

Chapter 8

Computational Strategies for Biological Interpretation of Metabolomics Data

Jianguo Xia

Abstract Biological interpretation of metabolomics data relies on two basic steps: metabolite identification and functional analysis. These two steps need to be applied in a coordinated manner to enable effective data understanding. The focus of this chapter is to introduce the main computational concepts and workflows during this process. After a general overview of the field, three sections will be presented: the first section will introduce the main computational methods and bioinformatics tools for metabolite identification using spectra from common analytical platforms; the second section will focus on introducing major bioinformatics approaches for functional enrichment analysis of metabolomics data; and the last section will discuss the three main workflows in current metabolomics studies, including the chemometrics approach, the metabolic profiling approach and the more recent chemo-enrichment analysis approach. The chapter ends with summary and future perspectives on computational metabolomics.

Keywords Metabolomics • Chemometrics • Metabolic profiling • Metabolite set enrichment analysis • Chemo-enrichment analysis

Abbreviations

AMDIS	Automated mass spectral deconvolution and identification system
BATMAN	Bayesian automated metabolite analyzer for NMR
GC-MS	Gas chromatography mass spectrometry

J. Xia

Institute of Parasitology, and Department of Animal Science, McGill University,
Sainte Anne de Bellevue, QC H9X 3V9, Canada

Department of Microbiology and Immunology, McGill University,
Montreal, QC H3A 2B4, Canada

e-mail: jeff.xia@mcgill.ca

© Springer International Publishing AG 2017

A. Sussulini (ed.), *Metabolomics: From Fundamentals to Clinical Applications*,
Advances in Experimental Medicine and Biology,
DOI 10.1007/978-3-319-47656-8_8

191

CSF	Cerebral spinal fluid
GO	Gene ontology
GSEA	Gene set enrichment analysis
LC-MS	Liquid chromatography mass spectrometry
MSEA	Metabolite set enrichment analysis
NIST	National Institute of Standards and Technology
NMR	Nuclear magnetic resonance
PCA	Principal component analysis
PLS-DA	Partial least squares discriminant analysis
OPLS-DA	Orthogonal partial least squares discriminant analysis
ORA	Overrepresentation analysis
PCR	Polymerase chain reaction

8.1 Introduction

Measuring metabolites and interpreting their biological relevance within the contexts of different experimental conditions are the primary objective in metabolomics researches. To achieve this objective, two basic steps need to be performed: metabolite identification and functional analysis, with the former providing the necessary inputs for the latter operation. These two steps need to be executed in a coordinated manner to promote efficient biological understanding. However, significant challenges remain in both steps.

The ultimate goal of metabolomics is to achieve comprehensive and high-throughput metabolome measurement. This goal is hampered by at least three major obstacles: (1) small compounds have diverse chemical properties, making it difficult to assay many metabolites simultaneously using a single analytical platform; (2) there is no effective amplification technique available to facilitate detection of low-abundance metabolites (such as using PCR for DNA molecules); and (3) many metabolites lack unique spectral signatures to allow unambiguous compound assignment. Nuclear magnetic resonance (NMR) spectroscopy and gas or liquid chromatography coupled with mass spectrometry (GC- or LC-MS) are commonly used in combination to improve the metabolome coverage. Metabolite identification is mainly performed by searching the spectral features against a reference spectral library. However, searching a comprehensive spectral database often leads to many potential hits with similar matching scores, and researchers often need to manually choose the most probable identities based on the context and domain knowledge. This step represents a key bottleneck in current metabolomics studies. Better algorithms and more context-specific databases are needed to enable high-throughput and high-accurate metabolite identifications.

Knowing compound identities is the first step toward biological interpretation of metabolomics data. The conventional procedure after this step involves manually looking up the metabolites of interest in different compound databases, reading relevant literature, and finally synthesizing the information into a justifiable biological

“story” based on the overall information obtained. This approach is subjective and time-consuming. Over the past decade, many computer-assisted data interpretation strategies have been developed. Among them, functional enrichment analysis using a predefined knowledge database has gained wide acceptance in omics data interpretation. The basic idea is to shift the unit of analysis from a single molecule to groups of functionally related molecules (i.e., those within the same pathway or biological process). This approach directly connects statistical significance with biological interpretation. More advanced algorithms have also been recently implemented that are able to integrate the dependencies and connectivities among different molecules to further reveal the biological insight and to improve system understanding.

Based on their strategies in dealing with metabolite identification and functional analysis, current metabolomics workflows can be summarized into three general categories: the chemometrics approach (also known as untargeted metabolomics), the metabolic profiling approach (also known as targeted or quantitative metabolomics), and the chemo-enrichment analysis approach (Fig. 8.1). The chemometrics approach focuses on identifying and interpreting a subset of spectral features that are found to have changed significantly during the experimental studies; the metabolic profiling approach aims to comprehensively characterize all metabolites in the spectra before subsequent statistical and functional analysis; and the more recent chemo-enrichment analysis approach directly maps spectral features into metabolic pathways/networks and then tests the enrichment of the collective chemical signals generated from these biological processes, which largely avoids the time-consuming step for accurate compound assignment.

