University of Alberta

Human Metabolite Identification through Web-based Applications

by

Ronghong Li

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Fall 2013
Edmonton, Alberta

Abstract

High throughput bio-technologies in chemistry experiments generate a huge amount of data. These data are awaiting to be analyzed for knowledge discovery.

In this work, we have developed a web-based resource MyCompoundID for compound identification. Our base database contains 8,021 metabolite substrates imported from Human Metabolome Database, and we adopt 76 the most commonly encountered biotransformations collected from the literature. We first expand the database to include all the pseudo metabolic products for up to two reactions, which are filtered by multiple levels of restrictions we specified. Using MyCompoundID for compound identification, either through mass queries or MS/MS spectrum queries, can identify many more unknown metabolites than using the existing works.

Acknowledgements

First and foremost I would like to thank my supervisors, Dr. Guohui Lin and Dr. Liang Li, for their assistance and feedback during my master studies. I am indebted to my two supervisors as they led me into this chemo-informatics research field and proposed many interesting projects and gave me a chance to work in this field. I would like to give my sincere thanks to PDF Dr. Jianjun Zhou, and a graduate student in my lab, Dr. Yi Shi, who help me start the MyCompoundID project.

Tao Huan, a PhD student of Dr. Liang Li in Chemistry Department, deserves my sincere thanks for his collaboration on digging out metabolic reactions from the literature. Without this collaboration, the Restriction on Structures in MyCompoundID is impossible to implement.

I am also grateful to the developers of the open source library Chemistry Development Kit. They have implemented certainly great data structures and wonderful file IO interfaces. Without their open source code, it would take much longer to implement the Restriction on Structures in MyCompoundID.

Lastly, I would like to thank my family for their years of support, without which this work would not have been possible.

Table of Contents

List of Tables

List of Figures

List of Symbols

+formula	reaction adding the structure presented by the formula
−formula	reaction deleting the structure presented by the formula
A_i	the i-th amino acid in a small peptide
ASCII	American standard code for information interchange
Atom_1–Atom_2	Atom_1 and Atom_2 connected by a single bond
Atom_1=Atom_2	Atom_1 and Atom_2 connected by a double bond
C	carbon atom
CDK	Chemistry Development Kit
CG-MS	gas chromatography mass spectrometry
Da	Dalton, the atomic mass unit
FT-MS	Fourier transform mass spectrometer
H	hydrogen atom
HMDB	Human Metabolome Database
idf	inverse document frequency
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC-MS	liquid chromatography mass spectrometry
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MVC	Model-View-Controller framework
MyCompoundID	My Compound Identification
NCBI	National Center for Biotechnology Information
N	nitrogen atom
NMR	Nuclear Magnetic Resonance
O	oxygen atom
p	a peak in a spectrum
P	a set of peaks
P	phosphor atom
ppm	parts per million
s	a mass spectrum
S	a set of mass spectra
S_2	the set of MS/MS spectra for di-peptides
S	sulfur atom
SMILES	Simplified Molecular-Input Line-Entry System
tf	term frequency
TOF-MS	time of flight mass spectrometer

List of Terms

- **Accurate Mass**
  Experimentally determined mass of an ion of known charge.

- **adduct ion**
  Ion formed by the interaction of a precursor ion with one or more atoms or molecules to form an ion containing all the constituent atoms of the precursor ion as well as the additional atoms from the associated atoms or molecules.

- ***b*-ion**
  Fragment ion containing the peptide N-terminus formed upon dissociation of a peptide ion at the peptide backbone C-N bond.

- **charge number,**z
  Absolute value of charge of an ion divided by the value of the elementary charge (e) rounded to the nearest integer.
  See also m/z.

- **chemical ionization (CI)**
  Formation of a newion in the gas phase by the reaction of a neutral with an ion. The process may involve transfer of an electron, a proton, or other charged species between the reactants.

- **dalton, Da**
  See *unified atomic mass unit.*

- **Fourier transform mass spectrometry,FT-MS**
  Mass spectrometry technique in which m/zvalues are represented by frequencies of ion motion and mass spectra are generated by Fourier transform mathematical operations from time domain transients produced by image current detection.

- **fragmentation reaction**
  Reaction of an ion that results in two or more fragments of which at least one is an ion.

- **fragment ion**
  Product ion that results from the dissociation of a precursor ion.

- **gas chromatography-mass spectrometry (GC-MS)**
  gas chromatography/mass spectrometry (GC/MS)
  Technique by which a mixture is separated into individual components by gas chromatography, followed by detection with a *mass spectrometer.*
  component in a single chromatographic peak.

- **ion**
  Atomic, molecular, or radical species with a non-zero net electric charge [21].

- **liquid chromatography-mass spectrometry (LC-MS)**
  liquid chromatography/mass spectrometry (LC/MS)

Technique by which a mixture of analytes is separated into individual components by liquid chromatography (typically high-performance liquid chromatography), followed by detection with a *mass spectrometer*.

- **mass peak, peak** (in mass spectrometry)
  Localized region of relatively intense detector response in a mass spectrum when ions of a specified *m/z* are detected. If resolving power is insufficient two or more components of similar *m/z* may contribute to one unresolved mass peak.
  Note 1: Although mass peaks are often associated with particular ions, the terms peak and ion should not be used interchangeably.
  Note 2: Care should be used to distinguish mass spectrum peaks from chromatographic peaks in GC-MS and LC-MS [14, 21].

- **mass spectrometer**
  Instrument that measures the *m/z* values and abundances of gas-phase ions.

- **mass spectrometry**
  Study of matter through the formation of gas-phase ions that are characterized using mass spectrometers by their mass, charge, structure, and/or physicochemical properties [14, 21].

- **mass spectrometry/mass spectrometry (MS/MS)**
  tandem mass spectrometry
  Acquisition and study of the spectra of the product ions or precursor ions of *m/z* selected ions, or of precursor ions of a selected neutral mass loss [14].

- **mass spectrum**
  Plot of the relative abundances of ions forming a beam or other collection as a function of their *m/z* values. [14, 21].
  Note: The term is a misnomer because it is *m/z* rather than mass that is the independent variable in a *mass spectrum.*

- ***m/z***
  Deprecated: mass-to-charge ratio, Thomson.
  Abbreviation representing the dimensionless quantity formed by dividing the ratio of the mass of anion to the unified atomic mass unit, by its charge number (regardless of sign). The abbreviation is written in italicized lowercase letters with no spaces.

- **time-of-flight mass spectrometer (TOF-MS)**
  Mass spectrometer that separates ions by *m/z* in a field-free region after acceleration through a fixed accelerating potential. Ions of the same initial translational energy and different *m/z* require different times to traverse a given distance in the field-free region [14, 21].

- **unified atomic mass unit,** u
  Non-SI unit of mass defined as one-twelfth of the mass of one atom of ^{12}C at rest in its ground state and equal to $1.660538921(73) \times 10^{-27}$ kg where the digits in parentheses indicate the estimated uncertainty in the final two digits of the value. Equivalent to the dalton (Da) unit.

- y**-ion**
  Fragment ion containing the peptide C-terminus formed upon dissociation of a peptide ion at the peptide backbone C-N bond.

Preface

Cheminformatics, a.k.a. chemoinformatics or chemical informatics, is an emerging field of study where computing techniques are applied for solving a wide range of problems in chemistry science. This interdisciplinary research pools knowledge and expertise from computing science, chemistry, biology, medical sciences and beyond. Typically, for an underlying problem, batches of chemistry experiments are designed and conducted, through which a significant amount of numerical data are collected; one then seeks to analyze these data and draw biological and chemical conclusions based on the analytical results. The large amount of collected data in general cannot be routinely examined without computer programs. In cheminformatics, computer scientists design programs that conveniently help chemists to filter, to sort, to search, and to analyze their raw data, and enable chemists to achieve their research goals within a reasonable time frame.

Metabolomics is a sub-area of studies in cheminformatics, and it studies the chemical processes involving metabolites. Among others, identifying the metabolites extracted in a sample is a critical step in metabolomics. Nevertheless, due to many difficulties, both methods for metabolite identification and construction of metabolite libraries are under studied; the situation becomes even worse if compared with the current fast developing high throughput bio-technologies in chemistry experiments, which generates huge amounts of data awaiting to be scientifically analyzed. In other words, analytical tools are not developed as rapidly as the hardware technologies, and metabolomics has become an efficiency bottleneck in chemistry, biological and medical research.

Motivated by the practical needs, in this thesis work we aim to develop a pipeline of tools for metabolomics research. We first construct a web-based database to organize and publish both raw data and interpreted data out of high throughput chemistry experiments, in particular the mass spectral data in our case; due to the collaborative nature, this web-based database centralizes the research results produced at multiple labs at different regions of the world. We then design algorithms for data analysis, including single compound identification and whole sample metabolite profiling; we analyze theoretically the algorithms to suit for the largest range of chemical experiment instruments and protocols. These algorithms are implemented into web-based services that can be accessed world-wide. With our pipeline of tools, it is expected that the raw data generated out of chemical experiments can be used to the largest extent, and in the most efficient way. Besides using our tools for their own scientific discovery, users can also contribute to our metabolite library by confirmation of the predicted compounds being identified in their work, together with their interpreted mass spectra.

Thesis Layout

We build a pipeline of web-based tools for metabolite identification, with possible extension to sample metabolite profiling in the near future.

In Chapter 1, we give a brief introduction to some basic concepts and background knowledge pertaining to our research, and the related work. Our project description is presented in Chapter 2, where we state our objectives and solutions to the problems inside the project. The science leading to our web-based metabolite database, and the detailed construction are presented in Chapter 3. In particular in Section 3.6, we present how to generate the theoretical MS/MS spectrum for a metabolite in our library, and the spectrum matching algorithm together with certain scoring schemes. We then describe in detail the metabolite search engine based on a compound mass in Chapter 4. Experimental results on metabolite identification through mass queries and spectrum queries are presented in Chapter 5, and discussed. We conclude with a number of other features of our web-based database and services, and their future extension to sample metabolite profiling.

Chapter 1

Introduction

Metabolomics is a sub-area of studies in cheminformatics, and it studies the chemical processes involving metabolites. More specifically, metabolomics systematically studies the unique chemical fingerprints that a cellular process leaves behind, aiming to profile its small molecule metabolites [7]. These metabolites are considered as the end products of cellular processes, and the biological fluids from which the metabolites are extracted could reflect the health condition of an individual. In practice, metabolic profiling has been used to detect the physiological changes caused by toxic insult of a chemical or mixture of chemicals [23], and to determine the phenotype changes caused by the genetic manipulation. The latter has a very important application in food industry, for example, to determine the phenotypic changes in a genetically modified plant intended for human consumption, so as to predict the function of unknown genes by comparison with the metabolic perturbations caused by deletion or insertion of a known gene [25, 3].

There are two technological platforms used in metabolomics, which are nuclear magnetic resonance (NMR) spectroscopy [20, 22], and mass spectrometry [8, 17]. Although the unique structural information about the metabolites could be retrieved by NMR, NMR suffers from limitations in sensitivity and chemical resolution. In contract, although less-conclusive structural information are provided by mass spectrometry, the sensitivity and large dynamic range of mass spectrometry allows for the detection of many more metabolites in a single experiment [32].

Among others, identifying the metabolites extracted from a sample is a critical step in metabolomics. Nevertheless, due to many difficulties, both methods for metabolite identification and construction of metabolite libraries are under studied; the situation becomes even worse if compared with the current fast developing high throughput technologies in chemistry experiments, which generates huge amounts of data awaiting to be scientifically analyzed. In other words, analytical tools are not developed as rapidly as the hardware technologies, and metabolomics has become an efficiency bottleneck in chemistry, biological and medical research [2].

Metabolite identification can be done by other experiments. As a continuously improving high throughput biotechnology, mass spectrometry presents unique advantages in metabolite identification. At the forefront, identifying individual molecules in the complex mixtures is accurate with high confidence [24]. (In this work, we use metabolite identification, molecule identification, and compound identification interchangeably.) For example, by using appropriate standards, a molecule can be identified based on its mass with great precision by using a combination of chromatography followed by mass spectrometry. Such a level of precision is even more significant when one considers the fact that glucose is the same molecule when measured from a bacterium to a fly to a human [24]. Another advantage of metabolomics through mass spectrometry is that mass spectral experiments allow for quantitative interpretation, which is difficult in other technologies such as mRNA (gene expression) profiling or protein (gene translation) profiling [24]. The measured metabolite quantities enable the research to adopt further statistical methodologies and databasing approaches. Last but not least, metabolomics through mass spectrometry is believed a more direct measure for a disease state or the action of a drug, because that disease states result ultimately from a change in the biochemistry of a system and most drugs act at the level of biochemistry.

1.1 Mass Spectrometry

Mass spectrometry is both the science and art of displaying the spectra of the masses of a sample of material. It is used for determining the elemental composition of a sample, the masses of particles and of molecules, and for elucidating the chemical structures of molecules, such as peptides and other chemical compounds. With the help of charging field and magnetic field, mass spectrometry works by ionizing chemical compounds in the sample to generate charged molecules or molecule fragments and measuring their mass-to-charge ratios (a.k.a. m/z) [27]. The mass spectrometer, which is the instrument used in mass spectrometry, typically consists of three components: ion source, mass analyzer, and detector [6]. Some portion of the sample are converted into ions by the ionizer. These charged molecules are sorted by the mass analyzer based on their mass-to-charge ratio. Finally the detector measures the value of an indicator quantity and thus provides data for the mass spectra. After the mass spectrometry, the molecules with the different mass-to-charge ratios are recorded and separated.

The mass spectrum is a (relative) intensity vs. mass-to-charge ratio plot representing the molecules or molecule fragments in the sample in a histogram format [21]. The x-axis of a mass 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 by the unified atomic mass unit and by its charge number (positive absolute value) [21, 29]. The y-axis of a mass spectrum represents signal intensities of the ions.

In the spectrum, each peak represents the detection of a kind of molecules or a group of different kinds of molecules. With the information of mass-to-charge ratio by x-axis and the intensity by y-axis, the molecular weight of molecule(s) could be determined. After this, the molecules could be fragmented by applying energy onto them. Another round of mass spectrometry experiment will be run on these fragments to get another mass spectrum, which is also called MS/MS spectrum. This MS/MS spectrum reveals the structure information of the molecule. This process is also called MS/MS spectrometry or tandem mass spectrometry [21].

1.2 Related Work

Metabolomics is a rapidly evolving discipline that is involved in systems biology studies and disease biomarker discoveries [15, 28]. Prior to our work, several online chemical molecule databases had been developed to facilitate the metabolomics research. The National Center for Biotechnology Information (NCBI) maintains the largest database, PubChem [1], of over 100 million chemical molecules, as well as their substance description, and activities against biological assays. These molecules are small in that each contains fewer than 1,000 atoms and 1,000 chemical bonds. The problem with this database in metabolite identification is that the returned possible metabolites predicted by PubChem is too many. The exact metabolites could not be identified from that many possible suggestions from PubChem.

Another important database for metabolomics research is Kyoto Encyclopedia of Genes and Genomes (KEGG), which is a collection of online databases of genomes, enzymatic pathways and biological chemicals. KEGG chemical database [16] contains 16,907 low molecular mass compounds. Both of the above two major libraries contain chemicals of all sorts, including synthetic compounds, and are widely used in compound identification processes. This database is smaller than PubChem. In metabolite identification process, this small database could not be sufficient to propose enough possible metabolites, which would result in lots of unidentified metabolites in the spectral features.

The Human Metabolome Database (HMDB) [31] is a freely available web-based database containing up to 40,250 small molecule metabolites found in the human body, both water-soluble and lipid-soluble, and their detailed information. HMDB is another useful source for metabolite identification, and in fact it is intended to be used for more applications in metabolomics, such as clinical chemistry, biomarker discovery and general education. Compared with PubChem and KEGG, HMDB has two major advantages in human metabolite identification. The first one is that HMDB only contains the human metabolites and related other small molecules, it is a very specific database designed for human metabolite identification. The other advantage of HMDB is that the HMDB supports MS/MS spectra searching, although the number of entries in its MS/MS library is small, but it is expanding over time as more experimental spectra are available. HMDB is the mainly used human metabolite database for identification up to now.

There are some other databases less commonly used than the ones mentioned above. One is called METLIN [26], which is a repository of over 75,000 endogenous and exogenous metabolites that includes metabolites from essentially any living creature, whether it is bacteria, plants, human/animal, etc.. Just like KEGG, METLIN contains lots of metabolites from any species, but less human metabolites than HMDB.

Another minor used database in chemical identification is called MassBank [13]. MassBank is the first public repository of mass spectral data for sharing them among scientific research community. MassBank data are useful for the chemical identification and structure elucidation of chemical compounds detected by mass spectrometry. Although it contains many MS/MS spectra, it is not very good for human metabolite identification. This reason is similar to the PubChem, which is that it contains all sorts of chemical molecules, and is not a very specific database for human metabolites.

