MyCompoundID MS/MS Search and DnsID through Web-based Applications

by

Chenqu Tang

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

Master of Science

Department of Computing Science

University of Alberta

©Chenqu Tang, 2015

Abstract

The MS/MS spectrum of a compound can be manually interpreted to understand its structure. However, given the fact that manual interpretation of fragmentation spectra is time-intensive and often impractical, libraries containing spectra information have been developed to provide reliable source of metabolite identification.

In this work, by applying in silico fragmentation approach, a predicted MS/MS spectrum of a compound was created by compiling a list of fragment ions generated based on chemical bond cleavage of the compound structure. We develop a MS/MS search program which allows a user to search an experimental MS/MS data against our MyCompoundID database which contains 383,000 simulated MS/MS compounds for spectral match. A search program, DnsID, has been developed in MyCompoundID for automated identification of dansyl labeled metabolites. These methods allow user to narrow down the candidate-list which generated from MS search into one or a few unique structures.

Acknowledgements

This thesis concludes the work in my MS.C life for the past two year. Here I would like to take this opportunity to express my sincere appreciation to my supervisors and coworkers.

First, I'm deeply grateful to my supervisors, Dr. Guohui Lin and Dr. Liang Li, for their assistance and suggestion during my studies. This thesis would not have been possible without the help, support and patience of my principal supervisor, Prof. Guohui Lin, not to mention his advice and unsurpassed knowledge of Computer Science and Bioinformatics. The good advice, support and friendship of my second supervisor, Prof. Liang Li, has been invaluable on both an academic and a personal level, for which I am extremely grateful.

I would also like to thank Ronghong Li, who helps me start the MyCompoundID project and implement certainly great data structures. Tao Huan, a PhD student in Dr. Liang Li's group, who helps me dig out metabolic chopping reaction rules and gives me many constructive suggestions. Without their collaboration, the implementation of MyCompoundID MS/MS search program is impossible.

Finally, I would like to thank my family for their unconditional love and support during the last two years. I would not have been able to complete this thesis without their continuous love and encouragement.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Metabolomics is the scientific study of chemical processes involving metabolites. Specifically, metabolomics is the systematic study of the unique chemical fingerprints that specific cellular processes leave behind, the study of their small-molecule metabolite profiles [1]. The metabolome refers to the complete set of small-molecule chemicals found within a biological sample. Also, it represents the collection of all metabolites in a biological cell, tissue, organ or organism, which are the end products of cellular processes [5]. Since the metabolic profiles are context dependent, and vary in response to a variety of factors including environment and environmental stimuli, health status, disease and a myriad of other factors, it can be used to detect the physiological changes caused by toxic insult of a chemical, and reveal the set of gene products being produced in the cell, data that can represent one aspect of cellular function. When combined with genomic transcriptomic and proteomic studies, metabolomics can also help in interpretation and understanding of many complex biological processes. Indeed, metabolomics is now widely recognized as being a cornerstone to all of systems biology [9].

There are many detection methods used in metabolomics. The most widely used techniques are Mass spectrometry (MS) and Nuclear magnetic resonance (NMR). Mass spectrometry (MS)-based metabolomics has been developed quite dramatically in the past decades. However, the metabolite identification of the MS data is still a bottleneck. MS coupled with chromatographic separation techniques, is a key analytical approach for high-throughput analysis of small molecules. To achieve the structural information in MS experiment, collision-induced-ionization has been developed to break down a compound and record the compound's fragments. The matching of experimental MS/MS spectrum with a reference MS/MS spectrum provides confidence of the compound identification.

1.1 MS/MS Spectra

1.1.1 Introduction

MS/MS spectrometry is both the science and the art of displaying the spectra of the mass and structure information of a sample of material. The x-axis of MS/MS spectrum represents a relationship between the mass of a given ion and the number of elementary charges that it carries. This is written as the IUPAC standard m/z to denote the quantity formed by dividing the mass of an ion in the unified atomic mass unit and by charge number. The y-axis of the spectral represents signal intensities of the ions. MS/MS spectrum is produced using a tandem mass spectrometer, each peak represents one fragment's mass and intensity information. MS/MS spectrum also can be manually interpreted, often against a possible chemical structure, to confirm or disapprove a structural assignment. For these reasons, MS/MS spectrum can be used to reveal the structure information of the molecule [8].

1.1.2 Related Work

Metabolite MS/MS spectral library are available from spectral libraries such as MassBank [3], HMDB [16], Metlin [11]. However, the metabolites with reference spectra available are very limited and it turns out to be an issue for the metabolite identification of MS/MS spectra that can't match with any available reference spectra. This problem can be solved by the accurate prediction of compound structure from its experimental MS/MS data.

There are several approaches to achieve the goal of computational MS/MS annotation. The first approach is the rule-based fragmentation spectrum prediction. This method works based on the hypothesis that by applying fragmentation rules to chemical structures, it is possible to generate predicted spectra which is similar to their real MS/MS spectra. In practice, such rules are curated from MS literature.

Nowadays, there are two major commercial tools that predict MS fragmentation based on rules: Mass Frontier (Thermo Scientific, Waltham, USA), ACD/MS Fragmenter (Advanced Chemistry Labs, Toronto, Canada). However, these mentioned commercial softwares haven't published their algorithmic details. The second approach is combinatorial fragmentation. This approach aims at explaining the peaks in a measured spectrum rather than simulates the fragmentation spectrum of a given compound in the rule-based fragmentation. Metfrag is an example of this combinatorial fragmentation approach. It can be applied to a metabolite database to find the compound that best explains the experimental spectrum [10].

1.2 Retention Time for Metabolite Identification

Chemical isotope labeling (CIL) liquid chromatography mass spectrometry (LC-MS) is an enabling technology that provides accurate quantification of metabolites with high submetabolome coverage. It is based on rational design of the labeling reagents to target a class of metabolites sharing the same functional group to improve metabolite separation, detection and quantification [2]. Metabolite identification remains to be one of the major analytical challenges. The first path of metabolite identification often involves the search of accurate mass and MS/MS spectrum of a given peak against a compound library for possible match [14]. Several compound libraries containing accurate masses and MS/MS spectra information have been developed. One major limitation of this approach is that not all metabolites can produce a sufficient number of fragment ions for library search. On the other hand, using mass search alone can lead to many possible structure candidates.

The interval between the instant of injection and the detection of the component is known as the retention time. Because retention times vary with identity of the component, they can be used to identify component. According to these reasons, retention time (RT) of metabolites can be another important piece of information. However, RT can vary greatly, depending on a number of factors including LC setup [15], column type and elution conditions used, and thus is not commonly used as a search parameter in a publicly available compound library. RT match is often performed at the final stage of confirming a metabolite identity using an authentic standard. By spiking a standard to a sample or running identical LC-MS conditions for the standard and sample, retention time can then be compared [4].

1.3 The Motivation

Metabolomics research has advanced rapidly in the last decade, but the metabolite identification remains analytical challenge. Firstly, all the currently available MS/MS prediction needs huge amount of calculation. The calculation takes quite a bit of time to complete one structure annotation and thus not suitable for LC-MS/MS experiment. Secondly, all the currently available computational MS/MS annotation programs only work on metabolites that are known in the metabolome library. For the potentially existing metabolites that do not exist in the metabolome library, the computational interpretation approach does not work. Last but not the least, there still is no effective and efficient compound identification algorithm to take full advantages of mass spectral features.

For these consideration and to satisfy the practical needs, we first developed an evidence based metabolome database: MyCompoundID [6]. It consisting two sub datasets: 8,300 known human metabolites from HMDB and 383,000 metabolites which generated from the first one after one reaction. We then developed a web-based pipeline of tools to identifying unknown metabolites in metabolome profiling. Our program allows user to upload MS/MS spectral data and search against the MyCompoundID databses for metabolite identification.

Chapter 2

MS/MS Search Construction

In this chapter, we will introduce our MS/MS chopping reaction algorithm and database construction. The paper [1] has been submitted to Analytical Chemistry according to this work.

2.1 Searching Inputs and Outputs

Our search program let the user to input the precursor ion and experimental MS/MS fragmentation peak list. The precursor ion is first to search against the database. A list of candidate molecules with molecular mass the same as the precursor mass would be isolated out. Then, the experimental MS/MS fragmentation peak list is used to compare with the predicted MS/MS peak list of all the candidate molecules in the candidate list. A score is then assigned to each of the comparison to evaluate the similarity between experimental and predicted MS/MS spectra. The work flow is shown in Figure 2.1.

2.2 Preparation for Chopping Algorithm

2.2.1 Chemical Graph

Since our chopping reaction is based on the chemical graph, we give this definition at the very beginning of the chapter. In chemical graph theory, a molecular graph

[1] T. Huan, C. Tang, R. Li, Y. Shi, G. Lin and L. Li. MyCompoundID MS/MS Search: Metabolite Identification Using a Library of Predicted Fragment-Ion-Spectra of 383,000 Possible Human Metabolites. *Analytical Chemistry*

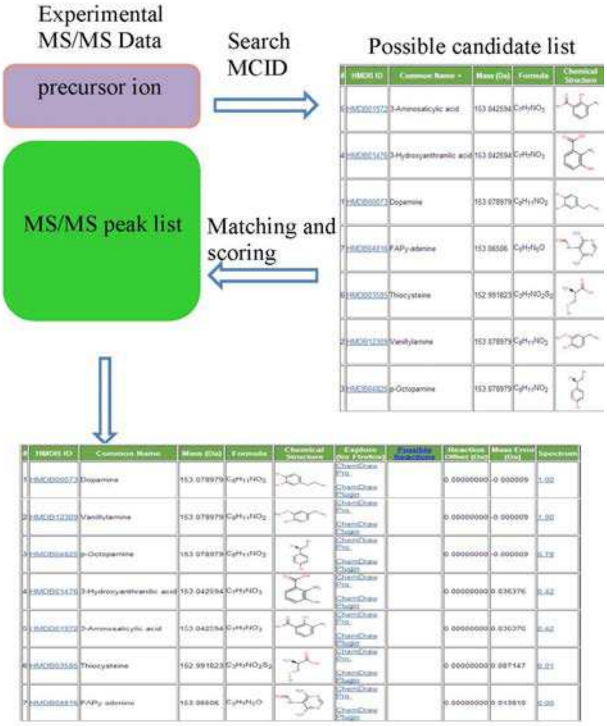


Figure 2.1: Work flow of MS/MS search.

is a representation of the structural formula of a chemical compound in terms of graph theory. A chemical graph is a labeled graph whose vertices correspond to the atoms of the compound and edges correspond to chemical bonds. Its vertices are labeled with the kinds of the corresponding atoms and edges are labeled with the types of bonds. The hydrogen-deleted molecular graph which we used as our data structure is the molecular graph with hydrogen vertices deleted. There is a limitation of this data structure that is this molecular graph does not contain any information about the 3D arrangement of the bonds so that we can not distinguish conformational isomers.

2.2.2 Data Structure and API

According to the chemical graph in the above subsection, we modify the data structure in Chemistry Development Kit [7] to convert chemistry molecule to a special graph structure. The following figures give detail information of our data structure. Our chopping algorithm is based on this data structure.

Atom structure

Usage: Represents the idea of a chemical atom

Normal Constructor: Atom(String elementname)
Constructs an Atom from a String containing an element symbol

This structure have the following important methods:

Function	Returns Type	Description
getMass()	Double	Returns the mass of this atom
getName()	String	Returns the name of this atom
getHCount()	Integer	Returns the hydrogen count of this atom

Figure 2.2: Atom structure.

Bond structure

Usage: Implements the concept of a bond between two atoms

Normal Constructor: Bond(Atom a1, Atom a2,int Order)
Constructs a bond with a given order between two given atoms

This structure have the following important methods:

Function	Return Type	Description
atoms()	Iterable<Atom>	Returns the iterator to atoms making up this bond
contains(Atom a)	Boolean	Returns true if the given atom participates in this bond
getOrder()	integer	Returns the bond order of this bond
setAtom(Atom[] a)	void	Sets an Atom in this bond
setOrder(int o)	void	Sets the bond order of this bond

Figure 2.3: Bond structure.

2.3 Chopping Algorithm

2.3.1 Molecule Ring Detection

In chemical graph, a cyclic compound is a compound in which a series of atoms is connected to form a loop or ring [12]. Since the ring detection is very important in our chopping algorithm, so we give the ring detection algorithm (Algorithm 1) based on the chemical graph data structure described above. Actually, this algorithm can be explained by the following steps which maybe easier to understand:

AtomContainer structure

Usage: Base class for all chemical objects that maintain a list of Atoms and related bonds.

Normal Constructor: AtomContainer()
Constructs an empty AtomContainer.

This structure have the following important methods:

Function	Return Type	Description
addAtom(Atom a)	void	Adds an atom to this container
addBond(int a,int b,int o)	void	Adds a bond to this container.
atoms()	Iterable<Atom>	Returns an Iterable for looping over all atoms in this container
bonds()	Iterable<Bond>	Returns an Iterable for looping over all bonds in this container
contains(Atom a)	boolean	True, if the AtomContainer contains the given atom object
contains(Bond b)	boolean	True, if the AtomContainer contains the given bond object
getBond(int number)	Bond	Get the bond at position number in [0, ..]
getAtom(int number)	Atom	Get the atom at position number in [0, ..]

Figure 2.4: AtomContainer structure.

1. The input atom is the start atom, and if the target is also this atom we can check if this input atom is in the ring.
2. Access all bonds connected to this atom.
3. Check whether we already visited this atom before. If not, add the connected atom to the path.
4. If the connected atom is our target then return true. If not, we can recursively visit the neighbors of this connected atom and try to find our target. If we successfully find the target, then return true, the path still contains all the atoms which lead us from start atom to the target atom.
5. If we can not find the target atom, we remove this atom and bond from the path.
6. After visiting all the neighbor atoms, if we still can not find the target, then we return false. The path should be empty.

Algorithm 1 Ring detection algorithm

```
Input: AtomContainer molecule, Atom atom, Atom target, AtomContainer path.
Output: boolean value whether the atom can connected to the target, and the
path will contain all atoms we visited to get to the target.

function DFS(AtomContainer m, Atom atom, Atom target, AtomContainer path)
atom.visited=true
for Bond b in m.getConnectedBonds(atom) do
    Atom connected = b.getConnectedAtom()
    if connected.visited is false then
        path.addAtom(connected)
        path.addBond(b)
        if connected is target then
            return true;
        else
            if DFS(m, connected, target, path) then
                return true;
            else
                path.removeAtom(connected)
                path.removeBond(b)
            end if
        end if
    end if
end for
return false
```

2.3.2 Benzene Ring Detection

Benzene is a special organic chemical compound with the molecular formula C_6H_6. Its molecule is composed of 6 carbon atoms joined in a ring, with one hydrogen atom attached to each carbon atom. Since the benzene ring has some unique characters, so we need to detect whether the compound has a benzene ring. The following algorithm is used to detect benzene ring (Algorithm 2) and it can can be explained by the following steps:

1. The input is an Atom-Container, using algorithm 1 we achieve all the rings in this molecule.
2. Filter rings obtained by the first step which contain less or more than 6 carbon.
3. Check whether there is any atom in the ring which is not carbon. If not, go to the next step.

4. Go though all the bonds in the ring. If the sum degree is not equal to 9, return false.

5. If the ring satisfies all constraints described above, this ring has a benzene structure.

Algorithm 2 Benzene ring detection algorithm

```
Input: AtomContainer molecule
Output: Whether this molecule has a benzene ring.

function isBenzenering(AtomContainer m)
List<AtomContainer> ringset=m.getRingset().
for AtomContainer container in ringset do
    if container.getatomcount() is not 6 or one of the atoms is not carbon then
        Continue;
    end if
    Set BondSum equals to 0
    Iterable<Bond> iter=container.bond()
    while iter.hasNext() do
        BondSum =BondSum+iter.next().degree
    end while
    if BondSum != 9 then
        return false;
    end if
end for
return true;
```

2.3.3 Chopping Algorithm

First, we will introduce some definition which is related to our chopping algorithm.

