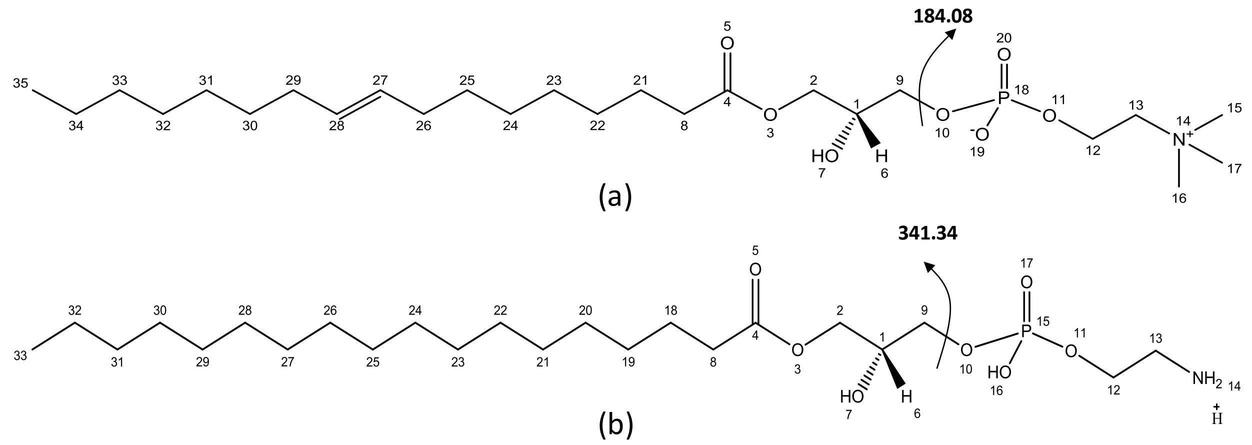


Figure 8.1. (a) Lysophosphocholine. (b) Lysoglycerophosphoethanolamine.

The ionization process for the PC lipids in (a) starts, at the given pH, with the molecule being a zwitterion: a neutral molecule, with the oxygen O, atom number 19, having a negative charge and nitrogen N, atom number14, having a positive charge. The ESI protonates the oxygen (atom 19), leaving the positive nitrogen (atom 14) the only charged atom. This makes the molecule an ion with charge +1.

The ionization process is different for the PE in (b). Here the oxygen, atom 16, already has a hydrogen atom at the given pH. The nitrogen, atom 14, in the amine (NH_2) is not charged. ESI protonates the NH_2 group in the PE, giving the molecule a +1 charge.

The prediction of protonation/deprotonation in charging a molecules involves finding the atom most likely to gain a proton in positive charging or finding the atom most likely to lose a proton in negative charging. Making such model is a large undertaking involving extensive calculations/modeling of molecules using methods drawn from electronic structure theory or molecular dynamics. This is beyond the scope of this thesis. Instead a simpler algorithm was used in predicting which molecule fragments are neutral and which are charged.

8.1 ANN Ion Prediction

A second artificial neural network (ANN) in MetISIS predicts which fragment carries the charge when a molecule fragments. This ANN uses the same input vectors as the ANN predicting bond cleavage temperatures and was initially also trained by the genetic algorithm, but having two ANNs doubled the number of weights to be optimized which in turn slowed the training significantly. The training of this ANN was therefore taken offline using the backpropagation algorithm [Werbos, 1974; Parker, 1982; LeCun, 1985; Rumelhart, 1986; Werbos, 1994]. This ANN was trained to predict on which sides of a bond cleavage the ions and neutrals would appear for all possible first fragmentations of the intact precursor lipids. The fragments and true charges were defined from the experimental spectra used in training.

The ANN predicting bond's cleavage temperatures was first used to predict every bond's propensity to cleave. A lower predicted temperature suggests a higher propensity for a bond to cleave. Selecting only the fragment pairs resulting from the most likely bond cleavages, these fragments were matched to the corresponding m/z ions in their observed spectra. If a fragment's *m*/*z* matched a peak in the experimental CID spectrum, the fragment was assumed to have a charge; if not, the fragment was the neutral. Using these vector encoded fragments pairs as inputs and the ion/neutral labels as outputs, the ANN was trained to predict the which fragment had the charge. Additional details describing this prediction model is given in (Schrom, et al., 2011).

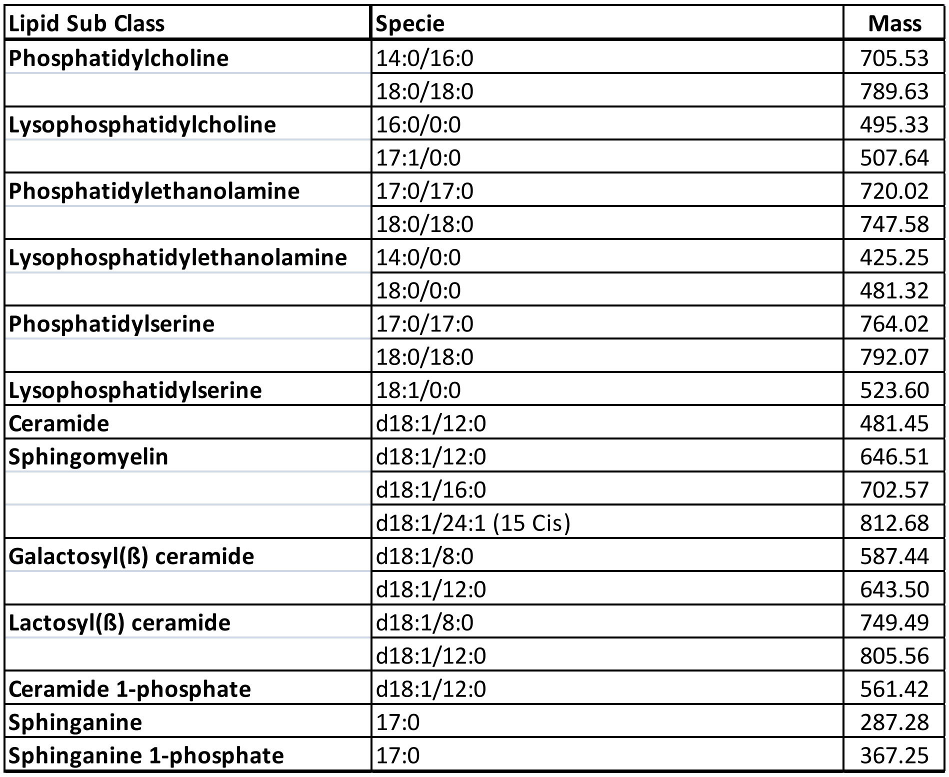
9 LIPID STANDARDS TESTS

The MetISIS algorithm is generic to all organic molecules containing atoms from the set {C,H,O,N,S,P}. To predict the propensities for bonds to cleave, it needs to be trained with pairs of know molecules and their observed spectra. In the typical use of mass spectrometry in metabolimics, samples are analyzed that have a large number of different metabolites and other molecules. There are two reasons why data from these samples are difficult to use in training MetISIS: first, there is much uncertainty in metabolite identifications from these samples (the absence of good tools is of course why MetISIS is being developed). Second, a mass spectrometer for the MS/MS step will select all precursor ions that have a certain m/z within certain margins. All species with similar m/z are thus fragmented together, making spectra often composites of many species.

To avoid contaminated spectra with uncertain identifications of precursor ions, MetISIS was trained with metabolites standards. A "standard" is a sample containing only one known species. A small set of 22 unique lipid standards were analyzed in an LTQ ion trap mass spectrometer to generate training data. When MetISIS was fully trained on these lipids, a second set of 46 unique lipid standards was purchased and analyzed in the same instrument to serve as a test set.

The 22 training and 46 test lipid standards were purchased from Avanti Polar Lipids Inc. (Alabaster, AL). A detailed list of the training and test lipid standards is provided in Table 9.1 and Table 9.2, respectively. A working standard of 1-10 pmol/µL was prepared for each lipid standard in chloroform/methanol/300 mM aqueous ammonium acetate (30/65.5/3.5).

Table 9.1. Training lipids.



Lipid Sub Class	Specie	Mass
Phosphatidylcholine	14:0/16:0	705.53
	18:0/18:0	789.63
Lysophosphatidylcholine	16:0/0:0	495.33
	17:1/0:0	507.64
Phosphatidylethanolamine	17:0/17:0	720.02
	18:0/18:0	747.58
Lysophosphatidylethanolamine	14:0/0:0	425.25
	18:0/0:0	481.32
Phosphatidylserine	17:0/17:0	764.02
	18:0/18:0	792.07
Lysophosphatidylserine	18:1/0:0	523.60
Ceramide	d18:1/12:0	481.45
Sphingomyelin	d18:1/12:0	646.51
	d18:1/16:0	702.57
	d18:1/24:1 (15 Cis)	812.68
Galactosyl(ß) ceramide	d18:1/8:0	587.44
	d18:1/12:0	643.50
Lactosyl(ß) ceramide	d18:1/8:0	749.49
	d18:1/12:0	805.56
Ceramide 1-phosphate	d18:1/12:0	561.42
Sphinganine	17:0	287.28
Sphinganine 1-phosphate	17:0	367.25

Table 9.2. Testing lipids.

Lipid Sub Class	Specie	Mass
Phosphatidylcholine	14:0/14:0	677.50
	16:1/16:1	729.53
	16:0/16:0	733.56
	17:0/17:0	761.59
	18:3/18:3 (Cis)	777.53
	18:2 (9, 12 Cis)/18:2 (9, 12 Cis)	781.56
	18:1 (9 Cis)/18:1 (9 Cis)	785.59
	20:4 (5, 8, 11, 14 Cis)/20:4 (5, 8, 11, 14 Cis)	829.56
	20:1 (11 Cis)/20:1 (11 Cis)	841.66
	23:0/23:0	929.78
Lysophosphatidylcholine	14:0/0:0	467.30
	15:0/0:0	481.32
	16:0/0:0	495.33
	17:0/0:0	509.35
	18:1/0:0	521.35
	18:0/0:0	523.36
Phosphatidylethanolamine	12:0/12:0	579.39
	15:0/15:0	663.48
	16:1 (9 Cis)/16:1 (9 Cis)	687.48
	16:0/18:1 (9 Cis)	717.53
	18:0/18:1 (9 Cis)	745.56
Lysophosphatidylethanolamine	14:0/0:0	425.25
	16:0/0:0	453.29
	18:1 (9 Cis)/0:0	479.30
Phosphatidylserine	12:0/12:0	623.38
	14:0/14:0	679.44
	16:0/18:2 (9, 12 Cis)	759.47
	18:0/18:1 (9 Cis)	789.55
	18:0/18:2 (9, 12 Cis)	787.54
Lysophosphatidylserine	16:0/0:0	497.28
	18:0/0:0	525.31
Ceramide	d18:1/18:0	565.54
	d18:1/24:0	649.64
	d18:1/17:0	551.53
	d18:1/20:0	593.57
	d18:1/22:0	621.61
Sphingomyelin	d18:1/17:0	716.58
	d18:1/18:1 (9 Cis)	728.58
Galactosyl(ß) ceramide	d18:1/16:0	699.56
	d18:1/24:1 (15 Cis)	809.67
Lactosyl(ß) ceramide	d18:1/16:0	861.62
	d18:1/24:0	973.74
Ceramide 1-phosphate	d18:1/8:0	505.35
	d18:1/16:0	617.48
	d18:1/18:1 (9 Cis)	643.49
	d18:1/24:0	729.60

9.1 Mass Spectrometric Analysis of Lipids

Mass spectrometric analysis was performed using a linear ion trap (LTQ; Thermo Scientific, San Jose, CA) operated in positive ion mode. Samples were delivered to the mass spectrometer through a 100 cm capillary of 150 µm internal diameter and 360 µm outer diameter at a flow rate of 0.5 µL/min. The ion spray voltage, capillary voltage and capillary temperature were set to 2.2 kV, 49 V and 200 °C, respectively. Full scan spectra of each lipid standard were first obtained to determine the *m/z* of the precursor ion. The parameters for CID were set as follows: isolation width (*m/z*) of 3 Da, normalized collision energy of 30%, activation Q of 0.18 and activation time of 30 msec.

9.2 Lipid Database

This research used LIPID Metabolites And Pathways Strategy (LIPID MAPS) structure database (LMSD) dated March 24, 2010 and available from www.lipidmaps.org. The database contains molecular data for 22,396 lipids.

9.3 *In Silico* Spectral Library

To test MetISIS, the contents of the LIPID MAPS database was used, and, to ensure correct hits were possible, the 45 test lipids (Table 9.2) were added to this database. Also, from the more than 22k lipids in LIPID MAPS, only those lipids with atoms in {C,H,O,N,S,P} and with masses ≤1100 Da were used (1100 Da is the upper bound of the lipids of interest in our research). These 18,399 filtered lipids were processed, with 300 replicates of each, by MetISIS to produce *in silico* spectra for a spectral library, the contents of which will be compared to the experimental spectra of the 45 test lipids. The collision energy in MetISIS was set at 30% for all lipids.

Generating *in silico* spectra requires computationally expensive MC simulations, about one minute per spectrum. We partition spectra onto multiple threads in this task.

9.4 Database Screening

A rank test was performed with 45 lipids not included in the training set but which were selected from the same lipid classes/subclasses as those in the training set. The test of each lipid proceeded by first finding the subset of lipids in the *in silico* spectral library that matched the experimental mass of the precursor ion within ±500 ppm. Next, the *in silico* spectra for these subset lipids were compared to the experimental test spectrum and Pearson R-square scores were generated. The subset lipids were sorted in descending order based on these scores.

Table 9.3 shows the results from screening the spectral library with test lipid PS (18:0/18:1) observed at *m/z* 790.5 ($[M+H]^+$). The rank list shows that the first four hits, true positives, have high R-squares, 0.993 to 0.996. Starting with the fifth hit, the R-squares fall rapidly, 0.117 to 0.000, and corresponds to false positives.

Table 9.3. Rank list for test lipid PS (18:0/18:1).

R-square	Molecule	Mass	Configuration	Formula	TruePos
0.996	LMGP03010025	789.552	PS(18:0/18:1(9Z))	C42H80NO10P	Y
0.994	LMGP03010019	789.552	PS(18:1(9Z)/18:0)[U]	C42H80NO10P	Y
0.993	LMGP03010034	789.552	PS(18:1(9Z)/18:0)	C42H80NO10P	Y
0.993	LMGP03010012	789.552	PS(18:0/18:1(9Z))[U]	C42H80NO10P	Y
0.170	LMGP01011144	789.625	PC(24:0/12:0)[U]	C44H88NO8P	N
0.119	LMGP01010616	789.625	PC(16:0/20:0)	C44H88NO8P	N
0.087	LMGP01010468	789.625	PC(13:0/23:0)	C44H88NO8P	N
0.084	LMGP01010549	789.625	PC(15:0/21:0)	C44H88NO8P	N
0.081	LMGP01010511	789.625	PC(14:0/22:0)	C44H88NO8P	N
0.080	LMGP01010617	789.625	PC(16:0/20:0)[U]	C44H88NO8P	N
0.068	LMGP01010422	789.625	PC(11:0/25:0)	C44H88NO8P	N
0.061	LMGP01011085	789.625	PC(22:0/14:0)	C44H88NO8P	N
0.054	LMGP01010449	789.625	PC(12:0/24:0)	C44H88NO8P	N
0.053	LMGP01010748	789.625	PC(18:0/18:0)[U]	C44H88NO8P	N
0.053	LMGP01010974	789.625	PC(19:0/17:0)[U]	C44H88NO8P	N
0.044	LMGP01010402	789.625	PC(10:0/26:0)[U]	C44H88NO8P	N
0.043	LMGP01010713	789.625	PC(17:0/19:0)[U]	C44H88NO8P	N
0.042	LMGP01010747	789.625	PC(18:0/18:0)[S]	C44H88NO8P	N
0.039	LMGP01011066	789.625	PC(21:0/15:0)[U]	C44H88NO8P	N
0.036	LMGP01010450	789.625	PC(12:0/24:0)[U]	C44H88NO8P	N
0.034	LMGP01011002	789.625	PC(20:0/16:0)	C44H88NO8P	N
0.033	LMGP01011168	789.625	PC(25:0/11:0)[U]	C44H88NO8P	N
0.025	LMGP01011125	789.625	PC(23:0/13:0)[U]	C44H88NO8P	N
0.016	LMGP01010006	789.625	PC(18:0/18:0)	C44H88NO8P	N
0.008	LMGP01020059	789.661	PC(O-16:0/21:0)[U]	C45H92NO7P	N
0.007	LMGP01020080	789.661	PC(O-17:0/20:0)	C45H92NO7P	N
0.000	LMGP02010071	789.625	PE(19:0/20:0)[U]	C44H88NO8P	N
0.000	LMGP02010070	789.625	PE(18:0/21:0)[U]	C44H88NO8P	N
0.000	LMGP02010256	789.625	PE(16:0/23:0)[U]	C44H88NO8P	N
0.000	LMGP02010214	789.625	PE(22:0/17:0)[U]	C44H88NO8P	N
0.000	LMGP02020017	789.661	PE(O-18:0/22:0)	C45H92NO7P	N
0.000	LMGP02010209	789.625	PE(21:0/18:0)[U]	C44H88NO8P	N
0.000	LMGP02020016	789.661	PE(O-18:0/22:0)[U]	C45H92NO7P	N
0.000	LMGP02010255	789.625	PE(17:0/22:0)[U]	C44H88NO8P	N

The 45 test lipids within ± 500 ppm resulted in a total of 808 candidates against the *in-silico* spectral library (a candidate is a hit against one molecule in the database that has a mass within a mass margin of the observed precursor mass).