1.3 The Motivation

The above mentioned databases, PubChem, KEGG and HMDB, all serve as great references for metabolite identification using the data generated by the high throughput analytical techniques such as liquid chromatography mass spectrometry (LC-MS). Metabolomics research has advanced rapidly in the last decade, but the major analytical challenge remains in metabolite identification of the detected spectral features [5, 12]. In fact, only a very small portion of the spectral features observed in LC-MS can be identified as known compounds [11]. There are two main difficulties, one is that current reference databases are limited as they contain only a small fraction of potential metabolites and their bio-transformational products, and the other is that there is no effective and efficient compound identification algorithm to take full advantages of mass spectral features. To satisfy the practical needs, we aim to develop a web-based pipeline of tools to facilitate the metabolite identification in this thesis, allowing the extension to sample metabolite profiling in the near future.

Chapter 2

The Problems

The NCBI PubChem is an online metabolite database which contains over 100 million (small) chemical molecules. PubChem is considerably larger than the other two major databases KEGG and HMDB, which contain only tens of thousands of compounds. However, it is not necessarily true that using PubChem as reference database one can identify more compounds from the same set of spectral features. Increasing the size of the reference database has seemingly reached its limit in compound identification. We strongly believe that efforts should be put into new approach development to speed up compound identification and to better use the spectral features.

In this thesis, we aim to build a pipeline of novel web-based tools for compound identification, with possible extension to sample metabolite profiling in the near future. Trying to balance the database expanding and efficient compound identification, we design algorithms to better interpret collected spectral features and to mine more hidden but useful information for the identification purposes. We examine the performance of these algorithms substantially and implement them into web-based services.

2.1 The Unique Position

The mass or mass list of a compound is collected by mass spectrometry (including Gas-chromatography mass spectrometry GC-MS and Liquid-chromatography mass spectrometry LC-MS). The unique advantage of such collected mass data is that their error range is much smaller than the mass of a proton, and thus can be immediately used in matching against a compound library. On the other hand, multiple protocols have been developed in tandem mass spectrometry (MS/MS), which enable us to observe the functional groups in a target compound. This motivates us to develop compound identification algorithms through spectrum matching between the theoretical (or predicted) spectrum and the experimental (or interpreted) spectrum.

2.2 The Objectives

The HMDB (Human Metabolome Database) contains about 40 thousand small molecule metabolites found in the human body. Among them, 8,021 endogenous metabolites are considered as metabolite substrates; and many other database entries are products of metabolic (or biotransformation) reactions on these substrates. Indeed, in biological systems, each compound could be involved in some metabolic reactions to produce different metabolic products. Consequently, it is highly possible that the samples analyzed in metabolomics research experiments contain both the substrates and their metabolic products after one or several metabolic reactions.

The compounds that have been identified and documented in the literature constitute only a small fraction of these metabolite substrates and their plausible metabolic products; while the majority is to be uncovered. In the evidence-based metabolome library we have constructed for My Compound Identification (subsequently called MyCompoundID), we take a database construction approach different from the others by collecting only the 8,021 metabolite substrates but also the known commonly encountered metabolic reactions. We examine carefully for all metabolic reactions on their applicabilities. Based on these two types of entities, we are able to generate theoretically all plausible compounds for identification purpose. It is expected that using MyCompoundID, one can identify significantly more compounds for the same set of spectral data. Our way of database construction can maintain the database dynamically, with any level of extendibility. Another advantage is that our database entries, the metabolic products in particular, are labeled as confirmed and plausible; the users can choose to search against only the confirmed entries, and when they identified a plausible compound we can change the label to confirmed.

2.3 Hypothesis

Our hypothesis is as follows:

By including all of known metabolites as well as the metabolic products of some commonly encountered metabolic reactions in the library, many unknowns that are structurally related to the known metabolites can potentially be identified, which could not be identified by any of the existing databases.

The purpose of this work is to investigate the above hypothesis by implementing algorithms and tools to simulate the metabolic reactions in the metabolism to generate the possible metabolic products.

2.4 Solutions to the Problems

In this thesis, we have decided to include all known metabolite substrates and their metabolic products after one or two commonly encountered biotransformation reactions in our web-based database. It is expected that using our database as reference, many unknown compounds that are structurally related to the known metabolite substrates can be identified. The ultimate goal of this work is to achieve the level of compound identification that could not be achieved by any existing databases.

A mass spectral experiment typically generates a set of multiple spectra, each of which is essentially a list of spectral features — masses and intensities of the detected compounds (and/or compound fragments). Similar to peptide and protein identification through mass spectrometry based proteomics, the detected masses are the only clue to compound identification. Our MyCompoundID is designed to efficiently and accurately search for metabolite substrates and their metabolic products that match a query mass.

There are 8,021 metabolite substrates in our web-based database, as well as 76 common biotransformation reactions. Theoretically there could be infinitely many metabolic products by applying even only these 76 reactions. We take a novel approach to resolve the potential efficiency and storage issues. On one hand, one or a series of multiple biotransformations on a metabolite substrate results in a product, of which the mass change can be calculated exactly and instantly; this implies that the inverse task of identification is trivial. However, our true identification task is non-trivial since there could be a large number of series of biotransformations that lead to the same mass change. It is thus simply infeasible to pre-process to generate all (infinitely many) metabolic products for identification purpose.

We first note that each biotransformation requires certain specific sub-structure in the molecule. Therefore, simply predicting metabolic products by assuming that every compound is compatible to any metabolic reaction gives rise to a lot of false positives. The effective way to exclude these false positives, and is implemented into MyCompoundID database, is to verify that the base compound has the specific substructure. However, as one can imagine, verifying molecular structure is a very time consuming step. Also, when a biotransformation can happen on a base compound, it could happen at multiple places giving rise multiple metabolic products, which could in turn require long processing time and much space for storing. To overcome this time and space challenge in MyCompoundID, the structure of the base compounds are verified only when necessary, so are the multiple metabolic products and their structures. One typical necessary moment for structure verification is invoked when the compound has its mass matching a query mass submitted by a user. Such an implementation not only saves storage space, but also saves the pre-processing time; and in this way, the project development time has been shortened tremendously whereby the computing resources are used in a more efficient and balanced manner.

Chapter 3

Compound Database Construction

Our compound database for identification essentially is built on two parts. The first part is the compound library, our base database, which consists of 8,021 known human endogenous metabolites taken from HMDB (Human Metabolome Database) [31]; the second part is the biotransformation library which consists of 76 commonly encountered metabolic reactions extracted from literature.

All chemical compounds are made up of atoms of various types, and every atom consists of a certain number of protons, which determines the atom type, a certain number of neutrons and a certain number of electrons. In this work, we adopt Dalton (Da) as the unit for all molecular masses, and the following constants for calculating the mass for a compound:

- proton mass = 1.00727638Da,
- neutron mass = 1.0086649156Da, and
- electron mass = 0.0005446623Da.

The mass of every compound is simply calculated as the linear sum of the masses of all protons, neutrons and electrons therein. In the sequel, we drop the mass unit Da for simplicity.

3.1 Searching Inputs and Outputs

Our system is aimed to identify metabolites based on the only information of MS spectra and MS/MS spectra. In the mass spectrum, the masses of the metabolites are easily calculated based on the mass-to-charge ratio, where the charge on each metabolites is known. Thus, a spectrum could be interpreted as a mass list. Our database takes the mass list as the input, and return different return outputs for MS search and MS/MS search. In the MS search, each mass peak represents a metabolites, the output for MS search is a list of possible metabolites for each mass peak input; while in the MS/MS search, each mass peak represents a fragment of a metabolites and the whole peak list represents that metabolite, the output for MS/MS search is a list of possible metabolites for the whole list of mass peaks.

3.2 The Base Database

The first part of the base database consists of 8,021 known human endogenous metabolites taken from HMDB [31]. For each of them, we have calculated its mass. These 8,021 masses are plotted in Figure 3.1, where the x-axis is the mass in Da and the y-axis represents the number of metabolites having the same mass. One can see that this distribution is rather discrete due to the small numbers of protons, neutrons and electrons in these metabolites. The maximum mass among them is 3457.700304; the maximum number of compounds sharing the identical mass is 28.

The 76 commonly encountered metabolic reactions we extracted from literature are shown in Table 3.1, where the exact change in atom composition (or the chemical formula), the exact mass difference caused, and the common name for each reaction are listed. Among them, one is paired up with its inverse reaction. We apply these 76 reactions on the 8,021 metabolite substrates in the base database to generate plausible (or pseudo) metabolic products. Theoretically, this enables us to expand the compound library indefinitely.

Table 3.1: The 76 commonly encountered metabolic reactions in literature and used in MyCompoundID. A reaction and its inverse are paired up for clarity.

#	Reaction	Mass Difference (Da)	Description
1	$-H_2$	−2.015650	dehydrogenation
2	$+H_2$	2.015650	hydrogenation
3	$-CH_2$	−14.015650	demethylation
4	$+CH_2$	14.015650	methylation
5	$-NH$	−15.010899	loss of NH
6	$+NH$	15.010899	addition of NH
7	$-O$	−15.994915	loss of oxygen
8	$+O$	15.994915	oxidation

Table 3.1: (Continued)

#	Reaction	Mass Difference (Da)	Description
9	$-NH_3$	−17.026549	loss of ammonia
10	$+NH_3$	17.026549	addition of ammonia
11	$-H_2O$	−18.010565	loss of water
12	$+H_2O$	18.010565	addition of water
13	$-CO$	−27.994915	loss of CO
14	$+CO$	27.994915	addition of CO
15	$-C_2H_4$	−28.031300	loss of C_2H_4
16	$+C_2H_4$	28.031300	addition of C_2H_4
17	$-C_2H_2O$	−42.010565	deacetylation
18	$+C_2H_2O$	42.010565	acetylation
19	$-CO_2$	−43.989830	loss of CO_2
20	$+CO_2$	43.989830	addition of CO_2
21	$SO_3H \rightarrow SH$	−47.984745	sulfonic acid to thiol
22	$SH \rightarrow SO_3H$	47.984745	thiol to sulfonic acid
23	$-C_2H_3NO$	−57.021464	loss of glycine
24	$+C_2H_3NO$	57.021464	glycine conjugation
25	$-SO_3$	−79.956817	loss of sulfate
26	$+SO_3$	79.956817	sulfate conjugation
27	$-HPO_3$	−79.966333	loss of phosphate
28	$+HPO_3$	79.966333	addition of phosphate
29	$-C_4H_3N_3$	−93.032697	loss of cytosine
30	$+C_4H_3N_3$	93.032697	addition of cytosine
31	$-C_4H_2N_2O$	−94.016713	loss of uracil
32	$+C_4H_2N_2O$	94.016713	addition of uracil
33	$-C_3H_5NOS$	−103.009186	loss of cysteine
34	$+C_3H_5NOS$	103.009186	cysteine conjugation
35	$-C_2H_5NO_2S$	−107.004101	loss of taurine
36	$+C_2H_5NO_2S$	107.004101	taurine conjugation
37	$-C_5H_4N_2O$	−108.032363	loss of thymine
38	$+C_5H_4N_2O$	108.032363	addition of thymine
39	$-\ (C_5H_5N_5\ -\ H_2O)$	−117.043930	loss of adenine
40	$+\ (C_5H_5N_5\ -\ H_2O)$	117.043930	addition of adenine
41	$-C_3H_5NO_2S$	−119.004101	loss of S-cysteine
42	$+C_3H_5NO_2S$	119.004101	S-cysteine conjugation
43	$-C_5H_8O_4$	−132.042260	loss of D-ribose
44	$+C_5H_8O_4$	132.042260	addition of D-ribose
45	$-C_5H_3N_5$	−133.038845	loss of guanine
46	$+C_5H_3N_5$	133.038845	addition of guanine
47	$-C_7H_{13}NO_2$	−143.094629	loss of carnitine
48	$+C_7H_{13}NO_2$	143.094629	addition of carnitine
49	$-C_5H_7NO_3S$	−161.014666	loss of N-acetyl-S-cysteine
50	$+C_5H_7NO_3S$	161.014666	addition of N-acetyl-S-cysteine

Table 3.1: (Continued)

#	Reaction	Mass Difference (Da)	Description
51	$-C_6H_{10}O_5$	−162.052825	loss of hexose
52	$+C_6H_{10}O_5$	162.052825	addition of hexose
53	$-C_6H_8O_6$	−176.032090	loss of glucuronic acid
54	$+C_6H_8O_6$	176.032090	addition of glucuronic acid
55	$-C_{10}H_{12}N_2O_4$	−224.079708	loss of thymidine
56	$+C_{10}H_{12}N_2O_4$	224.079708	addition of thymidine
57	$-C_9H_{11}N_3O_4$	−225.074957	loss of cytidine
58	$+C_9H_{11}N_3O_4$	225.074957	addition of cytidine
59	$-C_9H_{10}N_2O_5$	−226.058973	loss of uridine
60	$+C_9H_{10}N_2O_5$	226.058973	addition of uridine
61	$-C_{16}H_{30}O$	−238.229665	loss of palmitic acid
62	$+C_{16}H_{30}O$	238.229665	addition of palmitic acid
63	$-C_6H_{11}O_8P$	−242.019158	loss of glucose-6-phosphate
64	$+C_6H_{11}O_8P$	242.019158	addition of glucose-6-phosphate
65	$-C_{10}H_{11}N_5O_3$	−249.086190	loss of adenosine
66	$+C_{10}H_{11}N_5O_3$	249.086190	addition of adenosine
67	$-C_{10}H_{11}N_5O_4$	−265.081105	loss of guanosine
68	$+C_{10}H_{11}N_5O_4$	265.081105	addition of guanosine
69	$-C_{10}H_{15}N_3O_5S$	−289.073244	loss of glutathione
70	$+C_{10}H_{15}N_3O_5S$	289.073244	addition of glutathione
71	$-C_{10}H_{15}N_3O_6S$	−305.068159	loss of S-glutathione
72	$+C_{10}H_{15}N_3O_6S$	305.068159	addition of S-glutathione
73	$-C_{12}H_{20}O_{10}$	−324.105650	loss of di-hexose
74	$+C_{12}H_{20}O_{10}$	324.105650	addition of di-hexose
75	$-C_{18}H_{30}O_{15}$	−486.158475	loss of tri-hexose
76	$+C_{18}H_{30}O_{15}$	486.158475	addition of tri-hexose

Roughly according to the nature of the reaction, the above 76 metabolic reactions can be grouped into four categories:

- **double bond reactions**: in each of which a double bond is removed from or formed in the target compound (for example Reaction #1 and #11);
- **chopping reactions**: in each of which a group of atoms in the target compound is chopped off (for example Reaction #3);
- **adding reactions**: in each of which a group of atoms is added onto the target compound (for example Reaction #4);
- **substitution reactions**: in each of which a group of atoms in the target

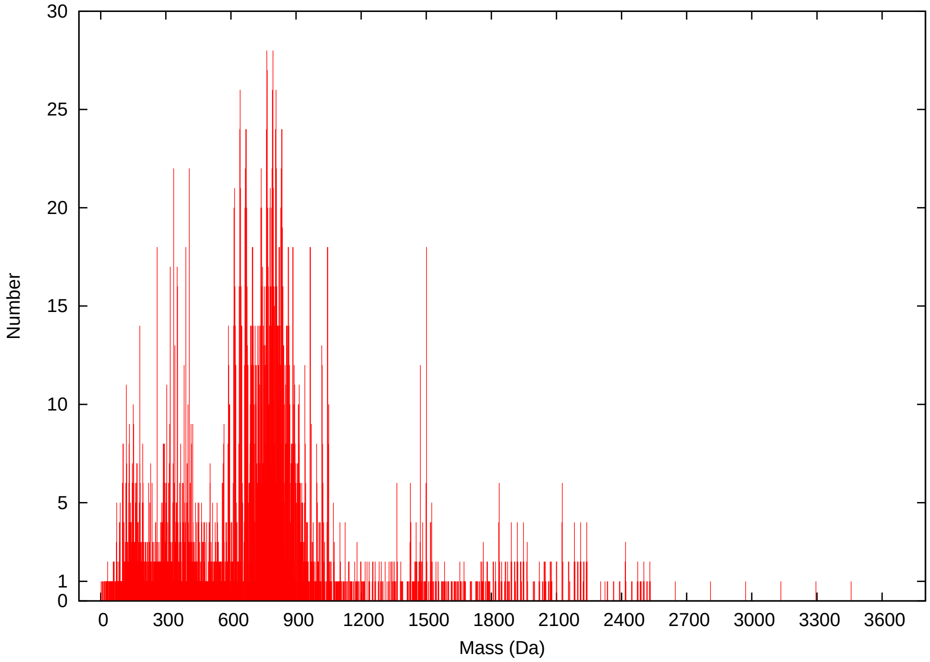


Figure 3.1: The mass distribution of 8,021 metabolite substrates in the base database of MyCompoundID. The x-axis is the mass in Da and the y-axis represents the mass frequency.

compound is substituted with another group of atoms (for example Reactions #21 and #22).

3.3 Database Expanding through Metabolic Reactions

In a mass query, MyCompoundID accepts a mass value, or a list of mass values in the batch mode, as the input. Such a value can come from the detected spectral features with a tolerance threshold. MyCompoundID returns those among the 8,021 metabolite substrates in the base database, and/or the metabolic products of the 8,021 metabolite substrates after certain numbers of metabolic reactions, such that their masses match with the query within the specified tolerance threshold.