Definition 2.3.1. Terminal-atom is an atom which has only one edge connects to another atom.

Definition 2.3.2. Ring-set is a set which contains all the ring structure in a molecule.

Definition 2.3.3. Split-able-bond is a bond which doesn't connect to the terminal-atom and shouldn't be a double bond in the ring-set.

We then proceed to chop the compound by the following steps (Algorithm 3):

Algorithm 3 K-layer chopping algorithm

```
Input: AtomContainer molecule
Output: MS/MS fragments

function Klayerchopping(AtomContainer m)
List<Fragments> fragments.
List<Bond> bondset=m.getSplitBond().
for Atom atom in m.getTerminalatom() do
    if atom is hetero then
        Fragment f = split(m,atom);
        fragments.add(f);
    end if
end for
int K;
if Bondset.size()<=40 then
    K = 4
end if
if Bondset.size()>40 and Bondset.size()<=60 then
    K = 3
else
    K = 2
end if
fragments.add(split(m,bondset,k)).
return  fragments;
```

1. Iterate all the terminal-atoms, if the atom is hetero (O,N,P,etc), then we chop this atom and add these two fragments' mass in our mass-set.

2. Check the size of the split-able-bonds, if it less than 40 we do 4-layer chopping, between 40-60 then 3-layer chopping, more than 60 then 2-layer chopping.

3. Get all the split-able-bonds in the ring-set and the benzene ring-set.

4. Get all the fragments after step 3, we do linear K-layer chopping according to the size of their split-able-bonds.

The reason we use different layers chopping for different compounds is because sometimes the metabolite is huge, it's too time-consuming for us to do the 4-layer chopping and too many fragment ion masses will be created. Figure 2.5 shows an example of ring chopping, while Figure 2.6 shows an example of layer chopping.

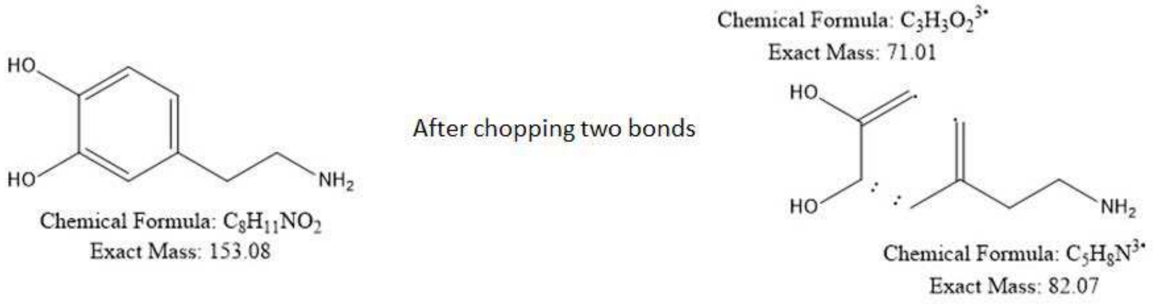

HO
HO
NH2
Chemical Formula: $C_8H_{11}NO_2$
Exact Mass: 153.08
After chopping two bonds
Chemical Formula: $C_3H_3O_2^{3\bullet}$
Exact Mass: 71.01
HO
HO
NH2
Chemical Formula: $C_5H_8N^{3\bullet}$
Exact Mass: 82.07

Figure 2.5: An example of ring chopping.

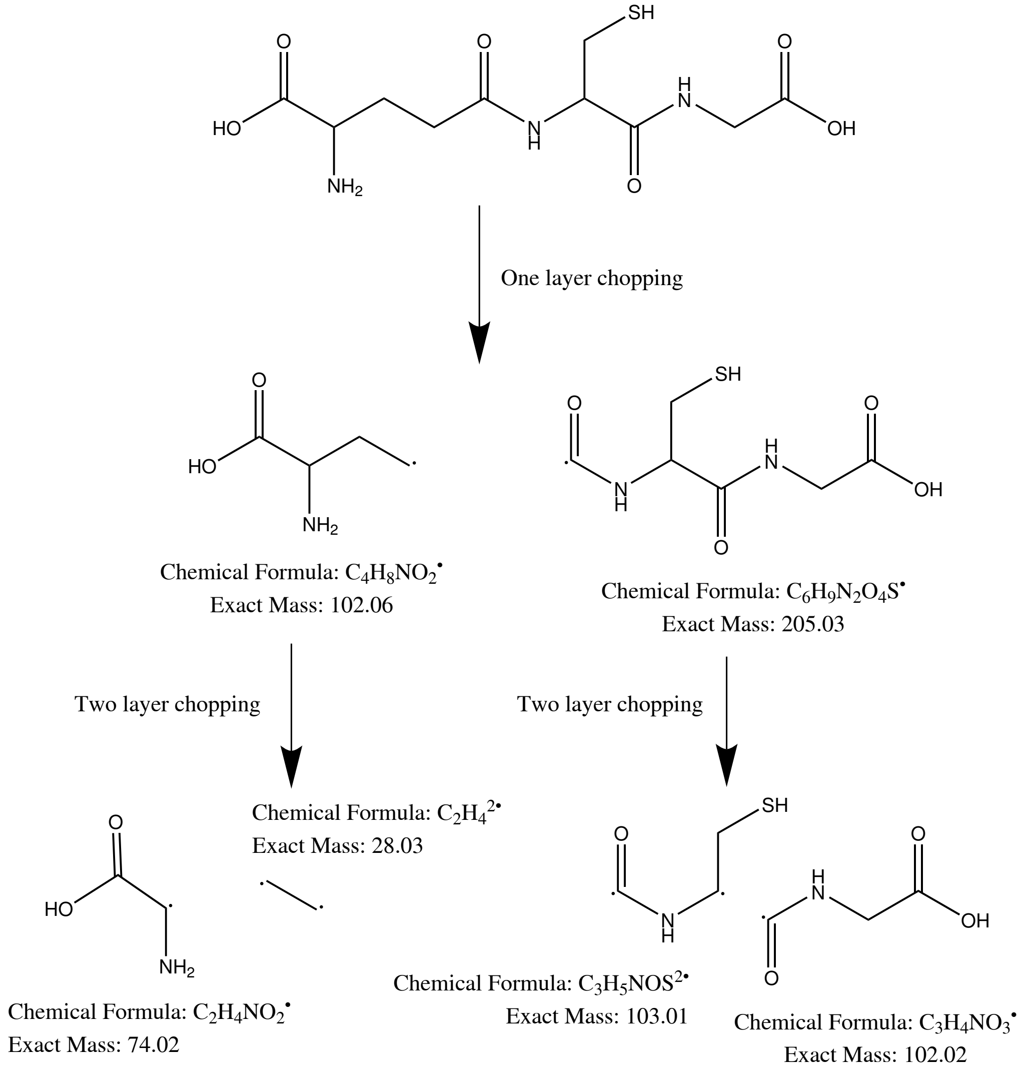

SH
O
O
O
HO
N H
H N
OH
NH2
O
One layer chopping
O
HO
NH2
Chemical Formula: $C_4H_8NO_2^{\bullet}$
Exact Mass: 102.06
SH
O
O
N H
H N
OH
O
Chemical Formula: $C_6H_9N_2O_4S^{\bullet}$
Exact Mass: 205.03
Two layer chopping
Two layer chopping
O
HO
NH2
Chemical Formula: $C_2H_4NO_2^{\bullet}$
Exact Mass: 74.02
Chemical Formula: $C_2H_4^{2\bullet}$
Exact Mass: 28.03
SH
O
N H
Chemical Formula: $C_3H_5NOS^{2\bullet}$
Exact Mass: 103.01
O
H N
OH
O
Chemical Formula: $C_3H_4NO_3^{\bullet}$
Exact Mass: 102.02

Figure 2.6: An example of linear chopping.

2.4 Scoring Scheme

After the chopping algorithm, we can get the predicted MS/MS fragments. An algorithm was developed to gauge the similarity between the experimental MS/MS data and the predicted MS/MS sepectrum. The equation for the comparison is shown below. A weight is calculated for each comparison by the dot product of the matched m/z's and intensities. A m/z tolerance is set to determine if the experimental m/z is matched with the predicted m/z. For providing the optimal match score, the exponent for m/z is 3 and for intensity is 0.6, which was taken from the literature [13]. A score is calculated by normalization against the maximum weight in all the candidates. A higher score indicates a better match between the experimental MS/MS and the predicted MS/MS.

$$
\begin{aligned}
weight_i &= < \overrightarrow{m/z} >^3 \cdot < \overrightarrow{Int} >^{0.6} \\
Socre_i &= \frac{1}{\max(weight)} weight_i \\
< \overrightarrow{m/z} > &: the\ matched\ list\ of\ m/z \\
< \overrightarrow{Int} > &: the\ matched\ list\ of\ intensities \\
i \quad &: index\ number\ of\ compounds\ in\ candidate\ set
\end{aligned}
$$

Besides the match score, a fit score is used to quantify how well the experimental fragmentation are matched to the predicted spectrum. The fit score is defined as:

$$
\begin{aligned}
fit\_score &= \frac{< \overrightarrow{m/z} > \cdot < \overrightarrow{Int} >}{< \overrightarrow{M} > \cdot < \overrightarrow{I} >} \\
< \overrightarrow{m/z} > &: the\ matched\ list\ of\ m/z \\
< \overrightarrow{Int} > &: the\ matched\ list\ of\ intensities \\
< \overrightarrow{M} > &: the\ experimental\ list\ of\ m/z \\
< \overrightarrow{I} > &: the\ experimental\ list\ of\ intensities
\end{aligned}
$$

A higher fit score will be generated if all or most of the experimental fragment ion peaks are explained by the predicted spectrum. The fit score calculation considers the experimental peaks' march quality, while the match score calculation does not. The match score is useful for ranking the mass-matched metabolite candidates and the fit score is useful for judging the quality of a match with the predicted spectrum.

2.5 Database

We use MySQL as database to store our metabolite compounds information. In this section, we will introduce the structure and capacity of our database.

2.5.1 Zero Reaction Database

We have two tables for the zero-reaction database, Table 2.1 has three fields, hmdb_id, mw, ms, while Table 2.2 contains the detailed information such as the chemical formula and common names. We make a connection between the two tables through the foreign key hmdb_id. This table contains 8021 zero-reaction metabolites and the .mol files can be downloaded from HMDB. Figure 2.7 shows the mass distribution of 8021 standard metabolites in our zero reaction database. The x-axis represents the mass in Da while the y-axis represents the number of compounds in the round number.

Table 2.1: Myid_mw table

Name	Type	Usage
hmdb_id	varchar(25)	HMDB.No as a primary key
mw	double$(12, 6)$	molecular mass
ms	blob	molecular MS/MS spectrum

Table 2.2: Myid_detail table

Name	Type	Usage
hmdb_id	varchar(25)	HMDB.No as foreign key
formula	varchar(255)	Proportions of atoms that constitute a particular compound
common_name	varchar(255)	Common Names of Chemical Compounds

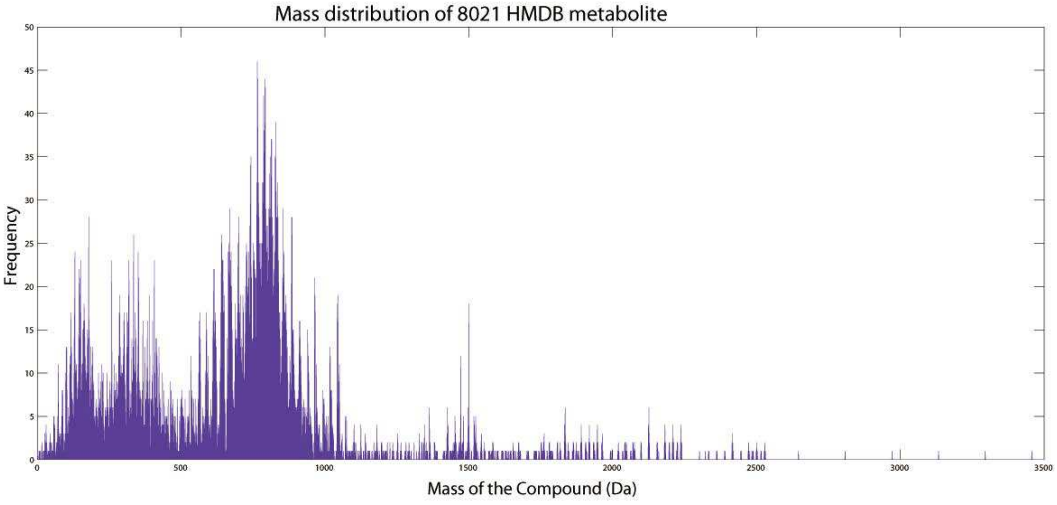


Figure 2.7: Zero reaction 8021 compound mass distribution graph.

2.5.2 One Reaction Database

Based on 76 commonly encountered metabolic reactions, we extended the zero-reaction database to one-reaction database. The structure of the table is shown in Table 2.3, where each table entry is for one compound. This database contains 383,000 one-reaction metabolites. In Table 2.3, a compound has a unique react_id, for example, ”HMDB00003_35” which reveals this compound is the product of HMDB00003 after #35 reaction in the possible reactions table which is in our web-site. The ms column contains all the fragments’ mass which is generated by our chopping algorithm while the isShow reveals whether the compound has been validated by our users and indicates whether the entry is going to be searched against. Figure 2.8 shows the mass distribution of 383,000 one-reaction metabolites in our one reaction database.

Table 2.3: One_reaction table

Name	Type	Null	Key	Default	Usage
react_id	varchar(25)	No	PRI		unique key for the compound
mw	double(12, 6)	No	MUL	Null	molecular mass
ms	medium-blob	Yes		Null	molecular MS/MS spectrum
isShow	tinyint(1)	Yes		0	whether to show this compound

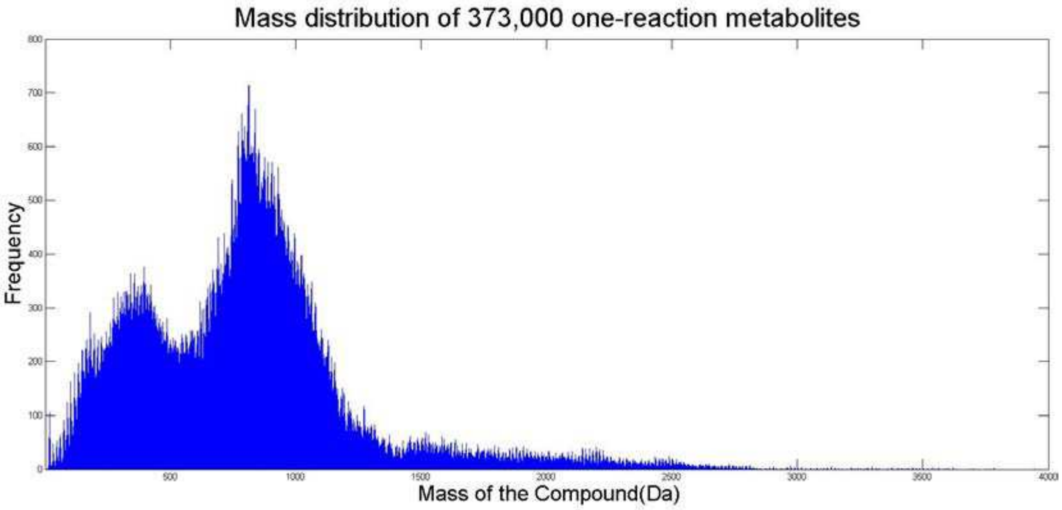


Figure 2.8: 373,000 one-reaction metabolites mass distribution.