This chapter is organized into three sections. The first section introduces the main computational approaches for metabolite identification from common analytical platforms (Fig. 8.1, Step 1); the second section describes the three main bioinformatics approaches for functional enrichment analysis (Fig. 8.1; Step 2); and the last

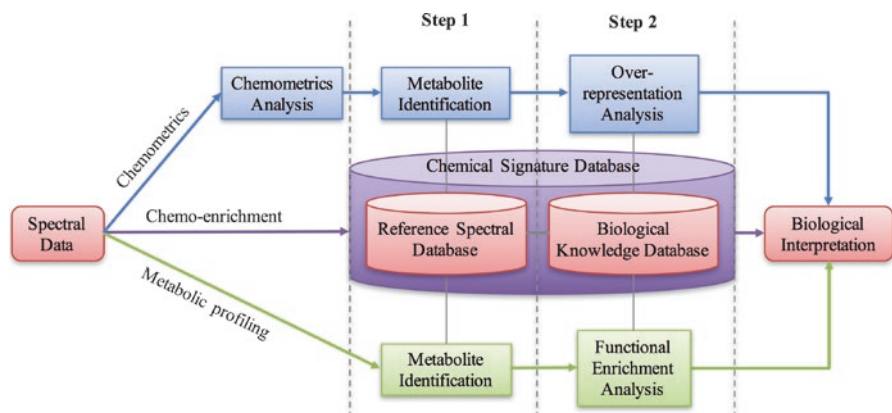


Fig. 8.1 The diagram summarizes the three computational strategies for metabolomics data interpretation: the chemometrics approach (*top*), the metabolic profiling (*bottom*), and the chemo-enrichment analysis (*middle*). The dotted lines delineate the two major steps in the process: metabolite identification and functional analysis. Note that these two steps are integrated into a single one in the chemo-enrichment approach

section compares the three metabolomics workflows for biological interpretation. Each section is further organized under subtitles describing the computational concepts, the available bioinformatics tools, and their main features.

8.2 Metabolite Identification Methods

Although it is possible to determine the identity of a single metabolite *de novo* through labor-intensive NMR or MS-based methods, this approach is generally infeasible in metabolomics in which hundreds to thousands of compound species are measured simultaneously. In practice, compound identification is based on matching features from sample spectra against a reference spectral database, and a closely matched hit will be considered as the putative identity of the corresponding spectral peaks. However, many metabolites do not produce unique, detectable signatures in their NMR or MS spectra to permit unambiguously determination of their identities. The situation is further complicated by peak shifts and overlaps typical in the spectra of complex biological samples. Direct database search tends to yield high percentage of false positives, and further labor-intensive manual refinement is usually necessary. To improve the efficiency of metabolite identification, two general computational strategies have been employed: (1) limiting the search space to only those biologically and biochemically possible candidates by developing more context-specific spectral databases, and (2) improving the peak assignment algorithms by incorporating prior knowledge based on spectral dependencies, biochemical connectivities and biological relationships.

8.2.1 Compound Identification from NMR Spectra

Proton NMR spectroscopy has been widely used in metabolomics studies involving human biofluids. Multiple small-molecule metabolites can be measured simultaneously without prior separation, which greatly simplifies the sample preparation requirements. NMR spectra are highly reproducible, and samples analyzed from one spectrometer will generate near-identical results to those measured on other types of spectrometers. These features have made NMR spectroscopy a platform of choice for large-scale collaborative metabolomics projects.

The Chenomx NMR Suite (Chenomx, Canada) is a widely used metabolomics tool for processing and profiling one-dimensional (1D) proton NMR spectra. The main feature of Chenomx is the integration of a powerful interactive visualization interface with a reference spectral library for over 600 metabolites that are detectable by NMR in common biofluids. Metabolite identification and quantification are achieved through manual peak fitting against those reference spectra. Another widely used commercial tool is the AMIX software package (Bruker Biospin GmbH, Germany), which offers similar features. The company has recently implemented a

software (FoodScreener) that supports automated high-throughput targeted metabolomics profiling for wine, honey, and juice using defined spectra libraries.

Compared to commercial tools, public bioinformatics tools for NMR-based metabolomics tend to focus on spectral alignment, binning and batch processing [1, 2]. They usually lack user-friendly interface or comprehensive spectra libraries to support manual compound identification. As public NMR spectra libraries become increasingly available [3, 4], this situation has begun to change. For instance, the Bayesian automated metabolite analyzer for NMR (BATMAN) is an R package designed for deconvolution and quantification of metabolites from 1D proton NMR spectra of complex mixtures [5, 6]. The Bayesian model incorporates characteristic peak patterns of metabolites and also accounts for peak shifts commonly seen in NMR spectra of biological samples. BATMAN can compute relative concentrations of the compounds together with associated uncertainty estimates using a Markov chain Monte Carlo algorithm. The procedure is computationally intensive and usually requires hours of CPU time to process a single spectrum of common biofluids. Bayesil is a web-based tool that supports automated phasing, referencing, baseline correction, metabolite identification, and quantification for 1D proton NMR metabolomics spectra [7]. The algorithm is implemented based on probabilistic graphical models and a prior knowledge of probable biofluid compositions with built-in support for cerebral spinal fluid (CSF), serum, and plasma. Compared to BATMAN, Bayesil can process a spectrum in a few minutes with high precision and recall. For excessively overlapped NMR spectra of complex biofluid mixtures, two-dimensional (2D) NMR is often used to help resolve spectra ambiguities for metabolite identification purpose. The Bruker AMIX package (Bruker Biospin GmbH, Germany) can also support 2D NMR analysis. The Java desktop application MetaboMiner and the R package rNMR are two public bioinformatics tools for metabolite identification from 2D NMR spectra [8, 9].