3.3.1 The Purposes

Apparently one can generate all metabolic products on the fly for every query, yet one might also imagine that such on-the-fly generating process is time consuming and unnecessary and perhaps can be replaced by expanding the base database to include the metabolic products after a limited number of reactions. Nevertheless, the ending database should not be too large to be housed in a normal-size hard drive. In other words, the limited number shall be small.

Mathematically, the 8,021 metabolite substrates in the base database together with the 76 common metabolic reactions could generate up to 609,596 pseudo (or plausible) metabolic products as the results of one reaction; and could generate up to 46,329,296 pseudo metabolic products as the results of two reactions, such as acetylation followed by dehydrogenation. We have decided in this work to expand the database by generating all pseudo metabolic products as the results of one or two reactions. We develop several levels of database filtering algorithms to eliminate the theoretically impossible products from the nearly 50 million pseudo products. Our experiments have shown that this scheme of trading time with space works effectively.

In this section, we present the first level of filtering algorithms to restrict the compounds to which a specific reaction applies. The next level of more powerful filtering algorithms based on compound structures are presented in the next section. The detailed implementation is described in Sections 3.4 and 3.5.

3.3.2 The Strategy

Chemical reactions vary much from one to another, since different combinations of electron shifting, double bond shifting, ring formation and breaking and so on, can be involved. Besides known rules, there could be outlying situations where a reaction unexpectedly happens, and outlying situations a reaction could theoretically happen but never actually happens. As such, it is very challenging to implement a small set of rules to cover all reactions. On the other hand, describing too many rules into the filtering algorithms leads to a long verifying time.

In this work, we decide to implement a small set of the most common rules on the 76 metabolic reactions. Consequently, we expand the compound library by building up a superset of the metabolic products, which unavoidably contains false positive predictions. Nevertheless, our common rules ensure that (hopefully) no actual metabolic products escape from the library. Later on, on top of this superset, more restricted rules are invoked during the query processes, whereby impossible products are removed. Our strategy basically does a sufficient product generation during the database construction, and distributes the impossible product elimination to the query processes. Over time, the pseudo metabolic products survived in MyCompoundID can be deemed more and more reliable.

3.3.3 Restriction on Atoms

All compounds in this work are made up of a combination of atoms C, H, N, O, S and P, *i.e.* $C_aH_bN_cO_dS_eP_f$ containing a carbon atoms, b hydrogen atoms, etc. Each of the 76 metabolic reactions specifies the net change in the number of atoms of each type, and thus spells exactly the atom combination in the ending product.

Trivially, a compound containing x atoms of one type cannot lose more than x such atoms through reactions. This is exactly the following Restriction 1, which is implemented as a filtering step in our database expanding.

Restriction 1 (Restriction on Atoms) *A compound cannot lose more atoms of a specific type than it has.*

Recall that theoretically there could be up to 609,596 pseudo metabolic products as the results of one reaction and there could be up to 46,329,296 pseudo metabolic products as the results of two reactions. The above seemingly trivial Restriction 1 reduces the numbers to 413,907 and 12,536,913, respectively. Considering the fact that Restriction 1 does nothing to "adding" reactions, the reduction from 609,596 pseudo one-reaction products to 413,907 eliminates 64.2% "chopping" reactions. Indeed, we examined the base database and found out that only 380 (out of 8,021) substrates contain atom S, and only these substrates can participate the 9 reactions involving an S (Reactions #21, #22, #25, #33, #35, #41, #49, #69, #71).

3.3.4 Restriction on Bonds

All atoms in a compound are connected via chemical bonds. For ease of presentation we define a *chopping* metabolic reaction to be one that breaks a single chemical bond in a compound to produce a pair of fragments. Subsequently, the compound loses one fragment, which is said *chopped off* from the target compound. Clearly, the group of chopped atoms specified by a chopping reaction are not randomly distributed in the target compound, but need to be connected and they as a group must be connected to the rest via a single chemical bond.

The above background knowledge is described as the second restriction — Restriction on Bonds —- stated as follows:

Restriction 2 (Restriction on Bonds) *A compound can only lose a terminal group of connected atoms by breaking exactly a single chemical bond.*

Implementing Restriction 2 as a filtering step and applying it further reduces the number of false positives by chopping reactions and substitution reactions. For instance, the total number of one-reaction pseudo metabolic products is reduced to 346,784, or a 61.5% reduction. Applying this filtering step, essentially every chemical bond in the target compound needs to be checked, and thus is time consuming. We did not run it on two-reaction pseudo metabolic products.

3.3.5 Restrictions on Structures

Chemical reactions are essentially molecular structure changes. Every metabolic reaction in our 76 ones has its unique characteristics on where it can happen. This requires the target compound to have the substructure specified by the reaction. In this sense, the above Restrictions on Atoms and Restrictions on Bonds are only coarse implementations, yet they are common to all reactions and they can filter efficiently.

These coarse implementations could pass some false positive one-reaction metabolic products, which give rise to a huge number of false positive two-reaction metabolic products and even more afterwards. In other words, eliminating a false positive one-reaction metabolic product can have a big impact. We therefore carefully examine each of the 76 metabolic reactions and characterize its applicable substructures. The observations are implemented into a series of reaction rules, to be detailed in Section 3.5. In general, a chopping metabolic reaction or a substitution metabolic reaction looks for a very specific group of atoms which are connected in a very specific way in the target compound; the group of atoms in an adding metabolic reaction looks for a specific local structure in the target compound. Otherwise, such a reaction does not apply. Overall, the set of reaction specific rules are based on compound structures, and they form the following Restriction 3:

Restriction 3 (Restriction on Structures) *Every metabolic reaction requires a very specific local structure in the target compound.*

To apply Restriction on Structures as a filtering step, we verify the two-dimensional structures of all 8,021 metabolite substrates in the base database. While leaving the details of the verification to the next section, as a result we further reduce the total number of one-reaction metabolic products to 206,811. Again, such a structure verification process is time consuming, and we did not apply them on two-reaction products. Nevertheless, the verification process is invoked on the search results returned for a user query. In other words, two-reaction metabolic products are verified by Restriction on Structures when they become potential identified compounds.

3.3.6 Summary

The following Table 3.2 summarizes the numbers of compounds during the process of expanding MyCompoundID library. The ending library before any query process contains 8,021 metabolite substrates, 206,811 one-reaction metabolic products, and 12,536,913 two-reaction metabolic products, or 12,751,745 compounds in total readily for search.

Table 3.2: Summary of the MyCompoundID compound library. Each entry represents the number of compounds.

Restrictions	Two-reaction products	One-reaction products	Substrates
–	46,329,296	609,596	8,021
Atoms	12,536,913	413,907	
Bonds	–	346,784	
Structures	–	206,811	

Table 3.3: Summary of search time in MyCompoundID using 1,000 masses. The search was repeated on the set of metabolic products for 0, 1, and 2 reactions respectively. The collected time in the second and last columns are on the base database and the expanded database, respectively. All time are in CPU seconds.

	Base Database	Expanded Database
Disk space	866KB	480MB
0 reaction	3.4s	3.4s
1 reaction	17.6s	6.8s
2 reactions	589.2s	79.5s

In MyCompoundID, users have options to perform a single mass search or a batch search for multiple masses. Users can also specify whether they only want to search against metabolite substrates, or one-reaction pseudo products, or two reaction pseudo products. The reference search library thus contains 8,021, or 206,811, or 12,536,913 compounds, respectively. Table 3.3 summarizes the search time in all three cases. A total of 1,000 (randomly generated) masses were used. The time in the middle column were collected when the base database (around 866KB disk space) is not expanded beforehand, but pseudo metabolic products were generated during the queries; the time in the last column were collected on expanded databases (around 480MB disk space). The results show that space trades very well with search time.

Figure 3.2 plots the logarithm (base 10) of the numbers of pseudo metabolic products up to two reactions. As expected, the number of metabolic products increases exponentially in the number of the reactions. We distinguish the two cases on whether the three filtering steps are applied. Clearly seen from the plot that the filtering steps are effective and they successfully remove a big portion of false positives.

Figure 3.3 plots the average search CPU time for processing one mass query, where the difference is shown between using only the base database of 8,021 metabolite substrates and using the expanded database of 12,751,745 compounds up to 2 reactions. Again as expected, the difference between the search CPU time increases dramatically along with the number of reactions, demonstrating a successful trading space for time.

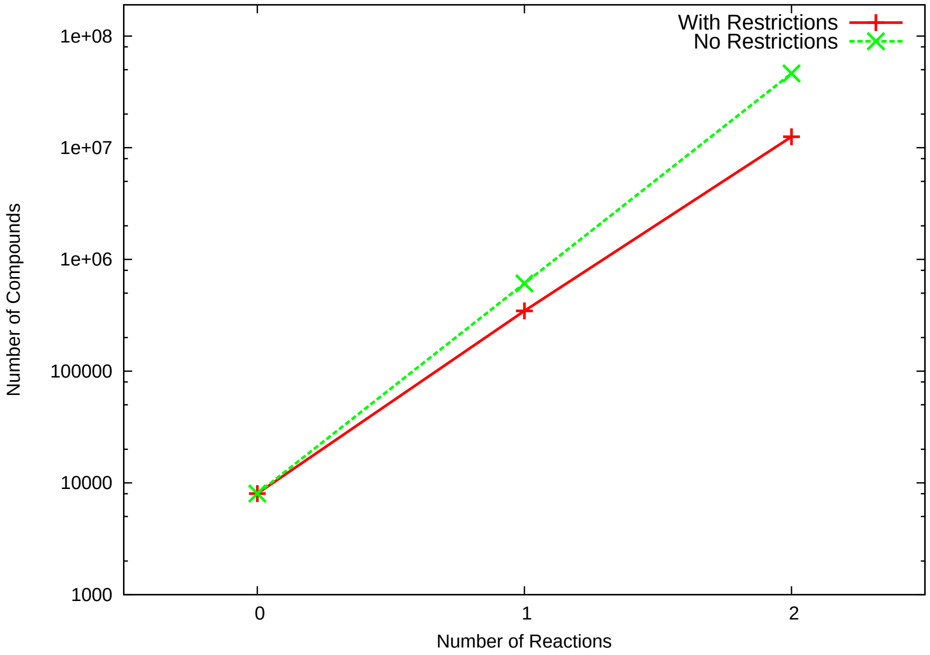


Figure 3.2: Comparison on the logarithm of the numbers of pseudo metabolic products up to 2 reactions, with and without applying the filtering steps, respectively.

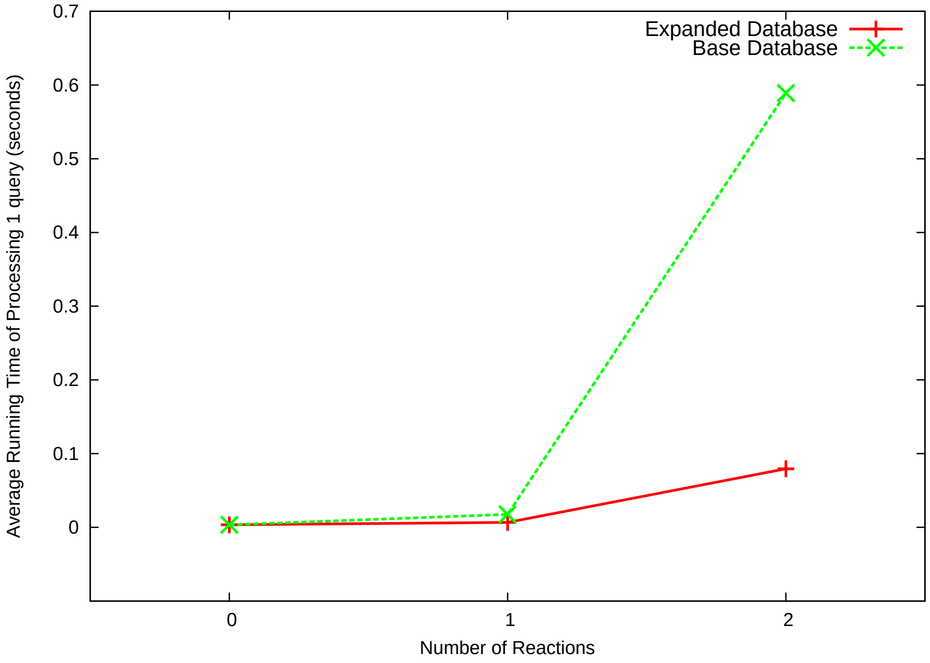


Figure 3.3: The average search CPU time for processing one mass query, using the base database and the expanded database up to two reactions. The average is taken over 1,000 randomly selected masses.

3.4 Pseudo Metabolic Product Filtering Algorithms

In this section, we present the detailed algorithmic processes to implement the rules for filtering impossible metabolic products, as summarized in Restrictions 1–3 in Section 3.3.

3.4.1 Restriction on Atoms

Restriction 1 on Atoms tells that a compound cannot lose in a reaction more atoms of one type than it has. This rule is efficient to check, and thus is applied on both metabolite substrates and one-reaction metabolic products.

For each one of such compounds, we retrieve its chemical formula, which tells exactly the numbers of atoms of all six types in the compound. We then proceed to determine whether a reaction is applicable to the compound as follow:

i) if the reaction is an "adding" reaction, then it is applicable and the product is generated by returning its chemical formula;

ii) if the reaction is a "chopping" reaction, then it is applicable only if the compound contains an equal or greater number of atoms of each type;

iii) if the reaction is a "substitution" reaction, such as Sulfonic acid to Thiol (#21 in Table 3.1) or Adenine to water (#39 in Table 3.1), then it is treated the same as "chopping" the first group off the compound. In other words, a substitution reaction is considered as a chopping reaction followed by an adding reaction.

The above has been implemented as Algorithm 1, where only the 8,021 metabolite substrates are processed. It can be directly migrated to process all one-reaction products.

As mentioned in Section 3.3.3, Algorithm 1 reduces the pseudo one-reaction products from 609,596 to 413,907, and reduces the pseudo two-reaction products from 46,329,296 to 12,536,913, respectively.

3.4.2 Restriction on Bonds

Restriction 2 specifies a more restricted rule for chopping reactions where only a terminal group of connected atoms can be chopped off through breaking a single chemical bond. It is expected to enhance the filtering performance, but one should bear in mind that verifying chemical bonds in the target compound is more time

Algorithm 1 Filtering algorithm by Restriction on Atoms.

```
applicable[8021][76] := **false**; mass[8021][76] := $-1$;
**for** compound $c$ in 8,021 metabolite substrates **do**
   **for** reaction $r$ in 76 reactions **do**
      **if** $r$.type == adding **then**
         applicable[$c$.index][$r$.index] := **true**;
      **else if** $r$.type == chopping **or** $r$.type == substitution **then**
         $o$ := $r$.chopOffPart;
         applicable[$c$.index][$r$.index] := **true**;
         **for** atomType $at$ in $o$.atomTypes **do**
            **if** $c$.countAtomsOf($at$) $<$ $o$.countAtomsOf($at$) **then**
               applicable[$c$.index][$r$.index] := **false**;
            **end if**
         **end for**
      **end if**
      **if** applicable[$c$.index][$r$.index] == **true** **then**
         mass[$c$.index][$r$.index] = $c$.mass + $r$.massOffset;
      **end if**
   **end for**
**end for**
```

consuming than simply checking the number of atoms. This rule is applied on top of Restriction 1 to further reduce false positive one-reaction pseudo metabolic products, but not applied for reducing the two-reaction products.

For each one-reaction pseudo metabolic product generated by a compound and a reaction and passed filtering Algorithm 1, we proceed as follows:

i) if the reaction is an “adding” reaction, then it is still applicable;

ii) if the reaction is a “chopping” reaction, then the compound is examined by breaking every single chemical bond to generate two fragments and checking whether any one of the two fragments is identical to the reaction, *i.e.* the fragment and the reaction have identical chemical formula; if affirmative, then the reaction is applicable to the compound;

iii) if the reaction is a “substitution” reaction, then it is treated the same as “chopping” off the first group off the compound; and again such a reaction is deemed a chopping followed by an adding.

The above has been implemented as Algorithm 2 to filter only the one-reaction pseudo metabolic products. Theoretically one can also migrate it to reduce the two-reaction products, but we did not do so as it can take very long time.

Algorithm 2 Filtering algorithm by Restriction on Bonds.

```
applicable[8021][76] := **false**; mass[8021][76] := $-1$;
**for** compound $c$ in 8,021 metabolite substrates **do**
   **for** reaction $r$ in 76 reactions **do**
      **if** $r$.type == adding **then**
         applicable[$c$.index][$r$.index] := **true**;
      **else if** $r$.type == chopping **or** $r$.type == substitution **then**
         $o$ := $r$.chopOffPart;
         **for** bond $b$ in $c$.getSingleBonds **do**
            ($f_1$, $f_2$) := $c$.breakIntoTwoFragmentsAtBond($b$);
            **if** $f_1$.formula == $o$.formula **or** $f_2$.formula == $o$.formula **then**
               applicable[$c$.index][$r$.index] := **true**;
               break;
            **end if**
         **end for**
      **end if**
      **if** applicable[$c$.index][$r$.index] == **true then**
         mass[$c$.index][$r$.index] = $c$.mass + $r$.massOffset;
      **end if**
   **end for**
**end for**
```

Algorithm 2 reduces the 413,907 pseudo one-reaction products passed by Algorithm 1 down to 346,784.