2.5.3 Database Indexing

Our databases use HMDB.No and React.ID as primary key and the mass information can be seen as a data in the table. For every mass search, our program should return all candidates that have mass within the mass tolerance range. We compare 3 different database-structures' performance, B-Tree, Hash, Index File which can be implemented by adding a second file associated with a data file of mass information. First, we randomly pick 1500 entries and use their HMDB.No or React.ID to do key search, then we use their exact mass to do mass search, the last step is range search we use 0.005Da as mass tolerance and search the candidates in our database. Table 2.4 and 2.5 show the performance of these database-structures, for zero-reaction and one-reaction respectively.

Table 2.4: Average performance of zero reaction database in millisecond

	B-Tree	Hash	Index File
Key Search	1,563	1,203	803
Data Search	78,156	56,998	605
Range Search	8,886	83,568	3,489

Table 2.5: Average performance of one reaction database in millisecond

	B-Tree	Hash	Index File
Key Search	2,589	1,688	1,896
Data Search	356,911	256,389	1,568
Range Search	268,546	508,129	8,566

According to the tables above, the B-tree and Hash databases performed key and data searches with similar efficiency. However, the Hash database performed range search much slower than the B-tree database. The index file shows the best average performance on all 3 types of searches compared to B-tree and Hash database implementations. The marked performance increase for the data and range search in the index file is due to the implementation of a secondary database in which the key-value pairs of the primary are swapped. Due to the frequent use of range search in our program, the index file is the most appropriate type for our database.

2.5.4 Database Capacity

In this subsection, we will describe capacity of our database. Figure 2.9 shows the capacity for different databases in our program. The x-axis represents the number of reactions while the y-axis represents the capacity on a logarithmic scale in Gigabyte. The MS database is used in MS search, from the graph we can see that the capacity of MS/MS version 2.0 database is much larger than the version 1.0, the reason is that we implement the ring chopping algorithm and the K-layer chopping algorithm which generate more fragments than the old chopping algorithm.

Since database generation is time-consuming and we need large space to store the information, the two-reaction MS/MS database haven't been implemented yet. From the graph, the database capacity for one-reaction MS/MS version 2.0 is about 40.56GB and we have 76 common reactions. According to the above information, we can predict the capacity for two-reaction MS/MS database is about $40GB * 76 \approx$ 3TB.

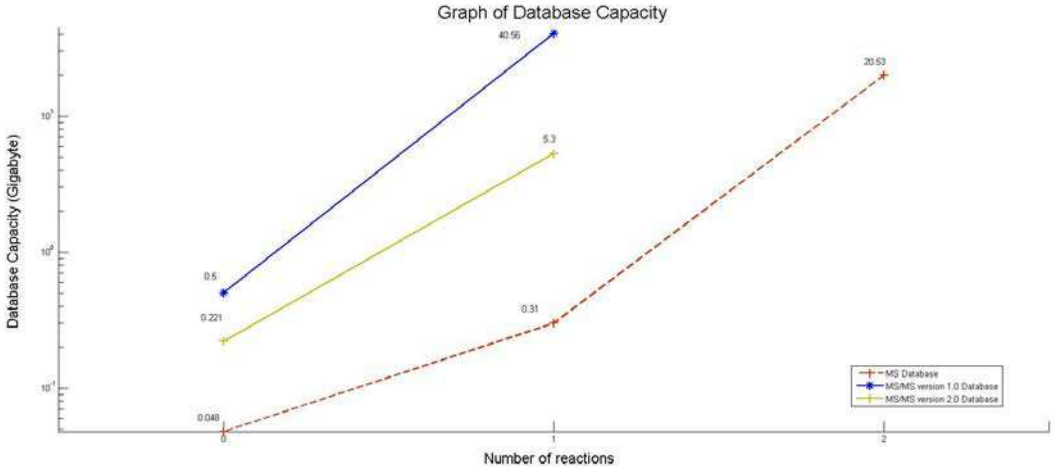


Figure 2.9: Database capacity graph

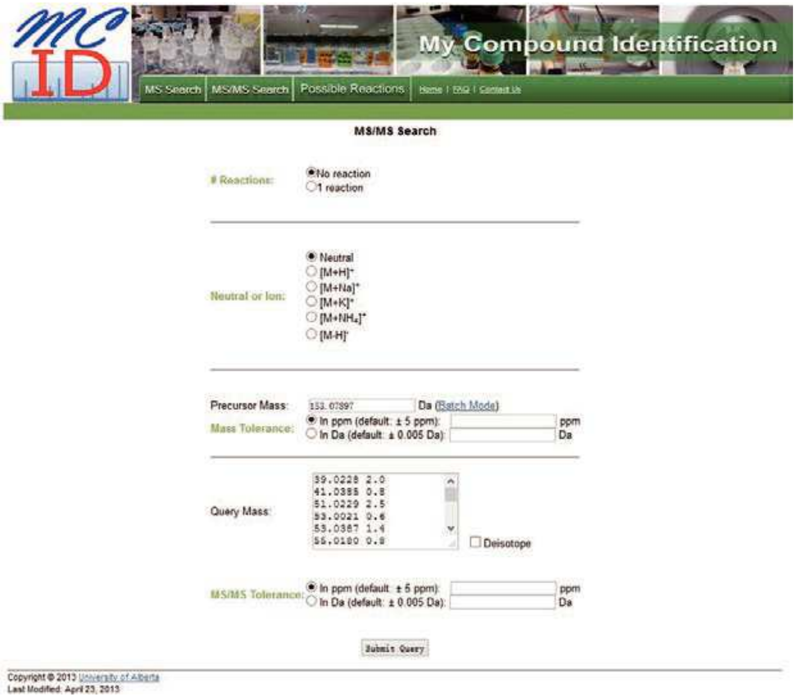
2.6 Web Services

MyCompoundID Version 2.0 was set up for users on May 1, 2015. This chapter presents the main framework design and the display of results.

2.6.1 Search Parameters

In our search page, the user should input 7 parameters for our search program. They are Reaction type, Neutral or ionized, Precursor mass, Mass tolerance type and range, Query mass, MS/MS tolerance type and range and Deisotype option. We put the most common default value for users, while the user can also choose to use our batch search model which allows uploading a csv file containing all the MS/MS information from LC-MS/MS experiments. The following Figure 2.10 and Figure 2.11 show the search interfaces for the these two models respectively.

We then give some detail information about the parameters. Firstly, the reaction type means the user can select either zero-reaction metabolite database or one-reaction metabolite database. Secondly, the user has the option to search for neutral molecules, $[M + H]^+$ ions, $[M - H]^-$ ions and the most common adducts obtained by the ionization process. Then the user should enter a single mass and mass error threshold in either parts Daltons (Da) or per million (ppm). The default is 0.005 Da or 5 ppm, which is the typical mass accuracy readily achievable by high resolution instruments. Lastly, the user should enter MS/MS query and also error threshold in two ways. The ”Deisotope” option shows whether the user wants to filter the isotope result.



MC ID
My Compound Identification
MS Search | MS/MS Search | Possible Reactions | Home | FAQ | Contact Us

MS/MS Search

# Reactions: ● No reaction ○ 1 reaction

Neutral or Ion: ● Neutral ○ $[M+H]^+$ ○ $[M+Na]^+$ ○ $[M+K]^+$ ○ $[M+NH_4]^+$ ○ $[M-H]^-$

Precursor Mass: 153.07897 Da (Batch Mode)

Mass Tolerance: ● In ppm (default: ± 5 ppm): ppm ○ In Da (default: ± 0.005 Da): Da

Query Mass:

```
39.0228 2.0
41.0385 0.8
51.0229 2.5
53.0021 0.6
53.0387 1.4
55.0180 0.8
```

□ Deisotope

MS/MS Tolerance: ● In ppm (default: ± 5 ppm): ppm ○ In Da (default: ± 0.005 Da): Da

Submit Query

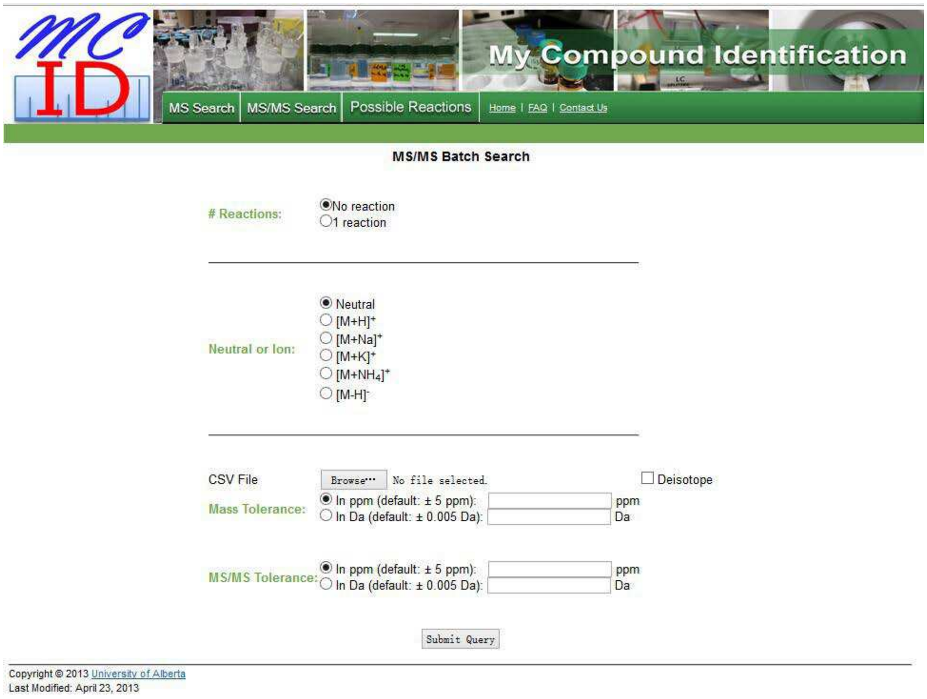
Copyright © 2013 University of Alberta
Last Modified: April 23, 2013

Figure 2.10: MS/MS single search page.

2.6.2 Search Result Display

After the user clicks the ”Submit Query” button, the result page will display (Figure 2.12). The result page for MS/MS search have two tables, the first table shows the detail of the users’ input data and parameters, while the second table shows the candidate list which is sorted by score from high to low. For this particular search, there were three possible hits of matches. The highest score of the candidate is the best match and most likely to be the true structure of the experiment data.

There are 11 columns in the result table (Figure 2.12). This table can be ordered using every column as key. The second one ”HMDB ID” also provides the link for the metabolite substrate to HMDB. Column ”Spectrum” shows the score of the candidate compounds which can be clicked by the user to show how the experiment data match the simulated one.



MS/MS Batch Search

# Reactions: ● No reaction ○ 1 reaction

Neutral or Ion: ● Neutral ○ $[M+H]^+$ ○ $[M+Na]^+$ ○ $[M+K]^+$ ○ $[M+NH_4]^+$ ○ $[M-H]^-$

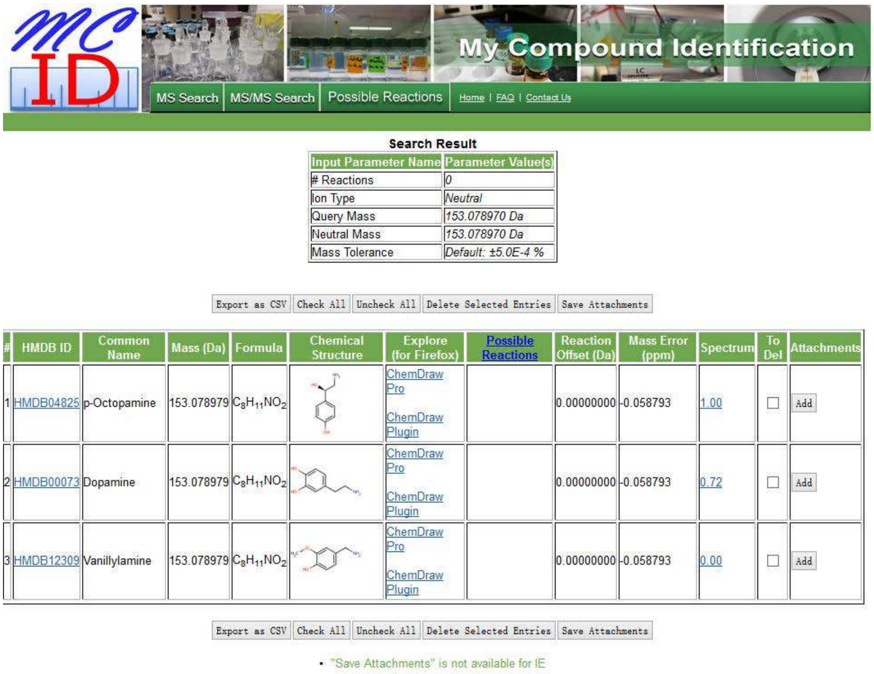
CSV File: Browse… No file selected. □ Deisotope

Mass Tolerance: ● In ppm (default: ± 5 ppm): ppm ○ In Da (default: ± 0.005 Da): Da

MS/MS Tolerance: ● In ppm (default: ± 5 ppm): ppm ○ In Da (default: ± 0.005 Da): Da

Copyright © 2013 University of Alberta
Last Modified: April 23, 2013

Figure 2.11: MS/MS batch search page.



Search Result

Input Parameter Name	Parameter Value(s)
# Reactions	0
Ion Type	Neutral
Query Mass	153.078970 Da
Neutral Mass	153.078970 Da
Mass Tolerance	Default: ±5.0E-4 %

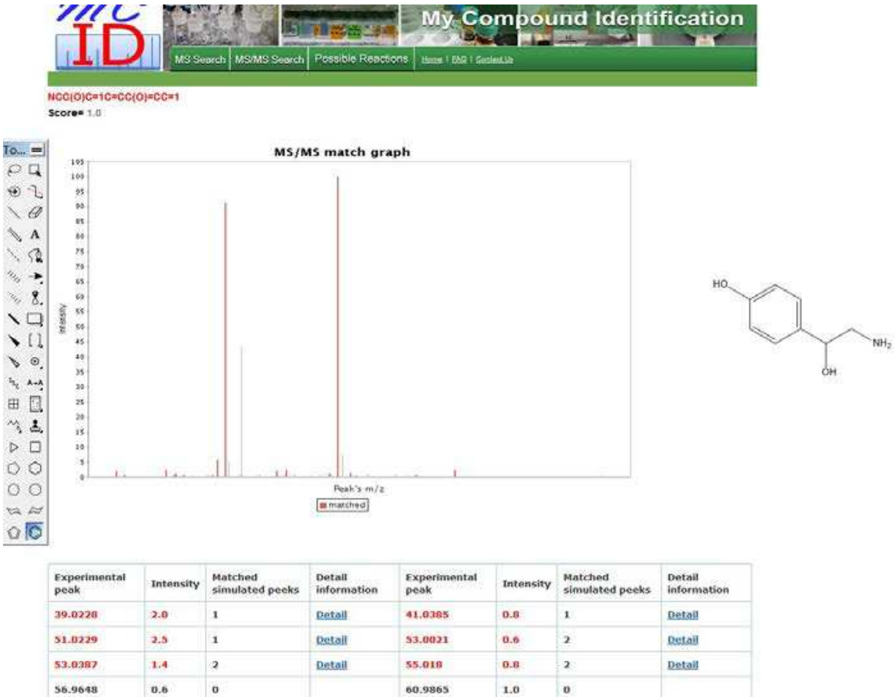
#	HMDB ID	Common Name	Mass (Da)	Formula	Chemical Structure	Explore (for Firefox)	Possible Reactions	Reaction Offset (Da)	Mass Error (ppm)	Spectrum	To Del	Attachments
1	HMDB04825	p-Octopamine	153.078979	$C_8H_{11}NO_2$		ChemDraw Pro ChemDraw Plugin		0.00000000	-0.058793	1.00	□	Add
2	HMDB00073	Dopamine	153.078979	$C_8H_{11}NO_2$		ChemDraw Pro ChemDraw Plugin		0.00000000	-0.058793	0.72	□	Add
3	HMDB12309	Vanillylamine	153.078979	$C_8H_{11}NO_2$		ChemDraw Pro ChemDraw Plugin		0.00000000	-0.058793	0.00	□	Add

- "Save Attachments" is not available for IE

Figure 2.12: Search result page-1.

