Development of computational analysis tools for natural products research and metabolomics

（天然物科学およびメタボロミクスのための計算解析ツールの開発）

２０１５

Ahmed Mohamed Mohamed

Dedication

To my parents, Mohamed and Enas

This thesis is the culmination of their years of hard efforts

Abstract

Metabolic analysis in living organisms is important for understanding biological systems, having wide applications ranging from therapeutics, drug discovery and biotechnology. For example, metabolic profiles can be used to identify biomarkers for early disease prognosis. In natural products research, bioactive secondary metabolites are considered to be new drug leads. In biotechnology, metabolic engineering is routinely used for optimizing metabolite production.

Depending on applications, metabolic analysis can be carried out by one of two paradigms: network analysis and metabolite identification. Firstly, network analysis investigates metabolic networks to systematically identify active metabolic pathways and metabolite production patterns. Hence, metabolic network analysis is used for biomarker discovery and metabolite production optimization. Secondly, in metabolite identification paradigm, the presence and concentration of individual metabolites are investigated. For example, discovery of new drug leads from natural products involves structure determination of individual metabolites with promising bioactivities or novel chemical scaffolds.

Despite the importance of metabolic analysis, necessary computational tools are still lacking. The technological advances increased the amount of experimental data that can be collected, making manual analysis challenging. For example, analysis of genome-scale metabolic networks with thousands of metabolites is manually infeasible, and requires computational tools. In natural products research, integration of computational tools with spectral databases are needed for rapid identification of known compounds. Also, software tools for online processing of NMR measurements, a central technique for metabolite identification, are still lacking. Easy-to-use computational tools enable researchers to quickly analyze and interpret experimental data, reducing cost and effort.

In this thesis, I explore computational methods and tools needed for different paradigms for metabolic analysis, presenting two novel tools, NetPathMiner and NMRPro. First, I present NetPathMiner, a software in R framework, for identification of active metabolic pathways based on gene expression. Second, I review computational resources for rapid identification of natural products identifying the need for software tools for processing nuclear magnetic resonance (NMR) spectra. Finally, I present NMRPro, a web component for online interactive processing of NMR spectra. I discuss each topic briefly below.

NetPathMiner is a general framework for mining, from genome-scale networks, paths that are related to specific experimental conditions. NetPathMiner interfaces with various input formats including KGML, SBML and BioPAX files and allows manipulation of networks in three different forms: metabolic, reaction and gene representations. NetPathMiner ranks active paths and applies clustering and classification to the ranked paths for easy interpretation, providing static and interactive visualizations of networks and paths.

Rapid identification of previously isolated compounds in an automated manner, called dereplication, steers researchers toward novel findings, thereby reducing the time and effort for identifying new drug leads. Dereplication identifies compounds by comparing processed experimental data with those of known compounds, and so, diverse computational resources, such as databases and tools to process and compare compound data, are necessary. Automating the dereplication process through the integration of computational resources has always been an aspired goal for natural product research. To increase the utilization of current computational resources for natural products, I provided an overview of the dereplication process, and then listed useful resources, categorizing them into databases, methods and software tools and further explained them from a dereplication perspective. Finally, I discussed the current challenges to automating dereplication and proposed solutions.

Finally, I present NMRPro, an integrated web component for interactive processing and visualization of NMR spectra. Web applications are well used

recently because they are platform-independent and easy to extend through reusable web components. Although available web applications can analyze NMR spectra, they still lack essential processing and interactive visualization functionalities. Incorporating NMRPro into current web applications enables easy-to-use online interactive processing and visualization.

In conclusion, I surveyed the current status of computational tools for metabolic analysis and presented two novel tools, which can be considered as building blocks for automating research in natural products and metabolomics.

Publication Notes

The content of this thesis is based on three scientific publications, which appeared *Bioinformatics* (2 papers) and *Briefings in Bioinformatics* (1 review paper) journals.

Publication list

A. Mohamed, T. Hancock, C. H. Nguyen, H. Mamitsuka, NetPathMiner: R/Bioconductor package for network path mining through gene expression. *Bioinformatics* **30**, 3139-3141 (2014).

A. Mohamed, C. H. Nguyen, H. Mamitsuka, Current status and prospects of computational resources for natural product dereplication: a review. *Briefings in bioinformatics*, bbv042 (2015).

A. Mohamed, T. Hancock, C. H. Nguyen, H. Mamitsuka, NMRPro: An integrated web component for interactive processing and visualization of NMR spectra. *Bioinformatics* ***(in revision)***.

Contents

Abstract i

Publication Notes iv
Publication list iv

Contents v

List of Figures vii

List of Tables viii

Chapter 1 Introduction 1
1.1. Background 1
1.1.1. Hierarchy of cellular systems 1
1.1.2. Types of metabolites 2
1.2. The need for software tools for metabolic analysis 4
1.2.1. Metabolic path mining from gene expression 5
1.2.2. Processing NMR data for metabolite identification 6
1.3. Thesis organization 7

Chapter 2 NetPathMiner: R/Bioconductor package for network path mining through gene expression 8
Chapter Summary 8
2.1. Introduction 9
2.2. Input to network path mining 12
2.2.1. Network: 12
2.2.2. Gene expression matrix 14
2.3. Workflow of NetPathMiner 15
2.3.1. Pathway File Processing (Step 1 in Figure 2.1) 16
2.3.2. Network Manipulation (Step 2 in Figure 2.1) 17
2.3.3. Weighting the network (Step 3 in Figure 2.1) 20
2.3.4. Path Ranking (Step 4 in Figure 2.1) 20
2.3.5. Paths Clustering and Classification (Step 5 in Figure 2.1) 23
2.3.6. Visualization (Step 6 in Figure 2.1) 23
2.4. Additional functionalities 26
2.4.1. Analysis of Signaling Networks: 26
2.4.2. Integration with other R packages 26
2.5. Conclusion 26

Chapter 3 Current status and prospects of computational resources for natural product dereplication 28
Chapter Summary 28
3.1. Introduction 29
3.2. Overview of natural products compound identification 31
3.3. Databases 33
3.3.1. General databases (Table 3.2): 33
3.3.2. Natural products-specific databases (Table 3.3): 37

3.4. Methods and Software 37
3.4.1. Spectral preprocessing 43
3.4.2. Compound identification 45
3.5. Future Perspectives 50
3.5.1. Enriching databases using automated machine leaning methods: 50
3.5.2. Developing software suite from building blocks: 50
3.5.3. Integrating different spectral types: 51
3.5.4. Sorting databases for efficient search: 51
Chapter 4 NMRPro: An integrated web component for interactive processing and visualization of NMR spectra 52
Chapter Summary 52
4.1. Introduction 53
4.2. Web applications as medium for scientific development 54
4.2.1. Current status of web applications for NMR data 55
4.3. Software architecture of NMRPro 57
4.3.1. Challenges for developing web application for NMR 57
4.3.2. Design considerations for NMRPro 57
4.4. Subcomponents of NMRPro 59
4.4.1. Python Package 59
4.4.2. Django App 60
4.4.3. SpecdrawJS 61
4.5. Availability and Installation 63
4.6. Conclusion 64
Chapter 5 Conclusions 65
Acknowledgements 67
References 68

List of Figures

FIGURE 1.1 OVERVIEW OF CELLULAR SYSTEMS: FROM GENOMES TO METABOLITES 1
FIGURE 1.2 CATEGORIES AND GOALS OF METABOLIC ANALYSIS 2
FIGURE 2.1 GENERAL WORKFLOW AND MODULES OF NETPATHMINER 11
FIGURE 2.2 EXAMPLES OF METABOLIC NETWORKS IN DIFFERENT REPRESENTATIONS. 12
FIGURE 2.3 GENE EXPRESSION MATRIX. ROWS ARE GENES AND COLUMNS ARE SAMPLES. SAMPLES CAN BE DIVIDED INTO GROUPS ACCORDING TO THE EXPERIMENTAL CONDITIONS. 15
FIGURE 2.4 EXAMPLES OF REACTIONS WITH NO ASSOCIATED GENES 19
FIGURE 2.5 NETPATHMINER PATH VISUALIZATION. CARBOHYDRATE METABOLISM NETWORK EXTRACTED FROM REACTOME, AND ANALYZED WITH NETPATHMINER. TOP 100 PATHS WERE EXTRACTED, AND GROUPS INTO 3 CLUSTERS. PATHS PLOTTED ON DIFFERENT NETWORK REPRESENTATIONS AND COLORED BY CLUSTER MEM-BERSHIP. (A) METABOLITE-REACTION BIPARTITE REPRESENTATION. (B) REACTION NETWORK REPRESENTATION PLOTTED USING THE SAME LAYOUT AS A. (C) THE UNDERLYING GENE NETWORK OF CARBOHYDRATE METABOLISM, PLOTTED USING THE SAME LAYOUT. (D) PATHS (ROWS) AND THEIR COMPONENTS (COLUMNS), COLORED BY CLUSTER MEMBERSHIP. (E) PROBABILITIES THAT EACH PATH BELONGS TO ITS ASSIGNED CLUSTER. 24
FIGURE 2.6 CARBOHYDRATE METABOLIC NETWORK IN GENE REPRESENTATION, WITH VERTICES COLORED BY SUBCELLULAR COMPARTMENT (PLOTTED IN R). 25
FIGURE 2.7 CYTOSCAPE PLOTS FOR THE CARBOHYDRATE METABOLIC NETWORK IN GENE REPRESENTATION, WITH VERTICES COLORED BY SUBCELLULAR COMPARTMENT. 25
FIGURE 3.1 COMPOUND IDENTIFICATION IN NATURAL PRODUCTS WITHOUT AND WITH DEREPLICATION. 32
FIGURE 3.2 SPECTRAL PREPROCESSING. ^{1}H NMR SPECTRA OF CHOLESTEROL AND STIGMASTEROL, TWO COMMON AND STRUCTURALLY SIMILAR NATURAL COMPOUNDS, ARE USED FOR DEMONSTRATION. THE RAW NMR FILES WERE DOWNLOADED FROM HMDB (79) AND CONVERTED TO JCAMP-FORMAT DX USING MESTRENOVA. BASELINE ESTIMATION WAS PERFORMED IN R USING 3RD ORDER POLYNOMIAL FITTING. THE BASELINE-CORRECTED SPECTRA WERE STACKED, AND THEN ALIGNED USING MESTRENOVA, SHOWING HIGHER SIMILARITY (PEASRON'S CORRELATION OF 0.423) THAN BEFORE ALIGNMENT (0.288). 41
FIGURE 3.3 DATA REDUCTION OF SPECTRA. ^{1}H AND ^{13}C NMR SPECTRA OF CAMPHOR, A NATURAL COMPOUND, DEMONSTRATE THE EFFECT OF EACH DATA REDUCTION METHOD ON DIFFERENT TYPES OF SPECTRA. PEAK PICKING REDUCES THE ^{13}C SPECTRUM TO A FEW PEAKS (2A), BUT FAILS WITH THE ^{1}H SPECTRUM (1A) AS RESONANCE COUPLING GENERATES NUMEROUS OVERLAPPING MULTIPLET PEAKS. BINNING PRODUCES IN A LARGE VECTOR (1532 BIN) IN THE ^{13}C SPECTRUM AND A SMALL ONE IN THE ^{1}H SPECTRUM (47 BINS). BOTH SPECTRA ARE REDUCED TO RELATIVELY FEW NODES WHEN REPRESENTED AS TREES. 42
FIGURE 4.1 COMPONENT ARCHITECTURE OF NMRPRO. 59
FIGURE 4.2 DATA EXCHANGE PROTOCOL BETWEEN SERVER AND CLIENT-SIDES, AS MANAGED BY DJANGO SUBCOMPONENT. 60
FIGURE 4.3 SPECDRAWJS VISUALIZATION. A) 1D NMR DATASET. B) 2D NMR SPECTRUM 62

List of Tables

TABLE 1.1 CHAPTER CONTENTS 7
TABLE 2.1 FUNCTIONALITIES OF CURRENT NETWORK PATH MINING TOOLS 10
TABLE 2.2 DIFFERENCES BETWEEN MAJOR PATHWAY FILE FORMATS. 16
TABLE 3.1 DIFFERENCES BETWEEN COMPOUND IDENTIFICATION IN NATURAL PRODUCTS RESEARCH AND METABOLOMICS. 31
TABLE 3.2 GENERAL CHEMICAL DATABASES. 35
TABLE 3.3 NATURAL PRODUCTS-SPECIFIC DATABASES. 36
TABLE 3.4 ANALYSIS FLOW OF SPECTRA FROM ACQUISITION TO COMPOUND IDENTIFICATION. 39
TABLE 3.5 SOFTWARE TOOLS WITH A POTENTIAL ROLE IN DEREPLICATION. 40
TABLE 4.1 COMPARISON OF SOFTWARE CAPABILITIES WITH EXISTING WEB-BASED APPLICATIONS. 56
TABLE 4.2 COMPARISON OF NMRPRO WITH EXISTING FRAMEWORKS 57
TABLE 4.3 FUNCTIONALITIES AVAILABLE IN EACH SPECDRAWJS CONFIGURATION. 61

Chapter 1

Introduction

Chapter Contents

1.1. Background 1
1.1.1. Hierarchy of cellular systems 1
1.1.2. Types of metabolites 2
1.1.2.1. Analysis of primary metabolites 3
1.1.2.2. Analysis of secondary metabolites 4
1.1.2.3. Analysis of recombinant metabolites 4
1.2. The need for software tools for metabolic analysis 4
1.2.1. Metabolic path mining from gene expression 5
1.2.2. Processing NMR data for metabolite identification 6
1.3. Thesis organization 7

1.1. Background

1.1.1. Hierarchy of cellular systems

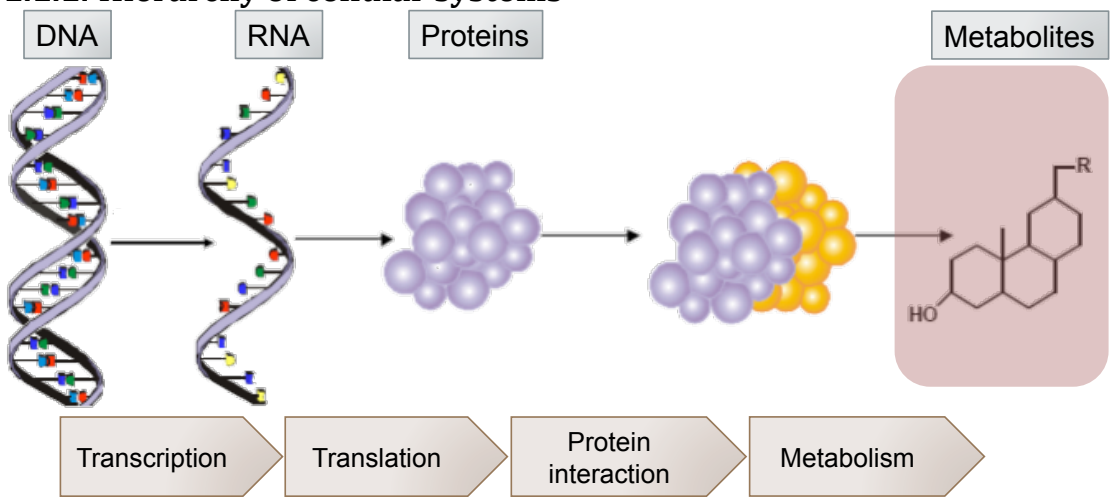

Figure 1.1 Overview of cellular systems: from genomes to metabolites

The genetic code contained in the cells of all living organisms dictate their behavior, from survival to reproduction. This genetic material is stored as a chain of deoxyribonucleic acids (DNA), of which only a small fraction is transcribed as ribonucleic acid (RNA) (*1, 2*). Then, coding RNA is translated into proteins, the functional building blocks of the cell. Proteins activate or inhibit

each other using post-translational modifications through highly regulated and complex interaction network. Activated proteins with biochemical activities, referred to as enzymes, control the production and consumption of metabolites. Therefore, analysis metabolic processes is challenging because of the numerous interactions involved in controlling metabolic activity (Figure 1.1).

1.1.2. Types of metabolites

Naturally, living organisms produce two types of metabolites: 1) Primary metabolite, which are essential for the survival of the organism (*3, 4*). Examples of primary metabolites include energy molecules, such as adenosine triphosphate (ATP), and amino acids that are later used as building blocks for peptides and proteins. 2) Secondary metabolites, which give the organism competitive advantages but are not essential for survival. Secondary metabolites are prevalent in microbial organisms, plants, marine animals, in which they involved in interspecies defense (*5, 6*). In addition to naturally produced metabolites, recombinant DNA technologies allow the production of xenobiotics for biotechnological purposes (*7*). Analysis of each of these types has different goals and requires the use of different analytical methods, as shown in Figure 1.2.

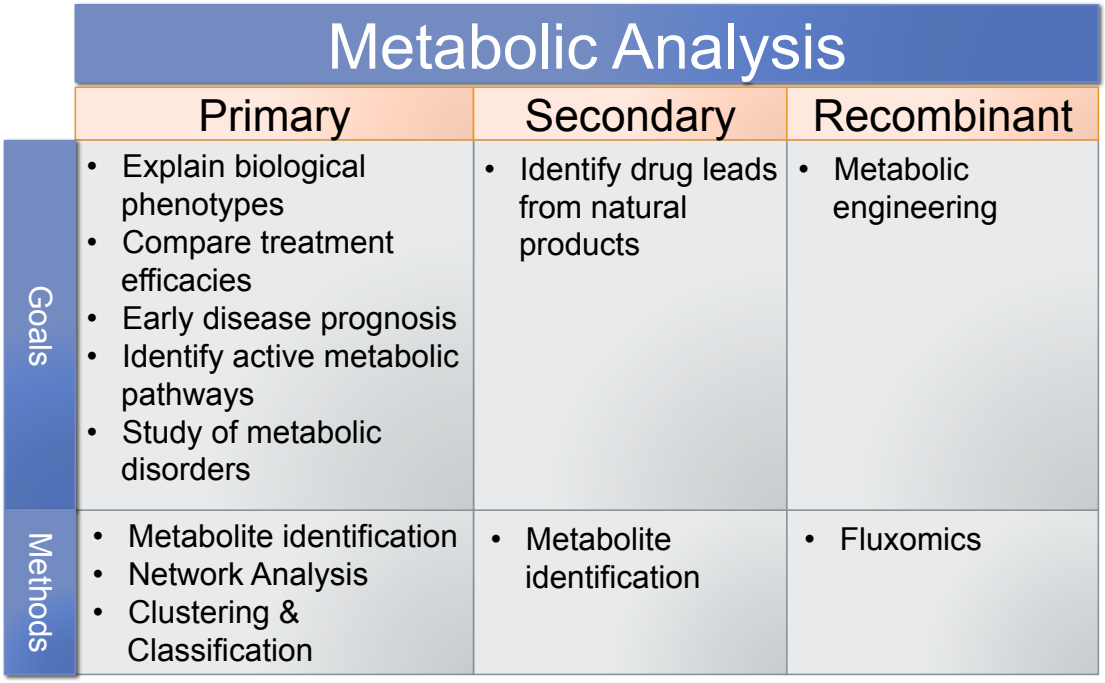

Figure 1.2 Categories and goals of metabolic analysis

1.1.2.1. Analysis of primary metabolites

Primary metabolites are involved in cellular growth, development or reproduction. Therefore, any impairment in the production of primary metabolites directly affects the normal function of the organism. Because they are associated with normal biological functions, analysis of primary metabolites enhances our understanding of how the biological system works, and helps explain the observed phonotypes systematically (*8*).

An important example for the analysis of primary metabolites is the use of machine learning techniques to identify metabolite production patterns that are associated with different experimental conditions. Identification of metabolites that are associated with certain drug treatment outcome can classify patients who likely to respond to treatment (*9*). Alternatively, metabolic biomarkers can be used for early disease prognosis and progression (*10*).

Analysis of primary metabolites is often done systematically by one of two methods: metabolite identification or network analysis. First, in metabolite identification method, the presence or absence, as well as the concentrations of individual metabolites are directly measured. Collective experimental measurement of metabolites is referred to as metabolic profiling or shortly metabolomics. Briefly, samples of biological fluids are analyzed using spectroscopic techniques such as nuclear magnetic resonance (NMR), liquid chromatography–mass spectrometry (LC-MS), Gas chromatography–mass spectrometry (GC-MS) and measured spectra are matched against a database containing spectra of known metabolites.

Second, network analysis method aims to discern the metabolic state of the whole metabolic network, thereby identifying which metabolites are present. Unlike metabolite identification, network analysis considers the interactions between the metabolic system and other cellular systems, providing a more holistic approach. Therefore, metabolic activity can be characterized with experimental measurements of higher-order systems, such as gene expression.

1.1.2.2. Analysis of secondary metabolites

Secondary metabolites are involved in plant or microbial defense against predators, and hence many secondary metabolites posses potent bioactivities. The study of bioactive secondary metabolites with the goal of identifying new drug leads is referred to as natural products research. Over the last centaury, natural products research has fueled the drug discovery pipeline with novel scaffold that are not easily accessible through combinatorial chemistry (*11*).

Currently, identification of bioactive natural products involves the isolation and purification of individual compounds followed by a myriad of spectral measurements including multi-dimensional NMR and mass spectrometry (MS). Finally the acquired spectra are carefully interpreted by experts to elucidate the chemical structure of a single metabolite.

Despite the similarity between metabolomics and natural products research, the latter still relies on individual metabolite identification rather than systematic. This is in part due to the scarcity of databases containing spectra of known natural products as well as accurate methods for spectral matching.

1.1.2.3. Analysis of recombinant metabolites

Recombinant metabolites play an important role in industrial production of biopharmaceuticals. The metabolic analysis of recombinant metabolites is referred to as metabolic engineering, in which the goal is to reconstruct metabolic pathways in order to maximize the product of desired metabolites (*7*). Metabolic network reconstruction and analysis of reaction fluxes offer a computational modeling tool to optimize metabolite production (*12*).

1.2. The need for software tools for metabolic analysis

The recent technological advances enabled genome-wide experimental measurement to be acquired at reduced cost and time. Also, the study of biological systems has uncovered previously unknown complexity. As a result, holistic analysis of large datasets on complex biological models is becoming manually infeasible. Computational tools are needed for data processing, modeling, analysis and visualization.

From the wide applications of metabolic analysis, this thesis focuses on two aspects where limited computational tools were available: 1) Metabolic path mining from gene expression, and 2) processing NMR data for metabolite identification. I discuss each point in detail below.

1.2.1. Metabolic path mining from gene expression

Network analysis is an important method for the analysis of cellular primary metabolism, in part because of three main reasons: 1) Metabolites are identified systematically, and therefore giving a more holistic model of the biological system. 2) The ability of incorporating prior knowledge by using metabolic networks that are constructed by human curators from literature. 3) The ability to infer the presence / absence of metabolites from easier-to-measure systems, such as transcription. Gene expression values can be used to identify the metabolic activity when analyzed within a network context.