3.4.3 Restriction on Structures

In Restriction 3 on Structures, the two-dimensional structure of a compound is fully examined to determine whether a metabolic reaction is applicable to it. Here the two-dimensional structure refers to the full details for each atom its adjacent atoms and the chemical bonds connecting them. The structure is converted into a special weighted graph, in which the vertices represent the atoms and the edges represent the chemical bonds. Every vertex has its atom type, and every edge weight represents the order of the chemical bond.

Restriction 3 says that every metabolic reaction requires a specific local structure in the target compound to be applicable. Seeking such a local structure is a special case of the general *graph isomorphism* problem [10].

We first define this special local structure with respect to a metabolic chopping reaction as following:

Definition 1 (Target Subgraph and Its Remaining Subgraph) *In the target* *graph, a* target subgraph *or* target local structure *is a subgraph which*

- *i) is isomorphic to the two-dimensional structure of the chopping reaction, and*
- *ii) is connected to the rest of the graph via a single edge.*

Deleting the target subgraph from the graph gives rise to the remaining subgraph *associated with the target subgraph.*

With respect to a metabolic adding reaction, a *target vertex* or *target local structure* in the target graph is simply a vertex to which the adding reaction can attach the two-dimensional structure of the reaction chemical group.

Restriction 3 on Structures thus checks for whether or not there are target subgraphs or target vertices in the graph, and can be detailed as follows:

- i) if the graph contains target subgraphs or target vertices, the reaction is applicable;
- ii) if the reaction is an "adding" reaction, the pseudo metabolic products are generated by attaching the reaction chemical group to one target vertex at a time;
- iii) if the reaction is a "chopping" metabolic reaction, the pseudo metabolic products are generated by deleting a reaction chemical group from the target vertices one at a time;
- iv) if the reaction is a "substitution" metabolic reaction, the pseudo metabolic products are generated by replacing one target subgraph at a time with the replacement chemical group.

Rules in Restriction 3 are implemented into three different algorithms according to the type of reaction is involved, to be detailed in the next section of "metabolic reaction simulation system". One clearly expects that these new filtering algorithms are more time consuming than Algorithms 1 and 2, but should be more powerful in eliminating false positives. This is indeed true. After applying these three new filtering algorithms, the total number of one-reaction pseudo metabolic products reduces to 206,811.

3.5 Metabolic Reaction Simulation System

In this section we describe the *metabolic reaction simulation system*, which aims to implement Restriction 3 into three different filtering algorithms for three types of reactions, respectively. On top of Algorithms 1 and 2, these three new filtering algorithms further reduce the total number of one-reaction pseudo metabolic products from 346,784 to 206,811. They also generate all possible two-dimensional structures for each such product, and the total number of structures for these 206,811 metabolic products is 2,135,541, or about 10 structures per product. One obvious reason is that an adding reaction can be applied to multiple atoms in a metabolite substrate.

Theoretically applying (all) these filtering algorithms on two-reaction pseudo metabolic products is trivial — one only needs to replace the 8,021 metabolite substrates with the 206,811 one-reaction products (and their 2,135,541 two-dimensional structures). However, this needs a huge CPU time that one might not be affordable.

3.5.1 The System

The *metabolic reaction simulation system* consists a set of utilities for simulating the metabolic reactions. Each metabolic reaction simulation in the system works on the two-dimensional structures of compounds, to simulate the reaction processes similar to the reality. This system is designed to generate the pseudo metabolic products based on Restriction 3 on Structures.

3.5.2 Use of Two-Dimensional Structures

To determine whether a metabolic reaction is applicable to a compound, the two-dimensional structure of the compound is examined. Essentially every metabolic reaction requires some local structure features in the target compounds to be applicable. For the simplest case of an adding reaction, where a chemical group is added to the target compound, the chemical group normally has only one or two specific atoms each of which is able to form a chemical bond between itself and the target compound. This requires in turn a few very specific functional groups in target compound structure. If no *match* is found, the adding reaction is not applicable to the target compound.

For example, Reaction #30 in Table 3.1 is “addition of cytosine” (+$C_4H_3N_3$), in which the cytosine molecule looks for a carbon atom in the target compound with a single bond to an oxygen and a single bond to a hydrogen, and this carbon must not be in a benzene ring. Using only filtering Algorithms 1 and 2, implementing Restrictions on Atoms and Bonds, "addition of cytosine" would be applicable to all compounds, generating a lot of false positives.

The above explains the advantages of using two-dimensional structures in the filtering process. The following presents some insights on the choice of two-dimensional structure but not three-dimensional structure, the latter clearly contains more structural information than the former. Consider a chopping reaction. The target compound is examined by breaking a single chemical bond in its two-dimensional structure, to generate two (connected) fragments; each fragment together with its inherited two-dimensional structure is then compared against the chemical group together with its two-dimensional structure, and "applicable" is called when both the chemical formula match and the structure match are confirmed. In general, chemical formula matching can be verified rather easily, but structure matching is a special case of the *graph isomorphism* [10] which is deemed a hard computational problem. At least one factor makes structure matching problem even more difficult, which is, sometimes, a two-dimensional structure can give rise to multiple three-dimensional structures. In reality, chemical compounds fold in the three-dimensional space, some of which are *stereoisomers*, *i.e.* isomeric molecules that have the same chemical formula and sequence of bonded atoms (constitution) but differ in the three-dimensional orientations of their atoms in space.

Nevertheless, we choose to implement two-dimensional structure match but not three-dimensional structure match, based on the following a few observations and the compound identification purposes we set at the very beginning. Firstly, there are compounds having lots of stereoisomers, and some reactions are applicable to all possible stereoisomers. For example, Reaction #51 in Table 3.1 is "loss of hexose" ($-C_6H_{10}O_5$). Hexose has many stereoisomers, all of which are acceptable in compound identification. Three-dimensional structure comparison takes CPU time and disk space, but totally unnecessary in MyCompoundID. Other reactions involving multiple stereoisomers include Reactions #71, #73, and #75 in Table 3.1. Secondly, most compounds do not have no stereoisomers, or even some do, only one stereoisomer is biologically active in human metabolism. For example, for carnitine, only the L-carnitine is biological active [19].

In summary, because of the above reasons, we decided not to consider the stereoisomers during compound identification. This essentially means that two-dimensional structures of compounds and reaction chemical groups are sufficient, and they make our filtering algorithms more efficient.

3.5.3 SMILES Conversion and CDK Library

The *Simplified Molecular-Input Line-Entry System*, or SMILES, is a specification in the form of one-line notation for describing the two-dimensional structure of a chemical compound using short ASCII strings. In the literature, there are algorithms developed to ensure that the same SMILES is generated for a compound disregard the order of atoms in the two-dimensional structure. Indeed, an SMILES string specifies a compound structure without losing any structural information, and is known as having the smallest size comparing with all other formats describing the same two-dimensional structure. SMILES files can be concatenated together in a text file where each line represents a compound structure. Comparing two two-dimensional structures becomes comparing two strings in the SMILES format, which is much easier.

In our work, all the compounds are saved in SMILES format. Thus the disk space for storing all compounds and their two-dimensional structures is reduced to the minimum possible, and the total number of disk accesses for a query is also reduced.

The *Chemistry Development Kit* library, or CDK, is used in two-dimensional structure optimization for data structure and IO support. CDK IO library supports many common chemistry file formats, such as MDLV2000 and SMILES, which are adopted in our MyCompoundID project. The basic data structures in CDK for two-dimensional structure optimization include Atom, Bond, and Molecule objects, all of which are good abstractions for a real chemistry structure. Essentially, a molecule is a container which consists of bonds and atoms; a bond connects two atoms, and is marked as 'SINGLE' or 'DOUBLE' (or others); an atom has a charge, symbol, and other properties.

We develop our filtering algorithms using CDK library, which is quite convenient now as we adopt the CDK data structures to describe the two-dimensional structures. In the next five subsections and Section 3.6, we present the new filtering algorithms for adding reactions, chopping reactions, substitution reactions, double bond reactions, and fragmentation, respectively. All these algorithms are implemented using CDK library.

3.5.4 Algorithm for SMILES Generation

A well implemented utility for generating a SMILES string for a given compound provided by CDK is used in our work. The implementation by CDK is accomplished by a method called CANGEN, which was proposed by Weininger [30] in 1989. CANGEN is a combination of two separated algorithms, which are CANON and GENES. CANON is used in first to label a graph, which represents the molecular structure, with canonical labels, while GENES selects one vertex as the root in the graph and generates the unique SMILES notation by traversing the graph starting with the selected root in a specific order based on the canonical labels.

CANON

The CANON algorithm canonically labels a molecular graph by using an unambiguous function. The canonical label of each vertex is initialized with the information of the its atom information as well as its neighbors'. To avoid the calculation overflow, the canonical labels are replaced by their ranks after sorting. Generating unique SMILES notations requires the unambiguous canonical labels, which is achieved by the extended connectivity and the breaking ties functions described later.

The CANON algorithm is shown in Algorithm 3.

Algorithm 3 CANON Algorithm (Weininger [30])

```
1: Set atomic vector to initial invariants. Go to step 3.
2: Set vector to product of primes corresponding to neighbors' ranks.
3: Sort vector, maintaining stability over previous ranks.
4: Rank atomic vector.
5: If not invariant partitioning, go to step 2.
6: On first pass, save partitioning as symmetry classes.
7: If highest rank is smaller than number of nodes, break ties, go to step 2.
8: ... else done.
```

Four core functions in CANON are described as following:

(a) Initial Graph Invariant Order. Graph theoretical invariants are properties of graphs that are independent of the way a graph is ordered. Six invariants as following are considered: (1) number of connections, (2) number of non-hydrogen bonds, (3) atomic number, (4) sign of charge, (5) absolute charge, and (6) number of attached hydrogens. The atomic vector is initialized as the linear combination of these invariants. For example, the invariants of methyl carbon in pentane (CCCCC) are 1,01,06,0,0,6, respectively, which give the linear description as 10106003, while

Table 3.4: Canonical labeling procedure

Step	canonical labels
1	10106003-20206002-20206002-20206002-10106003
3,4	1-2-2-2-1
2	9-13-18-13-9
3,4	1-2-3-2-1
7	1-4-6-4-2
2	49-173-98-173-2
3,4	1-3-4-3-2
2	25-53-50-58-25
3,4	1-3-5-4-2

the one of methylene carbon is 2026002.

(b) Rank Equivalence. The values in a invariant set are used to order the vertices in the spanning tree, so the exact values are not necessary but the ranks of them are important. To avoid the computing overflow, those values are replaced by small numbers starting with 1 which represent their rankings. Take pentane as an example, the initial invariants 10106003-20206002-20206002-20206002-10106003 become 1-2-2-2-1.

(c) Extended Connectivity Using an Unambiguous Function. There are two types of carbons but three symmetry classes in pentane. To differentiate the symmetry classes, an unambiguous function is used on on each vertex. In this function, each rank is replaced by its corresponding prime starting with 2 and then replaced by the product of all its neighbors' replaced ranks. For example, the pentane's ranks are 1-2-2-2-1, whose corresponding primes are 2,3,3,3,2, and after this function, the ranks in pentane are 9-13-18-13-9.

(d) Breaking Ties. If there is a tie, the algorithm doubles the rank of each vertex, and reduces the value of the first atom, which is tied and with the lowest value in rank, by one. The ranks of pentane is changed from 1-2-3-2-1 to 1-4-6-4-2.

Example on Pentane The canonical labels for pentane are generated as the following procedure with respect to the CANON algorithm in Table 3.4

GENES

GENES is the algorithm which traverses the molecular graph to generate a unique SMILES notation.

The lowest canonically numbered atom vertex is selected as the initial node, or as the root of the graph. the algorithm uses DFS to traverse the rooted graph with high priority to select the vertex with low canonical number at the fork. The SMILES notation generation of cyclic and polycyclic structures could be uniquely calculated by run DFS twice.

3.5.5 Reaction Specifications

Each metabolic reaction needs a clear specification on what the local structure in target compound is required and when it happens what the resultant local structure is. We implement all reactions in the Metabolic Reaction Simulation System. In the following, we define the detailed specification for each type of metabolic reactions.

Chopping Reactions

Each chopping reaction must specify the two-dimensional structure of the chemical group to be chopped off from the target compounds. It optionally specifies to which atoms in the target compound the chopped chemical group attaches.

Adding Reactions

Each adding reaction must specify the following three aspects: 1) the two-dimensional structure of the chemical group to be added to the target compounds, 2) the atom in the added chemical group which will form a bond to a target vertex in the target compounds, and 3) the target vertices in the target compounds.

The atom in the chemical group to be added is usually called a *reaction site*. Most adding reactions have only one reaction site each, while only a few of them have two reaction sites each. In the Metabolic Reaction Simulation System, the reaction sites in an adding reaction are marked using a single electron, which enables us to access the sites directly. The adding reactions are classified into several subtypes, depending on their reaction sites.

Substitution Reactions

In the Metabolic Reaction Simulation System, a substitution reaction is specified as a chopping reaction followed by an adding reaction. These two consecutive reactions must work on the same target vertex, *i.e.* one cannot chop a chemical group from one place in the target compound while adds another chemical group to some other place. For otherwise, they should be regarded as two separate reactions.

Bond Order Reactions

Each bond order reaction specifies the chemical bond between a pair of atoms in the target compound, and transforms the single bond to a double bond, or the other way around. Such a reaction does not apply alone to any target compound, but is always binded to one of the above three types of reactions.

3.5.6 The Algorithm for Adding Reactions

Given an adding reaction, which specifies the two-dimensional structure of the chemical group to be added to the target compounds and its reaction site(s), our algorithm first retrieves the two-dimensional structure of a target compound. The adding reaction also specifies the kinds of target vertices in the target compounds. Our algorithm subsequently identifies all the target vertices in the target compound that match the specification.

For example, Reaction #24 in Table 3.1 is "glycine conjugation" (+C_2H_3NO). Its (only) reaction site is the nitrogen atom and a target vertex must be a carbon with a double bond to an oxygen in the target compound.

We realize that determining the reaction sites for the chemical group in an adding reaction is often easy, but could be challenging to specify the target vertices. In general, the atoms in a target compound have different abilities to form a bond with another atom, and their neighboring atoms (through a bond) can also change such ability. The rules of thumb on the target vertices in the target compounds include:

a) all carbon atoms for adding reactions with a small chemical group;

b) the carbon with a double bond to an oxygen for adding reactions involving an amine group and/or a hydroxyl group;

c) the oxygen for adding reactions involving sulfate and phosphate;

d) the sulfur for adding reactions involving thiol (-SH);

e) the carbon with a single bond to an oxygen for some adding reactions involving nitrogen and/or a hydroxyl group;

f) and specially for Reaction #62 in Table 3.1 ("addition of palmitic acid", +$C_{16}H_{30}O$), the target compound has to have an amine group for attacking the carboxyl group in palmitic acid.

Algorithm 4 Filtering algorithm by Restriction on Structures: adding reactions.

```
$c$ := input(target compound);
$r$ := input(reaction);
if $r$.type == adding then
    $addingType$ := $r$.type.addingType;
    $compoundList$ := markActiveAtom($c$, $addingType$);
    $pseudoProductsList$ = connect($compoundList$, $r$.molecule);
    $pseudoProductsList$ = removeMarkers($pseudoProductsList$);
    $pseudoProductsList$ = checkDuplicates($pseudoProductsList$);
end if
```

```
FUNCTION markActiveAtom($c$, $addingType$)
$compoundList$ := new List;
for Atom $a$ in $c$.atoms do
    if $a$.type == $addingType$.activeAtomType then
        $isTarget$ := checkNeighbors($a$, $addingType$);
        if $isTarget$ == true then
            Compound $c'$ = markOnAtom($a$);
            $compoundList$.add($c'$);
        end if
    end if
end for
RETURN $compoundList$;
```

In the filtering Algorithm 4 we design based on Restriction on Structures for adding reactions, we call both the atoms at the reaction sites and the target vertices *active atoms*. Using the two-dimensional structure, the active atoms in the target compound are identified by iteratively checking the atom type of all atoms not in a ring followed by verifying their neighboring atoms as well as the bonds between them. All the active atoms in the target compound are marked, and each is attached with the chemical group in the adding reaction, respectively, to generate a temporary structure saved in a temporary list.

Next, for each temporary structure saved in the temporary list, a single bond is formed between the two active atoms, one in the reaction group and the other in the target compound. The corresponding markers are then removed, as well as extra neighboring hydrogen atoms to the two active atoms. This gives a pseudo metabolic product associated with its two-dimensional structure. Lastly, duplicate pseudo products are examined and removed by using a hash table of SMILES strings of all the pseudo products.

In Algorithm 4, the function *checkNeighbors* inside function *markActiveAtom* requires the *addingType*, since in different *addingType*s the active atoms are affected by different neighbors.

3.5.7 The Algorithm for Chopping Reactions

Compared with adding reactions, chopping reactions are much easier to implement. In general there are much less rules and these rules are mostly reaction independent.