Figure 9.1 shows the distributions of the true and the false positive R-square scores for these hits. Clearly, most true positives have high scores and false positive have low scores. Observe that hits were counted as true positives if they only varied in chirality, locations of double bonds in the fatty acids, or by the distribution of the correct total number of carbons over two fatty acids—information that cannot be determined by simple CID MS/MS analyses in positive mode.

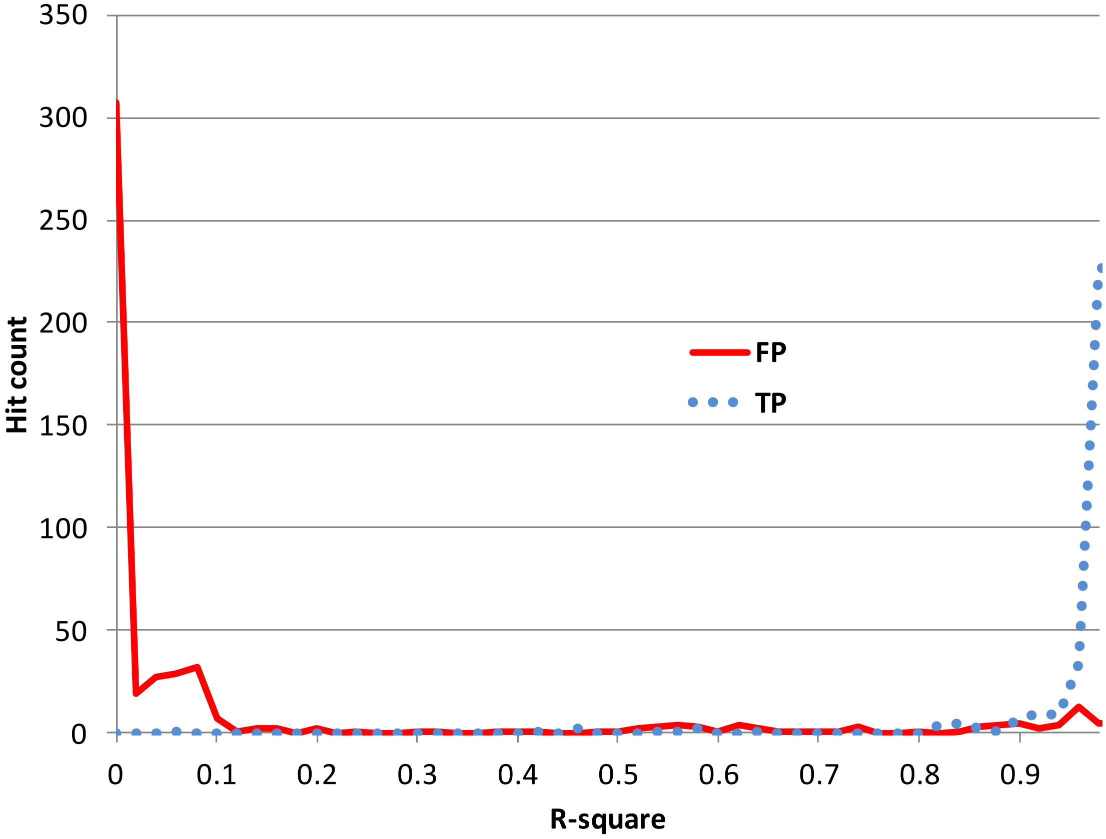


Figure 9.1. Distributions of True and False Positives.

The test ranked 40 of the 45 test lipids at the top position and five at the second position. These five test lipids were ester lipids which each had one ether lipid from the same class ranked above it. For example, the ester test lipid PC (18:0/0:0), mass 523.3638 Da, ranked second after a false positive identification of ether PC (O-19:0/0:0), mass 523.4002 Da. These two masses differ by 70 ppm, which is not enough for the linear ion trap to separate. The confusion between ester and ether lipids is a result of the training set not having ether lipids.

Examining the experimental spectra of LPC and PC ester and ether lipids, it appears that the algorithm should learn to separate these subclasses by the relative intensities of the 184 Da ion from the head group (discussed in the Supplement). Incorporating the five new ether lipids into the training required the vector lengths to increase from 627 to 738 observed indices—a significant increase. This means that, having only used the 627 indices when the ether lipid *in silico* spectra were generated for the library, many of the atom types and bond orders did not contribute to prediction of correct ion intensities.

10 COMPARISON: METISIS TO METFRAG

We compared the performance of MetISIS in ranking candidate spectra to that of MetFrag (Wolf, 2010; Hildebrandt, 2011). The LipidMaps database is too large to load for the online MetFrag application. This was solved by giving MetFrag a "database" of only the candidates from LipidMaps to rank that were within ± 500 ppm mass margins of the precursor adduct mass of the test lipid. Our initial test of MetISIS found an average of 18 candidates in LipidMaps for each test lipid. For larger lipids, the online MetFrag timed out the user before all candidates were processed. We solved this by reducing the candidates to an average of 8 candidates for each test lipid, or a total of 360 candidates.

Both MetISIS and MetFrag compare observed spectra against candidate spectra and return rank lists with scores in the 0 to 1 range, representing the similarities between the observed spectra and their respective candidate spectra. MetFrag normalizes each rank list such that the top ranked candidate receives a score of 1.0. MetISIS provides the users the actual scores (R-squares) from comparing the observed spectrum against each candidate *in silico* spectrum. The actual scores give the users the option to not trust any ranked candidate if its score is low because the observed spectrum is poor due to, for example, the spectrum containing more than one species or the true candidate is not present in the database. This information is lost to the users if the scores are normalized.

Using the 45 test lipids in Table 9.2, MetISIS ranked 40 of the correct candidates at the top and the remaining 5 in second positions. Table 10.1 shows that MetFrag ranked only 21 lipids at the top and 8 at the second position. The remaining 16 lipids ranked in the 3rd and 4th positions. MetISIS thus performed significantly better at ranking the candidate spectra as correct identifications.

Table 10.1. Top ranks of correct identifications of 45 test lipids for MetISIS and MetFrag.

MetISIS		**MetFrag**	
Rank	**Count**	**Rank**	**Count**
1	40	1	21
2	5	2	8
N/A	N/A	3	10
N/A	N/A	4	6

Figure 10.1 shows the distributions of the scores MetISIS and MetFrag assign to the 360 candidates for the 45 lipids. For each algorithm the scores are divided into the true (TP) and false (TN) candidates. The figure shows that MetISIS assigns significantly more very low scores to false candidates and assigns significantly fewer high scores to false candidates compared to MetFrag. MetISIS and MetFrag give similar high scores to true candidates (the curves overlap on the right side of the graph), but a third curve overlapping the first two (on the right side) shows that MetFrag also assigning a significant number of high scores to false candidates.

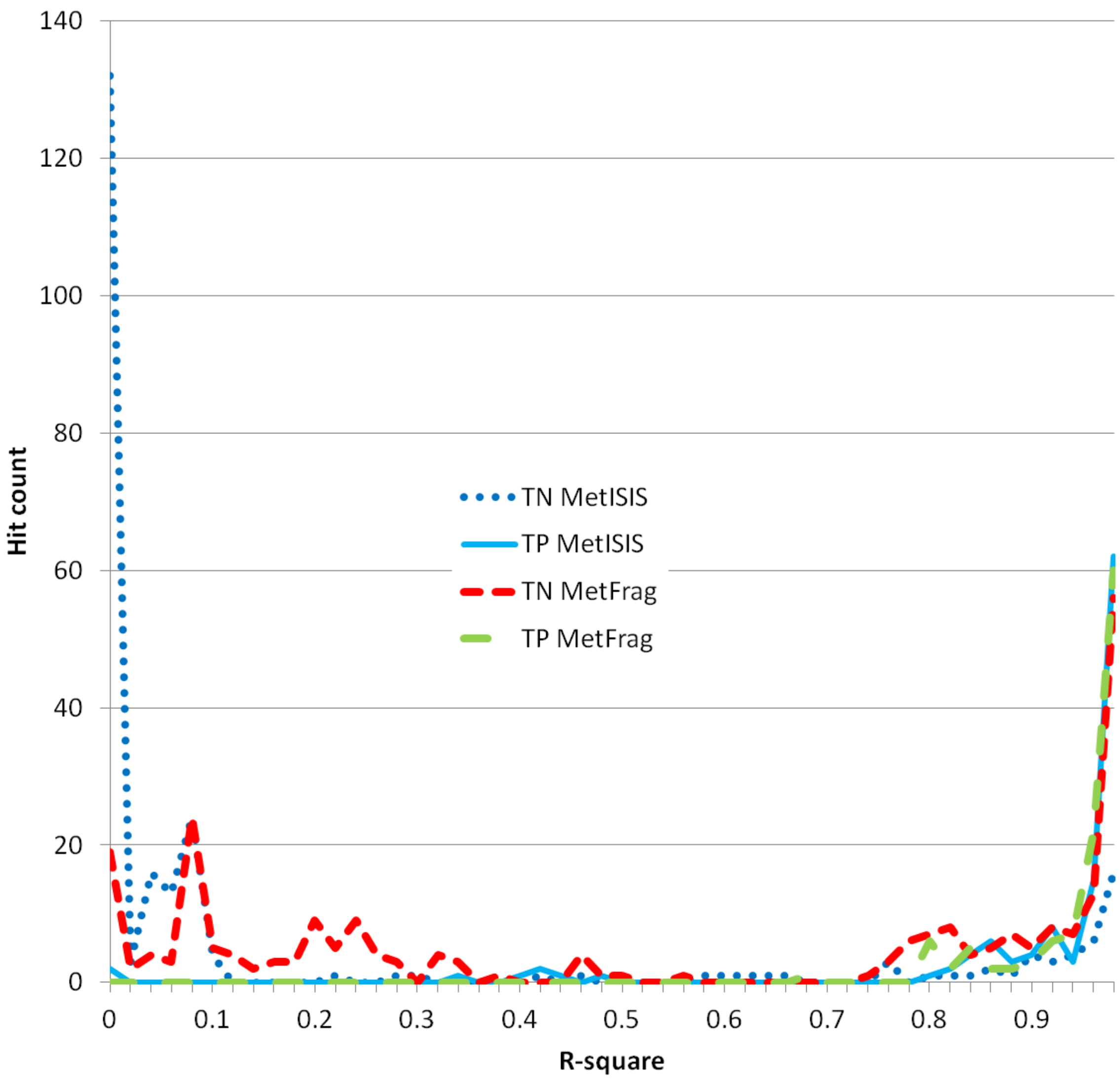


Figure 10.1. Distributions are MetISIS and MetFrag scores for rank list candidates. The scores are separated into those from true and false candidates.

That MetFrag has a high false positive rate can be seen in Figure 10.1. We set an arbitrary 0.5 cutoff level as if each algorithm performed binary classifications: a low score predicts a false candidate and a high score predicts a true candidate (Figure 10.1 suggests that any cutoff between 0.3 and 0.7 would yield similar results). MetFrag calls 132 false positives (MetISIS 47) of the 244 false candidates, giving MetFrag a low specificity of 0.459 compared to 0.807 for MetISIS.

Table 10.2. Statistics on MetISIS and MetFrag assuming a binary classification with a cutoff set at score 0.5.

MetISIS	**Candidate**		**Sensitivity**	0.931
	True ID	**False ID**	**Specificity**	0.807
True Call	108	47	**Positive predictive value**	0.697
Falls Call	8	197	**Negative predictive value**	0.961

MetFrag	**Candidate**		**Sensitivity**	1.000
	True ID	**False ID**	**Specificity**	0.459
True Call	116	132	**Positive predictive value**	0.532
Falls Call	0	112	**Negative predictive value**	1.000

MetISIS processes spectra much faster than the online MetFrag application, which required approximately one hour per identification; each observed spectrum was compared against an average of 8 candidate spectra. In a speed test, MetISIS identified 3400 observed spectra per minute; each compared to an average of 35 candidate spectra. Thus, only MetISIS is viable in high-throughput analytical environments and when high accuracy is needed. Supplement shows ranks and scores for all lipids tested in MetISIS and MetFrag.

11 LIPID SCREENING TEST

The test described here reflects the true conditions at which lipid metabolites are identified in contrast to the performance test in Chapter 9 where lipid standards were used. With lipid standards, the true identities are known, while in actual usage of MetISIS, the objective is of course to determine the identities. Consequently, rather than reporting on correct identification statistics, we discuss what can be observed in the screening results.

11.1 Sample preparation

Female BALB/cJ mice of age ~ 8 weeks old were purchased from the Jackson Laboratory (Bar Harbor, ME) and were euthanized with 70/30 CO_2/O_2 to harvest the femurs. To extract lipid species out of bone marrow, the femurs were first centrifuged at 12,000 g for 5 min to release the bone marrow. Immediately after centrifugation, 100 µl of freshly prepared 100 mM NH_4HCO_3 was added, followed by short agitation with a vortex mixer and kept at 5 °C for 15 min; 400 µl of - 20°C cold $CHCl_3/CH_3OH$ (2:1, v/v) was added afterwards. After votexing, the mixture was kept at 5 °C for 15 min, followed by 12,000 g centrifugation for 8 min. The bottom organic layer was transferred out and dried down in a speed-vac prior to reconstitution of the residue in 150 µl of methanol and subjected to LC-MS/MS analysis.

11.2 LC-MS/MS Analysis of Lipids

Each sample analyzed in a mass spectrometer (MS) typically contains hundreds of different species. To not overwhelm the MS instrument with all species at once, the samples are first separated in time by liquid chromatography (LC). The LC elutes the species according to their hydrophobicities over a period of time, for example, 30-60 minutes, and as the species are eluted from the LC, they are injected into the MS instrument.

To analyze the extracted lipids from bone marrow, 5 μl of reconstituted sample was injected onto 50 cm x 75 μ (i.d.) capillary column in-house packed with Jupiter 5 μ C_{18} particles (Phenomenex). LC separation was carried out on an in-house built dual column LC system under constant pressure (10,000 psi) with a gradient of mobile phase B from 0-100% in 90 min (mobile phase A: 50/50 H_2O/CH_3OH with 10 mM ammonium acetate; mobile phase B: 50/50 CH_3OH/CH_3CN with 10 mM ammonium acetate). The effluent from LC separation was electrospray ionized (ESI) in positive mode with the ion spray voltage, capillary voltage and capillary temperature were set to 2.2 kV, 48 V and 200 °C, respectively. MS data was acquired on a LTQ-Orbitrap mass spectrometer (ThermoFisher, Bremen, Germany). The survey MS scan was acquired with resolution of 100,000, followed by data dependent low resolution CID-MS/MS scans for the top 5 most intense ions. The parameters for CID were set as follows: isolation width (*m/z*) of 2 Da, normalized collision energy of 30%, activation Q of 0.18 and activation time of 30 msec. To avoid repeated fragmentation of the same ions, dynamic exclusion was applied if the same ions were selected within one minute.

11.3 Data Processing

The LC/MS/MS experiment described above yielded 5,648 CID spectra, which were matched against *in silico* spectra generated by MetISIS for the lipids in LIPIDMAPS database (www.lipidmaps.org). The first step in this process is to find the subset of database entries that have the same mass as the precursor ion within the accuracy of the LTQ-Orbitrap, which is substantially greater than that of a LTQ-linear ion trap. Consequently, the 500 ppm mass margin used in the screening test described in Chapter 8 is not appropriate for the experimental data being discussed in this chapter.

Figure 11.1 shows the number of entries in the LIPIDMAPS database for matching to the 5,648 experimental CID spectra as a function of mass margin. The curve shows an elbow at 4-7 ppm that corresponds approximately to the accuracy of the instrument (Gross, 1994). At 20 ppm, only 2,141 candidates are found in LIPIDMAPS, which is significantly less than the 5,648 spectra used in the search.