Figure 2.13 shows the MS/MS match graph and the table. Firstly, for the graph, the y-axis represents the signal intensity of MS/MS peaks while the x-axis represents the peaks' mass which also in the table below. The red peaks in the graph are the matched experimented fragments while the grey peaks are not matched. On the right side of the graph we add the ChemDraw plug-in which allows users to fragment the compound by themselves to validate our results. There are 4 columns in the result table, the first two columns corresponding to the peaks in the graph one by one. The third column represents the number of structures which generated by our chopping algorithm match this peak.

After clicking on the "Detail" button, Figure 2.14 will display this table shows how each peak matched by our chopping algorithm and the user can also check the structure by ChemDraw.



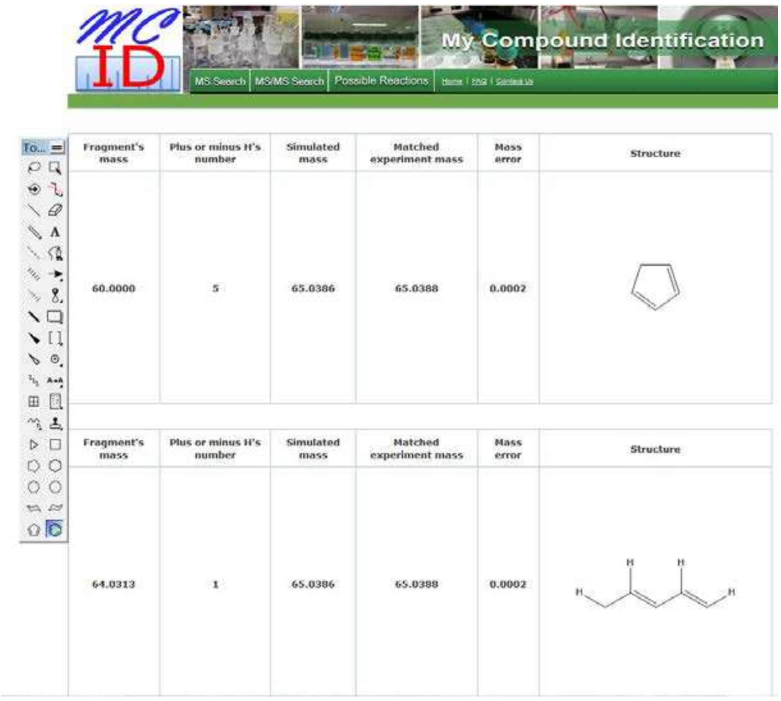
NCC(O)C=1C=CC(O)=CC=1

Score= 1.0

Experimental peak	Intensity	Matched simulated peeks	Detail information	Experimental peak	Intensity	Matched simulated peeks	Detail information
39.0228	2.0	1	Detail	41.0385	0.8	1	Detail
51.0229	2.5	1	Detail	53.0021	0.6	2	Detail
53.0387	1.4	2	Detail	55.018	0.8	2	Detail
56.9648	0.6	0		60.9865	1.0	0	

Figure 2.13: Search result page-2.

In addition to single spectrum search, a user can upload a CSV file generated from LC-MS/MS analysis of a sample to MCID MS/MS for batch mode search. The file format used is shown in the tutorial of our web-site. Batch mode search is useful for examining all the possible matches in a metabolomic profiling experiment. A



Fragment's mass	Plus or minus H's number	Simulated mass	Matched experiment mass	Mass error	Structure
60.0000	5	65.0386	65.0388	0.0002	

Fragment's mass	Plus or minus H's number	Simulated mass	Matched experiment mass	Mass error	Structure
64.0313	1	65.0386	65.0388	0.0002	

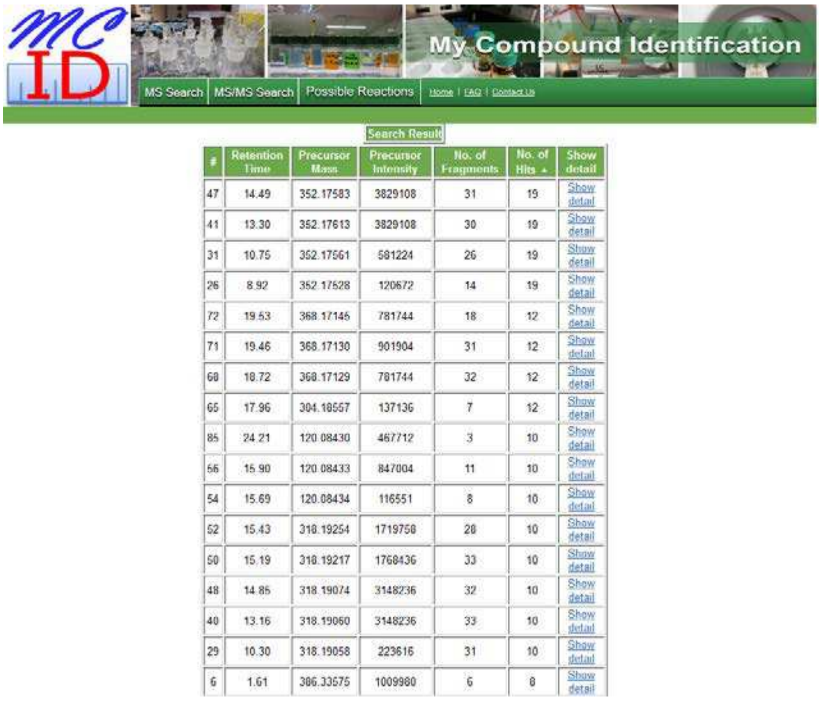
Figure 2.14: Search result page-3.

partial screenshot of the batch mode search results is shown in Figure 2.15. In this case, MS/MS spectra of metabolites were acquired from a human urine sample using LC-QTOF-MS. As Figure 2.15 shows, the summary table lists information on retention time, precursor mass, number of matched metabolites, matching scores with links, matching quality. Again, by clicking the ”detail” of a match, several levels of information can be displayed for manual inspection of the match.

2.6.3 Web Server Framework

We use web MVC frameworks to build this web server. Tomcat and Apache are used as the web server container. JSP, Jquery and HTML are used to build the View, while Java-Servlets is used as the Controller and Java is the main programming language which is used to build the Model. Finally, the MySQL is used as the database. Figure 2.16 shows the framework of our web server.

In the future, as the number of users increases, we will build a distributed system and a load balancer to distribute work to different servers which will be built in different countries. This distributed system haven’t been implemented yet, but we



#	Retention Time	Precursor Mass	Precursor Intensity	No. of Fragments	No. of Hits ▲	Show detail
47	14.49	352.17583	3829108	31	19	Show detail
41	13.30	352.17613	3829108	30	19	Show detail
31	10.75	352.17561	581224	26	19	Show detail
26	8.92	352.17528	120672	14	19	Show detail
72	19.53	368.17146	781744	18	12	Show detail
71	19.46	368.17130	901904	31	12	Show detail
68	18.72	368.17129	781744	32	12	Show detail
65	17.96	304.18557	137136	7	12	Show detail
85	24.21	120.08430	467712	3	10	Show detail
56	15.90	120.08433	847004	11	10	Show detail
54	15.69	120.08434	116551	8	10	Show detail
52	15.43	318.19254	1719758	28	10	Show detail
50	15.19	318.19217	1768436	33	10	Show detail
48	14.85	318.19074	3148236	32	10	Show detail
40	13.16	318.19060	3148236	33	10	Show detail
29	10.30	318.19058	223616	31	10	Show detail
6	1.61	386.33575	1009980	6	8	Show detail

Figure 2.15: Batch result-1.

will do that in the near future.

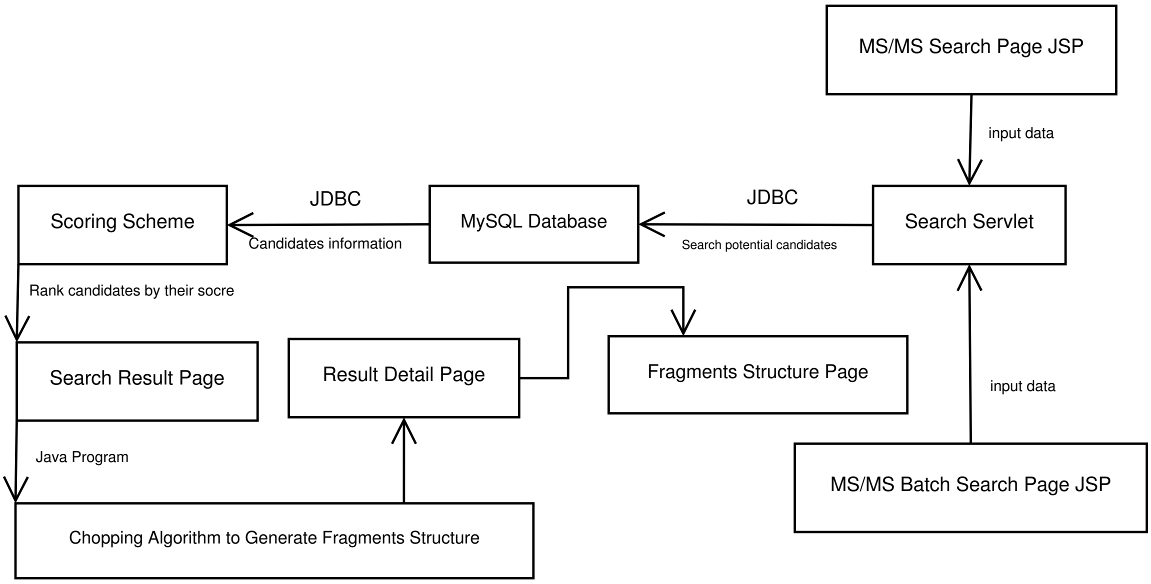


Figure 2.16: Web server framework.

2.7 Result of MS/MS Search Program

2.7.1 MS/MS Search of Standards

To evaluate the performance of MCID MS/MS search, we searched the MS/MS spectra of 50 human metabolite standards against the predicted MS/MS spectral library. These metabolites were randomly picked in order to cover as many different types of compounds as possible. These MS/MS data were searched using both zero reaction and one reaction. Figure 2.17 shows the search results generated, while Table 2.6 shows more detail information of these standards. For the zero-reaction search, an average of 5.6 compounds were mass-matched to a standard, while MS/MS search resulted in an average of 1.3 matches. For the one-reaction search, an average of 16.7 compounds were matched to a standard when accurate mass search alone was used. With MS/MS search, an average of 1.4 compounds were matched to a standard. Both zero-reaction and one-reaction search results indicate that the MS/MS search provides a great improvement over a mass-only library search.

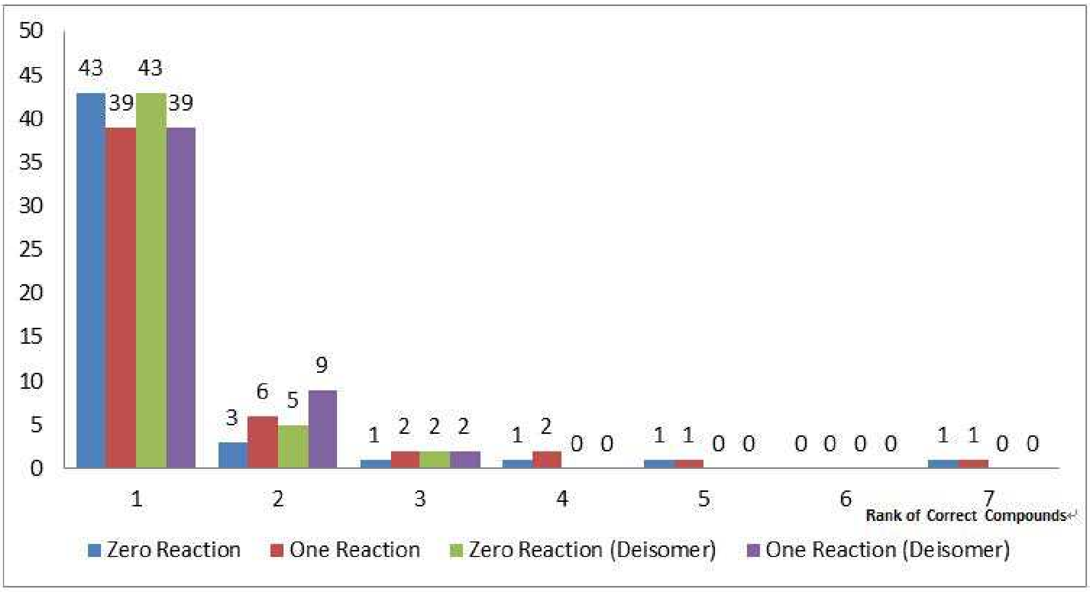


Figure 2.17: Validation result.