Algorithm 5 Filtering algorithm by Restriction on Structures: chopping reactions.

```
$c$ := input(target compound);
$r$ := input(reaction);
if $r$.type == chopping then
    List $productList$ := new List;
    List $ringBondList$ := markBondInRings($c$);
    $rm$ := $r$.molecule;
    $reactionSMILES$ = getSMILES($rm$);
    for Bond $b$ in $c$.bonds do
        if $b$ not in $ringBondList$ then
            ($f_1$, $f_2$) := breakAtBond($c$, $b$);
            if hasSameAtomNum($f_1$, $rm$) then
                if $f_1$.getSMILES == $reactioinSMILES$ then
                    $productList$.add($f_2$);
                    continue;
                end if
            else if hasSameAtomNum($f_2$, $rm$) then
                if $f_2$.getSMILES == $reactioinSMILES$ then
                    $productList$.add($f_1$);
                    continue;
                end if
            end if
        end if
    end for
end if
```

Algorithm 5 is the filtering algorithm based on Restriction on Structures where the reaction is chopping off a chemical group from the target compounds. Basically, our algorithm walks through the two dimensional structure of a compound to break one breakable bond at a time. As a result, two fragments are generated, and for each of them the SMILES string is generated from the SMILES string for the target compound, and compared against the SMILES string of the chemical group specified by the reaction. If a match is found, then the other fragment is returned as the pseudo metabolic product of this chopping reaction. Note that our algorithm checks for all breakable bonds and generates a list of such pseudo metabolic products.

Breakable bonds can be specified by the chopping reaction. When the specification is absent, the following rules generally apply: 1) all single bonds are breakable; 2) bonds in rings are not breakable; and 3) bonds connecting a hydrogen is not breakable.

We would like to point out that the most time consuming part in Algorithm 5 is generating the SMILES strings for fragments. In our implementation, for the two fragments resulting from one bond breaking, their SMILES strings are generated only if the chemical formula of the group specified by the reaction matches the formula of one of the fragments. This simple Restriction by Atoms seems to work very well as a pre-screening process.

3.5.8 The Algorithm for Substitution Reactions

A substitution metabolic reaction is logically a chopping reaction followed by an adding reaction, under the constraint that the chemical group must be added to the target compound at the atom from which the other chemical group was chopped off.

There can be two possible ways to implement a substitution metabolic reaction in the Metabolic Reaction Simulation System, using the modules of adding and chopping. If we were to implement using the chopping Algorithm 5, then essentially all breakable bonds in the target compound need to be examined and the two-dimensional structures of resultant fragments have to be compared with the chopping group. This is too time consuming for large target compounds. The actual filtering algorithm we implement is shown in Algorithm 6, which uses the adding Algorithm 4. Basically, our algorithm runs very the same as Algorithm 4 to search for active atoms in the target compound, with a modified definition of an active atom being one that has a single bond connecting to the chopping group. In fact, Algorithm 6 differs from Algorithm 4 only at the definition of an active atom. Again, determining whether there is a chopping group connected to the active atom is firstly by Restriction on Atoms followed by SMILES string matching if needed.

3.5.9 Algorithms for Bond Order Reactions

Double bond utilities are a collection of algorithms implementing 1) adding a double bond, 2) removing a double bond, and 3) rearranging double bonds.

When a double bond needs to be added to a target compound, we first generate a list of single bonds which can potentially change to a double bond. In general,

Algorithm 6 Filtering algorithm by Restriction on Structures: substitution reactions.

```
$c$ := input(target compound);
$r$ := input(reaction);
if $r$.type == substitution then
   addingType := $r$.type.addingType;
   compoundList := markActiveAtom($c$, addingType);
   pseudoProductsList = connect(compoundList, $r$.molecule);
   pseudoProductsList = removeMarkers(pseudoProductsList);
   pseudoProductsList = checkDuplicates(pseudoProductsList);
end if
```

```
FUNCTION markActiveAtom($c$, addingType)
compoundList := new List;
for Atom $a$ in $c$.atoms do
   if $a$.type == addingType.activeAtomType then
      isTarget := checkNeighbors($a$, addingType);
      if isTarget == true then
         if addingType belongs Substitution then
            Atom activeAtom := $a$.connectedActiveAtom;
            Compound $c'$ = chopOffTarget($c$, $a$.targetStructure);
            $c'$ = markOnAtom(activeAtom);
         else
            Compound $c'$ = markOnAtom($a$);
         end if
         compoundList.add($c'$);
      end if
   end if
end for
RETURN compoundList;
```

these single bonds must not be in benzene rings. For every such candidate bond, its two ending atoms must be bonded with free hydrogens (for "dehydrogenation" as Reaction #1 in Table 3.1), or one ending atom is bonded with a hydroxyl group (for "loss of water" as Reaction #11 in Table 3.1). Associated with each bond in the list, a pseudo product and its SMILES string are generated.

These pseudo products are not final because we might have to do double bond rearrangement. Theoretically, after a double bond is formed out of a single bond (as described for example in the above), the resultant structure might not be stable to exist. Typically, when two carbon atoms are connected via a double bond and one of them is also connected to an OH group, *i.e.* a local structure of C=C–OH, the hydrogen in the OH group will move to the other ending carton atom and the double bond moves to connect the carbon and the oxygen, *i.e.* the local structure changes to CH–C=O. Our algorithm walks through each product in the list to fix the positions for double bonds, as well as finalize their SMILES strings.

Removing a double bond from a compound, or more precisely transferring the double bond into a single bond, is easier implement, where the candidates are those double bonds not residing in benzene rings. For each candidate, the two ending atoms are now connected by a single bond, and each of them is attached with a free hydrogen. The corresponding pseudo product is generated, as well as the SMILES string.

3.6 Compound Fragmentation and Spectrum Matching

In this section we describe the algorithms for compound fragmentation to generate a theoretical MS/MS spectrum. As a prototype, we present a spectrum matching algorithm for di-/tri-peptide identification using their experimental MS/MS spectra generated through ToF mass spectrometry.

In an MS experiment, the masses of the compounds can be detected accurately. Compounds of different masses can then be separated and each separated compound can be fragmented to have the fragment masses detected. This multiple steps of mass selection, known as tandem mass spectrometry (MS/MS), generate the MS/MS spectrum for a compound or a set of compounds of identical mass (the latter case is rare). Since fragmentation is not arbitrary, but through breaking certain (yet not fully understood) chemical bonds in the compound, the fragments of a compound make up the fingerprint of that compound. Therefore, MS/MS spectrum match is able to identify the true compound structure, more powerful and reliable than the mass match.

3.6.1 The Fragmentation Algorithm

The theoretical MS/MS spectrum is the collection of all possible fragment masses of a compound. In the rest of this section, we use a fragment and a fragment mass interchangeably. We have decided to generate all fragments of a compound by breaking one possible chemical bond in the compound at a time. Consequently, the two-dimensional structure of the target compound is needed.

The following rules specify which chemical bonds are breakable and/or which chemical bonds are unbreakable. 1) Every carbon–carbon single bond not in a ring structure is breakable, and when broken two fragments are generated. 2) Heteroatoms (sulfur, oxygen, nitrogen, phosphor and halogen elements) could be easily protonated. The chemical bond between a heteroatom and a Carbon is breakable, and when broken two fragments are generated (one hydrogen added onto the heteroatom, and one hydrogen added to the carbon). Note that the resultant fragment containing the heteroatom, or any fragment containing a heteroatom, can be further fragmented by the same rule.

Our fragmentation algorithm first breaks all carbon–carbon single bonds one by one, then proceeds to analyze every heteroatom in the molecule, to generate the largest set of possible fragments. For a general compound, the number of peaks in the theoretical MS/MS spectrum can be an order of magnitude more than the number of peaks in an experimental MS/MS spectrum for the same compound, which makes the spectrum matching a challenging task. Nevertheless, on a subset of compounds that are di-/tri-peptides, their numbers of peaks in theoretical MS/MS spectra are not too large compared against the ones in experimental spectra. Basically, due to the very well studies on peptide MS/MS spectra, their fragmentation patterns are well understood. A peptide is a linear chain of amino acids and its fragmentation is amino acid based (instead of atom based). Associated with one amino acid there are only about a dozen fragments, among which several of them can be frequently observed in typical MS/MS spectra. The theoretical MS/MS spectra we have generated for these 8,400 peptides have 100% coverage, which is defined as the ratio between the number of common peaks and the number of peaks in the experimental spectrum for the target compound.

3.6.2 The Spectrum Matching Algorithm

The MS/MS spectrum matching algorithm we develop in the following is for di-/tri-peptide identification. It serves as a prototype for general compound identification.

We assemble a database to include the theoretical MS/MS spectra for all 8,400 peptides, where each theoretical spectrum contains all possible fragments of the peptide. To develop a scoring scheme for matching an experimental spectrum against a theoretical spectrum, we adopt the well known "term frequency – inverse document frequency" (*tf–idf*) concept from information retrieval. Basically, every peak (the mass of a fragment) is regarded as a "term" (or keyword) and a theoretical spectrum is taken as a document. Term frequency *tf* refers to the number of times the term occurs in a document, while the inverse document frequency *idf* is a measure of whether the term is common or rare across all documents. *tf–idf* is a numerical statistic which reflects how important a term is to a document in a collection of corpus. The *tf–idf* value of a term increases proportionally to the number of times the term appears in a document, but is offset by the frequency of the term in the corpus, which helps control the fact that some terms are generally more common than the others.

Let S be the set of all 8,400 theoretical MS/MS spectra, and P be the set of all peaks across all spectra in S. For every peak $p \in P$ and every spectrum $s \in S$, the frequency of term p in document s is $tf(p, s)$, defined as

$$tf(p, s) = \begin{cases} w(p, s), & \text{if } p \in s, \\ 0, & \text{if } p \notin s, \end{cases} \tag{3.1}$$

where $w(p, s)$ is the weight of peak p in spectrum s depending on the ion type for p. Note that in general different types of fragment ions have different probabilities to be generated. When such probabilities are absent, $w(p, s)$ is set to constant 1.

The inverse document frequency of p in the corpus S, denoted as $idf(p, S)$, is defined as the log of the inverse ratio of documents containing term p:

$$idf(p, S) = \log \frac{|S|}{|\{s \in S : p \in s\}|} \tag{3.2}$$

Theoretically speaking, if the fragment p is an ion type easily formed in a spectrum s, then $w(p, s)$ should be high, indicating more likely p is related to s; consequently, if a fragment p is strongly related to a spectrum s, then the $tf(p, s)$ value is high (equal to $w(p, s)$); also, if the fragment is only weakly related to the other spectra, then the $idf(p, S)$ value is high (denominator is small). The score of matching an experimental spectrum s' and a theoretical spectrum s is $tf\text{–}idf(s', s)$,

$$tf\text{–}idf(s', s) = \sum_{p \in s'} tf(p, s) \times idf(p, S). \tag{3.3}$$

Our preliminary experiments using real MS/MS spectra shows that Equation (3.3) performs excellently for di-peptide identification. For tri-peptides, Equation (3.3) does not perform ideally. One of the main observations is that, for each real spectrum s', the scores $tf\text{–}idf(s', A_1A_2)$ and $tf\text{–}idf(s', A_2A_1)$ are usually close to each other. For di-peptide identification, the score for the true di-peptide is always higher than the score for the reverse di-peptide, since all the peaks in the experimental spectrum contributes to the score. For tri-peptide identification, assuming the true target is $A_1A_2A_3$, theoretically only half of the peaks contribute to scoring A_1A_2 and A_2A_1, and thus score $tf\text{–}idf(s', A_2A_1)$ can be significantly larger than $tf\text{–}idf(s', A_1A_2)$. Nevertheless, when this happens, score $tf\text{–}idf(s', A_2A_3)$ stands out significantly. In summary, using Equation (3.3) for tri-peptide identification is not ideal, but one should seeks for a combination of scores $tf\text{–}idf(s', A_1A_2A_3)$, $tf\text{–}idf(s', A_1A_2)$ and $tf\text{–}idf(s', A_2A_3)$.

Since all the peaks in the experimental spectrum s' contribute to score $tf\text{–}idf(s', s(A_1A_2A_3))$, here $s(A_1A_2A_3)$ is the theoretical spectrum for tri-peptide $A_1A_2A_3$, if indeed the target is a tri-peptide, then about half of the peaks can be used for identify the prefix di-peptide and about half peaks can be used for identify the suffix di-peptide. However, usually identification of prefix and suffix di-peptides are not done both well at the same time, but at least one of them can receive a high score; note that identifying either the prefix or the suffix implies identifying the whole tri-peptide. We therefore decide to take a linear combination of $\mathit{tf}\text{–}\mathit{idf}(s', A_1A_2A_3)$ and $\max\{\mathit{tf}\text{–}\mathit{idf}(s', A_1A_2), \mathit{tf}\text{–}\mathit{idf}(s', A_2A_3)\}$. Since if the target is a di-peptide, we use about 150% of the peaks in score calculation and therefore finalize the match score between s' and $s(A_1A_2A_3)$ as

$$\begin{aligned} \mathit{tf}\text{–}\mathit{idf}(s', s(A_1A_2A_3)) \quad = \quad & \tfrac{2}{3} \times \textstyle\sum_{p \in s'} \mathit{tf}(p, s(A_1A_2A_3)) \times \mathit{idf}(p, S) \\ + \quad & \tfrac{2}{3} \times \max\Big\{ \textstyle\sum_{p \in s'} \mathit{tf}(p, s(A_1A_2)) \times \mathit{idf}(p, S_2), \\ & \qquad \textstyle\sum_{p \in s'} \mathit{tf}(p, s(A_2A_3)) \times \mathit{idf}(p, S_2) \Big\} \end{aligned} \tag{3.4}$$

where S_2 is the set of all 400 di-peptide theoretical MS/MS spectra.

3.7 Data Storage and Database Design

In this work, we use MySQL to store our metabolite compounds and support our web-based databases.

3.7.1 The Base Database

The molecular weight is the key value for compound identification.

The 8,021 metabolite substrates we extracted from HMDB are stored in two tables, Table 3.5 (*myid_mw*) and Table 3.6 (*myid_details*). As seen, *myid_mw* contains only two fields, *hmdb_id* and *mw*. For each substrate, more detailed information such as chemical formula and common names are stored in *myid_details*. The two tables are linked through *hmdb_id*.

Table 3.5: The *myid_mw* table for 8,021 metabolite substrates in MyCompoundID.

Field	Type
hmdb_id	varchar(25)
mw	double(12,6)

Table 3.6: The *myid_details* table for 8,021 metabolite substrates in MyCompoundID.

Field	Type
hmdb_id	varchar(25)
formula	varchar(255)
common_name	varchar(255)

3.7.2 The Expanded Database

We create two tables to store the metabolic products resulted from one reaction and two reactions, respectively. They are named *one_reaction* and *two_reaction*, respectively. The form of the tables is the same, as shown in Table 3.7, where each table entry is for one product. In *one_reaction* table, a product has name *hmdb_id–react_id*, which essentially tells that the product is the result of Reaction #*id* on substrate *hmdb_id*. In *two_reaction* table, a product has name *hmdb_id–react_id1–react_id2*, which essentially tells that the product is the result of Reaction #*id1* followed by Reaction #*id2* on substrate *hmdb_id*. For example, "HMDB00001–8" is the one-reaction product of HMDB00001 getting oxidation (Reaction #8 in

Table 3.1); “HMDB00002–8–4” is the result of HMDB00002 getting oxidation first and then followed by methylation (Reaction #4 in Table 3.1). Each table entry has a field of *mw*, recording the molecular weight of the product.

Table 3.7: The MySQL tables for one- and two-reaction products in MyCompoundID, called *one_reaction* and *two_reaction*, respectively.

Field	Type	Null	Key	Default
hmdb_id–react_id [*–react_id*]	varchar(25)	NO	PRI	
mw	double(12,6)	YES	MUL	
verified_T	tinyint(1)	YES		0
verified_E	tinyint(1)	YES		0
show	tinyint(1)	YES		1

Recall that three restrictions have been used to generate and filter the one- and two-reaction metabolic products to expand our database, as presented in Sections 3.4 and 3.5. Generally speaking, the pseudo metabolic products generated by applying Restriction 1 on Atoms and/or Restriction 2 on Bonds form a superset of those can be generated in reality. However, the pseudo metabolic products generated by the Metabolic Reaction Simulation System, *i.e.* by applying Restriction 3 on Structures, could leave out some real metabolic products. The reason is that the metabolic reaction rules used in the Metabolic Reaction Simulation System are not complete and could be too restrict, and thus some metabolic products can exist in reality but will not be generated by our rules. We therefore have decided to keep the superset generated by Restrictions 1 and 2; nevertheless we create a field *verified_T* to record whether the product passes Restriction 3. The product table evolves to have in total five fields, among which *verified_E* denotes whether the product has been experimentally validated by users and *show* indicates whether the entry is going to be searched against (Table 3.7).

For each pseudo metabolic product generated by applying Algorithms 1 and 2 (the superset), the default value for (*verified_T*, *verified_E*, *show*) is $(0, 0, 1)$. If it passes either of Algorithms 4, 5 and 6, *verified_T* changes to 1; if fails, *i.e.* Restriction 3 on Structures says it is impossible, *show* changes to 0. Recall that for one-reaction products, Restriction 3 on Structures is applied, but not for two-reaction products. Two-reaction products are “verified” by Algorithms 4, 5 and 6 on the fly, meaning that those returned as hits for a user query are verified in the Metabolic Reaction Simulation System. This way, all two-reaction products in the superset can be potentially theoretically verified, and false positives can be marked using the *show* attribute.