One network analysis method, network path mining, is particularly useful in metabolic analysis. Network path mining takes a genome-scale network along with gene expression values, and enumerates, from within all possible paths in the network, a list of linear paths that are highly activated. Within the context of metabolic networks, linear paths represent metabolic cascades.

Despite importance of metabolic path mining and biologically intuitive meaning of its output, an easy-to-use software tools were lacking. Software tools are particularly needed for metabolic path mining, because the size and complexity of genome-scale networks warrants manual analysis infeasible. Moreover, since path mining relies on path enumeration, thousands of paths are given as output. Effective clustering and visualization of output paths via a software tool is needed.

I developed NetPathMiner, which is an R package for mining active metabolic paths from genome-scale networks based on gene expression. NetPathMiner allows easy incorporation of prior knowledge by constructing networks from various pathway file formats. Also, NetPathMiner handles genome-scale network

analysis efficiently, and provides interactive visualizations for networks and output paths.

1.2.2. Processing NMR data for metabolite identification

Metabolite identification from experimental measurements is widely used in both metabolomics and natural products research. Unlike inference methods, metabolite identification provides direct evidence for the presence or absence as well as the concentration of a particular metabolite in the measured sample.

Among spectroscopic techniques used for metabolite identification is NMR, which provides detailed information about the structural features of the measured metabolites. Moreover, NMR allows the structure determination of novel metabolites, particularly useful in natural products research, in which the goal is to identify new structural scaffolds.

Because of the nature of experimental technique, measured NMR spectra are not interpretable before they pass through a series of processing steps. Processing NMR spectra handles two issues: 1) transform the data from instrument-specific readings to human readable formats, and 2) correct the variations and artifacts present in the spectra due to inadvertent experimental conditions.

To first identify the lacking points in the computational processing NMR spectra, I surveyed current computational resources for dereplication of natural products. Dereplication is a technique for rapid identification of previously known metabolites from a natural extract, thereby reducing the time and effort to discover novel metabolites. I discussed three important resources: databases, processing methods and software. The literature survey revealed two major shortages: 1) scarcity of free-to-use spectral databases and 2) lack of easy-to-use free tools for processing NMR spectra.

To address the identified shortage in software tools, I developed NMRPro, which is a web component for interactive processing and visualization of NMR spectra. NMRPro provides a web-based solution for processing NMR spectra, which allows easy sharing of raw and processed spectra between collaborators.

Moreover, distributing the software as a web component enables its integration into current web servers.

1.3. Thesis organization

This thesis consists of five chapters, three main chapters besides introductory and conclusion chapters (Table 1.1). The current chapter provides a background on the field of metabolic analysis and its current goals and challenges. Chapter 2 discusses NetPathMiner, which is a tool for metabolic network analysis. Chapter 3 presents a survey of the currently available computational tools for metabolite identification in natural products research, identifying several lacking resources including easy-to-use online NMR processing software. In Chapter 4, I present NMRPro as a tool to overcome the current lack in interactive processing software of NMR spectra. NMRPro can be considered as a building block for web-based software for analysis of metabolomics and natural products data. Finally, Chapter 5 present a thesis summary and future remarks.

Table 1.1 Chapter contents

	Chapter 2	**Chapter 3**	**Chapter 4**
Metabolite Type	Primary	Secondary	Primary, Secondary
Metabolic analysis method	Network analysis	Metabolite identification	Metabolite identification
Description	Software tool	Survey	Software tool
Data	Gene expression	NMR spectra	NMR spectra

Chapter 2

NetPathMiner: R/Bioconductor package for network path mining through gene expression

Chapter Contents

Chapter Summary **8**
2.1. Introduction **9**
2.2. Input to network path mining **12**
2.2.1. Network: 12
2.2.1.1. Network representation 12
2.2.1.2. Network origin 13
2.2.2. Gene expression matrix 14
2.3. Workflow of NetPathMiner **15**
2.3.1. Pathway File Processing (Step 1 in Figure 2.1) 16
2.3.1.1. Network attributes: 17
2.3.2. Network Manipulation (Step 2 in Figure 2.1) 17
2.3.2.1. Network representations 17
2.3.2.2. Network Editing 18
2.3.3. Weighting the network (Step 3 in Figure 2.1) 20
2.3.4. Path Ranking (Step 4 in Figure 2.1) 20
2.3.4.1. Probabilistic Shortest-path Method: 21
2.3.4.2. P-value Method: 22
2.3.5. Paths Clustering and Classification (Step 5 in Figure 2.1) 23
2.3.6. Visualization (Step 6 in Figure 2.1) 23
2.4. Additional functionalities **26**
2.4.1. Analysis of Signaling Networks: 26
2.4.2. Integration with other R packages 26
2.5. Conclusion **26**

Chapter Summary

NetPathMiner is a general framework for mining, from genome-scale networks, paths that are related to specific experimental conditions. NetPathMiner interfaces with various input formats including KGML, SBML and BioPAX files and allows for manipulation of networks in three different forms: metabolic, reaction and gene representations. NetPathMiner ranks the obtained paths and applies Markov model-based clustering and classification methods to the ranked

paths for easy interpretation. NetPathMiner also provides static and interactive visualizations of networks and paths to aid manual investigation.

2.1. Introduction

Mining subnetworks from genome-scale biological networks is an important step in biological data analysis, because as their size and complexity increase, manual analysis becomes infeasible. Such networks are highly modular and span over several biological processes, and hence, only certain parts of the network are activated under a particular biological condition. Therefore, given biological experimental data along with a genome scale network, active subnetwork detection remains a non-trivial step in data analysis and mining.

Numerous methods for active subnetwork detection using experimental data have been described in the literature (*13-16*), where these methods provide various output formats. Taking metabolic network analysis as an example, active metabolic subnetworks inferred from gene expression data can be expressed as node clusters (*17, 18*), or as a set of linear paths (*19, 20*). I focus on linear paths, which are particularly useful by carrying an intuitive meaning, as in metabolic reaction paths and signaling cascades.

Currently, network path mining is hampered by two main challenges: *i*) Constructing genome scale networks from curated pathway databases and *ii*) visualization of output paths. Genome scale metabolic networks can be constructed by connecting individual pathways from available databases. Several online databases have tried to catalogue biological knowledge into human interpretable pathways representations, such as KEGG (*21*), Reactome (*22*), BioCyc (*23*) and Pathway Commons (*24*). Although such data are readily accessible though various standard formatted files, such as KGML, SBML and BioPAX, two main issues arise. Firstly, each file may represent a particular pathway or a biological process, rather than the full network, and therefore, several files have to be concatenated to obtain the genome scale network. Secondly, network preprocessing, such as removing nodes with missing annotations, may be necessary for efficient path mining. The second challenge to network path mining is visualization. Network path mining can output a large

number of active paths, in some cases thousands, making their visualization and biological interpretation strenuous.

Table 2.1 Functionalities of current network path mining tools

	PathRanker	rBiopaxParser	PathView
Input network format	KGML	BioPAX	KGML
Supported network types	Metabolic	Metabolic & Signaling	Metabolic & Signaling
Network representation conversion	Limited	✗	✗
Path extraction	✓	✗	✗
Visualization	Paths only	Networks only	Networks only

Despite the importance and wide applicability of network path mining in biology, a universal tool implementing the process flow of biological path mining in R/Bioconductor is still at short (Table 2.1). Hancock and colleagues previously developed PathRanker, an R package for mining metabolic pathways from gene expression data (*20*). However, PathRanker is limited only to metabolic networks constructed from KGML files, restricting its use to KEGG metabolic pathways. Moreover, the absence of a standard format to represent network objects in R hinders integrated network-based analysis. Another tool, *Pathview* (*25*), provides ways to integrate and visualize KEGG metabolic and signaling pathways in data analysis through powerful attribute mapping functions. However, *Pathview* lacks network path mining methods and is also limited to KGML formatted files. *rBiopaxParser* is also an R package parsing and visualization of BioPAX formatted files, although current visualization functions is limited to regulatory networks (*26*). Other packages such as *KEGGgraph* and *graphite* available on Bioconductor are limited to specific databases or limited to particular types of pathways (*27-29*).

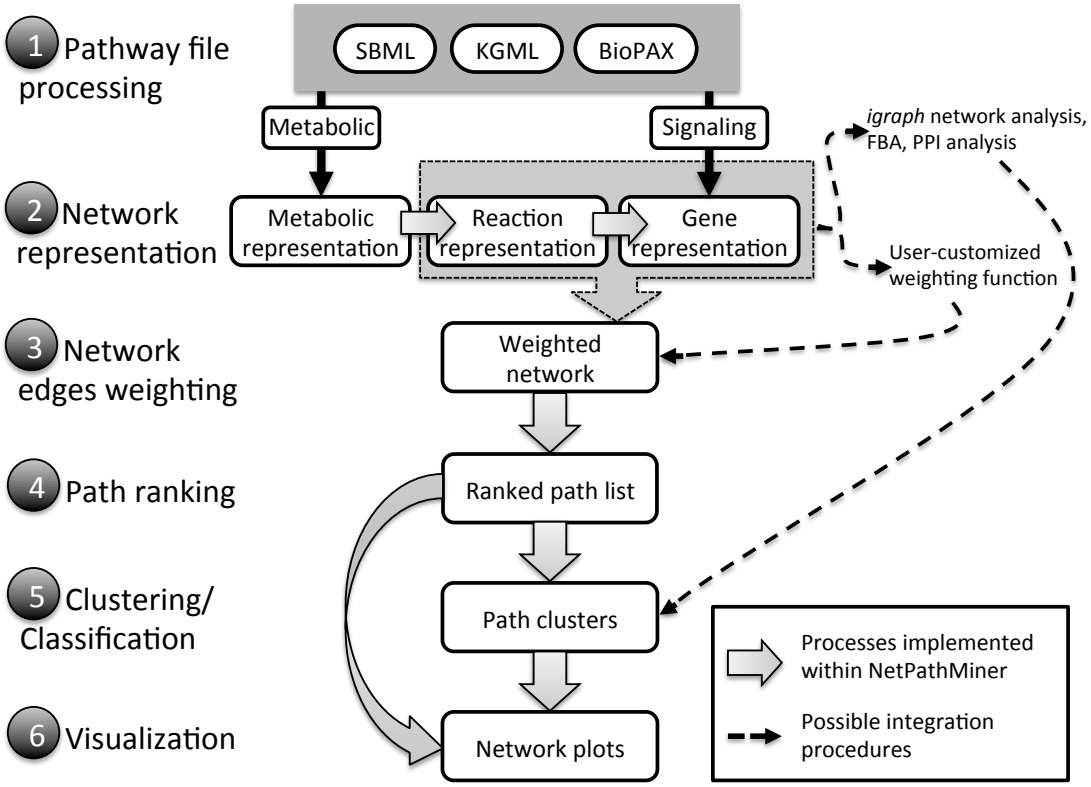

Figure 2.1 General workflow and modules of NetPathMiner

This chapter presents NetPathMiner; a general framework for network path mining in R. NetPathMiner provides several functions for full network construction using different pathway file formats, enabling its utility to most common pathway databases. Borrowing and extending upon path mining methods presented in PathRanker, NetPathMiner enables a flexible module-based process flow for network path mining and visualization (Figure 2.1), where each step can be replaced by user-customized functions. Network representation using *igraph* (*30*) allows integrated network analysis. Finally, visualization of output paths is achieved by combining clustering methods (*31, 32*) and plotting functions in *igraph* package.

The rest of chapter starts by describing the inputs for network path mining., followed by discussion of each module of NetPathMiner in step-by-step fashion. The chapter concludes by comparing the performance NetPathMiner with existing software.

2.2. Input to network path mining

Network path mining takes two inputs; network structure and gene expression matrix, and produces linear paths as output. This section discusses the characteristics of both inputs.

2.2.1. Network:

Because the network acts as the guide map in which the subsequent analyses occur, it is important to address the different ways a network can represent the biological system (*metabolism, in our case*), and the methods to obtain such networks.

2.2.1.1. Network representation

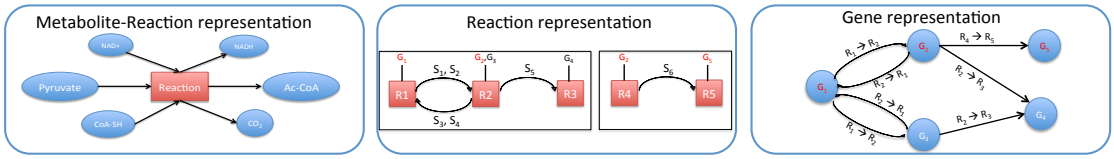

Figure 2.2 Examples of metabolic networks in different representations.

Metabolic networks provide a graph representation for the biological system. Metabolic network is consisted of two main components, nodes and edges. Nodes represent biological entities, such as proteins, reactions or metabolites. Edges connect between nodes to indicate the relationship between different entities. As shown in Figure 2.2, metabolic networks can have different representations, depending on what the nodes and edges represent. Metabolite-reaction networks are bi-partite graphs, i.e. containing two types of nodes, metabolites and reactions. The edges indicate whether a metabolite is consumed or produced by a reaction. Reaction networks contain only reaction nodes, in which edges connect between successive reactions. Finally, gene networks expand each reaction node to its catalyzing genes. Expanding a reaction network into genes can result in ambiguities when reactions are catalyzed by more than one gene. Additionally, connected gene networks can be obtained from disconnected reaction networks if genes are participating in several reactions **(**Figure 2.2**)**.

2.2.1.2. Network origin

Networks can either be obtained based on prior knowledge from literature, called curated (*21, 22*), or directly inferred from experimental data (*33*). I discuss here the characteristics of each network type an how it affect the subsequent path mining.

Curated networks are stored in pathway data givesbases, such as KEGG (*21*), Reactome (*22*) and BioCyc (*23*), which are constructed by extracting the biological knowledge from the literature. Curated networks extracted from these sources are highly annotated and reliable, as these networks have been under extensive revision and annotation. However, curated networks do not specify the conditions under which the network structure is valid. Moreover, networks are also confined only to well-studied genes, covering only a small portion of the genetic landscape. For example, the most extensive human PPI network constructed from the literature covers only 49% of all proteins in Swiss-Prot database (*34*). Moreover, it is estimated that current interaction maps covers less that 10% of all potential protein interactions (*35*). Curated networks, therefore, are information rich and reliable, however they can be out of context and with low coverage.

Networks inferred from specific experimental measurements (*33, 36*) predict interactions that may be present under particular experimental conditions. The experimental data provide "*context*" to the constructed networks and as a consequence, such networks have different structures under different experimental conditions. Examining such models gives insight into the dynamic nature of biology, and how a living organism copes different environmental stresses with few genetic elements (*37*). Although constructed networks provide more coverage than curated networks, they usually suffer from poor reliability, as high throughput techniques may produce noisy experimental measurements. For example, interaction networks constructed by mass spectrometry techniques have been found to vary greatly across experiments (*38*).

Network path mining takes genome-scale curated networks as input, and uses experimental data (*in our case gene expression*) to weight the network. As

discussed, curated networks are limited to well-studied genes and pathways, and therefore, a significant portion of the gene expression measurements is not omitted.

2.2.2. Gene expression matrix

The technological advances over the last decade have enabled high throughput and accurate measurement of gene expression profiles. The development of DNA microarray (*39*) followed by the more recent RNA-seq (*40*) and SAGE (*41*) technologies allowed accurate quantification of RNA transcripts with reasonable effort and cost. The genome-scale measurement of expression provides a snapshot of the cellular state, allowing deep investigation of the biological processes, including metabolic analysis.

Inference of metabolic activity from gene expression measurements traverses multiple biological systems (*42*). While gene expression values represent RNA transcription, metabolic activity cannot be inferred directly therefrom. First, the transcribed RNA is translated into proteins, which act as enzymes. These enzymes are then activated through post-transcriptional modifications. Then, activated enzymes control metabolic flux, affected by interplay of metabolite concentration, enzyme levels, and reaction fluxes in a highly connected network.

Metabolic activity can be inferred from coordinated gene expression. Metabolism is a dynamic and coordinated activity, whose behavior differs in organisms, organs, tissues, subcellular location and external environment conditions (*43*). Therefore, under each specific condition, only portions of the possible paths would be preferentially co-regulated, and thus would include highly correlated gene expression (*19*). Extracting top correlated paths from a list of all possible paths based can be considered as most active metabolic paths.

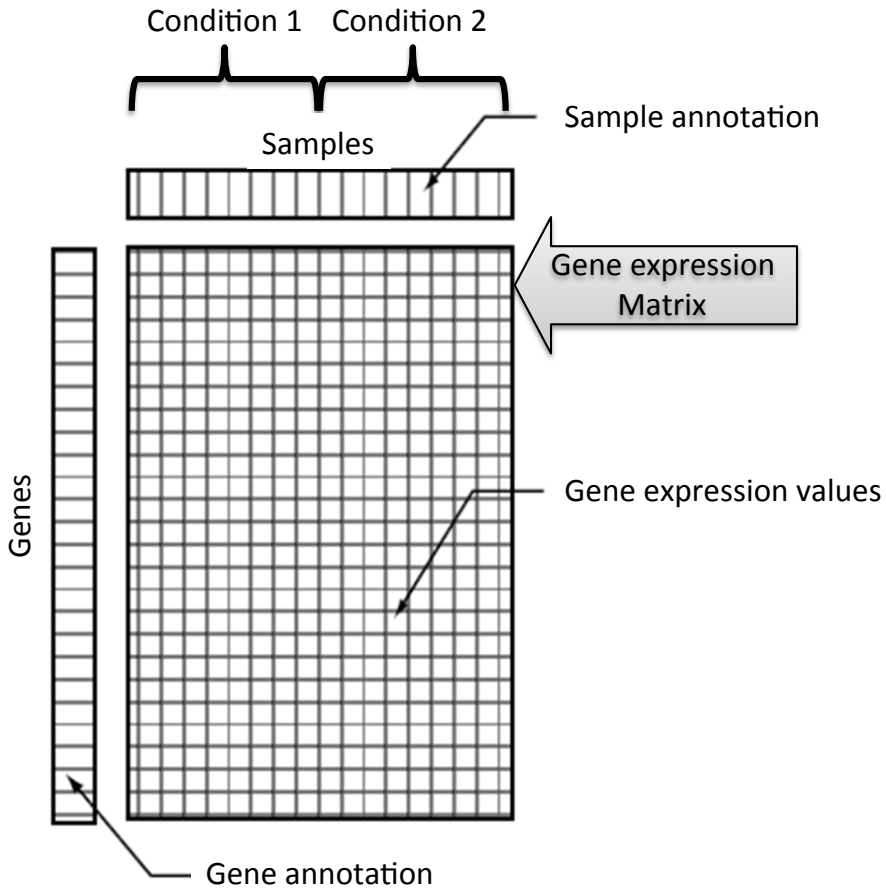

Figure 2.3 Gene expression matrix. Rows are genes and columns are samples. Samples can be divided into groups according to the experimental conditions.

Figure 2.3 shows an example of a gene expression matrix used as input to network path mining. Gene expression measurements (*rows*) for each sample (*columns*) are provided as numerical values. The correlations between adjacent genes (with respect to the input network) are used to weight the edges of the network. Using a weighted network, top *k* active paths are then extracted.

2.3. Workflow of NetPathMiner

NetPathMiner package contains several functions necessary to automate network path mining. The basic flow chart is presented in Figure 2.1. Although the process flow is optimized for metabolic network analyses, NetPathMiner can be also applied similarly to other types of networks such as signaling and regulatory pathways.

2.3.1. Pathway File Processing (Step 1 in Figure 2.1)

NetPathMiner gives the user the option to choose from the available pathway databases by supporting the commonly used pathway file formats: KGML, SBML and BioPAX. Table 2.1 summarizes the differences between various file formats. For detailed discussion about different standard formats, I refer to (*44*). KGML files are specific to KEGG database, where each KEGG file contains information about a single pathway in a particular species. A list of KGML files can be supplied to NetPathMiner where they are combined into a single network. In contrast, SBML and BioPAX, each file may contain one or more pathways. For example, Reactome (*22*) offers database download in a single file both in SBML and BioPAX formats, which may be provided also as input to NetPathMiner. Such files are large, and their parsing tends to be slow if done in R. With the exception of BioPAX parsing, all the text parsing for KGML and SBML files is carried out using efficient C++ libraries for speed optimization. For BioPAX formatted files, I opted to make use of functions provided in *rBiopaxParser* (*26*).

Table 2.2 Differences between major pathway file formats.

Features	KGML	SBML	BioPAX
Number of pathways per file	One	One	One or more
Are metabolic reactions distinct?	Yes	No	Yes
Transport reactions	No	Yes [a]	Yes
Reaction kinetics	No	Yes	No
Cellular location	No	Yes	Yes
MIRIAM annotations	No	Yes	Yes
Databases	KEGG	Reactome, Biomodels, Recon X	PID, Reactome, BioCyc, Biocarta, WikiPathways

[a] Transport reactions are detected indirectly from reaction description.

Similar pathways may be represented differently due to the discrepancy in how different databases and file formats represent the data. To alleviate the effect of such discrepancies between file formats on the constructed networks, the user can choose whether to parse pathways as metabolic or signaling networks. Metabolic reactions are represented as “reactions” in both SBML and KGML formats, while in BioPAX, they are termed “biochemical reactions” to discriminate them from other transport and assembly reactions. The resulting

network here is given as bipartite graph, where metabolite and reactions are represented as different vertex types. In contrast, when pathways are parsed as signaling pathways, the output network is a gene network. Signaling pathways can be parsed from KGML format using "relation" attribute between different proteins, and from BioPAX format as a "control" class. SBML doesn't provide a way to differentiate between metabolic reactions and other types of reactions, thus, signaling pathways are first parsed as metabolic bipartite graph, which is then converted to a gene network.

I choose *igraph* package to represent all our constructed network objects in R to efficiently handle large graphs commonly encountered in biology and to allow NetPathMiner to integrate with other network analysis tools. *igraph* is an R package that contains a comprehensive set of functions for analysis of complex networks (*30*). Besides being able to handle large graphs efficiently, commonly encountered in biology, *igraph* representations allow the integration of NetPathMiner with other analytical methods in the package. In addition, igraph objects provide a standard format for network objects facilitating future development.

2.3.1.1. ***Network attributes:***

NetPathMiner uses MIRIAM identifiers (*45*) to standardize annotation attributes across different file formats. NetPathMiner attempts to extract most of the vertex attributes available in each file format, such as Uniprot, kegg.compound, GO, ChEBI identifiers using URI syntax. Moreover, the user can provide additional attribute names, where the parser searches for such attributes, and fetches them. Moreover, NetPathMiner also implements an attribute fetcher using BridgeDb web service (*46*) to convert between different MIRIAM annotations.