Since all the predicted MS/MS spectra were stored in the server, the MCID

HMDB No.	Common name	Score rank/Total hits (zero-reaction library)	Score rank/Total hits (one-reaction library)
HMDB00034	Adenine	1/5	1/19
HMDB00053	Androstenedione	3/12	2/33
HMDB00073	Dopamine	1/6	1/22
HMDB00121	Folic acid	1/1	1/7
HMDB00123	Glycine	2/3	1/18
HMDB00125	Glutathione	1/4	1/5
HMDB00159	L-Phenylalanine	1/14	1/40
HMDB00161	L-Alanine	1/4	1/32
HMDB00235	Thiamine	1/6	2/15
HMDB00235_2	Thiamine	1/6	2/15
HMDB00244	Riboflavin	1/1	1/6
HMDB00262	Thymine	1/6	1/6
HMDB00271	Sarcosine	1/4	1/45
HMDB00294	Urea	1/8	3/12
HMDB00303	Tryptamine	1/13	1/8
HMDB00306	Tyramine	1/13	1/8
HMDB00518	Chenodeoxycholic acid	2/18	7/44 (2/10)*
HMDB00562	Creatinine	1/3	1/1
HMDB00688	Isovalerylcarnitine	1/3	1/16
HMDB00688_2	Isovalerylcarnitine	1/3	3/16
HMDB00696	L-Methionine	1/6	1/8
HMDB00954	trans-Ferulic acid	1/13	1/35
HMDB01044	2'-Deoxyguanosine	1/4	1/32
HMDB01129	N-Acetylmannosamine	2/8	1/20
HMDB01389	Melatonin	1/3	2/6
HMDB01431	Pyridoxamine	1/8	1/16
HMDB01904	3-Nitrotyrosine	1/5	1/8
HMDB02064	N-Acetylputrescine	1/11	4/8
HMDB04816	FAPy-adenine	7/18 (3/8)*	5/8 (2/4)*
HMDB04825	p-Octopamine	1/2	1/6
HMDB29865	Umbelliferone	1/1	1/3
HMDB00064	Creatine	1/17	1/4
HMDB00168	L-Asparagine	1/24	1/10
HMDB00192	L-Cystine	1/1	1/4
HMDB00214	Ornithine	1/23	1/12
HMDB00239	Pyridoxine	1/8	1/26
HMDB00251	Taurine	1/5	1/2
HMDB00289	Uric acid	1/9	1/1
HMDB00292	Xanthine	1/19	1/3
HMDB00299	Xanthosine	1/4	1/9
HMDB00300	Uracil	1/9	1/1
HMDB00575	DL-Homocystine	1/10	1/4
HMDB00641	L-Glutamine	5/15 (2/5)*	2/17
HMDB00670	Homo-L-arginine	1/13	2/8
HMDB00687	L-Leucine	1/8	1/18
HMDB00715	Kynurenic acid	1/5	1/6
HMDB00719	L-Homoserine	4/8 (2/4)*	4/18 (2/6)*
HMDB00725	4-Hydroxyproline	1/17	1/52
HMDB00881	Xanthurenic acid	1/5	1/10
HMDB00883	L-Valine	1/8	1/19

Table 2.6: Summary of MS/MS search results for 50 metabolite standards using the zero-reaction and one-reaction libraries in MCID

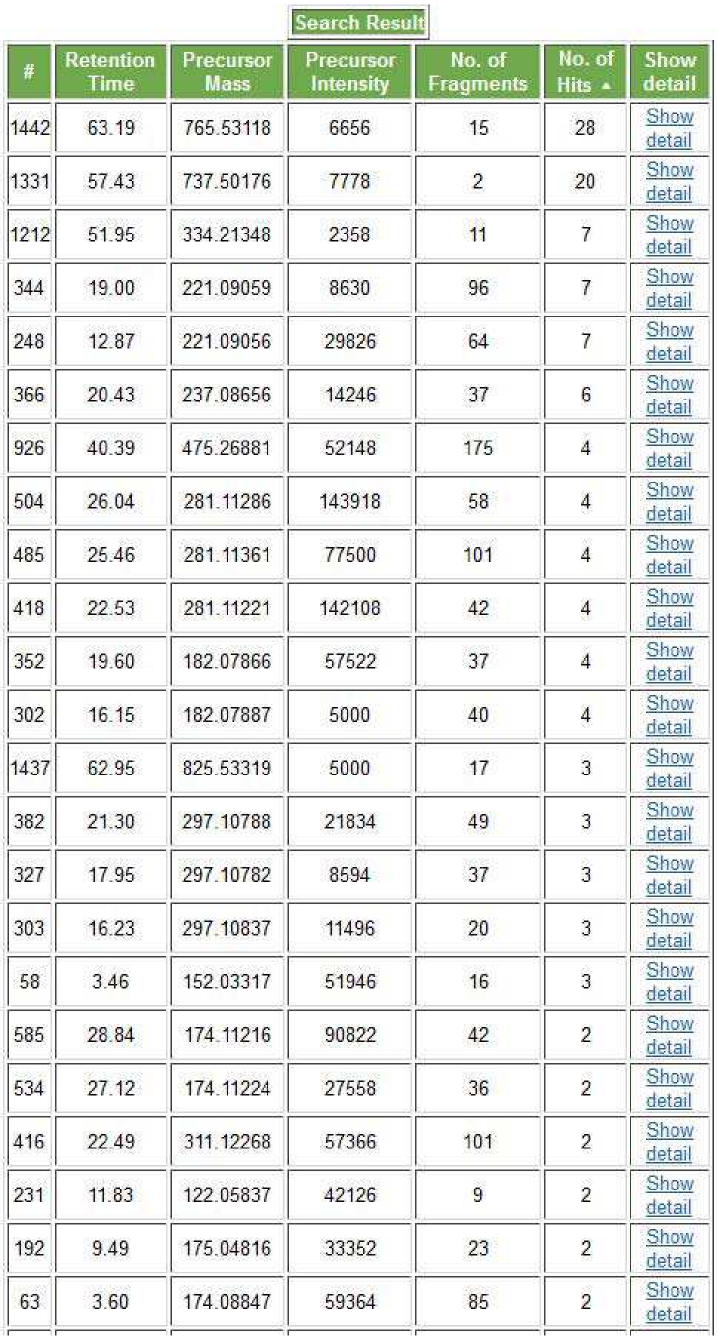
MS/MS searching was very fast. For the 50 standards , the search time was less than 2.5 s per compound. Because the automated MS/MS search can remove a lot of false matches generated from the mass-search alone, only a few top candidates

* in Table 2.6 represents another rank after considering isomers as one group

need to be manually inspected to confirm or disapprove a match. As Figure 2.17 shows, for the zero-reaction search, 43, 3 and 1 out of 50 , gave the correct identity as the top, 2nd and 3rd ranked match, respectively. There was only 2 match below the 3rd rank. Even for the one-reaction search, 38, 6 and 2 out of 50, gave the correct identity as the top, 2nd and 3rd ranked match, respectively. Only 4 had the correct match below the 3rd rank. These results suggest that all or most of the metabolites could be correctly identified as one of the top three candidates from the MS/MS search. Thus, only these candidates need to be inspected manually for match confirmation, greatly improving the metabolite identification efficiency. Table 2.6 shows the detailed rank for these 50 metabolites for both zero reaction and one reaction database. For those metabolites matched below the top 3, Table 2.6 also provides another rank after considering isomers into one group.

2.7.2 MS/MS Search of Urine Metabolites

To demonstrate the utility of MCID MS/MS search for real world applications, a human urine sample was used to test our program. After uploading the MS/MS data file to the MCID MS/MS website and entering the precursor ion mass tolerance of 5 ppm and fragment ion mass tolerance of 5 ppm, the program performed a batch mode search. Figure 2.18 shows a partial screenshot of the search result page using the zero-reaction spectral library. Figure 2.19 shows an example of urine sample's MS/MS spectra in MCID. Table 2.7 lists the identified metabolites, the last column shows the availability of standard MS/MS spectra in the Bruker library (Yes or No). In this case, using the MS/MS spectra collected from the three LC-MS/MS runs to search the zero-reaction library, there were 45 matches. For the ones matched with zero-reaction metabolites, we have tried to perform a cross-validation of these matches using the Bruker human metabolite MS/MS spectral library. This Bruker library of about 800 metabolite standards was created in the same QTOF instrument as the one used for running the urine sample. Thus, excellent fragmentation pattern match of the urine metabolite and library metabolite is expected, which should in turn provide high accuracy for validation of a MCID MS/MS search result of a urine metabolite. From the table, we can see there are 84.4% metabolites which are in the Bruker library. Thus, the MCID MS/MS search result was cross-validated.



Search Result

#	Retention Time	Precursor Mass	Precursor Intensity	No. of Fragments	No. of Hits ▲	Show detail
1442	63.19	765.53118	6656	15	28	Show detail
1331	57.43	737.50176	7778	2	20	Show detail
1212	51.95	334.21348	2358	11	7	Show detail
344	19.00	221.09059	8630	96	7	Show detail
248	12.87	221.09056	29826	64	7	Show detail
366	20.43	237.08656	14246	37	6	Show detail
926	40.39	475.26881	52148	175	4	Show detail
504	26.04	281.11286	143918	58	4	Show detail
485	25.46	281.11361	77500	101	4	Show detail
418	22.53	281.11221	142108	42	4	Show detail
352	19.60	182.07866	57522	37	4	Show detail
302	16.15	182.07887	5000	40	4	Show detail
1437	62.95	825.53319	5000	17	3	Show detail
382	21.30	297.10788	21834	49	3	Show detail
327	17.95	297.10782	8594	37	3	Show detail
303	16.23	297.10837	11496	20	3	Show detail
58	3.46	152.03317	51946	16	3	Show detail
585	28.84	174.11216	90822	42	2	Show detail
534	27.12	174.11224	27558	36	2	Show detail
416	22.49	311.12268	57366	101	2	Show detail
231	11.83	122.05837	42126	9	2	Show detail
192	9.49	175.04816	33352	23	2	Show detail
63	3.60	174.08847	59364	85	2	Show detail

Figure 2.18: Urine sample search result.

Table 2.7: Urine sample identified metabolites

HMDB ID	Common Name	Mass (Da)	Score	Validation
HMDB07951	PC	765.530857	1.00	Yes
HMDB00212	N-Acetylgalactosamine	221.089939	1.00	Yes
HMDB00238	Sepiapterin	237.08619	1.00	Yes
HMDB11479	LysoPE	475.269892	0.67	Yes
HMDB04326	2'-O-Methyladenosine	281.112405	1.00	Yes
HMDB03331	1-Methyladenosine	281.112405	1.00	Yes
HMDB00107	Galactitol	182.07904	1.00	No
HMDB00247	Sorbitol	182.07904	0	Yes
HMDB05862	2-Methylguanosine	297.10732	1.00	No
HMDB01563	1-Methylguanosine	297.10732	1.00	Yes
HMDB00517	L-Arginine	174.111676	1.00	No
HMDB03416	D-Arginine	174.111676	1.00	Yes
HMDB01961	1,7-Dimethylguanosine	311.12297	1.00	No
HMDB02994	Erythritol	122.05791	1.00	No
HMDB02023	Ethyladipic acid	174.08921	1.00	No
HMDB00893	Suberic acid	174.08921	0.92	Yes
HMDB06509	Nervonyl carnitine	102.128274	1.00	Yes
HMDB05041	Donepezil	379.214744	1.00	Yes
HMDB13122	LysoPC(P-18:0)	507.368877	1.00	Yes
HMDB06294	16-hydroxy hexadecanoic acid	271.22732	1.00	Yes
HMDB06059	20-Carboxyleukotriene B4	366.20424	1.00	Yes
HMDB11603	4-(Methylnitrosamino)	207.100777	1.00	Yes
HMDB02171	Glycylprolylhydroxyproline	285.132472	1.00	Yes
HMDB05764	Melanostatin	284.184841	1.00	Yes
HMDB06059	20-Carboxyleukotriene B4	366.20424	1.00	Yes
HMDB05768	Kyotorphin	337.175005	1.00	No
HMDB00670	Homo-L-arginine	188.127326	1.00	No
HMDB06790	Galactosylglycerol	254.10017	1.00	Yes
HMDB03950	7-Methylinosine	283.104246	0.95	Yes
HMDB00389	2'-Deoxysepiapterin	221.091275	1.00	Yes
HMDB06357	cis-2-Methylaconitate	188.03209	1.00	Yes
HMDB01410	2-Amino-4-oxo-6	267.06037	0.01	Yes
HMDB00824	Propionylcarnitine	217.131409	1.00	Yes
HMDB00157	Hypoxanthine	136.038511	1.00	Yes
HMDB01185	S-Adenosylmethionine	399.145066	1.00	Yes

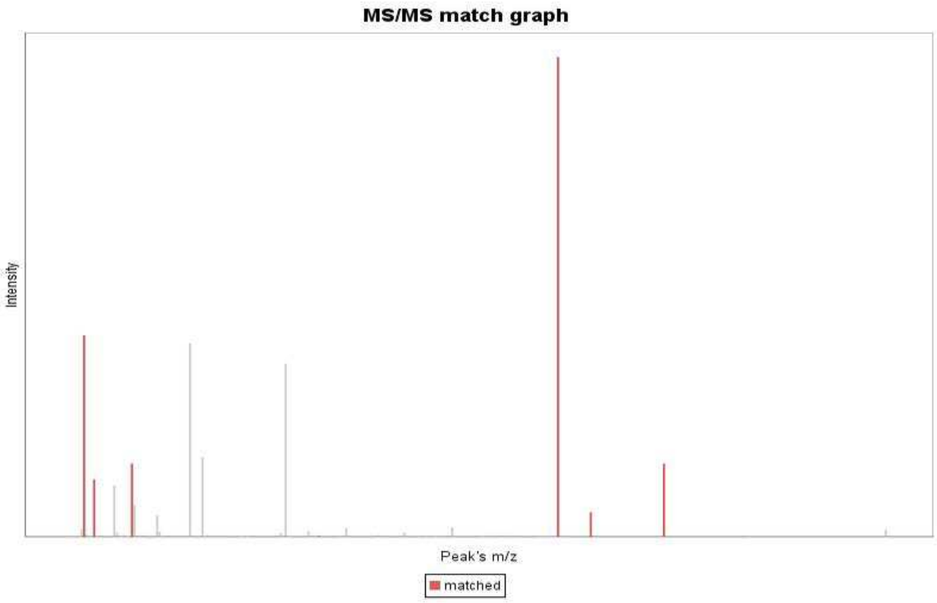


Figure 2.19: S-Adenosylmethionine MS/MS spectra.

2.8 Conclusions and Future Work

In this work, we have developed a web server for MS/MS metabolite identification based on accurate MS and MS/MS search using a comprehensive library of predicted spectra of human metabolites. In this program, we have an efficient method of predicting fragments ions using heteroatom-initiated bond breakage rules and then applied it to all the possible human metabolites in the MCID databases to generate a predicted MS/MS spectral library. An automated MS/MS search program was developed that allows a user to search an experimental MS/MS spectra against this MCID MS/MS spectral library for spectral match. The search results could be manually interpreted for possible metabolite identification. This MCID MS/MS web server allows a user to narrow down the possible metabolite structures and thus guide the synthesis of chemical standards for eventual structure confirmation.

For the web services, we use the MVC model which enables quick updating and modifications. In the future, when the number of users increases, we will implement a distributed system and a load balancer to control the work load.

The major limitation of the MCID MS/MS search, compared to a standard M-

S/MS library, is that the accuracy of any match to a predicted spectrum is not as high as a match between the experimental spectrum and a standard spectrum. Thus, we feel that MCID MS/MS search provides a complimentary tool to the standard MS/MS library search. Also we feel that with more experimental data collected we can use advanced machine learning algorithms to learn more chopping rules from the data and develop a better scoring scheme.

Chapter 3

DNS Library Construction

The paper has been submitted to Analytical Chemistry according to this work.

3.1 DnsID Program for Metabolite Identification

Identification of labeled metabolites in a sample is usually done in two steps. The first step is to run the RT calibration program mixture in LC-MS to produce the retention time information for all the calibrants. The next step is to run a real sample under the same LC-MS condition as those used for running the calibrants. The two data files are then uploaded to the DnsID program which is hosted at http://mcid.cs.ualberta.ca:8080/Compound_MRT/. In DnsID, the retention times of all the labeled metabolites detected in the sample are first corrected using the retention time information obtained from the RTcal. The program then compares the accurate mass and corrected retention time of individual unknown metabolite to those in the Dns-library for possible match. If a tandem mass spectrometer is available, MS/MS spectrum of a matched metabolite can be generated and searched against the standard MS/MS spectra in the Dns-library for further confirmation of the metabolite identity. This program's work flow is shown in Figure 3.1. Figure 3.2 shows the distribution of metabolites in our library, the x-axis represents the retention time in Min while the y-axis represents the mz_light of compounds in Da.

