UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

RICARDO ROBERTO DA SILVA

**Anotação probabilística de perfis de metabólitos obtidos por cromatografia
líquida acoplada a espectrometria de massas**

Ribeirão Preto
2014

RICARDO ROBERTO DA SILVA

Anotação probabilística de perfis de metabólitos obtidos por cromatografia líquida acoplada a espectrometria de massas

Tese apresentada ao Programa de Pós-Graduação em Genética da Faculdade de Medicina de Ribeirão Preto, da Universidade de São Paulo, para obtenção do Título de Doutor em Ciências.

Área de Concentração: Genética

Orientador: Prof. Dr. Ricardo Zorzetto
Nicolliello Vêncio

Ribeirão Preto

2014

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Silva, Ricardo Roberto.

Anotação probabilística de perfis de metabólitos obtidos por cromatografia líquida acoplada a espectrometria de massas / Ricardo Roberto da Silva; orientador Ricardo Zorzetto Nicoliello Vêncio. - Ribeirão Preto, 2014.

90 f. : il.

Tese (Doutorado)--Universidade de São Paulo, 2014.

1 – Bioinformática. 2 – Metabolômica. 3 – Estatística Bayesiana. 4 - Espectrometria de Massas

SILVA, R. R. Anotação probabilística de perfis de metabólitos obtidos por cromatografia líquida acoplada a espectrometria de massas. Tese apresentada ao Programa de Pós-Graduação em Genética Faculdade de Medicina de Ribeirão Preto, da Universidade de São Paulo, para obtenção do Título de Doutor em Ciências.

Aprovado em:

Banca Examinadora

Prof(a). Dr(a). _____ Instituição: _____
Julgamento: _____ Assinatura: _____

Prof(a). Dr(a). _____ Instituição: _____
Julgamento: _____ Assinatura: _____

Prof(a). Dr(a). _____ Instituição: _____
Julgamento: _____ Assinatura: _____

Prof(a). Dr(a). _____ Instituição: _____
Julgamento: _____ Assinatura: _____

Prof(a). Dr(a). _____ Instituição: _____
Julgamento: _____ Assinatura: _____

Agradecimentos

Ao Prof. Dr. Ricardo N. Z. Vêncio pela orientação, pelas constantes discussões, por todo o apoio e incentivo oferecido durante o desenvolvimento deste projeto, mas principalmente por me ensinar a contestar os dogmas ou verdades estabelecidas e sempre almejar um alto padrão de desempenho e qualidade para no meu trabalho. Muito obrigado, espero levar estas lições para minha vida profissional.

À colega Madeleine Ernst pela colaboração, pela confiança e pelas oportunidades oferecidas.

Ao Dr. Fabien Jourdan pela paciência e empenho na discussão das análises de grafos. Ao Dr. Fabien Letisse e ao Dr. Emilien L. Jamin pela ajuda nos conceitos de química e sugestões para o formato de saída das análises.

Ao *The European Bioinformatics Institute* (EBI) por catalisar a cooperação com nossos colaboradores franceses, em especial ao Prof. Dr. Reza Salek pela constante disposição em auxiliar.

Ao Prof. Dr. Norberto Peporine Lopes pela confiança e pelas oportunidades oferecidas.

Gostaria de agradecer o Prof. Aposentado pelo IFUSP Kiyomi Koide pela honra de poder aprender um pouco de espectrometria com alguém que participou do desenvolvimento experimental de um espectrômetro de massas na USP.

Ao Prof. Dr. Carlos Alberto Labate pelas oportunidades oferecidas e aos colegas do Laboratório Max Feffer de Genética de Plantas – ESALQ – USP, Fabricio Edgar de Moraes, Dra Ilara Budzinski e Dra Simone Guidetti-Gonzalez pela geração dos dados de cana-de-açúcar utilizados nesta tese.

Aos colegas integrantes, ex-integrantes e vizinhos do Laboratório de Processamento de Informação Biológica (LabPIB), Ricardo C. Waldemarin, Diego Martinez Salvanha, Leonardo Rippel, Bruna Renata Silva Corrêa, Bruno Della Vecchia, Marcos Abraão S. Fonseca, Gabriela Guardia, Gabriela Persinoti, Julio Cesar O. Garcia, Atlas Khan, Eliane Z. Traldi e Danilo C. Almeida-e-Silva pela amizade, convivência e colaborações.

Aos colegas Felipe ten Caten, José Vicente Gomes Filho, Marjorie Cornejo Pontelli pela amizade e convivência.

Novamente ao amigo Diego Martinez Salvanha pela constante disposição em auxiliar, pela colaboração no desenvolvimento de ferramentas de visualização para este trabalho e principalmente pelos exemplos de sabedoria que vivenciei ao seu lado.

Ao amigo Ricardo C. Waldemarin pelas consultorias particulares em computação, pela parceria no blog e nas competições, e pelo exemplo de paixão pura pela ciência.

Às secretárias do departamento Susie, Silvia, pela atenção e pelos auxílios prestados.

Agradeço à assistência social do campus da USP de Ribeirão Preto pela utilização da moradia estudantil e pela oportunidade de aprendizado da convivência em grupo.

Agradeço ao Departamento de Genética e à Faculdade de Medicina de Ribeirão Preto pela estrutura oferecida para minha formação na pós-graduação.

Agradeço à CAPES pela bolsa de Doutorado no início do programa e a FAPESP pela continuidade da bolsa de Doutorado oferecida para a realização deste projeto e todo o suporte financeiro.

À CAPES e FAEPA pelo apoio financeiro dado ao projeto.

À minha esposa Lívia, meu esteio, meu porto seguro por todo o amor e companheirismo.

Ao meu irmão Diego pelo exemplo de coragem, e a lição de erguer a cabeça e encarar os desafios da vida de frente.

Aos meus pais Carlos Roberto da Silva e Maria Aparecida Moreira da Silva, pelas lições de honestidade, humildade e perseverança que universidade nenhuma ensina e que vão nortear minha vida em todo lugar.

Aos meus grandes amigos Diego “de Nilda” e Charles por mostrarem que nem os mais altos morros de Minas separam uma grande amizade.

A toda a minha família que sempre suportou as minhas decisões e sempre esteve ao meu lado a todo momento.

A todos que de forma direta ou indireta colaboraram com a realização deste trabalho. Muito obrigado!

Este trabalho é dedicado a consciência coletiva formada pela rede mundial de computadores, onde não há recompensa material, mas marcas que ecoam pela eternidade.

RESUMO

SILVA, R. R. **Anotação probabilística de perfis de metabólitos obtidos por cromatografia líquida acoplada a espectrometria de massas**. 2014. 90 f. Tese (Doutorado) - Faculdade de Medicina de Ribeirão Preto, da Universidade de São Paulo, Ribeirão Preto, 2014.

A metabolômica é uma ciência emergente na era pós-genômica que almeja a análise abrangente de pequenas moléculas orgânicas em sistemas biológicos. Técnicas de cromatografia líquida acoplada a espectrometria de massas (LC-MS) figuram como as abordagens de amostragem mais difundidas. A extração e detecção simultânea de metabólitos por LC-MS produz conjuntos de dados complexos que requerem uma série de etapas de pré-processamento para que a informação possa ser extraída com eficiência e precisão. Para que as abordagens de perfil metabólico não direcionado possam ser efetivamente relacionadas às alterações de interesse no metabolismo, é estritamente necessário que os metabólitos amostrados sejam anotados com confiabilidade e que a sua inter-relação seja interpretada sob a pressuposição de uma amostra conectada do metabolismo. Diante do desafio apresentado, a presente tese teve por objetivo desenvolver um arcabouço de *software*, que tem como componente central um método probabilístico de anotação de metabólitos que permite a incorporação de fontes independentes de informações espectrais e conhecimento prévio acerca do metabolismo. Após a classificação probabilística, um novo método para representar a distribuição de probabilidades *a posteriori* em forma de grafo foi proposto. Uma biblioteca de métodos para o ambiente R, denominada **ProbMetab (Probilistic Metabolomics)**, foi criada e disponibilizada de forma aberta e gratuita. Utilizando o *software* ProbMetab para analisar um conjunto de dados *benchmark* com identidades dos compostos conhecidas de antemão, demonstramos que até 90% das identidades corretas dos metabólitos estão presentes entre as três maiores probabilidades. Portanto, pode-se enfatizar a eficiência da disponibilização da distribuição de probabilidades *a posteriori* em lugar da classificação simplista usualmente adotada na área de metabolômica, em que se usa apenas o candidato de maior probabilidade. Numa aplicação à dados reais, mudanças em uma via metabólica reconhecidamente relacionada a estresses abióticos em plantas (Biossíntese de Flavona e Flavonol) foram automaticamente detectadas em dados de cana-de-açúcar, demonstrando a importância de uma visualização centrada na distribuição *a posteriori* da rede de anotações dos metabólitos.

Palavras-chave: Bioinformática, Metabolômica, Estatística Bayesiana, Espectrometria de Massas.

ABSTRACT

SILVA, R. R. **Probabilistic annotation of metabolite profiles obtained by liquid chromatography coupled to mass spectrometry**. 2014. 90 f. Tese (Doutorado) - Faculdade de Medicina de Ribeirão Preto, da Universidade de São Paulo, Ribeirão Preto, 2014.

Metabolomics is an emerging science field in the post-genomic era, which aims at a comprehensive analysis of small organic molecules in biological systems. Techniques of liquid chromatography coupled to mass spectrometry (LC-MS) figure as the most widespread approaches to metabolomics studies. The metabolite detection by LC-MS produces complex data sets, that require a series of preprocessing steps to ensure that the information can be extracted efficiently and accurately. In order to be effectively related to alterations in the metabolism of interest, is absolutely necessary that the metabolites sampled by untargeted metabolic profiling approaches are annotated with reliability and that their relationship are interpreted under the assumption of a connected metabolism sample. Faced with the presented challenge, this thesis developed a software framework, which has as its central component a probabilistic method for metabolite annotation that allows the incorporation of independent sources of spectral information and prior knowledge about metabolism. After the probabilistic classification, a new method to represent the *a posteriori* probability distribution in the form of a graph has been proposed. A library of methods for R environment, called ***ProbMetab*** (**Probabilistic Metabolomics**), was created and made available as an open source software. Using the ProbMetab software to analyze a set of benchmark data with compound identities known beforehand, we demonstrated that up to 90% of the correct metabolite identities were present among the top-three higher probabilities, emphasizing the efficiency of *a posteriori* probability distribution display, in place of a simplistic classification with only the most probable candidate, usually adopted in the field of metabolomics. In an application to real data, changes in a known metabolic pathway related to abiotic stresses in plants (Biosynthesis of Flavone and Flavonol) were automatically detected on sugar cane data, demonstrating the importance of a view centered on the posterior distribution of metabolite annotation network.

Keywords: Bioinformatics, Metabolomics, Bayesian Statistics, Mass Spectrometry.

Lista de ilustrações

Figura 1 - Representação esquemática de um espectrômetro de massas.....	22
Figura 2 - Estrutura dos dados obtida pelos detectores em espectrometria de massas.....	25
Figura 3 - Efeito do aumento da resolução na diferenciação de íons formados pelo padrão isotópico do peptídeo bombesina (código UniProt P84214).....	26
Figura 4 - Decomposição de um espectro de massas em diferentes componentes de sinais detectados.....	29
Figura 5 - Etapas da análise de dados em dados de cromatografia líquida acoplada a espectrometria de massas.....	32
Figura 6 - Representação esquemática da atribuição de picos de massa para fórmulas de compostos conhecidos, considerando sua inter-relação em vias metabólicas através de um modelo probabilístico.....	43
Figura 7 - Exemplo da codificação das reações metabólicas de compêndios de reações bioquímicas caracterizadas.....	46
Figura 8 - Representação conceitual da visualização de uma rede para anotação probabilística do metaboloma.....	52
Figura 9 - Representação esquemática do escopo do pacote ProbMetab na análise de dados de espectrometria.....	55
Figura 10 - Captura de tela do arquivo de saída no formato tabular gerado pelo pacote ProbMetab.....	57
Figura 11 - Representação da tabela de classificação de picos final exportada como uma rede ponderada por correlações sobreposta a uma rede de reações bioquímicas.....	65
Figura 12 - Visão parcial da rede de possíveis reações bioquímicas e correlações (valor absoluto maior que 0.7) sobrepostas, geradas pelo pacote ProbMetab.....	69
Figura 13 - Representação da via metabólica da Biossíntese de Flavonóides (map00941) automaticamente exportada do KEGG para o Cytoscape pelo pacote ProbMetab.....	71
Figura 14 - Espectros selecionados dos maiores (acima) e menores (abaixo) valores absolutos da razão intensidade seca/intensidade em irrigação.....	75

Lista de tabelas

Tabela 1 - Exemplos de bases de dados do metabolismo.....	16
Tabela 2 - Principais tecnologias utilizadas para medição do metaboloma. Adaptado de (WISHART, 2010).....	19
Tabela 3 - Performance de espectrômetros de massa modernos na quantificação de pequenas moléculas. Extraída de (ZHOU et al., 2012).....	23
Tabela 4 - Átomos e suas proporções isotópicas relevantes para espectrometria de massas. Adaptado de (SMEDSGAARD, 2007).....	27
Tabela 5 - Íons e agrupamentos mais frequentemente formados em ESI. Adaptado de (SMEDSGAARD, 2007).....	28
Tabela 6 - Os quatro níveis de confiança para identificação de metabólitos definidos pela Iniciativa de Padrões da Comunidade de Metabolômica. Adaptado de (DUNN et al., 2012)..	36
Tabela 7 - Anotação de picos para o pré-processamento com pacote xcms.....	59
Tabela 8 - Comparação da adição de diferentes fontes de informação ao modelo probabilístico.....	60
Tabela 9 - Exemplos da atribuição de probabilidades para diferentes componentes do modelo, com a identidade correta destacada em negrito.....	61
Tabela 10 - Proporção cumulativa da posição da identidade correta em diferentes modelos.....	63
Tabela 11 - Visão parcial da tabela de probabilidades exportada pelo pacote ProbMetab, mostrando em amarelo as possíveis identidades associadas à síntese de Flavonóides.....	73
Tabela 12 - Espectros selecionados dos maiores (acima) e menores (abaixo) valores absolutos da razão intensidade seca/intensidade em irrigação mostrando a fórmula sugerida pela anotação do íon.....	74

Sumário

1-INTRODUÇÃO.....	12
1.1-O metaboloma.....	12
1.2-Medição experimental do metaboloma.....	16
1.3-Aquisição e representação de dados em espectrometria de massas.....	23
1.4-Pré-processamento de dados em espectrometria de massas.....	29
2-HIPÓTESE DO TRABALHO.....	38
3-OBJETIVOS.....	39
3.1-Objetivos específicos.....	39
4-MÉTODOS.....	40
4.1-Anotação dos íons.....	40
4.2-Descrição do Modelo Probabilístico.....	41
4.3-Representação do metabolismo.....	49
4.4-Algoritmo de amostragem e implementação.....	54
5-RESULTADOS E DISCUSSÃO.....	55
5.1-Construção de uma biblioteca para o ambiente R.....	55
5.2-Validação e desempenho do modelo.....	58
5.3-Representação abrangente dos resultados.....	63
5.4-Aplicação a dados originais: análises de espectros de massa de cana-de-açúcar.....	66
6-CONCLUSÕES E PERSPECTIVAS.....	77
6.1-Implementação e validação do modelo.....	77
6.2-Busca automatizada de alterações no metabolismo.....	79
7-REFERÊNCIAS BIBLIOGRÁFICAS.....	81
8-ANEXOS.....	90

1- INTRODUÇÃO

1.1- O metaboloma

A palavra metabolismo vem do Grego *metabole* e significa mudança ou transformação (ROESSNER, 2007). O metabolismo pode ser dividido em primário e secundário. O metabolismo primário é relacionado primordialmente a energia e produção de estruturas essenciais para a célula e compartilha muitas similaridades entre espécies, com muitos metabólitos amplamente difundidos na natureza. O metabolismo secundário está relacionado com a produção de metabólitos mais especializados, alguns únicos para determinadas células, tecidos ou espécies e requerem genes ou vias metabólicas completas especificamente para sua produção (ROESSNER, 2007). Estes metabólitos têm suas funções muitas vezes desconhecidas, atuando em vias metabólicas que, em última instância, determinam o papel de sua unidade biológica.

O metaboloma pode ser definido como o conjunto completo de pequenas moléculas (<1500 Da), os metabólitos, encontrados em uma célula, órgão ou organismo específico (WISHART et al., 2012). O sequenciamento de genomas completos de um grande número de espécies tem possibilitado a estimativa do número de metabólitos em seus respectivos metabolomas. Porém, a falta de anotações completas para estes genomas impede que muitas reações sejam conhecidas o que torna as estimativas apenas aproximações distantes. Um organismo pode potencialmente catalisar centenas de reações, entretanto, as reações são reguladas no espaço e no tempo. As três principais formas de regulação são: controle dos níveis enzimáticos, controle da atividade enzimática e controle da captação, e transporte (VOET; VOET, 2011).

Reações bioquímicas são executadas para transformar e modificar a natureza química dos metabólitos. Frequentemente estas reações só são possíveis pela presença de catalisadores específicos, as enzimas, estruturas protéicas altamente especializadas. Um catalisador aumenta a taxa ou velocidade em que uma reação química ocorre, sem ser modificado pela mesma no decorrer do processo (VOET; VOET, 2011). As enzimas modificam as taxas de

uma reação, mas não afetam seu equilíbrio, atuando na diminuição da barreira energética de uma reação, aumentando assim a fração de moléculas que possuem energia suficiente para atingir o estado de transição (ROESSNER, 2007).

O controle dos níveis enzimáticos é complexo e envolve diferentes etapas de regulação, como a síntese protéica e a taxa de degradação da enzima. Os metabólitos podem interagir com as enzimas e exercerem controles positivos e negativos, regulando a utilização de outros metabólitos na mesma via. Muitas enzimas existem em conformações ativa ou inativa. Estas enzimas são frequentemente proteínas de múltiplas subunidades, com sítios alostéricos para ligação e ativação/inativação. Uma das principais formas de controle do fluxo de metabólitos se dá pela separação de reações metabólicas em diferentes compartimentos, que promovem a separação espacial e temporal (VOET; VOET, 2011).

Uma forma de compartimentalização das reações ocorre quando agregados de enzimas envolvidas em uma via metabólica se formam, os quais são denominados canais de substratos (*substrate channeling*) ou *metabolons*. Os complexos enzimáticos permitem a transferência direta de uma série de intermediários biossintéticos entre sítios catalíticos sem a liberação dos metabólitos intermediários no solvente celular (AGAPAKIS et al., 2012). É imprescindível conhecer os mecanismos de compartimentação do metabolismo, pois estes são responsáveis por delimitar que vias se interconectam, por quais metabólitos e por consequência definir a topologia da rede do metabolismo global.

Os níveis dos metabólitos são controlados pela regulação e atividade de enzimas. Modelos de reconstrução do metabolismo em escala genômica podem ser utilizados para extrair padrões gerais do metabolismo. Nielsen (2003) reportou o padrão de distribuição do número de metabólitos envolvidos em reações bioquímicas catalogadas para seis reconstruções. Nos seis genomas avaliados cerca de 30% dos metabólitos conhecidos são envolvidos em apenas duas reações, enquanto 12% participam em mais que dez reações e cerca de 4% em mais que 20 reações. Além disso, muitas reações envolvem mais que um único substrato e um único produto (mais de 67% das reações catalogadas, por exemplo, para a levedura *Saccharomyces cerevisiae*) garantindo um alto grau de conectividade do metaboloma. Devido ao alto nível de acoplamento de uma grande diversidade de reações, pequenas perturbações no proteoma podem resultar em mudanças significativas no nível de muitos metabólitos, uma vez que os

produtos de algumas reações são substratos de outras e as mudanças nos níveis são propagadas rapidamente na rede (NIELSEN, 2003; ROESSNER, 2007).

Uma característica geral do metabolismo, quando descrito como uma rede complexa, é seu caráter de “mundo pequeno”, isto é, dois metabólitos quaisquer em um sistema podem ser conectados por caminhos relativamente curtos entre os metabólitos intermediários, o que habilita que mudanças no nível de um dado metabólito atinjam outros metabólitos na rede de uma maneira rápida (WAGNER; FELL, 2001).

A natureza convoluta do metabolismo celular, onde o mesmo metabólito participa de várias vias metabólicas diferentes dificulta o relacionamento direto entre um gene e um metabólito (KIND et al., 2009). A determinação de limites biológicos, ou seja, a capacidade de um organismo ou célula produzir um determinado composto, é essencial para a determinação do metaboloma. Entretanto, é preciso considerar o potencial catalítico celular e as conexões às demais células ou ambiente externo. Estes limites permitiriam o posicionamento do composto em uma hierarquia de organização de célula, tecido, órgão, organismo e táxon, que habilitariam pesquisas restritas a uma unidade biológica, reduzindo o trabalho de inspeção manual na busca de metabólitos a partir de dados espectrais, como veremos abaixo.

Na avaliação de repositórios que compilam os esforços de reconstrução do metabolismo como MetaCyc (*Encyclopedia of Metabolic Pathways*) ou KEGG (*Kyoto Encyclopedia of Genes and Genomes*) (KANEHISA; GOTO, 2000; CASPI et al., 2011), existe a constatação que o número de entidades bioquímicas dos tipos enzima, reação e composto crescem mais rápido que o número de vias metabólicas, indicando a necessidade de conectar as novas entidades à rede do metabolismo (FIEHN et al., 2011; ALTMAN et al., 2013). Cerca de metade dos compostos catalogados no KEGG não estão ligados a genes ou enzimas. Embora alguns destes possam participar de reações químicas espontâneas, o mais provável é que a sua catalise enzimática simplesmente não tenha sido anotada, ou não seja conhecida (FIEHN et al., 2011).

Segundo Fiehn et al., (2011) a diversidade do metaboloma está diretamente relacionada a promiscuidade das enzimas quanto aos seus substratos e reações catalisadas. A especificidade do substrato para muitas enzimas não é clara e permanece sem comprovação científica.

Muitas enzimas ainda necessitam de uma rigorosa caracterização funcional, pois são apenas caracterizadas de forma ampla, como por exemplo “citocromo P450” ou “oxidoredutases”, que são classes funcionais vagas para o posicionamento em vias metabólicas realizadas em reconstruções do metabolismo (FIEHN et al., 2011).

A elucidação estrutural *ab initio* de pequenas moléculas é um processo complexo que requer um grande investimento de tempo e recursos, que é sempre efetuada utilizando um conjunto de métodos independentes, tais como espectroscopia de ressonância magnética nuclear (RMN) mono (1D) e bidimensional (2D) ou espectroscopia de infravermelho e cristalografia de raios-X, espectrometria de massas e outros métodos espectroscópicos (PAVIA et al., 2008; KIND; FIEHN, 2010a). É importante que a informação de espectros seja catalogada e facilmente disponibilizada para cada composto. Esta prática permite que informações acumuladas por diferentes experimentos, potencialmente de diferentes grupos de pesquisa, possam sugerir estruturas para compostos observados experimentalmente que não possuam candidatos nos bancos de dados.

Na **Tabela 1** podemos observar o sumário de alguns bancos de dados representativos para o metabolismo. Entre as bases de dados mais abrangentes, os bancos de dados KEGG e MetaCyc possuem conteúdo informativo complementar. De maneira geral, o KEGG possui mais compostos catalogados, enquanto que o MetaCyc possui mais reações e vias metabólicas. Além disso, ambas bases de dados possuem vias metabólicas exclusivas (ALTMAN et al., 2013).

Tabela 1 – Exemplos de bases de dados do metabolismo. Estatísticas consultadas em 14/02/2014.

Base de dados de compostos	Compostos	Vias metabólicas	Reações	Espectros	Organismos
KEGG	17151	455	9485	-	2989
MetaCyc	10965	2097	11409	-	2460
HMDB	5044 ^a	306 ^c	-	1163 ^d	1
ECMDB	2564 ^b	125 ^c	3145	415 ^d	1

^aNúmero total de metabólitos detectados e quantificados;

^bNúmero total de metabólitos que possuem proteínas associadas;

^cVias metabólicas do KEGG associadas aos compostos catalogados;

^dTotal de metabólitos associados a espectros de massa/espectros de ressonância magnética nuclear, incluindo acoplamento a diferentes métodos de separação e diferentes fontes de ionização.

Quando buscamos bases de dados com informações experimentais os recursos são geralmente mais dispersos e escassos. Mesmo organismos amplamente estudados e de fácil manipulação experimental como a bactéria *Escherichia coli* possuem um pequeno número de espectros depositados. Isso dificulta a interpretação de dados espectrais obtidos em novos experimentos, que acabam por ser comparados a bases de dados espectrais genéricas. Por outro lado, iniciativas de bases de dados abrangentes como *The Human Metabolome Database* (HMDB) (WISHART et al., 2007) e *The Escherichia coli Metabolome Database* (ECMDB) (GUO et al., 2013), que armazenam dados espectrais, concentrações dos metabólitos sob diferentes condições e conexões a bases de dados de vias metabólicas (**Tabela 1**) representam marcas iniciais importantes para a constituição de um catálogo abrangente de um metaboloma específico.

1.2- Medição experimental do metaboloma

Dois termos são comumente utilizados para definir a análise do subconjunto observável do metaboloma: perfil de metabólitos (*metabolite profiling*) e impressão digital metabólica (*metabolic fingerprint*). O perfil de metabólitos tem por objetivo a identificação e quantificação de uma ampla gama de metabólitos em um dado sistema. Esta abordagem pode

ser realizada de forma direcionada (*targeted analysis*), onde um grupo selecionado de metabólitos é analisado com mais detalhes utilizando técnicas dedicadas de extração/detecção ou de forma não direcionada (*untargeted analysis*) onde se busca maximizar a amostragem de metabólitos por um desenho experimental específico. Outra abordagem comumente utilizada é a de impressão digital metabólica, em que se almeja a detecção de metabólitos em larga escala, com protocolos práticos para extração/detecção. Esse método não prevê a identificação inicial das identidades dos metabólitos, mas sim com as assinaturas que os picos identificados fornecem do evento do metabolismo sob investigação (NIELSEN, 2007).

O genoma, o transcrito e o proteoma são compostos de biopolímeros. Quatro diferentes nucleotídeos (com o nucleotídeo Timina (T) presente apenas no genoma e o nucleotídeo Uracila (U) presente apenas no transcrito) constituem o genoma e transcrito. O proteoma por sua vez é constituído por 22 aminoácidos. As técnicas utilizadas nas análises desses *omas* são baseadas em sua estrutura regular, como, por exemplo o método de sequenciamento de Sanger. Por outro lado, o metaboloma é composto por uma grande variedade de estruturas químicas, com uma ampla gama de propriedades físico-químicas. As moléculas variam de espécies iônicas simples a produtos naturais complexos. Esta complexidade torna a amostragem completa do metaboloma, através de um único procedimento experimental, virtualmente impossível (NIELSEN, 2007).

Uma rápida taxa de transformação (*turnover*) representa uma segunda camada de complexidade para o metaboloma. Muitos metabólitos estão presentes em baixas concentrações e possuem acelerados fluxos de transformação, se convertendo em outras espécies de metabólitos em escalas de frações de segundos. Diante disso é necessário paralisar o metabolismo (*quench*) rapidamente antes da análise. Basicamente, cada análise do metaboloma necessita de uma definição detalhada dos objetivos pretendidos, e das técnicas de extração e análise disponíveis para delimitar a abrangência do metabolismo amostrado (VILLAS-BOAS, 2007). Como resultado da complexidade do metaboloma em diversidade de propriedades químicas e ampla gama de concentrações, a otimização de métodos de extração é de extrema importância. Protocolos de extração em metabolômica são, de modo geral, desenvolvidos para um determinado organismo, ou mais especificamente, para uma determinada estrutura celular (VILLAS-BOAS, 2007).

Os metabólitos podem ser divididos em três classes, quanto a sua solubilidade: 1) metabólitos solúveis em água ou compostos polares; 2) metabólitos insolúveis em água ou compostos não-polares e 3) metabólitos voláteis. Não existem técnicas disponíveis, capazes de extrair todas as classes de metabólitos simultaneamente. A escolha de um método eficiente para a extração de metabólitos intracelulares requer o entendimento de (i) a estrutura dos limites celulares (paredes e membranas celulares), que são a primeira e principal barreira a ser quebrada; (ii) a natureza química dos metabólitos; (iii) as fontes de perdas (especialmente seu impacto na recuperação subsequente dos metabólitos) (VILLAS-BOAS, 2007).

O primeiro passo na preparação de amostras é a rápida paralisação do metabolismo concomitantemente ou imediatamente após a coleta da amostra. A etapa subsequente demanda que os metabólitos sejam separados do conteúdo celular e ou do meio circundante (meio de cultura, por exemplo). O método de extração tem por objetivo romper a estrutura celular liberando o máximo possível de metabólitos no seu estado original, de uma maneira quantitativa para um meio de extração definido. Nesta etapa, a diversidade química e a abrangência de quantidades (de 10^{-12} mol a 10^{-2} mol) impossibilitam uma amostragem completa de todos os metabólitos, e considerações acerca das espécies-alvo para o estudo devem a ser feitas. A perda pela degradação química constitui uma importante fonte de variabilidade, por exemplo, metabólitos termo ou fotossensíveis podem ser rapidamente degradados se a preparação não é realizada de forma adequada (VILLAS-BOAS, 2007).

Alguns dos fatores importantes a serem levados em conta previamente à análise são: polaridade; pKa (ácido, alcalino, neutro); concentração (passível de ser detectado); detectabilidade (cromóforos, ionização, outros); volatilidade; matriz (interferência de substâncias extraídas em conjunto) (SMEDSGAARD, 2007).