2.3.2. Network Manipulation (Step 2 in Figure 2.1)

2.3.2.1. Network representations

Network representation involves how the biological information is incorporated in the network structure. I explain below the different representations available.

Metabolic network is a series of chemical reactions, in which a gene or a set of genes catalyze each reaction. Each chemical reaction consists of substrates (chemical compounds consumed in the reaction), products (compounds produced) and annotated genes.

NetPathMiner provides three network representations for metabolic networks: **i)** ***Metabolic representation*** which is as a directed bipartite graph G(V,E) and V = {M **υ** R} where *M* and *R* are sets of metabolites and reactions, respectively. Reaction vertices *R* represent the transition events themselves, and the direction of an edge *e* (*r* , *m*) indicates whether metabolite *m* is a substrate or a product. **ii)** ***Reaction representation*** deletes metabolite *M* vertices, retaining them as edge attributes between reactions. **iii)** ***Gene representation*** expands reaction vertices into their catalyzing gene(s). Since certain genes may participate in several reactions, separate gene vertices are created for each reaction that they participate in.

2.3.2.2. Network Editing

NetPathMiner implements several network-editing functions to amend those provided by *igraph*. NetPathMiner provides functions to delete vertices and expand gene complexes.

*2.3.2.2.1. Ubiquitous metabolites:*Ubiquitous metabolites, such as currency compounds (ATP, CO_2) and reaction cofactors are prevalent in metabolic networks. However, connecting reactions through these metabolites may not be biologically meaningful. NetPathMiner can either remove ubiquitous metabolites, or create separate vertices for each reaction they participate in.

2.3.2.2.2. Reactions with missing genes

NetPathMiner relies on the gene annotations of reaction nodes to find correlated paths, and therefore reaction nodes with no annotated genes represent discontinuity in the genetic component of the network structure. There are three main reasons for a reaction node to have no associate genes: 1) Spontaneous reactions (Figure 2.4a), 2) Translocation reactions (Figure 2.4b), which transport metabolites across cellular membranes, however involve no chemical

modification and 3) Missing annotations. NetPathMiner allows the user to detect spontaneous and translocation reactions and remove them, without affecting the biological interpretation (Figure 2.4).

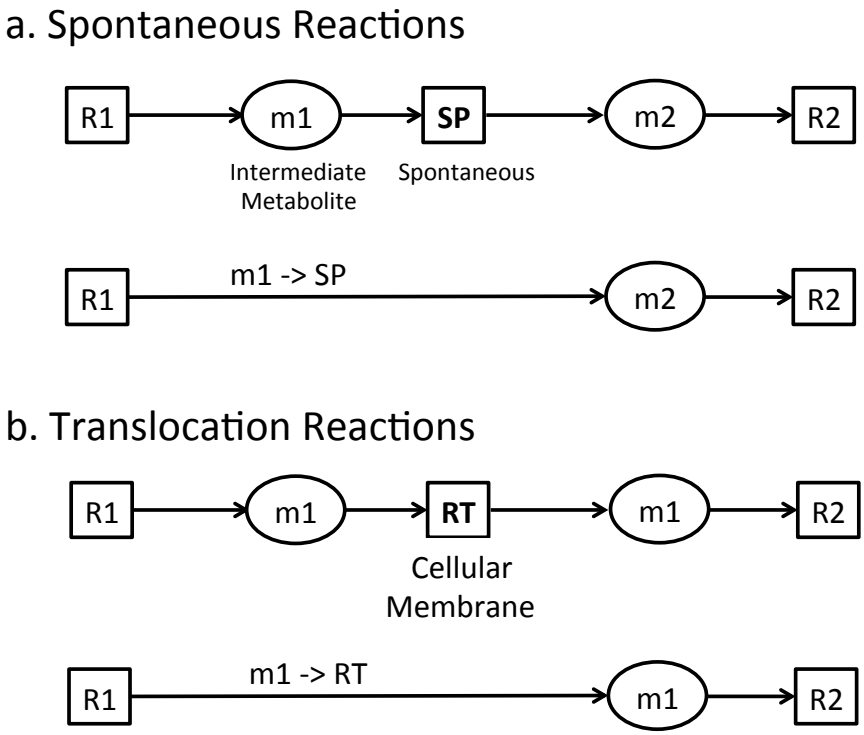

Figure 2.4 Examples of reactions with no associated genes

2.3.2.2.3. Vertex expansion and contraction:

NetPathMiner provides functions to expand / contract vertices by their annotation attributes, useful in expanding protein complexes. Vertex expansion can be utilized to unify annotations used in networks from different databases. For example, to compare Reactome networks with KEGG ones, metabolite vertices can be expanded to their KEGG compound annotations. Vertex contraction, on the contrary, can be used to examine interactions between sets of vertices, such as pathways and gene sets. For example, contracting vertices by their pathway annotations yields a network in which pathways are vertices and edges represent their crosstalk. Similar technique can be used to investigate metabolite transport between cellular compartments.

2.3.3. Weighting the network (Step 3 in Figure 2.1)

In this step, edges on the provided network are assigned weights according to Pearson correlation of gene expression. Gene expression profiles should be provided as a numeric matrix, where rows represent genes and columns represent biological samples. Importantly, in the gene expression matrix, gene IDs must match the IDs annotated in the input network. Moreover, biological samples can be further labeled into categories (control/treatment, alive/deceased), where edge weights are computed for each label separately. Aside from the provided weighting function, users can provide edge weights computed from a customized function without altering the rest of the process flow.

2.3.4. Path Ranking (Step 4 in Figure 2.1)

Path ranking functions attempts to find a set of paths of node/edge sequences (paths) maximizing edge weights. Generally, paths are extracted between two sets of nodes, starting nodes, denoted as **S**, and target nodes, denoted as **T**. By default, NetPathMiner uses all entry and exit compounds of the metabolic network as starting and target nodes, respectively. However, **S** and **T** can be specified by the user as input.

Currently, two methods for path ranking are implemented in NetPathMiner, "*shortest.path*" or "*p.value*" returning outputs a list of k-most probable paths, or a list of paths passing a p-value cutoff, respectively. Path ranking functions can be used independently or as part of the described process flow. In both cases, all what is required is a weighted igraph object, and functions can return a ranked path list.

NetPathMiner ranks paths from networks by one of the two methods, probabilistic shortest-path and p-value methods. Both statistical methods rank paths by their edge weights, in which paths with larger edge weights are ranked higher.

2.3.4.1. Probabilistic Shortest-path Method:

Given a weighted network, the method identifies top *K* paths between sets of start **s** and end **t** vertices. The probabilistic shortest-path method is described in detail in a previous paper (*19*), and was implemented in a previous package PathRanker (*31*). Briefly, the method considers the empirical cumulative distribution function (ECDF) of all edges to probabilistically rank the edges in the network. For each edge $e \in E$, the probability of an edge weight is given by **(1)**:

$$prob(e) = P_{ECDF}(e) \tag{1}$$

where $P_{ECDF}(e)$ is probability of getting an edge weight of less than or equal to that of e from the empirical distribution of all edge weights in the Network. Therefore, for a path p consisting of a sequence of n edges will be:

$$prob(p) = \prod_{i=1}^{n} P_{ECDF}(e_i) \tag{2}$$

Here I set **s** and **t** sets as entry and exit nodes of the network, allowing the enumeration and ranking of paths across the network. To formulate the problem as a shortest path problem **(2)** is redefined as:

$$score(\pi) = \sum_{i=1}^{n} -\log\left(P_{ECDF}(e_i)\right) \tag{3}$$

If π is the path p score, the shortest path problem can be solved by minimizing the value of *score(π)*. Computationally, K-shortest paths are enumerated by Yen-Lawler algorithm, which uses dynamic programming to solve the problem in polynomial time (*47, 48*). If π is the path p score, the shortest path problem can be solved by minimizing the value of *score(π)*. Computationally, K-shortest paths are enumerated by Yen-Lawler algorithm, which uses dynamic programming to solve the problem in polynomial time (*47, 48*).

When ranking paths using the shortest-path method (*20*), two parameters can be tuned by users. The first parameter is number of returned paths *K*. While a large *K* will increase the computation time significantly, limiting the returned path list to a few paths will not recover all correlated paths over the network. From

previous real data experiment, I concluded that K=1,000-10,000 is reasonable for genome-wide metabolic network analysis, and can be decreased for smaller networks (*20*). The second parameter is the minimum returned path length. Since the shortest-path based ranking will tend to return very short paths, often biologically uninteresting, setting a minimum path length allows the investigation of longer, biologically relevant paths. However, increasing the threshold for returned path length will also increase the computation time.

2.3.4.2. *P-value Method:*

The probabilistic method described above aims to minimize path scores, and therefore is biased to shorter paths. The p-value method presented in (*49*) corrects the path length dependency by reformulating the problem into a p-value minimization problem to find paths of which the sum of edge weights are significantly larger than those of random paths of similar lengths.

For sets of start **T** and end **T** vertices, finding paths with minimum p-value relies on a two-step algorithm. For each $s \in \mathbf{T}$ and $t \in \mathbf{T}$ vertices, first, a list shortest paths of all possible lengths is enumerated. Second, calculating the p-value for this list to identify the most significant path between *s* and *t*.

P-values of paths are estimated based on the empirical distributions of path scores of similar lengths (simply the sum of their edge weights). The empirical distributions are estimated by randomly sampling paths from the network. Paths of increasing lengths are randomly sampled using Metropolis sampling algorithm (*50*), and the probability of path scores are stored as reference to compute the p-values for shortest path list. For detailed discussion about the method and algorithm, I refer the supplementary methods in (*49*).

Users using p-value method can set a p-value cutoff, in which paths with lower p-values are extracted. Since the method corrects for path length dependency, setting a minimum path length threshold is unnecessary. However, a maximum path length can be set to limit the computation time.

2.3.5. Paths Clustering and Classification (Step 5 in Figure 2.1)

Network path mining functions return a large number of paths, hampering their manual investigation. For example, to uncover most of the correlated paths in a full metabolic network, 1,000-10,000 paths should be extracted. To facilitate the analysis of such large number of paths, I include path clustering methods in NetPathMiner package.

To cluster extracted paths according to their structure, *pathCluster* function utilizes the 3M Markov mixture model (*32*), which identifies *M* key functional components by using the Markov structure of all extracted paths. With a user-specified *M* as an input, paths can be grouped into *M* clusters according to their underlying functional structure, making their analysis more feasible. Alternatively, when it is interesting to find paths that are specific to a certain biological condition, NetPathMiner *pathClassifier* function uses a supervised version of the 3M model to identify a set of paths that can be used to classify a particular response label (*31*). Both clustering and classification methods are adopted from our previous package, PathRanker. For detailed discussion of the methodology I refer to (*20*).

2.3.6. Visualization (Step 6 in Figure 2.1)

NetPathMiner provides both static and interactive visualizations of ranked paths using annotation information and machine learning techniques, making manual investigation easier. Figure 2.5a-c show a visualization example of different graph representations using the output of the last step, allowing users to examine metabolic regulation at different biological system levels. Visualization function matches vertices across all input representations and plots them using the same layout. To make visualization of a huge number of paths clearer, NetPathMiner assigns the same color to all paths in each obtained cluster, and assigns the same color to vertices within the same cellular compartment (Figure 2.6). Figure 2.5d-e show vertices in each path as well as probability of each path belonging to clusters.

NetPathMiner maximizes the use of annotation attributes in network visualization to enhance manual investigation. Figure 2.6 shows a gene network

where vertices in the same cellular compartment have the same color, and drawn closer to each other in the layout.

NetPathMiner also supports interactive visualization in Cytoscape by either exporting networks in GML format or using *RCytoscape* (*51*), which allows thorough investigation of vertex annotations and full customization of network colors and layout. Exporting networks to Cytoscape, allows the integration with its functions and plugins. Figure 2.7 shows the same network in Figure 2.6, visualized in Cytoscape, using the same layout, allowing the user to interactively select and investigate individual vertices or edges.

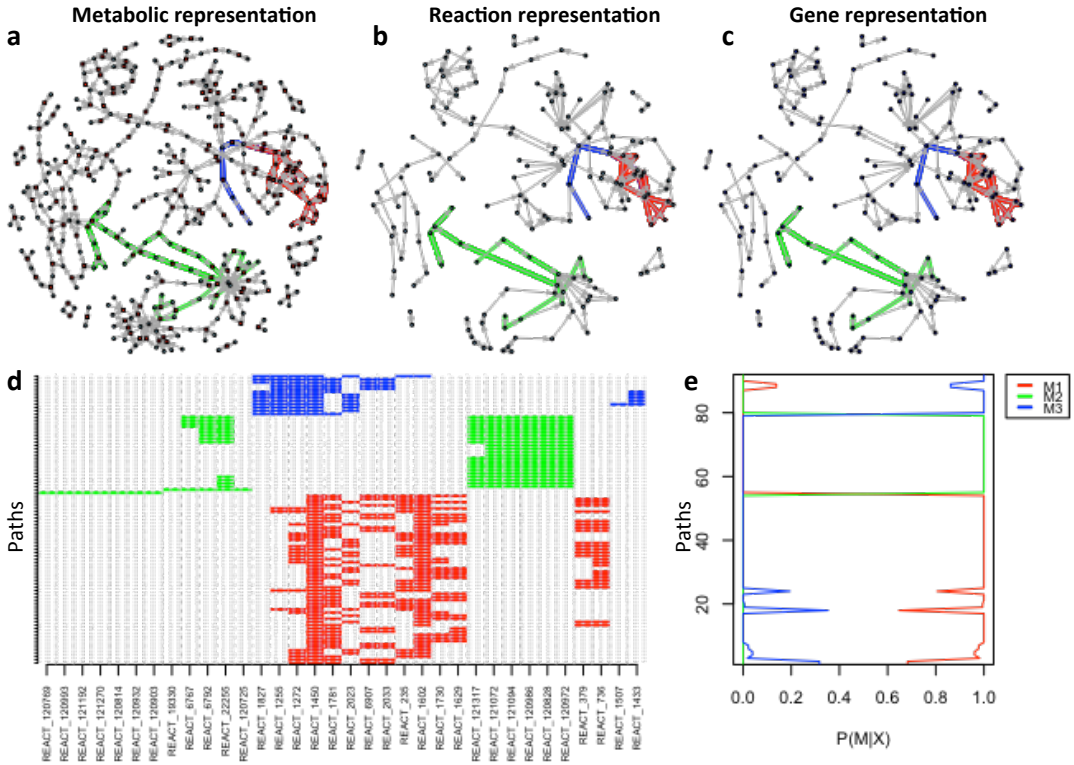

Figure 2.5 NetPathMiner path visualization. Carbohydrate metabolism network extracted from Reactome, and analyzed with NetPathMiner. Top 100 paths were extracted, and groups into 3 clusters. Paths plotted on different network representations and colored by cluster mem-bership. (a) Metabolite-reaction bipartite representation. (b) Reaction network representation plotted using the same layout as a. (c) The underlying gene network of carbohydrate metabolism, plotted using the same layout. (d) Paths (rows) and their components (columns), colored by cluster membership. (e) Probabilities that each path belongs to its assigned cluster.

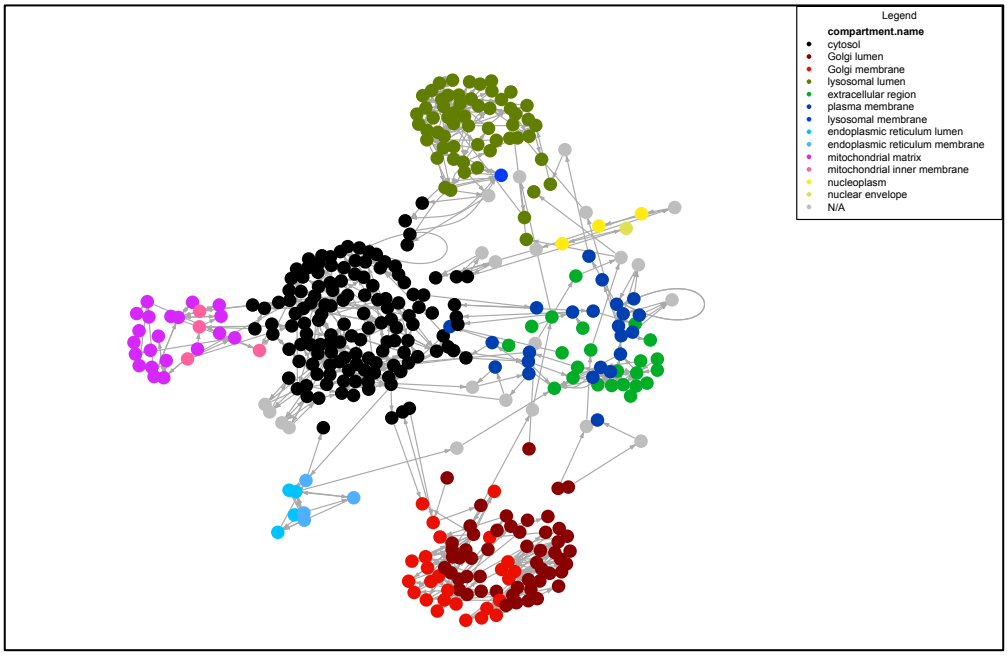

Figure 2.6 Carbohydrate metabolic network in gene representation, with vertices colored by subcellular compartment (plotted in R).

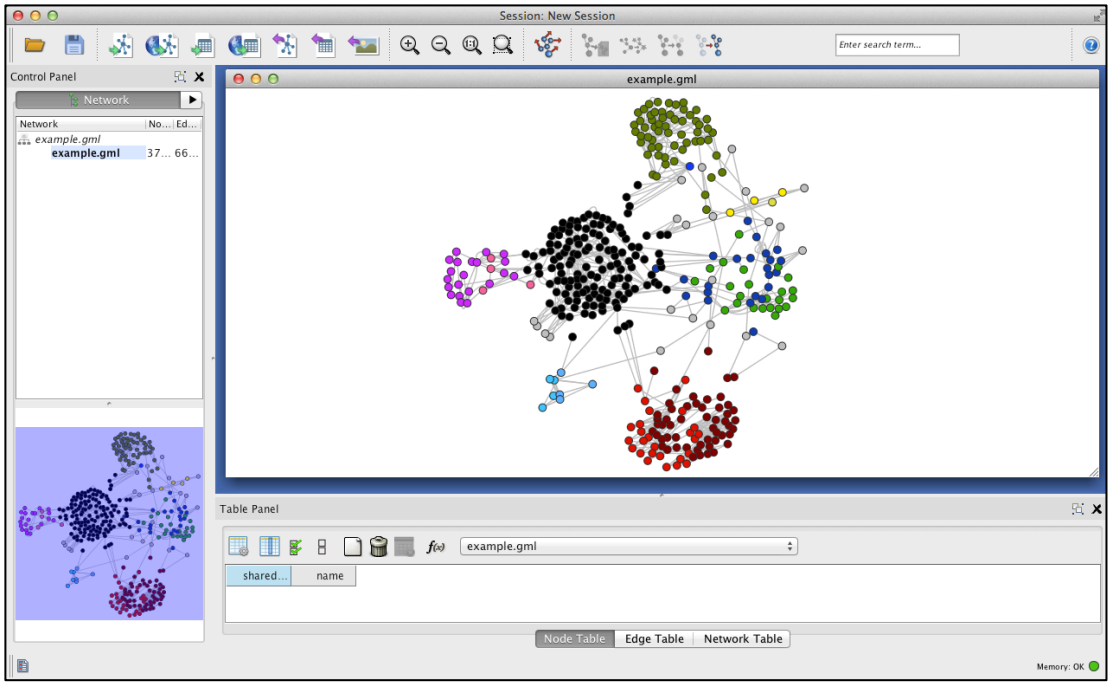

Figure 2.7 Cytoscape plots for the Carbohydrate metabolic network in gene representation, with vertices colored by subcellular compartment.

2.4. Additional functionalities

In addition to metabolic network path mining, NetPathMiner provides additional functionalities that are helpful to general network analysis. This section discusses some of these functionalities.

2.4.1. Analysis of Signaling Networks:

While NetPathMiner focuses on metabolic network analysis, the concept of network path mining is also applicable to signaling networks, in which case the extracted linear paths represent signaling cascades. Signaling networks describes the interactions between genes as directed graph. NetPathMiner constructs signaling networks from two types of biochemical reactions, signaling and metabolic reactions. First, signaling reactions include activation/inhibition, transcription regulations, where edges are directed from the activator to the activated gene (and similarly from regulator to regulated). Second, genes catalyzing successive metabolic reactions are considered to interact through a metabolite, where one gene produce the metabolite and the other gene consume it. NetPathMiner represents signaling networks as *gene representations.*

2.4.2. Integration with other R packages

Although NetPathMiner represents networks as igraph objects, it provides functions to convert networks to graphNEL objects (*52*), offering direct integration with a wide range of R packages in Bioconductor (*29*). Moreover, NetPathMiner implements functions to generate gene sets using vertex annotations in a network. For example, from a genome scale network, *getGeneSets* function can generate a list of pathways and vertices belonging to each pathway for direct integration with gene set enrichment analysis (GSEA) methods (*53*). In some GSEA methods, where network structure is factored (*54*), *getGeneSetNetworks* functions can be used instead.

2.5. Conclusion

I present NetPathMiner, an easy-to-use R package for network path mining. NetPathMiner constructs genome scale network from most common pathway file formats, overcoming the current database specificity. NetPathMiner also provides different visualizations for output paths, facilitating manual

investigations. I emphasize that functions in NetPathMiner can be fully integrated with other network analysis procedures. Future developments include providing the package as a web application for a wider audience. With this R package, I hope to ease the challenges faced by biologists in network path mining, enhancing its applicability in biological data mining.

Chapter 3

Current status and prospects of computational resources for natural product dereplication

Chapter Contents

Chapter Summary **28**
3.1. Introduction **29**
3.2. Overview of natural products compound identification **31**
3.3. Databases **33**
3.3.1. General databases (Table 3.2): 33
3.3.2. Natural products-specific databases (Table 3.3): 37
3.4. Methods and Software **37**
3.4.1. Spectral preprocessing 43
3.4.1.1. File format conversion: 43
3.4.1.2. Baseline correction: 44
3.4.1.3. Alignment: 44
3.4.1.4. Software summary for spectral preprocessing: 45
3.4.2. Compound identification 45
3.4.2.1. Data reduction: 45
3.4.2.2. Spectral comparison: 46
3.4.2.3. Searching databases 48
3.4.2.4. Software summary for compound identification: 49
3.5. Future Perspectives **50**
3.5.1. Enriching databases using automated machine leaning methods: 50
3.5.2. Developing software suite from building blocks: 50
3.5.3. Integrating different spectral types: 51
3.5.4. Sorting databases for efficient search: 51

Chapter Summary

Research in natural products has always enhanced drug discovery by providing new and unique chemical compounds. However, recently, drug discovery from natural products is slowed down by the increasing chance of re-isolating known compounds. Rapid identification of previously isolated compounds in an automated manner, called dereplication, steers researchers toward novel findings, thereby reducing the time and effort for identifying new drug leads.