T. Huan, Y. Wu, C. Tang, G. Lin, L. Li. DnsID in MyCompoundID for Rapid Identification of Dansylated Amine- and Phenol-containing Metabolites in Metabolomics. *Analytical chemistry*

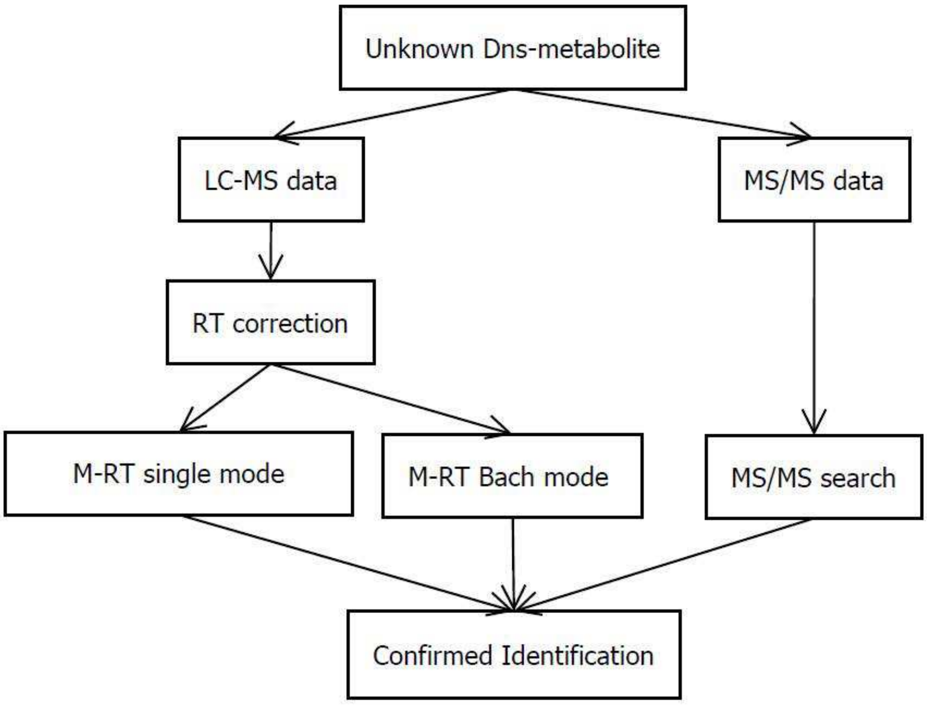


Figure 3.1: Work flow of Dns library

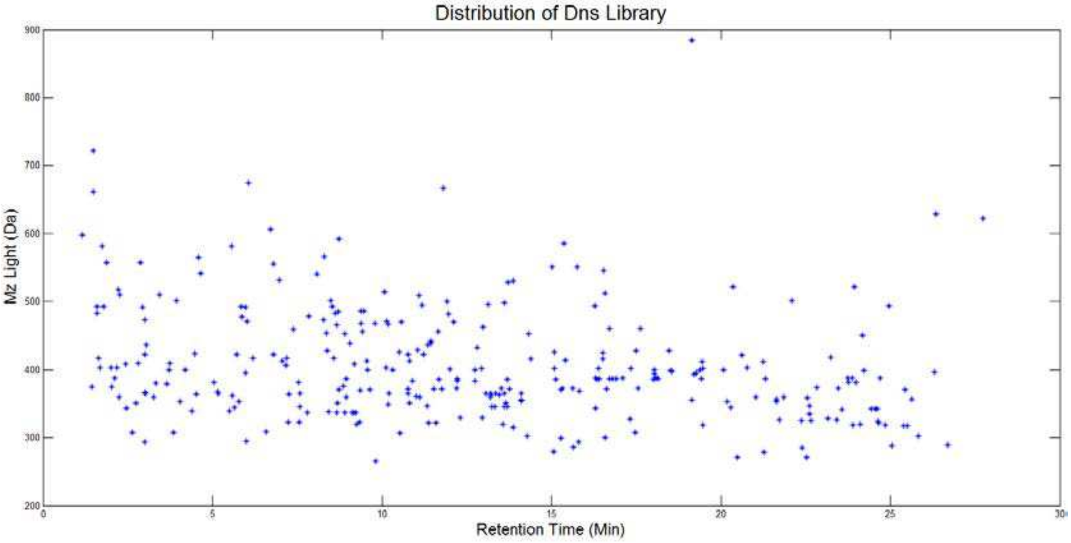


Figure 3.2: Distribution of Dns library

3.2 Retention Time Calibration Algorithm

In building the Dns-library, RT of individual Dns-metabolite has been normalized to those of RT calibration. Thus, the purpose of RT calibration is to correct any

RT shifts of metabolites in the LC-MS dataset generated in a user's laboratory and those of the Dns-library. This is accomplished by using a set of carefully chosen RT calibrants with their RT data already stored in the Dns-library. Algorithm 4 shows the RT calibration algorithm. This algorithm can be explained by the following steps:

1. Check the type of calibration file and read the corresponding system file.
2. Crate windows for each item in the calibration file.
3. Iterate all the items in the system file and put potential candidates which satisfy all constraints into the window.
4. Find the candidate in the window which has the highest intensity value. This candidate will be the matched one corresponding to the calibration file.
5. Use linear RT correction to calculate the predicted RT shift.

The "Type" in the algorithm means the user can choose the system files which contain 10 and 22 Dns-standards respectively see Table 3.1 and 3.2. Then the RT calibration works by dividing the whole LC chromatogram into 10 or 22 time intervals. Except the first and last time intervals, all the other 9 or 21 intervals were bracketed by two reference standards from the RTcal. In each interval, the RT differences of the two reference metabolites between the sample LC-MS run and the stored RTcal data are calculated $(\triangle t_a, \triangle t_{a+1})$. Then, a linear RT correction is applied to calculate the predicted RT shift $(\triangle t)$ (Figure 3.3). To correct the RT shift of any peak within the interval, the predicted RT shift $(\triangle t)$ is used to subtract from the original RT (t_o) and generate a corrected RT (t_c), see the following equation for detail.

$$
\begin{aligned}
&t_c = t_o + \Delta t_a + \frac{(t_o - t_a) * (\Delta t_{a+1} - \Delta t_a)}{t_{a+1} - t_a} \\
&t_a \in Standards = \{t_1, t_2, ...t_{m-1}\} \quad m \in \{10, 22\} \\
&\Delta t_a \in Shifts = \{\Delta t_1, \Delta t_2, ...\Delta t_{m-1}\} \quad m \in \{10, 22\} \\
&where\ t_o >= t_a\ and\ t_o <= t_{a+1}
\end{aligned}
$$

Algorithm 4 RT calibration algorithm

```
Input: List sample_RT,List Mass,boolean Type, Double input_RT
Output: Double output_RT

function RT_Calibration(List sample_RT,List Mass,boolean Type, Double input_RT)
if Type then
    List<metabolite> system = systemfile(1).
else
    List<metabolite> system = systemfile(2).
end if
List calibrated_RT.
for i in range(0,rt.size()) do
    List<metabolite> window.
    for j in range(0,system.size()) do
        if Math.abs(Mass.get(i)-system.get(j).getmass())<=0.005 and Math.abs(sample_RT.get(i)-system.get(j).getRT())<=60 then
            window.add(system.get(j))
        end if
    end for
    Double max_intensity;
    metabolite matched;
    for metabolite m in window do
        if m.intensity > max_intensity then
            matched= m;
            max_intensity = m.intensity;
        end if
    end for
    calibrated_RT.add(matched.getRT());
end for
Double output_RT=input_RT;
for i in range(0,calibrated_RT.size()-1) do
    if input_RT>=calibrated_RT.get(i) and input_RT<=calibrated_RT.get(i+1) then
        output_RT=system.get(i).getRT()+(input_RT-calibrated_RT.get(i))/(calubrated_RT.get(i+1)-calibrated_RT.get(i))*(system.get(i+1).getRT()-system.get(i).getRT())
    end if
end for
return output_RT
```

Table 3.1: 10 Dns-standards in system file

Index	Retention Time(Sec)	Mz_Light(Da)	Mz_heavy(Da)
1	211	408.1700	410.1773
2	245	339.1009	341.1082
3	320	353.1166	355.1239
4	448.8	323.1060	325.1133
5	624.8	349.1217	351.1290
6	676.9	383.1094	385.1167
7	806.9	399.1373	401.1446
8	937.1	354.0702	356.0775
9	1,053.2	307.1111	309.1184
10	1,360.8	324.5953	326.6020

Table 3.2: 22 Dns-standards in system file

Index	Retention Time(Sec)	Mz_Light(Da)	Mz_heavy(Da)
1	146.4	408.1700	410.1773
2	264	339.1009	341.1082
3	303	381.1115	383.1188
4	347.4	353.1166	355.1239
5	395.4	309.0903	311.0976
6	454.2	323.1060	325.1133
7	534.6	337.1216	339.1289
8	610.8	349.1216	351.1289
9	653.4	383.1094	385.1167
10	764.4	399.1373	401.1446
11	823.2	528.1799	530.1872
12	846.6	354.0702	356.0775
13	880.2	251.0849	253.0922
14	949.2	368.0859	370.0932
15	1,048.2	307.1111	309.1184
16	1,148.4	355.1475	357.1548
17	1,262.4	360.1138	362.1211
18	1,343.4	285.1162	287.1235
19	1,359.0	307.1111	309.1184
20	1,479.0	322.1058	324.1131
21	1,537.8	356.1315	358.1388
22	1,602.0	289.0767	291.840

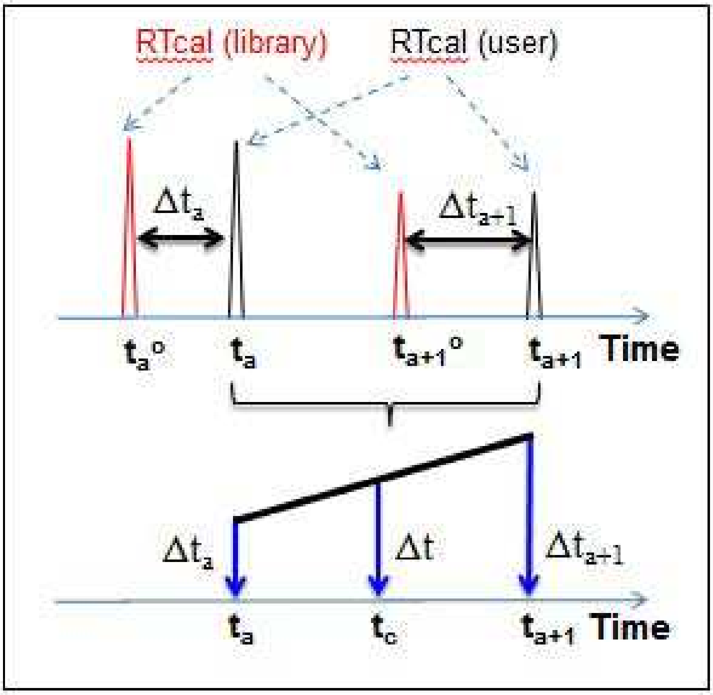


Figure 3.3: Interval Linear Calibration

The performance of the RT calibration method is illustrated in Figure 3.4 where retention time correlations of different LC-MS experiments before and after applying RT calibration are shown. In this case, 20 standards have been selected from the library with retention time span over the entire metabolite elution window. Figure 3.4 also shows the RT correlation plots of the 20 standards from the data obtained by LC-FTICR-MS and those in the Dns-library. Before applying the RT calibration, there is a near-constant shift to a higher RT for the LC-FTICR-MS data. The RT shift can be as large as 4.8 min. Although the RT shift becomes smaller at high organic elution region, the shift is still greater than 0.5 min (30 s). Nevertheless, even with these large RT variations, after applying the RT calibration, an excellent linear correlation between the corrected RT and the library RT can be obtained, with a slope of 1.0079. For all these 20 standards, the RT shift after calibration is below 15 s, which is the typical RT tolerance threshold we use for performing DnsID M-RT search. This example illustrates that the RT calibration method is able to correct for RT shifts found in different LC-MS setups.

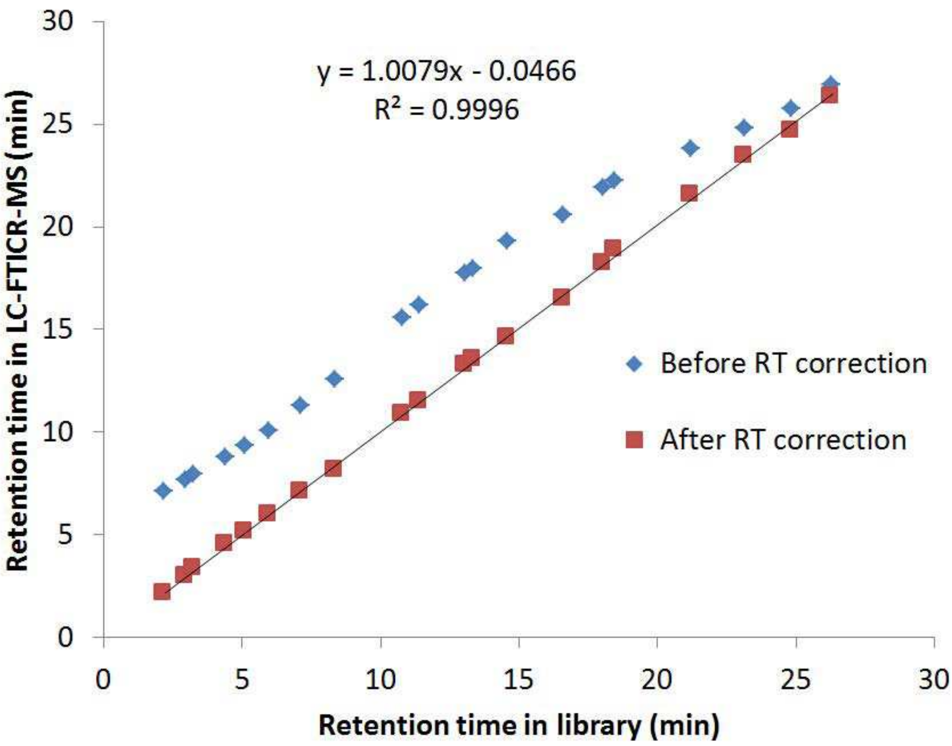


Figure 3.4: Retention times of RTcal obtained by LC-FTICR-MS vs. those in the Dns-library.

3.3 Search Algorithm

For the DnsID library, we have two search options for the users, one is M-RT search and the other one is MS/MS search. For the M-RT search, DnsID automatically performs RT calibration using the calibration file against data in our Dns-library, then we use RT tolerance and mass tolerance to filter the metabolites and display the potential candidates to the user (see Algorithm 5). This algorithm can be explained by the following steps:

1. Check the type of calibration file and achieve the corresponding system file.

2. Use Algorithm 4 to calculate the RT shift in the calibration file and linear correct the RT.

3. Check the type of mass tolerance and calculate the corresponding tolerance.

Algorithm 5 MRT search algorithm

```
Input: File Calibration, Double mass, Double RT, Boolean Type, Boolean isPPM, Double tolerance_m, Double tolerance_rt
Output: List<metabolite> result

function MRT_search(List sample_RT,List Mass,boolean Type, Double input_RT, Double tolerance_m, Double tolerance_rt)
if Type then
    List<metabolite> system = systemfile(1).
else
    List<metabolite> system = systemfile(2).
end if

List<metabolite> result =new List();
RT_c = RT_calibration(Calibration,RT,system);
List<metabolite> library= Dns_library();

if isPPM then
    tolerance_m = tolerance_m*mass*$10^{-6}$;
end if

for metabolite m in library do
    if Math.abs(mass-m.getMass())<=tolerance_m and Math.abs(m.getRT(i)-RT_c)<=tolerance_rt then
        result.add(m);
    end if
end for
return result;
```

4. Use mass tolerance and retention time tolerance to filter the metabolites in DNS library.

For the MS/MS search program, the dansyl compound MS/MS spectral library can be searched using an acquired MS/MS spectrum, then we use the MS/MS scoring scheme which is in the Chapter 2.4 to filter and rank the metabolites in Dns-library, see Algorithm 6. This algorithm can be explained by the following steps:

1. Achieve all the metabolites in DNS library.

2. Check the input MS tolerance type and MS/MS tolerance type and calculate the corresponding absolute value.

3. Use Algorithm 5 to filter the results and obtain the potential candidates.

4. Iterate all the potential candidates and use MS/MS scoring scheme to calculate the score for each candidates.

5. Sort the results by score.

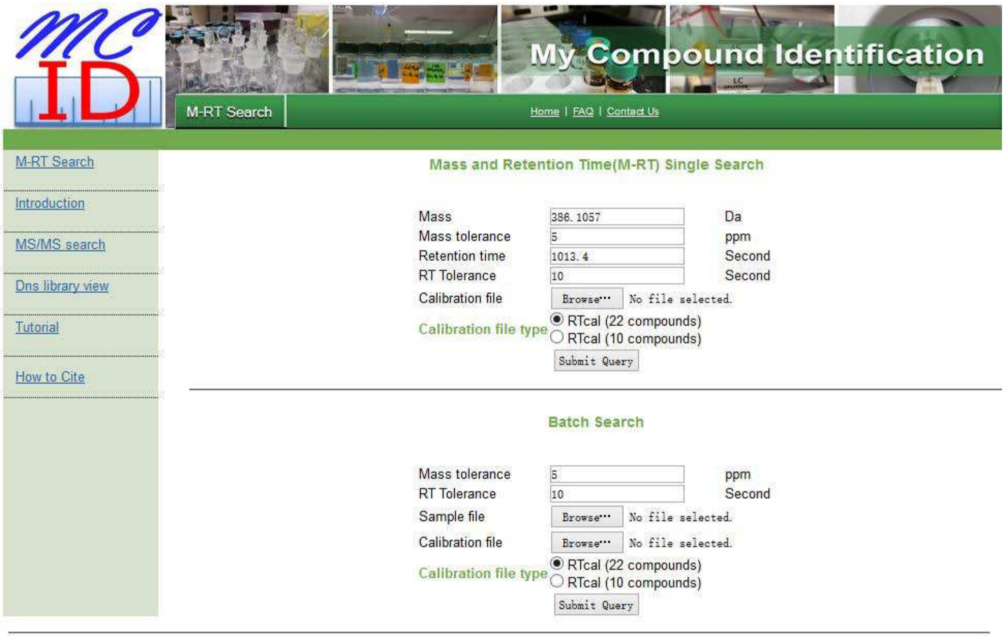
Algorithm 6 MRT MS/MS search algorithm

```
Input: Double mass, List<Double> Intensity, List<Double> MSMS, Double MSMS_tolerance, Boolean isPPM
Output: List<Double> scores;

function MRT_MSMS_search(List<Double> Intensity, List<Double> MSMS, Double mass, Double MSMS_tolerance, Boolean isPPM)
if isPPM then
    MSMS_tolerance = MSMS_tolerance * mass * $10^{-6}$.
end if

List<metabolite> library= Dns_library();

for metabolite m in library do
    List<Double> MSMS_m= m.getMSMS();
    List<Double> Match =new List();
    for Double fragments in MSMS do
        if there is a item in MSMS_m and let Math.abs(item-fragments)< MSMS_tolerance then
            Match.add(fragments);
        else
            Match.add(0);
        end if
    end for
    Double score=0;
    for int i in Range(0,Match.size()) do
        score= score+ Match.get$(i)^3$*Intensity.get$(i)^{0.6}$;
    end for
    scores.add(score);
end for
sort(scores);
return scores;
```

3.4 DnsID Web Services

3.4.1 M-RT Search

As Figure 3.5 shows, there are two modes of M-RT search. In the single search mode, a calibration file is first uploaded. Accurate mass of a Dns-metabolite of interest found in a sample is entered along with the mass tolerance. The retention time of the metabolite and its tolerance are then entered. The RT tolerance should be within the limit of RT variation which is typically within 15s and the mass error is within 10ppm. After submitting the query, a search result page is displayed (see Figure 3.6). It shows the matched compound name, HMDB number, several numeric parameters as well as external links to HMDB and KEGG. These links are useful to extract biological information about the matched metabolite. On the summary page, there is also a ”Show Detail” column which provides a link to the ion chromatogram and MS/MS spectrum of the Dns-standard (see Figure 3.7). The standard’s chromatogram is particularly useful for manual inspection of a match with a larger RT error (i.e., between 15 and 30s). A larger RT error is acceptable if this is due to relatively poor peak shape. Otherwise, the match may be false.



Mass and Retention Time(M-RT) Single Search

Mass	386.1057	Da
Mass tolerance	5	ppm
Retention time	1013.4	Second
RT Tolerance	10	Second
Calibration file	Browse… No file selected.	
Calibration file type	● RTcal (22 compounds) ○ RTcal (10 compounds)	

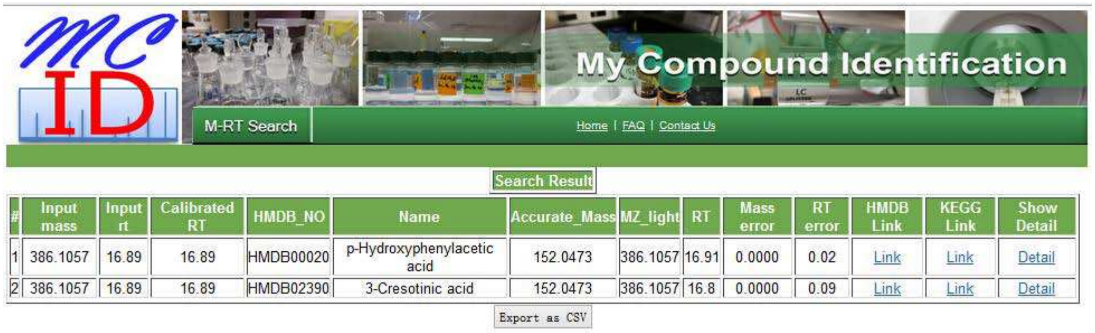
Submit Query

Batch Search

Mass tolerance	5	ppm
RT Tolerance	10	Second
Sample file	Browse… No file selected.	
Calibration file	Browse… No file selected.	
Calibration file type	● RTcal (22 compounds) ○ RTcal (10 compounds)	

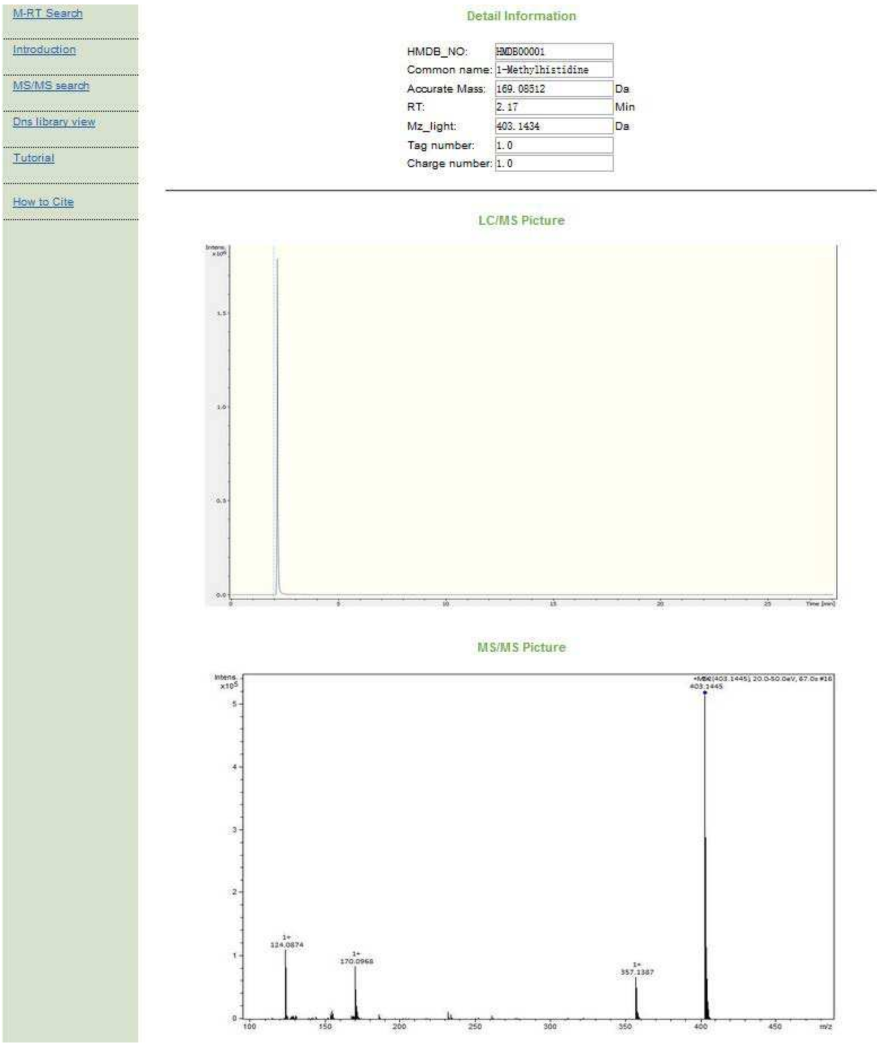
Submit Query

Figure 3.5: M-RT search page.



#	Input mass	Input rt	Calibrated RT	HMDB_NO	Name	Accurate_Mass	MZ_light	RT	Mass error	RT error	HMDB Link	KEGG Link	Show Detail
1	386.1057	16.89	16.89	HMDB00020	p-Hydroxyphenylacetic acid	152.0473	386.1057	16.91	0.0000	0.02	Link	Link	Detail
2	386.1057	16.89	16.89	HMDB02390	3-Cresotinic acid	152.0473	386.1057	16.8	0.0000	0.09	Link	Link	Detail

Figure 3.6: M-RT result page.



Detail Information

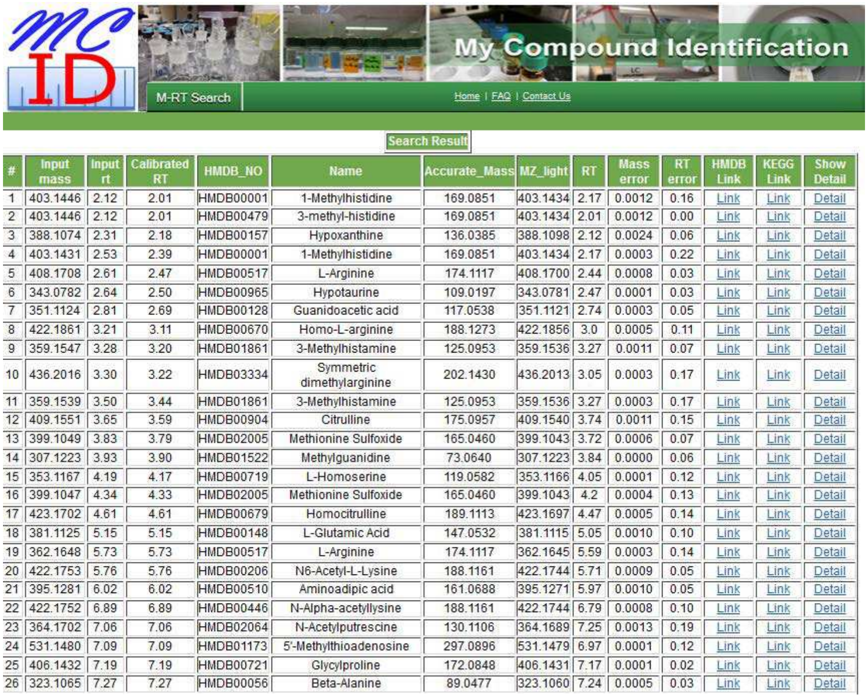
HMDB_NO: HMDB00001
Common name: 1-Methylhistidine
Accurate Mass: 169.08512 Da
RT: 2.17 Min
Mz_light: 403.1434 Da
Tag number: 1.0
Charge number: 1.0

LC/MS Picture

MS/MS Picture

Figure 3.7: M-RT result detail.

For untargeted metabolite identification, the batch mode M-RT search can be used. In this case, both the calibration file and the CSV file of a sample LC-MS run are uploaded (see Figure 3.5). The mass tolerance and retention time tolerance are also entered. The search result page displays all the matches that can be sorted according to an individual parameter. Figure 3.8 shows a partial list of matches from the analysis of a dansyl labeled human urine sample. The mass error and RT error of each match are shown in the search result. On the search result page, there is an option of saving the search results as a CSV file to the user's computer. This file can be opened locally by Excel or other program for presentation or further processing.

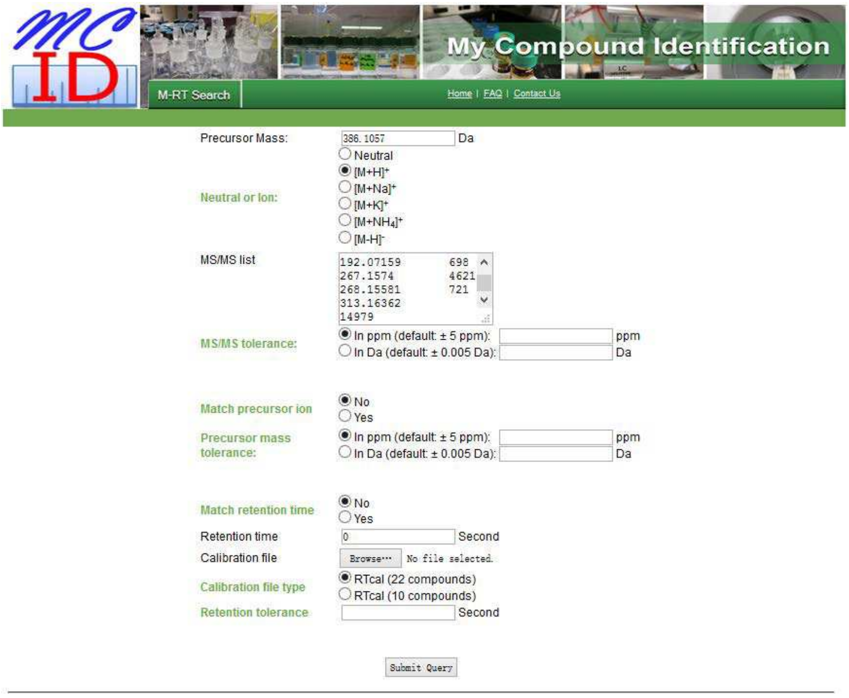


#	Input mass	Input rt	Calibrated RT	HMDB_NO	Name	Accurate_Mass	MZ_light	RT	Mass error	RT error	HMDB Link	KEGG Link	Show Detail
1	403.1446	2.12	2.01	HMDB00001	1-Methylhistidine	169.0851	403.1434	2.17	0.0012	0.16	Link	Link	Detail
2	403.1446	2.12	2.01	HMDB00479	3-methyl-histidine	169.0851	403.1434	2.01	0.0012	0.00	Link	Link	Detail
3	388.1074	2.31	2.18	HMDB00157	Hypoxanthine	136.0385	388.1098	2.12	0.0024	0.06	Link	Link	Detail
4	403.1431	2.53	2.39	HMDB00001	1-Methylhistidine	169.0851	403.1434	2.17	0.0003	0.22	Link	Link	Detail
5	408.1708	2.61	2.47	HMDB00517	L-Arginine	174.1117	408.1700	2.44	0.0008	0.03	Link	Link	Detail
6	343.0782	2.64	2.50	HMDB00965	Hypotaurine	109.0197	343.0781	2.47	0.0001	0.03	Link	Link	Detail
7	351.1124	2.81	2.69	HMDB00128	Guanidoacetic acid	117.0538	351.1121	2.74	0.0003	0.05	Link	Link	Detail
8	422.1861	3.21	3.11	HMDB00670	Homo-L-arginine	188.1273	422.1856	3.0	0.0005	0.11	Link	Link	Detail
9	359.1547	3.28	3.20	HMDB01861	3-Methylhistamine	125.0953	359.1536	3.27	0.0011	0.07	Link	Link	Detail
10	436.2016	3.30	3.22	HMDB03334	Symmetric dimethylarginine	202.1430	436.2013	3.05	0.0003	0.17	Link	Link	Detail
11	359.1539	3.50	3.44	HMDB01861	3-Methylhistamine	125.0953	359.1536	3.27	0.0003	0.17	Link	Link	Detail
12	409.1551	3.65	3.59	HMDB00904	Citrulline	175.0957	409.1540	3.74	0.0011	0.15	Link	Link	Detail
13	399.1049	3.83	3.79	HMDB02005	Methionine Sulfoxide	165.0460	399.1043	3.72	0.0006	0.07	Link	Link	Detail
14	307.1223	3.93	3.90	HMDB01522	Methylguanidine	73.0640	307.1223	3.84	0.0000	0.06	Link	Link	Detail
15	353.1167	4.19	4.17	HMDB00719	L-Homoserine	119.0582	353.1166	4.05	0.0001	0.12	Link	Link	Detail
16	399.1047	4.34	4.33	HMDB02005	Methionine Sulfoxide	165.0460	399.1043	4.2	0.0004	0.13	Link	Link	Detail
17	423.1702	4.61	4.61	HMDB00679	Homocitrulline	189.1113	423.1697	4.47	0.0005	0.14	Link	Link	Detail
18	381.1125	5.15	5.15	HMDB00148	L-Glutamic Acid	147.0532	381.1115	5.05	0.0010	0.10	Link	Link	Detail
19	362.1648	5.73	5.73	HMDB00517	L-Arginine	174.1117	362.1645	5.59	0.0003	0.14	Link	Link	Detail
20	422.1753	5.76	5.76	HMDB00206	N6-Acetyl-L-Lysine	188.1161	422.1744	5.71	0.0009	0.05	Link	Link	Detail
21	395.1281	6.02	6.02	HMDB00510	Aminoadipic acid	161.0688	395.1271	5.97	0.0010	0.05	Link	Link	Detail
22	422.1752	6.89	6.89	HMDB00446	N-Alpha-acetyllysine	188.1161	422.1744	6.79	0.0008	0.10	Link	Link	Detail
23	364.1702	7.06	7.06	HMDB02064	N-Acetylputrescine	130.1106	364.1689	7.25	0.0013	0.19	Link	Link	Detail
24	531.1480	7.09	7.09	HMDB01173	5'-Methylthioadenosine	297.0896	531.1479	6.97	0.0001	0.12	Link	Link	Detail
25	406.1432	7.19	7.19	HMDB00721	Glycylproline	172.0848	406.1431	7.17	0.0001	0.02	Link	Link	Detail
26	323.1065	7.27	7.27	HMDB00056	Beta-Alanine	89.0477	323.1060	7.24	0.0005	0.03	Link	Link	Detail

Figure 3.8: M-RT batch result.