Diante da grande diversidade de propriedades físico-químicas dos metabólitos, existe a necessidade de se utilizar diferentes técnicas analíticas, assim como a combinação de diferentes técnicas (ZHANG et al., 2012). Dentre as técnicas analíticas amplamente difundidas na comunidade de metabolômica se destacam as chamadas “técnicas hífenadas” representadas pelas técnicas de cromatografias (Gasosa - GC ou Líquida - LC) acopladas à espectrometria de massas (MS), e a ressonância magnética nuclear (NMR). Na **Tabela 2**, as principais características destas técnicas são brevemente discutidas de forma comparativa realçando as

características dos dados produzidos pela técnica de LC-MS, objeto de estudo da presente tese.

Tabela 2 - Principais tecnologias utilizadas para medição do metaboloma. Adaptado de (WISHART, 2010).

Tecnologia	Vantagens	Desvantagens
GC-MS	<ul style="list-style-type: none"> -Robusta, boa reprodutibilidade -De baixo custo -Requer pequenas quantidades de amostra -Boa sensibilidade -Grande disponibilidade de <i>software</i> e bancos de dados para identificação de metabólitos -Excelente resolução na separação 	<ul style="list-style-type: none"> -Amostras não são recuperadas -Requer derivatização das amostras -Requer separação -Lenta (20/30 mim/amostra)
LC-MS	<ul style="list-style-type: none"> -Excelente sensibilidade -Grande disponibilidade de combinações de colunas/solventes -Requer muito pequenas quantidades de amostra -Maior abrangência amostral do metaboloma 	<ul style="list-style-type: none"> -Amostras não são recuperadas -Pequena eficiência quantitativa -Alto custo -Lenta (20/30 mim/amostra) -Resolução da separação e reprodutibilidade baixas -Menor disponibilidade de base de dados de espectros
NMR	<ul style="list-style-type: none"> -Quantitativa -Não destrutiva -Rápida (2-3 min/amostra) -Não requer derivatização -Não requer separação -Detecta todas as classes orgânicas -Permite a identificação de novos compostos -Compatível com líquidos e sólidos -Grande disponibilidade de <i>software</i> e bancos de dados para identificação de metabólitos 	<ul style="list-style-type: none"> -Baixa sensibilidade -Alto custo -Não detecta íons inorgânicos -Não pode detectar compostos não protonados -Requer grandes quantidades de amostra

Segundo Beltran et al., (2012) ^1H NMR é uma técnica caracterizada por uma elevada reprodutibilidade, análises rápidas e robustas, e potencial para quantificação dos metabólitos analisados. Além disso, não requer uma extensa preparação da amostra, e existem bases de dados públicas disponíveis para identificação de metabólitos. A limitação inerente da NMR, no entanto, é a sua baixa sensibilidade, o que dificulta a detecção de metabólitos importantes, presentes em baixas concentrações (BELTRAN et al., 2012).

Um grande número de estudos em metabolômica utiliza métodos de separação antes da análise por espectrometria. Usualmente a cromatografia consiste em uma fase estacionária representada pela coluna cromatográfica e uma fase móvel. O processo de separação é realizado através da constante alimentação da coluna (fase estacionária) por uma fase móvel (líquido/gás), após inserção da amostra. O processo de separação é dinâmico e pequenas diferenças no coeficiente de distribuição determinam o tempo que diferentes compostos permanecerão na fase estacionária (THEODORIDIS et al., 2013).

A cromatografia gasosa é apropriada para amostras voláteis, para amostras cujos compostos possam ser volatilizados por derivatização ou quando um alto poder de separação é requerido. A derivatização é uma técnica utilizada em química, que altera a estrutura de um composto, modificando suas propriedades físicas, como o ponto de ebulição, por exemplo. O acoplamento da cromatografia gasosa a ionização por elétrons e espectrometria de massas é adequada para identificação e quantificação de compostos (KIND; FIEHN, 2010b).

A Cromatografia Líquida de Alto Desempenho (HPLC - *High-Performance Liquid Chromatography*) é um método de separação versátil que permite a separação de compostos de uma ampla gama de polaridades através da eluição isocrática (a composição do solvente aquoso permanece constante durante a separação) ou gradiente de eluição (a composição do solvente aquoso muda durante a separação). Acetonitrila, metanol e tetrahidrofurano são os solventes orgânicos utilizados mais comumente. O gradiente de eluição isocrático é preferido para amostras simples (i.e., menos que 10 componentes). A cromatografia líquida é utilizada no complemento de compostos não voláteis, em extratos complexos, onde a derivatização não pode ser utilizada e pode ser acoplada a uma infinidade de detectores: fotodiodos, fluorescência, eletroquímicos, NMR. Além desses, os espectrômetros de massa apresentam grande destaque (SMEDSGAARD, 2007).

A história da espectrometria de massas começa com o físico Sir Joseph John Thomson, agraciado com o prêmio Nobel de 1906 pela descoberta do elétron (RAYLEIGH, 1941). No início do século 20, Sir J. J. Thomson construiu o primeiro espectrômetro de massas, então denominado *parabola spectograph*, no qual íons eram separados pela diferença em sua trajetória parabólica em campos eletromagnéticos e a detecção ocorria quando os íons atingiam uma tela fluorescente ou placa fotográfica. Mais tarde, o discípulo de Thomson, Francis W. Aston da Universidade de Cambridge, desenhou um espectrômetro de massas que aumentou o poder de resolução das massas, permitindo o estudo de isótopos. Durante o mesmo período, A. J. Dempster da Universidade de Chicago também potencializou a resolução e criou a primeira fonte de ionização por elétron. Essas pesquisas desenvolveram forte fundamentação para espectrometria o que permitiu uma ampla gama de aplicações para diversas áreas do conhecimento (SIUZDAK, 2006).

Embora todos os espectrômetros de massas utilizem um analisador de massas, nem todos os analisadores operam de maneira similar, alguns separam íons pelo espaço enquanto outros separam íons pelo tempo. Em termos genéricos, um analisador de massas mede íons em fase gasosa com respeito a sua razão massa carga (m/z), onde a carga é produzida pela adição ou perda de próton(s) (SIUZDAK, 2006). Um espectrômetro de massas é tipicamente composto de três partes principais: fonte de ionização, analisador de massas e detector (**Figura 1**). Enquanto a fonte de ionização converte as moléculas da amostra em íons, o analisador separa os íons de acordo com as características físicas da molécula e o detector registra o sinal amplificado do íon.

Assim como técnicas de extração, diferentes técnicas de ionização são complementares quanto a eficiência de ionização para moléculas diversas, apresentando eficiência variável, por exemplo, quanto ao peso molecular e polaridade do composto (ZHOU et al., 2012). Em razão da grande diversidade de formas químicas dos metabólitos, a utilização de diferentes fontes de ionização é frequentemente requerida para analisar matrizes biológicas. A ionização por elétron (EI – *Electron Ionization*) a 70 eV resulta em um espectro de massas consistente e rico em fragmentos, devido à energia de ionização constante selecionada. Estes espectros de massas podem ser facilmente utilizados para a busca em uma biblioteca de espectros de massas, uma vez que representam um padrão de fragmentação dependente da estrutura da

molécula, que é altamente reprodutível por diferentes grupos de pesquisa. A ionização por elétron é comumente utilizada acoplada a configurações de GC-MS. A grande desvantagem de espectros de massa obtidos por EI é a baixa abundância ou ausência do íon molecular (KIND; FIEHN, 2010b).

A introdução de ionização por *electrospray* (ESI) foi um grande avanço para a análise de biomoléculas grandes e intactas. A técnica de ESI é atualmente o método de ionização de escolha para LC-MS em muitos laboratórios em todo o mundo (KIND; FIEHN, 2010b). Classificada como ionização fraca, ESI produz um grande número de íons através da troca de cargas na solução e frequentemente formam espectros com baixa fragmentação e com grande abundância de íons protonados/desprotonados, que auxiliam a identificação inicial do metabólito (SMEDSGAARD, 2007).

Figura 1 - Representação esquemática de um espectrômetro de massas, onde a formação de íons é representada pela ionização por *electrospray* (ESI), a separação pelo analisador quadrupolo e a detecção pelo multiplicador de íons dínodo contínuo. Após a separação um sinal é convertido e armazenado pelo *software* do equipamento. Adaptado de (SILVERSTEIN et al., 2005; HOFFMANN; STROOBANT, 2007).

O segundo componente de um espectrômetro de massas é o analisador. De acordo com Kind and Fiehn, (2010) espectrômetros de massa são normalmente concebidos para fins

analíticos específicos: espectrômetros de massa *ion trap* como versáteis, para análises de MSⁿ, por exemplo; espectrômetros de massa quadrupolo como equipamento de uso intenso para diversas aplicações, em geral, com baixo custo e resolução satisfatória; espectrômetros de massa triplo quadrupolo como instrumentos muito sensíveis para a análise de compostos alvo, e instrumentos de transformada de Fourier para medições que exigem alto poder de resolução e de alta precisão em massa (KIND; FIEHN, 2010b). A **Tabela 3** traz um resumo das principais características de analisadores de massa amplamente difundidos.

Tabela 3 - Performance de espectrômetros de massa modernos na quantificação de pequenas moléculas. Extraída de (ZHOU et al., 2012).

MS	Poder de Resolução ^a	Acurácia	Taxa de Scans	Limite de Detecção ^c
QqQ	Unidade	50 ppm ^b	Moderada	10 ⁻¹⁵ g-10 ⁻⁹ g(SMR)*
QIT	10.000	50 ppm	Moderada	10 ⁻¹⁵ g-10 ⁻⁹ g(SMR)
LIT	10.000	50 ppm	Rápida	10 ⁻⁹ g(SMR)
IT-TOF/TOF	20.000	3 ppm	Rápida/ Moderada	10 ⁻⁹ g(<i>full scan</i>)
TOF-alta-res.	60.000	2 ppm	Rápida	10 ⁻⁹ g(<i>full scan</i>)
Q-Orbitrap/LTQ-Orbitrap	140.000	2 ppm	Moderada	10 ⁻¹⁵ g-10 ⁻⁹ g(<i>full scan</i>); SMR compatível
LTQ-FTICR/Q-FTICR	1.000.000	≤ 1 ppm	Baixa/Baixa	10 ⁻⁹ g(<i>full scan</i>)

^aO poder de resolução depende do intervalo de valores de m/z e velocidade de *scan* do instrumento. Os valores apresentados foram obtidos para m/z 400.

^bPartes por milhão (ppm).

^cA sensibilidade depende fortemente da eficiência de ionização do composto na fonte de ionização.

*SRM (*Selected Reaction Monitoring*).

1.3- Aquisição e representação de dados em espectrometria de massas

Existem diferentes tipos de detectores, e sua escolha depende do desenho do instrumento, bem como da aplicação para a qual o mesmo será empregado. Diferentes estratégias são utilizadas para detectar os íons, entretanto, esta detecção é sempre baseada em sua carga,

massa ou velocidade. A interface entre um espectrômetro de massas e um computador que registra o sinal elétrico gerado pelos íon separados pelo analisador de massas é composta por um *hardware* que recebe e transforma o sinal analógico capturado (**Figura 1**). Esta interface funciona como uma conversora de sinal analógico para digital (ACD *Analogue-to-digital converter*). O volume de dados direcionado pelo ACD para o sistema de aquisição não tem relação direta com o número de íons detectados, mas é proporcional ao sinal dos íons coletados em um espectro (HOFFMANN; STROOBANT, 2007). Como o número de íons deixando o analisador de massas é muito pequeno, uma amplificação do sinal é necessária para obter-se um registro utilizável do íon (**Figura 1**) (HOFFMANN; STROOBANT, 2007).

Quando os dados são coletados, são transferidos para um computador, que relaciona este sinal a uma razão massa-carga (m/z). As tabelas de calibração são tipicamente preparadas por meio da análise de uma amostra conhecida e cálculo de uma relação entre os valores medidos de m/z e a massa monoisotópica verdadeira. Na maioria dos casos, uma curva de calibração polinomial é obtida através de picos conhecidos, suavizando pequenos erros (HOFFMANN; STROOBANT, 2007; SMEDSGAARD, 2007). Quando a escala de massa é adicionada aos dados, obtemos um espectro de massas bruto ou, espectro de massas contínuo, como mostrado na (**Figura 2**).

Na maioria dos casos, os dados contínuos são processados, os picos de massa são detectados e o resultado é mostrado como uma barra no centro de um valor de m/z e com uma altura correspondente a contagem da corrente (**Figura 2**). Estes espectros de barra são normalmente referidos como espectros em modo centróide e são geralmente normalizados para o pico mais alto do espectro. Há uma redução considerável do tamanho do arquivo de dados no cálculo de espectros de massa em modo centróide, com muito pouca perda de informação (SMEDSGAARD, 2007).

Figura 2 - Estrutura dos dados obtida pelos detectores em espectrometria de massas. Dados brutos em modo contínuo a esquerda. Na maioria dos casos os dados contínuos são convertidos em dados em modo centróide detectando a posição do centróide do pico e sua altura ou área, representação a direita. Adaptado de (SMEDSGAARD, 2007).

Um analisador de massas ideal seria capaz de distinguir íons com qualquer diferença real para a razão m/z , mas, como qualquer medição experimental, um espectro de massas é sujeito a incertezas advindas de variações aleatórias das medidas. A medição de vários íons com valores de m/z muito próximos gera a forma de um pico em um espectro de massas que é centralizado (**Figura 2**). Um pico de massas é, portanto, o máximo local de um sinal produzido pelo detector, que representa os íons de alguma entidade química (LANGE, 2008).

A **Figura 3** mostra o espectro de massas hipotético do padrão isotópico de íons do peptídeo bombesina, hipoteticamente obtido por analisadores com diferentes capacidades de separação da razão m/z . Um termo comumente utilizado para capacidade de separação é a resolução. Um analisador de baixa resolução ($R=1000$), não pode discriminar os três íons e somente um pico é observado no espectro de massas. A uma resolução um pouco maior ($R=2500$), os íons começam a ser diferenciados. Os picos podem ser claramente observados a uma resolução de 5000, e os sinais são resolvidos em um nível em que é possível separar o sinal real dos menores níveis de ruído experimental tipicamente observados (discutido abaixo) a uma resolução de 10000. Exemplos de equipamentos que proporcionam diferentes resoluções são exemplificados na **Tabela 3**.

Figura 3 - Efeito do aumento da resolução na diferenciação de íons formados pelo padrão isotópico do peptídeo bombesina (código UniProt P84214). O padrão isotópico foi simulado utilizando a ferramenta *Isotopica* (FERNANDEZ-DE-COSSIO et al., 2004). Adaptado de (LANGE, 2008).

A **Figura 3** mostra que a correspondência entre picos no espectro e íons formados pelos componentes é de um para um, apenas se o espectrômetro de massas é capaz de resolver os íons de componentes diferentes. Quanto maior a resolução, mais estreitos são os picos, e os mesmos são mais bem separados no espectro. Uma das definições mais utilizadas para resolução é a da largura máxima do pico na metade da intensidade máxima (SILVERSTEIN et al., 2005) (FWHM – *Full Width Half Maximum*):

$$R_{FWHM} = \frac{m}{\Delta m}$$

onde m é a posição da maior m/z do pico e Δm seu valor de FWHM.

Com resolução suficiente não se possibilita apenas a definição de moléculas, mas também o padrão isotópico de uma dada molécula. Isótopos são átomos do mesmo elemento que diferem na massa, uma vez que possuem um número de nêutrons diferentes, e o mesmo número de prótons e elétrons. A massa exata de um composto é dada pela soma dos isótopos mais frequentes de cada átomo (**Tabela 4**).

Tabela 4 - Átomos e suas proporções isotópicas relevantes para espectrometria de massas. Adaptado de (SMEDSGAARD, 2007).

Elemento	Isótopo	Abundancia (%)	Massa baseada no padrão ^{12}C
H, hidrogênio	^1H	99.985	1.007825
C, carbono	^{12}C	98.93	12.000000
	^{13}C	1.07	13.003355
N, nitrogênio	^{14}N	99.632	14.003074
	^{15}N	0.368	15.000109
O, oxigênio	^{16}O	99.557	15.994915
	^{17}O	0.038	16.999132
	^{18}O	0.205	17.999160
P, fósforo	^{31}P	100	30.973762
S, enxofre	^{32}S	94.93	31.972071
	^{33}S	0.76	32.971459
	^{34}S	4.26	33.967867
Cl, cloro	^{35}Cl	75.78	34.968853
	^{37}Cl	24.22	36.965903

O ruído em um espectro de massas pode ser classificado em químico ou aleatório (LANGE, 2008) (**Figura 4**). O ruído químico tem um padrão fixo, que se manifesta em razões m/z específicas. Este resulta da análise de espécies carregadas que não correspondem diretamente ao metabólito, como, por exemplo, íons ou adutos de sais que são formados na solução do *electrospray*, espécies geradas eletroquimicamente, ou espécies neutras presentes na atmosfera ao redor do spray de ESI que são carregadas na fase gasosa pela transferência de um próton (**Tabela 5**). Constituição de solvente e tampão, pKa, pH, propriedades de doadoras e aceptoras de prótons das substâncias e a acidez em fase gasosa podem influenciar a formação de adutos. Diferentes produtos de adição também podem resultar em diferentes caminhos de fragmentação (KIND; FIEHN, 2010b). A diversidade e ocorrência de diferentes formas iônicas vêm sendo catalogadas (KELLER et al., 2008), e devem ser utilizada para uma efetiva identificação dos metabólitos.

Tabela 5 - Íons e agrupamentos mais frequentemente formados em ESI. Adaptado de (SMEDSGAARD, 2007).

	ESI modo positivo		ESI modo negativo	
	Estrutura	Mudança na massa nominal (m/z)	Estrutura	Mudança na massa nominal (m/z)
Adutos	$[M+H]^+$	+1	$[M-H]^-$	-1
	$[M+NH_4]^+$	+14	$[M+Cl]^-$	+35
	$[M+H_2O+H]^+$	+19	$[M+CHOO]^-$	+45
	$[M+Na]^+$	+23	$[M+CH_3COO]^-$	+59
	$[M+CH_3CN+H]^+$	+42	$[M+HSO_4]^-$	+97
	$[M+CH_3CN+Na]^+$	+64	$[M+H_2PO_4]^-$	+97
	$[M-H+2Na]^+$	+45		
	$[M-(n-1)H+nNa]^+$	+23n-1		
Fragmentos	$[M-H_2O+H]^+$	-17	$[M-H_2O-H]^-$	-19
	$[M-H_2O+Na]^+$	+5	$[M-H_3PO_4-H]^-$	-98
	$[M-CO_2+H]^+$	-27		
	$[M-CO_2+Na]^+$	-5		
Multímeros	$[2M+H]^+$	2m + 1		
	$[2M+H_2O+H]^+$	2m + 19	$[2M-H]^-$	2m-1
	$[2M+NH_4]^+$	2m + 14		
	$[2M+Na]^+$	2m + 23		

M representa um íon com massa m . Em geral agrupamentos com moléculas do solvente são esperados. Para moléculas maiores (> 1000 Da) íons com múltiplas cargas (e) devem ser levados em conta, posicionados na razão de sua massa no espectro, m/e .

Ruído eletrônico é qualquer fonte de interferência indesejável cuja ocorrência não seja correlacionada com sinais específicos, sendo uniformemente distribuído por todas as razões m/z (**Figura 4**). Esse ruído é presumivelmente assumido como derivado do detector de um instrumento de MS, entretanto, quando ESI é acoplado à cromatografia líquida, o ruído químico pode ser muito abundante no início e no final do processo de eluição, e por este fato a linha base de ruído é frequentemente observada na forma apresentada na **Figura 4 - C** (LANGE, 2008). Ambos os tipos de ruído podem confundir ou mimetizar sinais de interesse (LANGE et al., 2006). A **Figura 4** ilustra a composição aditiva de um espectro de massas com

os picos dos metabólitos, com ruídos sistemático e aleatório.

Figura 4 - Decomposição de um espectro de massas em diferentes componentes de sinais detectados.

1.4- Pré-processamento de dados em espectrometria de massas

A extração e detecção simultânea de metabólitos por LC-MS produz conjuntos de dados

complexos, que requerem uma série de etapas de pré-processamento, para que a informação possa ser extraída com eficiência e precisão.

Uma vez que não existem métodos ou aplicações individuais capazes de extrair toda a informação relevante de conjuntos de dados multidimensionais gerados pelas ciências ômicas, novos algoritmos e aplicações são constantemente desenvolvidos para aperfeiçoar as análises. Plataformas de análises abrangentes, com a filosofia de “*one stop shop*” têm ganhado notoriedade no campo da metabolômica. Destaque para XCMS online (TAUTENHAHN; PATTI; et al., 2012), uma das ferramentas de acesso livre mais utilizada em metabolômica, que vem constantemente adicionando funcionalidades a um servidor web que fornece ampla documentação e suporte. Outra família de ferramentas também hospedada no centro de espectrometria do Instituto de Pesquisas Scripps é o banco de dados METLIN (TAUTENHAHN; CHO; et al., 2012), que disponibiliza ferramentas de busca em interfaces web, bem como acesso remoto ao seu banco de dados. Outro excelente repositório de ferramentas é a plataforma MetaboAnalist (XIA et al., 2009, 2012) que disponibiliza uma interface web com inúmeras opções de plataformas experimentais (GC/LC-MS, NMR, etc), acesso a diferentes bancos de dados (KEGG, HMDB, etc) e análises estatísticas que compreendem desde o pré-processamento a diferentes algoritmos de classificação, redução de dimensionalidade e inferência. A plataforma MetaboAnalist também apresenta uma iniciativa exemplar para o universo de *software* livre: todo o código e instrução para a instalação do servidor web é acessível caso o usuário queira replicar o servidor em sua própria estrutura computacional. Ainda, extensa documentação está disponível para cada funcionalidade, que além de realizar as análises geram relatórios padrão com breves explicações sobre o princípio dos algoritmos utilizados, auxiliando a interpretação dos resultados. São também notórios por sua extensão e usabilidade os *softwares* MZmine (PLUSKAL et al., 2010) e IDEOM (CREEK et al., 2012). Como uma aplicação da amplamente difundida planilha de cálculos, o Microsoft Excel, IDEOM implementa macros que utilizam algoritmos externos do ambiente R e apresentam os dados de maneira organizada e eficiente. O *software* IDEOM é também interligado a uma grande família de ferramentas presentes no projeto mzMatch (SCHELTEMA et al., 2011).

Muitas outras aplicações para uma ampla gama de análises vêm sendo disponibilizadas

para a comunidade de metabolômica, uma análise detalhada das características e especificidades de muitas ferramentas é descrita em (CASTILLO et al., 2011; SUGIMOTO et al., 2012). A fim de delinear o escopo do desafio das análises de dados abordados na presente tese, trataremos brevemente das principais etapas do pré-processamento e do produto destas etapas, que é o material de estudos desta tese. Ilustramos as etapas do pré-processamento com os métodos implementados por *xcms* e *MZmine 2*.

De acordo com Smith et al., (2006) no pré-processamento de dados de separação e espectrais, duas estratégias gerais podem ser utilizadas: (1) Dividir o sinal em intervalos (*bins*), incorporando todos os dados em um perfil de reconhecimento para cada amostra. (2) Identificar e quantificar individualmente picos (*features*), descartando dados não considerados parte do pico. Segundo os mesmos autores, variância e *bias*, não atribuídas a diferenças reais na abundância dos metabólitos, geralmente reduzem o poder destes procedimentos. Assim, um importante objetivo de um método de pré-processamento deve ser a redução da variância não biológica (SMITH et al., 2006).

O algoritmo de detecção de picos utilizado pelo *xcms* baseia-se na divisão dos dados de LC/MS em intervalos de uma fração de uma unidade de massa ($0,1 \text{ m/z}$) e, em seguida, realizam os cálculos subsequentes em intervalos individuais na dimensão do tempo cromatográfico, uma vez que esta dimensão possui um grande número de dados por pico. Dentro de cada intervalo, o sinal é determinado tomando a intensidade máxima em cada ponto de tempo no intervalo. Os autores demonstraram que o filtro aplicado para detecção do pico produzia melhorias consistentes para a relação SNR (Razão Sinal Ruído), mesmo se a largura do pico modelo for 1,5 a 4 vezes a largura do pico do sinal, indicando que o desempenho do filtro não é muito sensível a variações na largura do pico (**Figura 5 - A**).

Figura 5 - Etapas da análise de dados de cromatografia líquida acoplada a espectrometria de massas. Os passos essenciais constituem: identificação e filtro dos picos, agrupamento ao longo das réplicas experimentais e correção dos tempos de retenção. Após o pré-processamento uma etapa de agrupamento de formas redundantes de um composto é requerida. Adaptado de (TAUTENHAHN et al., 2008; KUHLE et al., 2011; TAUTENHAHN; PATTI; et al., 2012).

A geração de intervalos de massa diminui a precisão da determinação dos picos de massa.

Smith et al., (2006) argumentam que se os limites dos intervalos são escolhidos arbitrariamente, o sinal de um dado metabólito pode ser subdividido entre dois intervalos adjacentes, como exemplificado na **Figura 5 - C** pelos limites dos intervalos (representados pelas linhas pontilhadas em vermelho). Adicionalmente, múltiplos sinais de metabólitos independentes podem ser contabilizados para um intervalo único, dificultando a diferenciação de metabólitos. Para compensar a perda de precisão, a massa de cada pico é calculada a partir dos espectros originais de alta resolução. Em primeiro lugar, a massa de alta resolução é determinada em cada espectro contendo o pico. Em segundo lugar, a massa total do pico é calculada como uma média ponderada de todas as massas de resolução completa, usando intensidades como os pesos. O algoritmo de detecção de pico, portanto, lida com baixa resolução, alta resolução, e dados em modo centróide de uma maneira flexível e robusta (SMITH et al., 2006).

A primeira versão do *software* MZmine definia um conjunto de métodos de pré-processamento e visualização (KATAJAMAA et al., 2006), entretanto, seu desenho inicial não era modular dificultando a adição de novos métodos levando ao desenvolvimento de uma segunda versão mais modular, o MZmine 2 (PLUSKAL et al., 2010). Desenvolvido utilizando a tecnologia Java, o *software* MZmine 2 é portátil e possui uma interface intuitiva e de fácil utilização.

Na primeira etapa do pré-processamento disponível para MZmine 2 cada espectro é processado individualmente e convertido para pares m/z e valores de intensidade (isto é, cada espectro de massas é transformado em centróide). O *software* MZmine 2 fornece diferentes algoritmos para detecção de picos: O algoritmo de limiar recursivo (*recursive threshold*) é baseado em (KATAJAMAA et al., 2006). Este método reduz o número de falsos positivos evitando a detecção de ruído. O algoritmo de transformação *Wavelet* é particularmente adequado para dados sujeitos a ruído. O algoritmo de “Massa exata” assume espectros de alta qualidade (alta resolução e baixo ruído) e determina o centro de cada m/z utilizando o paradigma de FWHM. Valores de m/z são colocados no meio de uma linha, que cruza o pico na metade da intensidade máxima.

O *software* MZmine 2 utiliza um algoritmo que conecta valores de m/z consecutivos que compreendem várias aquisições de sinal independentes (*scans*) em um cromatograma. Um cromatograma abrangendo um dado intervalo de tempo é construindo para cada valor de m/z

(com uma tolerância definida pelo usuário). Cada cromatograma é então deconvolvido em picos cromatográficos individuais. Vários algoritmos adicionais também estão disponíveis como *plug-ins* para selecionar picos levando em conta reprodutibilidade, ruído e formato do pico.

Outra tarefa importante no pré-processamento de dados para análises comparativas é combinar os picos que representam o mesmo analito de diferentes amostras (**Figura 5 - C**) (SMITH et al., 2006). Existem métodos que realizam a correspondência dos picos gerando primeiro um cromatograma médio, cujos picos são usados para definir os centros dos grupos. Amostras individuais são então comparadas com a lista de referência. Segundo Smith et al., (2006) este algoritmo requer que o desvio no tempo de retenção de amostra para amostra não seja maior do que o tempo entre dois picos adjacentes.

Uma maneira de superar as limitações impostas pelo desvio nos tempos de retenção de amostra para amostra é a utilização de um algoritmo de alinhamento. De acordo com Smith et al., (2006) uma técnica frequentemente utilizada envolve o emprego de um pequeno número de padrões internos em todas as amostras antes da aquisição dos dados. Durante o pré-processamento de dados, os picos correspondentes aos padrões são identificados e utilizados para deslocar linearmente o tempo de retenção de cada uma das amostras adquiridas. Existe uma série de desvantagens importantes para este método. Primeiro, por que se supõe que os desvios no tempo de retenção sejam lineares. Segundo, por que este requer um passo adicional de preparação da amostra e a adição de produtos químicos que podem mascarar a presença de outros analitos relevantes experimentalmente (SMITH et al., 2006).

O método de alinhamento do tempo de retenção disponível para xcms corrige simultaneamente os tempos de retenção de todas as amostras em um único passo. A capacidade de fazer isso depende da existência inicial de uma correspondência grosseira dos picos em grupos razoáveis (**Figura 5 - C, D**). Esta correspondência inicial é possível por causa da separação de massas das amostras pela cromatografia e a relativamente baixa complexidade de cada cromatograma do íon base individual (SMITH et al., 2006).

A partir do agrupamento inicial, o algoritmo identifica tipicamente centenas de grupos de picos reprodutíveis (picos que apresentam padrões de intensidades correlacionadas em diferentes amostras) para os quais muito poucas amostras não têm picos atribuídos e muito poucas amostras possuem mais do que um pico atribuído. Tais grupos reprodutíveis têm uma

elevada probabilidade de serem adequadamente atribuídos e podem ser utilizados como padrões temporários (SMITH et al., 2006).