Dereplication identifies compounds by comparing processed experimental data to those of known compounds, and so diverse computational resources such as databases and tools to process and compare compound data are necessary. Automating the dereplication process through the integration of computational resources has always been an aspired goal of natural product researchers. To increase the utilization of current computational resources for natural products, this chapter first provides an overview of the dereplication process, and then lists useful resources, categorizing into databases, methods and software tools and further explaining them from a dereplication perspective. Finally, the chapter concludes by discussing the current challenges to automating dereplication and proposed solutions.

3.1. Introduction

Natural products have been a precious resource for drug discovery and lead identification (*55-57*). 75% of all FDA approved small molecules are either natural compounds or derivatives therefrom (*11*). The potential of natural products in drug discovery can be attributed to their unique structural scaffolds and high complexity, creating diverse biological screening libraries (*58*). Besides being attractive drug leads, the complexity of natural products and high content of stereogenic atoms increase protein binding selectivity (*59*), allowing natural products to be used in ligand design, particularly fragment-based drug design (*60*).

Despite the potential of natural products, there are two main factors that limit their role in recent drug discovery and lead identification research: i) time-consuming identification of active compounds: The general manner of experimental design for identifying natural products remained unchanged throughout the past decades. That is, it requires time-consuming purification and inefficient manual interpretation of compound NMR spectra by experts. ii) Repetitive effort for identifying known compounds. While it is estimated that more than 250,000 natural compounds have already been isolated (*61, 62*), incorporation of such knowledge to enhance drug discovery is still not fully exploited.

To overcome these two factors, one promising approach is dereplication, which is the early identification of known compounds without time-consuming manual structure elucidation (*63, 64*). Putative compounds are obtained by comparing preliminary spectral data to spectral databases of known compounds (This review mainly focuses on NMR spectra, while methods, software and databases of NMR spectra can be applied to other types of spectra, such as mass spectrometry). Early detection of known compounds and their reported and potential biological activities help researchers to focus their efforts toward novel findings (*65*). While the idea of dereplication is decades old (*66*), it has gained more attention recently with the increased sensitivity in analytical instruments (*64*), which allows structure elucidation at nanomole scales (*67-69*). In addition, coupling of ultrasensitive instrument such as capillary NMR and high-resolution MS with chromatography allows pre-isolation compound identification (*70-72*), which significantly reduces time and effort.

Despite instrumental advances that are useful for compound identification, computational tools for dereplication are still at a developing stage. Fortunately, natural products and metabolomics share common compound identification techniques, and they are said to be "two sides of the same coin" (*73*). Focusing on detecting dynamic metabolite changes in biological fluids, research in metabolomics spurred simultaneous development of accurate computational methods for fast and high throughput identification of compounds from complex biological mixtures. However, the small but significant differences between natural products and metabolomics prevent the direct cross-utilization of computational resources.

Table 3.1 shows the differences between compound identification in natural products and metabolomics. From data perspectives, there are particularly three key differences: 1) Natural products reference libraries are larger in size than those of metabolomics, increasing the computational demand to search through these libraries, and the lower quality of spectral data poses concern on the reliability of results. 2) Compound identification in metabolomics relies on "landmark" peak detection (*74*), often obtainable from proton-based NMR

spectra such as 1H and TOCSY (*75, 76*). However, due to structural diversity and spectral complexity of natural products, the identification of natural products often requires inclusion of carbon-based NMR measurements, such as 13C and HSQC spectra (*73, 77, 78*). 3) Metabolomics samples are complex biological mixtures where the goal is to both identify and quantify metabolites. However, quantitative analysis of mixtures is not the current focus of dereplication.

I review the current status of computational resources that are or could be used as building blocks to automate dereplication and how they can fit in the current experimental design. I discuss the overlaps and differences in computational demands of dereplication and compound identification in metabolomics. I start by a brief overview of the experimental design of dereplication, followed by detailed discussion on three computational aspects of dereplication: databases, methods and software. I finally conclude with future perspectives.

Table 3.1 Differences between compound identification in natural products research and metabolomics.

	Natural products research	Metabolomics
Reference library size	Large (>250,000) (*61*)	Small (few 1,000s) (*79*)
Quality of reference spectra	Low (*73*)	High (*73*)
Types of spectra	Both proton & carbon-based (*77, 78*)	Mainly proton-based (*80*)
Structural complexity	Complex (*81, 82*)	Simple
Sample purity	Purified or semi-purified compounds (*73*)	Complex biological fluid mixtures (*73, 80*)
Spectral comparison	Pairwise	Pairwise or multiple (time-series)
Overall goal	Compound identification	Compound identification and quantification

3.2. Overview of natural products compound identification

Figure 3.1 shows compound identification in natural products without and with dereplication. The standard experimental design for natural product identification starts with purification of bioactive compounds using bioassay-guided fractionation from natural extracts (Figure 3.1 Ia, Ib). Measured full

spectral data of the purified compounds are manually interpreted for deducing the compound structure (Figure 3.1 Ic, Id), which is then used for literature inquiry (Figure 3.1 Ie). With the increasing chance of isolating known compounds, the time and cost are becoming unacceptable. Dereplication utilizes prior knowledge of previously isolated compounds for early identification to minimize human intervention. Ideally, preliminary experimental data, such as source organism, bioactivity and measured spectra, are used to filter compounds that are either previously reported or lacking drug-like characteristics.

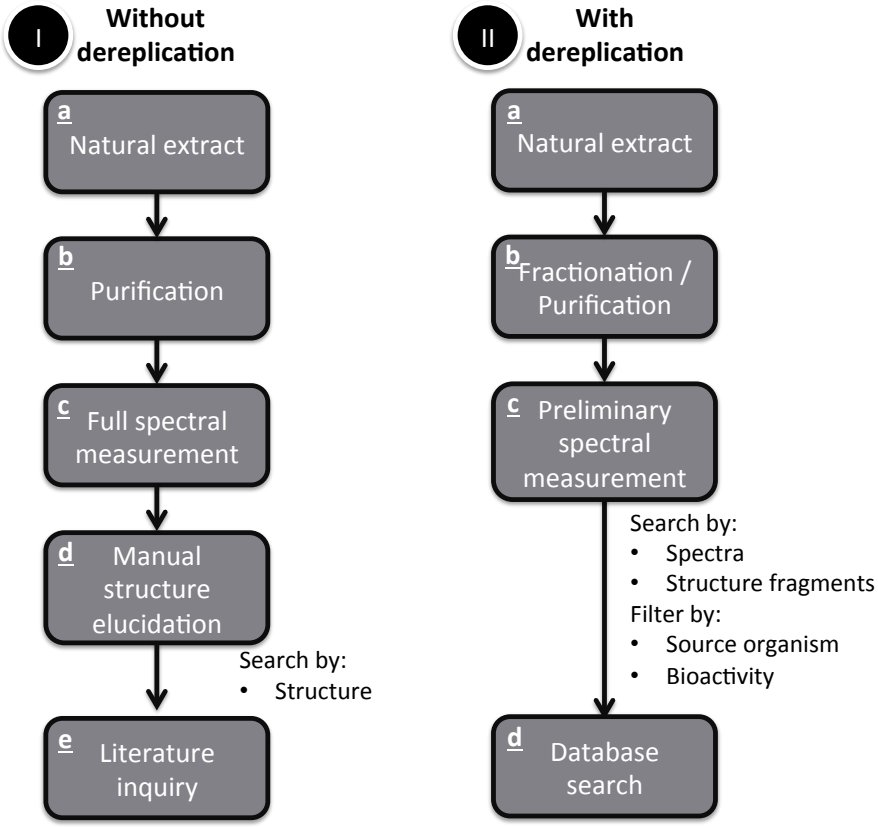

Figure 3.1 Compound identification in natural products without and with dereplication.

For researchers to integrate dereplication in their experimental design, they need a full software suite for automatic NMR processing and analysis that is linked to a reference database for dereplication. The reference database should provide a wide coverage of previously isolated natural compounds with their source organisms and reported / predicted bioactivities. A database query

should be carried out with a sophisticated method for compound matching integrating different types of spectral information.

Three components are needed to develop a complete dereplication software: i) databases to act as reference libraries, ii) spectral processing and searching methods to query databases and iii) software tools for spectral preprocessing and analysis. I discuss each component, identifying available resources and their current shortcomings where further research is needed. In Section III, I introduce available databases, discussing their coverage, deposited data, and relevant query options. In Section IV, I describe different methods as well as software tools for spectral preprocessing and compound identification.

3.3. Databases

The integration of chemoinformtics modeling in drug design motivated the development of numerous databases listing chemical compounds with their biological and physical properties. Databases relevant to natural products are already reviewed (*83-85*), while I discuss them here from a dereplication perspective. I divide available databases into general and natural product-specific databases (Table 3.2 and Table 3.3, respectively), and score each database with seven criteria that are important for dereplication: 1) Coverage of known natural compounds, 2) Availability of bioactivity data, 3) Availability of source organism data, 4) Searchability over compounds by measured compound spectra, 5) Programmatic access through web services or application programming interfaces (APIs), 6) Free availability to use, and 7) Free availability to download. Tables 2 and 3 demonstrate that no available databases satisfy all seven criteria for an ideal dereplication database. Below, I discuss these databases in terms of coverage, data content, spectral searchability and access.

3.3.1. General databases (Table 3.2):

I include fifteen chemical databases as general databases according to the following criteria: i) Cover more than 10% of already isolated natural products; around 20,000 compounds. ii) Contain at least 40,000 entries including both

synthetic and natural compounds. iii) Contain information useful in dereplication, such as bioactivity, source organism or spectra.

Regarding coverage, five databases contain more than 10 million entries. General databases provide wide coverage of natural compounds, with eleven databases containing more than 20,000 natural compounds (roughly 10% of already isolated compounds). Despite their wide coverage, searching is not easy to use for dereplication because synthetic compounds are among search candidates. Seven databases have natural compounds annotation, which allows users to limit their search to natural products only.

Dereplication relevant data-contents are two: bioactivity and source organism. Eleven databases include biological activity. PubChem (*86*), ChEBML (*87*) and BindingDB (*88*) databases contain detailed bioactivity information such as biological mechanism and protein targets, which can be used, in conjunction with spectral information, to enhance compound identification (*89*). Regarding source organism, only two databases, ChEBI (*90*) and Reaxys (*91*), contain this information.

While spectral searchability is important in dereplication, searching compounds by spectral data is not the focus of general databases, and only NMRShiftDB (*92*), CSEARCH (*93*) and SpecInfo (*94*) have this ability. Compounds in all fifteen general databases are searchable by similarity of structures or substructures; however this search has strong limitations for dereplication, where molecular structures are unknown.

There are three ways to access general databases: 1) manual access, 2) access via database download or 3) programmatic access. Twelve databases can be accessed manually for free and nine of them are freely downloadable. Ten databases provide APIs to access the data though programs, which enable their integration to user-customized analysis flows. However, programmatic access has limitations for dereplication because either necessary data or query options are lacking.

Table 3.2 General chemical databases.

Database	Website (*http://*)	Coverage		Data Content				Free?		Score
		# NPs	# Compounds	Bioactivity (type)	Source Organism	Spectral Searchability	Programmatic Access	Use	Download	
BindingDB (*88*)	*www.bindingdb.org*	NA	>450k	•(protein binding)			•	•	•	5
ChEBI (*90*)	*www.ebi.ac.uk/chebi/*	>25k	>42k	•(all)	•		•	•	•	5
ChemBank (*95*)	*chembank.broadinstitute.org*	NA	>800k	•(all)			•	•	•	5
Chembl (*87*)	*www.ebi.ac.uk/chembl/*	24K	>600k	•(all)			•	•	•	5
ChemIDplus	*chem.sis.nlm.nih.gov/chemidplus/*	>9k	>400k	•(all)				•		2
ChemSpider (*96*)	*www.chemspider.com*	>660K	>14M	•(all)			•	•	•	5
CSEARCH (*93*)	*nmrpredict.orc.univie.ac.at/*	NA	>450k			•		•		3
NCI	*cactus.nci.nih.gov/ncidb2.2/*	NA	>250k	•			•	•	•	5
NIAID ChemDB	*chemdb.niaid.nih.gov*	>9k	>130k	•(allergy, infectious diseases)				•		2
NMRShiftDB (*92*)	*nmrshiftdb.nmr.uni-koeln.de*	NA	>42k			•		•	•	4
PubChem (*86*)	*pubchem.ncbi.nlm.nih.gov*	NA	>30M	•(all)			•	•	•	5
Reaxys (*91*)	*www.reaxys.com/reaxys*	>200k	>10M	•(all)	•		•			4
SciFinder	*scifinder.cas.org*	NA	>90M	•(all)			•			3
SpecInfo (*94*)	*www.wiley-vch.de/stmdata/specinfo.php*	3.5k	>500k			•				1
ZINC (*97*)	*zinc.docking.org*	>180k	>20M					•	•	3

#NPs: Number of natural product compounds.

Table 3.3 Natural products-specific databases.

Database	Website (*http://*)	Coverage	Data Content		Spectral Searchability	Programmatic Access	Free?		Score
		# Compounds	Bioactivity	Source Organism			Use	Download	
AntiBase (*98*)	*www.wiley-vch.de/stmdata/antibase.php*	>40k	•(all)	•	•				4
BACTIBASE (*99*)	*bactibase.pfba-lab-tun.org*	220	•(all)	•			•	•	4
CamMedNP (*100*)	*NA*	2.5k		•			•	•	3
ConMedNP (*101*)	*NA*	3.2k		•			•	•	3
Dictionary of marine NP	*dmnp.chemnetbase.com*	>30k	•(all)	•					3
Dictionary of NP	*dnp.chemnetbase.com*	>250k	•(all)	•					3
HeteroCycles	*www.heterocycles.jp/newlibrary/natural_products/ structure*	>58k	O(anti-microbial)	•			•		4
Marinlit	*pubs.rsc.org/marinlit/*	>24k	•(all)	•	•				4
NAPROC-13 (*102*)	*c13.usal.es*	>20k			•		•		3
NPACT (*103*)	*crdd.osdd.net/raghava/npact/*	1574	•(anti-cancer)				•		2
NuBBE (*104*)	*nubbe.iq.unesp.br/portal/nubbedb.html*	640	•(anti-microbial)	•			•	•	4
PhytAMP (*105*)	*phytamp.pfba-lab-tun.org*	273	•(anti-microbial)	•			•	•	4
SuperNatural (*106, 107*)	*bioinformatics.charite.de/supernatural*	>350k	•(all)				•		3
TCM database (*108*)	*tcm.cmu.edu.tw*	>20k	•(traditional Chinese medicine)	•			•	•	5
UDNP (*109*)	*pkuxxj.pku.edu.cn/UNPD*	230k		•			•	•	4

O: Limited data.

3.3.2. Natural products-specific databases (Table 3.3):

I raise fifteen databases that catalogue molecules isolated from natural origins only, excluding those limited to primary metabolites, as those are relevant only to metabolomics. In terms of coverage, nine specific databases exceed 20,000 entries. Because of the coverage limitation, it is better to use multiple specific databases for reliable dereplication. Some specific databases have limited coverage because they focus on: i) particular compound features such as compound class (PhytAMP (*105*) and BACTIBASE (*99*)) or bioactivity (NPACT (*103*)) and ii) particular compound origins such as compounds from a particular family of source organisms (CamMedNP(*100*) and ConMedNP (*101*)) or geographic location (NuBBE (*104*) and TCM (*108*)).

Despite their limited coverage, specific databases contain bioactivity and source organism information, useful in dereplication. Eleven specific databases contain bioactivity data. Typical examples are NuBBE (*104*) and NPACT (*103*), which provide effective compound concentrations of different bioactivities for each entry. All specific databases have source organism information, except for SuperNatural (*106, 107*), NPACT (*103*) and NAPROC-13 (*102*).

Spectral searchability is limited in specific databases because of the scarcity of spectral data. Only three databases have spectral searchability, and only one database, NAPROC-13 (*102*), is freely accessible but limited to 13C spectra only.

Regarding database access, eleven specific databases can be manually searched, seven of which are freely downloadable. Specific databases are usually in-house developed and all of them do not provide programmatic access to the data, limiting automatic search and integration to other software.

3.4. Methods and Software

This section describes computational methods and software tools used as parts of natural product dereplication process. Table 3.4 summarizes two main steps of dereplication: spectral preprocessing, and compound identification. First, spectral preprocessing involves reformatting and denoising of the acquired spectra to alleviate the instrumental and experimental discrepancies (*80, 110*).

Second, compound identification uses preprocessed spectra and compares them to a reference database. To realize automatic and fast dereplication, each step needs to be carried out efficiently with minimal human intervention. Table 3.5 lists, to the best of our knowledge, currently available software tools for these steps, comparing the tools according to functionalities.

Note that while I focus here on software for spectral preprocessing and compound identification, natural product dereplication needs additional tools to manage and visualize chemical structures and spectra. For example, structures of chemical compounds are usually represented as SDF or MOL files, and software tools, such as Open Babel toolbox (*111*) and ChemmineR (*112*), rcdk (*113*) and Rcpi (*114*) R packages, are needed to handle these files and pass the data to the dereplication software for processing or visualization. For result visualization, Java and JavaScript libraries, such as JSpecView (*115*), JSME (*116*), MarvinJS (*117*), can offer in-browser chemical structure and spectral visualization for web applications.

Table 3.4 Analysis flow of spectra from acquisition to compound identification.

Spectral Preprocessing	**File Format Conversion**	• JCAMP-DX • NMRPipe • Sparky
	Baseline Correction	1. Baseline recognition • Derivative functions • Wavelet-based 2. Baseline modeling • Polynomial • Regression • Smoothing 3. Baseline subtraction
	Alignment	• FFT alignment • Multiple-dimension
Compound Identification	**Data Reduction**	❖ Peak lists • Peak picking ⭘ Numerical vectors • Binning • Feature Extraction o Sliding window o PCA ➢ Trees
	Spectral Comparison	❖ Peak lists • Tanimoto coefficient • Jaccard similarity ⭘ Numerical vectors • Correlation-based o Dot product o Pearson's correlation o Spearman's correlation o Weighted cross-correlation o Partial and semi-partial correlation • Distance-based o Absolute value distance o Euclidean distance ➢ Trees • Tree-based comparison
	Database Search	• Identity search • Ranking search • Interpretative search

Table 3.5 Software tools with a potential role in dereplication.

				Spectral Preprocessing						Compound Identification		
Software	**Software Type**	**Spectra Type**	**GUI**	**File Format Conversion**	**Baseline Correction**	**Alignment**	**Peak Picking**	**Binning**	**Feature Extraction**	**Free?**	**Score**
ACD Labs	Desktop	NMR (1D,2D), MS	•	•	•	•	•				4
Automics (*118*)	Desktop	NMR	•	•	•	•	•	•	•	•	7
BATMAN (*119*)	R package	NMR			•	•		•		•	4
ChemoSpec(*120*)	R package	Any			•		•			•	3
Chenomx NMR suite	Desktop	NMR (1D,2D)	•	•	•	•	•	•			5
cuteNMR	Desktop	NMR	•	•	•		•			•	4
MestreNova	Desktop	NMR (1D,2D), MS	•	•	•	•	•				4
mSPA (*121*)	R package	Any				•				•	2
MVAPACK(*122*)	Octave package	NMR (1D,2D)		•	•	•	•	•	•	•	7
mylims.org (*123*)	Web	NMR, MS		•	•		•			•	4
Nmrglue (*124*)	Python package	NMR (1D,2D)		•	•		•			•	4
Nmrpipe (*125*)	Desktop	NMR		•	•		•	•		•	5
NMRS (*126*)	R package	NMR		•						•	2
PERCH	Desktop	NMR (1D,2D)	•	•	•	•	•				4
rnmr (*127*)	R package	NMR (2D)	•	•	•		•	•		•	5
speaq (*128*)	R package	NMR				•	•			•	2

GUI: Graphical User Interface.

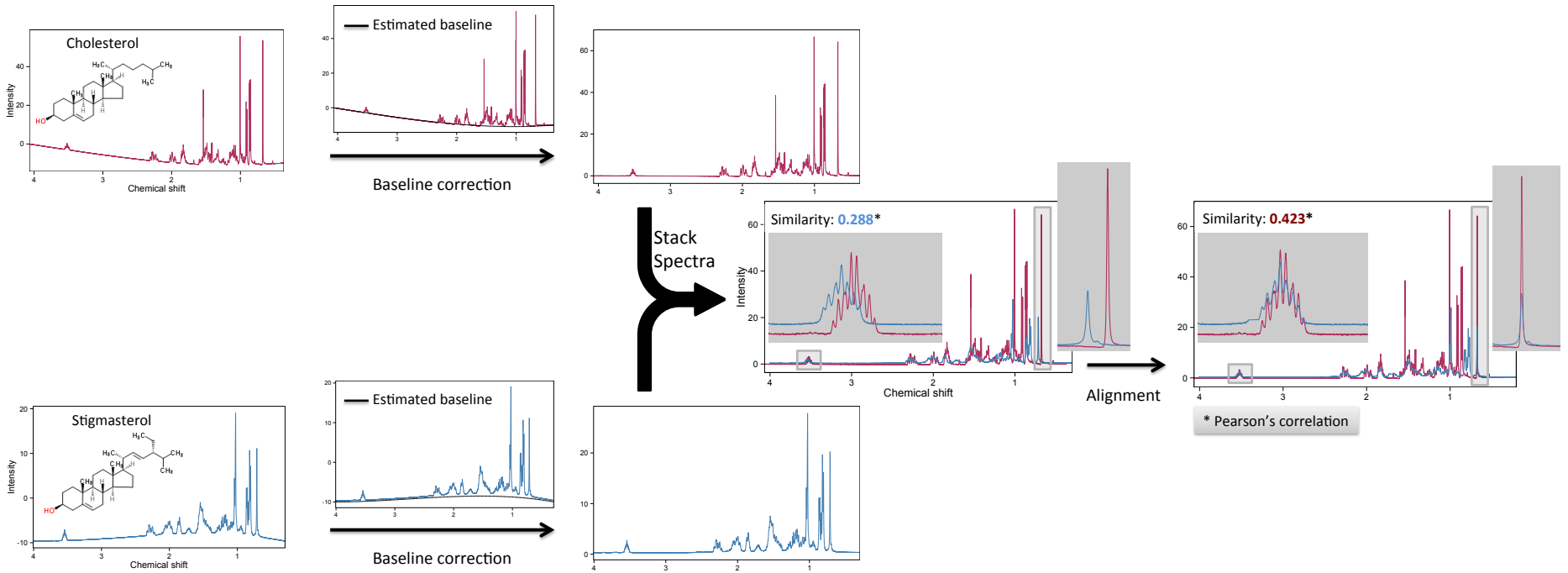

Figure 3.2 Spectral preprocessing. ^{1}H NMR spectra of cholesterol and stigmasterol, two common and structurally similar natural compounds, are used for demonstration. The raw NMR files were downloaded from HMDB (*79*) and converted to JCAMP-format DX using Mestrenova. Baseline estimation was performed in R using 3rd order polynomial fitting. The baseline-corrected spectra were stacked, and then aligned using Mestrenova, showing higher similarity (Peasron's correlation of 0.423) than before alignment (0.288).