3.4.2 DnsID MS/MS Search

In our DnsID, the dansyl compound MS/MS spectral library can be searched using an acquired MS/MS spectrum. There are two options of MS/MS search. The first option is to enter the precursor ion mass and the fragment ion masses with user-defined mass tolerances, while the 2nd option is to enter the fragment ion masses with a mass tolerance without specifying the precursor ion mass (see Figure 3.9). The latter is useful to find related metabolites having similar core structure and fragment ions. An example is shown in Figure 3.10. In this case, the unknown metabolite matches to Dns library based on the fragment ions only, we filter the metabolites which have score 0.



Precursor Mass: 386. 1057 Da

Neutral or Ion:
○ Neutral
● $[M+H]^+$
○ $[M+Na]^+$
○ $[M+K]^+$
○ $[M+NH_4]^+$
○ $[M-H]^-$

MS/MS list

```
192.07159        698
267.1574         4621
268.15581        721
313.16362
14979
```

MS/MS tolerance:
● In ppm (default: ± 5 ppm): ppm
○ In Da (default: ± 0.005 Da): Da

Match precursor ion
● No
○ Yes

Precursor mass tolerance:
● In ppm (default: ± 5 ppm): ppm
○ In Da (default: ± 0.005 Da): Da

Match retention time
● No
○ Yes

Retention time 0 Second

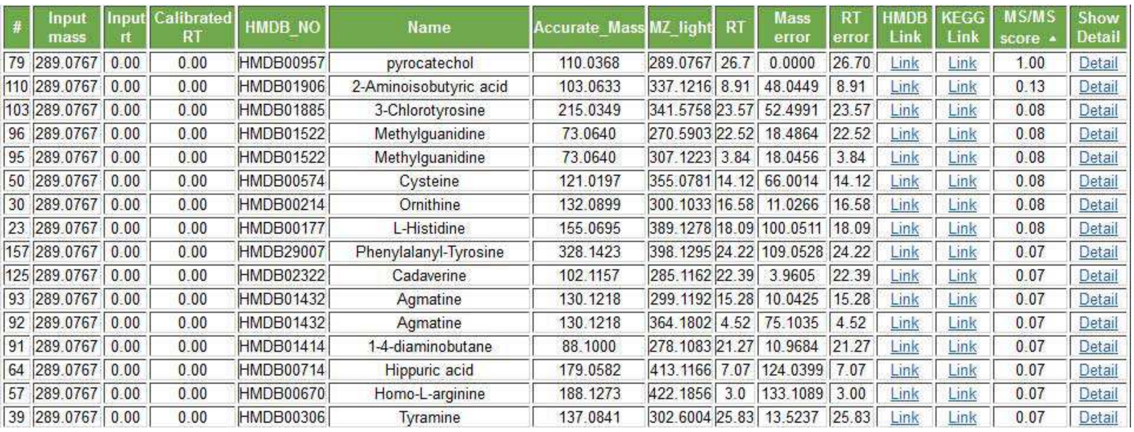
Calibration file Browse... No file selected.

Calibration file type
● RTcal (22 compounds)
○ RTcal (10 compounds)

Retention tolerance Second

Submit Query

Figure 3.9: M-RT MS/MS search page.



#	Input mass	Input rt	Calibrated RT	HMDB_NO	Name	Accurate_Mass	MZ_light	RT	Mass error	RT error	HMDB Link	KEGG Link	MS/MS score ▲	Show Detail
79	289.0767	0.00	0.00	HMDB00957	pyrocatechol	110.0368	289.0767	26.7	0.0000	26.70	Link	Link	1.00	Detail
110	289.0767	0.00	0.00	HMDB01906	2-Aminoisobutyric acid	103.0633	337.1216	8.91	48.0449	8.91	Link	Link	0.13	Detail
103	289.0767	0.00	0.00	HMDB01885	3-Chlorotyrosine	215.0349	341.5758	23.57	52.4991	23.57	Link	Link	0.08	Detail
96	289.0767	0.00	0.00	HMDB01522	Methylguanidine	73.0640	270.5903	22.52	18.4864	22.52	Link	Link	0.08	Detail
95	289.0767	0.00	0.00	HMDB01522	Methylguanidine	73.0640	307.1223	3.84	18.0456	3.84	Link	Link	0.08	Detail
50	289.0767	0.00	0.00	HMDB00574	Cysteine	121.0197	355.0781	14.12	66.0014	14.12	Link	Link	0.08	Detail
30	289.0767	0.00	0.00	HMDB00214	Ornithine	132.0899	300.1033	16.58	11.0266	16.58	Link	Link	0.08	Detail
23	289.0767	0.00	0.00	HMDB00177	L-Histidine	155.0695	389.1278	18.09	100.0511	18.09	Link	Link	0.08	Detail
157	289.0767	0.00	0.00	HMDB29007	Phenylalanyl-Tyrosine	328.1423	398.1295	24.22	109.0528	24.22	Link	Link	0.07	Detail
125	289.0767	0.00	0.00	HMDB02322	Cadaverine	102.1157	285.1162	22.39	3.9605	22.39	Link	Link	0.07	Detail
93	289.0767	0.00	0.00	HMDB01432	Agmatine	130.1218	299.1192	15.28	10.0425	15.28	Link	Link	0.07	Detail
92	289.0767	0.00	0.00	HMDB01432	Agmatine	130.1218	364.1802	4.52	75.1035	4.52	Link	Link	0.07	Detail
91	289.0767	0.00	0.00	HMDB01414	1-4-diaminobutane	88.1000	278.1083	21.27	10.9684	21.27	Link	Link	0.07	Detail
64	289.0767	0.00	0.00	HMDB00714	Hippuric acid	179.0582	413.1166	7.07	124.0399	7.07	Link	Link	0.07	Detail
57	289.0767	0.00	0.00	HMDB00670	Homo-L-arginine	188.1273	422.1856	3.0	133.1089	3.00	Link	Link	0.07	Detail
39	289.0767	0.00	0.00	HMDB00306	Tyramine	137.0841	302.6004	25.83	13.5237	25.83	Link	Link	0.07	Detail

Figure 3.10: M-RT MS/MS search result page.

3.4.3 Web Server Framework

The web server framework of DnsID is relatively simple. The current Dns-library consists of 315 metabolites and should be expandable in the future, because of the size of library we don't use database to store the information of metabolites which maybe implemented with the size of Dns-library growing up. The framework is also similar to the MCID MS/MS project (See Figure 3.11).

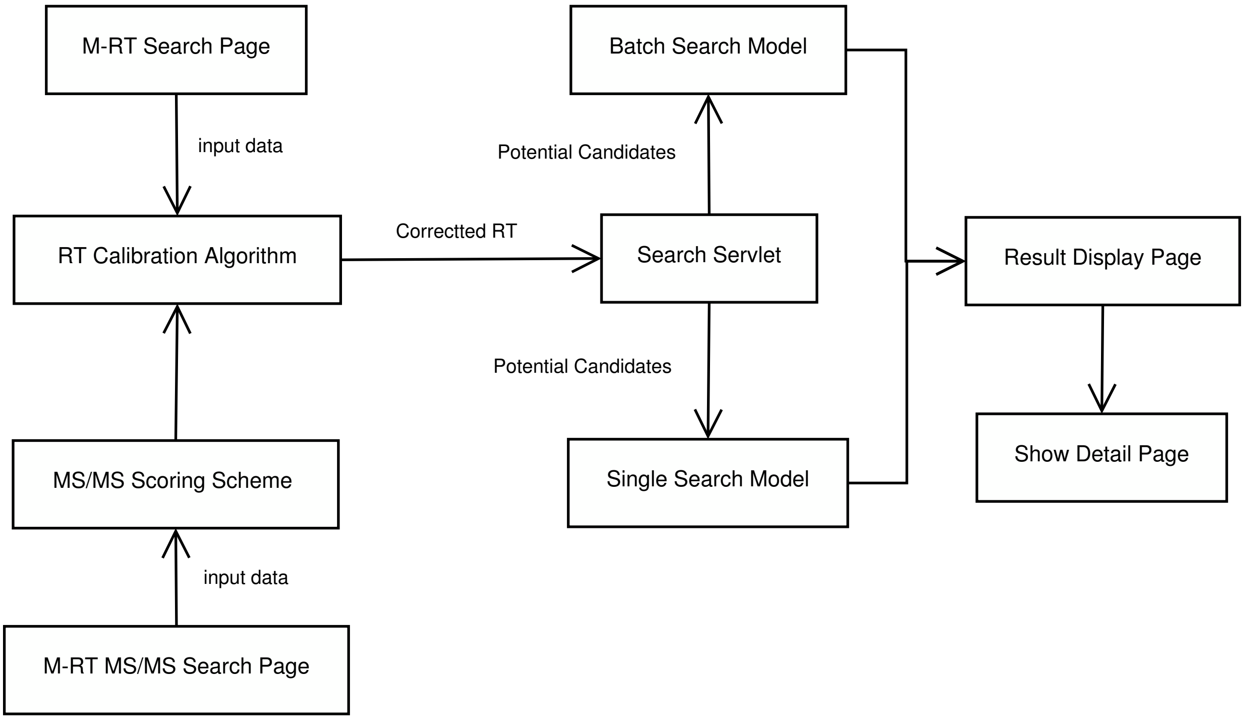


Figure 3.11: M-RT web server framework.

3.5 Conclusions and Future Work

We have developed a dansyl standards library and a library search program, DnsID, for rapid identification of metabolites in dansylation LC-MS targeting the analysis of the amine/phenol submetabolome. For each Dns-metabolite, accurate mass, MS/MS spectrum and retention time are included. To overcome the problem of RT shifts often found in LC-MS data sets collected using different experimental conditions, a RT calibration method based on the use of a mixture of 10 or 22

Dns-standards eluted across the entire retention space in RPLC has been developed. The retention time of Dns-metabolite in Dns-library is a normalized RT against the RTcal, allowing comparison of the library values with those of a sample obtained under slightly different LC-MS conditions after applying the RT calibration. This library along with the DnsID search program is freely accessible from http://mcid.cs.ualberta.ca:8080/Compound_MRT/. We demonstrate that DnsID can be used to perform M-RT search for metabolite identification with high confidence in the LC-MS data obtained from a dansyl labeled human urine sample. MS/MS spectral search can be used to provide additional confidence for metabolite identification. In addition, MS/MS search can be used to find related metabolites including derivatives and isomers of the library standards.

Bibliography

[1] B. Daviss. Growing pains for metabolomics. *The Scientist Magzine*, 19:25–28, 2005.

[2] L. S. Ettre. Iupac nomenclature for chromatography iupac recommendations. *Pure and Applied Chemistry*, 65:819–872, 1993.

[3] H. Horai, M. Arita, S. Kanaya, T. Ni, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, and K. Aoshima. Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45:703–714, 2010.

[4] H. Jenkins, H. Hardy, M. Beckmann, J. Draper, A. R. Smith, J. Taylor, O. Fiehn, R. Goodacre, R. J. Bino, R. Hall, J. Kopka, G. A. Lane, B. M. Lange, J. R. Liu, P. Mendes, B. J. Nikolau, S. G. Oliver, N. W. Paton, S. Rhee, U. RoessnerTunali, K. Saito, J. Smedsgaard, L. W. Sumner, T. Wang, S. Walsh, E. Wurtele, and D. Kell. A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology*, 22:1601–1606, 2004.

[5] K. W. Jordan, J. Nordenstam, G. Y. Lauwers, D. A. Rothenberger, K. Alavi, M. Garwood, and L. L. Cheng. Metabolomic characterization of human rectal adenocarcinoma with intact tissue magnetic resonance spectroscopy. *Diseases of the Colon Rectum*, 52:520–525, 2010.

[6] L. Li, R. Li, J. Zhou, A. Zuniga, A. Stanislaus, Y. Wu, T. Huan, J. Zheng, Y. Shi, D. S. Wishart, and G. Lin. Mycompoundid: using an evidence-based metabolome library for metabolite identification. *Analytical Chemistry*, 53:41–48, 2014.

[7] C. Ludwig and M. R. Viant. Two-dimensional resolved nmr spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis*, page 111, 2010.

[8] A. McNaught and A. Wilkinson. *Compendium of Chemical Terminology*, volume 1669. Blackwell Science Oxford, 1997.

[9] J. Nicholson, J. Lindon, and E. Holmes. Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological nmr spectroscopic data. *Xenobiotica*, page 1181, 1999.

[10] B. J. Pine. *Mass Customization: The New Frontier in Business Competition*. Harvard Business Review Press, 1992.

[11] C. Smith, G. O'Maille, E. Want, C. Qin, S. Trauger, T. Brandon, D. Custodio, R. Abagyan, and G. Siuzdak. Metlin: a metabolite mass spectral database. *Therapeutic Drug Monitoring*, 27:747–751, 2005.

[12] M. B. Smith and P. March. *Advanced Organic Chemistry*. John Wiley., Inc., 6 edition, 2008.

[13] S. E. Stein and D. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 9:859–866, 1994.

[14] W. C. Still, M. Kahn, and A. J. Mitra. Safe and efficient flash chromatography equipment for the research. *Journal of Organic Chemistry*, 43:2923–2925, 1978.

[15] M. Wilchek and I. Chaiken. *An Overview of Affinity Chromatography*. Humana Press, 2007.

[16] D. Wishart, D. Tzur, C. Knox, R. Eisner, A. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, and S. Sawhney. Hmdb: the human metabolome database. *Nucleic Acids Research*, 35:D521–D526, 2007.