Para cada grupo, o algoritmo calcula a mediana do tempo de retenção e o desvio da mediana para cada amostra em um dado grupo. Uma vez que os grupos de picos reprodutíveis são geralmente distribuídos uniformemente ao longo de porções significativas do perfil cromatográfico, um contorno detalhado de desvios não-lineares do tempo de retenção pode ser construído para cada amostra (SMITH et al., 2006).

A aplicação original do MZmine implementava um algoritmo para o alinhamento dos tempos de retenção (*join aligner*) que executa as seguintes tarefas: i) criar uma lista de picos vazia; ii) alinhar cada pico de outras listas de picos (amostras) ao melhor candidato da lista, utilizando uma janela de alinhamentos de duas dimensões representada por tolerâncias de m/z e tempo de retenção fornecidas pelo usuário; iii) se nenhum candidato dentro dos limites estabelecidos é encontrado, uma nova entrada é criada na lista. Esta abordagem não é capaz de levar em conta os desvios não lineares dos tempos de retenção e por isso foi atualizada na versão 2 do aplicativo.

O *software* MZmine 2 introduz o algoritmo iterativo não determinístico RANSAC (*Random Sample Consensus*) (FISCHLER; BOLLES, 1981). Em cada iteração, um conjunto aleatório de pontos dos dados é selecionado e um modelo é ajustado para estes pontos. Os dados remanescentes são testados contra o modelo ajustado e se os valores se ajustam bem, são considerados parte do modelo. Finalmente o modelo é testado e ao final das iterações o modelo com maior número de pontos ajustados é considerado o melhor.

Uma vez que os picos são detectados, separados do ruído e alinhados entre as réplicas experimentais, conjuntos de dados gerados por LC-MS possuem padrões de íons que são compostos por isótopos, fragmentos e adutos (**Figura 5 - E**) formados durante o processo de ionização (**Tabela 5**). Formas redundantes de um mesmo íon dificultam a interpretação dos espectros de massas. Ferramentas de busca web como METLIN ou a interface amigável do MZmatch 2 permitem a busca de padrões, como formas iônicas previamente catalogadas em experimentos similares (**Tabela 5**), sob a supervisão do usuário. Uma busca automatizada por padrões de ionização está disponível por meio de diferentes algoritmos para o ambiente de desenvolvimento R (SCHELTEMA et al., 2009; ALONSO et al., 2011; KUHLE et al., 2011). Estas aplicações utilizam diferentes algoritmos, que são baseados em princípios comuns, os

quais poderiam ser resumidos em: i) determinação de grupos de picos em intervalos de tempo de retenção; ii) correlação da forma do pico; iii) diferenças de massa relacionadas a formas iônicas conhecidas; iv) correlação de um padrão de picos ao logo de réplicas experimentais. A utilização destes padrões para formação de grupos de íons não depende de um íon mono protonado e diminui consideravelmente o número de íons em uma análise (SCHELTEMA et al., 2009).

A determinação da massa exata de uma molécula através destas regras é essencial para a anotação dos metabólitos. O termo anotação do metabólito será utilizado neste texto seguindo como referência os padrões estabelecidos pela comunidade de metabolômica para identificação de pequenas moléculas (SUMNER et al., 2007). A inferência da massa exata pelos padrões encontrados nos espectros, como descrito acima, é utilizada para uma primeira comparação e sugestão da identidade para os metabólitos, na ausência de compostos químicos purificados para comparação (**Tabela 6**).

Tabela 6 - Os quatro níveis de confiança para identificação de metabólitos definidos pela Iniciativa de Padrões da Comunidade de Metabolômica. Adaptado de (DUNN et al., 2012).

Nível	Confiança na identificação	Nível de evidência
1	Metabólito identificado com confiança	Comparação de 2 ou mais propriedades independentes com um composto purificado em condições analíticas idênticas
2	Metabólitos anotados putativamente	Baseado em propriedades físico-químicas e/ou similaridade espectral com bibliotecas espectrais, sem referência a padrões químicos autênticos
3	Classes de Metabólitos putativamente anotadas	Baseado em propriedades físico-químicas de uma classe de compostos, ou similaridade espectral a uma classe de compostos
4	Metabólitos desconhecidos	Mesmo não identificado ou discriminado em classes, o metabólito pode ainda ser diferenciado e quantificado baseando se em dados espectrais

A metabolômica vem gradualmente sendo sistematizada para delimitar a extensão do metaboloma, o escopo e a abrangência de cada plataforma experimental para amostragem do metaboloma, e as melhores abordagens para relacionar o estado dos metabólitos amostrados

aos fenótipos de interesse. Para que a continuidade desta evolução seja possível, a comunidade científica posiciona como principal gargalo a identificação dos picos de massa com confiabilidade (DUNN et al., 2012). Para que as abordagens de perfil metabólico não direcionado possam ser efetivamente relacionadas às alterações de interesse no metabolismo, é estritamente necessário que os conjuntos de dados sejam interpretados sob a pressuposição de uma amostra conectada do metabolismo. Sob esta pressuposição, a estrutura dos dados encerra a informação das inter-relações entre os metabólitos, e os picos de massa não são mais tratados como informação indireta de entidades químicas dissociadas e sim como uma representação de um subconjunto da rede do metabolismo.

Diante deste desafio posicionado pela comunidade de metabolômica realizamos na presente tese o desenvolvimento de um arcabouço de *software*, que tem como componente central um método probabilístico de anotação de metabólitos (ROGERS et al., 2009), que permite a incorporação de fontes independentes de informações espectrais e conhecimento prévio acerca do metabolismo do organismo, ou mais especificamente, de um evento do metabolismo amostrado. Nós demonstramos em conjuntos de dados reais que a implementação é eficiente e utiliza novas proposições metodológicas para inferência da massa exata dos metabólitos, coleção e formatação de bancos de dados com contexto biológico e representação dinâmica das redes com informações associadas.

2- HIPÓTESE DO TRABALHO

A hipótese de trabalho da presente tese é que a distribuição de probabilidades *a posteriori* completa das atribuições pico de massa para possíveis metabólitos representa as incertezas envolvidas. Diante desta suposição, disponibilizar ao especialista uma lista de possíveis identidades e suas probabilidades, em lugar de somente o candidato mais provável, possibilitaria uma classificação de melhor qualidade. Ainda, buscou-se demonstrar que tal distribuição *a posteriori* tem sua utilidade potencializada quando é representada como um grafo, em um ambiente de *software* que permite a navegação dinâmica dos dados.

3- OBJETIVOS

A presente tese teve por objetivo o desenvolvimento de um arcabouço de *software* flexível, disponibilizando um método probabilístico que incorpora o conhecimento prévio do metabolismo e informações de fontes experimentais independentes na forma de verossimilhança. Após a determinação de um ranqueamento de probabilidades para a classificação, novas formas de representação e exploração dos dados, utilizando grafos, foram propostas.

3.1- Objetivos específicos

- i) Propor métodos para incorporação prática de novas fontes de informação e cálculo das probabilidades de atribuição de um metabólito a um composto;
- ii) Criar formatos para representação dos resultados, com conexões práticas a dados espectrais e bases de dados biológicas;
- iii) Propor métodos para representação das distribuições de probabilidade *a posteriori* em redes que possibilitem a inferência de eventos metabólicos nas condições estudadas.

4- MÉTODOS

4.1- Anotação dos íons

Propor a massa exata é essencial para identificação dos metabólitos amostrados pela busca em bibliotecas de compostos conhecidos ou para o cálculo da composição elementar dos compostos (KUHL et al., 2011). Para gerar um conjunto de massas exatas propostas para a anotação dos compostos utilizamos o cálculo da “massa neutra” do metabólito como:

$$m_{neutra} = (m/z \cdot z) \pm m_{aduto}$$

ou

$$m_{neutra} = (m/z \cdot z) + m_{isotopo}$$

onde a soma ou subtração é definida pelo modo de ionização e m_{aduto} é a massa do aduto de ionização. Este aduto é selecionado por um conjunto de regras fixadas pelo usuário, no caso do *software* CAMERA, ou pela aplicação no caso do *software* mzMatch. A massa $m_{isotopo}$ é adicionalmente buscada dentro de um intervalo de tempo definido, correspondendo a diferença de massa e proporção de intensidade dada pela massa do isótopo de um átomo específico e frequência natural dos isótopos (**Tabela 4**).

Uma vez que o pacote CAMERA checa as regras para os adutos de uma tabela de diferenças de massa de referência, e deixa sem anotações os picos de massa que não atendem as restrições do algoritmo, adicionamos a opção de checar a heurística de íons na forma $[M+/-H]^{+/-}$ para os íons não anotados.

Além da predição da massa neutra do metabólito pela estrutura interna do espectro, outra abordagem utilizada para aumentar a cobertura de íons amostrados, e reunir mais evidências para determinação da massa neutra, é a utilização da combinação de modos de ionização. O *software* CAMERA utiliza uma tabela adicional de regras para busca nos dados brutos como, por exemplo: buscar $[M+Na]^+$ no espectro do modo positivo e $[M-H]^-$ no modo negativo, caso

exista um padrão nos dois espectros, a anotação do pico é estabelecida. Uma vez que é difícil antecipar qual será o par de íons presente em ambos os espectros, no presente trabalho estabelecemos um intervalo de massas e um intervalo de tempo, para os quais os íons preditos nas anotações prévias são confrontados, e se um valor M comum é encontrado, uma das anotações é descartada evitando a redundância na anotação dos compostos.

De posse de uma tabela de anotação no formato:

Massa	Tempo de retenção	Isótopo	Aduto	Traço	Pseudo grupo
x_1	t_1	0	0	ref_1	pc_1
x_m	t_m	1	0	Ref_m	pc_m
...	...			...	
x_M	t_M	0	N	ref_M	pc_M

onde, para um conjunto de massas neutras e seus respectivos tempos de retenção sugeridos pela sequência de etapas descrita acima, podemos encontrar o pico mais intenso de um padrão isotópico, representado pelo índice 0 nas colunas *Isótopo* e *Aduto*, seguida pelas moléculas com um ou mais átomos do isótopo minoritário representado pelos índices de um 1,...,N e 0 nas respectivas colunas. Quando os índices são 0 e 1,...,N nas colunas *Isótopo* e *Aduto*, temos a presença de um aduto putativo, cuja forma pode ser obtida pela referência à anotação original. As colunas *Traço* e *Pseudo grupo* fazem referência as tabelas originais para os arquivos de saída dos *softwares* xcms e CAMERA, respectivamente, e permitem o acesso direto aos dados brutos, para representações e inferências posteriores.

De posse de uma tabela de anotação dos íons podemos fazer a busca pela massa exata em bancos de dados, utilizando um modelo de classificação que leve em conta a pressuposição de uma amostra conectada, como apresentado na próxima seção.

4.2- Descrição do Modelo Probabilístico

Um modelo propondo a incorporação de reações entre os metabólitos na forma de possíveis transformações químicas foi proposto por Rogers et al., (2009). Os autores

introduzem a elegante ideia de utilizar um conjunto de reações químicas conhecidas entre compostos como forma de aumentar o poder de discriminação entre candidatos a estarem presentes na amostra, uma vez que certas combinações de compostos detectados juntos fariam mais sentido do que outras. As reações retiradas, por exemplo, de um compêndio manualmente inspecionado de reações bioquímicas, são utilizadas para que se atribua uma maior *crença* ou probabilidade de se classificar um metabólito como um composto que participe de um maior número de reações com os outros compostos presentes na mesma amostra do experimento (**Figura 6**).

Figura 6 - Representação esquemática da atribuição de picos de massa para fórmulas de compostos conhecidos, considerando sua inter-relação em vias metabólicas através de um modelo probabilístico. Com base apenas na massa do composto, a massa 315.0761 tem dois candidatos igualmente prováveis, Geranyl difosfato ($C_{10}H_{20}O_7P_2$) e (+)-Bornil-difosfato ($C_{10}H_{20}O_7P_2$). A atribuição de uma identidade para a massa 247.0135 ($C_5H_{12}O_7P_2$ – Isopentenil difosfato) aumenta a probabilidade de se atribuir o composto $C_{10}H_{20}O_7P_2$ para a massa 315.0761, pois Geranyl difosfato e Isopentenil difosfato são relacionados por uma reação bioquímica conhecida, como demonstrado no mapa da via metabólica do KEGG *Terpenoid backbone biosynthesis* (map00900). De forma semelhante, essa probabilidade pode ser aumentada em face da evidência do composto (2Z,6Z)-Farnesil difosfato ($C_{15}H_{28}O_7P_2$) como identidade para massa 405.1218.

Em um experimento de metabolômica onde foram encontrados, após pré-processamento que inclui etapas como calibração, normalização, remoção de ruído, entre outras, um conjunto de M picos no espectro massa/carga, denominados, por simplicidade, apenas massas: $\{x_1, x_2, x_3, \dots, x_M\}$. Dada uma lista de C compostos químicos conhecidos, de massas exatas $\{y_1, y_2, y_3, \dots, y_C\}$, o objetivo de uma identificação é atribuir cada pico de massa m a um único composto c , portanto uma relação unívoca $m \mapsto c$, com $m \in \{1, \dots, M\}$, $c \in \{1, \dots, C\}$.

Para expressar formal e matematicamente o que quer dizer que a relação unívoca $m \mapsto c$ foi estabelecida, utiliza-se um vetor booleano para cada pico: $\mathbf{z}_m = (z_{1,m}, z_{2,m}, z_{3,m}, \dots, z_{C,m})$ tal que $\mathbf{z}_m \in \{1,0\}^C$ e $z_{1,m} + z_{2,m} + \dots + z_{C,m} = 1$. A soma unitária garante que se uma das componentes do vetor $\mathbf{z}_m$ for 1, todas as outras devem ser 0. Ainda, a restrição similar $z_{c,1} + z_{c,2} + \dots + z_{c,M} = 1$ não existe pois é interessante permitir a flexibilidade de atribuir-se um mesmo composto c a mais de um pico de massa m uma vez que é possível que haja redundância nos picos de massa. Justamente a posição c do vetor $\mathbf{z}_m$ em que $z_{c,m} = 1$ sinaliza que a relação entre o pico m foi feita com o composto c .

Obter um modelo bayesiano completo para todas as probabilidades de todos os M picos conjuntamente seria exageradamente complexo. O Amostrador de Gibbs é um algoritmo que permite simular amostras que se comportam como especificado por uma distribuição de probabilidades sem exigir que se conheça a mesma (SUESS; TRUMBO, 2010). A teoria do Amostrador de Gibbs garante que conhecendo as distribuições marginais da conjunta é possível simular valores proporcionais a conjunta $f(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M \mid x_1, x_2, \dots, x_M)$ fazendo:

$$\begin{aligned}
\mathbf{z}_1^{(1)} &\leftarrow f(\mathbf{z}_1 \mid \mathbf{z}_2^{(0)}, \mathbf{z}_3^{(0)}, \dots, \mathbf{z}_{M-1}^{(0)}, \mathbf{z}_M^{(0)}, x_1, x_2, \dots, x_M) \\
\mathbf{z}_2^{(1)} &\leftarrow f(\mathbf{z}_2 \mid \mathbf{z}_1^{(1)}, \mathbf{z}_3^{(0)}, \dots, \mathbf{z}_{M-1}^{(0)}, \mathbf{z}_M^{(0)}, x_1, x_2, \dots, x_M) \\
\mathbf{z}_3^{(1)} &\leftarrow f(\mathbf{z}_3 \mid \mathbf{z}_1^{(1)}, \mathbf{z}_2^{(1)}, \dots, \mathbf{z}_{M-1}^{(0)}, \mathbf{z}_M^{(0)}, x_1, x_2, \dots, x_M) \\
&\dots \\
\mathbf{z}_M^{(1)} &\leftarrow f(\mathbf{z}_M \mid \mathbf{z}_1^{(1)}, \mathbf{z}_2^{(1)}, \mathbf{z}_3^{(1)}, \dots, \mathbf{z}_{M-1}^{(1)}, x_1, x_2, \dots, x_M) \\
\mathbf{z}_1^{(2)} &\leftarrow f(\mathbf{z}_1 \mid \mathbf{z}_2^{(1)}, \mathbf{z}_3^{(1)}, \dots, \mathbf{z}_{M-1}^{(1)}, \mathbf{z}_M^{(1)}, x_1, x_2, \dots, x_M) \\
\mathbf{z}_2^{(2)} &\leftarrow f(\mathbf{z}_2 \mid \mathbf{z}_1^{(2)}, \mathbf{z}_3^{(1)}, \dots, \mathbf{z}_{M-1}^{(1)}, \mathbf{z}_M^{(1)}, x_1, x_2, \dots, x_M) \\
&\dots \\
\mathbf{z}_M^{(2)} &\leftarrow f(\mathbf{z}_M \mid \mathbf{z}_1^{(2)}, \mathbf{z}_2^{(2)}, \mathbf{z}_3^{(2)}, \dots, \mathbf{z}_{M-1}^{(2)}, x_1, x_2, \dots, x_M) \\
\mathbf{z}_1^{(3)} &\leftarrow f(\mathbf{z}_1 \mid \mathbf{z}_2^{(2)}, \mathbf{z}_3^{(2)}, \dots, \mathbf{z}_{M-1}^{(2)}, \mathbf{z}_M^{(2)}, x_1, x_2, \dots, x_M) \\
&\dots \\
\mathbf{z}_M^{(3)} &\leftarrow f(\mathbf{z}_M \mid \mathbf{z}_1^{(3)}, \mathbf{z}_2^{(3)}, \mathbf{z}_3^{(3)}, \dots, \mathbf{z}_{M-1}^{(3)}, x_1, x_2, \dots, x_M) \\
\mathbf{z}_1^{(4)} &\leftarrow f(\mathbf{z}_1 \mid \mathbf{z}_2^{(3)}, \mathbf{z}_3^{(3)}, \dots, \mathbf{z}_{M-1}^{(3)}, \mathbf{z}_M^{(3)}, x_1, x_2, \dots, x_M) \\
&\dots \\
\mathbf{z}_{M-1}^{(B-1)} &\leftarrow f(\mathbf{z}_{M-1} \mid \mathbf{z}_1^{(B-1)}, \mathbf{z}_2^{(B-1)}, \mathbf{z}_3^{(B-1)}, \dots, \mathbf{z}_M^{(B-1)}, x_1, x_2, \dots, x_M) \\
\mathbf{z}_M^{(B)} &\leftarrow f(\mathbf{z}_M \mid \mathbf{z}_1^{(B)}, \mathbf{z}_2^{(B)}, \mathbf{z}_3^{(B)}, \dots, \mathbf{z}_{M-1}^{(B)}, x_1, x_2, \dots, x_M)
\end{aligned}$$

onde, o símbolo $\leftarrow$ é apenas uma notação que significa sortear um número aleatório segundo a lei de probabilidade dada por f e o índice sobrescrito, que varia de 1 a B, indica o número de simulações.

Após um número B suficientemente grande de ciclos, a teoria do Amostrador de Gibbs garante que os vetores numéricos obtidos: $S = \{ \{ \mathbf{z}_1^{(1)}, \dots, \mathbf{z}_M^{(1)} \}, \{ \mathbf{z}_1^{(2)}, \dots, \mathbf{z}_M^{(2)} \}, \dots, \{ \mathbf{z}_1^{(B)}, \dots, \mathbf{z}_M^{(B)} \} \}$ são governados pela lei de probabilidade conjunta desconhecida $f(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M \mid x_1, x_2, \dots, x_M)$.

Com a possibilidade de utilizar distribuições marginais para estimar uma distribuição conjunta, Rogers et al., (2009) apontam que a novidade do seu trabalho foi encontrar uma forma de escrever uma probabilidade *a priori* que codificasse a conexão entre os compostos candidatos por meio de reações químicas. A codificação é dada pela matriz binária W, uma matriz quadrada da relação dos compostos, onde compostos relacionados por alguma reação química recebem o valor:

$w_{ij}=1$ se existe uma reação química do tipo $i \rightleftharpoons j$ convertendo a massa y_i em y_j .

Na modificação proposta no presente trabalho foi introduzida a utilização de conexões presentes em bancos de dados de reações bioquímicas (**Figura 7**), em particular o banco KEGG (KANEHISA; GOTO, 2000), uma vez que a utilização de reações químicas genéricas poderia levar a conexões espúrias (BREITLING et al., 2006). Entretanto, a proposta de implementação pretende ser transparente e flexível de forma que o usuário possa fornecer uma matriz de conexões como a exemplificada acima, construída da forma que o mesmo julgar mais conveniente.

Figura 7 - Exemplo da codificação das reações metabólicas de compêndios de reações bioquímicas caracterizadas. No exemplo substrato, produto e intermediários são relacionados por uma matriz binária.

Partindo do pressuposto de que as identidades dos compostos, seriam mais provavelmente

a de compostos conectados por reações químicas, os autores enunciaram a seguinte probabilidade *a priori*

$$p(z_{cm}=1|Z^{(-m)}, W, \delta) = \frac{\delta + \beta_{cm}}{C\delta + \sum_{c'} \beta_{c'm}}$$

onde, $Z^{(-m)}$ é uma notação simplificada para $\{z_1, \dots, z_{m-1}, z_{m+1}, \dots, z_M\}$ colocando cada um dos vetores como colunas de uma matriz Z , δ delta representa um hyperparâmetro ou “parâmetro suavizador” garantindo que não se obtenha uma probabilidade 0 caso não haja conexão entre fórmulas e ajustando o balanço entre conexão e proximidade de massas, C representa o número de compostos, e β_{cm} é o número de formas diferentes que existem para se obter o composto c a partir de todos os outros compostos não- c que foram já atribuídos aos picos de massa.

Esta formulação utiliza a contagem do número de reações para aumentar a probabilidade de um dado composto ser atribuído a uma dada massa. A probabilidade de um dado composto ser atribuído a uma massa será tanto maior quanto maior for o número de reações em que estão envolvidos os outros compostos sendo atribuídos pela presente configuração da matriz Z . É dessa forma que o modelo acaba privilegiando atribuições que, em conjunto, fazem mais sentido do que tomando somente os resultados experimentais e de ruído dos picos de massa (**Figura 6**).

Desta forma, a cada iteração do algoritmo de amostragem (ver a descrição do Amostrador de Gibbs abaixo) a matriz é rearranjada para atribuir as identidades de cada metabólito, com favorecimento aos compostos que possuem um número maior de reações, levando a um rearranjo global da matriz Z , que privilegia o cenário de uma amostra conectada.

Uma função $L(.)$ de verossimilhança em um modelo bayesiano deve ser escolhida de forma a refletir a incerteza da medição de uma massa x_m para o pico m , supondo-se que o pico m é de fato o composto c . Em outras palavras, um modelo de erro ou ruído para a medição de uma massa pelo equipamento, ou seja, a probabilidade do pico m ser de fato o composto c sem utilizar o conhecimento prévio do especialista sobre o fenômeno, somente os dados obtidos experimentalmente.

No que se refere a $L(.)$, Rogers et al., (2009) escolheram um modelo gaussiano de ruído,

amplamente utilizado na literatura: $x_m \sim N(y_c, y_c^2 \sigma^2)$ com a variância proporcional a média. Para facilitar a representação e para utilizar a precisão do equipamento γ , mais comum de ser informada do que a variância $\sigma^2 = \gamma^{-1}$, os autores promoveram uma simples mudança de variáveis nesta densidade, tornando-a mais enxuta: $x_m \sim N(y_c, y_c^2 \sigma^2) \Leftrightarrow (x_m / y_c) \sim N(1, \gamma^{-1})$. Os valores numéricos obtidos com qualquer uma das duas opções são estritamente os mesmos, evidentemente.

A distribuição de probabilidade f do modelo pode ser formalizada como

$$f(\vec{z}_m | Z^{(-m)}, \mathbf{x}, \boldsymbol{\theta}) \propto L(x_m | Z, \mathbf{x}) \cdot p(\vec{z}_m | Z^{(-m)}, \mathbf{x}, \boldsymbol{\theta})$$

onde, $L(\cdot)$ e $p(\cdot)$ representam as funções de verossimilhança e probabilidade *a priori* propostas por (ROGERS et al., 2009), apresentadas em detalhes anteriormente, e o vetor $\boldsymbol{\theta}$ codifica os parâmetros necessários, também descritos anteriormente.

Expandindo o modelo de verossimilhança $L(\cdot)$ em termos multiplicativos de acordo com (ROGERS et al., 2010), foram implementadas as funções para predição do tempo de retenção (CREEK et al., 2011) e predição do padrão isotópico (WEBER et al., 2011) que representam fontes independentes de informação, na forma:

$$L(x_m | Z, \mathbf{x}, \boldsymbol{\theta}) \propto L_N(x_m | Z, \mathbf{x}, \boldsymbol{\theta}_N) \cdot L_{rt}(x_m | Z, \mathbf{x}, \boldsymbol{\theta}_{rt}) \cdot L_{iso}(x_m | Z, \mathbf{x}, \boldsymbol{\theta}_{iso})$$

onde, $L_N(\cdot)$ representa o modelo ruído, $L_{rt}(\cdot)$ representa o modelo dos erros de predição para o tempo de retenção e $L_{iso}(\cdot)$ é o modelo de filtro de fórmulas por padrão isotópico predito. Novamente os vetores $\boldsymbol{\theta}_*$ representam os parâmetros apropriados para cada função, como descrito acima.

Recentemente (CREEK et al., 2011) utilizaram um modelo de regressão linear, com os descritores das características físico-químicas dos compostos como variáveis independentes e o tempo de retenção com variável dependente, para realizar a predição do tempo de retenção e auxiliar a anotação de metabólitos. Em sua abordagem, compostos candidatos aos metabólitos que apresentassem um erro de predição do tempo de retenção maior que 35% eram descartados, reduzindo o número de candidatos a serem analisados. Para permitir a

incorporação dos dados de forma semelhante, a função de verossimilhança $L_{rt}(\cdot)$ retorna valores crescentes à medida que os erros de predição dos tempos de retenção decrescem.

Weber et al., (2011) demonstraram, para plataforma de espectrometria de massas Orbitrap (Thermo Fisher), que a abundância relativa de $^{13}\text{C}^{12}\text{C}_{(n-1)} / ^{12}\text{C}$ tem um desvio de sua proporção teórica que é dependente da razão SNR de cada pico de massas. Utilizando este conhecimento podemos introduzir um filtro como um termo multiplicativo ao modelo probabilístico. Para tanto, basta tornar a função $L_{iso}(\cdot) < 1$ para fórmulas de compostos candidatos aos metabólitos que possuem o número de carbonos fora do intervalo predito pela abordagem descrita por (WEBER et al., 2011).

Seguindo a mesma notação, propomos uma forma alternativa a formulação da função de verossimilhança proposta por (ROGERS et al., 2009) como um modelo $L_N(\cdot)$ representando o modelo proposto por (BÖCKER et al., 2009), com a seguinte forma:

$$L_N(x_m|Z, \mathbf{y}, w) \propto 1_\omega \left(1 - \Phi \left(\frac{1}{\sigma_\omega} (|x_m / y_c - 1| - \mu_\omega) \right) \right)$$

onde, 1_ω representa uma função indicadora para o conjunto ω composto por todos os j compostos candidatos na base de dados para os quais $|x_m - y_j| \leq w$, μ_ω e σ_ω representam a média e o desvio padrão para todas as massas do vetor $\mathbf{y}$ listadas em ω , Φ representa a função cumulativa da distribuição normal, e w é o intervalo admitido da diferença de massas indicada pelo especialista. O conjunto ω inibe que compostos altamente conectados na rede tenham sua probabilidade *a posteriori* inflada em sua atribuição à massas que apresentam o erro da massa medida altamente improvável para o equipamento utilizado no experimento.

4.3- Representação do metabolismo

A visualização de redes biológicas vem sendo sistematicamente estudada (SUDERMAN; HALLETT, 2007; MERICO et al., 2009). Estes estudos têm como principais objetivos: representar um grande volume de interações permitindo simultaneamente a inspeção manual e o acesso à informações da topologia da rede e permitir a identificação de módulos (vias metabólicas) dentro das redes que possam ser associados a eventos biológicos de interesse.

Ferramentas populares como *Pathway projector* e *Pathos* (KONO et al., 2009; LEADER et al., 2011) se baseiam em topologias de referência, inferidas sob a pressuposição de conservação das vias entre organismos (KANEHISA; GOTO, 2000), enquanto outros como *MetaNetter* e *MetaMapp* (JOURDAN et al., 2008; BARUPAL et al., 2012) permitem a inferência em redes utilizando o conhecimento das reações bioquímicas que compõem o metabolismo, mas produzem redes de difícil navegação e dissociadas do conhecimento clássico das vias metabólicas.

Para tratar os problemas citados na visualização e exploração do metabolismo, propomos um algoritmo para representação da distribuição *a posteriori* da anotação de metabólitos em redes metabólicas utilizando grafos, de maneira que a informação seja apresentada de forma condensada e seja dinamicamente navegável.

Um grafo é uma representação de dados que utiliza um conjunto de nós e um conjunto de arcos que conecta estes nós. Os nós são entidades de interesse e os arcos representam relações entre estas entidades (HUBER et al., 2007). Um grafo é especificado por um conjunto de nós V e um conjunto de arcos E . Cada elemento de E contém um par de elementos u, v de V . O relacionamento modelado pelos arcos pode ser dicotômico, o arco está presente ou ausente, ou pode se considerar uma representação genérica de E representada por uma função $f : V \times V \rightarrow F$ com contradomínio F discreto ou contínuo. De acordo com Heuber et al., (2007) se $F \subset \mathbb{R}$, então o valor $f(u,v)$ é denominado peso do arco entre u e v .

De posse do referencial teórico da representação por grafos podemos definir um algoritmo de representação onde os nós correspondem a todos os possíveis candidatos de uma dada massa e os arcos, do tipo dicotômico, como descrito acima, estão presentes se existe pelo menos uma reação bioquímica conhecida entre dois possíveis candidatos dos nós em questão.