8.2.2 *Compound Identification from GC-MS Spectra*

GC-MS offers a high degree of chromatographic resolution and reproducibility. The platform is suitable for measuring volatile, low-molecular mass (<500 Da), and thermally stable compounds such as sugars, fatty acids, and amino acids. For large and polar compounds, chemical derivatization is often employed to improve their volatility and thermal stability. The most commonly used ionization technique in GC-MS is electron ionization, which is very robust and reproducible. The characteristic mass spectral fragmentation patterns can be used to build a spectral library for metabolite identification.

Many software tools are available for metabolite identification and quantification from GC-MS-based metabolomics data. The automated mass spectral deconvolution and identification system (AMDIS) coupled with the National Institute of Standards and Technology (NIST) database is probably the most widely used software package for GC-MS data analysis [10]. The AnalyzerPro (SpectralWorks, UK) and ChromaTOF (LECO, USA) are the two widely used commercial tools for processing and profiling

the GC-MS spectra for metabolomics studies. Compared to NMR-based metabolomics data, more public bioinformatics tools are available for GC-MS spectral processing, deconvolution, alignment, as well as compound identification. Popular tools include BinBase [11], MetaQuant [12], MetabolomeExpress [13], MetaboliteDetector [14], TagFinder [15], *etc.* With the availability of public GC-MS spectral databases [16, 17] and our improved knowledge on the metabolite compositions of common biofluids such as CSF, serum, and urine [18–20], the GC-MS-based metabolomics is expected to be the most promising platform to deliver automated compound identification and quantification for a broad range of biofluids.

8.2.3 Compound Identification from LC-MS Spectra

Compared to GC-MS, LC-MS typically has lower chromatographic resolution and reproducibility. However, LC-MS techniques can access a much broader mass range (100–2000 Da) because volatilization or derivatization is not necessary. LC-MS is also a better choice for separating and identifying polar and nonvolatile compounds. Electrospray ionization and atmospheric pressure chemical ionization are the two most common ionization methods used in LC-MS. Both techniques will generate a molecular ion whose mass can be searched against a spectral database of known metabolites for possible identification. However, due to the finite mass accuracy of the MS equipment and the large number of potential formulas, using mass information alone is usually insufficient for metabolite identification [21].

To address this issue, many bioinformatics tools employ extra information to improve peak assignment and metabolite identification from LC-MS metabolomics data. One approach incorporates known chemical reactions among candidate compounds based on the metabolic pathways/networks to improve annotation, as certain combinations would make more biochemical sense when they are detected together. For instance, the MI-Pack and the ProbMetab are able to use the metabolic pathway information obtained from MetaCyc or KEGG to improve metabolite identification [22, 23]. The second approach takes into consideration of the dependency structures of multiple peaks (isotopologues, adducts, molecular fragments, and multiply charged ions) derived from each metabolite in a LC-MS spectrum to improve peak annotation. The MetAssign tool has implemented this approach [24]. The core algorithms used in these tools are based on graphical models, with most of them using a Bayesian approach to perform probabilistic annotation of metabolites.

8.3 Functional Analysis Approaches

Most metabolites can potentially participate in multiple functional roles within a biological system, and it is difficult to pinpoint the biological processes responsible for the profiles observed in a metabolomics experiment. A biological process is

typically made of a group of molecules. If a biological process is changed in a study, the molecules involved should have a higher potential to be identified as significant by the omics platform. Motivated by this concept, functional analysis has shifted the unit of analysis from a single molecule to a group of functionally related molecules. Instead of testing a single gene or metabolite, researchers now directly evaluate whether a group of molecules (representing a biological process) is consistently changed (enriched). This approach greatly simplifies the omics data interpretation and is more sensitive in detecting subtle but consistent changes occurred in a biological process.

The functional analysis requires two components: a knowledge database defining functionally related molecule groups and a statistical algorithm to perform enrichment tests. The popular gene set enrichment analysis (GSEA) tool is shipped with a comprehensive collection of gene sets in the form of Molecular Signature Database (MSigDB), which greatly facilitates the subsequent development of tools for enrichment analysis [25, 26]. In metabolomics, except the public metabolic pathway databases such as KEGG [27] or MetaCyc [28], a comprehensive collection of functionally related metabolite groups was unavailable until very recently. The first large collection of metabolite sets appeared in 2010 with the publication of the MSEA tool containing >6000 groups of metabolites based on pathways, diseases, genetic variants, and cellular compartments [29]. The other useful resource is the ConceptMetab database containing >16,000 biologically defined metabolite sets developed based on GO, KEGG, and Medical Subject Headings [30]. The ongoing developments of ontologies for systematic metabolite annotations are expected to greatly facilitate the development of enrichment analysis tools for metabolomics [31, 32]. Below I will introduce the three main categories of statistical approaches for functional analysis for metabolomics data: over-representation analysis (ORA), metabolite set enrichment analysis (MSEA), and metabolic pathway/network analysis.