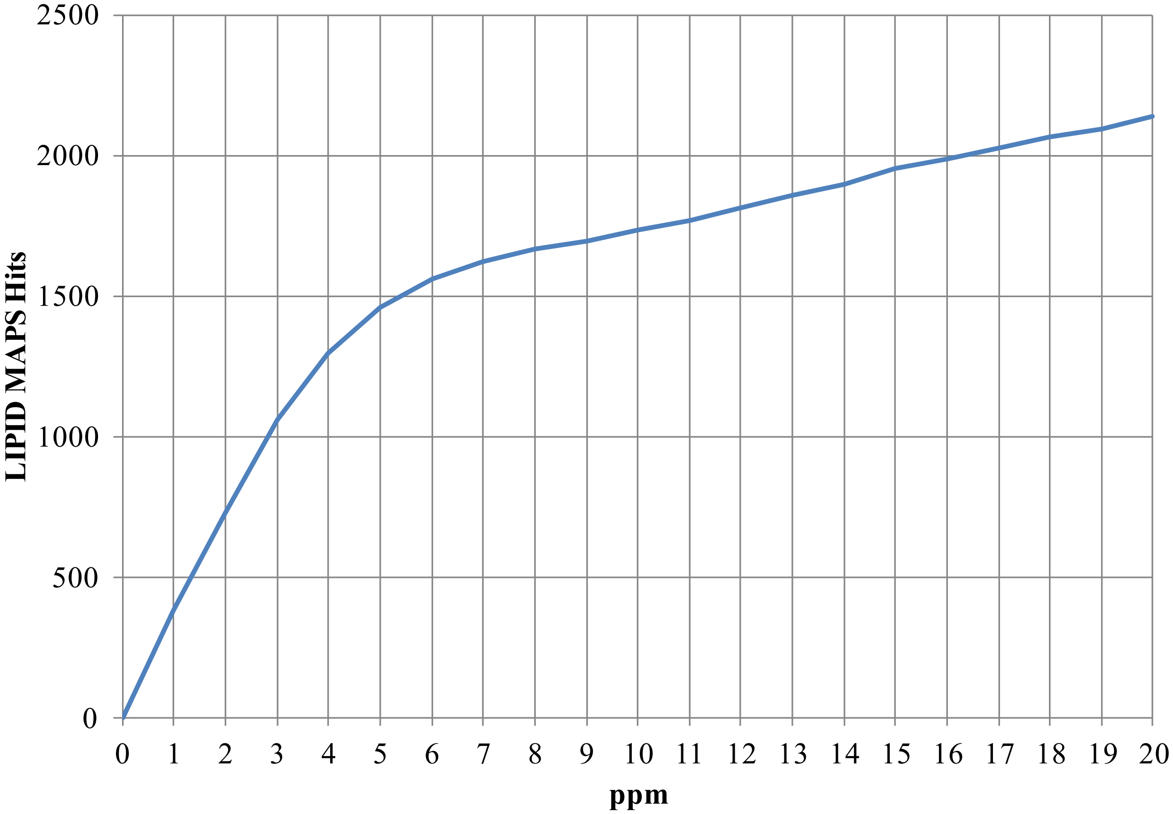


Figure 11.1. Number of LIPID MAPS Hits at Increasing Mass Margins: 0 to 20 ppm.

Figure 11.2 shows the number of LIPIDMAPS candidate identifications for mass margins from 0 to 1500 ppm. Even a wide mass margin of 1,500 ppm found only 4,529 candidates for matching CID spectra. The curve in Figure 11.2 change to near linear above 300 ppm. 300 ppm corresponds to 0.06 Da mass errors for 200 Da molecules and 0.33 Da mass errors for 1100 Da molecules. These mass errors are less than the mass of hydrogen (1.00794 Da) for all molecules processed by MetISIS.

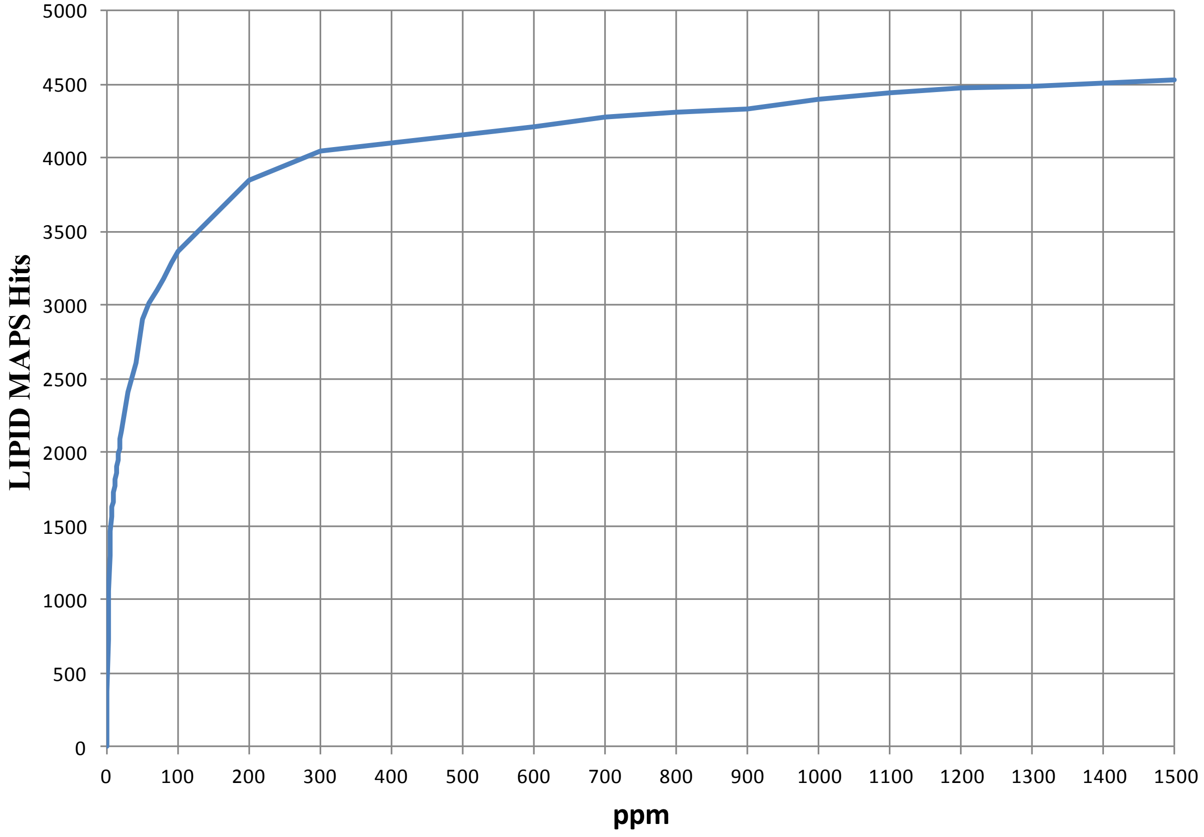
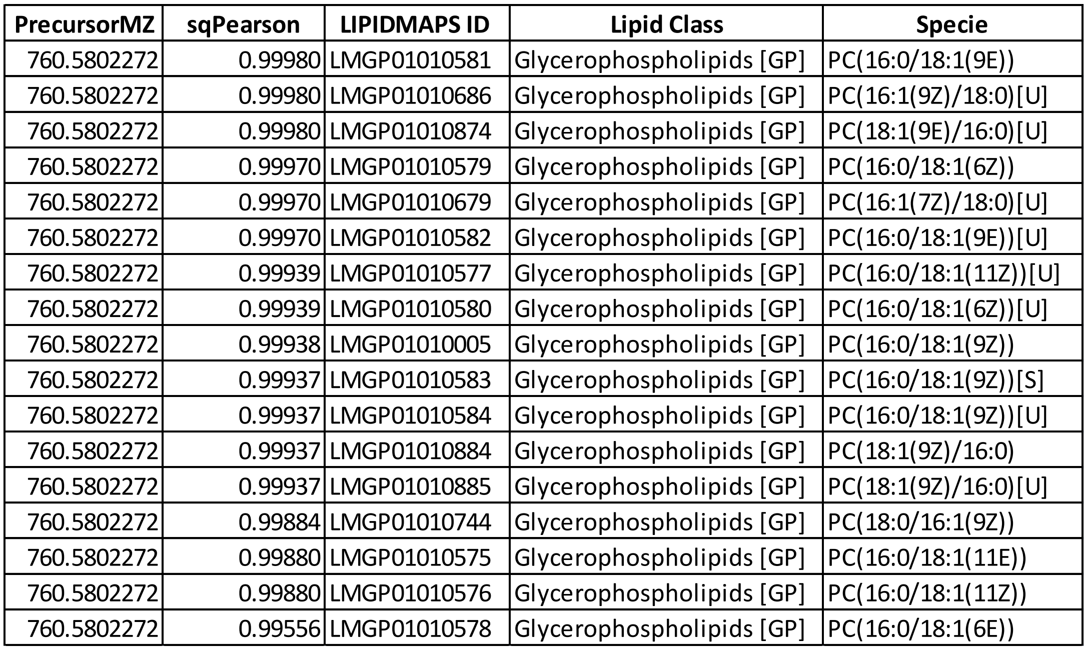


Figure 11.2. Number of LIPID MAPS Hits at Increasing Mass Margins: 0 to 1500 ppm.

Witt a mass margin of ±8 ppm, a set of 1,601 candidates in LIPIDMAPS were found for comparison to 346 experimental CID spectra, or an average of approximately 4.6 candidates per spectrum. The processor-ion masses of most of the 5,648 experimental CID spectra could not be matched to any entry in the LIPIDMAPS database with a ±8 ppm mass margin.

Table 11.1 shows 17 candidates found for one specific CID spectrum. Every candidate has one 16 carbon fatty acid tail, one 18 carbon fatty acid tail, and exactly one bond is saturated between the two tails. Clearly MetISIS associates this CID spectrum with PC 34:1 lipid (i.e., a total of 34 carbons and one saturated bond between the two fatty acid tails). All candidates have similar high (> 0.99) Pearson correlation coefficients between the observed CID spectrum and the *in silico* spectrum generated by MetISIS. The small variations in correlation coefficients are due to small variations in the *in silico* spectra generated by the non-deterministic Monte Carlo algorithm in MetISIS.

Table 11.1. LIPIDMAPS Hits for One Spectrum.



PrecursorMZ	sqPearson	LIPIDMAPS ID	Lipid Class	Specie
760.5802272	0.99980	LMGP01010581	Glycerophospholipids [GP]	PC(16:0/18:1(9E))
760.5802272	0.99980	LMGP01010686	Glycerophospholipids [GP]	PC(16:1(9Z)/18:0)[U]
760.5802272	0.99980	LMGP01010874	Glycerophospholipids [GP]	PC(18:1(9E)/16:0)[U]
760.5802272	0.99970	LMGP01010579	Glycerophospholipids [GP]	PC(16:0/18:1(6Z))
760.5802272	0.99970	LMGP01010679	Glycerophospholipids [GP]	PC(16:1(7Z)/18:0)[U]
760.5802272	0.99970	LMGP01010582	Glycerophospholipids [GP]	PC(16:0/18:1(9E))[U]
760.5802272	0.99939	LMGP01010577	Glycerophospholipids [GP]	PC(16:0/18:1(11Z))[U]
760.5802272	0.99939	LMGP01010580	Glycerophospholipids [GP]	PC(16:0/18:1(6Z))[U]
760.5802272	0.99938	LMGP01010005	Glycerophospholipids [GP]	PC(16:0/18:1(9Z))
760.5802272	0.99937	LMGP01010583	Glycerophospholipids [GP]	PC(16:0/18:1(9Z))[S]
760.5802272	0.99937	LMGP01010584	Glycerophospholipids [GP]	PC(16:0/18:1(9Z))[U]
760.5802272	0.99937	LMGP01010884	Glycerophospholipids [GP]	PC(18:1(9Z)/16:0)
760.5802272	0.99937	LMGP01010885	Glycerophospholipids [GP]	PC(18:1(9Z)/16:0)[U]
760.5802272	0.99884	LMGP01010744	Glycerophospholipids [GP]	PC(18:0/16:1(9Z))
760.5802272	0.99880	LMGP01010575	Glycerophospholipids [GP]	PC(16:0/18:1(11E))
760.5802272	0.99880	LMGP01010576	Glycerophospholipids [GP]	PC(16:0/18:1(11Z))
760.5802272	0.99556	LMGP01010578	Glycerophospholipids [GP]	PC(16:0/18:1(6E))

Figure 11.3 shows the Pearson R-square correlation values for the 1601 candidate identifications found for the 346 spectra. The R-square values are sorted in descending order. The candidates that have high R-square values and are lipid types that MetISIS has been trained on are believed to be correct identifications. Many candidates with low R-squared values, for example, diglyceride or triglyceride lipids are in classes that MetISIS was not yet trained to recognize; hence expanding the training set for MetISIS is desirable. Training based on experimental data for 73 lipids is discussed in section 11.4.

Results of screening tests on lipid standards (Chapter 8) showed R-squared values very near one or very near zero and few intermediate values. Figure 11.3 shows that about 30% of all candidates in the current screening tests have R-squared values between 0.1 and 0.9.

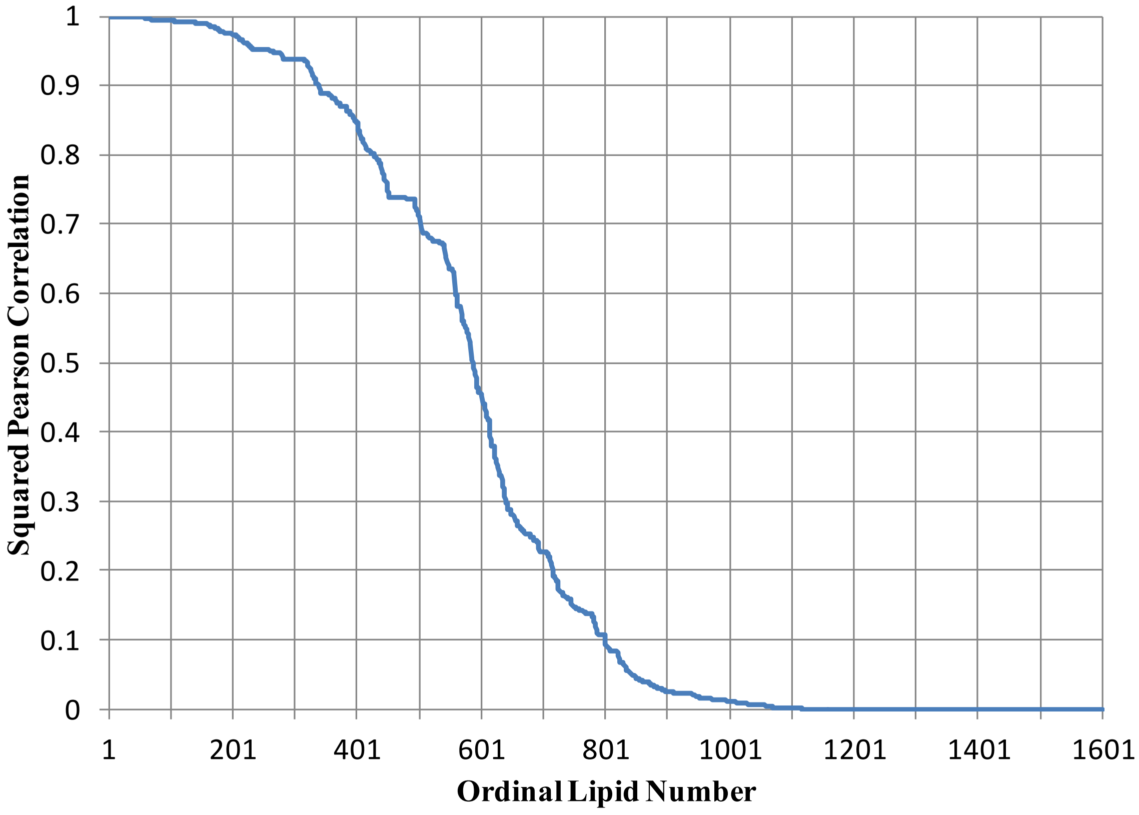


Figure 11.3. Squared Pearson Correlations of Candidate Identifications Sorted in Descending Order.

The explanation for this difference is believed to be that with standard lipids, each spectrum was ensured to be due to only one lipid species in the CID ion trap. In a sample with a large number of different species, more than one lipid species can be in the CID ion trap if they have similar *m/z* values. The MS configuration used to obtain the CID spectra being screened here selected all precursor ions in a 2 *m/z* window. This means that the MS instrument could select multiple lipid species for simultaneous fragmentation, and consequently, the product ions observed in spectra then consist of a mix of ions from several species. A mixed spectrum from, for example, two species can match two different candidates but neither of them will be a good match nor a poor match.

The confusion with mixed spectra is not unique to metabolomics but occur also in identifications in proteomics where only 20-25% of spectra are identified with significance as a know species. The quality of spectra resulting in low match scores can also result from low concentrations of species in samples. Low concentrations can cause low signal-to-noise ratios in the spectra.

Higher concentrations of species can increase the significance of identifications even when the individual spectra are of good quality. If a species occur in high concentrations, that species may elute over a longer period of time from the LC, and in turn be selected more often for fragmentations in the MS/MS step. Table 11.2 shows a sphingomyelin lipid (SM 40:2) identified from four different spectra (only the highest R-square candidate for each spectrum is shown). Seeing the same lipid identified multiple times increases the belief in that lipid being in the sample. The scan numbers (9588-10268) are also similar, suggesting that the four unknowns in the table eluted at similar times from the LC, since MS scan numbers are closely related to the LC elution time. The scan numbers for this MS experiment ranged from 6 to 17161.

Table 11.2. The Same Lipid Species Identified from Four Different Spectra.