3.7.3 Database Indexing

We have three tables storing the metabolite substrates (*myid_mw*), one-reaction pseudo products (*one_reaction*), and two-reaction pseudo products (*two_reaction*). For each mass query, compounds that have mass within the tolerance range must all be returned as output. Since the entries in our tables are not ordered in any specific way, all of them have to be checked by comparing their masses with the query mass, which is time consuming as the database grows large (millions of entries). Using 1,000 randomly picked masses, the search CPU time (in seconds) is shown in the second column of Table 3.8, for 0, 1, or 2 reactions respectively.

Table 3.8: Summary of search time using indexing in MyCompoundID using 1,000 masses. The search was repeated on the set of metabolic products for 0, 1, and 2 reactions respectively. The collected time in the second and last columns are without indexing and with indexing, respectively. All time are in CPU seconds.

	w/o Indexing	w/ Indexing
0 reaction	8.7s	3.4s
1 reaction	240.7s	6.8s
2 reactions	8,400s	79.5s

Search engine indexing collects, parses, and stores data to facilitate fast and accurate information retrieval. Indexes are costly in writing new entries into database and updating indexes. In MyCompoundID, the writing operation is not very common, which avoids this indexing cost. Since in MyCompoundID, the numerical values are needed to be indexed, two candidate index types supported by MySQL are under consideration: BTree and Hash indexes.

B+ tree, which is a variation of BTree [4], is selected by MySQL to index the data. A B+ tree is an n-ary tree with a variable but often large number of children per node. A B+ tree consists of a root, internal nodes and leaves [9]. A B+ tree can be viewed as a B-tree in which each node contains only keys (not pairs), and to which an additional level is added at the bottom with linked leaves. In Hash indexing, a hash table [4] is used to implement an associative array, a structure that can map keys to values. A hash table uses a hash function to compute an index into an array of buckets or slots, from which the correct value can be found. The structural differences between these two indexing data structures make them

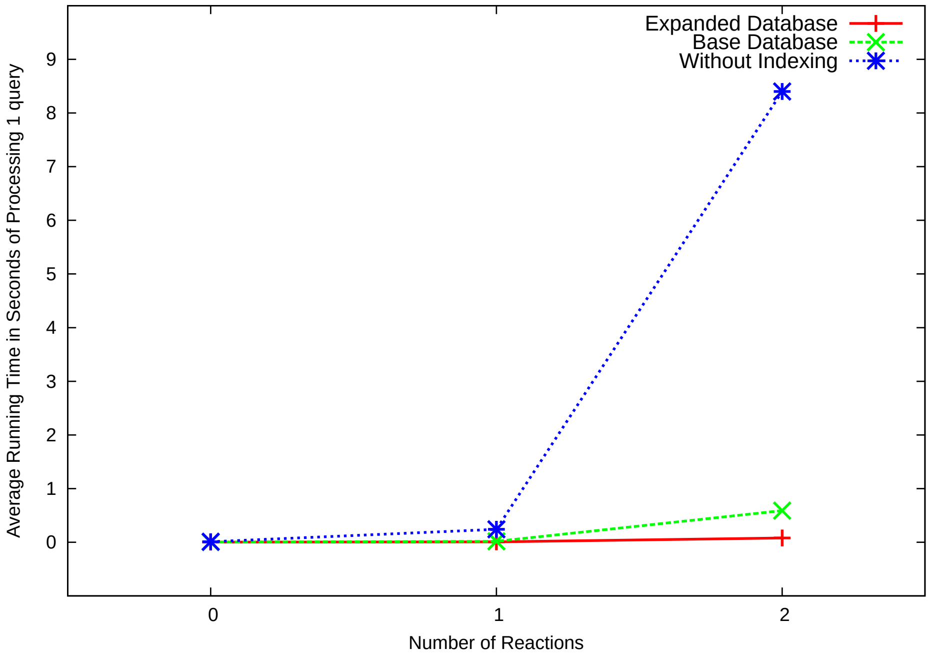


Figure 3.4: The average running time for processing 1 mass query and the comparison between without and with indexing the expanded database. The average is taken over 1,000 randomly selected masses.

be efficient in two different types of searching queries. Hash indexing is better in looking for exactly equal values while the BTree index is better in retrieving a list of entries in a range defined by a lower bound and upper bound. Thus, BTree is more suitable for MyCompoundID searching queries.

We later use BTree to index our tables. Using the same 1,000 masses for query, the search time decreases dramatically, shown in the last column of Table 3.8. The average search time per mass query is also plotted in Figure 3.4, where one sees an exponential growth in search time without indexing. Indeed, the average search time per mass query for 2 reactions on the indexed expanded database (*two_reaction*) is even less than the average search time per mass query for 2 reactions on the base database (*myid_mw*). That is, the linearly scanning the expanded database for mass matches takes a longer time than predicting pseudo metabolic products on the fly.

Chapter 4

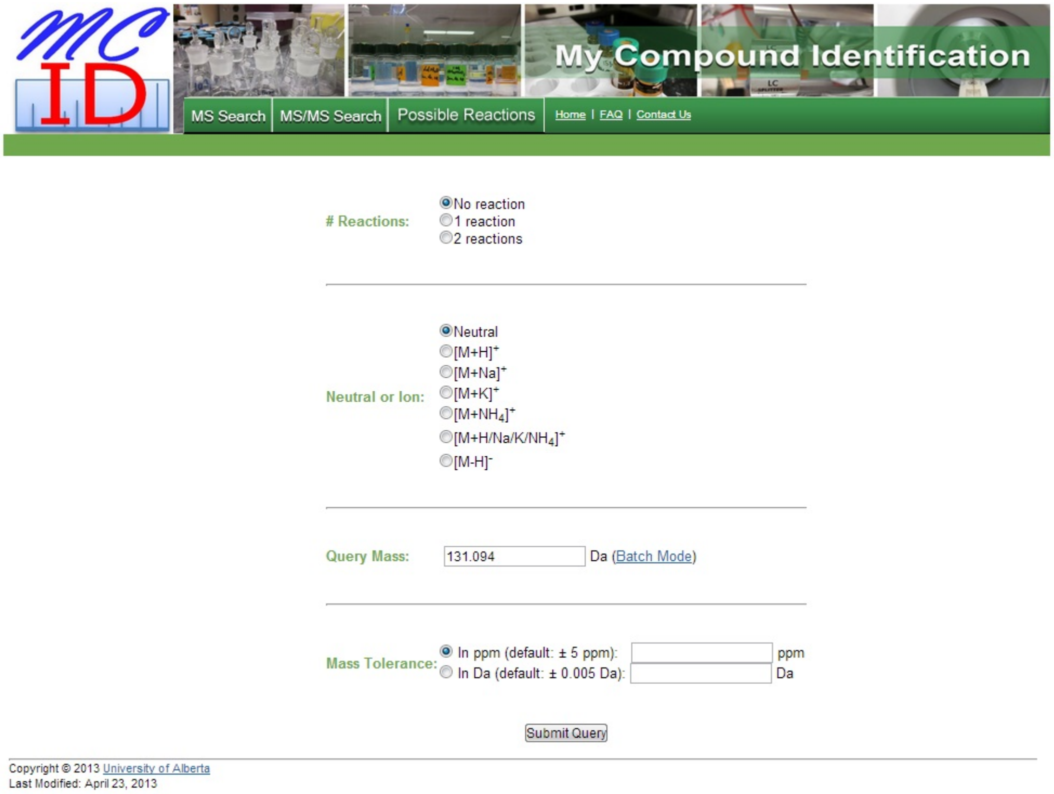
Web Services

Our work on compound identification, *My Compound Identification* or MyCompoundID, is implemented as web services through (`http://www.mycompoundid.org/`). MyCompoundID Version 1.0 was set up for services on February 7, 2013, as a supplement to "MyCompoundID: Using an evidence-based metabolome library for metabolite identification" [18] published in *Analytical Chemistry*, Volume 85, Pages 3401–3408, 2013. Version beta is under development and includes more functionalities such as spectral search. This chapter presents the main features of Version 1.0, and its framework design.

4.1 Search Parameters

In the Search front page, there are four parameters a user has to specify. They are "the (exact) number of reactions", "(search for) neutral or ionized (compounds)", "query mass(es)" and "mass tolerance" in either absolute or relative measure. The default values for three of them are set for the most common scenario, while the user needs to type in one query mass in *single search* mode or multiple masses in *batch search* mode. We note that in the batch search mode, the search options are the same for all query masses. Figures 4.1 and 4.2 contain the screen-shots of the search interfaces for the two modes respectively.

The following gives some detailed information on search parameters. Firstly, if the user wants to search against only those endogenous metabolites in HMDB (*i.e.* our base database), the "No reaction" should be selected; the other two values enable the search for metabolic products as results of one or two metabolic reactions. Secondly, searching for neutral compounds is set as default, but the user has options to search for one of the most common adducts obtained by various



Reactions: ● No reaction ○ 1 reaction ○ 2 reactions

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

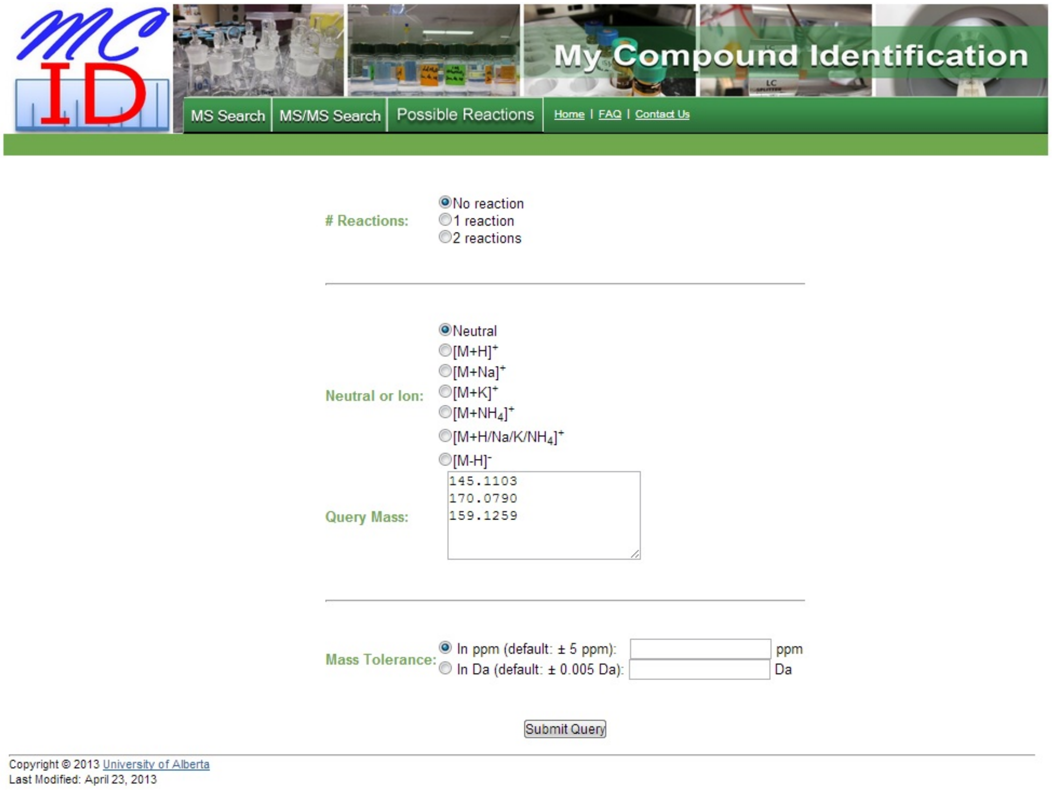
Query Mass: 131.094 Da (Batch Mode)

Mass 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 4.1: Single search interface in MyCompoundID.

ionization processes, such as sodium and ammonium adducts, or search for all ion types. Note that each ion type results in different mass shifting on the compound masses. Lastly, the mass tolerance threshold can be specified in two ways, absolute or relative. In the absolute measure, the unit is Dalton (Da) and 0.005Da is set as default; in the relative measure, the unit is part per million (ppm) and 5ppm is the default. These default values match the typical mass accuracy readily achievable by high resolution instruments such as an FT (Fourier Transform) or a TOF (Time-of-Flight) mass spectrometer. The mass calculation precision (evidenced by the mass distribution over all 8,021 metabolite substrates) suggests that adjusting this mass tolerance within a certain range has no effect on the search results; however, a too large mass tolerance assumes poor data quality on one hand, and results in hits that are difficult to interpret on the other hand.



# Reactions:	● No reaction ○ 1 reaction ○ 2 reactions
Neutral or Ion:	● Neutral ○ $[M+H]^+$ ○ $[M+Na]^+$ ○ $[M+K]^+$ ○ $[M+NH_4]^+$ ○ $[M+H/Na/K/NH_4]^+$ ○ $[M-H]^-$
Query Mass:	145.1103 170.0790 159.1259
Mass Tolerance:	● In ppm (default: ± 5 ppm): ______ ppm ○ In Da (default: ± 0.005 Da): ______ Da

Submit Query

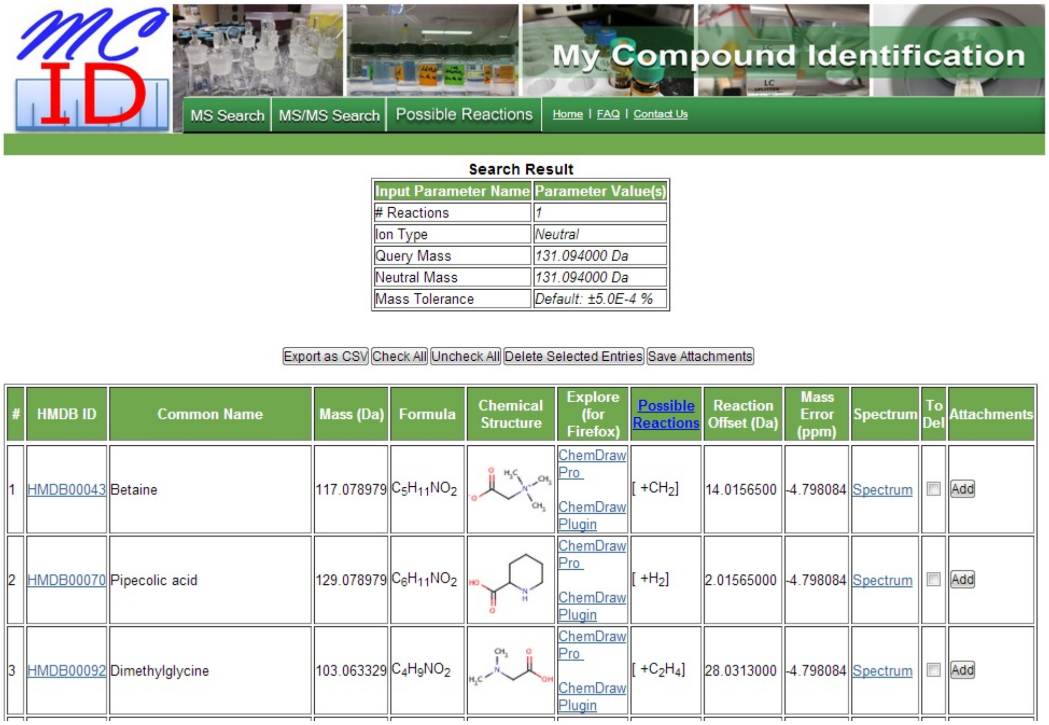
Copyright © 2013 University of Alberta
Last Modified: April 23, 2013

Figure 4.2: Batch search interface in MyCompoundID.

4.2 Searching Result Display

After the user submitted the query, the web page is replaced by the result page. The result page for a single query mass contains two tables (Figure 4.3), the first of which summarizes the query details and the second lists all the hits.

There are 12 columns in the result table (13 columns in beta version where a column displaying theoretical spectrum is added). The rows of the table can be ordered using every column into either ascending or descending order. The second column "HMDB ID" provides the link for the metabolite substrate to HMDB, which is the *MetaboCard* in HMDB containing the detailed information of the metabolite substrate, including biological relevance, synonyms, chemical structure, physical properties and, in some cases, experimental NMR spectra and/or MS/MS spectra. Columns 4–6 list the mass of the hit, the chemical formula, and the two-dimensional structure. The user can then use the next column "Explore" to view the structure in ChemDraw plug-in, or download the structure file in MDL format to view it in any other chemistry molecule structure software.