8.3.1 *Over-representation Analysis (ORA)*

The ORA approach is a traditional strategy for enrichment analysis. It starts with a list of metabolites of interest and tests whether certain metabolite groups appear more often than would be expected by random chance. This type of analysis can be performed using Fisher's exact test, a chi-square test, a hypergeometric test, or its binomial approximation. To perform ORA, researchers need to first perform a statistical comparison such as t-tests or ANOVA and then select significant metabolites using a certain threshold or criterion (i.e., adjusted p -values <0.05). Fold change values are also considered sometimes during the selection process.

The ORA approach is very flexible to use and is simple to implement. It has been implemented in many metabolomics tools and databases including MSEA, MBRole, MetaPA, IMPaLA, MPEA, BiNChE, and ConceptMetab [29–31, 33–36]. A common critic of the approach is related to its somewhat arbitrary threshold to decide

whether a metabolite is significant or not. For instance, different cutoffs sometimes lead to different interpretations, and ORA cannot be applied if no significant metabolites are found in a given study. Another limitation is that all metabolites are treated equally after the selection, ignoring their quantitative differences. Despite these shortcomings, ORA remains widely used in omics data interpretation [37].

8.3.2 *Metabolite Set Enrichment Analysis (MSEA)*

The MSEA approach has been developed to address the shortcomings associated with ORA. It directly tests the enrichment of functional groups using the complete concentration data without preselection of significant metabolites. The MSEA is named after the popular GSEA developed for gene expression data interpretation [26]. The original GSEA approach first uses a univariate method to rank all the genes and then tests whether the ranks in the gene set differ from a uniform distribution, using a weighted Kolmogorov-Smirnov test. The p-value for each gene set is calculated via permutation tests. Since then, many different variations of the GSEA have been developed with different performance characteristics [38]. For instance, the GlobalTest method has shown a general improved performance in terms of sensitivity, versatility, and computational efficiency and works especially well if most of the molecules within a group are associated with the phenotype in a modest way [38]. The algorithm is based on a generalized linear model to test whether a group of molecules is significantly associated with a specific phenotype [39].

Several bioinformatics tools have been implemented to support MSEA for metabolomics data. The web-based MSEA program (now part of MetaboAnalyst) is the first tool with such capacity to support functional analysis for quantitative metabolomics data [29, 40]. Like the original GSEA tool, it contains built-in libraries of defined metabolite sets associated with metabolic pathways, diseases, genetic variations, cellular compartments, *etc.* The GlobalTest algorithm is used for quantitative enrichment analysis directly from a metabolite concentration table. Another metabolomics tool with MSEA capacity is the MeltDB, which uses a modified GSEA method against the metabolite sets defined by the KEGG metabolic pathways [41]. With improved functional annotations for metabolite sets such as the ConceptMetab and metabolite ontologies [30, 31], more metabolomics tools with MSEA support will be developed in the near future.

8.3.3 *Metabolic Pathway and Network Analysis*

In the MSEA approach, groups of molecules labeled with biologically meaningful names are used to organize a large body of our current knowledge, making it a popular approach to aid in omics data interpretation. However, this “flat” representation of knowledge followed by enrichment tests based on group memberships ignores

the connectivities and dependencies among molecules as well as the inherent overlaps/hierarchies among different groups. For instance, changes at a central location within a pathway tend to have a larger impact on its overall functions compared to changes at the very downstream. Integrating the functional analysis with pathway/network topology analysis will help improve the accuracy in ranking the resulting list of biological processes.

In gene expression data analysis, the TopGO is probably the first method that integrates knowledge about relationships between different GO terms into calculating the statistical significances to increase the explanatory power of GO enrichment analysis [42]. The signaling pathway impact analysis (SPIA) is another approach that combines the evidence obtained from classical enrichment analysis with a novel type of evidence that utilize the pathway topology to measure the impact on a given pathway [43, 44]. Both approaches have been shown to provide increased sensitivity and specificity when compared to other methods based solely on enrichment analysis. Many more tools have been implemented to take into consideration of pathway topology for enrichment analysis of gene expression data [45]. Applications of similar approaches to metabolomics are currently hampered by two obstacles: firstly, metabolomics typically can only measure a small fraction of any given metabolic pathway at the moment, which greatly limits our ability to evaluate the impact on the overall pathway; secondly, the development of a hierarchical ontology system for metabolite annotation has not been well established to allow easy plug-in by different bioinformatics tools, as is the case of gene ontology system. Therefore, current metabolomics tools focus primarily on enrichment analysis and visualization of metabolic pathways. The web-based tool MetPA (now part of MetaboAnalyst) is the first tool that supports both enrichment analysis and topology analysis within the context of KEGG metabolic pathways [36]. The MetScape is another tool implemented as a Cytoscape plug-in that is able to incorporate prior knowledge of pathways and molecular interactions for metabolomics pathway analysis and network visualization [46].

8.4 Metabolomics Workflows for Biological Interpretation

As indicated in Fig. 8.1, current metabolomics workflows can be largely divided into three general categories based on their strategies in metabolite identification and functional analysis: chemometrics approach, metabolic profiling approach, and chemo-enrichment analysis approach. The chemometrics approach focuses on identifying and interpreting a subset of spectral features that are found to be important within the study. It is relatively high throughput, as only the significant features need to be characterized. This approach is widely used in exploratory metabolomics studies and for discovery of novel biomarkers. A main drawback associated with this approach is the difficulties in biological interpretation, as a limited number of compounds are usually insufficient to pinpoint the underlying biological processes. In contrast to the chemometrics approach, the metabolic profiling approach aims to

characterize all detectable metabolites from the spectral data before subsequent functional analysis. It generally yields better sensitivity, selectivity, and interpretability but is of very limited use for novel biomarker discovery. The main drawback associated with this approach is that the metabolite identification is usually time-consuming and labor intensive. The chemo-enrichment analysis approach has been recently developed to address the limitations associated with both chemometrics and metabolic profiling. It aims to estimate biological activities directly from the spectral features by mapping all possible metabolite matches to metabolic pathways/networks and then comparing the resulting profiles to identify the enriched biological processes.