Seja um conjunto de atribuição de picos de massa a possíveis compostos, com uma distribuição *a posteriori* como o sugerido acima:

Massa	Possíveis candidatos	Probabilidade	Intensidade amostra 1	...	Intensidade amostra N
x_1	y_1	$p_{1,1}$	I_{11}	...	I_{1N}
	y_2	$p_{1,2}$	...	...	...
	y_3	$p_{1,3}$	...	...	...
...	...	...	...	...	...
x_m	y_c	$p_{m,c}$	I_{m1}	...	I_{mN}
	y_{c+1}	$p_{m,c+1}$			
...	...	...			
x_M	y_C	$p_{M,C}$	I_{M1}		I_{MN}

Uma maneira de realizar uma verificação *post hoc* das conexões preditas pelo modelo probabilístico e uma visualização inspirada na natureza de vias metabólicas pode ser realizada sobrepondo a informação de reações a correlações. Assumindo que um pico de massa possa ter um ou mais compostos candidatos, e cada composto possa ter uma ou mais conexões a compostos candidatos de outras massas, a geração e inspeção de todas as redes de conexões seria inviável, assim como a inferência da sua distribuição de probabilidade conjunta no modelo probabilístico. Em vez de gerar todas as redes de conexão possíveis, tratamos cada massa como um nó, e todas as possíveis reações entre os candidatos para a massa de um dado nó com os candidatos para outra massa como um arco (**Figura 8**).

Figura 8 - Representação conceitual da visualização de uma rede para anotação probabilística do metaboloma. Os nós representam picos de massa e as arestas são desenhadas apenas se houver pelo menos uma reação bioquímica transformando um composto anotado em algum outro composto anotado. Nesta representação, m_i , m_k e m_j são picos de massa observados, as letras de A a D representam compostos candidatos para anotar m_i com probabilidades $p_a \leq \dots \leq p_d$. Analogamente, E a I são atribuídos com probabilidades p_e a p_i . Uma vez que existe uma reação bioquímica conhecida metabolizando o segundo melhor candidato para m_i , B, a I, uma aresta é então desenhada de m_i para m_j . Não há reações conhecidas de qualquer candidato de m_i para de qualquer candidato do m_k , portanto, nenhuma aresta é desenhada.

Essa abordagem simplificada diminui o número de grafos a serem analisados para apenas um, entretanto, os nós ainda possuem um alto grau de conectividade, dificultando a navegação e visualização. Como uma estratégia de priorização de conexões mais plausíveis, a sobreposição dos arcos (que representam todas as possíveis reações bioquímicas conhecidas) a informação de relações quantitativas entre os picos de massa que representam os nós, pode ser realizada quando temos acesso a experimentos com réplicas ou séries temporais. Medidas de associação linear entre variáveis são comumente utilizadas em metabolômica (STEUER, 2006). Dentre estas, a mais utilizada é a correlação de Pearson, dada por:

$$\rho_{ij} = \frac{\sum_{k=1}^n (I_{ki} - \bar{I}_i)(I_{kj} - \bar{I}_j)}{\sqrt{\sum_{k=1}^n (I_{ki} - \bar{I}_i)^2} \sqrt{\sum_{k=1}^n (I_{kj} - \bar{I}_j)^2}}$$

onde, I representa as intensidades de um par i,j de picos de massas, em n réplicas experimentais ou pontos em uma série temporal.

Em cenários onde possuímos mais amostras n que metabólitos m para os quais queremos estimar correlações parciais, estas podem ser facilmente obtidas (KRUMSIEK et al., 2011). Entretanto, na maioria dos problemas dos experimentos de metabolômica $m \gg n$ e não possuímos graus de liberdade suficientes para estimar a significância da correlação parcial entre i e j excluindo o efeito das outras $m-2$ variáveis. Por isso utilizamos a teoria apresentada em (SCHÄFER; STRIMMER, 2005) para *i*) calcular estimativas pontuais de correlação parcial para pequenas amostras experimentais; *ii*) estimar a taxa de falsos positivos das estimativas anteriores.

Com a criação do grafo de conexões bioquímicas e a sobreposição de correlações, podemos criar um grafo resumido, evitando o clássico problema de visualizações complexas em *hairball* (LONGABAUGH, 2012), e sobrepor informação ainda não modelada em um esquema de visualização dinâmica. Para navegar por estes dados criamos um formato compatível com uma ferramenta de representação e exploração de grafos amplamente difundida na bioinformática, o *software* Cytoscape (SHANNON et al., 2003).

Para integrar a exploração de redes de dados experimentais ao conhecimento tradicional de vias metabólicas implementamos o algoritmo de busca de vias (*Search Pathways*) disponível na ferramenta KEGG Mapper tools (KANEHISA et al., 2012). Esse algoritmo possibilita a busca de todas as vias em que cada composto candidato está envolvido, e a disponibilização da contagem do número de compostos em cada via. Com esta informação podemos criar mapas de vias de interesse e investigar o contexto metabólico de compostos anotados para um experimento de interesse. Estas representações serão amplamente utilizadas nas seções finais desta tese nas Figuras 11, 12 e 13.

4.4- Algoritmo de amostragem e implementação

De posse de um modelo de verossimilhança, a abordagem proposta por (ROGERS et al., 2009) estima a probabilidade *a posteriori* realizando iterações para calcular a probabilidade de cada massa ser atribuída a cada um dos compostos, dado que todas as outras massas tenham sido atribuídas a compostos que estão interconectados. A proposição da nova verossimilhança levou a conclusão de que esta deve ser calculada por funções externas ao simulador, proporcionando maior flexibilidade a incorporação das informações independentes ao modelo. Deste modo, as modificações ao modelo estão destacadas com sublinhado no que segue

Dados de Entrada: M massas mensuradas, $x_1, \dots, x_m, \dots, x_M$; C fórmulas empíricas potenciais com massas $y_1, \dots, y_c, \dots, y_C$; uma matriz de conexões W ; hiper parâmetros γ e δ ; matriz p com verossimilhanças dadas por $L(\cdot)$;

Dados de Saída: Um conjunto S de matrizes de atribuição $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \mathbf{Z}^{(3)}, \dots, \mathbf{Z}^{(B)}$.

Algoritmo:

Inicialize $\mathbf{Z} = \mathbf{Z}^{(0)}$ atribuindo randomicamente fórmulas para massas;

Calcule $\beta = \mathbf{WZ}^T$

para cada amostra s faça

para cada massa m faça

calcule $\beta_m = \beta - \mathbf{Wz}_m$

Determine o fator de normalização z_3 :

$$z_3 = \sum_{c=1}^C \underline{p}_m \cdot (\delta_3 + \beta_{mc})$$

Amostre um composto c de uma distribuição discreta

onde a probabilidade do c -ésimo composto é proporcional ao peso $q_c^{(s)}$:

$$z_3^{-1} \cdot \underline{p}_m \cdot (\delta_3 + \beta_{mc})$$

Insira a nova atribuição em $\mathbf{z}_m$

Calcule $\beta_m = \beta + \mathbf{Wz}_m$

fim

construa $\mathbf{Z}^{(s)}$ concatenando todos os $\mathbf{z}_m$ obtidos em colunas

fim

Algoritmo do amostrador de Gibbs, modificado de (ROGERS et al., 2009).

Para a implementação do modelo apresentado a linguagem de *script* R (R DEVELOPMENT CORE TEAM, 2010) foi escolhida, pela ampla disponibilidade de algoritmos para análise de dados de espectrometria de massas, bem como pela disponibilidade de código aberto, possibilitando amplo acesso à comunidade de bioinformática.

5- RESULTADOS E DISCUSSÃO

5.1- Construção de uma biblioteca para o ambiente R

Utilizando o modelo de atribuição de probabilidades apresentado na seção anterior, com a modificação no algoritmo para adição de informação e o modelo de representação gráfica proposto, foi construída uma biblioteca de métodos para o ambiente R, denominada **ProbMetab** (**Prob**abilistic **Metab**olomics). Seu escopo dentro da análise de dados de metabolômica é ilustrado na (Figura 9).

Figura 9 - Representação esquemática do escopo do pacote ProbMetab na análise de dados de espectrometria, evidenciando a integração em um ambiente de programação que proporciona diferentes soluções para análises associadas e integração de informações ao modelo.

Um dos grandes apelos do ambiente de programação R para colaboração científica se baseia no ferramental disponível para documentação, potencializando o desenvolvimento conjunto em um ambiente multidisciplinar e uma utilização mais eficiente para o usuário final. Para realizar a implementação dentro das recomendações do CRAN (*The*

Comprehensive R Archive Network), apresentando uma documentação padronizada, foram utilizadas ferramentas amplamente disponíveis através do conjunto de funções do pacote *devtools* (Wickham, 2011 - <https://github.com/hadley/devtools/wiki/>).

A primeira etapa das análises em dados de LC-MS faz uso de extensões do ambiente R para o pré-processamento dos espectros (SMITH et al., 2006), e filtragem das diferentes formas químicas dos íons (isótopos, junções e quebras) (SCHELTEMA et al., 2011; KUHL et al., 2011), como ilustrado na (**Figura 9**). Os métodos desenvolvidos até o momento recebem a lista final de íons não redundantes no formato do objeto CAMERA::xsAnnotate (arquivo de saída do *software* CAMERA). A etapa subsequente trata da busca em bancos de reações biológicas catalogadas, para modelar a interação entre os compostos. Como ponto de partida, os bancos de dados utilizados foram KEGG e MetaCyc (ALTMAN et al., 2013), por sua abrangência e disponibilidade de uma interface de acesso para aplicações pela web (*API – Application Programming Interface*).

Após a etapa de anotação do íon e busca a banco de dados passa-se a etapa do cálculo das probabilidades para atribuição de cada metabólito a uma lista de compostos. Utilizamos o modelo proposto por (ROGERS et al., 2010), com os modelos de verossimilhança para tempo de retenção, acurácia da massa e padrão isotópico. Sua incorporação ao modelo foi realizada pela multiplicação de funções de verossimilhança crescentes. O produto das funções de verossimilhança e a matriz de reações bioquímicas são utilizados como entrada para o modelo, e as probabilidades *a posteriori* são estimadas com base nas observações simuladas obtidas pelo Amostrador de Gibbs. Para obter um melhor desempenho no tempo de computação, o amostrador de Gibbs foi implementado na linguagem c++ e integrado ao R pela API do pacote Rcpp (EDDELBUETTEL; FRANÇOIS, 2011).

Uma vez obtido o ranqueamento probabilístico é necessária a representação eficiente dos resultados de longas listas de atribuição, que representam processos complexos do metabolismo. Para tratar a representação da forma mais abrangente possível desenvolvemos dois arquivos de saída. O primeiro arquivo consiste em uma tabela de probabilidades para cada possível atribuição de uma massa a um composto, disponibilizada no formato de uma matriz no ambiente R, que pode ser opcionalmente exportada no formato *html*. Esta tabela é exportada com informação adicional dos cromatogramas individuais para cada pico de massa, anotação do íon e ligação a interface web do banco de dados com informações adicionais dos

A matriz de classificação exportada pelo pacote ProbMetab apresenta as colunas: Massas Propostas (*Proposed Mass*); lista de compostos candidatos ranqueados (*Most Probable*

Compound); probabilidade de cada candidato (*Probability*) – o ranqueamento de probabilidades é interpretado como o ranqueamento mais provável de acordo com os pressupostos do modelo; entropia da distribuição de probabilidade entre os candidatos (*entropy* - da teoria da informação); opcionalmente, o valor-p do t-test/anova entre as amostras (*p-value*), e uma anotação condensada dos íons no formato: massa original # tempo de retenção original # padrão isotópico, se houver # aduto, se houver. O usuário pode optar por exportar uma tabela *html*, que será associado aos cromatogramas de íons extraídos (EIC - *Extracted Ion Chromatogram*) para todas as massas propostas, salvos para o diretório de trabalho do R.

É importante notar que a interpretação dos valores de probabilidade exportados é diferente da probabilidade de um pico de massa ser um dado composto, uma vez que o modelo não restringe o espaço de busca para o verdadeiro metaboloma, nem garante que um pico de massa se refere diretamente a massa de um metabólito. A probabilidade a que se refere a distribuição obtida é referente às opções disponíveis, isto é, aos eventos no espaço amostral definido para cada pico de massa.

A implementação, extensa documentação para instalação e utilização, conjuntos de dados e exemplos de aplicação, estão disponíveis no portal do LabPIB – Laboratório de Processamento de Informação Biológica para o livre acesso da comunidade de bioinformática (<http://labpib.fmrp.usp.br/methods/probmetab/>).

5.2- Validação e desempenho do modelo

Para testar a eficiência da implementação e o impacto das novas funções de verossimilhança e de representação, o pacote ProbMetab foi testado analisando dados públicos de *Trypanosoma brucei* (VINCENT et al., 2012). Paralelamente foi realizada a análise com o *software* IDEOM para comparação do número de picos filtrados e ilustração da necessidade de um ranqueamento dos compostos, e com o *software* MetSamp (<http://www.dcs.gla.ac.uk/inference/metsamp/>), implementação em MATLAB original de Rogers et al., (2009), para validação da implementação e comparação do tempo de computação das probabilidades.

O conjunto de dados de *T. brucei* foi, em sua publicação original, previamente analisado com o *software* IDEOM (CREEK et al., 2012), e, de acordo com o manual do mesmo “se a atribuição de um metabólito para um composto permanece ambígua, o primeiro composto do banco de dados é atribuído como a identidade do metabólito”, resultando em um ranqueamento estático, que pode não ser apropriado para diferentes cenários de análises.

Foi possível recuperar o mesmo número de picos que o *software* IDEOM recuperou utilizando o fluxo de pré-processamento do pacote mzMatch.R, entretanto, o ranqueamento fornecido pelo pacote ProbMetab permite uma classificação dinâmica que pode ser adaptada para diferentes contextos metabólicos. Utilizando o algoritmo para obtenção da massa neutra do composto (comb 2+ **Tabela 7**), proposto nesta tese, foi possível recuperar 93 compostos com identidade conhecida, entre os 127 previamente identificados (VINCENT et al., 2012).

Tabela 7 - Anotação de picos para o pré-processamento com pacote xcms.

	xcms			CAMERA			ProbMetab		
	Picos	Picos por amostra	Grupos de picos	Grupos	isótopos	adutos	íons propostos	isótopos	adutos
pos	557071	24220	15270	8701	1595	2819	2546	1081	1465
neg	425602	18504	13226	7506	1039	2596	2111	737	1374
comb 1							4789/4096	1081/737	3708/3359
comb 2							4408	1718	2690
comb 2 +							8707	1718	6989
comb 3							6714	276	6438

pos/neg – modo de aquisição; **comb1/2/3** – estratégias para combinar modos de aquisição: 1 – anotação gerada pela função combinexsAnnos do pacote CAMERA; 2 – anotação gerada pela função combMolIon do pacote ProbMetab, com parâmetro opcional para incluir íons não anotados (+); 3 – Integração ao pacote mzMatch, utilizando apenas as relações “bp” e “potential bp”.

A identificação baseada no tempo de retenção, com a utilização de compostos purificados, apresentada por Creek et al., (2011) é muito interessante no contexto de uma análise direcionada, onde um conjunto de compostos é eleito para monitoramento. Entretanto, esta abordagem tem as desvantagens do grande dispêndio de recursos financeiros, análises muito laboriosas, limitado número de compostos purificados disponíveis, além da dificuldade inerente da predição do tempo de retenção (KALISZAN, 2007).

Apesar das dificuldades citadas, a predição do tempo de retenção, como fonte de informação adicional, tem se mostrado promissora (CREEK et al., 2011; GORYŃSKI et al., 2013). O ranqueamento para o modelo “Apenas tempo de retenção” apresentado na **Tabela 8**

representa a avaliação do desempenho do modelo de retenção em um conjunto de compostos com identidade conhecida. A classificação baseada no erro da predição do tempo de retenção não se mostrou eficiente, com um baixo número de compostos classificados corretamente 19.2%, e um alto número de identidades incorretas 17%. Na **Tabela 9** temos como exemplo um caso particular em que o tempo de retenção foi a única informação capaz de distinguir a identidade correta para o pico de massa entre os compostos *N6-Acetyl-N6-hydroxy-L-lysine* e *N5-(L-1-Carboxyethyl)-L-ornithine*, que possuem a mesma fórmula molecular e não são diferenciados pelas suas reações putativas com os outros compostos na amostra. Este exemplo ilustra como diferentes fontes de informação podem auxiliar a classificação. Se a predição do tempo de retenção pudesse ser melhorada, ou utilizada para um subconjunto de metabólitos com menor erro de predição, a classificação final poderia ser melhorada.

Tabela 8 – Comparação da adição de diferentes fontes de informação ao modelo probabilístico.

Classe	Apenas tempo de retenção	Apenas massa	Conexões MetSamp	Conexões ProbMetab	Todos
Todos candidatos são equiprováveis	44 (56%)	68 (84,6%)	0 (0%)	0 (0%)	0 (0%)
Identidade correta é a maior probabilidade	15 (19.2%)	6 (7,7%)	47 (60,2%)	50 (64,1%)	49 (62,8%)
Identidade correta não é a maior probabilidade	17 (21.8%)	4 (5,1%)	31 (39,8%)	28 (35,9%)	29 (37,2%)
Total	78 (100%)	78 (100%)	78 (100%)	78 (100%)	78 (100%)

Apenas tempo de retenção – classificação baseada no erro do tempo de retenção predito; **Apenas massa** – Classificação baseada no padrão isotópico (quando presente) e verossimilhança das massas (função *erfc*); **Conexões MetSamp** - Modelo com verossimilhança para massas e conexões bioquímicas; **Conexões ProbMetab** - Modelo com verossimilhança para massas e conexões bioquímicas (implementação do R); **Todos** – classificação baseada no padrão isotópico (quando presente), verossimilhança das massas, erro do tempo de retenção predito e conexões bioquímicas.

Utilizando apenas a acurácia da massa (“Apenas massa” - **Tabela 8**) obtivemos uma baixa definição na classificação, apenas 7,7%, principalmente porque muitos dos candidatos são isômeros. Os picos incorretamente classificados pela acurácia da massa podem ser explicados por desvios maiores que os preditos, os quais ocorrem em baixa frequência em espectros de massas (ALLWOOD et al., 2013), além de desvios causados por parâmetros de pré-processamento subótimos.

Entre os compostos com identidades conhecidas para os quais foi realizada a predição do

número de carbonos baseada na proporção isotópica, apenas 30, de 93, possuíam um padrão isotópico putativo automaticamente recuperado utilizando os parâmetros padrão do pacote CAMERA. Entre estes, apenas três massas apresentaram compostos candidatos cujas fórmulas possuíam número de carbonos diferente. A função implementada para filtrar fórmulas possibilitou a discriminação de apenas uma fórmula, $C_9H_{15}N_4O_9P$ (*5-Amino-6-(5'-phosphoribosylamino)uracil*), que teve a sua probabilidade de ser atribuída a massa 354,056 diminuída, fixada como diminuição de 10 vezes, mas mantendo o candidato na lista de possíveis candidatos para inspeção manual (**Tabela 9**).

O modelo “Conexões ProbMetab” (conhecimento prévio do metabolismo) proporcionou o melhor resultado, considerando o compromisso entre identidades corretas 64,1% e um baixo número de identidades incorretas 35,9%, levando-se em conta que apenas o composto de maior probabilidade foi atribuído a massa. No modelo utilizando todos os componentes estudados (“Todos” na **Tabela 8**) proporcionou um desempenho pior, pela influência das atribuições incorretas do modelo de tempo de retenção.

Em uma análise detalhada das atribuições de todos os picos de massa de identidade conhecida foi possível observar que a classificação entre atribuição correta ou incorreta apresenta, para alguns picos de massa, uma diferença de probabilidade muito pequena, como por exemplo, a atribuição do composto *N6,N6,N6-Trimethyl-L-lysine* como identidade correta (**Tabela 9**). No cenário apresentado, o mais razoável seria supor que não existe informação suficiente modelada para fornecer uma única classificação correta, e informações adicionais em conjunto a inspeção manual são necessárias para determinar a melhor anotação a ser utilizada.

Tabela 9- Exemplos da atribuição de probabilidades para diferentes componentes do modelo, com a identidade correta destacada em negrito.

Nome do composto	Massa	Probabilidade associada a cada componente do modelo				Todos
		Apenas tempo de retenção	Apenas Massa	Conexões MetSamp	Conexões PtoMetab	
N5-(L-1-Carboxyethyl)-L-ornithine	204,1106330486	0,081	0,5	0,69	0,672	0,09
N6-Acetyl-N6-hydroxy-L-lysine		0,919	0,5	0,31	0,328	0,91
Phenolsulfonphthalein	354,0558718467	0,25	0,839	0,654	0,926	0,905
5-Amino-6-(5'-phosphoribosylamino)uracil		0,25	0,115	0,255	0,023	0,029
WIN56291		0,25	0,046	0,064	0,045	0,06
2-Caffeoylisocitrate		0,25	0	0,026	0,006	0,006
7,8-Diaminononanoate	188,1521095659	0,5	0,5	0,472	0,446	0,444
N6,N6,N6-Trimethyl-L-lysine		0,5	0,5	0,458	0,48	0,482
L-Histidine	155,0688293244	0,2	0,2	0,792	0,781	0,798
3-(Pyrazol-1-yl)-L-alanine		0,2	0,2	0,082	0,106	0,092
D-Histidine		0,2	0,2	0,043	0,034	0,036
Kininogen		0,2	0,2	0,041	0,039	0,043
Histidine		0,2	0,2	0,041	0,04	0,031

Os exemplos discutidos acima ressaltam que diferentes fontes de informação podem contribuir de diferentes maneiras para a classificação, e a interpretação dos resultados deve ser auxiliada com o máximo de informações disponíveis. O pacote ProbMetab permite a incorporação sistemática de informações, uma vez que sabemos como modelar uma fonte de informação do universo de metabólitos, que um determinado experimento de metabolômica pode amostrar. Levando em consideração a taxa de erro de todos os componentes do modelo, o usuário pode capitalizar o potencial da anotação ao interpretar os arquivos de saída com o ranqueamento de classificação resumido.

A implementação prévia do modelo probabilístico (ROGERS et al., 2009), MetSamp, uma aplicação para MATLAB (cujo código foi gentilmente cedido pelo autor Simon Rogers, após requisição pessoal) foi testada para os dados de *T. brucei* com o parâmetro precisão ajustado para 6.25×10^{12} (equivalente a desvio padrão igual a 4 ppm), e os vetores de massas x e de compostos candidatos y extraídos da matriz de anotação gerada previamente pelo pacote ProbMetab. Após o processamento, a matriz de contagem de atribuições foi exportada manualmente do MATLAB e utilizada para o cálculo das probabilidades apresentados na **Tabela 10**.

Na implementação na plataforma R colocamos em evidência a importância de informar o ranqueamento associado a informação ainda não modelada (padrão de adutos, definição do pico cromatográfico, correlação entre as intensidades dos picos de massa), estas informações são essenciais para fornecer ferramentas que permitam ao biologista decidir a melhor anotação para o pico de massa, selecionando os melhores candidatos potencialmente envolvidos na alteração metabólica sob estudo. Na **Tabela 10** é possível observar que, considerando todo o ranqueamento, ou seja a distribuição posterior completa, e não apenas a classificação final, até 90% das atribuições tem sua identidade correta entre as três probabilidades mais altas. A utilização deste ranqueamento associado às estratégias de visualização, apresentadas em sequência, possibilitam uma anotação de alta qualidade.

Tabela 10 - Proporção cumulativa da posição da identidade correta em diferentes modelos.

Posição da Identidade correta	Proporção acumulada de identidades corretas		
	MetSamp	ProbMetab	Todos
1	0,667	0,699	0,688
2	0,828	0,828	0,839
3	0,882	0,903	0,882
4	0,925	0,935	0,935
5	0,935	0,957	0,946
6	0,946	0,978	0,968

5.3- Representação abrangente dos resultados

O principal produto do modelo probabilístico é uma lista de atribuições ranqueadas, que representa nossa compreensão do experimento e do fenômeno biológico subjacente. As listas podem ser extensas e os usuários necessitam de uma representação apropriada para extrair informações de conjuntos de dados de larga escala. O pacote ProbMetab cria uma tabela com atribuições que pode ser utilizada para inferência dentro do ambiente R, bem como exportada, de forma associada a diferentes fontes de informação (**Figura 10**).

O pacote ProbMetab fornece um algoritmo para sobrepor informações associadas aos picos de massa e exportá-las para o *software* Cytoscape. Com a sobreposição das reações preditas pelo modelo a informações adicionais, podemos reunir evidências para confirmação da identidade do composto e suas vias associadas. Inspeccionando apenas compostos que possuem uma correlação em modulo maior que 0,75, podemos verificar como a sobreposição de informação pode ser utilizada observando os compostos para os quais conhecemos as identidades verdadeiras. Para os 32 picos de massa que possuíam correlações acima do limiar estabelecido, a identidade dos compostos foi confirmada por inspeção manual das reações preditas, como ilustrado na **Figura 11**. Dentro do subconjunto de 32 picos de massa, 20 (62,5%) apresentaram confirmação da identidade pela inspeção das reações. É possível observar que quando existe uma correlação entre um dado nó x e outro nó y as reações conhecidas entre as possíveis identidades dos nós levam a um número de possíveis identidades menor do que aqueles sugeridos pela comparação da massa exata ao banco de dados.

A representação das redes metabólicas em uma ferramenta de edição e navegação dinâmica é essencial para descoberta e interpretação de informações relevantes ao fenômeno estudado. A conexão entre informações sumarizadas, como atributos de nós e arcos, permite a utilização de ferramentas de busca padrão do Cytoscape, possibilitando a filtragem de vias metabólicas de interesse. Ainda, a representação com a versão *desktop* do Cytoscape, o pacote ProbMetab possibilita a representação sem a necessidade de deixar o ambiente do R e sem a instalação do Cytoscape, utilizando o Cytoscape Web (<http://labpib.fmrp.usp.br/methods/probmetab/#screencast>).

Figura 11 - Representação da tabela de classificação de picos final exportada como uma rede ponderada por correlações sobreposta a uma rede de reações bioquímicas. A principal informação importada para a janela do Cytoscape pode ser observada no painel de dados (*Data Panel*). Utilizando o algoritmo de ProbMetab o nó 425 (massa 155,06 na Tabela 7) possui cinco identidades possíveis e cada identidade possui um conjunto de possíveis reações com outros candidatos. Entretanto, as únicas reações que coincidiram com as correlações são reações do composto L-Histidina.

O conjunto de dados de *T. brucei* foi utilizado em sua publicação original para discutir o efeito da droga *Nifurtimox* (Figure S3) e os autores afirmam que *Nifurtimox* (identificada como o pico massa: 287,0577, RT: 5,25 minutos) foi observada em todas as amostras tratadas,

em adição o pico de massa: 257,0834, RT: 13,5 minutos, atribuído a um composto derivado da droga *Nifurtimox*. Na análise realizada pelo pacote ProbMetab os picos foram encontrados automaticamente (picos 105 e 599), mas nenhuma reação conhecida entre os compostos foi encontrada no banco de dados utilizado (KEGG). A análise do grafo mostra uma grande representação de compostos ligados a biossíntese de aminoácidos, processo de interesse investigado sob o efeito da droga no trabalho original.

5.4- Aplicação a dados originais: análises de espectros de massa de cana-de-açúcar

Para obter uma validação adicional do fluxo de análises implementadas no pacote ProbMetab em um experimento obtido por uma diferente plataforma experimental, em um organismo de maior complexidade e sem a utilizações de padrões purificados, dados de cana-de-açúcar foram analisados. Com as análises apresentadas para os dados de cana-de-açúcar ressaltamos como a escolha correta dos parâmetros de pré-processamento devem ser adaptados à plataforma experimental, e mostramos como a distribuição de probabilidade *a posteriori* pode ser utilizada em análises exploratórias para determinação de vias metabólicas alteradas sob as condições estudadas.

Como projeto associado ao Projeto Temático 2008/56100-5 Programa BIOEN-FAPESP “Metabolômica da cana-de-açúcar e descoberta de novas enzimas hidrolíticas para produção de biocombustíveis”, a presente tese utilizou dados gerados pelo Laboratório Max Feffer (<http://genfis40.esalq.usp.br/genfis/>) ESALQ-USP para exploração de vias metabólicas de interesse e posterior disponibilização do *software* para o grupo de pesquisa.

Foram coletadas amostras de folha +1 da cultivar SP80-3280 de plantas de cana-de-açúcar de 9 meses cultivadas a campo, no Apta de Jaú, em experimento de área irrigada (controle) e área não irrigada (sequeiro, tratamento de estresse hídrico). No dia da coleta o solo do tratamento de estresse hídrico apresentava umidade relativa de 50% da capacidade de campo.

Os metabólitos de amostras controle e em estresse hídrico foram extraídos a partir de 50 miligramas de material macerados em nitrogênio líquido. As amostras foram pesadas e armazenadas em tubos (2 ml). Foram utilizadas 5 réplicas biológicas para cada tratamento

(controle e estresse hídrico).

A extração foi realizada de acordo com a metodologia descrita por (VOS, DE et al., 2007), com algumas modificações. Foi adicionado 1 ml de tampão de extração (99,875% metanol e 0,125% ácido fórmico) e as amostras foram vortexadas por 10 segundos. As amostras foram sonicadas por 15 minutos em frequência máxima (40 kHz) à 20°C e depois foram centrifugadas por 20 minutos, 14.000 rpm a 20°C. O sobrenadante foi filtrado em filtro de 0,22 µm e transferido para novo tubo (vial).