SpectrumID	PrecursorMZ	sqPearson	scanNo	LIPIDMAPS ID	Lipid Class	Specie
3613	785.647868	0.99864	9637	LMSP03010071	Sphingolipids [SP]	SM(d16:1/24:1)
3598	785.648663	0.99840	9588	LMSP03010071	Sphingolipids [SP]	SM(d16:1/24:1)
3938	785.647912	0.91231	10268	LMSP03010072	Sphingolipids [SP]	SM(d18:1/22:1)
3937	785.647913	0.76189	10266	LMSP03010071	Sphingolipids [SP]	SM(d16:1/24:1)

A conservative estimate of the number of lipid species identified in this experiment is 36. Table 11.3 shows these lipids. Only candidates with R-squares greater than 0.99 were selected.

Five lipids were deleted from this set because they belonged to lipid classes not represented in the training data. Eleven of these lipids were observed more than once at R-squares greater than 0.99. Duplicate observations at R-squares below 0.99 were not counted.

Table 11.3. Identified Lipids from Experiment.

SpectrumID	PrecursorMZ	sqPearson	LIPIDMAPS ID	Class	Specie	Observations
3713	720.585551	0.99946	LMGP01080023	Glycerophospholipids [GP]	PC(16:0/O-16:0)	2
2866	706.533973	0.99899	LMGP01010530	Glycerophospholipids [GP]	PC(30:0)	1
2299	734.564624	0.99191	LMGP01010565	Glycerophospholipids [GP]	PC(32:0)	2
2779	732.549024	0.99910	LMGP01010490	Glycerophospholipids [GP]	PC(32:1)	2
3114	748.580624	0.94865	LMGP01010465	Glycerophospholipids [GP]	PC(33:0)	1
2695	744.548789	0.99966	LMGP01010543	Glycerophospholipids [GP]	PC(33:2)	1
2493	762.595738	0.91795	LMGP01011083	Glycerophospholipids [GP]	PC(34:0)	1
2777	760.580227	0.99980	LMGP01010581	Glycerophospholipids [GP]	PC(34:1)	2
2476	754.533271	0.99411	LMGP01010507	Glycerophospholipids [GP]	PC(34:4)	1
2786	770.564436	0.98138	LMGP01010611	Glycerophospholipids [GP]	PC(35:3)	1
2612	768.548495	0.99941	LMGP01010548	Glycerophospholipids [GP]	PC(35:4)	1
2506	780.549069	0.99974	LMGP01010634	Glycerophospholipids [GP]	PC(36:5)	1
2726	754.569535	0.96775	LMGP01020026	Glycerophospholipids [GP]	PC(O-15:0/20:4)	1
3787	734.600831	0.98973	LMGP01020032	Glycerophospholipids [GP]	PC(O-16:0/17:0)	2
2712	768.584822	0.99278	LMGP01020055	Glycerophospholipids [GP]	PC(O-16:0/20:4)	1
5013	804.679673	0.96932	LMGP01020061	Glycerophospholipids [GP]	PC(O-16:0/22:0)	2
4494	790.670398	0.98433	LMGP01020080	Glycerophospholipids [GP]	PC(O-17:0/20:0)	1
4843	818.703490	0.92426	LMGP01020083	Glycerophospholipids [GP]	PC(O-17:0/22:0)	1
3788	748.616582	0.99570	LMGP01020086	Glycerophospholipids [GP]	PC(O-18:0/16:0)	2
2999	796.616396	0.98986	LMGP01020102	Glycerophospholipids [GP]	PC(O-18:0/20:4)	1
2275	550.420504	0.99942	LMGP01040057	Glycerophospholipids [GP]	PC(O-18:0/O-3:1)	1
2538	506.357872	0.99677	LMGP01070012	Glycerophospholipids [GP]	PC(P-18:1/0:0)	1
3298	692.517872	0.96570	LMGP02010248	Glycerophospholipids [GP]	PE(32:0)	1
2872	716.517790	0.99203	LMGP02010042	Glycerophospholipids [GP]	PE(34:2)	1
2522	764.517255	0.99782	LMGP02010095	Glycerophospholipids [GP]	PE(38:6)	1
2747	740.518756	0.94473	LMGP20020008	Glycerophospholipids [GP]	PE(P-16:0/20:4)	1
5007	790.553882	0.98443	LMGP03010019	Glycerophospholipids [GP]	PS(36:1)	2
5151	788.538961	0.94553	LMGP03010030	Glycerophospholipids [GP]	PS(36:2)	2
3024	703.570071	0.99429	LMSP03010042	Sphingolipids [SP]	SM(34:1)	3
2927	729.585474	0.97631	LMSP03010051	Sphingolipids [SP]	SM(36:2)	1
3188	757.616458	0.99330	LMSP03010058	Sphingolipids [SP]	SM(38:2)	1
4297	773.647736	0.96236	LMSP03010067	Sphingolipids [SP]	SM(39:1)	1
3613	785.647868	0.99864	LMSP03010071	Sphingolipids [SP]	SM(40:2)	5
3170	783.633050	0.99457	LMSP03010070	Sphingolipids [SP]	SM(40:3)	1
3609	799.663520	0.98917	LMSP03010074	Sphingolipids [SP]	SM(41:2)	1
4371	813.679202	0.98921	LMSP03010007	Sphingolipids [SP]	SM(42:2)	1

Additional research will be required to determine how to optimally use all of the parameters associated with identifications. At this time, we do not have enough true labels for the experimental spectra to calculate sensitivities or specificities at different R-square thresholds. We believe that the number of observations and the scan numbers (shown in Table 11.2) should also be included in assigning significances to identifications.

11.4 Training Performance on 73 Standard Lipids

The training set discussed in this section included the 22 lipids shown in Table 9.1 plus the 46 lipids shown in Table 9.2, and the 5 ether lipids shown in Table 11.4. All lipids were purchased from Avanti Polar Lipids Inc. (Alabaster, AL).

Table 11.4. Ether lipids in training set.

Lipid Sub Class	Specie	Mass
Phosphatidylcholine	13:0e/13:0e	677.50
	18:1e/18:1e	729.53
	18(P)/18:1	733.56
	18(P)/20:4	761.59
Phosphatidylethanolamine	18(P)/20:4	579.39

The training performance with this set of 73 lipids is shown in Figure 11.4, where the lipids are sorted in descending order of the R-squared value between the experimental CID spectrum and that predicted by the trained ANN. The average R-square is 0.914, which is less than the average R-square of 0.97 obtained with the initial training set of 22 lipids. Figure 11.4 shows that most lipids were learned to a high R-square (50 had R-square greater than 0.9). The smallest R-square was obtained for ether PE lipid #73 that has an R-square near zero due to complex fragmentation pathways not yet implemented in MetISIS such as rearrangements and multiple bond cleavages to generate fragment ions. The experimental CID spectrum for lipid #72 (R-square of ~0.43) has an important peak that appears to require two bonds to cleave simultaneously; a functionality not yet implemented in MetISIS, since ion trap mass spectrometers usually perform single cleavages to generate ions. Ether phosphocholines (PCs) numbers 70 and 71 requires multiple bond cleavages to generate CID spectra, which contain 4 significant peaks. This training set of 73 lipids contained only 5 ether lipids and of these 70 and 71 were the only ones that have an ether in each of their two fatty acid hydrocarbon chains.

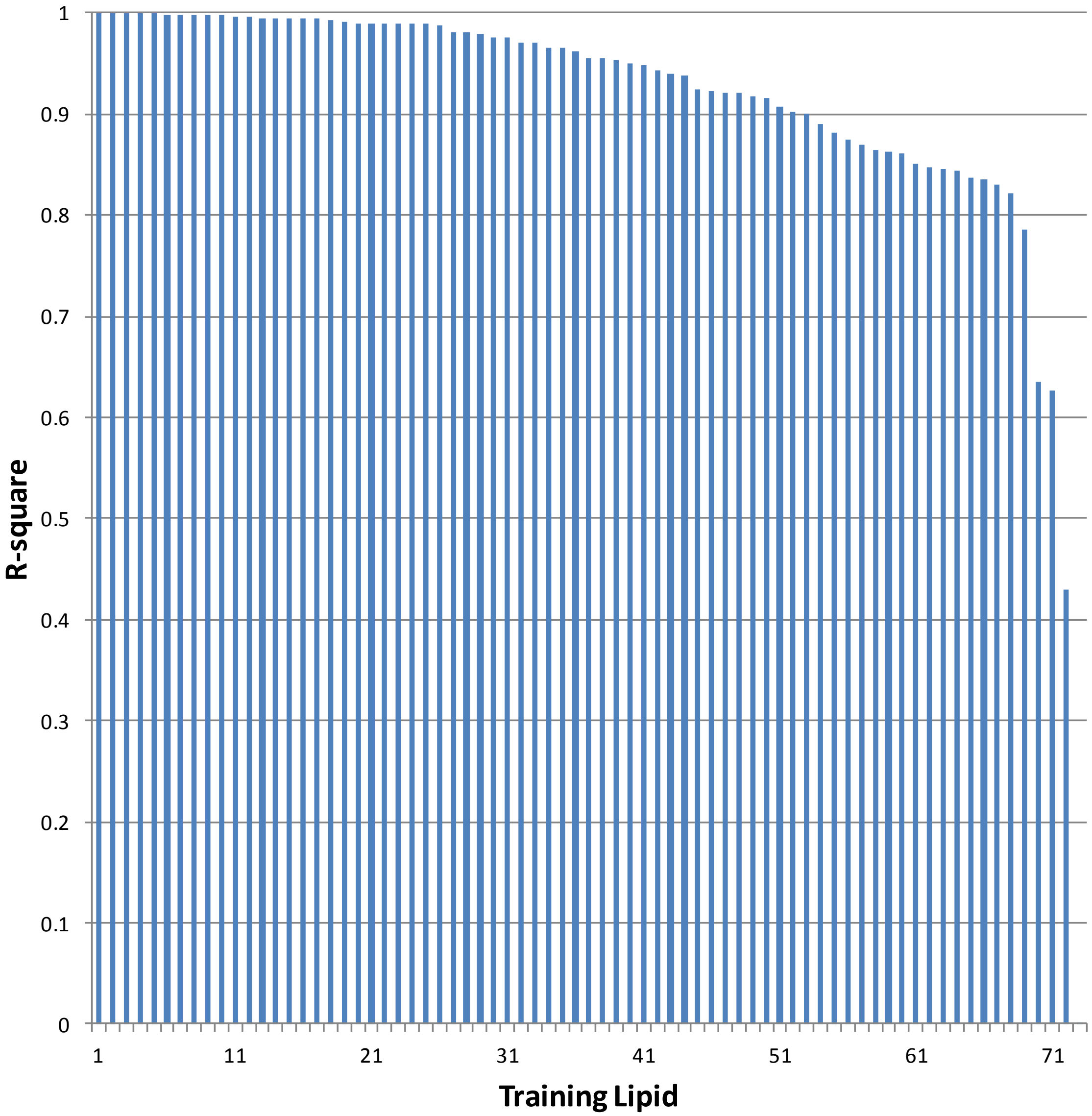


Figure 11.4. Accuracy of Predicting the CID Spectra of Training Lipids.

Increasing the training set to 73 lipids, required the input vectors to the artificial neural network (ANN) to increase from 1,260 (Table 6.1) to 1,486 elements that includes four additional inputs to aid the ANN predicting bond cleavage temperatures. (This ANN continued to use 8 hidden nodes.)

Table 11.5 shows the new inputs to the ANN. The total mass and total degrees of freedom were added to avoid having the ANN find these values from the two sides of the bond whose cleavage temperature is predicted. The other two additional inputs are flags indicating whether exactly one side of the bond was in a ring. In the absence of these flags, the ANN tends to cleave off hydroxide (OH) from rings in attempts to make water losses. The C-OH bonds with the carbon in the ring appear to be stronger than when the carbon is in a fatty acid where most water losses are observed.

Table 11.5. Reconfigured artificial neural network inputs.

Input	No of Inputs
Left tree encoding	738
Right tree encoding	738
Bond order	1
Cycle length	1
Total mass	1
Total degrees of freedom	1
Left tree mass	1
Right tree mass	1
Left tree degrees of freedom	1
Right tree degrees of freedom	1
Left atom in ring, right not	1
Right atom in ring, left not	1
	Total: 1486

Reconfiguring the ANN inputs to accommodate the new lipids in the training set, initially dropped the performance from an R-square of 0.97 for the 22 training lipids to ~0.8 for the 73 training lipids. Training was stopped after two months to generate this test with an average R-square of 0.915.

12 DISCUSSION

The MetISIS algorithm was developed to identify small molecules without any assumption of the type of molecules. The current configuration of MetISIS assumes molecules are composed of atoms from the set {C,H,O,N,S,P}, which is valid for most natural metabolite compounds.

To date only lipid metabolites have been used in training and testing of MetISIS, but no rules are used that assume lipids. Had lipid identification been the only target molecules, some rules could have been considered for the MetISIS algorithm.

Table 12.1 shows that many lipids like phospholipids (PC, SM, PE, PS, PI, PG, and PA) and galactolipids (MGDG and DGDG) are expected to have specific ions and neutral fragments in their CID spectra. Hence, simple rules can be included to either identify these lipids or add significance to lipid identifications if these ions and neutrals are observed. Information like this should be used to improve identifications and are used in other algorithms that focus on only lipids.

Table 12.1.Typical ions and neutral losses observed for lipids.

Scan mode	Precursor or neutral loss fragment	Lipid class detected
+	Prec 184	PC/LysoPC/SM
+	NL 141	PE/LysoPE
+	NL 185	PS
+	NL 277	PI
+	NL 189	PG
+	NL 115	PA
+	NL 179	MGDG
+	NL 341	DGDG
-	Prec 153	LysoPG

With MetISIS, we took a “purist” approach to avoid all rules because our intention was to make the algorithm generic to all molecules. Also, we believe that a rule based system cannot achieve the accuracy that is possible with MetISIS for a simple reason: not all rules can be made as we do not have all required knowledge for these rules. Table 5.2 showed that CID spectra of different variations of phosphocholine lipids all had the expected 184 Da peak, but the intensities of that fragment varies. Hence, rule-based approaches have to both identify all possible ions that can result from fragmentations and chemical reactions as well as determine at what intensities these ions occur. Both fragment sizes (m/z values) and their intensities have information that should be used in identifications. MetISIS uses both types of information through simulation of the CID process.

MetISIS was developed with metabolites in mind but nothing in the algorithm precludes it from identifying peptides, which are a special type of metabolite that has received a great deal of attention due to their close association with genomes. MetISIS now assumes at the most one charge in a molecule, while most peptides are larger than common metabolites and can take on multiple charges in the ionization step in mass spectrometry. Relatively minor logic added to MetISIS could make it ready for peptides.

Proteomics already has software tools like SEQUEST (Eng, 1994) and Mascot (Perkins, 1999) to identify peptides for species where protein coding DNA sequences or open reading frames (ORFs) are known. In the absence of a genome for the species of interest, *de novo* algorithms are used to elucidate a peptides' amino acid sequences. MetISIS applied to peptides would be neither a *de novo* approach nor one that searches genomes for ORFs. With MetISIS, peptides of interest (hypothetical or known to exist) could be placed in the *in silico* database as *in silico* spectra generated by MetISIS, which is searched to identify matches to observed CID spectra just like any other metabolite.

13 CONCLUSION

MetISIS was developed to generate *in silico* spectra of lipids for high throughput identifications in LC-MS-based non-peptide small-molecule studies. In the first test with lipids, the software appears to have significant sensitivity and specificity. Although the test was small with only a subset of lipid classes, MetISIS is expected to do well with other lipid classes and other metabolites as these are incorporated into the algorithm. The current training set has increased from 22 to 97 lipids that include more lipid classes and as more training exemplars are added, the algorithm is also expected to generalize better to new (untrained) metabolite classes.

Currently, the software only ranks database hits. An approach to improve the rank scores is to generate *in silico* spectra in both positive and negative MS modes. While some lipids only yield good spectra in one mode, many produce quality fragment ions in both modes for better rank scores. The algorithm has the capacity and versatility to be trained with either positive or negative mode spectra.

Also, to reduce the number of candidates for true positives, the rank lists can be shortened by using hybrid mass spectrometers like LTQ-Orbitrap or quadrupole-time-of-flight which have higher resolving powers and would allow narrower mass margins when screening *in silico* databases. Indeed, we have started the identification of experimental lipids using an LTQ-Orbitrap.

Two important additions to the MetISIS algorithm are currently in development. The first is modeling any rearrangement of atoms and bonds from bond cleavages. The second is to enable the algorithm to process different adducts—it now only accepts hydrogen adducts.

Modeling rearrangements provides individual atom charges that in turn provide the means to calculate fragment charges. Consequently, the ANN described to predict fragment charges may not be needed in the algorithm.

The algorithm presented here models a linear ion trap, a tandem in time instrument that typically generates ions from only primary fragmentations. However, the algorithm allows also secondary fragmentations. As we adopt the software to, for example, a tandem in space instrument like a triple quadrupole, only a small amount of program code need to be changed after the research needed to design a model for the new instrument.