8.4.1 The Chemometrics Approach

Chemometrics methods are a class of multivariate statistical methods heavily used in analytical chemistry and later metabolomics. These methods are especially useful for analysis and modeling of high-dimensional complex spectral data in untargeted metabolomics, where features (peaks or spectral bins) are highly correlated. The two most commonly used chemometrics methods are principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA). PCA aims to project a high-dimensional data into a low-dimensional space that captures the most variance of the data. The direction of projection is computed based on the data (X) only, without referring to the experimental conditions (Y). PCA is suitable for data overview and to understand the inherent patterns within the data. There is no guarantee that the directions of maximum variance will be the same as the directions of the variance associated with the experimental conditions. In contrast, PLS-DA aims to project a high-dimensional data X into a low-dimensional space that capture the most covariance between X and Y . It is often used to identify the spectral features that are different across experimental conditions. Orthogonal PLS-DA (OPLS-DA) is a variant of PLS-DA which uses orthogonal signal correction to maximize the explained covariance between X and Y on the first component, and the remaining components capture variance in X which is orthogonal to Y [47].

The chemometrics approach is composed of three general steps. A chemometrics method such as PLS-DA or OPLS-DA is first applied to analyze the spectral data to identify significant features associated with the experimental conditions. This step can be performed using several commercial or public tools. The SIMCA-P program (Umetrics, Sweden) is widely used by the metabolomics community. It offers excellent graphic capabilities and comprehensive analysis options for chemometrics methods including PCA, PLS/OPLS-DA, and SIMCA (soft independent modeling of class analogy). MetaboAnalyst is a web-based tool that supports comprehensive metabolomics data processing, normalization, and chemometrics analysis (PCA, PLS-DA, and more recently, Orthogonal PLS-DA [40, 48, 49]. For users who know how to program in R, many R packages are available for chemometrics analysis [50, 51]. After selection of significant spectral features, the second step is to perform

compound identification using the tools and resources as described in Sect. 8.2. In the third step, the list of identified metabolites will be subject to ORA to find out which pathways or biological processes are significantly enriched biological processes are significantly enriched for biological interpretation (Sect. 8.3).

8.4.2 *The Metabolic Profiling Approach*

Metabolic profiling is often used to validate and expand upon results obtained from untargeted analysis. It is also increasingly applied to study variations of metabolite concentrations in relatively well-characterized biofluids such as CSF, blood, urine, etc. Although the process of metabolite identification and quantification is currently a rate-limiting step, this approach offers several distinctive advantages. For instance, metabolic profiling significantly improves statistical power by reducing the number of features from 1000–10,000 of features peaks to hundreds of metabolites. The manual process also largely removes missing values and spectral noises, which greatly facilitates downstream statistical analysis and biomarker discovery.

The biggest advantage of metabolic profiling is the ease of data interpretation. The complete metabolite concentration table can be directly used for MSEA, metabolic pathway, or network analysis using the tools described in Sect. 8.3. The web-based tool MetaboAnalyst provides extensive functions for functional analysis and interpretation for data generated from metabolic profiling approach. Importantly, the metabolite concentration data is very compatible with other omics data and can be analyzed together to help pinpoint the biological pathways involved in the experimental conditions. There are several bioinformatics tools that provide support for integrated analysis of metabolomics data with transcriptomics data. For instance, the MetaCore (Thomson Reuters, USA) allows joint analysis and visual exploration within its comprehensive collections of pathway and network [52]. The public tools IMPaLA and MetScape can accept a list of metabolites and a list of genes for joint analysis and visualization on metabolic networks [34, 46]. INMEX is a web-based tool that supports statistical analysis and joint enrichment analysis for data sets from transcriptomics and metabolic profiling studies [53].

8.4.3 *The Chemo-enrichment Analysis Approach*

The chemo-enrichment analysis approach is a more recent strategy developed to facilitate high-throughput interpretation of metabolomics data generated from high-resolution LC-MS platforms. The key idea is to redefine the metabolite sets, metabolic pathways, or networks using the spectral features (i.e., m/z) of the corresponding metabolites and then test the enrichment of these “collective chemical signals” within the untargeted metabolomics data. Accurate compound identification is not necessary because errors (i.e., incorrect peak assignments) tend to will be randomly

distributed, while the true biological signals will be consistent, which can be detected by testing the enrichment of their collective chemical signals. The chemo-enrichment approach directly connects spectral features with biological interpretations without explicit compound identification. In practice, the metabolite identification is performed *post hoc* for those enriched biological processes of interest. The approach is useful in metabolomics studies for organisms with well-annotated metabolic pathways and networks.