Após a extração dos metabólitos, foi realizada análise de LC-MS (UPLC ACQUITY e QToF-Ultima, Waters) em triplicata. Para esta análise foi utilizada uma coluna cromatográfica de fase reversa (Acquity UPLC BEH C18 1,7 µm 2,1 x 100mm). Foram utilizados dois tampões: tampão A (H₂O + 0,1% ácido fórmico) e tampão B (acetonitrila + 0,1% ácido fórmico). Os gradientes de eluição utilizados foram: 95% A e 5% B (por 3 minutos), 75% A e 25% B (por 3 minutos), 5% A e 95% B (por 3 minutos), e 95% A e 5% B (por 4 minutos) para lavagem da coluna e condicionamento. O fluxo do gradiente utilizado foi de 0,5 ml/min.

A fonte de ionização foi *elektrospray* (ESI). Os espectros de massas foram adquiridos no modo positivo e modo V. O instrumento foi operado com voltagem de capilar de 3,0 kV e com voltagem de cone de 35 kV. As temperaturas da fonte e de dessolvatação foram de 150°C e 450°C. O fluxo do gás de dessolvatação foi de 550 l/h e do gás nebulizador (nitrogênio) de 50 l/h. O espectro de massa foi adquirido com razão massa/carga (m/z) de 100 a 2000 Da. A determinação dos picos de massa, alinhamento entre amostras e análises estatísticas foram realizadas no ambiente estatístico R, de acordo com o descrito no estudo de caso para o pacote ProbMetab (<http://labpib.fmrp.usp.br/methods/probmetab/#studycase>).

As análises potencializadas pelo pacote ProbMetab utilizam informação de diferentes fontes para auxiliar a compreensão das mudanças dinâmicas que ocorrem no metabolismo. A **Figura 12** ilustra a visão parcial de uma rede de correlações entre os picos de massa obtidos pelo fluxo de análises descrito anteriormente, condensando informações das reações e vias metabólicas em uma ferramenta que possibilita a manipulação dinâmica desta rede (SHANNON et al., 2003).

Com auxílio desta visualização foi possível observar que 31 dos picos de massa detectados têm como possíveis identidades compostos que participam da biossíntese de

Flavonóides, conhecida via do metabolismo secundário de plantas. A alteração dos níveis de Flavonóides é um marcador conhecido de estresses em plantas, e foi previamente descrito por conter a geração de espécies reativas de oxigênio (ROS), levando a inativação de enzimas antioxidantes, constituindo um sistema quelante secundário em plantas (PANDEY et al., 2013).

Figura 12 - Visão parcial da rede de possíveis reações bioquímicas e correlações (valor absoluto maior que 0,7) sobrepostas, geradas pelo pacote ProbMetab. Nós (em roxo) representam picos de massa com intensidade média significativamente diferente entre as condições de seca e irrigação. Arestas em vermelho representam correlação presente apenas na condição de irrigação (maior que o valor absoluto 0,7), verde apenas na condição de seca e azul em ambas as condições. Espessuras diferentes indicam se a diferença entre as correlações é significativa (mais espesso) ou não (menos espesso). O composto em destaque *Vitexin 2''-O-beta-L-rhamnoside*, possui a 4ª maior probabilidade, com uma pequena distância das 3 primeiras, mas é o composto mais frequente entre as reações conhecidas.

Uma vez detectada uma via metabólica de interesse, o pacote ProbMetab possibilita exportar uma visualização com o mesmo *layout* disponível no KEGG **Figura 13**, possibilitando uma referência à representação clássica do metabolismo, que também pode ser

acessada pela implementação do KEGG Mapper disponível no pacote.

Figura 13 - Representação da via metabólica da Biossíntese de Flavonóides (map00941) automaticamente exportada do KEGG para o Cytoscape pelo pacote ProbMetab, mostrando 13 compostos que figuram como identidade putativa para os picos encontrados na rede de reações/correlações. A – Algoritmo KEGG Mapper. B – Grafo com as posições reportadas no formato *kgml* (<http://www.kegg.jp/kegg/xml/>).

Entre os 31 picos, 27 tiveram a sua intensidade diminuída em condições de seca, *Intensity ratio* (razão da intensidade em seca/intensidade em irrigação) < 1 . Estudos prévios demonstraram que pequenas variações de intensidade dos metabólitos, de no mínimo uma a duas vezes (*Fold Change*) em condições contrastantes, são robustas a variação experimental (VINAIXA et al., 2012). No presente subconjunto de metabólitos foi observada variação de no máximo 5 vezes mais intenso na condição de irrigação (massa putativa 610,14) e 1.3 vezes mais intenso na condição de seca (massa putativa 550,09).

A **Tabela 11** realça as características do formato de representação resumido, explicado acima. Podemos notar que entre as possíveis identidades associadas a síntese de Flavonóides marcadas em amarelo, figuram isômeros para uma dada massa, e as massas propostas apresentam diferenças muito pequenas. Com o campo “Anotação do íon” podemos observar que os picos com massa proposta próxima possuem tempos de retenção diferentes (valor após o primeiro símbolo “#”) apontando para possibilidade de se tratarem de compostos diferentes, com características físico-químicas distintas. Podemos observar também que compostos como *Penicillin O* (C17405), um composto que não está associado a plantas no banco de dados utilizado, aparece como candidato pelo uso da base de dados genérica do KEGG, mas que é seguido por fitoquímicos como *3,7-Di-O-methylquercetin* (C01265) com probabilidade próxima. Se bancos de dados específicos para reações bioquímicas de plantas pudessem ser utilizados como fonte de informação para a distribuição *a priori*, evidentemente o modelo jamais consideraria C17405.

Tabela 11 - Visão parcial da tabela de probabilidades exportada pelo pacote ProbMetab, mostrando em amarelo as possíveis identidades associadas à síntese de Flavonóides.

Massa Proposta	Identidade mais provável	Nome da identidade mais provável	Probabilidade	Entropia	Razão das intensidades	Anotação do íon	p-valor t-test	Composto/vias
286,0439463083	C01514	Luteolin	0,162	0,955	0,349	287.0512#315.148#[48][M]+#[M+H-C6H10O5]+ 448.0,000017476	0,3015851606	C01514-00941;00944;01100;01110 C05903-00941;00944;01100;01110;01061
	C05903	Kaempferol	0,147					
	C17786	6-Demethoxycapillarisin	0,076					
	C10510	Orobol	0,074					
	C10036	Datiscetin	0,071					
	C10097	Isoscutellarein	0,07					
	C08576	Aureusidin	0,068					
	C12134	2'-Hydroxygenistein	0,067					
	C10041	Fisetin	0,066					
	C10184	Scutellarein	0,066					
	C17810	Citreorsein	0,065					
	C08720	Maritimetin	0,058					
	C07359	Oxazepam	0,009					
	C17405	Penicillin O	0,202					
330,0677718265	C01265	3',4',5-Trihydroxy-3,7-dimethoxyflavone	0,16	0,969	0,736	331.075#501.439#[70][M]+#	0,3015851606	C01265-00944
	C10193	Tricin	0,142					
	C03036	(+)-Bisdechlorogeodin	0,09					
	C17670	Aurantio-obtusin	0,086					
	C03040	(-)-Bisdechlorogeodin	0,084					
	C10424	Hildecarpin	0,082					
	C10033	Cirsiliol	0,082					
	C16754	Aflatoxin G2	0,074					
	C17405	Penicillin O	0,395					
	C01265	3',4',5-Trihydroxy-3,7-dimethoxyflavone	0,126					
	C10193	Tricin	0,112					
	C10033	Cirsiliol	0,069					
	C03036	(+)-Bisdechlorogeodin	0,065					
	C03040	(-)-Bisdechlorogeodin	0,064					
330,0660866171	C17670	Aurantio-obtusin	0,058	0,864	0,905	331.0734#381.2455#[71][M]+#[M+Na]+ 308.088	0,7463242681	C01265-00944
	C10424	Hildecarpin	0,056					
	C16754	Aflatoxin G2	0,054					
	C17405	Penicillin O	0,266					
	C01265	3',4',5-Trihydroxy-3,7-dimethoxyflavone	0,153					
	C10193	Tricin	0,14					
	C17670	Aurantio-obtusin	0,09					
	C10424	Hildecarpin	0,079					
	C03040	(-)-Bisdechlorogeodin	0,077					
	C03036	(+)-Bisdechlorogeodin	0,067					
	C16754	Aflatoxin G2	0,065					
	C10033	Cirsiliol	0,063					
	C06569	7a-Hydroxy-O-carbamoyl-deacetylcephalosporin C	0,728					
	C04608	Apigenin 7-O-beta-D-glucoside	0,052					
432,0957515893	C01714	Apigenin-6-C-glucoside	0,052	0,5	0,523	433.103#419.102#[169][M]+#[M+H-C2H4]+ 460.13	0,0049787785	C04608-00944 C01714-00944 C01460-00941;00944
	C01460	Vitexin	0,046					
	C09126	Genistein 7-O-beta-D-glucoside	0,03					
	C16911	Afzelin	0,022					
	C10420	Genistein 8-C-glucoside	0,022					
	C04609	Aerobacter aerogenes capsular polysaccharide	0,017					
	C10345	Emodin 8-glucoside	0,016					
	C01715	Kallikrein	0,016					

Considerando todo o experimento, não apenas possíveis identidades associadas à via de Biossíntese de Flavonóides, os compostos que apresentaram os maiores e menores valores para a razão intensidade seca/intensidade em irrigação são apresentados na **Tabela 12** e na **Figura 14**. É interessante notar a definição dos picos, e sua reprodutibilidade ao longo das amostras, indicativos da boa qualidade dos dados, bem como da eficiência do fluxo de análises automatizado, com parâmetros específicos para plataforma experimental e limiares de qualidade para seleção dos picos (<http://labpib.fmrp.usp.br/methods/probmetab>).

A identidade *Nanchangmycin* (código kegg C12047) é um composto anotado para o fungo *Streptomyces nanchangensis* e foi o único composto candidato para a massa proposta. O composto *N-(3-(S)-Hydroxybutyryl)homoserine lactone* (código kegg C11842) é associado a via map02020 - *Two-component system* de sinalização em microrganismos. Estes compostos muito provavelmente não correspondem as identidades corretas para os metabólitos, entretanto, a fórmula molecular pode ajudar a sugerir a composição elementar do metabólito, se este é eleito para futuras investigações em uma análise global do experimento. O composto *Ginsenoside Rg1* (código kegg C08946), consta apenas genericamente classificado como terpenóide, que é uma classe de compostos amplamente difundida em plantas que figuram como pigmentos fotossintéticos, hormônios, na síntese de polissacarídeos e como estrutura de membranas (MCGARVEY; CROTEAU, 1995). O composto *epsilon-Rhodomyconone* (código kegg C12425) é catalogado no metabolismo secundário de plantas e possui três reações bioquímicas conhecidas.

Tabela 12 - Espectros selecionados dos maiores (acima) e menores (abaixo) valores absolutos da razão intensidade seca/intensidade em irrigação mostrando a fórmula sugerida pela anotação do íon.

Massa Proposta	Identidade mais provável	Probabilidade	Razão da intensidade	p-valor t-test	Anotação do íon
866.544221320835	<i>Nanchangmycin</i>	1	87.596418	5.85979936129633e-06	884.5738#723.723#[489][M]+#[M+H+NH3]+ 866.544
800.480945360168	<i>Ginsenoside Rg1</i>	506	4.838134	5.68847300153053e-06	801.4882#725.899#[478][M]+#[M+Na]+ 778.502
187.082761136295	<i>N-(3-(S)-Hydroxybutyryl)homo:</i>	0.27	0.01404565	0.00532105216323475	170.0789#45.387##[M+H-H2O]+ 187.083
428.103479664795	<i>epsilon-Rhodomyconone</i>	1	0.06310671	0.000909057353815967	429.1108#353.655#[163][M]+#[M+H-H2O]+ 446.112

Figura 14 - Espectros selecionados dos maiores (acima) e menores (abaixo) valores absolutos da razão intensidade seca/intensidade em irrigação mostrando uma das fórmulas sugeridas pela anotação. Legendas: *F_I_Nature* em preto representam amostras de cana irrigadas e *F_S_Nature* em vermelho representam amostras de cana sob estresse de seca.

Com a implementação do algoritmo KEGG *Mapper*, podemos utilizar um filtro para o metabolismo de um organismo filogeneticamente próximo. Na presente análise utilizamos o milho (*Zea mays*) membro da família *Poaceae*, uma vez que não existem catálogos de vias disponíveis para cana. Com este filtro, as vias metabólicas com maior número de compostos candidatos foram a Biossíntese de Fenilpropanóides (map00940) com 8 identidades candidatas, Degradação de Compostos Aromáticos (map01220) com 7 identidades candidatas, Biossíntese de Flavonóides (map00941) com 11 identidades candidatas e Biossíntese da Flavona e Flavonóides (map00944) com 31 identidades candidatas.

6- CONCLUSÕES E PERSPECTIVAS

6.1- Implementação e validação do modelo

Os métodos disponibilizados pelo pacote ProbMetab nos permitiram concluir que a utilização da distribuição de probabilidades *a posteriori* associada a informações espectrais e de catálogos de vias metabólicas são eficientes em reduzir a incerteza na anotação de metabólitos. Esta redução foi alcançada pelo posicionamento das identidades corretas entre os candidatos mais prováveis para 90% dos compostos estudados. Esta classificação é uma alternativa à classificação simplista usualmente adotada na área de metabolômica, em que se usa apenas o candidato de maior probabilidade. Constatou-se que a constituição modular do pacote ProbMetab, com conexões flexíveis para anotação do íon, formatação de bancos de dados com contexto biológico, introdução de novas fontes de informação e representações eficientes da distribuição de probabilidades *a posteriori* formam um arcabouço de *software* robusto para a anotação de metabólitos e possibilita a exploração de conjuntos de dados multidimensionais com eficiência.

A natureza restrita da amostragem de um protocolo experimental específico, que compreende etapas desde a preparação das amostras ao processamento final dos dados, a complexidade da ionização e a falta do conhecimento da extensão do metaboloma têm limitado a metabolômica a um escopo mais restrito que o antecipado pelos seus propositores (OLIVER, 1998). Com o acúmulo de conhecimento em novas técnicas de processamento de dados e crescente disponibilização de novos compostos em bancos de dados públicos, torna-se possível delinear estratégias de análise que capitalizem esta informação para uma anotação de alta qualidade de metabólitos amostrados em experimentos de espectrometria de massas. Dividindo-se o processamento em três etapas: 1) Anotação do íon e busca pela massa exata; 2) Definição de um modelo probabilístico e estimação, e 3) Representação abrangente dos resultados, espera-se ressaltar a importância da incorporação de diferentes fontes de informação contidas no experimento, bem como do contexto biológico, potencializando a utilização do conhecimento disponível para anotação. Fornecer uma implementação de um *software* livre, gratuito e aberto, é essencial para a evolução do conhecimento, uma vez que a

comunidade interessada pode alterar o código fonte de tais aplicações, adicionando melhorias aos algoritmos e criando novas funcionalidades para o fluxo de análises.

Da perspectiva de incorporação de informação já existem fontes de conhecimento que são utilizadas de formas dissociadas, que se utilizadas em conjunto, poderiam potencializar a anotação, tais como especificidade de preparação da amostra (BELTRAN et al., 2012), o “defeito da massa” (*mass defect*) (SLENO, 2012) e MSⁿ (GAQUEREL et al., 2013). A medida que estas fontes de informação sejam sistematicamente armazenadas e disponibilizadas, a eficiência da anotação pode ser enormemente aumentada. Essas informações deveriam ser disponibilizadas em repositórios centralizados, que preferencialmente seriam acessados por interfaces programáticas. Bancos de dados como MetaboLights (STEINBECK et al., 2012), podem permitir esse compartilhamento de informações, atitude que foi historicamente decisiva para, por exemplo, a evolução da genômica (WATROUS et al., 2012; STEINBECK et al., 2012).

Como a natureza amostral de uma data técnica não direcionada de LC-MS nos impede de observar todos os compostos de uma via metabólica específica, a melhor maneira de representar tal amostra seria a superposição em uma rede metabólica. As redes de correlação têm sido utilizadas para mapear reações bioquímicas entre picos de massa e os defensores de sua utilização argumentam que as correlações parciais podem ser enviesadas por variáveis “ocultas” (e.g. metabólitos não amostrados, enzimas) (STEUER, 2006) ou relações não lineares (OPGEN-RHEIN; STRIMMER, 2007). Por outro lado, existem evidências demonstrando que as mesmas podem representar melhor a relação entre compostos, excluindo o efeito de outras variáveis (KRUMSIEK et al., 2011).

Com o desenvolvimento de um *software* que disponibiliza um método de anotação de metabólitos, por meio de um fluxo de análises que possibilita a incorporação de conhecimento prévio e informação espectral adicional, esperamos fornecer meios para sumarizar os resultados de uma série de análises necessárias para extração de conhecimento de conjuntos de dados multidimensionais e auxiliar o biólogo a identificar mudanças do metabolismo relevantes para o seu objeto de estudo.

6.2- Busca automatizada de alterações no metabolismo

A aplicação do pacote ProbMetab a dados reais de cana-de-açúcar, nos permitiu recuperar automaticamente alterações em uma via metabólica anteriormente reportada em resposta a estresses abióticos. Esta constatação nos permite concluir que a representação da distribuição *a posteriori* na forma de um grafo de conexões bioquímicas, possibilita a investigação de metabólitos importantes para estas vias e sua consequente seleção para experimentos adicionais.

Embora muitos estudos demonstrem ao nível transcricional que a produção de Flavonóides seja induzida em resposta a seca (PANDEY et al., 2013), Yang et al., (2007) demonstraram para *Glycyrrhiza inflata*, que, embora a produção aumente, o conteúdo total de Flavonóides diminui. Uma possível explicação para essa diminuição, também observada no experimento utilizado como caso de estudo nessa tese, poderia ser a modificação da forma identificável do Flavonóides pelo seu papel de quelante (YANG et al., 2007).

Uma análise mais detalhada das correlações diferenciais (FUKUSHIMA, 2013), introduzidas recentemente no fluxo de análises do pacote ProbMetab, pode fornecer evidências de como ocorre o fluxo de carbonos sob o estresse de seca, uma vez que uma correlação entre dois picos de massas que existe em uma condição e deixa de existir em outra, pode apontar para um deslocamento do fluxo de carbonos. Caldana et al., (2011) demonstram que, sob a mudança simultânea de condições de temperatura e luminosidade, plantas de *Arabidopsis thaliana* apresentaram grandes rearranjos das correlações entre os metabólitos mensurados. Os autores postularam um rearranjo no fluxo de carbonos, direcionando aminoácidos fornecidos pela degradação de proteínas para produção de energia. Este rearranjo foi corroborado pela observação paralela da expressão gênica relacionada (CALDANA et al., 2011).

O entendimento da partição dos carbonos sob condições de estresse é essencial para fornecer incrementos a produção em cana-de-açúcar. (HOFMANN; JAHUFER, 2011) demonstraram uma correlação negativa entre a produção de Flavonóides e o acúmulo de matéria seca em genótipos de trevo branco (*white clover*), e sugerem que os Flavonóides podem ser utilizados como marcadores bioquímicos para seleção de genótipos em programas

de melhoramento. Uma análise mais detalhada do metabolismo secundário de cana-de-açúcar pode sugerir biomarcadores importantes para um programa de melhoramento de múltiplos objetivos.

As análises automatizadas no pacote ProbMetab se mostraram eficientes em recuperar uma via metabólica reconhecidamente envolvida em estresses abióticos de plantas, apontando para estudos adicionais que estão sendo conduzidos para obtenção de uma identificação por MS/MS dos metabólitos anotados. Apenas uma pequena fração da rede de reações foi analisada, havendo ainda potencial para descoberta de outros fatores potencialmente impactantes nas alterações do metabolismo sob o estresse da seca.

7- REFERÊNCIAS BIBLIOGRÁFICAS

- AGAPAKIS, C. M.; BOYLE, P. M.; SILVER, P. A. Natural strategies for the spatial optimization of metabolism in synthetic biology. **Nature chemical biology**, v. 8, n. 6, p. 527–35, 2012. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. Disponível em: <<http://dx.doi.org/10.1038/nchembio.975>>. Acesso em: 9/1/2014.
- ALLWOOD, J.; WEBER, R.; ZHOU, J.; et al. CASMI—The Small Molecule Identification Process from a Birmingham Perspective. **Metabolites**, v. 3, n. 2, p. 397–411, 2013. Multidisciplinary Digital Publishing Institute. Disponível em: <<http://www.mdpi.com/2218-1989/3/2/397>>. Acesso em: 3/8/2013.
- ALONSO, A.; JULIÀ, A.; BELTRAN, A.; et al. AStream: An R package for annotating LC/MS metabolomic data. **Bioinformatics (Oxford, England)**, v. 27, n. 9, p. 1339–1340, 2011. Disponível em: <<http://bioinformatics.oxfordjournals.org/cgi/content/abstract/27/9/1339>>. Acesso em: 20/3/2011.
- ALTMAN, T.; TRAVERS, M.; KOTHARI, A.; CASPI, R.; KARP, P. D. A systematic comparison of the MetaCyc and KEGG pathway databases. **BMC Bioinformatics**, v. 14, n. 1, p. 112, 2013. Disponível em: <<http://www.biomedcentral.com/1471-2105/14/112>>. Acesso em: 27/3/2013.
- BARUPAL, D. K.; HALDIYA, P. K.; WOHLGEMUTH, G.; et al. MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. **BMC Bioinformatics**, v. 13, n. 1, p. 99, 2012. Disponível em: <<http://www.biomedcentral.com/1471-2105/13/99>>. Acesso em: 17/5/2012.
- BELTRAN, A.; SUAREZ, M.; RODRÍGUEZ, M. A.; et al. Assessment of compatibility between extraction methods for NMR- and LC/MS-based metabolomics. **Analytical chemistry**, v. 84, n. 14, p. 5838–44, 2012. American Chemical Society. Disponível em: <<http://dx.doi.org/10.1021/ac3005567>>. Acesso em: 12/3/2013.
- BÖCKER, S.; LETZEL, M. C.; LIPTÁK, Z.; PERVUKHIN, A. SIRIUS: decomposing isotope patterns for metabolite identification. **Bioinformatics (Oxford, England)**, v. 25, n. 2, p. 218–24, 2009. Disponível em: <<http://bioinformatics.oxfordjournals.org/content/25/2/218.full>>. Acesso em: 8/8/2013.
- BREITLING, R.; RITCHIE, S.; GOODENOWE, D.; STEWART, M. L.; BARRETT, M. P. Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. **Metabolomics**, v. 2, n. 3, p. 155–164, 2006. Disponível em: <<http://www.springerlink.com/content/x01808376167kk12>>. Acesso em: 30/6/2010.
- CALDANA, C.; DEGENKOLBE, T.; CUADROS-INOSTROZA, A.; et al. High-density kinetic analysis of the metabolomic and transcriptomic response of Arabidopsis to eight environmental conditions. **The Plant journal : for cell and molecular biology**, v. 67, n. 5, p. 869–84, 2011. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/21575090>>. Acesso em: 2/9/2011.

- CASPI, R.; ALTMAN, T.; DREHER, K.; et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. **Nucleic Acids Research**, v. 40, n. D1, p. D742–753, 2011. Disponível em: <<http://nar.oxfordjournals.org/cgi/content/abstract/40/D1/D742>>. Acesso em: 21/11/2011.
- CASTILLO, S.; GOPALACHARYULU, P.; YETUKURI, L.; OREŠIČ, M. Algorithms and tools for the preprocessing of LC–MS metabolomics data. **Chemometrics and Intelligent Laboratory Systems**, v. 108, n. 1, p. 23–32, 2011. Elsevier B.V. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0169743911000608>>. Acesso em: 18/7/2011.
- CREEK, D. J.; JANKEVICS, A.; BREITLING, R.; et al. Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. **Analytical chemistry**, 2011. American Chemical Society. Disponível em: <<http://dx.doi.org/10.1021/ac2021823>>. Acesso em: 4/11/2011.
- CREEK, D. J.; JANKEVICS, A.; BURGESS, K. E. V; BREITLING, R.; BARRETT, M. P. IDEOM: An Excel interface for analysis of LC-MS based metabolomics data. **Bioinformatics (Oxford, England)**, v. 28, n. 7, p. 1048–1049, 2012. Disponível em: <<http://bioinformatics.oxfordjournals.org/cgi/content/abstract/28/7/1048>>. Acesso em: 30/3/2012.
- DUNN, W. B.; ERBAN, A.; WEBER, R. J. M.; et al. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. **Metabolomics**, 2012. Disponível em: <<http://www.springerlink.com/content/0718581530254pg6/>>. Acesso em: 29/5/2012.
- EDDELBUETTEL, D.; FRANÇOIS, R. Rcpp: Seamless R and C++ integration. **Journal of Statistical Software**, v. 40, n. 8, p. 1–18, 2011.
- FERNANDEZ-DE-COSSIO, J.; GONZALEZ, L. J.; SATOMI, Y.; et al. Isotopica: a tool for the calculation and viewing of complex isotopic envelopes. **Nucleic acids research**, v. 32, n. Web Server issue, p. W674–8, 2004. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=441561&tool=pmcentrez&rendertype=abstract>>. Acesso em: 2/12/2013.
- FIEHN, O.; BARUPAL, D. K.; KIND, T. Extending biochemical databases by metabolomic surveys. **The Journal of biological chemistry**, v. 286, n. 27, p. 23637–23643, 2011. Disponível em: <<http://www.jbc.org/cgi/content/abstract/286/27/23637>>. Acesso em: 4/7/2011.
- FISCHLER, M. A.; BOLLES, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. **Communications of the ACM**, v. 24, n. 6, p. 381–395, 1981. ACM. Disponível em: <<http://dl.acm.org/citation.cfm?id=358669.358692>>. Acesso em: 7/11/2013.
- FUKUSHIMA, A. DiffCorr: an R package to analyze and visualize differential correlations in biological networks. **Gene**, v. 518, n. 1, p. 209–14, 2013. Disponível em: <<http://dx.doi.org/10.1016/j.gene.2012.11.028>>. Acesso em: 1/9/2013.
- GAQUEREL, E.; KUHL, C.; NEUMANN, S. Computational annotation of plant metabolomics profiles via a novel network-assisted approach. **Metabolomics**, 2013. Disponível em: <<http://link.springer.com/10.1007/s11306-013-0504-2>>. Acesso em: 4/3/2013.

GORYŃSKI, K.; BOJKO, B.; NOWACZYK, A.; et al. Quantitative structure–retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds. **Analytica Chimica Acta**, v. 797, p. 13–19, 2013. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0003267013011173>>. Acesso em: 25/9/2013.

GUO, A. C.; JEWISON, T.; WILSON, M.; et al. ECMDB: the E. coli Metabolome Database. **Nucleic acids research**, v. 41, n. Database issue, p. D625–30, 2013. Disponível em: <<http://nar.oxfordjournals.org/content/41/D1/D625.short>>. Acesso em: 23/9/2013.

HOFFMANN, E. DE; STROOBANT, V. Detectors and Computers. **Mass Spectrometry Principles and Applications**. 3rd ed., p.175–188, 2007. John Wiley & Sons Ltd.

HOFMANN, R. W.; JAHUFER, M. Z. Z. Tradeoff between biomass and flavonoid accumulation in white clover reflects contrasting plant strategies. (W. A. Peer, Ed.) **PloS one**, v. 6, n. 4, p. e18949, 2011. Public Library of Science. Disponível em: <<http://dx.plos.org/10.1371/journal.pone.0018949>>. Acesso em: 5/9/2013.

HUBER, W.; CAREY, V. J.; LONG, L.; FALCON, S.; GENTLEMAN, R. Graphs in molecular biology. **BMC bioinformatics**, v. 8 Suppl 6, n. Suppl 6, p. S8, 2007. Disponível em: <<http://www.biomedcentral.com/1471-2105/8/S6/S8>>. Acesso em: 9/3/2012.

JOURDAN, F.; BREITLING, R.; BARRETT, M. P.; GILBERT, D. MetaNetter: inference and visualization of high-resolution metabolomic networks. **Bioinformatics (Oxford, England)**, v. 24, n. 1, p. 143–5, 2008. Disponível em: <<http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/1/143>>. Acesso em: 23/8/2010.

KALISZAN, R. QSRR: quantitative structure-(chromatographic) retention relationships. **Chemical reviews**, v. 107, n. 7, p. 3212–46, 2007. American Chemical Society. Disponível em: <<http://dx.doi.org/10.1021/cr068412z>>. Acesso em: 18/5/2011.

KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic acids research**, v. 28, n. 1, p. 27–30, 2000. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract>>. Acesso em: 3/8/2011.

KANEHISA, M.; GOTO, S.; SATO, Y.; FURUMICHI, M.; TANABE, M. KEGG for integration and interpretation of large-scale molecular data sets. **Nucleic acids research**, v. 40, n. Database issue, p. D109–14, 2012. Disponível em: <<http://nar.oxfordjournals.org/content/40/D1/D109.abstract>>. Acesso em: 6/8/2013.

KATAJAMAA, M.; MIETTINEN, J.; ORESIC, M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. **Bioinformatics (Oxford, England)**, v. 22, n. 5, p. 634–6, 2006. Disponível em: <<http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btk039v1>>. Acesso em: 28/7/2010.

KELLER, B. O.; SUI, J.; YOUNG, A. B.; WHITTAL, R. M. Interferences and contaminants encountered in modern mass spectrometry. **Analytica chimica acta**, v. 627, n. 1, p. 71–81, 2008. Disponível em: <<http://dx.doi.org/10.1016/j.aca.2008.04.043>>. Acesso em: 14/3/2012.

KIND, T.; FIEHN, O. Advances in structure elucidation of small molecules using mass spectrometry. **Bioanalytical reviews**, v. 2, n. 1-4, p. 23–60, 2010a. Disponível em: <<http://www.springerlink.com/content/c1p83306688q7656/>>. Acesso em: 16/7/2012.