Bibliography

ACDLabs, http://www.acdlabs.com/products/adh/ms/ms_frag/

Aoki, K.F. and Kanehisa, M. (2005) Using the KEGG database resource. *Current Protocols in Bioinformatics*. John Wiley & Sons, Hoboken, New Jersey, Chapter 1, Unit 1.12.

Auberry, K.J., Kiebel, G.R., Monroe, M.E., Adkins, J.N., Anderson, G.A., and Smith, R.D. (2010) J Proteomics Bioinform, 3 (1), 1-4.

Armitage, J., et al. (1967) Documentation of Chemical Reactions by Computer Analysis of Structural Changes. J. Chem. Doc., 7 (4), pp 209–215.

Baba K, Enbutu I, Yoda M. Explicit representation of knowledge acquired from plant historical data using neural network, Proceedings of the International Joint Conference on Neural Networks (1990) (3), pp. 155–160, 1990.

Bach, R.D., *et al.* (1996) A Reassessment of the Bond Dissociation Energies of Peroxides. An *ab Initio* Study. *J. Am. Chem. Soc.*, **118**, 12758-12765.

Bersohn, M. and Esack, A. (1976) A computer representation of synthetic organic reactions. Computers & Chemistry 1(2): 103-108.

Bortz, A.B., *et al.* (1991) Theoretical foundations of dynamical Monte Carlo simulations. *J. Chem. Phys.*, **95**, 1090.

Brown AC, 1864, Transactions of the Royal Society of Edinburgh, 23,707–720 (1864).

Butlerov AM, 1861, Z. Chem,. 4(1861) 549.

Couper AS, 1858, Ann.Chim. Phys. 53(3) 1858) 469.

Cunningham Jr C, Glish GL, 2006, High Amplitude Short Time Excitation: A Method to Form and Detect Low Mass Product Ions in a Quadropole Ion Trap Mass Spectrometer, J. Am. Soc. Mass Spectrom., 17, 81-84.

Drahos, L., *et al.* (1999) Thermal Energy Distribution Observed in Electrospray Ionization. *J. Mass Spectrom.*, **34**, 1273-1379.

Drahos, L. and Vékey, K. (1999) Determination of the Thermal Energy and its Distribution in Peptides. *J. Am. Soc. Mass Spectrom.*, **10**, 323-328.

Eng, J.K., *et al.* (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976-989.

Euler L, 1736, Comm. Acad. Scie. Imp. Petropol., 8, 128.

Faulon J.L., *et al.* (2005) Enumerating Molecules. In Lipkowitz, K., Larter, R. and Cundari, T.R. (eds.), *Reviews in Computational Chemistry Vol. 21.* John Wiley & Sons, Hoboken, New Jersey.

Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. Science 246 (4926): 64–71.

Gabelica, V. and De Pauw, E. (2005) Internal Energy and Fragmentation of Ions Produced in Electrospray Sources. *Mass Spectrom. Rev.*, **24**, 566– 587.

Gabelica, V., *et al.* (2003) Calibration of Ion Effective Temperatures Achieved by Resonant Activation in a Quadropole ion trap. *Anal. Chem.*, **75**, 5152-5159.

Gillespie, D.T. (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.,* **22**, 403-434.

Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading Massachusetts.

Gross, M. L. (1994) Accurate Masses for Structure Confirmation. J. Am. Soc. Mass Spectrom 5 (2): 57

Hildebrandt, C., *et al.* (2011) Database supported candidate search for Metabolite identification. J. Integrative Bioinformatics 8(2).

Hill, D.W., *et al.* (2008) Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. *Anal. Chem.*, **80**, 5574-5582.

Holland, J.H. (1975) *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, Michigan.

Hornik K, Stinchcombe M, White H, (1989), Neural Networks 2, p. 359.

Laskin L., *et al.* (2000) Internal energy distributions resulting from sustained off-resonance excitation in FTMS. *Int. J. Mass Spectrom.*, **195**, 285-302.

LeCun, Y. (1985) Une procedure d'apprentissage pour reseau a seuil assymetrique, *Proc. Cognitiva '85: A la frontière de l 'intelligence Artificielle des Sciences de la Connaissance des Neuronsciences*, pp. 599-604.

Lederberg, J. (1987) How Dendral Was Conceived and Born. ACM Symposium on the History of Medical Informatics, 5 November 1987, Rockefeller University. New York: National Library of Medicine.

Lopez, L.L. *et al.* (1999) Automated Strategies for Obtaining Standardized Collisionally Induced Dissociation Spectra on a Benchtop Ion Trap Mass Spectrometer. *Rapid Commun. Mass Spectrom.*, **13**, 663-668.

Marcus, R.A. (1952) Unimolecular Dossociations and Free Radical Recombination Reactions. *J. Chem. Phys.*, **20**, 359.

Marzluff EM, Campbell S, Rodgers MT, and Beauchamp JL, 1994, Collisional Activation of Large Molecules is an Efficient Process, J. Am. Chem. Soc., 116, 6947-6948.

McLuckey S.A. and Goeringer, D.E. (1997) Slow Heating Methods in Tandem Mass Spectrometry. *J. Mass Spectrom.*, **32**, 461-474.

Meng, B. and Weinberg, W.H. (1994) Monte Carlo simulations of temperature programmed desorption spectra. *J. Chem. Phys.*, **100**, 5280.

Metz, T.O. *et al.* (2007) The future of liquid chromatography-mass spectrometry (LC-MS) in metabolic profiling and metabolomics studies for biomarker discovery. *Biomark. Med.*, **1**, 159-185.

Naban-Maillet, J. *et al.* (2005) Internal Energy Distribution in Electrospray onization. *J. Mass Spectrom.*, **40**, 1–8.

Pak, A. *et al.* (2008) Internal Energy Distribution of Peptides in Electrospray Ionization: ESI and Collision-induced Dissociation Spectra Calculation. *J. Mass Spectrom.*, **43**, 447-455.

Parker, D. (1982) *Learning-logic*, Invention Report S81-64, File 1, Office of Technology Licensing, Stanford University, Palo Alto, California.

Pennerath, F. et al. (2010) Graph-Mining Algorithm for the Evaluation of Bond Formability. J. Chem. Inf. Model., 50, 221–239.

Perkins, D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551-3567.

Priddy KL, Keller PE. Artificial Neural Networks: An Introduction, SPIE Press, Bellingham, WA, 2005, p. 1.

Pulfer M and Murphy RC, 2003, Electrospray Mass Spectrometry of Phospholipids. Mass Spectrometry Reviews, 2003, 22, 332-364.

Rosenstock, H.M. *et al.* (1952) Absolute rate theory for isolated systems and the mass spectra of polyatomic molecules. *Proc. Natl. Acad. Sci. USA*, **38**, 667-678.

Rumelhart, D.E. *et al.* (1986) Learning internal representations by error propagation, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Vol. 1: Foundations*, D.E. Rumelhart, J.L. McClelland, (eds.), MIT Press, Cambridge, Massachusetts, pp. 318-36.

Schietgat L. *et al.* (2008) An efficiently computable graph-based metric for the classification of small molecules. *Proc. of the 11th International Conference on Discovery Science* (LNAI 5525), pp. 197-209.

Schneider, B.B. *et al.* (2001) Collision-Induced Dissociation of Bradykinin Ions in the Interface Region of an ESI-MS. American Society for Mass Spectrometry. 12(7):772-9.

Schrom, B. *et al*. (2011) Charge Prediction of Lipid Fragments in Mass Spectrometry. *Proc. of Machine Learning and Applications and Workshops, 10th International Conference*, Dec 18-21.

Schwartz, J.C. *et al*. (2002) A two-dimensional quadrupole ion trap mass spectrometer. Journal of the American Society for Mass Spectrometry. Vol 13, Issue 6, June, p.659-669.

Shukla, A.K. and Futrell, J.H. (2000) Tandem mass spectrometry: dissociation of ions by collisional activation. *J. Mass Spectrom*., **35**, 1069-1090.

Sleno, S. and Volmer, D.A. (2004) Ion activation methods for tandem mass spectrometry. *J. Mass Spectrom*., **39**, 1091-1112.

Smith, C.A., *et al.* (2005) METLIN: a metabolite mass spectral database. *Ther. Drug Monit.,* **27**, 747-751.

Sylvester JJ, (1878), Nature, 17, 284.

Sztáray, J. (2009) *Modeling the Dissociation of Protonated Ions*. Ph.D. Dissertation. Institute of Structural Chemistry, Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary.

Vékey, K. (1996) Internal Energy Effects in Mass Spectrometry, *J. Mass Spectrom.*, **31**, 445-463.

Vleduts, G.E. (1963) Concerning One System of Classsification and Codification of Organic Reactions. Inf. Storage Retr., 1, 117-146.

Voter, A.F. (2005) Introduction to the Kinetic Monte Carlo Method, in Radiation Effects in Solids, edited by K. E. Sickafus and E. A. Kotomin (Springer, NATO Publishing Unit, Dordrecht, The Netherlands).

Wells, J.M. and McLuckey, S.A. (2005) Collision-Induced Dissociation (CID) of Peptides and Proteins. Methods in Enzymology, Vol. 402.

Werbos, P.J. (1974) *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD Thesis, Harvard University, Cambridge, Massachusetts.

Werbos, P.J. (1994) *The Roots of Backpropagation*. John Wiley & Son,. Hoboken, New Jersey.

Willett, P. (1979) Computer Techniques for the Indexing of Chemical Reaction Information. J. Chem. Inf. Comput. Sci., 19 (3), pp 156–158.

Williams, E.R, Kent D. Henry, Fred W. McLafferty, Jeffrey Shabanowitz and Donald F. Hunt, 1990, Journal of The American Society for Mass Spectrometry, Volume 1, Number 5, 413-416.

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

Young, W.M. and Elcock, E.W. (1966) Monte Carlo Studies of Vacancy Migration in Binary Ordered Alloys: I. *Proc. Phys. Soc.,* **89**, 735.

Zhang, Z. (2004) Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides. *Anal. Chem.*, **76**, 3908-3922.

SUPPLEMENT

Ranks and scores from MetISIS and MetFrag comparison.

							MetFrag		
Test Lipid	LipidMaps ID	Mass	Class	Specie	Correct Candidate	ISIS Score	Score	Rank	Best Ranked Correct ID
0	LMGP02010013	523.3274	Glycerophospholipids [GP]	PE(10:0/10:0)[U]	0	0.000	1.000	1	
0	LMGP02010101	523.3274	Glycerophospholipids [GP]	PE(10:0/10:0)	0	0.000	1.000	2	
0	LMGP02010092	523.3274	Glycerophospholipids [GP]	PE(9:0/11:0)[U]	0	0.000	0.987	3	
0	LMGP01050078	523.3638	Glycerophospholipids [GP]	PC(0:0/18:0)[U]	0	0.083	0.985	6	
0	LMGP01050077	523.3638	Glycerophospholipids [GP]	PC(0:0/18:0)[S]	0	0.103	0.985	5	
0	LMGP01050076	523.3638	Glycerophospholipids [GP]	PC(0:0/18:0)	0	0.096	0.985	4	
0	LMGP02010221	523.3274	Glycerophospholipids [GP]	PE(11:0/9:0)[U]	0	0.000	0.982	7	
0	LMGP01050028	523.3638	Glycerophospholipids [GP]	**PC(18:0/0:0)[U]**	1	0.951	0.975	10	
0	LMGP01050027	523.3638	Glycerophospholipids [GP]	**PC(18:0/0:0)[S]**	1	0.880	0.975	9	
0	**LMGP01050026**	**523.3638**	**Glycerophospholipids [GP]**	**PC(18:0/0:0)**	**1**	**0.916**	**0.975**	**8**	**4**
0	LMGP01060017	523.4002	Glycerophospholipids [GP]	PC(O-19:0/0:0)	0	0.957	0.972	11	
0	LMGP03050004	523.2910	Glycerophospholipids [GP]	PS(18:1(9Z)/0:0)[U]	0	0.921	0.907	13	
0	LMGP03050001	523.2910	Glycerophospholipids [GP]	PS(18:1(9Z)/0:0)	0	0.779	0.907	12	
0	LMGP01040011	523.4002	Glycerophospholipids [GP]	PC(O-10:0/O-9:0)[U]	0	0.060	0.116	14	
0	LMGP01040043	523.4002	Glycerophospholipids [GP]	PC(O-16:0/O-3:0)[U]	0	0.071	0.097	15	
0	LMGP01080008	523.3638	Glycerophospholipids [GP]	PC(2:0/O-16:0)[U]	0	0.073	0.097	16	
0	LMGP01020048	523.3638	Glycerophospholipids [GP]	PC(O-16:0/2:0)[U]	0	0.013	0.095	19	
0	LMGP01020047	523.3638	Glycerophospholipids [GP]	PC(O-16:0/2:0)[S]	0	0.013	0.095	18	
0	LMGP01020046	523.3638	Glycerophospholipids [GP]	PC(O-16:0/2:0)	0	0.015	0.095	17	
0	LMGP01040050	523.4002	Glycerophospholipids [GP]	PC(O-18:0/O-1:0)[U]	0	0.044	0.092	24	
0	LMGP01040049	523.4002	Glycerophospholipids [GP]	PC(O-18:0/O-1:0)[S]	0	0.049	0.092	23	
0	LMGP01040048	523.4002	Glycerophospholipids [GP]	PC(O-18:0/O-1:0)	0	0.058	0.092	22	
0	LMGP01040007	523.4002	Glycerophospholipids [GP]	PC(O-1:0/O-18:0)[U]	0	0.041	0.092	21	
0	LMGP01040006	523.4002	Glycerophospholipids [GP]	PC(O-1:0/O-18:0)	0	0.044	0.092	20	
1	LMGP01050073	467.3012	Glycerophospholipids [GP]	PC(0:0/14:0)	0	0.113	1.000	1	
1	LMGP02010102	467.2648	Glycerophospholipids [GP]	PE(8:0/8:0)	0	0.003	0.996	2	
1	LMGP02010103	467.2648	Glycerophospholipids [GP]	PE(8:0/8:0)[U]	0	0.002	0.996	3	
1	LMGP01060009	467.3376	Glycerophospholipids [GP]	PC(O-15:0/0:0)	0	0.919	0.992	4	
1	**LMGP01050012**	**467.3012**	**Glycerophospholipids [GP]**	**PC(14:0/0:0)**	**1**	**0.871**	**0.988**	**5**	**4**
1	LMGP01050013	467.3012	Glycerophospholipids [GP]	**PC(14:0/0:0)[U]**	1	0.880	0.988	6	
1	LMGP02060003	467.3376	Glycerophospholipids [GP]	PE(O-18:0/0:0)	0	0.772	0.953	7	
1	LMFA08020076	467.3036	Fatty Acyls [FA]	N-arachidonoyl tyrosine	0	0.584	0.836	8	
1	LMGP01011232	467.2648	Glycerophospholipids [GP]	PC(6:0/7:0)[U]	0	0.091	0.181	9	
1	LMGP01011247	467.2648	Glycerophospholipids [GP]	PC(8:0/5:0)[U]	0	0.091	0.177	10	
1	LMGP01020009	467.3012	Glycerophospholipids [GP]	PC(O-12:0/2:0)	0	0.012	0.161	11	
1	LMGP01020010	467.3012	Glycerophospholipids [GP]	PC(O-12:0/2:0)[U]	0	0.010	0.161	12	
1	LMGP01040021	467.3376	Glycerophospholipids [GP]	PC(O-14:0/O-1:0)	0	0.067	0.077	13	
1	LMGP01040022	467.3376	Glycerophospholipids [GP]	PC(O-14:0/O-1:0)[S]	0	0.050	0.077	14	
1	LMGP01040023	467.3376	Glycerophospholipids [GP]	PC(O-14:0/O-1:0)[U]	0	0.059	0.077	15	
2	LMSP02010002	481.4495	Sphingolipids [SP]	Cer(d18:1/12:0)	0	0.876	1.000	1	
2	**LMGP01050016**	**481.3168**	**Glycerophospholipids [GP]**	**PC(15:0/0:0)**	**1**	**0.906**	**0.989**	**2**	**2**
2	LMGP01050017	481.3168	Glycerophospholipids [GP]	**PC(15:0/0:0)[S]**	1	0.901	0.989	3	
2	LMGP01050117	481.3168	Glycerophospholipids [GP]	**PC(15:0/0:0)[U]**	1	0.936	0.989	4	
2	LMGP01060010	481.3532	Glycerophospholipids [GP]	PC(O-16:0/0:0)	0	0.923	0.984	5	
2	LMGP01060011	481.3532	Glycerophospholipids [GP]	PC(O-16:0/0:0)[S]	0	0.949	0.984	6	
2	LMGP01060012	481.3532	Glycerophospholipids [GP]	PC(O-16:0/0:0)[U]	0	0.910	0.984	7	
2	LMGP02050001	481.3168	Glycerophospholipids [GP]	PE(18:0/0:0)	0	0.306	0.923	8	
2	LMFA03020008	481.2498	Fatty Acyls [FA]	N-acetyl-LTE4	0	0.083	0.919	9	
2	LMGP01011238	481.2805	Glycerophospholipids [GP]	PC(7:0/7:0)	0	0.094	0.225	10	
2	LMGP01011239	481.2805	Glycerophospholipids [GP]	PC(7:0/7:0)[S]	0	0.092	0.225	11	
2	LMGP01011240	481.2805	Glycerophospholipids [GP]	PC(7:0/7:0)[U]	0	0.091	0.225	12	
2	LMGP01011233	481.2805	Glycerophospholipids [GP]	PC(6:0/8:0)	0	0.091	0.214	13	
2	LMGP01011234	481.2805	Glycerophospholipids [GP]	PC(6:0/8:0)[U]	0	0.090	0.214	14	
2	LMGP01011248	481.2805	Glycerophospholipids [GP]	PC(8:0/6:0)	0	0.095	0.212	15	
2	LMGP01011249	481.2805	Glycerophospholipids [GP]	PC(8:0/6:0)[U]	0	0.091	0.212	16	
2	LMGP01011269	481.2805	Glycerophospholipids [GP]	PC(9:0/5:0)	0	0.090	0.210	17	
2	LMGP01010403	481.2805	Glycerophospholipids [GP]	PC(10:0/4:0)	0	0.091	0.206	18	
2	LMGP01010443	481.2805	Glycerophospholipids [GP]	PC(12:0/2:0)	0	0.027	0.198	18	
2	LMGP01040084	481.3532	Glycerophospholipids [GP]	PC(O-8:0/O-8:0)	0	0.080	0.121	19	
2	LMGP01040085	481.3532	Glycerophospholipids [GP]	PC(O-8:0/O-8:0)[U]	0	0.080	0.121	20	
2	LMGP01040028	481.3532	Glycerophospholipids [GP]	PC(O-15:0/O-1:0)[U]	0	0.048	0.088	21	