Search Result

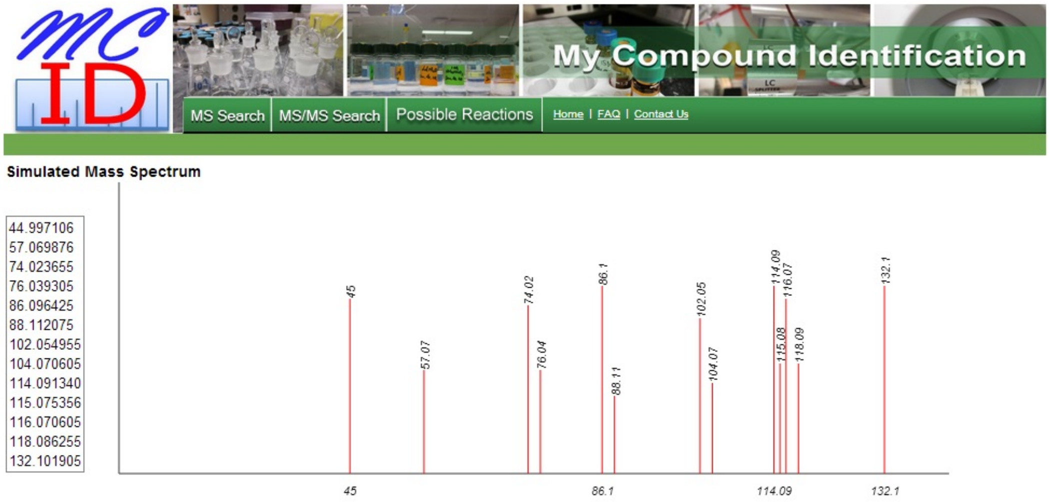
Input Parameter Name	Parameter Value(s)
# Reactions	*1*
Ion Type	*Neutral*
Query Mass	*131.094000 Da*
Neutral Mass	*131.094000 Da*
Mass Tolerance	*Default: ±5.0E-4 %*

#	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	HMDB00043	Betaine	117.078979	$C_5H_{11}NO_2$		ChemDraw Pro ChemDraw Plugin	[$+CH_2$]	14.0156500	-4.798084	Spectrum	□	Add
2	HMDB00070	Pipecolic acid	129.078979	$C_6H_{11}NO_2$		ChemDraw Pro ChemDraw Plugin	[$+H_2$]	2.01565000	-4.798084	Spectrum	□	Add
3	HMDB00092	Dimethylglycine	103.063329	$C_4H_9NO_2$		ChemDraw Pro ChemDraw Plugin	[$+C_2H_4$]	28.0313000	-4.798084	Spectrum	□	Add

Figure 4.3: The screen-shot of the search result page for query mass 131.094, with options for 1 reaction, neutral compounds, and mass tolerance at 5ppm. The webpage contains two tables, one summarizes the query and the other lists all the hits.

Column 8 is the reaction applied to the metabolite substrate to generate the hit, if applicable, and the next column shows the mass offset by this reaction. The "mass error" column indicates the mass difference between the query and the mass of the hit. The rows of the table is initially sorted in non-decreasing mass errors, and thus the top matches are displayed at the front. The next "to del" column provides a checkbox for the user to remove the particular hit/row from the table if the hit is believed *impossible*. In the last column, the user could add the local files which are related to the hit. When the user clicks the "Save Attachments" button, these files will be packed into a zipped archive and saved to a local folder specified by the user.

In the future when search by spectrum is enabled, a few more columns will be added to show the theoretical spectrum, the interpreted spectra, and the experimental (the query) spectrum. A sample webpage to display a spectrum is illustrated in Figure 4.4, where the mass list is in the left hand side and a figure shows the spectrum.

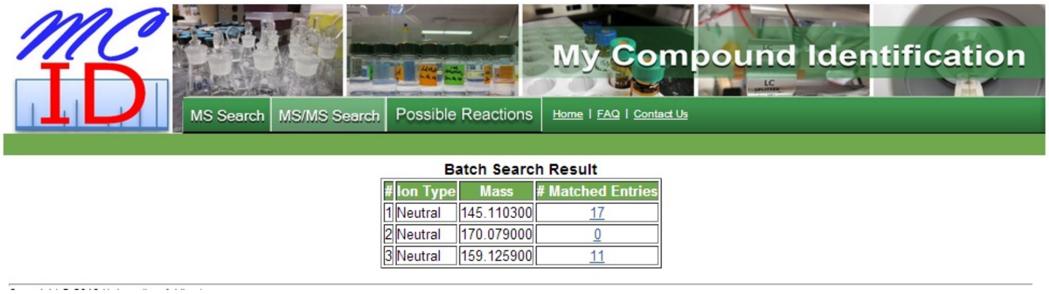


Simulated Mass Spectrum

44.997106
57.069876
74.023655
76.039305
86.096425
88.112075
102.054955
104.070605
114.091340
115.075356
116.070605
118.086255
132.101905

Figure 4.4: A sample webpage for display a compound spectrum in MyCompoundID.

When "all ions" option is selected in the single search mode and for the batch search mode, multiple result pages will be displayed, one for each mass. In this case, a summary webpage with links to all result pages is returned, see Figure 4.5, where the numbers of hits are reported respectively.



Batch Search Result

#	Ion Type	Mass	# Matched Entries
1	Neutral	145.110300	17
2	Neutral	170.079000	0
3	Neutral	159.125900	11

Figure 4.5: The summary webpage for a query with multiple masses. For each mass, the number of hits is indicated and the link to detailed result webpage is provided. This page is for batch search for neutral compounds with masses 145.1103, 170.0790, and 159.1259.

4.3 Web Server Framework Design

The logical flow of our MyCompoundID web server is relatively simple. Basically, the front page navigates different pages of search interfaces and documentation pages. Each search interface takes in search parameters and passes the query to search modules for data retrieve; then the search results are passed to a page to display to the user. That is, this follows the simple Model-View-Controller framework.

In the back-end, Tomcat is used as the container for the web server, JSP is used to build the View, Java Servlet is used to build the Controller, and Java is used to build the Model. We use MySQL for the database system.

The main pages in our web server are described in the following. The top page of the site map is "index.jsp", which navigates between Home, FAQ, Contact Us, Single Search Mode, Batch Search Mode, and Possible Reactions. Under the Home tab, pages Introduction, Workflow, Tutorial, Example, How to Cite, News and Updates are defined (see Figure 4.6 for the detailed site map).

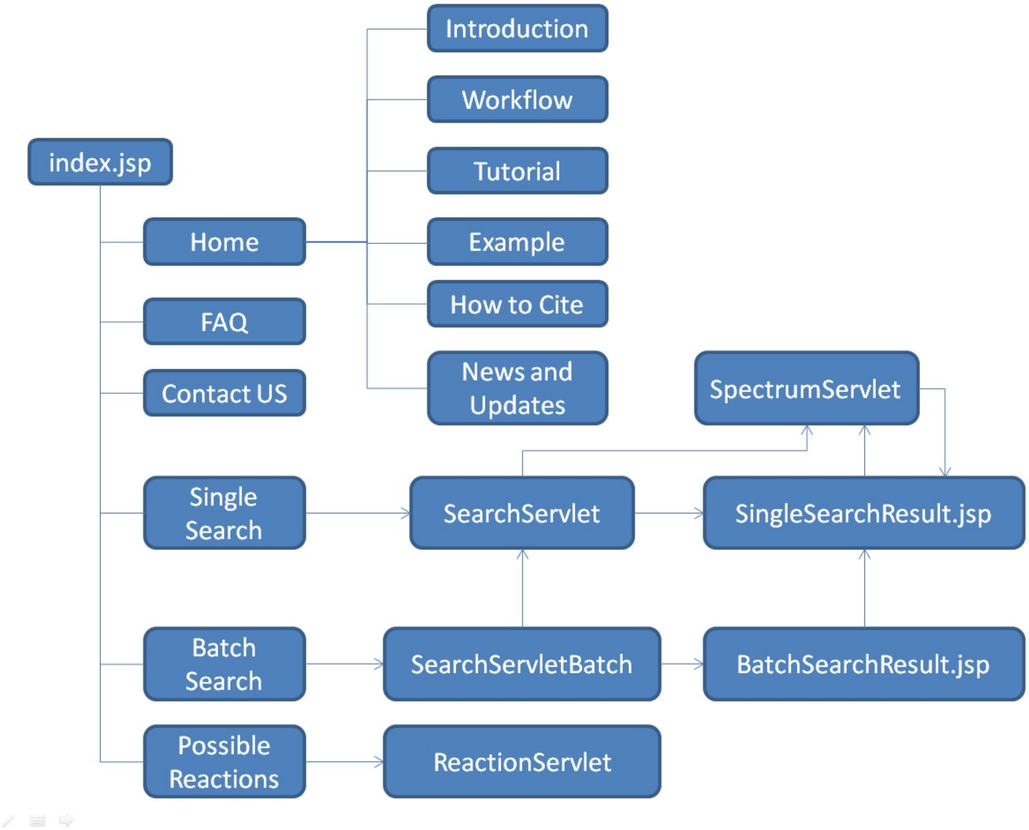


Figure 4.6: MyCompoundID site map.

For compound identification, the single search interface passes the query to "SearchServlet", and "SearchServlet" uses these parameters to call for "Search.java". "Search.java" constructs the MySQL query to search the database(s), retrieves the data, then returns the search result to "SearchServlet". "SearchServlet" passes the result to "SingleSearchResult.jsp" for wrapping into a table to display to the user.

When “all ions” option is selected in the single search mode and for the batch search mode, the query is passed to “SearchServletBatch” for parsing into individual mass queries, following by calling “SearchServlet” once for each individual mass query. All search results are saved in a temporary folder, and “BatchSearchResult.jsp” is invoked to generate the summary page and all individual search result pages.

“SpectrumServlet” can generate the theoretical MS/MS spectrum for a compound. This servlet can be invoked by the user inside “SingleSearchResult.jsp” when the user inquires such a spectrum for each hit compound. The detailed algorithms for generating an MS/MS spectrum are described in Section 3.6.

Chapter 5

Results and Discussion

For meaningful compound identification, an accurate mass of the compound must be obtained. This can be typically achieved by mass spectrometry. Clearly, a mass query returns all compounds that have masses within the tolerance range, which unavoidably include false positive hits. A more accurate way is to obtain the fragmentation, or the MS/MS, spectrum for the compound of interest, then this experimental spectrum can be used in spectrum search that not only satisfies the mass constraint, but also structure constraints.

In the next section, we present our results on human metabolite identification using mass queries, followed by spectral interpretation, using our web server MyCompoundID. Some advantages of MyCompoundID over existing chemical compound databases, PubChem and Kegg, are shown in Section 5.2. Section 5.3 presents results on compound and small peptide identification through spectrum match.

5.1 Identification through Mass Queries

Human urine and plasma were used as samples in this experiment. The sample was performed by a simple extraction to capture a small fraction of the metabolome. After a series of chemistry experiments, 347 compounds in urine and 116 compounds in plasma were extracted, each of which has both the accurate mass measured by TOF-MS and the MS/MS spectrum collected by QTrap-MS. Using the tolerance threshold 5ppm, all these masses are searched against HMDB and MyCompoundID. There are 900 metabolites in HMDB having MS/MS spectra information. To do the experiment on HMDB, all MS/MS spectra are searched against a library of 900 metabolite standards, and the number of metabolites in samples are recorded. To do the experiment on MyCompoundID, all the mass peaks in the MS spectra are searched against the MyCompoundID, and all the MS/MS spectra are manually interpreted and compared with the theoretical MS/MS spectra generated by MyCompoundID.

Table 5.1 summaries the search results. Against HMDB, only 8 metabolites were found in urine and 7 in plasma. Against MyCompoundID, followed by MS/MS spectral interpretation of individual matches, 14 metabolites in urine and 34 in plasma were identified when using option of 0 reaction, 41 metabolites in urine and 14 in plasma were identified when using option of exactly 1 reaction, and additionally 3 more metabolites in urine were identified when using option of exactly 2 reactions. That is, up to 2 reactions, we identified in total 58 metabolites in urine and 48 in plasma in MyCompoundID. These numbers are also plotted in Figure 5.1 for an easier view on the performance difference. Clearly, MyCompoundID significantly increases, over the standard library HMDB, the number of metabolites identifiable from bio-fluids [5].

Table 5.1: The experimental results on human metabolite identification using 347 compounds in urine and 116 compounds in plasma, searched against HMDB and MyCompoundID. The compound masses are used as queries and the search results are interpreted using their MS/MS spectra, respectively.

	Number of Compounds Identified	
	Urine (Total 347)	Plasma (Total 116)
MyCompoundID 0 reaction	14	34
MyCompoundID 1 reaction	41	14
MyCompoundID 2 reactions	3	0
MyCompoundID total	58	48
HMDB	8	7

In another separate experiment we want to identify metabolites in a human urine sample without solvent extraction. Using HMDB, only 23 metabolites were putatively identified by mass matching followed by MS/MS spectrum matching. The other accurate masses that do not have MS/MS spectrum matches were searched in MyCompoundID, and 63 more metabolites were identified with 0 reaction and 87 more metabolites with exactly 1 reaction. These extra metabolites identified in MyCompoundID were interpreted by their MS/MS spectra, respectively. We conclude that this second experiment once again demonstrates that MyCompoundID with the expanded database can be used to identify more putative metabolites from a bio-fluid.

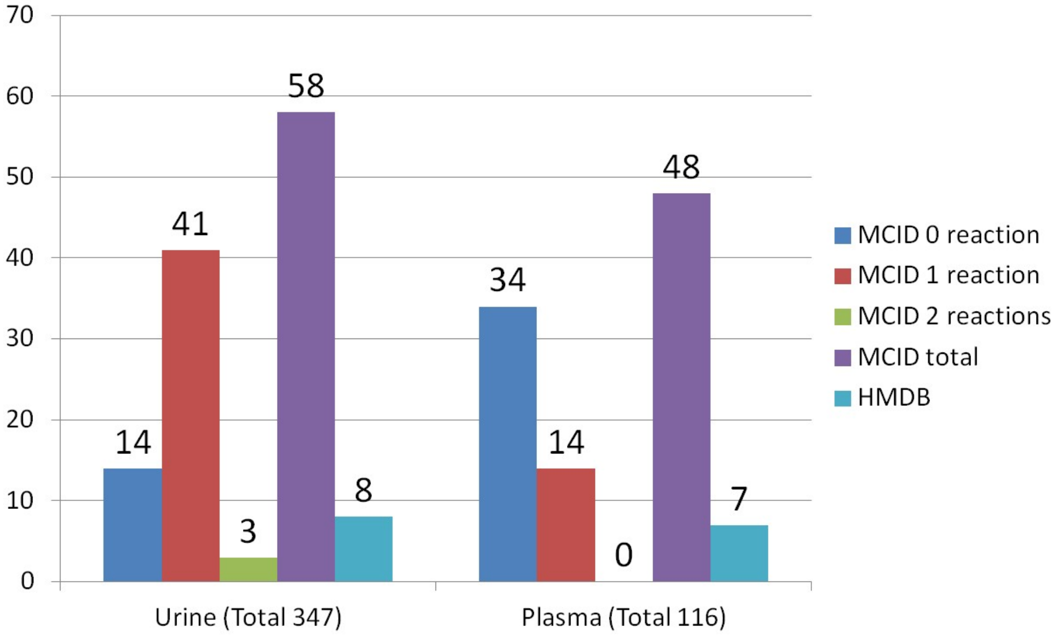


Figure 5.1: The comparison on the number of compounds identified through HMDB and MyCompoundID, using the masses 347 compounds in urine and 116 compounds in plasma for queries and search results interpreted by the corresponding MS/MS spectrum.

5.2 MyCompoundID versus PubChem and Kegg

PubChem is the largest compound library which has over 100 million entries. The Kegg has 16,907 low molecular mass compounds. These two libraries (or databases) contain all sorts of chemicals including synthetic compounds; while MyCompoundID databases are composed of human endogenous metabolites and their predicted metabolic products.

In the following comparison experiment, accurate masses of 83 putative one-reaction metabolic products in MyCompoundID were searched against both PubChem and Kegg using the same mass tolerance of 5ppm. For each mass, hundreds to thousands of hits were returned by PubChem, but only 29 out of the 83 masses have one or more structure or structural isomer matched with the proposed structure; only 23 out of the 83 masses have *correct* matches with the structures proposed by MyCompoundID. This experiment shows the potential advantages of MyCompoundID over ChemPub and Kegg for putative human metabolite identification.

5.3 Identification through Spectrum Queries

We have created a theoretical MS/MS spectrum for each compound using its two-dimensional structure to break every breakable bond. The rules defining breakable bonds are collected from literature. Nevertheless, we realize that at least for some compounds, a portion of their breakable bonds might not actually break in reality. This essentially implies that there could be too many false positive fragments in the theoretical MS/MS spectra we generated. These false positives will reduce the peak uniqueness and eventually spectrum matching confidence. On the other hand, one does not want to miss two many true peaks, since otherwise they will lower the matching scores against true target compounds resulting in identification errors.

In the following experiment, we have compared the generated theoretical M-S/MS spectra and the experimental MS/MS spectra for a set of eight compounds. Figure 5.2 shows the experimental MS/MS spectrum of Leucine and the theoretical MS/MS spectrum generated be MyCompoundID. There are 11 peaks in the experimental spectrum and 13 peaks in the theoretical spectrum, among which, the experimental spectrum and theoretical spectrum overlap with 3 peaks. These overlapped peaks are circled in the figure of each spectrum. As shown in Table 5.2, the number of peaks in each spectrum is counted, and the number of common peaks shared by the two spectra for one compound is obtained. The ratio between the number of common peaks and the number of peaks in the experimental spectrum is defined as the *coverage* of the theoretical spectrum, shown in the last column. Unfortunately, coverages for all eight theoretical MS/MS spectra are low.