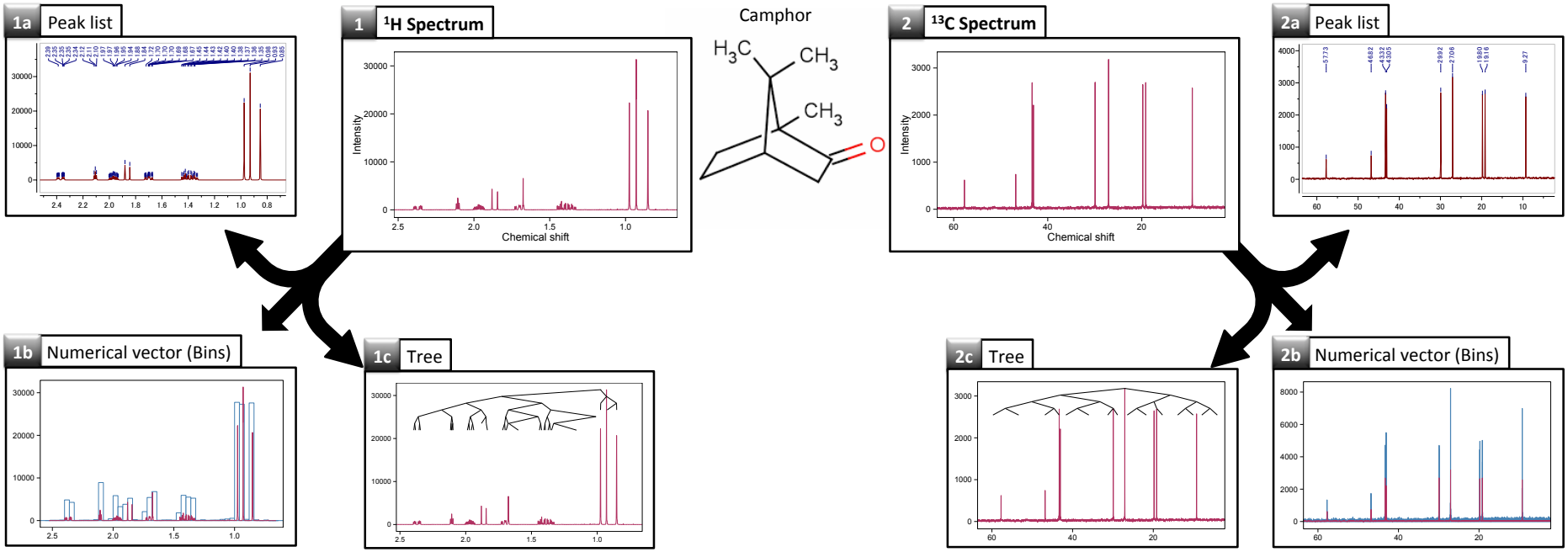

Figure 3.3 Data reduction of spectra. ^{1}H and ^{13}C NMR spectra of camphor, a natural compound, demonstrate the effect of each data reduction method on different types of spectra. Peak picking reduces the ^{13}C spectrum to a few peaks (2a), but fails with the ^{1}H spectrum (1a) as resonance coupling generates numerous overlapping multiplet peaks. Binning produces in a large vector (1532 bin) in the ^{13}C spectrum and a small one in the ^{1}H spectrum (47 bins). Both spectra are reduced to relatively few nodes when represented as trees.

3.4.1. Spectral preprocessing

I categorize preprocessing methods into three main steps: file format conversion, baseline correction and alignment. I first discuss each step, and demonstrate baseline correction and alignment on example spectra (1H NMR spectrum of stigmasterol in Figure 3.2). I finally summarize available software tools.

3.4.1.1. File format conversion:

While the acquired spectra are initially stored as proprietary data formats that are specific to each instrument, converting them to a common instrument-independent format ensures easier data exchange and wider compatibility. For NMR spectra, JCAMP-DX (*129*), NMRPipe (*125*) and Sparky (*130*) are among the most used file formats for describing spectral information of small molecules. JCAMP-DX (*129*) provides a simple and human-readable format, and allows additional labels to describe experimental conditions and parameters. However, representation of multi-dimensional NMR spectra in JCAMP-DX is not standardized. NMRPipe (*125*) and Sparky (*130*) have been used in web applications (*131, 132*) for their strong standardization and the ability to represent multi-dimensional NMR spectra.

Current NMR file formats mainly have the following three limitations for dereplication. First, current file formats do not contain structures of measured compounds, which prevent assigning spectral peaks to corresponding atoms. I have to include additional files for compound structure and peak assignment information (*133, 134*), which cannot be linked easily with spectral files. Second, 1D NMR and 2D NMR spectral data of the same sample cannot be linked with each other in current file formats. Third, current file formats are still insufficient to fully represent measurements and experimental parameters in high throughput studies (*133*). CCPN (*135, 136*) and STAR (*137-139*) provide different formats that can be used for high throughput studies, but are tailored for protein NMR experiments. A suitable file format for natural product dereplication is still needed to overcome the above three limitations.

3.4.1.2. Baseline correction:

Removal of baseline drifting is crucial to remove noise and artifacts resulting from different measurement conditions. Generally, baseline correction has three steps, baseline recognition, modeling and subtraction. First, baseline recognition distinguishes peak regions from baseline points, exploiting the fact that peak regions have higher variation in intensity. Higher variation regions are detected using spectrum derivatives (*140*) or wavelet transformation (*141-144*). Second, baseline modeling estimates a curve based on baseline points, by linear interpolation or non-linear approximations like polynomial fitting (*145, 146*), LOcally Weighted Scatterplot Smoothing (LOWESS) and quantile regressions (*147-151*) and Whittaker smoother (*141, 152*). Finally, in baseline subtraction, the estimated baseline curve is subtracted from the spectrum, leaving only the peak signals.

In natural product dereplication, baseline correction is a minor step compared to metabolomics because of two main differences: i) Since dereplication currently focuses on compound identification rather than quantification, accurate baseline estimation is less significant (*73*). ii) Dereplication is usually performed on purified compounds where spectra are less crowded than those of biological mixtures. Therefore, simple polynomial fitting is usually preferred for baseline correction, instead of more computationally demanding techniques such as LOWESS and quantile regressions and Whittaker smoother. In our example, the baseline is estimated as a third-order polynomial function (Figure 3.2).

3.4.1.3. Alignment:

Alignment of spectra is a process to alleviate the effect of experimental conditions on peak positions by shifting data points to match a reference spectrum (*80, 110*). Spectral alignment and relevant software tools are already reviewed in detail (*110*), and so I only describe alignment here briefly. Alignment is performed for quantitative comparison between multiple spectra of different samples that have similar chemical compositions, and therefore it is a standard manner for time-series NMR spectra in metabolomics. Using the same concept, alignment can be applied in dereplication when spectra for different fractions of the same extract are compared (*153*). Figure 3.2 shows how alignment removes

subtle chemical shift differences in the spectra of two structurally similar compounds, cholesterol and stigmasterol, increasing the overall similarity between the two spectra.

3.4.1.4. Software summary for spectral preprocessing:

Table 3.5 shows that out of sixteen currently available software, three steps in spectral preprocessing, i.e. file format conversion, baseline correction and alignment are implemented in twelve, thirteen and nine software tools, respectively, meaning that baseline correction is the most implemented. Six software tools, ACD Labs, Automics (*118*), Chenomx NMR suite, MestreNova, MVAPACK (*122*) and PERCH, implement all three steps, of which Automics and MVAPACK are freely available, making them most useful for spectral preprocessing. Six other tools implement two steps, and the remaining four tools (all are R packages) specialize in only one step.

3.4.2. Compound identification

For compound identification, preprocessed spectra are converted into different representations to be compared against reference spectra in a computationally efficient manner, to find compounds with the highest spectral similarity. In order to carry out compound identification, three steps are required: data reduction, spectral comparison and searching databases. I explain each of these three below.

3.4.2.1. Data reduction:

Spectral comparison of raw spectra needs long computation time because each spectrum has a large number of data points (more than 20,000 points for 1H NMR (*80*)), where each point has a position (chemical shift) and an intensity. To reduce computation time, I need methods to reduce data size without substantial loss of information. Data reduction transforms spectral data into peak lists, numerical vectors or trees. I describe the characteristics of each of these three representations.

3.4.2.1.1. Peak lists:

Spectra are reduced to peak lists by peak picking (*154-156*), which greatly simplifies the spectra to a handful of peak positions and their intensities. Limitations of peak picking arise if the spectrum contains broad or overlapped

peaks, such as crowded 1H NMR spectra, in which important peaks can be missed.

3.4.2.1.2. Numerical vectors:

Spectra can be reduced to numerical vectors of the same size by binning, sliding window or principal component analysis. First, binning (*157-159*) divides the spectrum into intervals and the total intensity in each interval is extracted. While binning keeps representative information about the spectrum, a peak may be split into two bins if the bin boundary lies on a peak center, which misrepresents the peak as shown in (Figure 3.3-2a). So, adaptive binning changes bin boundaries to prevent overlap with peak centers (*157, 159*). Second, sliding window divides the spectrum into fix-sized but overlapped intervals (*160*). Third, principal component analysis reduces spectra by transforming the original data space into a lower dimension space (*161*).

3.4.2.1.3. Trees:

A spectrum is transformed into a tree by assigning peaks to end nodes through recursively dividing the spectrum into subspectra at mass centers (*162, 163*) (Figure 3.3-3a,b). The resulting tree has spectra mass centers as branching nodes and peaks as end (leaf) nodes, which retains information about peak positions and as well as their hierarchy.

Two factors are important in data reduction for natural product dereplication: i) Type of measured spectra: NMR spectra vary in how sharp peaks are and the propensity for peaks to overlap. Sharp peaks in 13C NMR spectra are unlikely to overlap and so peak lists are suitable. In contrast, 1H NMR peaks tend to heavily overlap, especially that of complex mixtures and in condensed methylene regions, and so binning or trees are preferred. ii) Spectral comparison measure suitable for the representation (described in the next section).

3.4.2.2. Spectral comparison:

Spectra are compared using a similarity measure that reflects the structure similarity of the corresponding compounds. The choice of the similarity measure depends on the data representation, determined by the data reduction method

(described in the previous section). I discuss available similarity measures for each representation.

3.4.2.2.1. Peak lists:

When spectra are reduced to peak lists, which are of different sizes, they are represented as sets. Comparing two sets of peaks requires two steps: 1) Matching of set members, to produce one-to-one mappings between peaks of the query and reference sets. First, a list of matching candidates for each peak is narrowed to peaks whose positions lie within a defined threshold. A threshold can be either a fixed window (hard thresholding), which is chosen manually, or defined statistically using Bayesian (*164, 165*) and probability-based (*166, 167*) models (soft thresholding), which are more flexible. Second, matching peaks are chosen from the candidate list by either i) selecting the nearest peak, or ii) maximum bipartite matching (*168*), which maximizes the number of pairs between peaks of the two sets. 2) Measuring the overlap between two sets, which is computed by set similarity measures, typically Jaccard's similarity and Tanimoto's coefficient (*169-171*).

3.4.2.2.2. Numerical vectors:

Numerical vectors have the same dimension and a typical way to compare them uses a correlation or distance-based similarity measure such as inner product (*172-174*), Euclidean distance (*175-177*), difference in absolute value (*178, 179*). Among the three measures, inner product was reported to outperform the other two measures (*180*). Measures combining both correlation and distance-based similarities, such as partial correlation (*181*) and composite similarity measures (*182*), have been shown to perform better than a single measure (*183, 184*).

3.4.2.2.3. Trees:

Trees are compared by taking into account both peak positions (node position) and their hierarchy (children nodes) (*162, 163*).

3.4.2.2.4. Computational efficiency:

Applying spectral comparison to large databases requires efficient computation of similarity scores. The speed of computing similarity scores for each representation is affected by two factors: 1) Number of data points in spectral

representation and 2) Computational complexity of comparing two spectra. First, number of data points varies between spectra due to a) different spectral features or data reduction parameters. The number of data points in peak lists and trees depends on the number of peaks, while in numerical vectors, it is equal to the number of bins. The typical size of a natural product compound spectra is tens for peak lists, and 250 for numerical vectors (chemical shift range: 0~10ppm, bin size: 0.04ppm). Second, the computational complexity is determined by the number of computational operations for comparing two spectra of N data points, which is in the order of N^2 for peak lists, and N and $N\ logN$ for numerical vectors and trees, respectively. Theoretically, for a similar number of data points, numerical vectors are the fastest to compare, followed by trees and peak lists.

3.4.2.3. Searching databases

Three database search paradigms are useful in dereplication: 1) identity, 2) ranking and 3) interpretative; each search paradigm produces a different output format (*180, 185*). I explain each paradigm below.

3.4.2.3.1. Identity search:

Identity search returns a single compound with a spectrum that is equivalent to the query spectrum (*164, 180*). Identity search requires no manual investigation and so it can be very useful in automating dereplication. However, identity search has two limitations: 1) Searching small-coverage databases may return empty results, if the exact spectrum is not in the database. 2) Setting strict equivalence criteria may miss spectra that are affected by variations in experimental conditions or inadequate preprocessing.

3.4.2.3.2. Ranking search:

Ranking search returns a ranked list of compounds with spectra closest to that of the query by computing similarity scores the query spectrum and all spectra in the database (*178*). By investigating common substructures of highly similar compounds in the list, I can deduce chemical class or functional groups of the query compound. Similarity scores can also be computed using a subset of the query spectrum, allowing users to focus on distinctive peaks. One limitation for

ranking search is that deducing chemical classes and functional groups still requires manual investigation, which hampers automatic dereplication.

3.4.2.3.3. Interpretative search:

Interpretative search returns a list of matching fragments by assigning peaks from the query spectrum to connected fragments of reference compounds (*168, 186, 187*). The output fragments, which belong to different reference compounds, can then be combined to deduce the query compound structure and so interpretative search can identify novel compounds that are not included in the reference database. Currently, interpretative search is not applicable to 1H spectra because of the sensitivity of chemical shifts to spatial interactions (*168*) and because peak overlap prevents spectral peaks to be assigned to corresponding atoms.

3.4.2.4. Software summary for compound identification:

For data reduction, I focused on three spectral representations: i) peak lists obtained by peak picking ii) numerical vectors obtained by binning and feature extraction and iii) trees. Table 3.5 shows that peak picking is the most implemented method, available in thirteen out of sixteen software tools, followed by binning and then feature extraction, available in six and two tools, respectively. No software tools implement tree representation of spectra, however, the pseudocode is available (*162*). Automics (*118*) and MVAPACK (*122*) are the only tools implementing the three data reduction methods. rNMR (*127*), NMRPipe (*125*) and PERCH have both peak picking and binning functionalities.

Spectral comparison methods, such as inner product and partial correlation, are available in statistical software frameworks, such as R and Matlab.

Finally, among database search paradigms, ranking search is implemented in spectral databases, such as NMRShiftDB (*92*) and CSEARCH (*93*), because chemical class or functional groups can be deduced by investigating the ranked compound list.

3.5. Future Perspectives

Despite the abundance of computational resources that are useful for dereplication, I need to overcome several challenges to realize the aspired automation. I discuss four proposed solutions to existing challenges that can enhance the speed and quality of natural products dereplication results.

3.5.1. Enriching databases using automated machine leaning methods:

The deficiency of necessary data, namely measured spectra and source organisms, presents a challenge to the development of a dereplication database, being summarized into two points: 1) The scarcity of measured spectra prevents spectral searchability from producing reliable results. 2) The absence of source organisms data prevents their use to limit dereplication candidates. Two machine learning-derived approaches will provide a fast and automated way to add data to databases and complete missing data: spectral prediction and literature text mining. First, compound spectra can be predicted from existing spectra on the basis of compound structural similarity (*188*). Several machine learning algorithms have been proposed to predict NMR spectra (*189-191*), of which prediction accuracy increases with training data size (*192*). Similar algorithms have also been developed for other types of spectra, such as fragmentation pattern in MS spectra (*193-195*), ultraviolet spectra (UV) (*196*) and chromatographic retention index (*197-199*). Comparison and accuracy assessment of NMR prediction algorithms are reviewed in (*200-202*). Second, Text mining of chemical information (*203, 204*) can automatically extract compound associated data such as NMR assignments and source organisms from the literature.

3.5.2. Developing software suite from building blocks:

The wide use and integration of dereplication to current experimental design is hampered by the unavailability of open-source software to process NMR spectra, to link and to summarize information across all submitted spectra. While all steps for dereplication are implemented in software packages (**Table 5**), the dereplication process requires the use of different tools and familiarity of programming languages. To accelerate dereplication, a software suite combining

available software packages through a unified graphical interface that can be used intuitively by experimental researchers on natural products is needed.

3.5.3. Integrating different spectral types:

Relying on NMR data only for compound identification becomes insufficient as molecular complexity (*205, 206*) increases, as exemplified by fatty acids and peptides. The integration of different spectral data into dereplication can resolve structural ambiguities in these chemical classes. Several studies in dereplication showed promising results by integrating MS fragmentation with UV spectra (*207, 208*), and in combination with NMR spectra (*209, 210*). However, current studies have two limitations: 1) Other spectral types, such as chromatographic retention times, can differentiate between compounds that are otherwise similar. While these spectra utilized in metabolomics (*121, 181, 211*), they are not yet incorporated in dereplication. 2) Similarity scores between query and database compounds are calculated based on only one spectral type, and candidate structures are then filtered using the other spectra. Calculating similarity scores based on all available spectra is still lacking.

3.5.4. Sorting databases for efficient search:

Calculating similarity scores between a query spectrum and a database containing hundreds of thousands of spectra can computationally intensive. Classifying database compounds using molecular characteristics such as complexity (*206, 212*), common substructures (*70, 213*) have proved useful in efficient compound identification (*70, 71*) and mining of chemical databases (*214*). Applying similar strategies to spectral databases presents promising possibilities.

Chapter 4

NMRPro: An integrated web component for interactive processing and visualization of NMR spectra

Chapter Contents

Chapter Summary **52**
4.1. Introduction **53**
4.2. Web applications as medium for scientific development **54**
4.2.1. Current status of web applications for NMR data 55
4.3. Software architecture of NMRPro **57**
4.3.1. Challenges for developing web application for NMR 57
4.3.2. Design considerations for NMRPro 57
4.3.2.1. Client-side interactivity 58
4.3.2.2. Efficient transfer of data 58
4.3.2.3. Smooth display of multiple spectra 58
4.3.2.4. High extensibility for both server-side and client-side components 58
4.3.2.5. Using SpecdrawJS as standalone library 58
4.3.2.6. Integration into existing web applications 58
4.4. Subcomponents of NMRPro **59**
4.4.1. Python Package 59
4.4.2. Django App 60
4.4.3. SpecdrawJS 61
4.5. Availability and Installation **63**
4.6. Conclusion **64**

Chapter Summary

The popularity of using NMR spectroscopy in metabolomics and natural products has driven the development of an array of NMR spectral analysis tools and databases. Particularly, web applications are well used recently because they are platform-independent and easy to extend through reusable web components. Currently available web applications provide the analysis of NMR spectra. However, they still lack the necessary processing and interactive visualization functionalities. To overcome these limitations, I present NMRPro, a web

component that can be easily incorporated into current web applications, enabling easy-to-use online interactive processing and visualization. NMRPro integrates server-side processing with client-side interactive visualization through three parts: a python package to efficiently process large NMR datasets on the server-side, a Django App managing server-client interaction, and SpecdrawJS for client-side interactive visualization.

4.1. Introduction

Nuclear magnetic resonance (NMR) spectroscopy is indispensible for structure identification of chemical compounds, becoming an integral part of metabolomics and natural products studies. In metabolomics, NMR spectroscopy is increasingly used to identify and quantify metabolites present in biological samples (*215, 216*). In natural products, interpretation of NMR spectra allows the structure determination of complex compounds, leading to the discovery of new structural scaffold or potential drug leads (*217*).

As resolution of NMR spectra improves, more detailed information can be extracted allowing advanced quantitative analysis. The utilization of 1D 1H and 13C spectra as well as 2D spectra such as HSQC has enhanced the identification of trace metabolites with increasing sensitivity. Machine learning and pattern recognition techniques such as Principal Component Analysis (PCA) (*218*) and Partial Least Squares Discriminant Analysis (PLS-DA) (*219*) allowed NMR spectra to be used to classify biological samples, identifying otherwise undetectable biomarkers (*80, 220*).

Information in NMR spectra is extracted through multiple processing steps of analysis workflows, being different depending on applications. For example, metabolomics datasets are processed to overcome systematically occurring batch effects, by using techniques such as binning and spectra alignment. These techniques are sometimes not so simple, since chemical structure identification in natural products often requires sophisticated use of software and knowledge of NMR instrumentation. Processing and visualization of NMR spectra and then sharing such spectra for collaboration purposes, without prior expertise in

computer programming is certainly a current demand among experimental researchers (*217*).

This chapter presents NMRPro, an integrated web component for interactive processing and visualization of NMR spectra online. Users can input NMR spectra in raw formats, such as Bruker or NMRPipe, and then select from a wide range of processing and analysis functionalities, including apodization, Fourier transform, baseline and phase correction etc. NMRPro can be integrated into existing web applications and databases, and extended through plugins to suite the different needs for various web applications.

4.2. Web applications as medium for scientific development

The continuous advances in Web technologies offers a platform-independent highly interactive medium for the development of scientific application. In the past two decades, servers for web-based analysis and storage repositories for various scientific data have been developed. Genomic data repositories such as Entrez Gene (*221*), EBI's European Nucleotide Archive (*222*) and DNA Data Bank of Japan (DDJB) (*223*) became essential tools for experimental researchers because of their intuitive user interfaces and the utilization of web servers' capabilities in computationally intensive analyses. Other chemo-genomic applications such as ChEMBL (*87*), and chemical repositories such as PubChem (*86*) are more recent extensions to the toolbox of experimental researchers in various fields.