							MetFrag		
Test Lipid	LipidMaps ID	Mass	Class	Specie	Correct Candidate	ISIS Score	Score	Rank	Best Ranked Correct ID
3	LMGP02010285	495.2961	Glycerophospholipids [GP]	PE(9:0/9:0)[U]	0	0.000	1.000	1	
3	LMGP01050074	495.3325	Glycerophospholipids [GP]	PC(0:0/16:0)	0	0.112	0.992	2	
3	LMGP01050075	495.3325	Glycerophospholipids [GP]	PC(0:0/16:0)[U]	0	0.105	0.992	3	
3	**LMGP01050018**	**495.3325**	**Glycerophospholipids [GP]**	**PC(16:0/0:0)**	**1**	**0.884**	**0.972**	**4**	**3**
3	LMGP01050113	495.3325	Glycerophospholipids [GP]	**PC(16:0/0:0)[rac]**	1	0.908	0.972	7	
3	LMGP01050019	495.3325	Glycerophospholipids [GP]	**PC(16:0/0:0)[S]**	1	0.931	0.972	5	
3	LMGP01050020	495.3325	Glycerophospholipids [GP]	**PC(16:0/0:0)[U]**	1	0.962	0.972	6	
3	LMGP01060013	495.3689	Glycerophospholipids [GP]	PC(O-17:0/0:0)	0	0.967	0.969	8	
3	LMGP01011241	495.2961	Glycerophospholipids [GP]	PC(7:0/8:0)[U]	0	0.089	0.227	9	
3	LMGP01011250	495.2961	Glycerophospholipids [GP]	PC(8:0/7:0)[U]	0	0.090	0.219	10	
3	LMGP01020019	495.3325	Glycerophospholipids [GP]	PC(O-14:0/2:0)	0	0.016	0.103	11	
3	LMGP01020020	495.3325	Glycerophospholipids [GP]	PC(O-14:0/2:0)[U]	0	0.020	0.103	12	
3	LMGP01040005	495.3689	Glycerophospholipids [GP]	PC(O-1:0/O-16:0)[U]	0	0.062	0.091	16	
3	LMGP01040030	495.3689	Glycerophospholipids [GP]	PC(O-16:0/O-1:0)	0	0.063	0.091	13	
3	LMGP01040031	495.3689	Glycerophospholipids [GP]	PC(O-16:0/O-1:0)[S]	0	0.051	0.091	14	
3	LMGP01040032	495.3689	Glycerophospholipids [GP]	PC(O-16:0/O-1:0)[U]	0	0.066	0.091	15	
4	LMGP02010091	509.3118	Glycerophospholipids [GP]	PE(9:0/10:0)[U]	0	0.000	1.000	1	
4	LMGP02010272	509.3118	Glycerophospholipids [GP]	PE(10:0/9:0)[U]	0	0.000	0.989	2	
4	LMSP02010001	509.4808	Sphingolipids [SP]	Cer(d18:1/14:0)	0	0.893	0.963	3	
4	LMGP01060014	509.3845	Glycerophospholipids [GP]	PC(O-18:0/0:0)	0	0.962	0.929	4	
4	LMGP01060015	509.3845	Glycerophospholipids [GP]	PC(O-18:0/0:0)[S]	0	0.960	0.929	5	
4	LMGP01060016	509.3845	Glycerophospholipids [GP]	PC(O-18:0/0:0)[U]	0	0.972	0.929	6	
4	**LMGP01050024**	**509.3481**	**Glycerophospholipids [GP]**	**PC(17:0/0:0)**	**1**	**0.931**	**0.928**	**7**	**4**
4	LMGP01050025	509.3481	Glycerophospholipids [GP]	**PC(17:0/0:0)[U]**	1	0.886	0.928	8	
4	LMGP01011251	509.3118	Glycerophospholipids [GP]	PC(8:0/8:0)	0	0.077	0.255	9	
4	LMGP01011252	509.3118	Glycerophospholipids [GP]	PC(8:0/8:0)[S]	0	0.080	0.255	10	
4	LMGP01011253	509.3118	Glycerophospholipids [GP]	PC(8:0/8:0)[U]	0	0.078	0.255	11	
4	LMGP01010504	509.3118	Glycerophospholipids [GP]	PC(14:0/2:0)	0	0.033	0.219	12	
4	LMGP01040087	509.3845	Glycerophospholipids [GP]	PC(O-9:0/O-9:0)[U]	0	0.065	0.125	13	
4	LMGP01040010	509.3845	Glycerophospholipids [GP]	PC(O-10:0/O-8:0)[U]	0	0.061	0.115	14	
4	LMGP01020024	509.3481	Glycerophospholipids [GP]	PC(O-15:0/2:0)	0	0.011	0.091	15	
4	LMGP01020025	509.3481	Glycerophospholipids [GP]	PC(O-15:0/2:0)[U]	0	0.015	0.091	16	
4	LMGP01020004	509.3481	Glycerophospholipids [GP]	PC(O-1:0/16:0)	0	0.069	0.089	17	
4	LMGP01040068	509.3845	Glycerophospholipids [GP]	PC(O-2:0/O-16:0)[U]	0	0.081	0.088	18	
4	LMGP01020028	509.3481	Glycerophospholipids [GP]	PC(O-16:0/1:0)	0	0.026	0.084	19	
4	LMGP01040041	509.3845	Glycerophospholipids [GP]	PC(O-16:0/O-2:0)	0	0.049	0.084	20	
4	LMGP01040042	509.3845	Glycerophospholipids [GP]	PC(O-16:0/O-2:0)[U]	0	0.051	0.084	21	
4	LMGP01040047	509.3845	Glycerophospholipids [GP]	PC(O-17:0/O-1:0)[U]	0	0.056	0.083	22	
5	LMGP01050081	521.3481	Glycerophospholipids [GP]	PC(0:0/18:1(9E))[U]	0	0.094	1.000	2	
5	LMGP01050080	521.3481	Glycerophospholipids [GP]	PC(0:0/18:1(9E))	0	0.094	1.000	1	
5	LMGP01050083	521.3481	Glycerophospholipids [GP]	PC(0:0/18:1(9Z))[U]	0	0.094	1.000	4	
5	LMGP01050082	521.3481	Glycerophospholipids [GP]	PC(0:0/18:1(9Z))	0	0.086	1.000	3	
5	LMGP01050079	521.3481	Glycerophospholipids [GP]	PC(0:0/18:1(6Z))	0	0.083	0.996	5	
5	**LMGP01050033**	**521.3481**	**Glycerophospholipids [GP]**	**PC(18:1(9Z)/0:0)[U]**	**1**	**0.938**	**0.989**	**9**	**2**
5	LMGP01050030	521.3481	Glycerophospholipids [GP]	**PC(18:1(9E)/0:0)**	1	0.936	0.989	6	
5	LMGP01050115	521.3481	Glycerophospholipids [GP]	**PC(18:1(9E)/0:0)[U]**	1	0.935	0.989	11	
5	LMGP01050114	521.3481	Glycerophospholipids [GP]	**PC(18:1(9Z)/0:0)[rac]**	1	0.893	0.989	10	
5	LMGP01050031	521.3481	Glycerophospholipids [GP]	**PC(18:1(9)/0:0)[U]**	1	0.877	0.989	7	
5	LMGP01050032	521.3481	Glycerophospholipids [GP]	**PC(18:1(9Z)/0:0)**	1	0.865	0.989	8	
5	LMGP01050029	521.3481	Glycerophospholipids [GP]	**PC(18:1(6Z)/0:0)**	1	0.924	0.985	12	
5	LMGP01020148	521.3481	Glycerophospholipids [GP]	PC(O-16:1(9E)/2:0)[U]	0	0.015	0.060	13	
5	LMGP01020149	521.3481	Glycerophospholipids [GP]	PC(O-16:1(9Z)/2:0)[U]	0	0.011	0.060	14	
5	LMGP01020147	521.3481	Glycerophospholipids [GP]	PC(O-16:1(11Z)/2:0)	0	0.014	0.058	15	
5	LMGP01030009	521.3481	Glycerophospholipids [GP]	PC(P-16:0/2:0)	0	0.010	0.050	16	
5	LMGP01040061	521.3845	Glycerophospholipids [GP]	PC(O-18:1(9Z)/O-1:0)	0	0.051	0.040	18	
5	LMGP01040058	521.3845	Glycerophospholipids [GP]	PC(O-18:1(9E)/O-1:0)[U]	0	0.049	0.040	17	

							MetFrag		
Test Lipid	LipidMaps ID	Mass	Class	Specie	Correct Candidate	ISIS Score	Score	Rank	Best Ranked Correct ID
6	LMGP01040026	677.5723	Glycerophospholipids [GP]	PC(O-14:0/O-16:0)[U]	0	0.763	1.000	1	
6	LMGP01020011	677.5359	Glycerophospholipids [GP]	PC(O-14:0/15:0)	0	0.946	0.986	2	
6	LMGP02040012	677.5723	Glycerophospholipids [GP]	PE-NMe(O-16:0/O-16:0)	0	0.000	0.983	3	
6	**LMGP01010414**	**677.4996**	**Glycerophospholipids [GP]**	**PC(11:0/17:0)**	**1**	**0.989**	**0.975**	**4**	**3**
6	LMGP01010390	677.4996	Glycerophospholipids [GP]	PC(10:0/18:0)	1	0.989	0.970	5	
6	LMGP01010986	677.4996	Glycerophospholipids [GP]	PC(19:0/9:0)	1	0.980	0.964	6	
6	LMGP02020004	677.5359	Glycerophospholipids [GP]	PE(O-16:0/16:0)[U]	0	0.000	0.903	5	
6	LMGP02010251	677.4996	Glycerophospholipids [GP]	PE(17:0/14:0)[U]	0	0.000	0.888	7	
6	LMSP02010011	677.6686	Sphingolipids [SP]	Cer(d18:1/26:0)	0	0.000	0.836	8	
6	LMSP02020013	677.6686	Sphingolipids [SP]	Cer(d18:0/26:1(17Z))	0	0.000	0.804	9	
7	LMGP01040036	733.6349	Glycerophospholipids [GP]	PC(O-16:0/O-18:0)	0	0.915	1.000	1	
7	LMGP02020011	733.5985	Glycerophospholipids [GP]	PE(O-18:0/18:0)	0	0.000	0.972	2	
7	**LMGP01010397**	**733.5622**	**Glycerophospholipids [GP]**	**PC(10:0/22:0)**	**1**	**0.983**	**0.971**	**3**	**3**
7	LMGP02080006	733.5985	Glycerophospholipids [GP]	PE(18:0/O-18:0)[U]	0	0.000	0.970	4	
7	LMGP01011267	733.5622	Glycerophospholipids [GP]	PC(9:0/23:0)	1	0.995	0.969	5	
7	LMGP02010178	733.5622	Glycerophospholipids [GP]	PE(20:0/15:0)[U]	0	0.000	0.965	6	
7	LMGP02010349	733.5622	Glycerophospholipids [GP]	PE-NMe(17:0/17:0)[U]	0	0.000	0.957	7	
7	LMPK04000006	733.4612	Polyketides [PK]	Erythromycin	0	0.000	0.748	8	
8	LMGP01040051	761.6662	Glycerophospholipids [GP]	PC(O-18:0/O-18:0)[S]	0	0.877	1.000	1	
8	LMGP01020044	761.6298	Glycerophospholipids [GP]	PC(O-16:0/19:0)	0	0.952	0.964	2	
8	LMGP02010344	761.5935	Glycerophospholipids [GP]	PE-NMe(18:0/18:0)	0	0.000	0.949	3	
8	**LMGP01011123**	**761.5935**	**Glycerophospholipids [GP]**	**PC(23:0/11:0)[U]**	**1**	**0.988**	**0.942**	**4**	**4**
8	LMGP01010400	761.5935	Glycerophospholipids [GP]	PC(10:0/24:0)	1	0.987	0.939	5	
8	LMGP02020014	761.6298	Glycerophospholipids [GP]	PE(O-18:0/20:0)	0	0.000	0.875	6	
8	LMGP02010303	761.5935	Glycerophospholipids [GP]	PE(14:0/23:0)[U]	0	0.000	0.846	7	
8	LMGP03010017	761.5207	Glycerophospholipids [GP]	PS(18:1(9Z)/16:0)[S]	0	0.998	0.790	8	
8	LMGP03010007	761.5207	Glycerophospholipids [GP]	PS(16:0/18:1(11Z))	0	0.996	0.785	9	
9	**LMGP01010835**	**785.5935**	**Glycerophospholipids [GP]**	**PC(18:1(10E)/18:1(10E))[U]**	**1**	**0.981**	**1.000**	**1**	**1**
9	**LMGP01010864**	**785.5935**	**Glycerophospholipids [GP]**	**PC(18:1(5E)/18:1(5E))[U]**	**1**	**0.978**	**1.000**	**2**	
9	LMGP01010764	785.5935	Glycerophospholipids [GP]	PC(18:0/18:2(10Z,12Z))	1	0.987	0.983	3	
9	LMGP01010619	785.5935	Glycerophospholipids [GP]	PC(16:0/20:2(11E,14E))[U]	1	0.976	0.949	4	
9	LMGP03010018	785.5207	Glycerophospholipids [GP]	PS(18:1(9Z)/18:2(9Z,12Z))[U]	0	0.994	0.895	5	
9	LMGL05010003	785.6381	Glycerolipids [GL]	N/A	0	0.000	0.808	6	
9	LMSP0501AA21	785.6745	Sphingolipids [SP]	GlcCer(d18:0/22:0)	0	0.000	0.794	7	
9	LMGL00000127	785.6170	Glycerolipids [GL]	DGTA(18:1/22:4(10Z,13Z,16Z,19Z))	0	0.000	0.003	8	
10	LMGP02010086	929.7813	Glycerophospholipids [GP]	PE(24:0/25:0)[U]	0	0.000	1.000	1	
10	LMGP02010085	929.7813	Glycerophospholipids [GP]	PE(23:0/26:0)[U]	0	0.000	0.993	2	
10	**LMGP01011034**	**929.7813**	**Glycerophospholipids [GP]**	**PC(20:0/26:0)**	**1**	**0.002**	**0.965**	**3**	**2**
10	LMGP01011193	929.7813	Glycerophospholipids [GP]	PC(26:0/20:0)[U]	1	0.002	0.963	4	
10	LMPR04000020	929.7320	Prenol Lipids [PR]	bacteriohopane-,32,33,34-triol-35-(N-(9-cyclohexyl-nonanoyl))-glucosamine	0	0.000	0.000	5	
11	**LMGP01010918**	**781.5622**	**Glycerophospholipids [GP]**	**PC(18:2(15E,17E)/18:2(15E,17E))[U]**	**1**	**0.992**	**1.000**	**1**	**1**
11	LMGP01010897	781.5622	Glycerophospholipids [GP]	PC(18:1(9Z)/18:3(6Z,9Z,12Z))[U]	1	0.987	0.901	3	
11	LMGP01010949	781.5622	Glycerophospholipids [GP]	PC(18:3(6Z,9Z,12Z)/18:1(9Z))[U]	1	0.992	0.901	2	
11	LMGP01011049	781.5622	Glycerophospholipids [GP]	PC(20:4(5Z,8Z,11Z,14Z)/16:0)	1	0.988	0.812	5	
11	LMGP01010629	781.5622	Glycerophospholipids [GP]	PC(16:0/20:4(5E,8E,11E,14E))	1	0.991	0.811	4	
11	LMGP01010772	781.5622	Glycerophospholipids [GP]	PC(18:0/18:4(5Z,8Z,11Z,14Z))[U]	1	0.992	0.811	6	
11	LMGP01020081	781.5985	Glycerophospholipids [GP]	PC(O-17:0/20:4(5Z,8Z,11Z,14Z))	0	0.966	0.800	7	
12	**LMGP01010952**	**777.5309**	**Glycerophospholipids [GP]**	**PC(18:3(9E,11E,13E)/18:3(9E,11E,13E))**	**1**	**0.990**	**1.000**	**1**	**1**
12	LMGP01010953	777.5309	Glycerophospholipids [GP]	**PC(18:3(9E,11E,13E)/18:3(9E,11E,13E))[U]**	1	0.992	1.000	2	
12	LMGP01010954	777.5309	Glycerophospholipids [GP]	**PC(18:3(9Z,11E,13E)/18:3(9Z,11E,13E))**	1	0.994	1.000	3	
12	LMGP01010956	777.5309	Glycerophospholipids [GP]	**PC(18:3(9Z,12Z,15Z)/18:3(9Z,12Z,15Z))**	1	0.985	1.000	4	
12	LMGP01010957	777.5309	Glycerophospholipids [GP]	**PC(18:3(9Z,12Z,15Z)/18:3(9Z,12Z,15Z))[U]**	1	0.982	1.000	5	
12	LMGP01010950	777.5309	Glycerophospholipids [GP]	**PC(18:3(6Z,9Z,12Z)/18:3(6Z,9Z,12Z))[U]**	1	0.983	0.995	7	
12	LMGP01010951	777.5309	Glycerophospholipids [GP]	**PC(18:3(8E,10E,12E)/18:3(8E,10E,12E))[U]**	1	0.989	0.995	6	
12	LMGP01010960	777.5309	Glycerophospholipids [GP]	PC(18:4(9E,11E,13E,15E)/18:2(9Z,12Z))	1	0.988	0.888	8	
12	LMGP01010512	777.5309	Glycerophospholipids [GP]	PC(14:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z))	1	0.962	0.812	9	
12	LMGP01010513	777.5309	Glycerophospholipids [GP]	PC(14:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z))[U]	1	0.963	0.812	10	