There are a few tools that offer support for chemo-enrichment analysis. The *mummichog* is probably the first bioinformatics tool that implemented the concept [54]. It accepts two lists of spectral peaks (i.e., m/z values) – a significant peak list (i.e., those identified using t -tests) and a reference peak list (all features detected in the MS experiment). The significant peak lists are then searched against a database to find all potential matches to metabolic pathways and networks. The result is compared with those obtained based on peak lists randomly drawn from the reference peaks to compute statistical significance. The tool is available as a Python program. It has been recently implemented in the popular web-based tool XCMS Online to reach a broader audience [55]. MarVis-Pathway is a more recent stand-alone bioinformatics tool with chemo-enrichment analysis feature. It employs a hypergeometric-based approach to evaluate the enrichment of metabolic pathways directly from the untargeted metabolomics data [56].

8.5 Summary and Future Perspectives

This chapter introduces several key concepts and recent developments in computational strategies for metabolomics data interpretation. Compound identification constitutes a major bottleneck in current metabolomics studies. Accurate metabolite identification requires manual intervention and additional laboratory experiments. Advances in both analytical platforms and algorithms are making ways to enable high-throughput data interpretation. Integrating high-resolution analytics, context-specific reference spectral databases, together with advanced algorithms that incorporate chemical and biological information, we will be able to achieve accurate and high-throughput metabolite identification and biological interpretation.

Identification of metabolites (accurately or approximately) is a prerequisite for data interpretation. The list of compounds needs to be put into proper biological context by identifying their roles in metabolic pathways, their interconnectivity with other metabolites, links to genetic variations, or associations with pathophysiological conditions. The group-based functional enrichment analysis has been developed to address this issue. This is an active research area with a wide range of tools and implementations available. Given the current limitations of the knowledge databases and the statistical algorithms, the resulting enrichment p -values should be treated as a ranking system for data exploration and hypothesis generating rather than an absolute cutoff for decision-making purpose.

Compared to transcriptomics, metabolomics is closer to an organism's phenotype and is more sensitive to environmental perturbations. Small compounds represent the final products of complex interactions between the host genetics and environment. The metabolome includes both the endogenous metabolites produced directly by the host organism and the compounds derived from microbial, xenobiotic, dietary, and other exogenous sources. As a result, metabolomics is increasingly applied to study the impact of diet, gut microbiota, and environmental exposures. Developing novel bioinformatics tools and specialized knowledge databases to support these applications are the new frontiers in the current computational metabolomics.

References

1. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. Nmrpipe – a multidimensional spectral processing system based on Unix pipes. *J Biomol NMR*. 1995;6(3):277–93. doi:[10.1007/Bf00197809](https://doi.org/10.1007/Bf00197809).
2. Zhao Q, Stoyanova R, Du SY, Sajda P, Brown TR. HiRes – a tool for comprehensive assessment and interpretation of metabolomic data. *Bioinformatics*. 2006;22(20):2562–4. doi:[10.1093/bioinformatics/btl428](https://doi.org/10.1093/bioinformatics/btl428).
3. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, et al. HMDB 3.0—the human metabolome database in 2013. *Nucleic Acids Res*. 2013;41(Database issue):D801–7. doi:[10.1093/nar/gks1065](https://doi.org/10.1093/nar/gks1065).
4. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, et al. Metabolite identification via the Madison metabolomics consortium database. *Nat Biotechnol*. 2008;26(2):162–4. doi:[10.1038/nbt0208-162](https://doi.org/10.1038/nbt0208-162).
5. Hao J, Liebecke M, Astle W, De Iorio M, Bundy JG, Ebbels TMD. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat Protoc*. 2014;9(6):1416–27. doi:[10.1038/nprot.2014.090](https://doi.org/10.1038/nprot.2014.090).
6. Hao J, Astle W, De Iorio M, Ebbels TMD. BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics*. 2012;28(15):2088–90. doi:[10.1093/bioinformatics/bts308](https://doi.org/10.1093/bioinformatics/bts308).
7. Ravanbakhsh S, Liu P, Bjorndahl TC, Mandal R, Grant JR, Wilson M, et al. Accurate, fully-automated NMR spectral profiling for metabolomics (vol 10, e0124219, 2015). *Plos One*. 2015;10(7). doi:[10.1371/journal.pone.0132873](https://doi.org/10.1371/journal.pone.0132873).
8. Xia JG, Bjorndahl TC, Tang P, Wishart DS. MetaboMiner – semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics*. 2008;9:507. doi:[10.1186/1471-2105-9-507](https://doi.org/10.1186/1471-2105-9-507).
9. Lewis IA, Schommer SC, Markley JL. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn Reson Chem*. 2009;47:S123–6. doi:[10.1002/mrc.2526](https://doi.org/10.1002/mrc.2526).
10. Stein SE. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J Am Soc Mass Spectr*. 1999;10(8):770–81. doi:[10.1016/S1044-0305\(99\)00047-1](https://doi.org/10.1016/S1044-0305(99)00047-1).
11. Fiehn O, Wohlgemuth G, Scholz M. Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata. *Lect Notes Comput Sci*. 2005;3615:224–39. doi:[10.1007/11530084_18](https://doi.org/10.1007/11530084_18).
12. Bunk B, Kucklick M, Jonas R, Munch R, Schober M, Jahn D, et al. MetaQuant: a tool for the automatic quantification of GC/MS-based metabolome data. *Bioinformatics*. 2006;22(23):2962–5. doi:[10.1093/bioinformatics/btl526](https://doi.org/10.1093/bioinformatics/btl526).