Recently, the increasing use of JavaScript-based in web applications enabled the development of single-page applications, in which the web site consists of a single web page that is update instantaneously upon user interactions. The use of asynchronous JavaScript and XML (AJAX) technology to inject and update the contents of a web page enhances the overall user experience and allows to create richer visualization (*224-226*). Examples include BrowserGenome.org for the analysis and visualization of RNA-seq data (*227*).

A Web-based NMR analysis application has several advantages over conventional ones: 1) web is a platform-independent highly interactive

environment, which enables closer investigation of spectra through zooming and panning on multiple platforms including handled devices, and 2) existing web applications can be easily extended through integrating 'web components', such as JavaScript libraries and web services, to provide additional functionalities. For example, web application for spectral analysis and metabolite identification (*228, 229*), can add processing functionalities by simply integrating NMRPro. 3) Processing large datasets benefits from the web server computational capabilities, which are much more powerful that personal computers. 4) Easy sharing of raw and processed spectra. 5) Current NMR databases can benefit from the visualization functionality by displaying spectra interactively, instead of static images. 6) The software can be extended to educational purposes such as teaching NMR concepts.

4.2.1. Current status of web applications for NMR data

Online processing and interactive visualization of spectra are necessary functionalities for all NMR web applications (*217*). However, web applications for NMR analysis such as MetaboAnalyst (*228*), MetaboHunter (*229*) and COLMAR (*131*) require NMR spectra to be processed offline beforehand. Also, interactive investigation of NMR spectra in databases such as HMDB (*79*) and BMRB (*230*) requires raw spectra to be downloaded and visualized offline.

Although processing and interactive visualization of NMR spectra are needed for web applications, web components providing these functionalities are still lacking. In fact, previously used Java applet components, such as JSpecView (*115*) and Nemo, suffer from security concerns and require installation of additional software. Also, although the recently developed jsNMR (*231*) and SpeckTackle (*232*) offer JavaScript-based visualization, they have very limited processing functionalities.

I developed NMRPro to overcome the current lack in web-based software for processing NMR spectra, as shown in Table 4.1. Besides being an easy-to-integrate web component, NMRPro exceeds the processing functionalities of currently available web components and applications.

Table 4.1 Comparison of software capabilities with existing web-based applications.

Capabilities	NMRPro	jsNMR	SpeckTackle	MetaboAnalyst	Metabohunter	COLMAR
Description	Web component	Web component	Web component	Web application	Web application	Web application
Interactive visualization	✔	✔	✔	✖	✖	✖
Supported formats	Bruker, JSON, NMRPipe	Bruker, JSON	JSON	Tab-separated Files	Tab-separated Files	NMRPipe
Processing			✖	✖	✖	✖
Zero filling	✔	✖				
Apodization	✔	✖				
Fourier transform	✔	✔				
Phase correction	✔ (auto)	✔ (manual)				
Baseline correction	✔	✖				
Peak picking	✔	✖				

4.3. Software architecture of NMRPro

4.3.1. Challenges for developing web application for NMR

The development of web components for processing NMR spectra is hampered by three challenges caused by the large size of NMR spectra: 1) Processing of large datasets is computationally intensive, requiring server-side integration. 2) A compressed spectral format, required for efficient transfer across the Web, is lacking. 3) Visualization of large number of spectral data points presents a computational load on users' computers. Automatic reduction of data points is needed.

4.3.2. Design considerations for NMRPro

I designed NMRPro, an open-source easy-to-integrate web component for processing and visualization of NMR data, which is highly extensible to include new functionalities according to the needs of each application. NMRPro consists of three integrated parts, 1) Python package with extensible functionality plugins for server-side spectral processing, 2) Django App for spectral compression and managing communication between server- and client-sides, and 3) SpecdrawJS, a JavaScript library for visualization of 1D and 2D NMR datasets.

Table 4.2 Comparison of NMRPro with existing frameworks

	R Shiny	Bokeh	NMRPro
JavaScript lib.	NA	BokehJS	SpecdrawJS (extension of D3.js)
Spectral Compression	No	No	Yes
Data simplification	No	Server-side	Client-side
Programming language extensibility	R only	Python, R	Python, R, JavaScript
Library size [1]	NA	300 Kb	< 100 Kb
Framework	Undisclosed	Flask	Django

[1] Minified and Gzipped size of the library and its dependencies

I used Python-Django-JavaScript design to overcome the current challenges' for processing and visualizing NMR spectra on the web. Table 4.2 summarizes six key differences between NMRPro design and other frameworks. I discuss each one below.

4.3.2.1. Client-side interactivity

To generate interactive plots, NMRPro transfers the spectral data only once at the beginning of the display, instead of resending the data every time the user interacts with the plot. SpecdrawJS generates interactive plots from the data, which are used to update the plot instantaneously upon user interaction. So, sending data to the client-side is more advantageous than keeping them on server side, as done by R Shiny apps, which sends plots as static images.

4.3.2.2. Efficient transfer of data

NMRPro compresses original data for faster transfer from server-side to client-side. While data for line charts are commonly transferred as JSON X-Y format, their large size prohibits their use in transferring NMR spectra across the Web. For example, the size of a typical NMR spectrum with 16K points is ~500 Kb in JSON X-Y, compared to only ~16 Kb when compressed into PNG format.

4.3.2.3. Smooth display of multiple spectra

NMRPro provides data simplification on client-side to enable NMR datasets to be displayed smoothly in the browser. This is not currently available in BokehJS.

4.3.2.4. High extensibility for both server-side and client-side components

NMRPro python package can be extended using plugins that can integrate both python and R functions (examples given in the documentation). On the client-side, SpecDrawJS dependency on D3.js (a low level visualization library), allows easy extensibility.

4.3.2.5. Using SpecdrawJS as standalone library

The small size of SpecdrawJS, because of its minimal dependencies (Only D3.js), allows its use as a client-side-only library. This is particularly useful for displaying spectra in NMR databases, such as HMDB (*79*) and BMRB (*230*).

4.3.2.6. Integration into existing web applications

NMRPro uses the Django framework to easily integrate into current chemical web applications such as ChEMBL (*87*), and into educational platforms such as Edx (https://github.com/edx/).

4.4. Subcomponents of NMRPro

The general application architecture (Figure 4.1) consists of three main subcomponents, NMRPro python package, Django-NMRPro App and SpecdrawJS. Below, I discuss the role of each subcomponent.

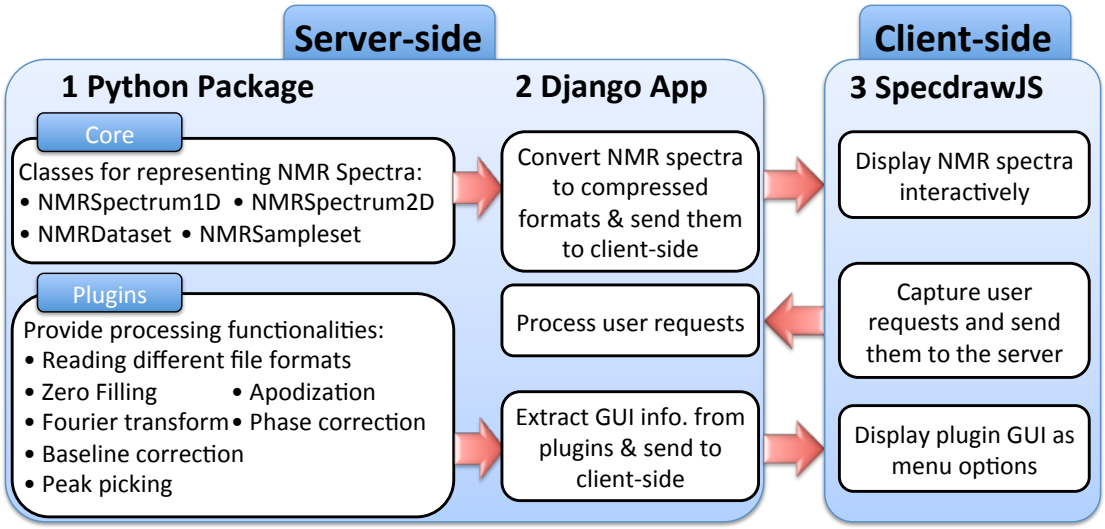

Figure 4.1 Component architecture of NMRPro.

4.4.1. Python Package

NMRPro python package consists of two main parts: python core, which provides object classes for representation of NMR spectra, and plugins, which provide different processing functionalities.

Python core provide four classes for programmatic representations of NMR spectra: 1D and 2D spectra, datasets and sample sets. All classes keep necessary information about the spectra and processing history. Processing history contains necessary functions to regenerate the processed spectrum from the raw one, increasing reproducibility.

Plugins contain functions for each of processing steps, where the input is NMR spectra along with processing parameters, and the output is the processed spectra. Each plugin also contains a GUI information entry, which is displayed in the web browser on the client-side, allowing the user to customize processing parameters. The plugin architecture allows extensibility of the application by

installing new plugins on the server, in which the GUI is updated automatically to match installed plugins.

NMRPro currently implemented plugins provides a wide range of functions, for time-domain and frequency-domain processing (Table 4.1). Each plugin allows automatic processing, in which the optimum algorithm is determined without user intervention, and customized processing. Customized processing provides users with list of comprehensive options covering most of the algorithms described in the literature. For example, apodization plugin contains 14 different window functions, and baseline correction contains 9 algorithms to estimate the baseline, extending the functionalities current software.

4.4.2. Django App

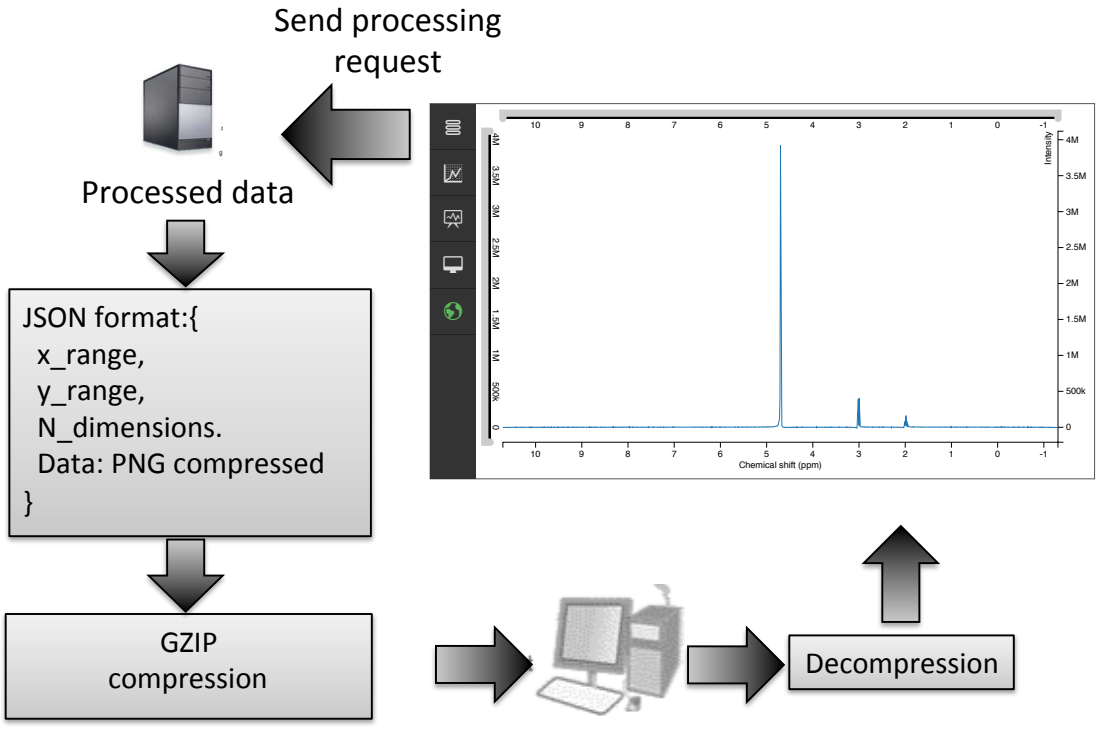

Figure 4.2 Data exchange protocol between server and client-sides, as managed by Django subcomponent.

Django framework enables the development software packages, 'Apps', that can be directly integrated into existing web applications, interfacing between python processing functionalities and client-side visualization. Django App is controls

the interaction between the server and client-side. The Django app has three roles in the API: 1) Efficient transfer of spectral to client-side. Since spectral data are too large to be sent to the web browser after each processing step, spectral data are first scaled down and then sent in a compressed format that can be read in the web browser, utilizing image compression (Figure 4.2). I chose PNG format for two reasons: lossless compression of data, and the ability to decompress PNG natively in all common browsers. 2) Management of user session data. While spectra are visualized in scaled down format on the client-side, calculation on the server-side are carried out using full-precision spectra. Django app stores and retrieves user spectra for processing on the server-side. 3) Aggregation of server-side plugins and sending their GUI to the client-side.

4.4.3. SpecdrawJS

Table 4.3 Functionalities available in each SpecdrawJS configuration.

Functionality	Static	Interactive	Full client-side	Connected
1D spectra	•	•	•	•
2D spectra	•	•	•	•
1D dataset	•	•	•	•
Sample sets (slides)		•	•	•
Zooming		•	•	•
Peak integration			•	•
Peak picking			• (Manual)	• (Manual, Threshold-based, CWT)
Save spectra (PNG Image, SVG)			•	•
Binning			•	•
Read NMR files			• (JCAMP-DX, PNG compressed)	• (Bruker, NMRPipe)
Spectral processing				•

CWT: continuous wavelet transform

SpecDrawJS is a platform-independent JavaScript library for visualization of 1D and 2D NMR spectra (Figure 4.3). SpecdrawJS can be used in four different configurations, summarized in Table 4.3: 1) Static view mode, in which spectra are rendered in-browser as scalable vector graphics (SVG) avoiding limited resolution of conventional images. 2) Interactive view mode allows users to zoom and pan across the spectra, and navigate between different slides. 3) Full

client-side mode provides opening locally stored files in JCAMP-DX or PNG compressed formats, peak picking and exporting spectra in different formats. 4) Connected mode provides GUI to all server-side functionalities listed in Table 4.1.

To enable visualization of NMR datasets, SpecdrawJS improves visualization performance by implementing two approaches: 1) Reducing the number of points in an NMR spectra using topology-preserving line simplification algorithm (*233, 234*). NMR spectra are reduced to the number of rendered pixels in the browser without affecting the perceived spectral shape. 2) Parallel programming using newly introduced web-worker technology.

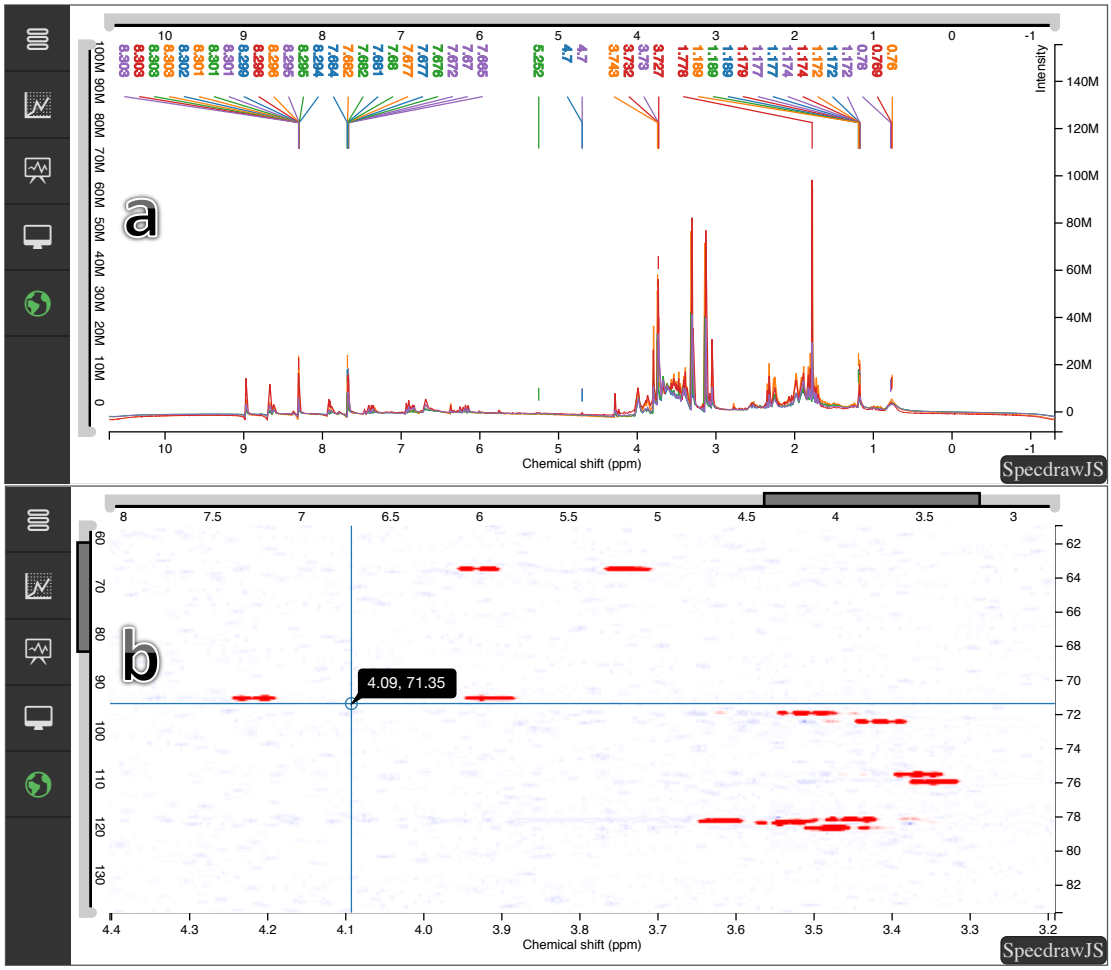

Figure 4.3 SpecdrawJS visualization. a) 1D NMR dataset. b) 2D NMR spectrum

4.5. Availability and Installation

NMRPro three subcomponents are available on public repositories. An introductory page, including live demo and instructions for installation, usage and plugin development is available at http://mamitsukalab.org/tools/nmrpro/. Below is a step-by-step instructions for installing NMRPro.

1. Install python 2.7 (https://www.python.org/downloads/release/python-2710/) and pip package manager.
2. From the terminal console, install NMRPro python package using the command:

```
pip install nmrpro
```

On Windows, from the command prompt:

```
python -m pip install nmrpro
```

3. Install the django-nmrpro App using the command:

```
pip install django-nmrpro
```

On Windows:

```
python -m pip install django-nmrpro
```

4. `pip` command automatically installs all necessary package dependencies.
5. There is no need to install SpecdrawJS separately since it is included in the Django App.

Once the Django App is installed, the user can integrate it into an existing Django project. To summarize the integration process, briefly:

1. If you do not have an existing Django project, first create one by following this tutorial (https://docs.djangoproject.com/en/1.8/intro/tutorial01/)
2. In `settings.py`, add `django_nmpro` to your `INSTALLED_APPS`.
3. In `urls.py`, add the following pattern:

```
url(r'^', include('django_nmrpro.urls')),
```

4. From the terminal console (command prompt on Windows), navigate to the projects home directory and run the web server using the command:

```
python manage.py migrate
```

5. Run the server using the command:

```
python manage.py runserver
```

6. To make sure that installation is successful, visit the URL:

http://127.0.0.1:8000/nmrpro_test/

Which should display 5 spectra from the Coffees dataset.

4.6. Conclusion

I presented NMRPro, an extensible web component that can be easily integrated in current web applications and databases, providing NMR processing and visualization functionalities. Future work is to extend NMRPro by implementing new plugins to add further functionalities such as covariance NMR and multivariate analysis for wider application in metabolomics and natural products.

Chapter 5

Conclusions

This study focused on computational tools for metabolomics and natural products research. Our initial literature review identified the lack of easy-to-use software in network path mining and processing NMR spectra. To address these two limitations, I developed two tools, NetPathMiner and NMRPro. I also conducted an comprehensive survey of computational resources for natural product dereplication.

I presented NetPathMiner as an R package that utilize gene expression measurements to infer activated parts of metabolic networks. NetPathMiner supported importing and constructing genome-scale metabolic networks through all major file formats, providing multiple representations for the constructed networks. Gene expression is used to weight network edges and then top *k* correlated paths are extracted. Because top correlated paths are enumerated from all possible paths, there tends to be up to thousands of output paths. NetPathMiner utilized clustering or classification to summarize paths according to their underlying functional components or their association with certain experimental conditions. Finally, paths are visualized on multiple network representations to facilitate the investigation of metabolic activity on multiple hierarchical levels.

I also surveyed current computational resources for rapid identification of natural products, dereplication. Dereplication requires the integration of diverse computational resources, namely, databases, methods and software. Reviewing the current databases indicated a scarcity of free-to-use databases that contain spectral data for previously isolated natural products. Also, a unified software tool with an easy-to-use interface to spectral processing and analysis was lacking.

Based on the results of the survey I presented NMRPro, which is a pluggable web component for interactive processing and visualization of NMR spectra. NMRPro can be easily integrated into existing web applications, and can be extended through NMRPro plugin architecture.

Acknowledgements

I would like to express my sincere gratitude and deepest sense of appreciation to Professor Hiroshi Mamitsuka, Bioinformatics Center, Institute for Chemical Research, Kyoto University, for his care and guidance throughout the entire period of study. This study would not have been possible without his extensive supervision and continued patience on my limited understanding and many shortcomings.

Sincere gratitude and thanks are also extended to Drs. Timothy Hancock and Canh Hao Nguyen for their care and relentless help during my study. Special thanks are due to my lab mates Drs. Masayuki Karasuyama and Makoto Yamada, Keiichiro Takahashi, Yayoi Natsume and Sohiya Yotsukura for their friendly and polite behavior. I offer my sincerest thanks to the faculty office staff for kind cooperation and providing valuable information about daily life in Japan.

I would like to acknowledge the Japan Society for the Promotion of Science, Japan (JSPS) for financial support during my stay in Boston, USA. I also would like to thank Rotary Yoneyama memorial foundation scholarship their financial support during my stay in Japan.

Special thanks are expressed to all Arab students in Osaka for their friendship, support and encouragement throughout my stay in Japan. I particularly thank Ahmed Haredy and Elias Tannous for being by side both personally and academically.

I would like to express the dearest thanks of all to my parents, to whom I am forever indebted. Also, thanks to my three younger sisters for the support and encouragement.

Finally, I would like to thank my lovely wife and constant source of happiness and hope, Ala, for her endless patience until this study was fruitfully finished. My appreciation for her support will last forever.

References

1. L. J. Collins, B. Schönfeld, X. S. Chen, in *Handbook of epigenetics: the new molecular and medical genetics*. (Academic, 2011), pp. 49-61.