KIND, T.; FIEHN, O. Advances in structure elucidation of small molecules using mass spectrometry. **Bioanalytical reviews**, v. 2, n. 1-4, p. 23–60, 2010b. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3015162&tool=pmcentrez&rendertype=abstract>>. Acesso em: 19/2/2013.

KIND, T.; SCHOLZ, M.; FIEHN, O. How large is the metabolome? A critical analysis of data exchange practices in chemistry. (M. Polymenis, Ed.) **PloS one**, v. 4, n. 5, p. e5440, 2009. Public Library of Science. Disponível em: <<http://dx.plos.org/10.1371/journal.pone.0005440>>. Acesso em: 7/8/2012.

KONO, N.; ARAKAWA, K.; OGAWA, R.; et al. Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. (R. K. Aziz, Ed.) **PloS one**, v. 4, n. 11, p. e7710, 2009. Public Library of Science. Disponível em: <<http://dx.plos.org/10.1371/journal.pone.0007710>>. Acesso em: 7/11/2013.

KRUMSIEK, J.; SUHRE, K.; ILLIG, T.; ADAMSKI, J.; THEIS, F. J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. **BMC systems biology**, v. 5, n. 1, p. 21, 2011. Disponível em: <<http://www.biomedcentral.com/1752-0509/5/21>>. Acesso em: 27/2/2013.

KUHL, C.; TAUTENHAHN, R.; BOTTCHE, C.; LARSON, T. R.; NEUMANN, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of LC/MS data sets. **Analytical chemistry**, 2011. American Chemical Society. Disponível em: <<http://dx.doi.org/10.1021/ac202450g>>. Acesso em: 29/11/2011.

LANGE, E. **Analysis of mass spectrometric data: peak picking and map alignment**, 2008. Thesis major: Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany, 2008. There is no corresponding record for this reference.

LANGE, E.; GRÖPL, C.; REINERT, K.; KOHLBACHER, O.; HILDEBRANDT, A. High-accuracy peak picking of proteomics data using wavelet techniques. **Pacific Symposium on Biocomputing**. **Pacific Symposium on Biocomputing**, p. 243–54, 2006. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/17094243>>. Acesso em: 4/12/2013.

LEADER, D. P.; BURGESS, K.; CREEK, D.; BARRETT, M. P. Pathos: a web facility that uses metabolic maps to display experimental changes in metabolites identified by mass spectrometry. **Rapid communications in mass spectrometry : RCM**, v. 25, n. 22, p. 3422–6, 2011. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3509215&tool=pmcentrez&rendertype=abstract>>. Acesso em: 22/10/2013.

LONGABAUGH, W. J. R. Combing the hairball with BioFabric: a new approach for visualization of large networks. **BMC bioinformatics**, v. 13, n. 1, p. 275, 2012. Disponível em: <<http://www.biomedcentral.com/1471-2105/13/275>>. Acesso em: 26/11/2013.

LOPES, C. T.; FRANZ, M.; KAZI, F.; et al. Cytoscape Web: an interactive web-based network browser. **Bioinformatics (Oxford, England)**, v. 26, n. 18, p. 2347–8, 2010. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2935447&tool=pmcentrez&rendertype=abstract>>. Acesso em: 11/11/2013.

- MCGARVEY, D. J.; CROTEAU, R. Terpenoid metabolism. **The Plant cell**, v. 7, n. 7, p. 1015–26, 1995. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=160903&tool=pmcentrez&rendertype=abstract>>. Acesso em: 9/12/2013.
- MERICO, D.; GFELLER, D.; BADER, G. D. How to visually interpret biological data using networks. **Nature biotechnology**, v. 27, n. 10, p. 921–4, 2009. Nature Publishing Group. Disponível em: <<http://dx.doi.org/10.1038/nbt.1567>>. Acesso em: 15/11/2013.
- NIELSEN, J. It is all about metabolic fluxes. **Journal of bacteriology**, v. 185, n. 24, p. 7031–5, 2003. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=296266&tool=pmcentrez&rendertype=abstract>>. Acesso em: 11/7/2012.
- NIELSEN, J. Metabolomics in Functional Genomics and Systems Biology. In: S. G. Villas-Boas; J. Nielsen; J. Smedsgaard; M. A. E. Hansen; U. Roessner-Tunali (Eds.); **Metabolome analysis: an introduction**. v. 24, p.3–17, 2007. John Wiley & Sons.
- OLIVER, S. Systematic functional analysis of the yeast genome. **Trends in Biotechnology**, v. 16, n. 9, p. 373–378, 1998. Disponível em: <[http://dx.doi.org/10.1016/S0167-7799\(98\)01214-1](http://dx.doi.org/10.1016/S0167-7799(98)01214-1)>. Acesso em: 8/5/2013.
- OPGEN-RHEIN, R.; STRIMMER, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. **BMC systems biology**, v. 1, n. 1, p. 37, 2007. Disponível em: <<http://www.biomedcentral.com/1752-0509/1/37>>. Acesso em: 11/3/2013.
- PANDEY, N.; RANJAN, A.; PANT, P.; et al. CAMTA 1 regulates drought responses in *Arabidopsis thaliana*. **BMC genomics**, v. 14, n. 1, p. 216, 2013. Disponível em: <<http://www.biomedcentral.com/1471-2164/14/216>>. Acesso em: 17/8/2013.
- PAVIA, D.; LAMPMAN, G.; KRIZ, G.; VYVYAN, J. **Introduction to Spectroscopy**. Cengage Learning, 2008.
- PLUSKAL, T.; CASTILLO, S.; VILLAR-BRIONES, A.; ORESIC, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. **BMC bioinformatics**, v. 11, n. 1, p. 395, 2010. Disponível em: <<http://www.biomedcentral.com/1471-2105/11/395>>. Acesso em: 15/6/2011.
- R DEVELOPMENT CORE TEAM. R: A Language and Environment for Statistical Computing. ,2010. Vienna, Austria. Disponível em: <<http://www.r-project.org/>>. .
- RAYLEIGH. Joseph John Thomson. 1856-1940. **Obituary Notices of Fellows of the Royal Society**, p. 587–609, 1941. JSTOR.
- ROESSNER, U. The Chemical Challenge of the Metabolome. In: S. G. Villas-Boas; J. Nielsen; J. Smedsgaard; M. A. E. Hansen; U. Roessner-Tunali (Eds.); **Metabolome analysis: an introduction**. v. 24, p.15–38, 2007. John Wiley & Sons.
- ROGERS, S.; SCHELTEMA, R. A.; BARRETT, M.; BREITLING, R. Bayesian Approaches for Mass Spectrometry-Based Metabolomics. **Handbook of Statistical Systems Biology**. p.467–476. Disponível em: <<http://onlinelibrary.wiley.com/doi/10.1002/9781119970606.ch24/summary>>. Acesso em: 21/3/2012.
- ROGERS, S.; SCHELTEMA, R. A.; GIROLAMI, M.; BREITLING, R. Probabilistic

assignment of formulas to mass peaks in metabolomics experiments. **Bioinformatics (Oxford, England)**, v. 25, n. 4, p. 512–8, 2009. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/19095699>>. .

SCHÄFER, J.; STRIMMER, K. An empirical Bayes approach to inferring large-scale gene association networks. **Bioinformatics (Oxford, England)**, v. 21, n. 6, p. 754–64, 2005. Disponível em: <<http://bioinformatics.oxfordjournals.org/content/21/6/754>>. Acesso em: 15/3/2013.

SCHELTEMA, R. A.; JANKEVICS, A.; JANSEN, R. C.; SWERTZ, M. A.; BREITLING, R. PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. **Analytical chemistry**, v. 83, n. 7, p. 2786–93, 2011. American Chemical Society. Disponível em: <<http://dx.doi.org/10.1021/ac2000994>>. Acesso em: 3/4/2012.

SCHELTEMA, R.; DECUYPERE, S.; DUJARDIN, J.; et al. Simple data-reduction method for high-resolution LC-MS data in metabolomics. **Bioanalysis**, v. 1, n. 9, p. 1551–7, 2009. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/21083103>>. .

SHANNON, P.; MARKIEL, A.; OZIER, O.; et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. **Genome research**, v. 13, n. 11, p. 2498–504, 2003. Disponível em: <<http://genome.cshlp.org/cgi/content/abstract/13/11/2498>>. Acesso em: 18/7/2011.

SHANNON, P. T.; GRIMES, M.; KUTLU, B.; BOT, J. J.; GALAS, D. J. RCytoscape: tools for exploratory network analysis. **BMC Bioinformatics**, v. 14, n. 1, p. 217, 2013. Disponível em: <<http://www.biomedcentral.com/1471-2105/14/217>>. Acesso em: 9/7/2013.

SILVERSTEIN, R. M.; WEBSTER, F. X.; KIEMLE, D. J. **Spectrometric identification of organic compounds**. 7th ed. John Wiley & Sons, 2005.

SIUZDAK, G. **The Expanding Role of Mass Spectrometry in Biotechnology**. McC Press, 2006.

SLENO, L. The use of mass defect in modern mass spectrometry. **Journal of mass spectrometry : JMS**, v. 47, n. 2, p. 226–36, 2012. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/22359333>>. Acesso em: 29/8/2013.

SMEDSGAARD, J. ANALYTICAL TOOLS. In: S. G. Villas-Boas; J. Nielsen; J. Smedsgaard; M. A. E. Hansen; U. Roessner-Tunali (Eds.); **Metabolome analysis: an introduction**. v. 24, p.83–145, 2007. John Wiley & Sons.

SMITH, C. A.; WANT, E. J.; O'MAILLE, G.; ABAGYAN, R.; SIUZDAK, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. **Analytical chemistry**, v. 78, n. 3, p. 779–87, 2006. American Chemical Society. Disponível em: <<http://dx.doi.org/10.1021/ac051437y>>. Acesso em: 1/7/2010.

STEINBECK, C.; CONESA, P.; HAUG, K.; et al. MetaboLights: towards a new COSMOS of metabolomics data management. **Metabolomics**, 2012. Disponível em: <<http://www.springerlink.com/content/7h5422241557mn14/>>. Acesso em: 27/9/2012.

STEUER, R. Review: on the analysis and interpretation of correlations in metabolomic data. **Briefings in bioinformatics**, v. 7, n. 2, p. 151–8, 2006. Disponível em:

<<http://bib.oxfordjournals.org/content/7/2/151.long>>. Acesso em: 18/3/2013.

SUDERMAN, M.; HALLETT, M. Tools for visually exploring biological networks. **Bioinformatics (Oxford, England)**, v. 23, n. 20, p. 2651–9, 2007. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/17720984>>. Acesso em: 12/11/2013.

SUESS, E. A.; TRUMBO, B. E. **Introduction to Probability Simulation and Gibbs Sampling with R**. Springer Verlag, 2010.

SUGIMOTO, M.; KAWAKAMI, M.; ROBERT, M.; SOGA, T.; TOMITA, M. Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. **Current bioinformatics**, v. 7, n. 1, p. 96–108, 2012. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3299976&tool=pmcentrez&rendertype=abstract>>. Acesso em: 13/11/2013.

SUMNER, L. W.; URBANCZYK-WOCHNIAK, E.; BROECKLING, C. D. Metabolomics data analysis, visualization, and integration. **Methods in molecular biology (Clifton, N.J.)**, v. 406, p. 409–36, 2007. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/18287705>>. Acesso em: 8/8/2013.

TAUTENHAHN, R.; BÖTTCHER, C.; NEUMANN, S. Highly sensitive feature detection for high resolution LC/MS. **BMC bioinformatics**, v. 9, n. 1, p. 504, 2008. Disponível em: <<http://www.biomedcentral.com/1471-2105/9/504>>. Acesso em: 16/7/2012.

TAUTENHAHN, R.; CHO, K.; URITBOONTHAI, W.; et al. An accelerated workflow for untargeted metabolomics using the METLIN database. **Nature Biotechnology**, v. 30, n. 9, p. 826–828, 2012. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. Disponível em: <<http://dx.doi.org/10.1038/nbt.2348>>. Acesso em: 10/9/2012.

TAUTENHAHN, R.; PATTI, G. J.; RINEHART, D.; SIUZDAK, G. E. XCMS Online: a web-based platform to process untargeted metabolomic data. **Analytical chemistry**, 2012. American Chemical Society. Disponível em: <<http://dx.doi.org/10.1021/ac300698c>>. Acesso em: 30/4/2012.

THEODORIDIS, G. A.; MICHPOULOS, F.; GIKA, H. G.; PLUMB, R. S.; WILSON, I. D. Liquid Chromatographic Techniques in Metabolomics. In: T. Hyotylainen; S. Wiedmer (Eds.); **Chromatographic Methods in Metabolomics**. p.64–83, 2013. Royal Society of Chemistry.

VILLAS-BOAS, S. G. Sampling and Sample Preparation. In: S. G. Villas-Boas; J. Nielsen; J. Smedsgaard; M. A. E. Hansen; U. Roessner-Tunali (Eds.); **Metabolome analysis: an introduction**. v. 24, p.39–82, 2007. John Wiley & Sons.

VINAIXA, M.; SAMINO, S.; SAEZ, I.; et al. A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. **Metabolites**, v. 2, n. 4, p. 775–795, 2012. Disponível em: <<http://www.mdpi.com/2218-1989/2/4/775/>>. Acesso em: 27/10/2012.

VINCENT, I. M.; CREEK, D. J.; BURGESS, K.; et al. Untargeted metabolomics reveals a lack of synergy between nifurtimox and eflornithine against *Trypanosoma brucei*. (P. A. Bates, Ed.) **PLoS neglected tropical diseases**, v. 6, n. 5, p. e1618, 2012. Public Library of Science. Disponível em: <<http://dx.plos.org/10.1371/journal.pntd.0001618>>. Acesso em:

27/7/2012.

VOET, D.; VOET, J. G. **Biochemistry**. John Wiley & Sons, 2011.

VOS, R. C. H. DE; MOCO, S.; LOMMEN, A.; et al. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. **Nature protocols**, v. 2, n. 4, p. 778–91, 2007. Nature Publishing Group. Disponível em: <<http://dx.doi.org/10.1038/nprot.2007.95>>. Acesso em: 11/6/2011.

WAGNER, A.; FELL, D. A. The small world inside large metabolic networks. **Proceedings. Biological sciences / The Royal Society**, v. 268, n. 1478, p. 1803–10, 2001. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1088812&tool=pmcentrez&rendertype=abstract>>. Acesso em: 19/11/2013.

WATROUS, J.; ROACH, P.; ALEXANDROV, T.; et al. Mass spectral molecular networking of living microbial colonies. **Proceedings of the National Academy of Sciences of the United States of America**, v. 109, n. 26, p. E1743–52, 2012. Disponível em: <<http://www.pnas.org/content/early/2012/05/08/1203689109>>. Acesso em: 6/8/2013.

WEBER, R. J. M.; SOUTHAM, A. D.; SOMMER, U.; VIANT, M. R. Characterization of isotopic abundance measurements in high resolution FT-ICR and Orbitrap mass spectra for improved confidence of metabolite identification. **Analytical chemistry**, v. 83, n. 10, p. 3737–43, 2011. American Chemical Society. Disponível em: <<http://dx.doi.org/10.1021/ac2001803>>. Acesso em: 16/7/2011.

WISHART, D. S. Computational approaches to metabolomics. **Methods in molecular biology (Clifton, N.J.)**, v. 593, p. 283–313, 2010. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/19957155>>. Acesso em: 31/12/2013.

WISHART, D. S.; JEWISON, T.; GUO, A. C.; et al. HMDB 3.0--The Human Metabolome Database in 2013. **Nucleic acids research**, p. gks1065–, 2012. Disponível em: <<http://nar.oxfordjournals.org/content/early/2012/11/16/nar.gks1065.abstract>>. Acesso em: 20/11/2012.

WISHART, D. S.; TZUR, D.; KNOX, C.; et al. HMDB: the Human Metabolome Database. **Nucleic acids research**, v. 35, n. Database issue, p. D521–6, 2007. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1899095&tool=pmcentrez&rendertype=abstract>>. Acesso em: 19/7/2011.

XIA, J.; MANDAL, R.; SINELNIKOV, I. V.; BROADHURST, D.; WISHART, D. S. MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis. **Nucleic Acids Research**, v. 40, n. W1, p. W127–133, 2012. Disponível em: <<http://nar.oxfordjournals.org/cgi/content/abstract/40/W1/W127>>. Acesso em: 3/5/2012.

XIA, J.; PSYCHOGIOS, N.; YOUNG, N.; WISHART, D. S. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. **Nucleic acids research**, v. 37, n. Web Server issue, p. W652–60, 2009. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2703878&tool=pmcentrez&rendertype=abstract>>. Acesso em: 11/7/2010.

YANG, Y.; HE, F.; YU, L.; et al. Influence of drought on oxidative stress and flavonoid production in cell suspension culture of *Glycyrrhiza inflata* Batal. **Zeitschrift für**

Naturforschung. C, Journal of biosciences, v. 62, n. 5-6, p. 410–6. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/17708448>>. Acesso em: 5/9/2013.

ZHANG, A.; SUN, H.; WANG, P.; HAN, Y.; WANG, X. Modern analytical techniques in metabolomics analysis. **The Analyst**, v. 137, n. 2, p. 293–300, 2012. The Royal Society of Chemistry. Disponível em: <<http://pubs.rsc.org/en/content/articlehtml/2012/an/c1an15605e>>. Acesso em: 17/12/2013.

ZHOU, B.; XIAO, J. F.; TULI, L.; RESSOM, H. W. LC-MS-based metabolomics. **Molecular bioSystems**, v. 8, n. 2, p. 470–81, 2012. The Royal Society of Chemistry. Disponível em: <<http://pubs.rsc.org/en/content/articlehtml/2012/mb/c1mb05350g>>. Acesso em: 25/11/2013.

8- ANEXOS

8.1-Artigo publicado

8.1.1-Material suplementar do artigo publicado

8.2-Artigo em processo de submissão

8.3-Revisão de literatura submetida para publicação

8.4-Histórico Janus

Artigo publicado

ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS based metabolomics

Ricardo R. Silva¹, Fabien Jourdan^{2,3}, Diego M. Salvanha^{1,4}, Fabien Letisse^{2,3,5}, Emilien L. Jamin^{2,3}, Simone Guidetti-Gonzalez⁶, Carlos A. Labate^{6,7} and Ricardo Z.N. Vêncio^{1*}

¹LabPIB, Department of Computing and Mathematics FFCLRP-USP, University of Sao Paulo, Ribeirao Preto, Brazil

²INRA UMR1331, Toxalim, Research Centre in Food Toxicology, Toulouse, France

³Universit de Toulouse, INSA, UPS, INP; LISBP, Toulouse, France

⁴Institute for Systems Biology, Seattle, Washington, USA

⁵CNRS, UMR5504, Toulouse, France

⁶Department of Genetics ESALQ-USP, University of Sao Paulo, Piracicaba, Brazil

⁷Laboratorio Nacional de Ciencia e Tecnologia do Bioetanol CTBE, Campinas, Brazil

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Summary: We present ProbMetab, an R package which promotes substantial improvement in automatic probabilistic LC-MS based metabolome annotation. The inference engine core is based on a Bayesian model implemented to: (i) allow diverse source of experimental data and metadata to be systematically incorporated into the model with alternative ways to calculate the likelihood function and; (ii) allow sensitive selection of biologically meaningful biochemical reactions databases as Dirichlet-categorical prior distribution. Additionally, to ensure result interpretation by system biologists, we display the annotation in a network where observed mass peaks are connected if their candidate metabolites are substrate/product of known biochemical reactions. This graph can be overlaid with other graph-based analysis, such as partial correlation networks, in a visualization scheme exported to Cytoscape, with web and stand alone versions.

Availability and Implementation: ProbMetab was implemented in a modular fashion to fit together with established upstream (xcms, CAMERA, AStream, mzMatch.R, etc) and downstream R package tools (GeneNet, Rcytoscape, DiffCorr, etc). ProbMetab, along with extensive documentation and case studies, is freely available under GNU license at: <http://labpib.fmrp.usp.br/methods/probmetab/>.

Supplementary information: Supplementary data (supporting tables and two detailed case studies) are available at Bioinformatics online and at ProbMetab website.

Contact: rvencio@usp.br

1 INTRODUCTION

Metabolomics is an emerging field of study in post-genomics, which aims at comprehensive analysis of small organic molecules in biological systems. Techniques of mass spectrometry coupled to liquid chromatography (LC-MS, Liquid Chromatography-Mass Spectrometry) stand out as dominant methods in metabolomics experiments.

*to whom correspondence should be addressed

Although computational strategies have been used to filter and annotate mass peaks in LC-MS experiments (Dunn *et al.*, 2012), these methods do not include the addition of external information into a mathematical model in a principled way. Recently, (Rogers *et al.*, 2009) put forward a proof-of-concept in which information incorporated to a probabilistic model provides better annotation (Breitling *et al.*, 2013). Their Bayesian model, by means of appropriate prior distribution selection, introduces the elegant idea of using a set of known chemical reactions among candidate compounds to improve annotation since certain combinations, detected together, would make more biochemical sense than others.

The state-of-the-art in probabilistic annotation established by (Rogers *et al.*, 2009) did not include: an integrative computational implementation; a practical connection to public biological databases such as KEGG or MetaCyc (Altman *et al.*, 2013); nor a network-based output visualization schema. Therefore, our contribution is to fulfill these specific needs allowing easy access to this powerful statistical model for all metabolomics bioinformatics community.

2 RESULTS AND CONCLUSION

The platform chosen for implementation of these ideas was the well-known and established R programming environment (R Development Core Team, 2010), which incorporates a wide range of analyzes including successful tools that perform preprocessing of spectral data required for metabolite annotation (Supplementary Fig. 1) (Smith *et al.*, 2006; Kuhl *et al.*, 2011).

Following (Rogers *et al.*, 2010) brief suggestion on how their previous method could be extended to incorporate additional experimental information and metadata, we implemented modifications to the likelihood term. Expanding the likelihood function L in multiplicative independent terms allows one to account for additional orthogonal (independent) information sources: $L = L_N \cdot L_{rt} \cdot L_{iso}$, where subindexes N , rt and iso stand for measurement noise model, retention time error model and isotope profile error

model, respectively. For a complete model's description we refer the interested reader to the supplemental material.

The main product of a probabilistic annotation is a list of compound candidates ranked by their probabilities (Supplementary Fig. 2). In order to easily navigate over ProbMetabs results, we display tabular and dynamic network outputs along with supporting information, which assists practitioners to ultimately decide on most parsimonious annotations instead of forcing them to simplistically rely on the top probability assignment. All mass peaks are viewed as graph's nodes. Edges between two nodes are drawn if any candidate compound assigned to the outgoing node can be metabolized to any candidate compound assigned to the incoming node by means of a known biochemical reaction (Supplementary Fig. 3). ProbMetab is capable of producing reaction graphs and export them as standard Cytoscape input files or broadcasting the necessary graph data and attributes (color, shapes, etc) directly to Cytoscape Desktop using RCytoscape (Shannon *et al.*, 2013). This information can be easily overlaid with other widely used systems biology strategies such as correlation or partial correlation networks. If a mass spectra time-series or biological replicates are available, ProbMetab uses third party packages integrated downstream to export correlation or partial correlation graphs, along with their intersection/difference with the reaction graph.

Alternatively, a biologist can visualize ProbMetab's results in a simplified searchable web interface. Our package has a function that is responsible to consume an online web-service, which checks and renders the broadcast results as a web-page. The visualization approach was developed taking advantage of the cytoscape.js library (Lopes *et al.*, 2010) and its dependencies and can be easily integrated or embedded into any html5 web application.

ProbMetab's documentation brings two detailed case studies in which all its features are explored. Moreover, to highlight integration with downstream and upstream third party R packages, data analysis examples mentioned are carried out from raw data, following through preprocessing until it reaches ProbMetab's specific point of action. We used publicly available data from *Trypanosoma brucei*, causative agent of sleeping sickness, and an original dataset from *Saccharum officinarum* (sugarcane), an important biofuel source, to illustrate several points in typical metabolomics analysis sections.

The *T. brucei* dataset, obtained from the mzMatch.R project website, was chosen because it presents a set of metabolites identified with the aid of internal control standard compounds, being specially suited for performance evaluation. With this validation dataset, we compare the MetSamp (<http://www.dcs.gla.ac.uk/inference/metsamp/>) implementation from (Rogers *et al.*, 2009) with ProbMetab's implementation and show that, the efficient R/c++ integrated function (Eddelbuettel and François, 2011) had a 3 fold running time improvement over the MATLAB implementation. For both implementations the higher probability candidate was the true identity in up to 60% of the metabolites. However, instead of reporting only the higher probability candidate identity as proposed by Rogers *et al.* (2009), we show that exporting the complete ranking in summarized visualizations, up to 90% of metabolites identities are among the top-three higher probabilities. The full or filtered ranking allows the experimenter to associate the candidates with additional information present in this output and attribute the correct identity.

The sugarcane dataset was chosen to exemplify differential expression of annotated metabolites in contrasting environmental perturbation. We successfully recovered changes in a known stress response pathway (flavone and flavonol biosynthesis), showing the importance of a network-centric visualization for metabolite annotation in order to track metabolism changes. The benchmark dataset confirms, as preconceived by Rogers *et al.* (2009), that a probabilistic model using orthogonal data and metadata yields better automatic mass peak annotation. The perturbation dataset shows that probabilistic annotation can produce otherwise impossible interpretation for differential network connectivity.

We implemented a method to annotate compounds in a computational framework that allows the introduction of prior knowledge and additional spectral information. With the R package ProbMetab, we provide ways to summarize the results of series of analysis needed to extract information from complex high dimensional mass spectrometry data, and help the experimenter to track metabolism changes in the process of interest.

ACKNOWLEDGEMENT

We thank Bruno M Della Vecchia (USP), Dr Oscar Yanes's team (Rovira i Virgili University) and Paul Shannon (Fred Hutchinson Cancer Research Center) for critical discussions. We thank EBI/EMBO's courses for catalyzing cooperation among coauthors.

Funding: Microsoft Research and FAPESP grant 09/53161-6. FAPESP Temático grant 08/56100-5 and fellowships 10/14926-4, 12/05392-1, 10/15417-6. CAPES through Pós-graduação em Genética FMRP-USP.

REFERENCES

- Altman, T., Travers, M., Kothari, A., Caspi, R., and Karp, P. D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, **14**(1), 112.
- Breitling, R., Achcar, F., and Takano, E. (2013). Modeling challenges in the synthetic biology of secondary metabolism. *ACS synthetic biology*, **2**(7), 373–8.
- Dunn, W. B., Erban, A., Weber, R. J. M., Creek, D. J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., and Viant, M. R. (2012). Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*.
- Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, **40**(8), 1–18.
- Kuhl, C., Tautenhahn, R., Bottcher, C., Larson, T. R., and Neumann, S. (2011). CAMERA: An integrated strategy for compound spectra extraction and annotation of LC/MS data sets. *Analytical chemistry*.
- Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., and Bader, G. D. (2010). Cytoscape web: an interactive web-based network browser. *Bioinformatics*, **26**(18), 2347–2348.
- Rogers, S., Scheltema, R. A., Girolami, M., and Breitling, R. (2009). Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics (Oxford, England)*, **25**(4), 512–8.
- Rogers, S., Scheltema, R. A., Barrett, M., and Breitling, R. (2010). Bayesian Approaches for Mass SpectrometryBased Metabolomics. In *Handbook of Statistical Systems Biology*, pages 467–476.
- Shannon, P. T., Grimes, M., Kutlu, B., Bot, J. J., and Galas, D. J. (2013). RCytoscape: tools for exploratory network analysis. *BMC Bioinformatics*, **14**(1), 217.
- Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry*, **78**(3), 779–87.

Material suplementar do artigo publicado

Supplemental Material - ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS based metabolomics

Ricardo R. Silva¹, Fabien Jourdan^{2,3}, Diego M. Salvanha^{1,4},
Fabien Letisse³, Emilien L. Jamin^{2,3}, Simone Guidetti-Gonzalez⁵,
Carlos A. Labate^{5,6} and Ricardo Z.N. Vêncio^{1*}

¹LabPIB, Department of Computing and Mathematics FFCLRP-USP, University of Sao Paulo, Ribeirao Preto, Brazil

²INRA UMR1331, Toxalim, Research Centre in Food Toxicology, Toulouse, France

³Université de Toulouse, INSA, UPS, INP; LISBP, Toulouse, France

⁴Institute for Systems Biology, Seattle, Washington, USA

⁵Department of Genetics ESALQ-USP, University of Sao Paulo, Piracicaba, Brazil

⁶Laboratorio Nacional de Ciencia e Tecnologia do Bioetanol CTBE, Campinas, Brazil

November 16, 2013

Abstract

1 Summary:

We present ProbMetab, an R package which promotes substantial improvement in automatic probabilistic LC-MS based metabolome annotation. The inference engine core is based on a Bayesian model implemented to: (i) allow diverse source of experimental data and metadata to be systematically incorporated into the model with alternative ways to calculate the likelihood function and; (ii) allow sensitive selection of biologically meaningful biochemical reactions databases as Dirichlet-categorical prior distribution. Additionally, to ensure result interpretation by system biologists, we display the annotation in a network where observed mass peaks are connected if their candidate metabolites are substrate/product of known biochemical reactions. This graph can be overlaid with other graph-based analysis, such as partial correlation networks, in a visualization scheme exported to Cytoscape, with web and stand alone versions.

2 Availability and Implementation:

ProbMetab was implemented in a modular fashion to fit together with established upstream (xcms, CAMERA, AStream, mzMatch.R, etc) and downstream R package tools (GeneNet, RCytoscape, DiffCorr, etc). ProbMetab, along with extensive documentation and case studies, is freely available under GNU license at: <http://labpib.fmrp.usp.br/methods/probmetab/>.