							MetFrag		
Test Lipid	LipidMaps ID	Mass	Class	Specie	Correct Candidate	ISIS Score	Score	Rank	Best Ranked Correct ID
13	LMGP01011036	841.6561	Glycerophospholipids [GP]	PC(20:1(11E)/20:1(11E))	1	0.989	1.000	1	1
13	LMGP01011038	841.6561	Glycerophospholipids [GP]	PC(20:1(11Z)/20:1(11Z))	1	0.982	1.000	2	
13	LMGP01011039	841.6561	Glycerophospholipids [GP]	PC(20:1(11Z)/20:1(11Z))[U]	1	0.980	1.000	3	
13	LMGP01011042	841.6561	Glycerophospholipids [GP]	PC(20:1(9E)/20:1(9E))	1	0.983	1.000	4	
13	LMGP01011043	841.6561	Glycerophospholipids [GP]	PC(20:1(9Z)/20:1(9Z))	1	0.983	1.000	5	
13	LMGP01011044	841.6561	Glycerophospholipids [GP]	PC(20:1(9Z)/20:1(9Z))[U]	1	0.989	1.000	6	
13	LMGP01011021	841.6561	Glycerophospholipids [GP]	PC(20:0/20:2(11Z,14Z))	1	0.986	0.977	7	
13	LMGP13010001	841.3891	Glycerophospholipids [GP]	CDP-DG(12:0/12:0)	0	0.000	0.830	8	
13	LMSP0501AA25	841.7371	Sphingolipids [SP]	GlcCer(d18:0/26:0)	0	0.000	0.789	9	
14	LMGP01011047	829.5622	Glycerophospholipids [GP]	PC(20:4(5E,8E,11E,14E)/20:4(5E,8E,11E,14E))	1	0.846	1.000	1	1
14	LMGP01011048	829.5622	Glycerophospholipids [GP]	PC(20:4(5E,8E,11E,14E)/20:4(5E,8E,11E,14E))[U]	1	0.842	1.000	2	
14	LMGP01011052	829.5622	Glycerophospholipids [GP]	PC(20:4(5Z,8Z,11Z,14Z)/20:4(5Z,8Z,11Z,14Z))	1	0.841	1.000	3	
14	LMGP01011053	829.5622	Glycerophospholipids [GP]	PC(20:4(5Z,8Z,11Z,14Z)/20:4(5Z,8Z,11Z,14Z))[U]	1	0.838	1.000	4	
14	LMGP01010947	829.5622	Glycerophospholipids [GP]	PC(18:2(9Z,12Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))	1	0.830	0.971	5	
14	LMGP01010948	829.5622	Glycerophospholipids [GP]	PC(18:2(9Z,12Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))[U]	1	0.817	0.971	6	
14	LMGP01011118	829.5622	Glycerophospholipids [GP]	PC(22:6(4Z,7Z,10Z,13Z,16Z,19Z)/18:2(9Z,12Z))[U]	1	0.840	0.964	7	
14	LMGP02010287	829.6561	Glycerophospholipids [GP]	PE(20:0/22:1(13Z))	0	0.000	0.945	8	
14	LMGP02010290	829.6561	Glycerophospholipids [GP]	PE(22:0/20:1(11Z))	0	0.000	0.944	9	
14	LMGP02010293	829.6561	Glycerophospholipids [GP]	PE(18:0/24:1(15Z))	0	0.000	0.935	10	
14	LMGP02010294	829.6561	Glycerophospholipids [GP]	PE(24:0/18:1(9Z))	0	0.000	0.929	11	
15	LMGP01010682	729.5309	Glycerophospholipids [GP]	PC(16:1(9E)/16:1(9E))	1	0.990	1.000	1	1
15	LMGP01010494	729.5309	Glycerophospholipids [GP]	PC(14:0/18:2(11Z,14Z))	1	0.983	0.957	2	
15	LMGP02030004	729.5672	Glycerophospholipids [GP]	PE(P-18:0/18:1(9Z))	0	0.000	0.923	3	
15	LMSP02050008	729.6036	Sphingolipids [SP]	CerP(d18:1/24:0)	0	0.000	0.872	4	
15	LMSP0501AA19	729.6119	Sphingolipids [SP]	GlcCer(d18:0/18:0)	0	0.000	0.772	5	
15	LMST03020541	729.4465	Sterol Lipids [ST]	(6R)-vitamin D3 6,19-[4-{2-(6,7-dimethoxy-4-methyl-3-oxo-3,4-dihydroquinoxalinyl)ethyl}-1,2,4-triazoline-3,5-dione] adduct / (6R)-cholecalciferol 6,19-[4-{2-(6,7-dimethoxy-4-methyl-3-oxo-3,4-dihydroquinoxalinyl)ethyl}-1,2,4-triazoline-3,5-dione] adduct	0	0.000	0.000	6	
15	LMGP02040013	729.5097	Glycerophospholipids [GP]	N/A	0	0.000	0.000	7	
16	LMGP01010571	745.5622	Glycerophospholipids [GP]	PC(16:0/17:1(9Z))	0	0.000	1.000	2	
16	LMGP02010322	745.5622	Glycerophospholipids [GP]	PE-NMe2(16:0/18:1(9Z))	0	0.000	1.000	1	
16	LMGP01020089	745.5985	Glycerophospholipids [GP]	PC(O-18:0/16:1(9Z))	0	0.000	0.999	3	
16	LMGP01020036	745.5985	Glycerophospholipids [GP]	PC(O-16:0/18:1(9E))[U]	0	0.000	0.996	5	
16	LMGP01020152	745.5985	Glycerophospholipids [GP]	PC(O-18:1(9Z)/16:0)	0	0.000	0.996	4	
16	LMGP02010051	745.5622	Glycerophospholipids [GP]	PE(18:1(9E)/18:0)[U]	1	0.995	0.860	6	4
16	LMGP02080008	745.5046	Glycerophospholipids [GP]	N/A	0	0.986	0.859	7	
16	LMGP02010312	745.5622	Glycerophospholipids [GP]	PE(16:0/20:1(11Z))	1	0.990	0.835	8	
17	LMGP01010383	579.3900	Glycerophospholipids [GP]	PC(10:0/11:0)[U]	0	0.000	1.000	1	
17	LMGP01020116	579.4264	Glycerophospholipids [GP]	PC(O-18:0/4:0)	0	0.000	0.976	2	
17	LMGP01020123	579.4264	Glycerophospholipids [GP]	PC(O-20:0/2:0)	0	0.000	0.973	3	
17	LMGP02010098	579.3900	Glycerophospholipids [GP]	PE(12:0/12:0)	1	0.989	0.907	4	2
17	LMGP02010245	579.3900	Glycerophospholipids [GP]	PE(13:0/11:0)[U]	1	0.994	0.888	5	
17	LMGP02010265	579.3900	Glycerophospholipids [GP]	PE(10:0/14:0)[U]	1	0.991	0.880	6	
17	LMST05020030	579.2536	Sterol Lipids [ST]	Taurochenodeoxycholic acid 7-sulfate	0	0.000	0.816	7	
17	LMGP01050053	579.4264	Glycerophospholipids [GP]	PC(22:0/0:0)	0	0.000	0.810	8	
18	LMGP02010316	663.4839	Glycerophospholipids [GP]	PE-NMe2(14:0/14:0)	0	0.000	1.000	1	
18	LMGP01010434	663.4839	Glycerophospholipids [GP]	PC(12:0/15:0)[U]	0	0.000	0.989	2	
18	LMGP01010389	663.4839	Glycerophospholipids [GP]	PC(10:0/17:0)[U]	0	0.000	0.980	3	
18	LMGP01010834	663.4839	Glycerophospholipids [GP]	PC(18:0/9:0)[U]	0	0.000	0.973	4	
18	LMGP02040004	663.5567	Glycerophospholipids [GP]	PE(O-16:0/O-16:0)	0	0.918	0.884	5	
18	LMGP02010234	663.4839	Glycerophospholipids [GP]	PE(13:0/17:0)[U]	1	0.980	0.862	6	4
18	LMGP02010022	663.4839	Glycerophospholipids [GP]	PE(18:0/12:0)[U]	1	0.979	0.850	7	
18	LMSP02010013	663.6529	Sphingolipids [SP]	Cer(d18:1/25:0)	0	0.000	0.806	8	
19	LMGP02010108	687.4839	Glycerophospholipids [GP]	PE(16:1(9Z)/16:1(9Z))	1	0.995	1.000	2	1
19	LMGP02010019	687.4839	Glycerophospholipids [GP]	PE(16:1(9Z)/16:1(9Z))[U]	1	0.993	1.000	1	
19	LMGP02010354	687.4839	Glycerophospholipids [GP]	PE(16:1(11Z)/16:1(11Z))	1	0.990	1.000	3	
19	LMGP02010356	687.4839	Glycerophospholipids [GP]	PE(16:1(5Z)/16:1(5Z))	1	0.995	0.999	4	
19	LMPK04000007	687.4194	Polyketides [PK]	Oleandomycin	0	0.000	0.798	5	

							MetFrag		
Test Lipid	LipidMaps ID	Mass	Class	Specie	Correct Candidate	ISIS Score	Score	Rank	Best Ranked Correct ID
20	LMGP01010534	717.5309	Glycerophospholipids [GP]	PC(15:0/16:1(7Z))[U]	0	0.000	1.000	1	
20	LMGP01010002	717.5309	Glycerophospholipids [GP]	PC(16:0/15:1(14))	0	0.000	0.999	2	
20	LMGP01020016	717.5672	Glycerophospholipids [GP]	PC(O-14:0/18:1(9Z))	0	0.000	0.998	3	
20	LMGP02010010	717.5309	Glycerophospholipids [GP]	PE(16:0/18:1(11Z))	1	0.993	0.961	4	3
20	LMGP02010099	717.5309	Glycerophospholipids [GP]	PE(18:1(9Z)/16:0)	1	0.989	0.960	5	
20	LMPK04000012	717.4663	Polyketides [PK]	Erythromycin B	0	0.000	0.772	6	
22	LMGP01060007	453.3219	Glycerophospholipids [GP]	PC(O-14:0/0:0)	0	0.463	1.000	1	
22	LMGP01050001	453.2855	Glycerophospholipids [GP]	PC(13:0/0:0)	0	0.295	0.999	2	
22	LMGP02050002	453.2855	Glycerophospholipids [GP]	PE(16:0/0:0)	1	0.976	0.977	3	3
22	LMPK04000035	453.3090	Polyketides [PK]	10-Deoxymethymycin	0	0.237	0.821	4	
22	LMGP01040017	453.3219	Glycerophospholipids [GP]	PC(O-12:0/O-2:0)	0	0.000	0.273	5	
22	LMGP01040020	453.3219	Glycerophospholipids [GP]	PC(O-13:0/O-1:0)[U]	0	0.000	0.272	6	
22	LMGP01011229	453.2492	Glycerophospholipids [GP]	PC(6:0/6:0)	0	0.000	0.014	7	
22	LMGP01040080	453.3219	Glycerophospholipids [GP]	PC(O-7:0/O-7:0)	0	0.000	0.000	8	
23	LMGP01070006	479.3376	Glycerophospholipids [GP]	PC(P-16:0/0:0)	0	0.609	1.000	1	
23	LMGP02050004	479.3012	Glycerophospholipids [GP]	PE(18:1(9Z)/0:0)	1	0.970	0.977	2	2
23	LMGP02050006	479.3012	Glycerophospholipids [GP]	PE(18:1(9Z)/0:0)[U]	1	0.954	0.977	3	
23	LMPK11000002	479.2672	Polyketides [PK]	Cytochalasin B	0	0.008	0.830	4	
24	LMGP03010028	679.4424	Glycerophospholipids [GP]	PS(14:0/14:0)	1	0.998	1.000	2	1
24	LMGP03010009	679.4424	Glycerophospholipids [GP]	PS(14:0/14:0)[U]	1	0.994	1.000	1	
25	LMGP03010015	623.3798	Glycerophospholipids [GP]	PS(12:0/12:0)[U]	1	0.991	1.000	1	1
25	LMGP03010027	623.3798	Glycerophospholipids [GP]	PS(12:0/12:0)	1	0.995	1.000	2	
26	LMGP03010014	787.5363	Glycerophospholipids [GP]	PS(18:1(9E)/18:1(9E))[U]	1	0.995	1.000	1	1
26	LMGP03010031	787.5363	Glycerophospholipids [GP]	PS(18:0/18:2(9Z,12Z))	1	0.994	0.922	3	
26	LMGP03010032	787.5363	Glycerophospholipids [GP]	PS(18:2(9Z,12Z)/18:0)[U]	1	0.993	0.922	2	
26	LMGP01010749	787.6091	Glycerophospholipids [GP]	PC(18:0/18:1(11E))[U]	0	0.986	0.347	4	
26	LMGP01010840	787.6091	Glycerophospholipids [GP]	PC(18:1(11Z)/18:0)	0	0.992	0.347	5	
26	LMGP01011037	787.6091	Glycerophospholipids [GP]	PC(20:1(11Z)/16:0)[U]	0	0.991	0.340	6	
26	LMGP01010618	787.6091	Glycerophospholipids [GP]	PC(16:0/20:1(11Z))[U]	0	0.991	0.339	7	
26	LMGP02010040	787.5152	Glycerophospholipids [GP]	PE(20:4(5Z,8Z,11Z,14Z)/20:4(5Z,8Z,11Z,14Z))[U]	0	0.000	0.000	9	
26	LMGP01080003	787.5516	Glycerophospholipids [GP]	N/A	0	0.988	0.000	8	
27	LMGP03010976	759.5050	Glycerophospholipids [GP]	PS(16:0/18:2)	1	0.985	1.000	1	1
27	LMGP01010679	759.5778	Glycerophospholipids [GP]	PC(16:1(7Z)/18:0)[U]	0	0.988	0.480	2	
27	LMGP01010874	759.5778	Glycerophospholipids [GP]	PC(18:1(9E)/16:0)[U]	0	0.983	0.477	3	
27	LMGP01010575	759.5778	Glycerophospholipids [GP]	PC(16:0/18:1(11E))	0	0.985	0.473	4	
27	LMGP01010744	759.5778	Glycerophospholipids [GP]	PC(18:0/16:1(9Z))	0	0.984	0.470	5	
27	LMGP01020077	759.6142	Glycerophospholipids [GP]	PC(O-17:0/18:1(9Z))	0	0.927	0.249	6	
27	LMGP02040017	759.5567	Glycerophospholipids [GP]	N/A	0	0.000	0.000	7	
28	LMGP01010402	789.6248	Glycerophospholipids [GP]	PC(10:0/26:0)[U]	0	0.988	1.000	1	
28	LMGP02020017	789.6611	Glycerophospholipids [GP]	PE(O-18:0/22:0)	0	0.000	0.893	2	
28	LMGP02010214	789.6248	Glycerophospholipids [GP]	PE(22:0/17:0)[U]	0	0.000	0.889	3	
28	LMGP03010025	789.5520	Glycerophospholipids [GP]	PS(18:0/18:1(9Z))	1	0.994	0.846	5	3
28	LMGP03010034	789.5520	Glycerophospholipids [GP]	PS(18:1(9Z)/18:0)	1	0.997	0.846	4	
28	LMGP01010974	789.6248	Glycerophospholipids [GP]	PC(19:0/17:0)[U]	0	0.993	0.300	6	
28	LMGP01020080	789.6611	Glycerophospholipids [GP]	PC(O-17:0/20:0)	0	0.975	0.298	7	
28	LMGP01010468	789.6248	Glycerophospholipids [GP]	PC(13:0/23:0)	0	0.982	0.280	8	
29	LMGP03050006	525.3067	Glycerophospholipids [GP]	PS(18:0/0:0)	1	0.984	1.000	2	1
29	LMGP03050003	525.3067	Glycerophospholipids [GP]	PS(18:0/0:0)[U]	1	0.979	1.000	1	
29	LMPK04000038	525.3302	Polyketides [PK]	Pikromycin	0	0.815	0.828	3	
30	LMGP03050002	497.2754	Glycerophospholipids [GP]	PS(16:0/0:0)	1	0.979	1.000	1	1
31	LMGP02010054	551.3587	Glycerophospholipids [GP]	PE(10:0/12:0)[U]	0	0.000	1.000	1	
31	LMGP02010225	551.3587	Glycerophospholipids [GP]	PE(12:0/10:0)[U]	0	0.000	0.992	2	
31	LMGP02010282	551.3587	Glycerophospholipids [GP]	PE(9:0/13:0)[U]	0	0.000	0.981	3	
31	None	551.5277	Sphingolipids [SP]	CER d18:1/17:0 CER	1	0.974	0.944	4	2
31	LMGP01020094	551.3951	Glycerophospholipids [GP]	PC(O-18:0/2:0)	0	0.000	0.184	7	
31	LMGP01020072	551.3951	Glycerophospholipids [GP]	PC(O-16:0/4:0)	0	0.000	0.158	5	
31	LMGP01010715	551.3587	Glycerophospholipids [GP]	PC(17:0/2:0)[S]	0	0.000	0.151	6	
31	LMGP01040070	551.4315	Glycerophospholipids [GP]	PC(O-20:0/O-1:0)	0	0.000	0.002	8	