2. G. Elgar, T. Vavouri, Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in genetics* **24**, 344-352 (2008).

3. N. R. Boyle, J. A. Morgan, Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii. *BMC systems biology* **3**, 1 (2009).

4. N. Irani, M. Wirth, J. van den Heuvel, R. Wagner, Improvement of the primary metabolism of cell cultures by introducing a new cytoplasmic pyruvate carboxylase reaction. *Biotechnology and bioengineering* **66**, 238-246 (1999).

5. J. Koricheva, S. Larsson, E. Haukioja, M. Keinänen, Regulation of woody plant secondary metabolism by resource availability: hypothesis testing by means of meta-analysis. *Oikos*, 212-226 (1998).

6. N. P. Keller, G. Turner, J. W. Bennett, Fungal secondary metabolism—from biochemistry to genomics. *Nature Reviews Microbiology* **3**, 937-947 (2005).

7. C. Smolke, *The metabolic pathway engineering handbook: Fundamentals*. (CRC press, 2009), vol. 1.

8. L. J. Sweetlove, T. Obata, A. R. Fernie, Systems analysis of metabolic phenotypes: what have we learnt? *Trends in Plant Science* **19**, 222-230 (2014).

9. C. Lerman, R. Tyndale, F. Patterson, E. P. Wileyto, P. G. Shields, A. Pinto, N. Benowitz, Nicotine metabolite ratio predicts efficacy of transdermal nicotine for smoking cessation*. *Clinical Pharmacology & Therapeutics* **79**, (2006).

10. D. S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai, A. L. Barabási, The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences* **105**, 9880-9885 (2008).

11. D. J. Newman, G. M. Cragg, Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of natural products* **75**, 311-335 (2012).

12. G. Plata, T.-L. Hsiao, K. L. Olszewski, M. Llinás, D. Vitkup, Reconstruction and flux-balance analysis of the Plasmodium falciparum metabolic network. *Molecular Systems Biology* **6**, 408 (2010); published online EpubSep 07 (10.1038/msb.2010.60).

13. E. Segal, H. Wang, D. Koller, Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* **19**, i264--i272 (2003).

14. I. Ulitsky, R. Shamir, Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics* **25**, 1158--1164 (2009).

15. E. Georgii, S. Dietmann, T. Uno, P. Pagel, K. Tsuda, Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics* **25**, 933--940 (2009).

16. T. Ideker, O. Ozier, B. Schwikowski, A. F. Siegel, Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S233-240 (2002).

17. D. Hanisch, A. Zien, R. Zimmer, T. Lengauer, Co-clustering of biological networks and gene expression data. *Bioinformatics* **18**, S145--S154 (2002).

18. J. P. Vert, M. Kanehisa, Extracting active pathways from gene expression data. *Bioinformatics* **19**, ii238--ii244 (2003).

19. I. Takigawa, H. Mamitsuka, Probabilistic path ranking based on adjacent pairwise coexpression for metabolic transcripts analysis. *Bioinformatics*

(Oxford, England) **24**, 250-257 (2008); published online EpubFeb 13 (10.1093/bioinformatics/btm575).

20. T. Hancock, I. Takigawa, H. Mamitsuka, Mining metabolic pathways through gene expression. *Bioinformatics (Oxford, England)* **26**, 2128-2135 (2010); published online EpubSep 01 (10.1093/bioinformatics/btq344).

21. H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research* **27**, 29--34 (1999).

22. G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, others, Reactome: a knowledgebase of biological pathways. *Nucleic acids research* **33**, D428--D432 (2005).

23. R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, P. D. Karp, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research* **40**, D742-753 (2012); published online EpubJan (10.1093/nar/gkr1014).

24. E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, C. Sander, Pathway Commons, a web resource for biological pathway data. *Nucleic acids research* **39**, D685-690 (2011); published online EpubJan (10.1093/nar/gkq1039).

25. W. Luo, C. Brouwer, Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**, 1830-1831 (2013); published online EpubJul 15 (10.1093/bioinformatics/btt285).

26. F. Kramer, M. Bayerlova, F. Klemm, A. Bleckmann, T. Beissbarth, rBiopaxParser--an R package to parse, modify and visualize BioPAX data. *Bioinformatics* **29**, 520-522 (2013); published online EpubFeb 15 (10.1093/bioinformatics/bts710).

27. J. D. Zhang, S. Wiemann, KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* **25**, 1470-1471 (2009); published online EpubJun 1 (10.1093/bioinformatics/btp167).

28. G. Sales, E. Calura, D. Cavalieri, C. Romualdi, graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics* **13**, 20 (2012)10.1186/1471-2105-13-20).

29. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, J. Zhang, Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004)10.1186/gb-2004-5-10-r80).

30. G. Csardi, T. Nepusz, The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, (2006).

31. T. Hancock, H. Mamitsuka, Active pathway identification and classification with probabilistic ensembles. *Genome informatics. International Conference on Genome Informatics* **22**, 30-40 (2010); published online EpubFeb (

32. H. Mamitsuka, Y. Okuno, A. Yamaguchi, Mining biologically active patterns in metabolic pathways using microarray expression profiles. *ACM SIGKDD Explorations Newsletter* **5**, 113--121 (2003).

33. J. M. Stuart, E. Segal, D. Koller, S. K. Kim, A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249--255 (2003).

34. G. Wu, X. Feng, L. Stein, A human functional protein interaction network and its application to cancer data analysis. *Genome Biology* **11**, R53 (2010)10.1186/gb-2010-11-5-r53).

35. A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease. *Nature Publishing Group* **12**, 56--68 (2011); published online EpubJanuary (

36. S. Bandyopadhyay, R. Kelley, N. J. Krogan, T. Ideker, Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Computational Biology* **4**, e1000065 (2008); published online EpubMay (10.1371/journal.pcbi.1000065).

37. B. Mlecnik, M. Scheideler, H. Hackl, J. Hartler, F. Sanchez-Cabo, Z. Trajanoski, PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic acids research* **33**, W633--W637 (2005).

38. A. Breitkreutz, H. Choi, J. R. Sharom, L. Boucher, V. Neduva, B. Larsen, Z. Y. Lin, B. J. Breitkreutz, C. Stark, G. Liu, others, A global protein kinase and phosphatase interaction network in yeast. *Science Signalling* **328**, 1043 (2010).

39. G. A. Churchill, Fundamentals of experimental design for cDNA microarrays. *Nature genetics* **32**, 490-495 (2002).

40. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63 (2009).

41. H. Matsumura, S. Reich, A. Ito, H. Saitoh, S. Kamoun, P. Winter, G. Kahl, M. Reuter, D. H. Krüger, R. Terauchi, Gene expression analysis of plant host–pathogen interactions by SuperSAGE. *Proceedings of the National Academy of Sciences* **100**, 15718-15723 (2003).

42. A. Hoppe, What mRNA abundances can tell us about metabolism. *Metabolites* **2**, 614-631 (2012).

43. F. Carrari, C. Baxter, B. Usadel, E. Urbanczyk-Wochniak, M. I. Zanor, A. Nunes-Nesi, V. Nikiforova, D. Centero, A. Ratzka, M. Pauly, L. J. Sweetlove, A. R. Fernie, Integrated analysis of metabolite and transcript levels reveals the

metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. *Plant Physiol* **142**, 1380-1396 (2006); published online EpubDec (Doi 10.1104/Pp.106.088534).

44. A. Bauer-Mehren, L. I. Furlong, F. Sanz, Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular Systems Biology* **5**, 290 (2009)10.1038/msb.2009.47).

45. N. Juty, N. Le Novere, C. Laibe, Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res* **40**, D580-586 (2012); published online EpubJan (10.1093/nar/gkr1097).

46. M. P. van Iersel, A. R. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B. R. Conklin, C. T. Evelo, The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* **11**, 5 (2010)10.1186/1471-2105-11-5).

47. J. Y. Yen, Finding the k shortest loopless paths in a network. *Management Science* **17**, 712-716 (1971).

48. E. L. Lawler, A procedure for computing the k best solutions to discrete optimization problems and its application to the shortest path problem. *Management Science* **18**, 401-405 (1972).

49. T. Hancock, N. Wicker, I. Takigawa, H. Mamitsuka, Identifying neighborhoods of coordinated gene expression and metabolite profiles. *PLoS ONE* **7**, e31345 (2012)10.1371/journal.pone.0031345).

50. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines. *The journal of chemical physics* **21**, 1087 (1953).

51. P. T. Shannon, M. Grimes, B. Kutlu, J. J. Bot, D. J. Galas, RCytoscape: tools for exploratory network analysis. *BMC Bioinformatics* **14**, 217 (2013)10.1186/1471-2105-14-217).

52. R. Gentleman, E. Whalen, W. Huber, S. Falcon, graph: A package to handle graph data structures. *R package*, (2009).

53. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550 (2005); published online EpubOct 25 (10.1073/pnas.0506580102).

54. S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, R. Romero, A systems biology approach for pathway level analysis. (2007).

55. J. W. Li, J. C. Vederas, Drug discovery and natural products: end of an era or an endless frontier? *Science* **325**, 161-165 (2009); published online EpubJul 10 (10.1126/science.1168243).

56. J. A. Beutler, Natural products as a foundation for drug discovery. *Current protocols in pharmacology / editorial board, S.J. Enna* **Chapter 9**, Unit 9 11 (2009); published online EpubSep (10.1002/0471141755.ph0911s46).

57. F. E. Koehn, G. T. Carter, The evolving role of natural products in drug discovery. *Nat Rev Drug Discov* **4**, 206-220 (2005); published online EpubMar (10.1038/nrd1657).

58. J. Berdy, Bioactive microbial metabolites. *The Journal of antibiotics* **58**, 1-26 (2005); published online EpubJan (10.1038/ja.2005.1).

59. P. A. Clemons, N. E. Bodycombe, H. A. Carrinski, J. A. Wilson, A. F. Shamji, B. K. Wagner, A. N. Koehler, S. L. Schreiber, Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc Natl Acad Sci U S A* **107**, 18787-18792 (2010); published online EpubNov 2 (10.1073/pnas.1012741107).

60. B. Over, S. Wetzel, C. Grutter, Y. Nakai, S. Renner, D. Rauh, H. Waldmann, Natural-product-derived fragments for fragment-based ligand discovery. *Nature chemistry* **5**, 21-28 (2013); published online EpubJan (10.1038/nchem.1506).

61. J. Buckingham, *Dictionary of natural products*. (CRC Press, 1993), vol. 6.

62. J. W. Blunt, M. H. G. Munro, 22 Is There an Ideal Database for Natural Products Research? *Natural Products: Discourse, Diversity, and Design*, 413 (2014).

63. J.-L. Wolfender, G. Marti, E. Ferreira Queiroz, Advances in Techniques for Profiling Crude Extracts and for the Rapid Identificationof Natural Products: Dereplication, Quality Control and Metabolomics. *Current organic chemistry* **14**, 1808-1832 (2010).

64. G. Lang, N. A. Mayhudin, M. I. Mitova, L. Sun, S. van der Sar, J. W. Blunt, A. L. Cole, G. Ellis, H. Laatsch, M. H. Munro, Evolving trends in the dereplication of natural product extracts: new methodology for rapid, small-scale investigation of natural product extracts. *Journal of natural products* **71**, 1595-1599 (2008); published online EpubSep (10.1021/np8002222).

65. W. H. Gerwick, B. S. Moore, Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chemistry & biology* **19**, 85-98 (2012); published online EpubJan 27 (10.1016/j.chembiol.2011.12.014).

66. M. L. Rosenblum, M. A. Gerosa, C. B. Wilson, G. R. Barger, B. F. Pertuiset, N. de Tribolet, D. V. Dougherty, Stem cell studies of human malignant brain tumors. Part 1: Development of the stem cell assay and its potential. *Journal of neurosurgery* **58**, 170-176 (1983); published online EpubFeb (10.3171/jns.1983.58.2.0170).

67. T. F. Molinski, Microscale methodology for structure elucidation of natural products. *Curr Opin Biotechnol* **21**, 819-826 (2010); published online EpubDec (10.1016/j.copbio.2010.09.003).

68. M. Halabalaki, K. Vougogiannopoulou, E. Mikros, A. L. Skaltsounis, Recent advances and new strategies in the NMR-based identification of natural products. *Current opinion in biotechnology* **25**, 1-7 (2014).

69. Y. Liu, M. D. Green, R. Marques, T. Pereira, R. Helmy, R. T. Williamson, W. Bermel, G. E. Martin, Using pure shift HSQC to characterize microgram samples of drug metabolites. *Tetrahedron Letters* **55**, 5450-5453 (2014).

70. J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira, P. C. Dorrestein, Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci U S A* **109**, E1743-1752 (2012); published online EpubJun 26 (10.1073/pnas.1203689109).

71. J. Y. Yang, L. M. Sanchez, C. M. Rath, X. Liu, P. D. Boudreau, N. Bruns, E. Glukhov, A. Wodtke, R. de Felicio, A. Fenner, W. R. Wong, R. G. Linington, L. Zhang, H. M. Debonsi, W. H. Gerwick, P. C. Dorrestein, Molecular networking as a dereplication strategy. *Journal of natural products* **76**, 1686-1699 (2013); published online EpubSep 27 (10.1021/np400413s).

72. M. E. Elyashberg, Identification and structure elucidation by NMR spectroscopy. *TrAC Trends in Analytical Chemistry*, (2015).

73. S. L. Robinette, R. Brüschweiler, F. C. Schroeder, A. S. Edison, NMR in metabolomics and natural products research: two sides of the same coin. *Accounts of chemical research* **45**, 288-297 (2011).

74. B. Wang, A. Fang, J. Heim, B. Bogdanov, S. Pugh, M. Libardoni, X. Zhang, DISCO: distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Analytical chemistry* **82**, 5069-5081 (2010).

75. D. S. Wishart, Quantitative metabolomics using NMR. *TrAC Trends in Analytical Chemistry* **27**, 228-237 (2008).

76. O. Beckonert, H. C. Keun, T. M. D. Ebbels, J. Bundy, E. Holmes, J. C. Lindon, J. K. Nicholson, Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols* **2**, 2692-2703 (2007).

77. K. A. Blinov, D. Carlson, M. E. Elyashberg, G. E. Martin, E. R. Martirosian, S. Molodtsov, A. J. Williams, Computer‐assisted structure elucidation of natural products with limited 2D NMR data: application of the StrucEluc system. *Magnetic Resonance in Chemistry* **41**, 359-372 (2003).

78. R. C. Breton, W. F. Reynolds, Using NMR to identify and characterize natural products. *Natural product reports* **30**, 501-524 (2013).

79. D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, A. Scalbert, HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res* **41**, D801-807 (2013); published online EpubJan (10.1093/nar/gks1065).

80. A. Smolinska, L. Blanchet, L. M. Buydens, S. S. Wijmenga, NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Anal Chim Acta* **750**, 82-97 (2012); published online EpubOct 31 (10.1016/j.aca.2012.05.049).

81. H. F. Ji, X. J. Li, H. Y. Zhang, Natural products and drug discovery. *EMBO reports* **10**, 194-200 (2009).

82. S. Dandapani, L. A. Marcaurelle, Grand challenge commentary: Accessing new chemical space for'undruggable'targets. *Nature chemical biology* **6**, 861-863 (2010).

83. M. Füllbeck, E. Michalsky, M. Dunkel, R. Preissner, Natural products: sources and databases. *Natural product reports* **23**, 347-356 (2006).

84. J. Blunt, M. Munro, M. Upjohn, in *Handbook of Marine Natural Products*. (Springer, 2012), pp. 389-421.

85. A. A. Lagunin, R. K. Goel, D. Y. Gawande, P. Pahwa, T. A. Gloriozova, A. V. Dmitriev, S. M. Ivanov, A. V. Rudik, V. I. Konova, P. V. Pogodin, Chemo-and bioinformatics resources for in silico drug discovery from medicinal plants beyond their traditional use: a critical review. *Natural product reports* **31**, 1585-1611 (2014).

86. Q. Li, T. Cheng, Y. Wang, S. H. Bryant, PubChem as a public resource for drug discovery. *Drug discovery today* **15**, 1052-1057 (2010); published online EpubDec (10.1016/j.drudis.2010.10.003).

87. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **40**, D1100-1107 (2012); published online EpubJan (10.1093/nar/gkr777).

88. T. Liu, Y. Lin, X. Wen, R. N. Jorissen, M. K. Gilson, BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research* **35**, D198-D201 (2007).

89. C. Roldán, A. de la Torre, S. Mota, A. Morales-Soto, J. Menéndez, A. Segura-Carretero, Identification of active compounds in vegetal extracts based on correlation between activity and HPLC–MS data. *Food chemistry* **136**, 392-399 (2013).

90. J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, C. Steinbeck, The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* **41**, D456-463 (2013); published online EpubJan (10.1093/nar/gks1146).

91. J. Goodman, Computer software review: Reaxys. *Journal of Chemical Information and Modeling* **49**, 2897-2898 (2009).

92. C. Steinbeck, S. Kuhn, NMRShiftDB -- compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* **65**, 2711-2717 (2004); published online EpubOct (10.1016/j.phytochem.2004.08.027).

93. H. Kalchhauser, W. Robien, CSEARCH: A computer program for identification of organic compounds and fully automated assignment of carbon-13 nuclear magnetic resonance spectra. *Journal of Chemical Information and Computer Sciences* **25**, 103-108 (1985).

94. A. Barth, SpecInfo: an integrated spectroscopic information system. *Journal of chemical information and computer sciences* **33**, 52-58 (1993).

95. K. P. Seiler, G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski, S. Norton, S. Brudz, J. P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N. J. Tolliday, S. L. Schreiber, P. A. Clemons, ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res* **36**, D351-359 (2008); published online EpubJan (10.1093/nar/gkm843).

96. . vol. 2014.

97. J. J. Irwin, B. K. Shoichet, ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **45**, 177-182 (2005); published online EpubJan-Feb (10.1021/ci049714+).

98. H. Laatsch, AntiBase, a Database for rapid dereplication and structure determination of microbial natural products. *Book AntiBase, a Database for rapid dereplication and structure determination of microbial natural products*, (2010).

99. R. Hammami, A. Zouhir, C. Le Lay, J. Ben Hamida, I. Fliss, BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC microbiology* **10**, 22 (2010)10.1186/1471-2180-10-22).

100. F. Ntie-Kang, J. A. Mbah, L. M. Mbaze, L. L. Lifongo, M. Scharfe, J. N. Hanna, F. Cho-Ngwa, P. A. Onguene, L. C. Owono Owono, E. Megnassan, W. Sippl, S. M. Efange, CamMedNP: building the Cameroonian 3D structural natural products database for virtual screening. *BMC complementary and alternative medicine* **13**, 88 (2013)10.1186/1472-6882-13-88).

101. F. Ntie-Kang, P. A. Onguéné, M. Scharfe, L. C. O. Owono, E. Megnassan, L. M. a. Mbaze, W. Sippl, S. M. N. Efange, ConMedNP: a natural product library from Central African medicinal plants for drug discovery. *RSC Advances* **4**, 409-419 (2014).

102. J. L. Lopez-Perez, R. Theron, E. del Olmo, D. Diaz, NAPROC-13: a database for the dereplication of natural product mixtures in bioassay-guided protocols. *Bioinformatics* **23**, 3256-3257 (2007); published online EpubDec 1 (10.1093/bioinformatics/btm516).

103. M. Mangal, P. Sagar, H. Singh, G. P. Raghava, S. M. Agarwal, NPACT: Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database. *Nucleic Acids Res* **41**, D1124-1129 (2013); published online EpubJan (10.1093/nar/gks1047).

104. M. Valli, R. N. dos Santos, L. D. Figueira, C. H. Nakajima, I. Castro-Gamboa, A. D. Andricopulo, V. S. Bolzani, Development of a natural products database from the biodiversity of Brazil. *Journal of natural products* **76**, 439-444 (2013); published online EpubMar 22 (10.1021/np3006875).

105. R. Hammami, J. Ben Hamida, G. Vergoten, I. Fliss, PhytAMP: a database dedicated to antimicrobial plant peptides. *Nucleic Acids Res* **37**, D963-968 (2009); published online EpubJan (10.1093/nar/gkn655).

106. M. Dunkel, M. Fullbeck, S. Neumann, R. Preissner, SuperNatural: a searchable database of available natural compounds. *Nucleic Acids Res* **34**, D678-683 (2006); published online EpubJan 1 (10.1093/nar/gkj132).

107. P. Banerjee, J. Erehman, B.-O. Gohlke, T. Wilhelm, R. Preissner, M. Dunkel, Super Natural II—a database of natural products. *Nucleic acids research*, gku886 (2014).

108. C. Y. Chen, TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One* **6**, e15939 (2011)10.1371/journal.pone.0015939).

109. J. Gu, Y. Gui, L. Chen, G. Yuan, H.-Z. Lu, X. Xu, Use of natural products as chemical library for drug discovery and network pharmacology. *PloS one* **8**, e62839 (2013).

110. T. N. Vu, K. Laukens, Getting your peaks in line: a review of alignment methods for NMR spectral data. *Metabolites* **3**, 259-276 (2013).

111. N. M. Olboyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open Babel: An open chemical toolbox. *J Cheminf* **3**, 33 (2011).

112. Y. Cao, A. Charisi, L. C. Cheng, T. Jiang, T. Girke, ChemmineR: a compound mining framework for R. *Bioinformatics* **24**, 1733-1734 (2008); published online EpubAug 1 (10.1093/bioinformatics/btn307).

113. R. Guha, Chemical informatics functionality in R. *Journal of Statistical Software* **18**, 1-16 (2007).

114. D.-S. Cao, N. Xiao, Q.-S. Xu, A. F. Chen, Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds, and their interactions. *Bioinformatics*, btu624 (2014).

115. R. J. Lancashire, The JSpecView Project: an Open Source Java viewer and converter for JCAMP-DX, and XML spectral data files. *Chemistry Central journal* **1**, 31 (2007)10.1186/1752-153X-1-31).

116. B. Bienfait, P. Ertl, JSME: a free molecule editor in JavaScript. *J Cheminform* **5**, 24 (2013); published online EpubMay 21 (10.1186/1758-2946-5-24).

117. F. Csizmadia, JChem: Java applets and modules supporting chemical database handling from web browsers. *Journal of Chemical Information and Computer Sciences* **40**, 323-324 (2000).

118. T. Wang, K. Shao, Q. Chu, Y. Ren, Y. Mu, L. Qu, J. He, C. Jin, B. Xia, Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. *BMC Bioinformatics* **10**, 83 (2009)10.1186/1471-2105-10-83).

119. J. Hao, W. Astle, M. De Iorio, T. M. Ebbels, BATMAN--an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* **28**, 2088-2090 (2012); published online EpubAug 1 (10.1093/bioinformatics/bts308).

120. B. A. Hanson, ChemoSpec: An R Package for Chemometric Analysis of Spectroscopic Data and Chromatograms (Package Version 1.61-3). (2013).