3 Supplementary information:

Supplementary data (supporting tables and two detailed case studies) are available at Bioinformatics online and at ProbMetab website.

*to whom correspondence should be addressed

4 Contact:

rvencio@usp.br

5 Introduction

Metabolomics is an emerging field of study in post-genomics, which aims at comprehensive analysis of small organic molecules in biological systems. Techniques of mass spectrometry coupled to liquid chromatography (LC-MS, Liquid Chromatography-Mass Spectrometry) stand out as dominant methods in metabolomics experiments.

The workflow in conventional mass spectrometry analysis generates a set of characteristic mass peaks associated with chromatographic retention times and requires a large labor investment for manual inspection. Also, recent advances in ultra high resolution mass spectrometers that enabled experimental accuracies of less than 1 ppm (part per million of one unit of mass) are still insufficient to distinguish among candidate compounds in high mass windows (Kind and Fiehn, 2006).

Although computational strategies have been used to filter and annotate mass peaks in LC-MS experiments (Dunn *et al.*, 2012), these methods do not include the addition of external information into a mathematical model in a principled way. Recently, (Rogers *et al.*, 2009) put forward a proof-of-concept in which information incorporated to a probabilistic model provides better annotation (Breitling *et al.*, 2013). Their Bayesian model introduces the elegant idea of using a set of known chemical reactions among candidate compounds to improve annotation since certain combinations, detected together, would make more biochemical sense than others.

The state-of-the-art in probabilistic annotation established by (Rogers *et al.*, 2009) did not include: an integrative computational implementation; a practical connection to public databases biological such as KEGG or MetaCyc (Altman *et al.*, 2013); nor a network-based output visualization schema. Therefore, our contribution is to fulfill these specific needs allowing easy access to this powerful statistical model for all metabolomics bioinformatics community (Fig. 1).

6 Methods

Aiming clarity, we refer the interest reader to the original work (Rogers *et al.*, 2009) and its subsequent extension (Rogers *et al.*, 2010) for mathematical details. Here we briefly recall the essential ingredients of their model.

The surjective mapping between mass peaks onto compounds $m \mapsto c$; $m \in \{1, \dots, M\}$, $c \in \{1, \dots, C\}$ and $M \leq C$; is represented by a column vector $\vec{z}_m \in \{1, 0\}^C$ such that $|\vec{z}_m| = 1$. The joint distribution of all mappings given the observed mass spectra is: $f(\vec{z}_1, \dots, \vec{z}_M | x_1, \dots, x_M, I_1(t), \dots, I_M(t), \theta_{M+1}, \dots, \theta_N)$, abbreviated by: $f(Z | \vec{x}, \vec{\theta})$. In this notation, Z concatenates all column vectors in a matrix; x_m is the m -th mass peak’s actual observed mass and $I_m(t)$ its measured intensity (possible a time-series and/or repeated experiments); and $\vec{\theta}$ encodes all other N necessary parameters such as $\vec{I}(t)$, spectrometer precision, known compounds in the biological database chosen, adjacency matrix of bioreaction graphs, hyperparameters, and so on. Gibbs sampling is performed as usual to access such complicated joint $f(\cdot)$ making use of its marginals: $f(\vec{z}_m | Z^{(-m)}, \vec{x}, \vec{\theta})$ where $Z^{(-m)}$ is Z lacking the m -th column relative to the mass peak assignment $\vec{z}_m$. These probability distributions, by Bayes rule, are proportional to: $L(x_m | Z, \vec{x}^{(-m)}, \vec{\theta}) \cdot p(\vec{z}_m | Z^{(-m)}, \vec{x}^{(-m)}, \vec{\theta})$. $L(\cdot)$ and $p(\cdot)$ represent the likelihood and *prior* functions, respectively, and $\vec{x}^{(-m)}$ means $\vec{x}$ lacking x_m .

As pioneered by (Rogers *et al.*, 2009), we still use a straightforward Dirichlet-categorical *prior* $p(\cdot)$ on the chemical reaction graph to enhance the odds of more connected compounds over less connected ones. However, we integrate this model to real biochemical networks/pathways databases such as KEGG or BioCyc (Altman *et al.*, 2013) in a sensible way in order to retain biologically meaningful *priors*. Although this integration is made by web-services provided by public databases, ProbMetab allows

in-house (possibly curated or specific) pathway databases to be used to inform the *prior* component of this Bayesian model.

In a subsequent review, (Rogers *et al.*, 2010) briefly suggested how their previous method could be extended to incorporate additional experimental information and metadata since the original formulation considered only a measurement noise model. We followed these leads expanding the likelihood function $L(\cdot)$ in multiplicative independent terms to account for additional orthogonal information sources:

$$L(x_m|Z, \vec{x}^{(-m)}, \vec{\theta}) \propto L_N(x_m|Z, \vec{\theta}_N) \cdot L_{rt}(x_m|Z, \vec{\theta}_{rt}) \cdot L_{iso}(x_m|Z, \vec{\theta}_{iso}) \quad (1)$$

where $L_N(\cdot)$ is the measurement noise model; $L_{rt}(\cdot)$ is the retention time error model and; $L_{iso}(\cdot)$ is the isotope profile error model. $\vec{\theta}_\star$'s are their respective appropriate parameters subset of $\vec{\theta}$.

We implemented computationally this improved likelihood function in ProbMetab with modularity in mind. This is important because depending on particular dataset, equipment setup, experimental conditions and so on, a given multiplicative term in $L(\cdot)$ can be dropped, or a new one included.

For example, there are isotope profile error models proposed for Orbitrap (Thermo Fisher) mass spectrometry platform that can take advantage of known natural relative abundance of carbons ^{12}C and ^{13}C . Taking into account measured peak intensities would improve the odds of a given compound over others if it fulfills an expected ratio (Weber *et al.*, 2011).

Additionally, one may include a retention time error model to take advantage of known differences in liquid chromatography elution time of chemically distinct compounds with similar mass. Models to predict retention time exist (Creek *et al.*, 2011) although still error prone. The decision to include or drop this term, $L_{rt}(\cdot)$, is up to the practitioner and ProbMetab is equally prepared to perform probabilistic annotations with both options.

The likelihood term $L_N(\cdot)$ is mandatory and was the sole term in (Rogers *et al.*, 2009) original model's $L(\cdot)$. This measurement noise model was introduced as a gaussian density with a fixed variance $1/\gamma$ given by the mass spectrometer's precision γ . Therefore, the likelihood $L_N(\cdot)$ was proportional to $\exp(-\frac{1}{2}\gamma(x_m/y_c - 1)^2)$ and the parameters $\vec{\theta}_N = (\vec{y}, \gamma)$, where $\vec{y} = (y_1, \dots, y_C)$ are the exact masses of compounds attributable to m , and the precision γ .

Experience shows that relying on nominal precision figures from vendors is not optimal since it can vary with several factors in an actual experimental setup and site (Böcker *et al.*, 2009). In order to overcome this limitation, we replaced $L_N(\cdot)$ following the ideas put forward by (Böcker *et al.*, 2009) when modeling isotope and noise error simultaneously but, here, considering only the noise part:

$$L_N(x_m|Z, \vec{y}, w) \propto 1_\omega \left(1 - \Phi \left(\frac{1}{\sigma_\omega} (|x_m/y_c - 1| - \mu_\omega) \right) \right) \quad (2)$$

where 1_ω is the indicator function for the set ω composed by all j candidate compounds in the database for which $|x_m - y_j| \leq w$; μ_ω and σ_ω are the average and standard deviation of all masses in $\vec{y}$ listed in ω ; Φ is the normal cumulative distribution function; and w is a researcher-defined mass difference limit.

The set ω , defined by the threshold w , is an improvement that inhibits the unwanted effect of letting highly connected compounds (hubs in the metabolic networks) overinfluence the *posterior* when using the originally proposed Dirichlet-categorical *prior*. It captures the practitioners' common understanding that, in real experimental setups, there are mass peaks that would never be confused beyond certain mass limit. The indicator function rules out unwanted compounds turning their $L_N(\cdot)$ to zero. Moreover, considering variance over ω instead of a static precision value γ , the model becomes data-centric instead of relying on equipment (ideal) specifications. These improvements come from taking (Rogers *et al.*, 2009) and (Böcker *et al.*, 2009) ideas together.

The product of a probabilistic model is a list of ranked assignments in which the probability values create a gradient among candidates. If a mass peak m has only one feasible candidate c (e.g. $|\omega| = 1$), there are no ties to be broken and $z_{cm} = 1$ with probability one. On the other hand, if no external information added is able to untie assignments from m to $c_1, \dots, c_{|\omega|}$, the probability values are all equal to the uninformative $1/|\omega|$. Also, even if there is a probability gradient among candidate compounds

for a given mass peak, it is conceivable that the true assignment could not be the 1st ranked but rather be among the top ranked compounds. Due to several reasons, ranging from modeling limitations to measurement noise, the rankings could be imprecise and a probabilistic annotation, instead of a simple single one-to-one assignment, would be best considered as a whole.

7 Results and Conclusion

The platform chosen for implementation of these ideas was the well-known and established R programming environment (R Development Core Team, 2010), which incorporates a wide range of analyzes including successful tools that perform preprocessing of spectral data required for metabolite annotation (Fig. 1) (Smith *et al.*, 2006; Kuhl *et al.*, 2011).

Figure 1: Schematic representation of ProbMetab package scope in spectral data analysis, showing the integration in a programming environment that provides different solutions for analysis and information integration associated with the model.

The main product of a probabilistic annotation is a list of compound candidates ranked by their probabilities (Fig. 2). In order to easily navigate over ProbMetab’s results, we display tabular and dynamic network outputs along with supporting information, which assists practitioners to ultimately decide on most parsimonious annotations instead of forcing them to simplistically rely on the top probability assignment. All mass peaks are viewed as graph’s nodes. Edges between two nodes are drawn if any candidate compound assigned to the outgoing node can be metabolized to any candidate compound assigned to the incoming node by means of a known biochemical reaction (Fig. 3). ProbMetab is capable of producing reaction graphs and export them as standard Cytoscape input files or broadcasting the necessary graph data and attributes (color, shapes, etc) directly to Cytoscape Desktop (Shannon *et al.*, 2013). This information can be easily overlaid with other widely used systems biology strategies such as correlation or partial correlation networks. If a mass spectra time-series or biological replicates are available, ProbMetab uses third party packages integrated downstream to export correlation or partial correlation graphs, along with their intersection/difference with the reaction graph.

Alternatively, a biologist can visualize ProbMetab’s results in a simplified searchable web interface. Our package has a function that is responsible to consume an online web-service, which checks and

Figure 2: Screenshot of ProbMetab’s tabular output showing how the probability ranking is associated to spectra and database weblinks.

renders the broadcast results as a web-page. The visualization approach was developed taking advantage of the cytoscape.js library (Lopes *et al.*, 2010) and its dependencies and can be easily integrated or embedded into any html5 web application.

ProbMetab’s documentation brings two detailed case studies in which all its features are explored. Moreover, to highlight integration with downstream and upstream third party R packages, data analysis examples mentioned are carried out from raw data, following through preprocessing until it reaches ProbMetab’s specific point of action. We used publicly available data from *Trypanosoma brucei*, causative agent of sleeping sickness, and an original dataset from *Saccharum officinarum* (sugarcane), an important biofuel source, to illustrate several points in typical metabolomics analysis sections.

The *T. brucei* dataset, obtained from the mzMatch.R project website, was chosen because it presents a set of metabolites identified with the aid of internal control standard compounds, being specially suited for performance evaluation. With this validation dataset we compare the MetSamp (<http://www.dcs.gla.ac.uk/inference/metsamp/>) implementation from (Rogers *et al.*, 2009) with ProbMetab’s implementation and show that, the efficient R/c++ integrated function (Eddelbuettel and François, 2011) had a 3 fold time improvement over the MATLAB implementation. For both implementations the higher probability candidate was the true identity in up to 60% of the metabolites. However, instead of reporting only the higher probability candidate identity as proposed by (Rogers *et al.*, 2009) we show that exporting the complete ranking in summarized visualizations, up to 90% of metabolites identities are among the top-three higher probabilities. The full or filtered ranking allow the experimenter to associate the candidates with additional information present in this outputs and attribute the correct identity.

The sugarcane dataset was chosen to exemplify differential expression of annotated metabolites in contrasting environmental perturbation. We successfully recovered changes in a known stress response pathway (flavone and flavonol biosynthesis), showing the importance of a network-centric visualization for metabolite annotation in order to track metabolism changes. The benchmark dataset confirms, as preconceived by (Rogers *et al.*, 2010), that a probabilistic model using orthogonal data and metadata yields better automatic mass peak annotation. The perturbation dataset shows that probabilistic annotation can produce otherwise impossible interpretation for differential network connectivity.

Figure 3: Conceptual representation of the network visualization for a hypothetical metabolome probabilistic annotation. Nodes are mass peaks and edges are drawn only if there is at least one biochemical reaction transforming an annotated compound into some annotated compound. Here, m_i , m_j and m_k are observed mass peaks. A to D are candidate compounds to annotate m_i with probabilities $p_a \leq \dots \leq p_d$. Analogously, E to I are assigned with probabilities p_e to p_i . Since there is a known biochemical reaction metabolizing m_i 's 2nd best candidate B to I , an edge is drawn from m_i to m_j . There are no known reactions from any m_i 's candidates to m_k 's candidates thus no edge is drawn.

We implemented a method to annotate compounds in a computational framework that allows the introduction of prior knowledge and additional spectral information. With the R package ProbMetab we provide ways to summarize the results of series of analysis needed to extract information from complex high dimensional mass spectrometry data, and help the experimenter to track metabolism changes in the process of interest.

Acknowledgement

We thank Bruno M Della Vecchia (USP), Dr Oscar Yanes's team (Rovira i Virgili University) and Paul Shannon (Fred Hutchinson Cancer Research Center) for critical discussions. We thank EBI/EMBO's courses for catalyzing cooperation among coauthors.

Funding Microsoft Research and FAPESP grant 2009/53161-6. FAPESP fellowships (2010/14926-4, 2012/05392-1, 2010/15417-6). We thank the Programa de Pos-graduação em Genética FMRP-USP for CAPES travel grants and initial funding awarded to RRS.

References

- Altman, T., Travers, M., Kothari, A., Caspi, R., and Karp, P. D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, **14**(1), 112.
- Böcker, S., Letzel, M. C., Lipták, Z., and Pervukhin, A. (2009). SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics (Oxford, England)*, **25**(2), 218–24.
- Breitling, R., Achcar, F., and Takano, E. (2013). Modeling challenges in the synthetic biology of secondary metabolism. *ACS synthetic biology*, **2**(7), 373–8.
- Creek, D. J., Jankevics, A., Breitling, R., Watson, D. G., Barrett, M. P., and Burgess, K. E. V. (2011). Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography-

- Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Analytical chemistry*.
- Dunn, W. B., Erban, A., Weber, R. J. M., Creek, D. J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., and Viant, M. R. (2012). Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*.
- Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, **40**(8), 1–18.
- Kind, T. and Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC bioinformatics*, **7**(1), 234.
- Kuhl, C., Tautenhahn, R., Bottcher, C., Larson, T. R., and Neumann, S. (2011). CAMERA: An integrated strategy for compound spectra extraction and annotation of LC/MS data sets. *Analytical chemistry*.
- Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., and Bader, G. D. (2010). Cytoscape web: an interactive web-based network browser. *Bioinformatics*, **26**(18), 2347–2348.
- Rogers, S., Scheltema, R. A., Girolami, M., and Breitling, R. (2009). Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics (Oxford, England)*, **25**(4), 512–8.
- Rogers, S., Scheltema, R. A., Barrett, M., and Breitling, R. (2010). Bayesian Approaches for Mass SpectrometryBased Metabolomics. In *Handbook of Statistical Systems Biology*, pages 467–476.
- Shannon, P. T., Grimes, M., Kutlu, B., Bot, J. J., and Galas, D. J. (2013). RCytoscape: tools for exploratory network analysis. *BMC Bioinformatics*, **14**(1), 217.
- Smith, C. A., Want, E. J., O’Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry*, **78**(3), 779–87.
- Weber, R. J. M., Southam, A. D., Sommer, U., and Viant, M. R. (2011). Characterization of isotopic abundance measurements in high resolution FT-ICR and Orbitrap mass spectra for improved confidence of metabolite identification. *Analytical chemistry*, **83**(10), 3737–43.

Artigo em processo de submissão

A metabolomic protocol for plant taxonomy by MALDI mass spectrometry

Madeleine Ernst^{1,†}, Denise B. Silva^{1,†}, Ricardo Silva^{2,†}, Marcelo Monge³, João Semir³, Ricardo Z. N. Vêncio², Norberto P. Lopes^{1,*}

¹Núcleo de Pesquisa em Produtos Naturais e Sintéticos, Department of Physics and Chemistry, Faculty of Pharmaceutical Sciences, University of São Paulo, Av. do Café, s/n, 14040-903 Ribeirão Preto, São Paulo, Brazil.

²Department of Computing and Mathematics, FFCLRP, University of São Paulo, Av. do Café, s/n, 14040-903 Ribeirão Preto, São Paulo, Brazil.

³Department of Plant Biology, Institute of Biology, UNICAMP, CP 6109, 13083-970 Campinas – SP, Brazil.

KEYWORDS

MALDI-TOF MS, metabolic fingerprinting, plant taxonomy, multivariate data analysis, R environment

ABSTRACT

Matrix- assisted laser desorption/ionization time-of flight mass spectrometry (MALDI-TOF MS) has been widely used for the identification and classification of microorganisms based on their proteomic fingerprints. However, the use of MALDI-TOF MS in plant research has been very limited. Due to the wide variety of secondary metabolites in the plant kingdom, metabolomics approaches have been proposed for the classification of plants using hyphenated or direct injection mass spectrometry methods or nuclear magnetic resonance. However, neither of these approaches has been used for the classification of plants from different genera, and MALDI-TOF MS has not yet been used as an analytical technique for this application. In the present study, a first protocol is proposed for metabolic fingerprinting by MALDI-TOF MS using three different MALDI matrices with subsequent multivariate data analysis by in-house algorithms in the R environment for the taxonomic classification of plants from different genera, families and orders. The primary advantage of MALDI-TOF MS, in addition to showing less ion suppression effects in compound mixtures compared with electrospray ionization mass spectrometry (ESI-MS), is its extremely rapid data acquisition, which facilitates the rapid classification of species with challenging botanical identification. Correct taxonomic classification is essential for the conservation of biodiversity. All plants analyzed were collected in areas of the Brazilian Savanna (*Cerrado*), which is classified as a “hotspot for conservation priority”. With the proposed protocol, classifications similar to the literature-based taxonomic classifications could be obtained. The present work therefore highlights the great potential of applying MALDI-TOF MS for the taxonomic classification of plants and, furthermore, provides a preliminary foundation for future research.

Revisão de literatura submetida para publicação

Mass Spectrometry in Plant metabolomics strategies: from analytical platforms to data acquisition and processing

Journal:	<i>Natural Product Reports</i>
Manuscript ID:	NP-REV-09-2013-070086
Article Type:	Review Article
Date Submitted by the Author:	03-Sep-2013
Complete List of Authors:	Lopes, Norberto; University of São Paulo, FCFRP Ernst, Madeleine; University of São Paulo, FCFRP Silva, Denise; FCFRP-USP, NPPNS Silva, Ricardo; FFCLRP-USP, Física Vencio, Ricardo; FFCLRP-USP, Física

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/npr

REVIEW

Mass Spectrometry in Plant Metabolomics Strategies: From Analytical Platforms to Data Acquisition and Processing

Madeleine Ernst^a, Denise Brentan Silva^a, Ricardo Silva^b, Ricardo Z.N. Vêncio^b and Norberto Peporine Lopes^{*a}

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
DOI: 10.1039/b000000x

Plant metabolomics is a relatively recent research field that has gained increasing interest in the past few years. Numerous review articles and guide books on this subject have been published. This review article focuses on the current limitations inherent to the so-called modern mass spectrometry techniques, especially in combination with electrospray ionisation (ESI), an ionisation technique that is most commonly applied to metabolomics studies by using modern mass spectrometry methods. As a possible alternative to ESI, perspectives on matrix-assisted laser desorption/ionisation (MALDI-MS) in metabolomics studies are introduced, and this method is still not widespread in the field. As in metabolomics studies, the results must always be interpreted in context with applied sampling procedures as well as data analysis, and different sampling strategies are also introduced. The importance of data analysis is illustrated in the metabolic network modelling example.

1	Introduction
2	Classification of Metabolomics Approaches
3	Setup of a Metabolomics Experiment
3.1	Sample Preparation
3.1.1	Harvesting
3.1.2	Processing before extraction
3.1.3	Extraction
3.1.4	Pre-analytical sample preparation
3.2	Metabolite Analysis
3.2.1	ESI source reaction: Ion generation and suppression
3.2.2	MALDI source: Ion generation and matrix effects
3.2.3	The use of fragmentation spectra (MS/MS and MS ⁿ) in metabolomics
3.2.4	Hyphenated Mass Spectrometry Methods: GC/LC-MS
3.2.5	Direct Injection MS (DIMS)
4	Preprocessing and Statistical Analysis of Plant Metabolomics Data
4.1	Introduction
4.2	Preprocessing Mass Spectrometric Metabolomics Data
4.3	Metabolic Network Modelling
5	Conclusions and Future Perspectives
6	Acknowledgments
7	References

1 Introduction

Rochfort¹ defines metabolomics as the measurement of all the metabolites; *i.e.*, the metabolome, in a given system (cell, tissue, or organism) under a given set of conditions, and similar

definitions are also found in other sources¹⁻⁵. Metabolomics is a relatively new research field^{1, 5-10}, however, after the introduction of the term “metabolome” in 1998 by Oliver and collaborators¹¹ publications on the subject increased exponentially. An international metabolomics society was founded in 2004, which today publishes its own journal⁶. The areas in which metabolomics techniques have been applied are numerous, and new applications are continuously being explored. Examples include metabolomics as a functional genomics tool¹²⁻¹³, applications in pharmaceutical and medical research¹⁴⁻³⁰, the discrimination of different genotypes and the effects of genetic manipulation³¹⁻³², natural product discovery²⁶⁻²⁷ and diverse applications in plant research³¹⁻³³ (list of examples inspired by Brown et al.³⁴, Shulaev³⁵, and Wishart³⁶). Based on a bibliographic search on the Chemical Abstract Databank, the number of papers in plant metabolomics increased as rapidly as research published in general metabolomics and comprises approximately 20% of the annual total papers published on metabolomics (Figure 1). Compared to animals, plants contain a remarkably wide variety of metabolites; the total number of metabolites present in the plant kingdom is estimated at 200,000 or more³⁷. For this reason, plant metabolomics has found applications in an especially broad variety of areas³⁸, such as the fingerprinting of geno- or ecotypes for taxonomic or biochemical purposes, comparisons of mutants and transgenic plants with their wild-types, the influence of a metabolite profile by external stimuli, interactions between plants and herbivores, developmental processes, quality control of medicinal herbs and activity determinations of medicinal plants³⁹. Numerous books containing protocols and reviews on plant metabolomics, for

example, the most recently published Handbook of Plant Metabolomics⁴⁰, as well as several review articles on the subject, have already been published, including in high impact journals^{1, 7, 9, 38-39, 41-51}.

The variety of plant metabolomics applications is very diverse, as are the techniques used for the analysis of plant metabolomes. Hyphenated methods, such as mass spectrometry coupled to gas chromatography or liquid chromatography (GC-MS and LC-MS), direct injection mass spectrometry (DIMS), nuclear magnetic resonance (NMR), capillary electrophoresis mass spectrometry (CE-MS), high-performance liquid chromatography with photodiode array detection (HPLC-PDA) and thin layer chromatography with UV detection (TLC-UV), among others, have been described^{7, 39, 43}. Out of these techniques, hyphenated methods such as GC-MS and LC-MS, DIMS and NMR are likely the most commonly applied methods. GC-MS was introduced in the early 1950s^{1, 9, 52} and is a traditional analytical technique used for the analysis of apolar, volatile, thermostable, low molecular weight compounds. Despite its restrictions, and because of the limitations of analysable compounds, it belongs to the most commonly applied techniques, primarily because of the predictability and reproducibility of fragmentation patterns and thus the availability of databases⁴³. As GC-MS is discussed in more detail in other articles published in this issue, we would like to focus here on the so-called modern techniques for the analysis of more polar compounds, such as LC-MS, DIMS and MALDI-MS, which were only very recently introduced to plant metabolomics research. Because the most common ionisation source applied in such techniques is electrospray ionisation, special emphasis shall be given to the limitations and considerations inherent to ESI. One alternative ionisation method to ESI is matrix-assisted laser desorption/ionisation (MALDI). It has only been introduced to plant metabolomics very recently⁵³⁻⁵⁵, only its potential and possible limitations shall be discussed. The present limitations and bottle-necks of metabolomics research do not solely depend on analytical platforms but also depend on sample preparation and data analysis. A detailed review of sources of influence during all steps of sample preparation and a variety of sample preparation methods are therefore also introduced. In data analysis, we concentrated on the pre-processing of mass spectrometric data, and we noted the current limitations of metabolic network modelling, which is highly dependent on the limitations of previously discussed sample preparation methods and metabolite analysis. Outside these limitations, various successful studies have been published in plant metabolomics research, which are cited throughout the text. It is therefore expected that significant and exciting progress will still be made in this area in the coming years.

2 The Classification of Metabolomics Approaches

Due to the lack of simple automated and sufficiently sensitive analytical strategies, the quantification of all metabolites in a given system is impossible at present^{3-5, 49, 56}. Out of this impossibility, several terms evolved to differentiate between different approaches of metabolome analyses. However, because metabolomics is still a developing research field, and because its terminologies are still evolving^{3, 57}, different sources give slightly different definitions. Table 1 lists some of the most common

terms.

Most sources (see Table 1) designate target analysis, metabolite profiling, metabolomics and metabolic fingerprinting as separate approaches to metabolome analysis. Villas-Bôas and collaborators⁵⁸ subdivided metabolomics analytical methodologies into only two categories, namely, target analysis and metabolite profiling. The authors define target analysis as a procedure that must include the identification and absolute quantification of selected metabolites. They define metabolite profiling as the rapid analysis of a large number of different metabolites in which quantification is not mandatory. They then subdivided metabolite profiling further into metabolic fingerprinting (analysis of the endometabolome = intracellular metabolites⁵⁹) and metabolic footprinting (analysis of the exometabolome = extracellular metabolites⁵⁹), a term that was introduced in 2003 by Allen and collaborators¹².

3 Setup of a Metabolomics Experiment

Metabolomics studies are carried out in stages^{4, 7, 43}. The analysis of data obtained from metabolomics experiments is complex. Different platforms and experimental designs produce data in various formats and units, making the exchange of metabolomics data difficult. Providing full information regarding the experimental design and sample preparation route for the development of an analytical methodology and an approach to data analysis that includes all such information and is simultaneously easy to handle is a challenging task. To integrate and distribute metabolomics research data to a broad research community, a metabolomics information standard is mandatory so that metabolite data and metadata (including detailed experimental information such as sample preparation routes, instrument settings, and analysis conditions) can be easily interpreted by any laboratory and the results can be verified outside the original source laboratory⁴¹. For example, Bino et al.⁴¹ have presented such an information standard, named MIAMET, or the Minimum Information About METabolomics experiment. This standard suggests the minimal information that must be presented in a study on metabolomics, including minimal statements about the experimental design, such as sampling, preparation, metabolite extraction, and derivatisation, a metabolic profiling design and metabolite measurements and specifications. A data model that supplements MIAMET for plant metabolomics has been developed at the University of Wales, Aberystwyth, UK, which is named ArMet, an Architecture for Metabolomics (see Jenkins et al.⁶⁴ and <http://www.armet.org/>). However, because metabolomics research is still in its beginnings, a universally applicable standard procedure has not yet been established, and single publications only include suggestions that will have to be refined in the future⁴¹. The adoption of this type of bottom-up community-generated standards played a crucial role in the success experienced by the gene expression field.

Table 1: Classification of metabolomics approaches

Term	Explanation
Metabolic fingerprinting 2, 38, 42, 56, 60-61	Rapid high throughput screening of all detectable analytes in a sample without mandatory identification. Large differences between samples are found by multivariate data analysis of the spectra from unidentified peaks. Eventually, prominent, sample class-defining metabolites are then identified in a further step. Metabolic fingerprinting can be performed before metabolite profiling to guide a research project. Sample preparation, separation and detection are therefore as simple and as fast as possible. Fiehn ⁴² mentions metabolic fingerprinting specifically in relation to analytical techniques without previous chromatographic separation, such as mass spectrometry, infrared spectroscopy or nuclear magnetic resonance spectroscopy.
Metabolite profiling/ Metabolic profiling 2, 38, 42, 56, 60-61	The identification and quantification of a number of predefined metabolites, which may be associated with the same pathway or belong to the same class of compounds. The analytical method may be matched to the specific chemical properties of the analysed substances. Spectra are then transformed into lists of metabolites and their concentrations. In contrast to metabolite fingerprinting, metabolite profiling also comprises the quantification of the analysed metabolites. The results obtained by metabolite profiling should ideally be independent of the applied analytical technique.
Metabolite target analysis 2, 38	Metabolites are selected (targeted) prior to analysis, for example, by optimised extraction or specific separation and/or detection, and analysis is focused only on these substances. The selection of metabolites can be based on previous knowledge or it can follow broad-scale metabolomics analyses.
Metabolome 3, 11, 38, 62-63	Oliver and collaborators ¹ introduced the term "metabolome" in 1998. Goodacre and collaborators ³ state that there is a disagreement in the scientific community on the exact definition of the metabolome; according to Hall ³⁸ , the metabolome comprises all small molecules present in living things. Larger molecules (> 1500 Da), typical amino acids and sugar polymers are excluded out of practicality in terms of extraction and detection. Hollywood and collaborators ⁶² complemented that the metabolome comprises all measurable metabolites under a particular physiological or developmental state of the cell and that these metabolites have a typical mass-to-charge ratio less than 3000 m/z. Beecher ⁶³ defines the metabolome as all small molecules that participate in general metabolic reactions and are essential for the normal living functions of a cell.
Metabolomics 2, 5, 38, 42, 45, 50, 56, 62	Some authors define metabolomics as the comprehensive, qualitative, and quantitative analysis of all metabolites in a biological system by high-throughput analytical strategies. However, others state that metabolomics is the measurement of as many metabolites as possible (not all of them because of the lack of appropriate detection methods).