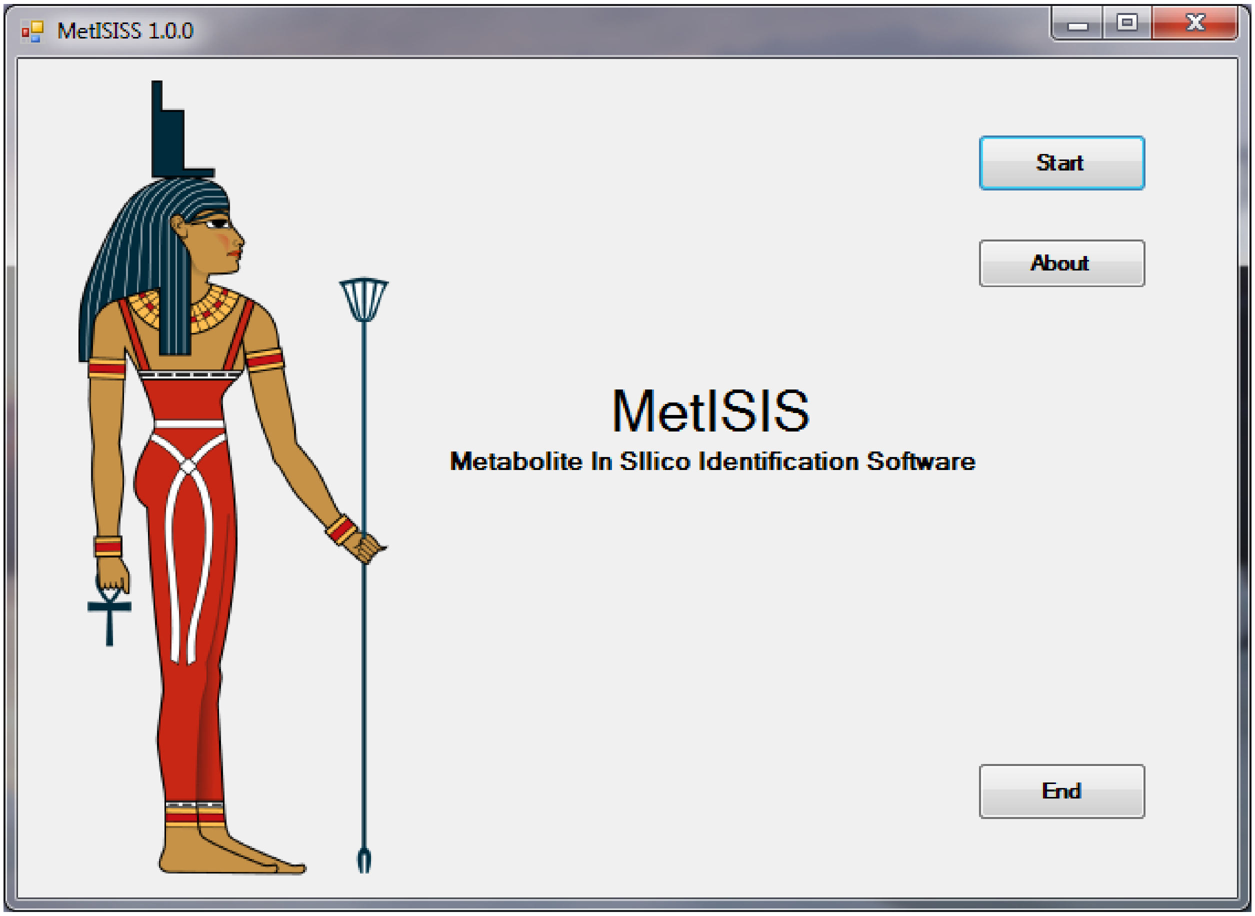
							MetFrag		
Test Lipid	LipidMaps ID	Mass	Class	Specie	Correct Candidate	ISIS Score	Score	Rank	Best Ranked Correct ID
32	LMSP02010006	565.5434	Sphingolipids [SP]	Cer(d18:1/18:0)	1	0.975	1.000	1	1
32	LMSP02020015	565.5434	Sphingolipids [SP]	Cer(d18:0/18:1(9Z))	1	0.995	0.690	2	
32	LMGP01020118	565.4107	Glycerophospholipids [GP]	PC(O-19:0/2:0)	0	0.000	0.580	3	
32	LMGP01010779	565.3744	Glycerophospholipids [GP]	PC(18:0/2:0)	0	0.000	0.519	4	
32	LMGP02010283	565.3744	Glycerophospholipids [GP]	PE(9:0/14:0)[U]	0	0.000	0.245	5	
32	LMGP01010380	565.3744	Glycerophospholipids [GP]	PC(10:0/10:0)	0	0.000	0.000	6	
32	LMGP01040009	565.4471	Glycerophospholipids [GP]	PC(O-10:0/O-12:0)[U]	0	0.000	0.000	10	
32	LMGP01040076	565.4471	Glycerophospholipids [GP]	PC(O-6:0/O-16:0)[U]	0	0.000	0.000	9	
32	LMGP02010266	565.3744	Glycerophospholipids [GP]	PE(10:0/13:0)[U]	0	0.000	0.000	8	
32	LMGP02010348	565.3744	Glycerophospholipids [GP]	PE-NMe(11:0/11:0)	0	0.000	0.000	7	
33	LMGP02010341	593.4057	Glycerophospholipids [GP]	PE-NMe(12:0/12:0)	0	0.000	1.000	1	
33	LMGP02010235	593.4057	Glycerophospholipids [GP]	PE(14:0/11:0)[U]	0	0.000	0.966	2	
33	LMGP02010264	593.4057	Glycerophospholipids [GP]	PE(10:0/15:0)[U]	0	0.000	0.964	3	
33	LMSP02010007	593.5747	Sphingolipids [SP]	Cer(d18:1/20:0)	1	0.976	0.955	4	3
33	LMPK12120407	593.1745	Polyketides [PK]	Tinctormine	0	0.653	0.767	5	
33	LMGP01010384	593.4057	Glycerophospholipids [GP]	PC(10:0/12:0)[U]	0	0.000	0.250	6	
33	LMGP01011220	593.4057	Glycerophospholipids [GP]	PC(4:0/18:0)	0	0.000	0.244	7	
33	LMGP01010831	593.4057	Glycerophospholipids [GP]	PC(18:0/4:0)[U]	0	0.000	0.243	8	
33	LMGP01040015	593.4784	Glycerophospholipids [GP]	PC(O-12:0/O-12:0)	0	0.000	0.126	9	
33	LMGP01040077	593.4784	Glycerophospholipids [GP]	PC(O-6:0/O-18:0)[U]	0	0.000	0.119	10	
34	LMGP02040010	621.5097	Glycerophospholipids [GP]	PE-NMe(O-14:0/O-14:0)	0	0.000	1.000	1	
34	LMGP02010059	621.4370	Glycerophospholipids [GP]	PE(13:0/14:0)[U]	0	0.000	0.881	2	
34	LMGP02010204	621.4370	Glycerophospholipids [GP]	PE(14:0/13:0)[U]	0	0.000	0.873	3	
34	LMGP02010241	621.4370	Glycerophospholipids [GP]	PE(10:0/17:0)[U]	0	0.000	0.845	4	
34	LMSP02010008	621.6060	Sphingolipids [SP]	Cer(d18:1/22:0)	1	0.989	0.814	5	3
34	LMGP01010386	621.4370	Glycerophospholipids [GP]	PC(10:0/14:0)[U]	0	0.000	0.282	6	
34	LMGP01011228	621.4370	Glycerophospholipids [GP]	PC(6:0/18:0)	0	0.000	0.266	7	
35	LMGP02010343	649.4683	Glycerophospholipids [GP]	PE-NMe(14:0/14:0)	0	0.000	1.000	1	
35	LMGP02010298	649.4683	Glycerophospholipids [GP]	PE(16:0/13:0)[U]	0	0.000	0.888	2	
35	LMGP02010228	649.4683	Glycerophospholipids [GP]	PE(11:0/18:0)[U]	0	0.000	0.874	3	
35	LMSP02010012	649.6373	Sphingolipids [SP]	Cer(d18:1/24:0)	1	0.987	0.859	4	3
35	LMSP02020011	649.6373	Sphingolipids [SP]	Cer(d18:0/24:1(15Z))	1	0.987	0.827	5	
35	LMGP01010388	649.4683	Glycerophospholipids [GP]	PC(10:0/16:0)	0	0.000	0.344	6	
35	LMGP01011243	649.4683	Glycerophospholipids [GP]	PC(8:0/18:0)	0	0.000	0.336	7	
35	LMGP01010833	649.4683	Glycerophospholipids [GP]	PC(18:0/8:0)[U]	0	0.000	0.334	8	
35	LMGP01040024	649.5410	Glycerophospholipids [GP]	PC(O-14:0/O-14:0)	0	0.000	0.203	9	
36	LMGP01010494	729.5309	Glycerophospholipids [GP]	PC(14:0/18:2(11Z,14Z))	0	0.000	1.000	1	
36	LMSP02050008	729.6036	Sphingolipids [SP]	CerP(d18:1/24:0)	1	0.975	0.926	2	2
36	LMGP02040013	729.5097	Glycerophospholipids [GP]	N/A	0	0.000	0.909	3	
36	LMSP0501AA19	729.6119	Sphingolipids [SP]	GlcCer(d18:0/18:0)	0	0.427	0.823	4	
36	LMST03020542	729.4465	Sterol Lipids [ST]	(6S)-vitamin D3 6,19-[4-{2-(6,7-dimethoxy-4-methyl-3-oxo-3,4-dihydroquinoxalinyl)ethyl}-1,2,4-triazoline-3,5-dione] adduct / (6S)-cholecalciferol 6,19-[4-{2-(6,7-dimethoxy-4-methyl-3-oxo-3,4-dihydroquinoxalinyl)ethyl}-1,2,4-triazoline-3,5-dione] adduct	0	0.001	0.820	5	
37	None	643.4941	Sphingolipids [SP]	CerP(18:1/18:1)	1	0.980	1.000	1	1
37	LMSP0501AA01	643.5023	Sphingolipids [SP]	GlcCer(d18:1/12:0)	0	0.827	0.000	2	
38	LMSP02050002	617.4784	Sphingolipids [SP]	CerP(d18:1/16:0)	1	0.955	1.000	1	1
39	None	505.3532	Sphingolipids [SP]	CerP(d18:1/8:0)	1	0.978	1.000	1	1
39	LMGP01070012	505.3532	Glycerophospholipids [GP]	PC(P-18:1(9Z)/0:0)	0	0.855	0.957	2	
40	LMSP03010029	728.5832	Sphingolipids [SP]	SM(d18:1/18:1(9Z))	1	0.930	1.000	1	1
40	LMGP04110001	728.4992	Glycerophospholipids [GP]	N/A	0	0.000	0.931	2	
40	LMGL02010284	728.6319	Glycerolipids [GL]	DG(22:0/22:4(7Z,10Z,13Z,16Z)/0:0)[iso2]	0	0.004	0.872	3	
40	LMGL02010291	728.6319	Glycerolipids [GL]	DG(22:2(13Z,16Z)/22:2(13Z,16Z)/0:0)	0	0.004	0.000	4	
40	LMGL02010277	728.6319	Glycerolipids [GL]	DG(22:1(13Z)/22:3(10Z,13Z,16Z)/0:0)[iso2]	0	0.004	0.000	3	
41	LMSP03010044	716.5832	Sphingolipids [SP]	SM(18:1/17:0)	1	0.868	1.000	1	1
41	LMPR01070130	716.5016	Prenol Lipids [PR]	Rhodopin beta-D-glucoside/ Rhodopin glucoside	0	0.002	0.828	2	
41	LMPR01070168	716.5016	Prenol Lipids [PR]	1'-OH-gamma-carotene glucoside/ (Carotenoids B-G)	0	0.001	0.814	3	
41	LMGL02010304	716.5380	Glycerolipids [GL]	DG(22:4(7Z,10Z,13Z,16Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z)/0:0)[iso2]	0	0.003	0.000	4	
43	LMSP0501AC01	699.5649	Sphingolipids [SP]	N/A	0	0.666	1.000	1	
43	LMSP0501AA03	699.5649	Sphingolipids [SP]	GlcCer(d18:1/16:0)	1	0.981	0.997	2	2

Test Lipid	LipidMaps ID	Mass	Class	Specie	Correct Candidate	ISIS Score	MetFrag		
							Score	Rank	Best Ranked Correct ID
44	LMFA07050029	809.1258	Fatty Acyls [FA]	Acetyl-CoA	0	0.091	1.000	1	
44	LMSP0501AC07	809.6745	Sphingolipids [SP]	N/A	0	0.630	0.949	2	
44	**LMSP0501AA08**	**809.6745**	**Sphingolipids [SP]**	**GlcCer(d18:1/24:1(15Z))**	**1**	**0.991**	**0.948**	**3**	**3**
44	LMGP01010904	809.5935	Glycerophospholipids [GP]	PC(18:1(9Z)/20:3(5Z,8Z,11Z))	0	0.000	0.497	4	
44	LMGP01010642	809.5935	Glycerophospholipids [GP]	PC(16:0/22:4(7Z,10Z,13Z,16Z))	0	0.000	0.251	5	
44	LMGP01010801	809.5935	Glycerophospholipids [GP]	PC(18:0/20:4(5E,8E,11E,14E))[U]	0	0.000	0.221	6	
44	LMGP01040096	809.6662	Glycerophospholipids [GP]	N/A	0	0.000	0.088	7	
45	**LMSP0501AD05**	**973.7429**	**Sphingolipids [SP]**		**1**	**0.497**	**1.000**	**2**	**1**
45	LMSP0509AA05	973.7429	Sphingolipids [SP]	N/A	1	0.443	1.000	3	
45	LMSP0501AB07	973.7429	Sphingolipids [SP]	N/A	1	0.431	1.000	1	
46	**LMSP0509AA01**	**861.6177**	**Sphingolipids [SP]**	**d18-1_16-0 Di Hex Cer**	**1**	**0.422**	**1.000**	**3**	**1**
46	LMSP0501AB03	861.6177	Sphingolipids [SP]	N/A	1	0.411	1.000	1	
46	LMSP0501AD01	861.6177	Sphingolipids [SP]	N/A	1	0.343	1.000	2	
46	LMPK05000001	861.5086	Polyketides [PK]	Erythromycin ethylsuccinate	0	0.051	0.858	4	
46	LMGP01011028	861.6248	Glycerophospholipids [GP]	PC(20:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z))	0	0.000	0.396	5	

SUPPLEMENT

MetISIS shareware software

The MetISIS application has been placed on an open access website http://omics.pnl.gov/ for all to down load and use for their own lipid metabolite data.


MetISISS 1.0.0
Start
About
MetISIS
Metabolite In SIlico Identification Software
End