121. S. Kim, A. Fang, B. Wang, J. Jeong, X. Zhang, An optimal peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry using mixture similarity measure. *Bioinformatics* **27**, 1660-1666 (2011); published online EpubJun 15 (10.1093/bioinformatics/btr188).

122. B. Worley, R. Powers, MVAPACK: a complete data handling package for NMR metabolomics. *ACS chemical biology* **9**, 1138-1144 (2014).

123. J. Wist, L. Patiny, Structural Analysis from Classroom to Laboratory. *Journal of Chemical Education* **89**, 1083-1083 (2012).

124. J. J. Helmus, C. P. Jaroniec, Nmrglue: an open source Python package for the analysis of multidimensional NMR data. *Journal of biomolecular NMR* **55**, 355-367 (2013).

125. F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, A. D. Bax, NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Journal of biomolecular NMR* **6**, 277-293 (1995).

126. J. L. Izquierdo, M. Orphaned, F. Depends Rwave, Package 'NMRS'.

127. I. A. Lewis, S. C. Schommer, J. L. Markley, rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magnetic resonance in chemistry : MRC* **47 Suppl 1**, S123-126 (2009); published online EpubDec (10.1002/mrc.2526).

128. T. N. Vu, D. Valkenborg, K. Smets, K. A. Verwaest, R. Dommisse, F. Lemiere, A. Verschoren, B. Goethals, K. Laukens, An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics* **12**, 405 (2011)10.1186/1471-2105-12-405).

129. A. N. Davies, P. Lampen, Jcamp-Dx for NMR. *Applied spectroscopy* **47**, 1093-1099 (1993).

130. T. D. Goddard, D. G. Kneller, Sparky—NMR assignment and integration software. *University of California, San Francisco*, (2006).

131. F. Zhang, R. Brüschweiler, Robust deconvolution of complex mixtures by covariance TOCSY spectroscopy. *Angewandte Chemie International Edition* **46**, 2639-2642 (2007).

132. S. L. Robinette, F. Zhang, L. Bruschweiler-Li, R. Brüschweiler, Web server based complex mixture analysis by NMR. *Analytical chemistry* **80**, 3606-3611 (2008).

133. D. V. Rubtsov, H. Jenkins, C. Ludwig, J. Easton, M. R. Viant, U. Günther, J. L. Griffin, N. Hardy, Proposed reporting requirements for the description of NMR-based metabolomics experiments. *Metabolomics* **3**, 223-229 (2007).

134. J. Downing, P. Murray-Rust, A. P. Tonge, P. Morgan, H. S. Rzepa, F. Cotterill, N. Day, M. J. Harvey, SPECTRa: the deposition and validation of primary

chemistry research data in digital repositories. *Journal of chemical information and modeling* **48**, 1571-1581 (2008).

135. W. F. Vranken, W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides, E. D. Laue, The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* **59**, 687-696 (2005); published online EpubJun 1 (10.1002/prot.20449).

136. F. Chignola, S. Mari, T. J. Stevens, R. H. Fogh, V. Mannella, W. Boucher, G. Musco, The CCPN Metabolomics Project: a fast protocol for metabolite identification by 2D-NMR. *Bioinformatics* **27**, 885-886 (2011); published online EpubMar 15 (10.1093/bioinformatics/btr013).

137. S. R. Hall, The STAR file: A new format for electronic data transfer and archiving. *Journal of Chemical Information and Computer Sciences* **31**, 326-333 (1991).

138. S. R. Hall, N. Spadaccini, The STAR file: Detailed specifications. *Journal of Chemical Information and Computer Sciences* **34**, 505-508 (1994).

139. N. Spadaccini, S. R. Hall, Extensions to the STAR File syntax. *J Chem Inf Model* **52**, 1901-1906 (2012); published online EpubAug 27 (10.1021/ci300074v).

140. W. Dietrich, C. H. Rüdel, M. Neumann, Fast and precise automatic baseline correction of one-and two-dimensional NMR spectra. *Journal of Magnetic Resonance (1969)* **91**, 1-11 (1991).

141. J. C. Cobas, M. A. Bernstein, M. Martin-Pastor, P. G. Tahoces, A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *Journal of magnetic resonance* **183**, 145-151 (2006); published online EpubNov (10.1016/j.jmr.2006.07.013).

142. Q. Bao, J. Feng, F. Chen, W. Mao, Z. Liu, K. Liu, C. Liu, A new automatic baseline correction method based on iterative method. *Journal of magnetic resonance* **218**, 35-43 (2012).

143. X. Shao, C. Ma, A general approach to derivative calculation using wavelet transform. *Chemometrics and Intelligent Laboratory Systems* **69**, 157-165 (2003).

144. X. Shao, W. Cai, Z. Pan, Wavelet transform and its applications in high performance liquid chromatography (HPLC) analysis. *Chemometrics and intelligent laboratory systems* **45**, 249-256 (1999).

145. D. E. Brown, Fully automated baseline correction of 1D and 2D NMR spectra using Bernstein polynomials. *Journal of Magnetic Resonance, Series A* **114**, 268-270 (1995).

146. F. Gan, G. Ruan, J. Mo, Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems* **82**, 59-65 (2006).

147. Y. Xi, D. M. Rocke, Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC bioinformatics* **9**, 324 (2008).

148. H. F. M. Boelens, R. J. Dijkstra, P. H. C. Eilers, F. Fitzpatrick, J. A. Westerhuis, New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. *Journal of chromatography A* **1057**, 21-30 (2004).

149. A. F. Ruckstuhl, M. P. Jacobson, R. W. Field, J. A. Dodd, Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer* **68**, 179-193 (2001).

150. Ł. Komsta, Comparison of several methods of chromatographic baseline removal with a new approach based on quantile regression. *Chromatographia* **73**, 721-731 (2011).

151. X. Liu, Z. Zhang, P. F. M. Sousa, C. Chen, M. Ouyang, Y. Wei, Y. Liang, Y. Chen, C. Zhang, Selective iteratively reweighted quantile regression for baseline correction. *Analytical and bioanalytical chemistry* **406**, 1985-1998 (2014).

152. Z.-M. Zhang, S. Chen, Y.-Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst* **135**, 1138-1146 (2010).

153. A. F. Tawfike, C. Viegelmann, R. Edrada-Ebel, in *Metabolomics Tools for Natural Product Discovery*. (Springer, 2013), pp. 227-244.

154. R. Koradi, M. Billeter, M. Engeli, P. Guntert, K. Wuthrich, Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *Journal of magnetic resonance* **135**, 288-297 (1998); published online EpubDec (10.1006/jmre.1998.1570).

155. L. Brodsky, A. Moussaieff, N. Shahaf, A. Aharoni, I. Rogachev, Evaluation of peak picking quality in LC-MS metabolomics data. *Anal Chem* **82**, 9177-9187 (2010); published online EpubNov 15 (10.1021/ac101216e).

156. C. Yang, Z. He, W. Yu, Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinformatics* **10**, 4 (2009)10.1186/1471-2105-10-4).

157. R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, J. C. Wilson, Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometrics and Intelligent Laboratory Systems* **85**, 144-154 (2007).

158. T. De Meyer, D. Sinnaeve, B. Van Gasse, E. Tsiporkova, E. R. Rietzschel, M. L. De Buyzere, T. C. Gillebert, S. Bekaert, J. C. Martins, W. Van Criekinge, NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical Chemistry* **80**, 3783-3790 (2008).

159. P. E. Anderson, D. A. Mahle, T. E. Doom, N. V. Reo, N. J. DelRaso, M. L. Raymer, Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data. *Metabolomics* **7**, 179-190 (2011).

160. A. Hinneburg, A. Porzel, K. Wolfram, *An evaluation of text retrieval methods for similarity search of multi-dimensional nmr-spectra*. (Springer, 2007).

161. J. Luts, J. B. Poullet, J. M. Garcia‐Gomez, A. Heerschap, M. Robles, J. A. K. Suykens, S. V. Huffel, Effect of feature extraction for brain tumor classification based on short echo time 1H MR spectra. *Magnetic Resonance in Medicine* **60**, 288-298 (2008).

162. A. M. Castillo, L. Uribe, L. Patiny, J. Wist, Fast and shift-insensitive similarity comparisons of NMR using a tree-representation of spectra. *Chemometrics and Intelligent Laboratory Systems* **127**, 1-6 (2013).

163. A. M. Castillo, A. Bernal, L. Patiny, J. Wist, A new method for the comparison of 1H NMR predictors based on tree-similarity of spectra. *Journal of cheminformatics* **6**, 1-6 (2014).

164. A. P. Singh, J. Halloran, J. A. Bilmes, K. Kirchoff, W. S. Noble, Spectrum identification using a dynamic Bayesian network model of tandem mass spectra. *arXiv preprint arXiv:1210.4904*, (2012).

165. J. Jeong, X. Shi, X. Zhang, S. Kim, C. Shen, Model-based peak alignment of metabolomic profiling from comprehensive two-dimensional gas chromatography mass spectrometry. *BMC Bioinformatics* **13**, 27 (2012)10.1186/1471-2105-13-27).

166. D. E. Green, Quantitation of cannabinoids in biological specimens using probability based matching GC/MS. *NIDA research monograph*, 70-87 (1976); published online EpubMay (

167. F. W. McLafferty, R. H. Hertel, R. D. Villwock, Probability based matching of mass spectra. Rapid identification of specific compounds in mixtures. *Organic Mass Spectrometry* **9**, 690-702 (1974).

168. S. Koichi, M. Arisaka, H. Koshino, A. Aoki, S. Iwata, T. Uno, H. Satoh, Chemical Structure Elucidation from 13C NMR Chemical Shifts: Efficient Data

Processing Using Bipartite Matching and Maximal Clique Algorithms. *Journal of chemical information and modeling* **54**, 1027-1035 (2014).

169. M. Levandowsky, D. Winter, Distance between sets. *Nature* **234**, 34-35 (1971).

170. B. Egert, S. Neumann, A. Hinneburg, in *Data Integration in the Life Sciences*. (Springer, 2007), pp. 139-155.

171. A. Hinneburg, B. Egert, A. Porzel, Duplicate detection of 2d-nmr spectra. *Journal of Integrative Bioinformatics* **4**, 53 (2007).

172. I. Beer, E. Barnea, T. Ziv, A. Admon, Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **4**, 950-960 (2004); published online EpubApr (10.1002/pmic.200300652).

173. B. L. Atwater, D. B. Stauffer, F. W. McLafferty, D. W. Peterson, Reliability ranking and scaling improvements to the probability based matching system for unknown mass spectra. *Analytical Chemistry* **57**, 899-903 (1985).

174. D. L. Tabb, M. J. MacCoss, C. C. Wu, S. D. Anderson, J. R. Yates, Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Analytical chemistry* **75**, 2470-2477 (2003).

175. J. Li, D. B. Hibbert, S. Fuller, J. Cattle, C. Pang Way, Comparison of spectra using a Bayesian approach. An argument using oil spills as an example. *Analytical chemistry* **77**, 639-644 (2005).

176. A. Linusson, S. Wold, B. Nordén, Fuzzy clustering of 627 alcohols, guided by a strategy for cluster analysis of chemical compounds for combinatorial chemistry. *Chemometrics and intelligent laboratory systems* **44**, 213-227 (1998).

177. R. K. Julian, R. E. Higgs, J. D. Gygi, M. D. Hilton, A method for quantitatively differentiating crude natural extracts using high-performance liquid chromatography-electrospray mass spectrometry. *Analytical chemistry* **70**, 3249-3254 (1998).

178. A. Tsipouras, J. Ondeyka, C. Dufresne, S. Lee, G. Salituro, N. Tsou, M. Goetz, S. B. Singh, S. K. Kearsley, Using similarity searches over databases of estimated< sup> 13</sup> C NMR spectra for structure identification of natural product compounds. *Analytica Chimica Acta* **316**, 161-171 (1995).

179. G. T. Rasmussen, T. L. Isenhour, The evaluation of mass spectral search algorithms. *Journal of Chemical Information and Computer Sciences* **19**, 179-186 (1979).

180. S. E. Stein, D. R. Scott, Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry* **5**, 859-866 (1994).

181. S. Kim, I. Koo, J. Jeong, S. Wu, X. Shi, X. Zhang, Compound Identification Using Partial and Semipartial Correlations for Gas Chromatography–Mass Spectrometry Data. *Analytical chemistry* **84**, 6477-6487 (2012).

182. I. Koo, X. Zhang, S. Kim, Wavelet-and fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry. *Analytical chemistry* **83**, 5631-5638 (2011).

183. H. Horai, M. Arita, T. Nishioka, in *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*. (IEEE, 2008), vol. 2, pp. 853-857.

184. I. Koo, S. Kim, X. Zhang, Comparative analysis of mass spectral matching-based compound identification in gas chromatography–mass spectrometry. *Journal of Chromatography A* **1298**, 132-138 (2013).

185. R. G. Sadygov, D. Cociorva, J. R. Yates, Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nature methods* **1**, 195-202 (2004).

186. S. Nachkova, S. Milenkova, P. Bozov, P. Penchev, Interpretive search in a 13 C-NMR spectral library of plant compounds.

187. P. N. Penchev, K.-P. Schulz, M. E. Munk, INFERCNMR: A 13C NMR Interpretive Library Search System. *Journal of chemical information and modeling* **52**, 1513-1528 (2012).

188. A. R. Katritzky, M. Kuanar, S. Slavov, C. D. Hall, M. Karelson, I. Kahn, D. A. Dobchev, Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chemical reviews* **110**, 5714-5789 (2010).

189. K. A. Blinov, Y. D. Smurnyy, T. S. Churanova, M. E. Elyashberg, A. J. Williams, Development of a fast and accurate method of< sup> 13</sup> C NMR chemical shift prediction. *Chemometrics and Intelligent Laboratory Systems* **97**, 91-97 (2009).

190. Y. Binev, J. Aires-de-Sousa, Structure-based predictions of 1H NMR chemical shifts using feed-forward neural networks. *Journal of chemical information and computer sciences* **44**, 940-945 (2004).

191. J. Aires-de-Sousa, M. C. Hemmer, J. Gasteiger, Prediction of 1H NMR chemical shifts using neural networks. *Analytical chemistry* **74**, 80-90 (2002).

192. Y. Binev, M. Corvo, J. Aires-de-Sousa, The impact of available experimental data on the prediction of 1H NMR chemical shifts by neural networks. *Journal of chemical information and computer sciences* **44**, 946-949 (2004).

193. M. Heinonen, A. Rantanen, T. Mielikäinen, J. Kokkonen, J. Kiuru, R. A. Ketola, J. Rousu, FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Communications in Mass Spectrometry* **22**, 3043-3052 (2008).

194. S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann, In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics* **11**, 148 (2010).

195. F. Allen, A. Pon, M. Wilson, R. Greiner, D. Wishart, CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Research*, gku436 (2014).

196. W. L. Fitch, M. McGregor, A. R. Katritzky, A. Lomaka, R. Petrukhin, M. Karelson, Prediction of ultraviolet spectral absorbance using quantitative structure-property relationships. *Journal of chemical information and computer sciences* **42**, 830-840 (2002).

197. C. T. Peng, Prediction of retention indices: V. Influence of electronic effects and column polarity on retention index. *Journal of Chromatography A* **903**, 117-143 (2000).

198. S. S. Liu, Y. Liu, D. Q. Yin, X. D. Wang, L. S. Wang, Prediction of chromatographic relative retention time of polychlorinated biphenyls from the molecular electronegativity distance vector. *Journal of separation science* **29**, 296-301 (2006).

199. L. Liao, H. Mei, J. Li, Z. Li, Estimation and prediction on retention times of components from essential oil of< i> Paulownia tomentosa</i> flowers by molecular electronegativity-distance vector (MEDV). *Journal of Molecular Structure: THEOCHEM* **850**, 1-8 (2008).

200. M. W. Lodewyk, M. R. Siebert, D. J. Tantillo, Computational prediction of 1H and 13C chemical shifts: A useful tool for natural product, mechanistic, and synthetic organic chemistry. *Chemical reviews* **112**, 1839-1862 (2011).

201. S. Kuhn, B. Egert, S. Neumann, C. Steinbeck, Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC bioinformatics* **9**, 400 (2008).

202. M. Elyashberg, K. Blinov, Y. Smurnyy, T. Churanova, A. Williams, Empirical and DFT GIAO quantum‐mechanical methods of 13C chemical shifts prediction: competitors or collaborators? *Magnetic Resonance in Chemistry* **48**, 219-229 (2010).

203. A. Tharatipyakul, S. Numnark, D. Wichadakul, S. Ingsriswang, ChemEx: information extraction system for chemical data curation. *BMC Bioinformatics* **13 Suppl 17**, S9 (2012)10.1186/1471-2105-13-S17-S9).

204. M. Vazquez, M. Krallinger, F. Leitner, A. Valencia, Text mining for drugs and chemical compounds: methods, tools and applications. *Molecular Informatics* **30**, 506-519 (2011).

205. S. H. Bertz, On the complexity of graphs and molecules. *Bulletin of mathematical biology* **45**, 849-855 (1983).

206. S. Nikolic, N. Trinajstic, I. M. Tolic, Complexity of molecules. *Journal of chemical information and computer sciences* **40**, 920-926 (2000).

207. T. El-Elimat, M. Figueroa, B. M. Ehrmann, N. B. Cech, C. J. Pearce, N. H. Oberlies, High-resolution MS, MS/MS, and UV database of fungal secondary metabolites as a dereplication protocol for bioactive natural products. *Journal of natural products* **76**, 1709-1716 (2013).

208. K. F. Nielsen, M. Månsson, C. Rank, J. C. Frisvad, T. O. Larsen, Dereplication of microbial natural products by LC-DAD-TOFMS. *Journal of natural products* **74**, 2338-2348 (2011).

209. D. Staerk, J. R. Kesting, M. Sairafianpour, M. Witt, J. Asili, S. A. Emami, J. W. Jaroszewski, Accelerated dereplication of crude extracts using HPLC-PDA-MS-SPE-NMR: quinolinone alkaloids of Haplophyllum acutifolium. *Phytochemistry* **70**, 1055-1061 (2009); published online EpubMay (10.1016/j.phytochem.2009.05.004).

210. C. A. Motti, M. L. Freckelton, D. M. Tapiolas, R. H. Willis, FTICR-MS and LC-UV/MS-SPE-NMR applications for the rapid dereplication of a crude extract from the sponge Ianthella flabelliformis. *Journal of natural products* **72**, 290-294 (2009); published online EpubFeb 27 (10.1021/np800562m).

211. L. C. Menikarachchi, S. Cawley, D. W. Hill, L. M. Hall, L. Hall, S. Lai, J. Wilder, D. F. Grant, MolFind: a software package enabling HPLC/MS-based identification of unknown chemical structures. *Analytical chemistry* **84**, 9388-9394 (2012).

212. R. P. Bywater, Membrane-spanning peptides and the origin of life. *Journal of theoretical biology* **261**, 407-413 (2009).

213. M. A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, H. Waldmann, Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proceedings of the National Academy of Sciences of the United States of America* **102**, 17272-17277 (2005).

214. J. Batista, J. Bajorath, Chemical database mining through entropy-based molecular similarity assessment of randomly generated structural fragment populations. *Journal of chemical information and modeling* **47**, 59-68 (2007).

215. N. V. Reo, NMR-based metabolomics. *Drug and chemical toxicology* **25**, 375-382 (2002).

216. H. C. Keun, T. J. Athersuch, Nuclear magnetic resonance (NMR)-based metabolomics. *Metabolic Profiling: Methods and Protocols*, 321-334 (2011).

217. A. Mohamed, C. H. Nguyen, H. Mamitsuka, Current status and prospects of computational resources for natural product dereplication: a review. *Briefings in bioinformatics*, bbv042 (2015).

218. R. Stoyanova, T. R. Brown, NMR spectral quantitation by principal component analysis. *NMR in Biomedicine* **14**, 271-277 (2001).

219. C. L. Gavaghan, I. D. Wilson, J. K. Nicholson, Physiological variation in metabolic phenotyping and functional genomic studies: use of orthogonal signal correction and PLS-DA. *FEBS letters* **530**, 191-196 (2002).

220. J. T. Brindle, H. Antti, E. Holmes, G. Tranter, J. K. Nicholson, H. W. L. Bethell, S. Clarke, P. M. Schofield, E. McKilligin, D. E. Mosedale, Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nature medicine* **8**, 1439-1445 (2002).

221. D. Maglott, J. Ostell, K. D. Pruitt, T. Tatusova, Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* **33**, D54-D58 (2005).

222. R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, The European nucleotide archive. *Nucleic acids research*, gkq967 (2010).

223. S. Miyazaki, H. Sugawara, K. Ikeo, T. Gojobori, Y. Tateno, DDBJ in the stream of various biological data. *Nucleic Acids Research* **32**, D31-D34 (2004).

224. J. Gómez, L. J. García, G. A. Salazar, J. Villaveces, S. Gore, A. García, M. J. Martín, G. Launay, R. Alcántara, N. D. T. Ayllón, BioJS: an open source JavaScript framework for biological data visualization. *Bioinformatics*, btt100 (2013).

225. N. Rego, D. Koes, 3Dmol. js: molecular visualization with WebGL. *Bioinformatics*, btu829 (2014).

226. K. Mukhyala, A. Masselot, Visualization of protein sequence features using JavaScript and SVG with pViz. js. *Bioinformatics* **30**, 3408-3409 (2014).

227. J. L. Schmid-Burgk, V. Hornung, BrowserGenome. org: web-based RNA-seq data analysis and visualization. *Nature methods* **12**, 1001-1001 (2015).

228. J. Xia, R. Mandal, I. V. Sinelnikov, D. Broadhurst, D. S. Wishart, MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic acids research* **40**, W127-W133 (2012).

229. D. Tulpan, S. Léger, L. Belliveau, A. Culf, M. Čuperlović-Culf, MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. *BMC bioinformatics* **12**, 400 (2011).

230. J. F. Doreleijers, S. Mading, D. Maziuk, K. Sojourner, L. Yin, J. Zhu, J. L. Markley, E. L. Ulrich, BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *Journal of biomolecular NMR* **26**, 139-146 (2003).

231. T. Vosegaard, jsNMR: an embedded platform-independent NMR spectrum viewer. *Magnetic Resonance in Chemistry* **53**, 285-290 (2015).

232. S. Beisken, P. Conesa, K. Haug, R. M. Salek, C. Steinbeck, SpeckTackle: JavaScript charts for spectroscopy. *Journal of cheminformatics* **7**, 17 (2015).

233. D. H. Douglas, T. K. Peucker, Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization* **10**, 112-122 (1973).

234. D. H. Douglas, T. K. Peucker, Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature. *Classics in Cartography: Reflections on Influential Articles from Cartographica*, 15-28 (2011).