13. Carroll AJ, Badger MR, Millar AH. The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC Bioinformatics*. 2010;11. doi:[10.1186/1471-2105-11-376](https://doi.org/10.1186/1471-2105-11-376).
14. Hiller K, Hangebrauk J, Jager C, Spura J, Schreiber K, Schomburg D. MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Anal Chem*. 2009;81(9):3429–39. doi:[10.1021/ac802689c](https://doi.org/10.1021/ac802689c).
15. Luedemann A, Strassburg K, Erban A, Kopka J. TagFinder for the quantitative analysis of gas chromatography – mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics*. 2008;24(5):732–7. doi:[10.1093/bioinformatics/btn023](https://doi.org/10.1093/bioinformatics/btn023).
16. Schauer N, Steinhäuser D, Strelkov S, Schomburg D, Allison G, Moritz T, et al. GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett*. 2005;579(6):1332–7. doi:[10.1016/j.febslet.2005.01.029](https://doi.org/10.1016/j.febslet.2005.01.029).
17. Kind T, Wohlgemuth G, Lee DY, Lu Y, Palazoglu M, Shahbaz S, et al. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal Chem*. 2009;81(24):10038–48. doi:[10.1021/ac9019522](https://doi.org/10.1021/ac9019522).
18. Wishart DS, Lewis MJ, Morrissey JA, Flegel MD, Jeroncic K, Xiong YP, et al. The human cerebrospinal fluid metabolome. *J Chromatogr B*. 2008;871(2):164–73. doi:[10.1016/j.jchromb.2008.05.001](https://doi.org/10.1016/j.jchromb.2008.05.001).
19. Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, et al. The human serum metabolome. *Plos One*. 2011;6(2). doi:[10.1371/journal.pone.0016957](https://doi.org/10.1371/journal.pone.0016957).
20. Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, Knox C, et al. The human urine metabolome. *Plos One*. 2013;8(9). doi:[10.1371/journal.pone.0073076](https://doi.org/10.1371/journal.pone.0073076).
21. Kind T, Fiehn O. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*. 2006;7. doi:[10.1186/1471-2105-7-234](https://doi.org/10.1186/1471-2105-7-234).
22. Weber RJM, Viant MR. MI-pack: increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. *Chemometr Intell Lab*. 2010;104(1):75–82. doi:[10.1016/j.chemolab.2010.04.010](https://doi.org/10.1016/j.chemolab.2010.04.010).
23. Silva RR, Jourdan F, Salvanha DM, Letisse F, Jamin EL, Guidetti-Gonzalez S, et al. ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics*. 2014;30(9):1336–7. doi:[10.1093/bioinformatics/btu019](https://doi.org/10.1093/bioinformatics/btu019).
24. Daly R, Rogers S, Wandy J, Jankevics A, Burgess KEV, Breitling R. MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics*. 2014;30(19):2764–71. doi:[10.1093/bioinformatics/btu370](https://doi.org/10.1093/bioinformatics/btu370).
25. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739–40. doi:[10.1093/bioinformatics/btr260](https://doi.org/10.1093/bioinformatics/btr260).
26. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–50. doi:[10.1073/pnas.0506580102](https://doi.org/10.1073/pnas.0506580102).
27. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*. 2014;42(Database issue):D199–205. doi:[10.1093/nar/gkt1076](https://doi.org/10.1093/nar/gkt1076).
28. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. 2016;44(D1):D471–80. doi:[10.1093/nar/gkv1164](https://doi.org/10.1093/nar/gkv1164).
29. Xia JG, Wishart DS. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res*. 2010;38:W71–7. doi:[10.1093/nar/gkq329](https://doi.org/10.1093/nar/gkq329).
30. Cavalcante RG, Patil S, Weymouth TE, Bendinskas KG, Karnovsky A, Sartor MA. ConceptMetab: exploring relationships among metabolite sets to identify links among biomedical concepts. *Bioinformatics*. 2016;32(10):1536–43. doi:[10.1093/bioinformatics/btw016](https://doi.org/10.1093/bioinformatics/btw016).
31. Moreno P, Beisken S, Harsha B, Muthukrishnan V, Tudose I, Dekker A, et al. BiNChE: a web tool and library for chemical enrichment analysis based on the ChEBI ontology. *BMC Bioinformatics*. 2015;16. doi:[10.1186/s12859-015-0486-3](https://doi.org/10.1186/s12859-015-0486-3).