Table 5.2: The comparison between the theoretical MS/MS spectra generated by our fragmentation algorithm and the experimental MS/MS spectra for eight compounds. The coverages of these theoretical MS/MS spectra are low as shown in the last column.

Compound	Experimental	Theoretical	Coverage(%)
Leucine	11	13	3 (27%)
Indoleacetic acid	7	9	2 (29%)
3-Methoxybenzenepropanoic acid	9	17	4 (44%)
L-Acetylcarnitine	6	27	4 (67%)
L-Tryptophan	10	14	6 (60%)
Adenosine	20	6	2 (10%)
L-Aspartyl-L-phenylalanine	11	25	6 (55%)
L-Octanoylcarnitine	8	41	7 (88%)

Table 5.3 shows the detailed peak matching or peak hit result of our fragmenta-

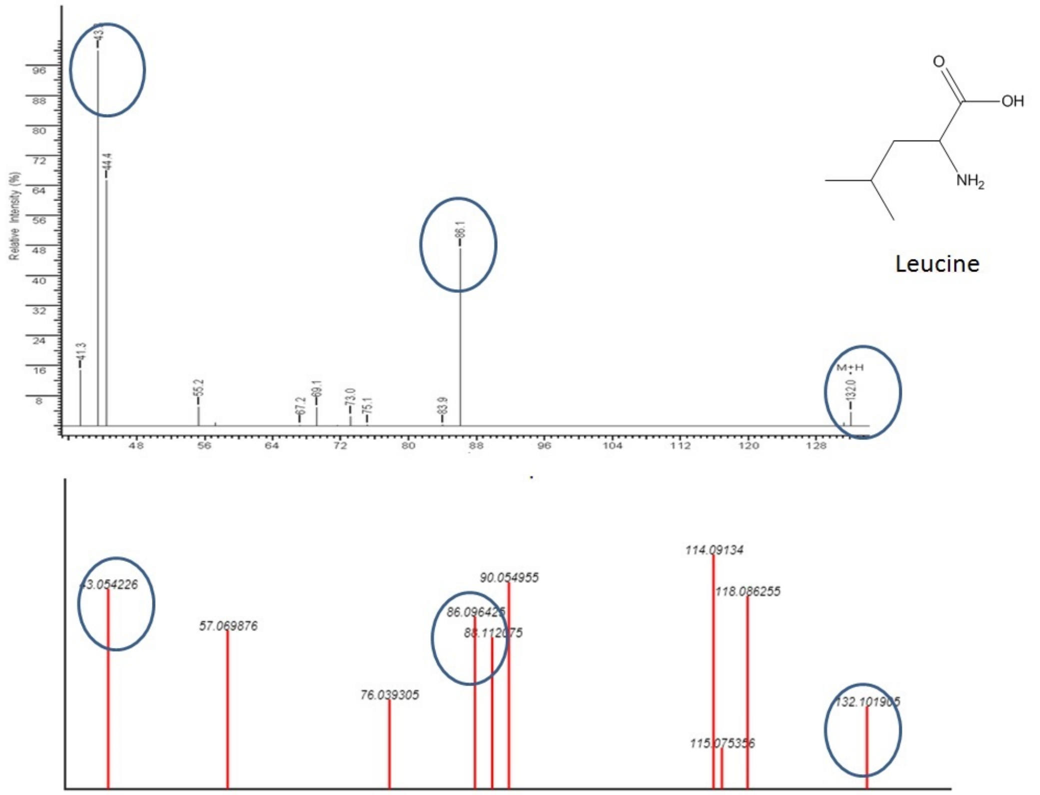


Figure 5.2: The comparison on the peaks in the experimental MS/MS spectrum (above) and the peaks in the theoretical MS/MS spectrum generated be MyCompoundID (below) of Leucine.

tion algorithm for those eight given experimental MS/MS spectra. From the table, we could see that the numbers of matched peaks between the experimental MS/MS spectrum and the one of each candidate metabolite are almost the same. This is very difficult for the simply matching peak counting based algorithms to return the desired metabolites based on the experimental MS/MS spectrum and the theoretical MS/MS spectrum MyCompoundID generated.

Interestingly, despite low coverage, these theoretical MS/MS spectra are very helpful for chemists to verify the identified compounds, or to assist chemists for semi-automated compound identification. Indeed, all eight compounds were successfully identified, and the identification process has been speeded up using the theoretical MS/MS spectra.

In the second experiment, experimental MS/MS spectra for four di-peptides (LW, WL, YG, GY) and four tri-peptides (GYA, FFF, WGG, YGG) are obtained.

Table 5.3: The number of peak hits in the theoretical MS/MS spectra generated by our fragmentation algorithm for the experimental MS/MS spectra for eight compounds.

Compound	Candidate No.	Avg. Hit	Target Hit
Leucine	6	3	3
Indoleacetic acid	2	1.5	2
3-Methoxybenzenepropanoic acid	1	4	4
L-Acetylcarnitine	1	4	4
L-Tryptophan	1	6	6
Adenosine	3	1.7	2
L-Aspartyl-L-phenylalanine	2	6.5	6
L-Octanoylcarnitine	1	7	7

Each of them is matched against every theoretical spectrum in our database of 8,400 di-/tri-peptides, and Equations (3.3) and (3.4) are used as the scoring function for matching against di-peptides and tri-peptides, respectively. We collected the ranks for the true target peptide and its reverse.

In the first setting, all $w(p, s)$ — which is the weight of peak p in spectrum s depending on the ion type for p — are set to 1. The identification results are listed in Table 5.4, where ranks in the second column are obtained by using Equation (3.3) to score all 8,400 peptides, and ranks in the last column are obtained by using Equation (3.3) to score the 400 di-peptides but using Equation (3.4) to score the 8,000 tri-peptides.

Table 5.4: The small peptide identification results through MS/MS spectrum match, in which the ion weights are uniformly at 1. A total of eight experimental MS/MS spectra are tested, using two scoring schemes. Each column shows the rank of the true target peptide, the rank of the reverse true target peptide, and the percentage of permuted peptides ranked higher than any other non-permuted peptides.

	Ranks and Coverage (%)	
	Equation (3.3)	Equations (3.3)+(3.4)
LW	1/2/100%	1/2/100%
WL	2/1/100%	2/1/100%
YG	1/2/100%	1/2/100%
GY	1/2/100%	1/2/100%
GYA	2/3/100%	1/2/100%
FFF	1/–/100%	1/–/100%
WGG	1/5/67%	1/3/100%
YGG	1/3/100%	1/3/100%

It happens that di-peptide WL could not identified (*i.e.* ranked to the top). A closer look into the experimental spectrum reveals that the experimental spectrum for WL does not seem sufficient to distinguish WL and LW. For example, the experimental spectrum for WL does not contain "critical" peaks that is in the theoretical spectrum for WL but not in the theoretical spectrum for LW. In other words, all the peaks in this experimental spectrum are common to WL and LW. Under the uniform ion weight factors, the score scheme Equation (3.3) is in favor of di-peptide LW (over WL).

When using Equation (3.3) to score tri-peptides, six out of eight peptides are identified (ranked the top). Using Equation (3.4) to score tri-peptides further identifies tri-peptide GYA; it also lifts the rank of tri-peptide GGW when identifying WGG.

In the second setting, we want to set the weights of the ion types differently according to how easy they can be formed. Since almost all y-ions present in our QToF MS/MS spectra, we end up with lifting only the weight of y-ion to 2 (*i.e.* twice the weights of the other types of ions). The experiment is repeated just the same as in the first setting. The identification results are listed in Table 5.5, where ranks in the second column are obtained by using Equation (3.3) to score all 8,400 peptides, and ranks in the last column are obtained by using Equation (3.3) to score the 400 di-peptides but using Equation (3.4) to score the 8,000 tri-peptides. Under this setting, di-peptide WL can be easily distinguished from LW using both scoring schemes.

Table 5.5: The small peptide identification results through MS/MS spectrum match, in which the ion weights are set according to how easy the ion can be formed. A total of eight experimental MS/MS spectra are tested, using two scoring schemes. Each column shows the rank of the true target peptide, the rank of the reverse true target peptide, and the percentage of permuted peptides ranked higher than any other non-permuted peptides.

	Ranks and Coverage (%)	
	Equation (3.3)	Equations (3.3)+(3.4)
LW	1/2/100%	1/2/100%
WL	1/2/100%	1/2/100%
YG	1/2/100%	1/2/100%
GY	1/2/100%	1/2/100%
GYA	2/3/100%	1/2/100%
FFF	1/1/100%	1/1/100%
WGG	1/5/67%	1/3/100%
YGG	1/3/100%	1/3/100%

5.3.1 Discussion

It appears that identifying di-peptides is much easier than identifying tri-peptides. One reason is that the numbers of peaks in the experimental MS/MS spectra for di-/tri-peptides are very close to each other, but the theoretical MS/MS spectra for tri-peptides contains around 50% more peaks than the counterparts.

Clearly every peak contributes differently to the match scores. We have also observed larger variations across tri-peptides than across di-peptides. It would be interesting to learn how to weight the importance of each individual peak, after a sufficient number of experimental MS/MS spectra have been collected.

The weights for different ion types seem to improve the performance of the match scoring schemes. But they are tested for only eight instances. Also, these weights are set roughly to match their probabilities of occurrence. Again it would be interesting to learn how to weight the ion types, after a sufficient number of experimental MS/MS spectra have been collected.

Chapter 6

Conclusions and Future Work

In this work, we present the MyCompoundID as a publicly accessible web server for metabolite identification. Identifying unknown metabolites in bio-fluids is challenging but the critical step towards metabolome profiling, which leads to biomarker discovery for various biological and medical applications.

MyCompoundID is able to use experimental data effectively to identify more metabolites than existing similar web applications such as HMDB, PubChem, and KEGG. The main support to achieving this success is the expanded database of pseudo one- and two-reaction metabolic products, generated from the 8,021 metabolite substrates using the 76 commonly encountered metabolic reactions in human. While the expanded database contains tens of millions of compounds, our implementation makes the best efforts on balancing CPU time consumption and the false positive rate. All two-reaction metabolic products are or will be validated structurally through the user query processes. This way, along the time, our database is left with more and more reliable pseudo metabolic products, and subsequently query processes become faster and faster.

The MyCompoundID web server is implemented in the well known MVC model, which increases the code re-usability and separation of concerns. Such a model also enables quick updating and modifications.

For compound identification through MS/MS spectrum search, so far our database contains only the theoretical spectrum for each compound (both substrates and metabolic products). The experimental spectra collected in various labs are empirical evidences for the compounds identified, and they are very important in understanding the fragmentation patterns for every specific compound. We will be implementing a functionality to accept the interpreted MS/MS spectra contributed by the users, and to construct a database of interpreted MS/MS spectra, besides the database of theoretical MS/MS spectra. We anticipate that such interpreted spectra would not only enrich our knowledge on metabolite chemistry, but also improve compound identification further.

For the di-/tri-peptide identification through MS/MS spectrum match, which is a prototype for the general compound identification through MS/MS spectrum match, our spectrum matching algorithm is based on weighted tf-idf. One reason for the huge success for this simple matching algorithm is that a theoretical MS/MS spectrum for a di-/tri-peptide contains only dozens of peaks. Nevertheless, the theoretical MS/MS spectrum for a metabolite can contains hundreds of peaks. Our preliminary experiment on a few spectra suggests that developing an effective scoring scheme for spectrum match is challenging; yet we believe with more experimental spectra collected we might be able to apply advanced machine learning algorithms in training some scoring schemes of good quality.

Bibliography

[1] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant. Pubchem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry*, 4:217–241, 2008.

[2] M. Brown, D. C. Wedge, R. Goodacre, D. B. Kell, P. N. Baker, L. C. Kenny, M. A. Mamas, L. Neyses, and W. B. Dunn. Automated workflows for accurate mass-based putative metabolite identification in lc/ms-derived metabolomic datasets. *Bioinformatics*, 27(8):1108–1112, 2011.

[3] K. P. Chiang, S. Niessen, A. Saghatelian, and B. F. Cravatt. An enzyme that regulates ether lipid signaling pathways in cancer annotated by multidimensional profiling. *Chemistry & biology*, 13(10):1041–1050, 2006.

[4] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, Cambridge, Massachusetts, 2001. Second Edition.

[5] Q. Cui, I. A. Lewis, A. D. Hegeman, M. E. Anderson, J. Li, C. F. Schulte, W. M. Westler, H. R. Eghbalnia, M. R. Sussman, and J. L. Markley. Metabolite identification via the Madison metabolomics consortium database. *Nature biotechnology*, 26:162–164, 2008.

[6] C. Dass. *Fundamentals of contemporary mass spectrometry*, volume 16. Wiley-Interscience, 2007.

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

[8] K. Dettmer, P. A. Aronov, and B. D. Hammock. Mass spectrometry-based metabolomics. *Mass spectrometry reviews*, 26(1):51–78, 2007.

[9] R. Elmasri. *Fundamentals Of Database Systems, 5/E*. Pearson Education India, 2008.

[10] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. W. H. Freeman and Company, San Francisco, 1979.

[11] K. Guo and L. Li. High-performance isotope labeling for profiling carboxylic acid-containing metabolites in biofluids by mass spectrometry. *Analytical chemistry*, 82:8789–8793, 2010.

[12] X. Han, K. Yang, and R. W. Gross. Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses. *Mass Spectrometry Reviews*, 31:134–178, 2011.

[13] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, et al. Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry*, 45(7):703–714, 2010.

[14] J. Inczédy, T. Lengyel, AM. Ure, and H. Freiser. *Compendium of analytical nomenclature*, volume 18. Blackwell Science Oxford,, UK, 1998.

[15] H. Jenkins, N. 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. Roessner-Tunali, K. Saito, J. Smedsgaard, L. W. Sumner, T. Wang, S. Walsh, E. S. Wurtele, and D. B. Kell. A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology*, 22:1601–1606, 2004.

[16] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[17] Z. Lei, D. V. Huhman, and L. W. Sumner. Mass spectrometry strategies in metabolomics. *Journal of Biological Chemistry*, 286(29):25435–25442, 2011.

[18] L. Li, R. Li, J. Zhou, A. Zuniga, A. E. Lewis-Stanislaus, Y. Wu, T. Huang, J. Zheng, Y. Shi, D. S. Wishart, and G. Lin. MyCompoundID: Using an evidence-based metabolome library for metabolite identification. *Analytical Chemistry*, 85:3401–3408, 2013.

[19] A. J. Liedtke, S. H. Nellis, L. F. Whitesell, and C. Q. Mahar. Metabolic and mechanical effects using L-and D-carnitine in working swine hearts. *American Journal of Physiology – Heart and Circulatory Physiology*, 243:H691–H697, 1982.

[20] C. Ludwig and M. R. Viant. Two-dimensional j-resolved nmr spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis*, 21(1):22–32, 2010.

[21] A. D. McNaught and A. Wilkinson. *Compendium of chemical terminology*, volume 1669. Blackwell Science Oxford, 1997.

[22] R. Powers. Nmr metabolomics and drug discovery. *Magnetic Resonance in Chemistry*, 47(S1):S2–S11, 2009.

[23] D. G. Robertson. Metabonomics in toxicology: a review. *Toxicological Sciences*, 85(2):809–822, 2005.

[24] J. Ryals. Metabolomics – an important emerging science. *Drug Discovery Metabolomics, Business Briefing: Pharmatech*, 2004.

[25] A. Saghatelian, S. A. Trauger, E. J. Want, E. G. Hawkins, G. Siuzdak, and B. F. Cravatt. Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry*, 43(45):14332–14339, 2004.

[26] C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak. Metlin: a metabolite mass spectral database. *Therapeutic drug monitoring*, 27(6):747–751, 2005.

[27] O. D. Sparkman. *Mass spectrometry desk reference.* Global View Pub, 2000.

[28] A. Sreekumar, L. M. Poisson, T. M. Rajendiran, A. P. Khan, Q. Cao, J. Yu, B. Laxman, R. Mehra, R. J. Lonigro, Y. Li, M. K. Nyati, A. Ahsan, S. Kalyana-Sundaram, B. Han, X. Cao, J. Byun, G. S. Omenn, D. Ghosh, S. Pennathur, D. C. Alexander, A. Berger, J. R. Shuster, J. T. Wei, S. Varambally, C. Beecher, and A. M. Chinnaiyan. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, 457:910–914, 2009.

[29] J. FJ. Todd. Recommendations for nomenclature and symbolism for mass spectroscopy. *Pure and applied chemistry*, 63:1541–1566, 1991.

[30] D. Weininger, A. Weininger, and J. L. Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, 1989.

[31] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M. A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. Macinnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, and L. Querengesser. Hmdb: the human metabolome database. *Nucleic Acids Research*, 35:D521–D526, 2007.

[32] Z. Zhu, A. W. Schultz, J. Wang, C. H. Johnson, S. M. Yannone, G. J. Patti, and G. Siuzdak. Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the metlin database. *Nature protocols*, 8(3):451–460, 2013.