32. Hastings J, Chepelev L, Willighagen E, Adams N, Steinbeck C, Dumontier M. The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. *Plos One*. 2011;6(10):e25513. doi:[10.1371/journal.pone.0025513](https://doi.org/10.1371/journal.pone.0025513).
33. Kankainen M, Gopalacharyulu P, Holm L, Oresic M. MPEA-metabolite pathway enrichment analysis. *Bioinformatics*. 2011;27(13):1878–9. doi:[10.1093/bioinformatics/btr278](https://doi.org/10.1093/bioinformatics/btr278).
34. Kamburov A, Cavill R, Ebbels TMD, Herwig R, Keun HC. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*. 2011;27(20):2917–8. doi:[10.1093/bioinformatics/btr499](https://doi.org/10.1093/bioinformatics/btr499).
35. Chagoyen M, Pazos F. MBRole: enrichment analysis of metabolomic data. *Bioinformatics*. 2011;27(5):730–1. doi:[10.1093/bioinformatics/btr001](https://doi.org/10.1093/bioinformatics/btr001).
36. Xia JG, Wishart DS. MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics*. 2010;26(18):2342–4. doi:[10.1093/bioinformatics/btq418](https://doi.org/10.1093/bioinformatics/btq418).
37. da Huang W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1–13. doi:[10.1093/nar/gkn923](https://doi.org/10.1093/nar/gkn923).
38. Tarca AL, Bhatti G, Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *Plos One*. 2013;8(11). doi:[10.1371/journal.pone.0079217](https://doi.org/10.1371/journal.pone.0079217).
39. Goeman JJ, Van De Geer SA, Van Houwelingen HC. Testing against a high dimensional alternative. *JRStatSocSerB(StatMethodol)*. 2006;68(3):477–93. doi:[10.1111/j.1467-9868.2006.00551.x](https://doi.org/10.1111/j.1467-9868.2006.00551.x).
40. Xia JG, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0-making metabolomics more meaningful. *Nucleic Acids Res*. 2015;43(W1):W251–7. doi:[10.1093/nar/gkv380](https://doi.org/10.1093/nar/gkv380).
41. Persicke M, Ruckert C, Plassmeier J, Stutz LJ, Kessler N, Kalinowski J, et al. MSEA: metabolite set enrichment analysis in the MeltDB metabolomics software platform: metabolic profiling of *Corynebacterium glutamicum* as an example. *Metabolomics*. 2012;8(2):310–22. doi:[10.1007/s11306-011-0311-6](https://doi.org/10.1007/s11306-011-0311-6).
42. Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*. 2006;22(13):1600–7. doi:[10.1093/bioinformatics/btl140](https://doi.org/10.1093/bioinformatics/btl140).
43. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, et al. A novel signaling pathway impact analysis. *Bioinformatics*. 2009;25(1):75–82. doi:[10.1093/bioinformatics/btn577](https://doi.org/10.1093/bioinformatics/btn577).
44. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, et al. A systems biology approach for pathway level analysis. *Genome Res*. 2007;17(10):1537–45. doi:[10.1101/gr.6202607](https://doi.org/10.1101/gr.6202607).
45. Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, et al. Methods and approaches in the topology-based analysis of biological pathways. *Front Physiol*. 2013;4. doi:[10.3389/fphys.2013.00278](https://doi.org/10.3389/fphys.2013.00278).
46. Karnovsky A, Weymouth T, Hull T, Tarcea VG, Scardoni G, Laudanna C, et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics*. 2012;28(3):373–80. doi:[10.1093/bioinformatics/btr661](https://doi.org/10.1093/bioinformatics/btr661).
47. Bylesjo M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J Chemometr*. 2006;20(8–10):341–51. doi:[10.1002/cem.1006](https://doi.org/10.1002/cem.1006).
48. Xia JG, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0-a comprehensive server for metabolomic data analysis. *Nucleic Acids Res*. 2012;40(W1):W127–33. doi:[10.1093/nar/gks374](https://doi.org/10.1093/nar/gks374).
49. Xia JG, Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res*. 2009;37:W652–60. doi:[10.1093/nar/gkp356](https://doi.org/10.1093/nar/gkp356).
50. Thevenot EA, Roux A, Xu Y, Ezan E, Junot C. Analysis of the human adult urinary metabolome variations with Age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *J Proteome Res*. 2015;14(8):3322–35. doi:[10.1021/acs.jproteome.5b00354](https://doi.org/10.1021/acs.jproteome.5b00354).
51. Edoardo G, Francesca C, Dimitrios S, Andrea S, Michela G, Jose MG-M, et al. muma, an R package for metabolomics univariate and multivariate statistical analysis. *Curr Metabolomics*. 2013;1(2):180–9. doi:[10.2174/2213235X11301020005](https://doi.org/10.2174/2213235X11301020005).
52. Jennen D, Ruiz-Aracama A, Magkoufopoulou C, Peijnenburg A, Lommen A, van Delft J, et al. Integrating transcriptomics and metabolomics to unravel modes-of-action of 2,3,7,8-tetrachlo

- rodibenzo-p-dioxin (TCDD) in HepG2 cells. *BMC Syst Biol.* 2011;5:139. doi:[10.1186/1752-0509-5-139](https://doi.org/10.1186/1752-0509-5-139).
53. Xia J, Fjell CD, Mayer ML, Pena OM, Wishart DS, Hancock RE. INMEX—a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res.* 2013;41(Web Server issue):W63–70. doi:[10.1093/nar/gkt338](https://doi.org/10.1093/nar/gkt338).
54. Li SZ, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, et al. Predicting network activity from high throughput metabolomics. *Plos Comput Biol.* 2013;9(7). doi:[10.1371/journal.pcbi.1003123](https://doi.org/10.1371/journal.pcbi.1003123).
55. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal Chem.* 2012;84(11):5035–9. doi:[10.1021/ac300698c](https://doi.org/10.1021/ac300698c).
56. Kaefer A, Landesfeind M, Feussner K, Mosblech A, Heilmann I, Morgenstern B, et al. MarVis-Pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data. *Metabolomics.* 2015;11(3):764–77. doi:[10.1007/s11306-014-0734-y](https://doi.org/10.1007/s11306-014-0734-y).