3.1 Sample preparation

The sample preparation has a direct influence on the resulting metabolomics experiments and is therefore an especially critical step. Sample preparation can basically be subdivided into the following four steps: 1) harvesting, 2) processing before extraction (for example, drying), 3) extraction and 4) pre-analytical sample preparation. The execution of all these steps

may not be necessary and can be avoided in some cases depending on the chosen analytical method or the physical-chemical properties of the analytes. In the following, the described steps are explained in more detail.

3.1.1 Harvesting

A large variety of metabolites are produced by plants, and they can be subdivided into primary and secondary metabolites. Secondary metabolites are uniquely related to signalling and plant defence. However, both secondary and primary metabolites suffer from significant environmental influence⁶⁵⁻⁶⁶. The seasonality, developmental stage and age, circadian rhythm, temperature, water availability, UV radiation, altitude, atmospheric composition, soil nutrients, and tissue damage may induce qualitative and/or quantitative variations in the metabolite composition^{10, 67}. The first difficulties therefore arise with the harvest of the plant material^{9, 68}. A plant metabolome may vary depending on the time or place of harvest^{7, 9, 10, 68}. The influence of environmental conditions is extremely relevant for plant material collected in the natural habitat of plants; however, it may be considered minor for plants cultivated in a greenhouse or in cell culture plants. According to Villas-Bôas and collaborators¹⁰, the principal source of variability during harvest is light. For this reason, the authors name photosynthesis as an influencing factor because it causes the intensity of metabolic processes to be heavily dependent on the availability of light. Additionally, the wavelength of the light also influences the metabolite profiles, mainly compounds related to UV protection, such as flavonoids, phenylpropanoids and others⁶⁹⁻⁷⁰, but the stimulation of biosynthesis of other compounds has also been reported⁷¹. Light does not reach each leaf to the same extent; upper leaves may therefore reveal different metabolite profiles than lower leaves of the same plant¹⁰. Furthermore, it should be emphasised that not only leaves but also subterranean parts of the plant exhibit light-dependent metabolite profiles⁶⁸. Accurate timing is an important factor that should be considered because the levels of metabolites (primary as well as secondary metabolites) vary throughout the day⁷²⁻⁷³. Diurnal changes in essential oils have been studied extensively; for example, the eugenol (sesquiterpene) concentration is reduced by 87% when comparing the harvesting times at midday and at 17 h⁷⁴. Variations in other substances have also been described, such as alkaloids⁷⁵⁻⁷⁶ and flavonoids⁷⁷. Moreover, the developmental stage of the plant at the time of harvest affects the metabolite profile^{9, 10, 68}. Verpoorte and collaborators⁹ therefore suggest measuring a series of samples harvested at different times of the day at distinct developmental stages and ages to determine the biological variation and to set standard conditions for the experiment. Younger tissues show a higher biosynthetic rate, which may exhibit an inverse allelochemical substance relationship⁷⁸⁻⁷⁹. Younger and older leaves therefore show differences in the metabolite profiles^{72, 80}, such as the higher accumulation of helenalin, a sesquiterpene lactone, in young *Arnica montana* leaves⁸¹. If only a small amount of plant material is needed, another possibility is to minimise the biological variability by harvesting all plant material at the same period of day/night within a very small time frame and to choose leaves or other parts of the plant that have been similarly exposed to light¹⁰.

⁴⁶ The significant influence of biological variance can be illustrated by unpublished data from Sumner and collaborators⁴⁹, which suggest an average biological coefficient of variation of 50% for *Medicago truncatula*, or by data published by Roessner⁵ and collaborators⁸², which present a wider biological than analytical variability in potato tubers analysed by gas chromatography-mass spectrometry. However, not only the developmental stage of the plant tissue influences the metabolite profile, but different plant organs also show significantly different metabolite profiles. For this reason, the separate analysis of different plant organs with separation by tissue age is fundamental to obtaining reliable results. Chemical differences can even be found within the same leaves (between veins and the rest)⁸³ or other plant parts, such as the glucosinolates distributed irregularly in *Arabidopsis thaliana* leaves⁸⁴ and flavonoids and hypericins in the leaves and petals of *Hypericum perforatum*⁸⁵, as evaluated by MALDI imaging.

3.1.2 Processing before extraction

²⁰ Harvesting stresses and wounds the plant and therefore causes alterations in the metabolome⁹. Changes in the plant's metabolism occur from seconds up to a few minutes; for this reason, harvesting must be performed as quickly as possible, and metabolism should be stopped immediately after harvesting^{9, 10}. According to Moritz and collaborators⁷, the most frequently used and, in their opinion, best current technique for this purpose is the freezing of plant material in liquid nitrogen. Samples can be stored at low temperatures until further use. Cells are destroyed³⁰ during freezing, and thawing may therefore induce all manner of biochemical conversions^{39, 86}. A suitable technique would be freeze-drying, which is based on lyophilisation and cryodesiccation. Water is removed by this technique, and the problems mentioned before are therefore minimised. Despite the³⁵ reduction of enzymatic activity, compound extraction may still be decreased because of the higher adsorption of some compounds in the plant tissue⁸⁷, which furthermore decreases the levels of some *Arabidopsis* metabolites during freeze-drying⁸⁸.

In addition to freezing or freeze-drying, the samples may also be⁴⁰ subjected to a drying process by an oven, air⁷² or by microwave heating, which destroys enzyme activities⁸⁹. The selection of a drying method should always consider the targeted molecules for analysis because some methods are incompatible with certain metabolites, such as volatile and thermo-unstable compounds,⁴⁵ among others.

Another point that is frequently discussed in plant metabolomics reviews is that each plant organ, tissue, or cell type contains different characteristic metabolites, and different external stimuli can lead to different metabolite profiles in the same tissues.⁵⁰ Ideally, single cells would have to be analysed. However, the analytical techniques employed in metabolomics still have limitations. Specific cell types and tissues must be used for extraction to obtain sufficient levels of all metabolites. The results obtained by traditional plant metabolomics experiments⁵⁵ therefore merely show a non-uniform average of the metabolite content distributed over different organs and tissues^{46, 68}. A first attempt at single cell metabolomics was published by Schad and collaborators⁹⁰ for *Arabidopsis thaliana*. Specific cell types or single cells have been analysed by laser dissection⁹¹⁻⁹⁴. Research

in this area has been designated micrometabolomics, and several recent reviews on the topic have been published^{53, 88}. By performing a metabolomics analysis of a single cell type, additional information that could not be shown by conventional metabolomics may be obtained. One example is a study of⁶⁵ *Catharanthus roseus*, in which three cell types in different parts of the same leaf were detected, each producing different compounds (irridoid, tryptamine and dimeric alkaloids)⁹⁵. Despite the fact that this technique has advantages, it is still very rudimentary and exhibits several limitations, including the ability⁷⁰ to perform routine analyses of many metabolites.

Further processing before extraction can include the pulverisation of the plant material into small particles to improve the extraction process. A mixer mill system or liquid nitrogen and manual homogenisation by mortar and pestle may be used for this⁷⁵ purpose³⁹. Non-homogeneous particle sizes of the pulverised plant material are another important point to consider because particle sizes can vary in relation to the hardness of different parts of the plant and the efficiency of the technique used for pulverisation. Homogenisation of particle sizes by sieving can⁸⁰ minimise experimental variation and improve reproducibility, but it also considerably increases the time used for sample preparation.

3.1.3 Extraction

⁸⁵ Another crucial step that influences the outcome of plant metabolomics experiments is extraction. The extraction of the plant material should occur within a very short time frame to prevent biochemical processes from occurring⁷. Because⁹⁰ metabolomics studies are usually large-scale experiments, extraction methods should be as simple and as fast as possible⁹⁶. Solvent extraction, vapour phase extraction, steam distillation, supercritical fluid extraction or ionic liquids may be used for extraction. At present, solvent extraction is most commonly used⁹⁵ in metabolomics experiments^{9, 39, 72, 97-98}. Steam distillation can be used for volatile compounds, and supercritical fluid extraction and ionic liquids are still uncommon because there is little experience in applying these methods to high-throughput experiments^{9, 72}. Degradation, modification, and loss of¹⁰⁰ metabolites during extraction should be minimised^{7, 9, 10}. However, there is no current method for the extraction of all metabolites without artefact formation or compound degradation^{7, 9, 99}. Plant tissue metabolites are highly diverse. Non-polar compounds (terpenoids and fatty acids from cell membranes),¹⁰⁵ medium polarity compounds (most of the secondary metabolites), and polar compounds (most of the primary metabolites, such as sugars and amino acids) may be present. There is no solvent that is able to extract all compounds at the same time^{9, 72}.

Metabolite extraction from plant tissue is labour-intensive, crucial¹¹⁰ for reproducibility, normally not automated and generally includes high costs and time consumption. Several critical factors must be considered during extraction, such as the ratio of solvent and plant material, solvent characteristics, the time of extraction, the temperature and the choice of an appropriate method for the¹¹⁵ desired goals. Several considerations have to be made prior to the selection of a solvent, including its inertness. The choice of a solvent is extremely important for the achievement of reliable results because it needs to be adequate for the metabolites

targeted for extraction and for the analytical method. The selectivity and polarity of the solvent, its boiling point (in case solvents need to be evaporated or are analysed by direct injection MS by ESI), toxicity and environmental considerations, interference with the analytical procedure and possible contaminants are all criteria that must be evaluated carefully before choosing an appropriate solvent⁹. Among solvent extraction methods, methanol and solvent mixtures that contain methanol are employed the most; furthermore, acidified solutions are common, and chloroform is often used for lipophilic compounds^{39, 72, 100}. Mixtures of organic solvents with water are common because they reduce the extraction of non-polar compounds, such as chlorophyll and fatty acids, which could be problematic due to their intensive adsorption on the reverse stationary phase. However, it should be noted that water must be avoided when GC-MS is used. Chloroform/methanol/water (2:1:1 v/v) is a common mixture for the extraction of non-polar and polar compounds⁷². Yuliana and collaborators¹⁰¹ recently developed a comprehensive extraction method, which is based on continuous extraction with a mixture of solvents of increasing polarity, and it thereby extracts non-polar and polar metabolites. However, this technique is more suitable for NMR studies because mixtures of non-polar and polar compounds cause problems in LC-MS (in the case of a reverse stationary phase) and GC-MS analyses, respectively. In addition, the extraction solvent should be miscible with the mobile phase used in the LC method, and the drying and resuspension of the extract for injection into the analytical system should be avoided. According to Moritz and Johansson⁷, metabolites are most commonly extracted by shaking the previously homogenised plant tissue at high or low temperatures in either a pure organic solvent for non-polar compounds or a mixture of solvents for polar compounds. Several techniques have been applied to accelerate solvent extraction, such as ultrasonic extraction, microwave and pressurised solvent extraction^{72, 39, 102-105}. Microwave extraction has the advantages that it increases the speed of extraction and the dissolution rate of metabolites (in relation to an increase in the internal temperature); however, this method is incompatible with some organic solvents because of the risk of explosion. For this reason, limitations should be verified prior to extraction^{102, 106-107}. Ultrasonic extraction is also rapid and simple and is one of the most commonly used methods for solvent extraction at present. Furthermore, this method does not have any limitations related to the solvent¹⁰⁸. A detailed description of extraction methods used in metabolomics experiments can be found in Villas-Bôas and collaborators¹⁰. Commonly used extraction methods continue to be more suitable for targeted analysis, and an ideal extraction method for metabolomics experiments has yet to be developed¹⁰. Even if all metabolites could be simultaneously extracted, limitations would still be found in the analytical technique because there is no current analytical technique that can identify all metabolites in a single analysis. Because a completely non-compound-specific sample preparation, as required in metabolomics experiments, is not possible, Oldiges and collaborators⁵⁷ suggested focusing on the reproducibility of the sample processing protocol. Preparations of the analysed samples must be identical for accurate comparisons, despite the fact that

some metabolites are excluded⁵⁷. A standard method for optimising the extraction methods used in metabolomics studies was proposed by Gullberg and collaborators¹⁰⁹.

3.1.4 Pre-analytical sample preparation

In some cases, pre-analytical procedures are required to prepare a sample for the analytical method. This preparation may avoid sample carryover in chromatographic systems, concentrate the analytes of interest or remove interfering substances. Nevertheless, this step must be simple and fast, and its necessity must be critically evaluated in relation to the employed extraction method and the selected analytical tool (GC-MS, LC-MS, NMR or others). The data quality, instrument repeatability, metabolite profiling quality and reproducibility are directly affected by the choice and execution of the sample preparation^{39, 72, 100, 110}. Non-polar compounds are generally analysed by GC-MS, but the presence of polar compounds in the extract may cause problems with the chromatographic column and the resolution of the separation. Samples obtained with solvents able to extract both, polar and non-polar compounds, must therefore be submitted to a pre-analytical process to remove polar compounds, or polar compounds must be submitted to prior derivatisation. This is also true for LC-MS analyses when non-polar compounds are present in the extract and a reverse stationary phase is used. This issue is less relevant in NMR analyses; however, signal overlap may occur if interfering substances are extracted to a great extent, such as sugars, fatty acids and others. In this case, sample cleanup is recommended.

Solid Phase Extraction (SPE) is a useful and simple method for further sample treatment. SPE cartridges are commercially available with different adsorbents, such as silica gel (normal phase), C2, C8, C18 and ionic exchange (cationic and anionic)¹¹¹, but new polymeric adsorbents have also been developed, mainly to reduce the limitations of conventional adsorbents¹¹²⁻¹¹⁶. Furthermore, combined or mixed-mode SPE is also used to improve the quality of the results, as in the analyses of phenolic acids from *Echinacea* species¹¹⁷ and quinolizidine alkaloids from *Lupinus albus*¹¹⁸. However, there is little precedent for their use in plant metabolomics.

The principles of SPE are the same as those of LC, and parameter optimisation is fundamental for a reproducible, fast and selective method¹¹³. It is sometimes possible to obtain cleaner samples and higher method robustness with fewer problems in the chromatographic column and mass spectrometer, primarily when direct infusion is employed. Furthermore, the effects related to ionisation suppression may also be improved^{39, 72, 110, 115, 119}. Beyond the elimination or reduction of interfering compounds, the concentration of compounds of interest can also occur, enabling better detection or even visualisation of compounds not detected before¹¹⁶. In a study of *Passiflora garckeii*, previous SPE improved the analysis of aromatic compounds by NMR and reduced the concentration of sugars in the sample¹²⁰. However, a disadvantage of using SPE may be the reduction of metabolite coverage or the loss of metabolites caused by the selectivity of the adsorbent¹¹⁹. The results obtained after SPE should therefore be carefully evaluated.

Liquid-liquid extraction or partitioning is also commonly applied in plant metabolomics. It is a widely used method in classical

phytochemical studies and may be used to remove interfering compounds or to concentrate metabolites of interest^{39, 72}. Examples include the enriched fraction of alkaloids by acidified solutions¹²¹ or the removal of non-polar compounds for the analysis of *O*-glucoside phenolic compounds and their aglycones in an olive oil study¹²². In another case, the *n*-butanol fraction, obtained by liquid-liquid extraction from beer, concentrated and improved the aromatic compound signals in NMR analyses¹²³. However, an analysis of the *n*-butanol fraction would not be possible by electrospray ionisation direct infusion mass spectrometry because of its influence on ionisation, and it would have to be previously dried. Recently, a molecular imprinted polymer prepared from acrylamide, ethylene glycol dimethacrylate and acetonitrile was used to extract gallic acid, *p*-hydroxybenzoic acid, vanillic acid and protocatechuic acid from *Melissa officinalis* extracts¹²⁴. Although this method has advantages, it has been poorly explored for plant metabolomics studies, but it may be a promising tool for metabolite target analysis.

3.2 Metabolite Analysis

A number of analytical methods have been applied in plant metabolomics research. The most common methods are mass spectrometry coupled to gas chromatography or liquid chromatography (GC-MS or LC-MS, respectively), direct injection mass spectrometry (DIMS) and nuclear magnetic resonance (NMR), but other methods, such as capillary electrophoresis mass spectrometry (CE-MS), high-performance liquid chromatography with photodiode array detection (HPLC-PDA) and thin layer chromatography with UV detection (TLC-UV), have been described^{7, 43}. An appropriate combination of analytical tools with respect to the analysed plant material is essential because each method has its own advantages and disadvantages⁷. Because one method can complement another, Bino and collaborators⁴¹ stated that a combination of LC, NMR, and MS systems will probably be used in the future as the number of quantifiable and identifiable metabolites can be increased. However, because of high costs, the technique would be restricted to only a few laboratories.

As mentioned before, some separation-based techniques coupled to mass spectrometry can be applied to metabolomics studies. Gas chromatography coupled to mass spectrometry by electron ionisation (EI) is an older and more traditional technique that has been used for a longer time and is already discussed in another article within this special issue. For this reason, it will not be discussed in more detail at this time. Other techniques, such as LC or capillary electrophoresis, can be coupled to mass spectrometry by atmospheric pressure ionisation. Without a doubt, the ionisation source that is most frequently used in metabolomics studies is electrospray ionisation (ESI). Despite its widespread use, it is still a relatively young technique, and for this reason, fundamental studies about reactions in the ionisation source, as well as fragmentation patterns (MS/MS or MSⁿ), are still being investigated.

In the following section, the advantages and disadvantages of the ESI source, reactions, analyser and fragmentation reactions by MS/MS or MSⁿ and the pros and cons for the use of hyphenated methods (GC/LC-MS) and DIMS are briefly described.

3.2.1 ESI source reactions: Ion generation and ion suppression

The history of ESI source development, with plenty of details, is found in a review written by Dr. John Fenn, who was awarded the Nobel Prize¹²⁵. The process can basically be summarised as follows. In an initial step, the solution containing the analyte crosses a metallic capillary, where an electric potential is applied (between 2 to 4.5 kV). This potential promotes a migration of charges to the interface of the capillary with the solution, which results in an electric double layer, and thus, droplets with charged surfaces are formed. The process of solvent evaporation starts within a chamber with a slight reduction of pressure under a nebuliser gas. With the reduction of droplet sizes, an approximation of the charges occurs, and consequently, the electrostatic repulsive forces are increased, which leads to a decrease in the surface tension of the droplets until rupture, a process denoted as a "Coulomb explosion". This rupture results in the liberation of ions into the gas phase¹²⁶⁻¹²⁸ by a spray of charged particles^{Error! Reference source not found.-131}, or in other words, an electrolytic current.

It is believed that ionisation occurs through acid-base reactions and/or coordination with metallic ions, principally Na⁺ and K⁺ in positive ion mode, in addition to other adducts, such as the addition of water or ammonium. Positive ionisation mode is used in a large majority of metabolomics studies, and the addition of H⁺ and Na⁺ ions must be balanced. This coordination depends on the number and spatial disposition of atoms with lone pairs and can also result in a total absence of protonation reactions, as can occur in some natural polyethers, which can show an almost cyclic conformation in which sodium (or another element of the family 1A) coordinates with the oxygens of tetrahydrofuran rings¹³²⁻¹³⁴. In these cases, the abiotic effects of the environment, for example potassium rich soils or soils with divalent metal cations, can lead to a higher coordination of this element. This coordination leads to a dislocation of the mass to charge ratio to values higher than expected¹³⁵⁻¹³⁶, which can in turn lead to incorrect data interpretation. Aside from this difficulty, little on this subject is discussed in literature, and a better known phenomenon is ion suppression. In metabolomics studies using the hyphenated system LC-ESI-MS, the co-elution of substances with a large difference in the ionisation potential can lead to the complete suppression of an ion, leading to an incorrect result indicating the non-appearance of a metabolite in question. In a recent review, Mol and collaborators¹³⁷ described this problem in detail, including diverse experimental procedures by full-scan-high-resolution analyses to evaluate and quantify this effect.

Without a doubt, the least discussed effect of source reactions in ESI is the possibility of forming cation radicals through the loss of one electron in positive ion mode and the formation of anion radicals by the capture of an electron in negative ion mode¹²⁸. Van Berkel and collaborators parameterised the function of the ESI source and proved the formation of oxidised and reduced species^{126-127, 138-139}. In a revision on the subject, the occurrence of such reactions for a series of low molecular weight compounds, such as isoquinolinic alkaloids, carotenoids, and quinones, among others¹⁴⁰, was described, suggesting that this phenomenon could

be much more amplified than expected. Thus, this process can lead to incorrect result interpretations because the method leads to the formation of molecular ions ($[M]^{+}$) instead of protonated ($[M+H]^{+}$) or cationised molecules ($[M+Na]^{+}$), and consequently, an ion with an anticipated odd mass appears as even and vice versa. A systematic study of carotenoids, xanthophylls and retinoids enabled the rationalisation of the phenomenon behind these polyenic structures¹⁴¹⁻¹⁴⁴. As can be observed in Figure 2, carotenoids, which possess a polyenic system and an absence of heteroatoms in their structure, exclusively form cation radicals in the capillary energy range. In total, the reactivity of carotenoids is increased so that it fragments rapidly at the source, and to observe the molecular ion, it is necessary to work with a capillary energy less than 2 kV. In contrast, xanthophylls, which possess the same polyenic skeleton with oxygen atoms in the final portions, can be ionised in a normal energy range of the ESI source because redox processes form $[M]^{+}$ ions as well as acid-base reactions, leading to the formation of $[M+H]^{+}$ ions. Finally, retinoids possess a shorter chain and heteroatoms, and in this case, the oxidation reaction is considerably decreased. The formation of molecular ions in positive ion mode through redox reactions motivated some authors to look for the formation of radical anions following the same logic. In this context, lapachol and other quinonic derivatives were investigated, and based on experimental data and theoretical calculations, the formation of cationic and anionic molecular ions was supported¹⁴⁵⁻¹⁴⁸. This phenomenon can serve as a further tool for structural determination; however, it can also lead to the falsification of metabolomics studies. In addition to the problems already discussed (such as the inversion of odd and even ions), it was also shown that this anionic species is much more reactive and can lead to additional reactions of the solvent, resulting in an anion molecule reaction with a chain extension through the formation of carbon-carbon bonds. For this reason, and despite its great importance, ESI is still inspires fundamental studies in order to turn metabolomics studies more robust. Furthermore, it is also known that the use of hyphenated systems, as well as the use of direct infusion mass spectrometry, still presents several limitations, such as ion suppression, radical species formation and cross reactions.

3.2.2 MALDI source: Ion generation and matrix effects

Out of a large variety of promising ionisation sources for metabolomics studies, MALDI (matrix-assisted laser desorption/ionisation) deserves special consideration. In the early 1960s, the technique of ionising organic molecules by irradiation with a high intensity laser pulse was developed. This process was successful in forming ions with different structures. Initially, this technique was named laser desorption/ionisation (LDI). Continuous research and development allowed its application for the volatilisation of non-volatile biopolymers and organic macromolecules¹⁴⁹. In 1987, Michael Karas and Franz Hillenkamp¹⁵⁰ successfully demonstrated an organic matrix for the analysis of high molecular weight compounds in LDI studies. The matrix had to show a strong laser light absorbance at the applied wavelength and high sublimation¹⁵¹⁻¹⁵³. In this historical

demonstration, the analyte was incorporated into the matrix at a low concentration, and it was immediately exposed to a high pulsed laser light. Volatile ions were produced at each laser pulse with high reproducibility. This demonstration formed the basis for the development of MALDI (matrix-assisted laser desorption/ionisation). Dr. Koichi Tanaka¹⁵⁴ demonstrated the application of MALDI for a set of biological macromolecules, for which he was awarded the Nobel Prize in Chemistry in 2002¹⁵⁵.

From this point on, MALDI was introduced to the market with great success, principally in proteomic research, and it experienced rapid growth, which, in the beginning, was only exceeded by the growth of mass spectrometry by ESI^{128, 156-158}. MALDI can be explained in the following three basic steps: (i) *formation of a "solid solution"*, in which analytes are completely isolated from each other through dilution in a homogeneous matrix for subsequent desorption; (ii) *excitation of the matrix*, in this step, a laser beam is applied to the matrix surface. Photons from the laser are absorbed by the chromophore parts of the matrix molecules, which causes rapid vibrational excitation followed by disintegration. Clusters of analyte molecules surrounded by the matrix and salt ions are ejected at the solution surface, producing the plume and subsequent evaporation of the matrix molecules to expose free analyte molecules to the gas phase; and (iii) *ionisation of the analyte*, in the desorbed matrix, cloud ionisation (plume) occurs via proton or cation transference from the photo-activated matrix molecules, which leads to the formation of characteristic ions $[M+X]^{+}$ ($X = H, Na, K, \text{etc.}$)¹⁵⁶⁻¹⁵⁸. In these steps, the process exhibits limitations for the analysis of organic micromolecules, such as natural and synthetic products, and pharmaceutical substances, among others. This limitation was caused by the matrix, which is cationised or anionised after desorption and extracted into the analyser. For this reason, the mass range between 0 and 500 m/z may have shown numerous matrix signals, complicating any analysis of low molecular weight compounds. Due to advances in laser beams, in addition to the possibility of direct ionisation and the use of ionic liquids¹⁵⁹⁻¹⁶⁰, grapheme flakes¹⁶¹ and others as matrices¹⁶²⁻¹⁶³, this problem could be minimised. Thus, an increase in spectra acquired from low molecular weight compounds by this technique has been observed in recent years, which makes it a very promising area of study.

The possibility of analysing micromolecules by MALDI-MS facilitated the understanding and stimulated comparisons of different fragmentation mechanisms. Despite the fact that the collisional activation process still occurs similar to that observed in ESI (discussed in more details in Section 3.2.3), the internal energy of the ion is not the same during the process of ionisation¹⁶⁴. However, the same cation radical formation effect of carotene in ESI was also observed with MALDI ionisation¹⁴¹. As cited before, MALDI-TOF-MS was extensively applied for proteomics studies and for biological markers with a peptide origin. Despite the fact that lipids were rarely investigated by MALDI-TOF-MS, it was recently suggested that these structures are clinically and biologically as relevant as proteins for the analysis of cellular samples, which are usually present in much lower concentrations. From 2005 on, the use of MALDI in combination with thin layer chromatography (TLC-MALDI) for lipid analysis has come into use¹⁶⁵. This progress in the structural

determination was also supported by continuous advances in the understanding of some fragmentation mechanisms in the gas phase, leading to the popularisation of this combined technique. Pioneering studies on the lipid composition of sheep

mesenchymal stem cells compared to lipids extracted from human erythrocytes that applied TLC-MALDI supplied essential contributions to stem cell research¹⁶⁵. In another study employing this technique, the characterisation of various pork brain lipids, heterogeneous lipopolysaccharides (endotoxins, LPS) of gram-negative bacteria and glycosphingolipids was also possible¹⁶⁶⁻¹⁶⁸.

In the field of metabolomics, studies realised by Dr. Dorrestein and collaborators allowed a large advance in the study of the characterisation of metabolite exchanges between microorganisms in the cellular communication system (characterising metabolic exchange factors from microbial systems involved in cell-to-cell communication)¹⁶⁹. In his works, the intensive use of mass spectrometric techniques can be observed, such as MALDI imaging, which was popularised with the proteomic studies of biomarkers relating to various diseases¹⁷⁰. In the present issue, three other articles describe this subject in detail, but in summary, the principle of this technique is based on the spatial acquisition capacity by MALDI with good resolution and the correlation of the ion intensity with a colour scale. Thus, the tissue is previously photographed, and the intensity data of the ions are superimposed with the bombarded area, which indicates the quantity of a determined analyte in a specific area. The success of this procedure stimulated its use in plant tissue analysis¹⁷¹. An interesting point is that this transition was very rapid, without having a previous broad distribution of the analysis of plant extracts by MALDI-MS. In relation to metabolomics, this technique only began to appear in 2007. Li and collaborators¹⁷² evaluated the occurrence of a series of biomolecules by using the direct excitation of banana, grape and other fruits. One interesting observation was the high occurrence of clusters with potassium, which (as discussed above) can imply false data interpretation. In another study, Fraser and collaborators¹⁷³ carried out a systematic investigation of the use of MALDI in metabolomics studies of carotenoids, including the analysis of crude extracts. Recently, our research group dedicated a portion of the ongoing research to the advantages and disadvantages of using MALDI for the analysis of crude plant extracts with the aim of creating metabolic fingerprints for chemotaxonomy. Initial data showed that MALDI may generate a large quantity of data (ions) that must be carefully organised. This investigation started with the analysis of polar extracts from *Lychnophora* species, which showed serious problems with correct botanical classification and a high incidence of flavonoids and terpenoids in their secondary metabolism¹⁷⁴. Figures 3 and 4 show the analysis spectra for crude extracts, and it is possible to observe a collection of ions with or without the addition of any matrix substance. Despite the success in generating data, this detailed analysis also showed the significant impact of applied parameters, such as the laser intensity, frequency, focus, and number of shots applied on one raster spot between many others resulting in a significant change of the observed ions, which demonstrates that direct ionisation by MALDI-MS also has its limitations and must be further investigated (unpublished data). Therefore, this technique may be employed to analyse the

secondary metabolites in extracts, tissues and single plant cells⁵³⁻⁵⁵. Error! Reference source not found., demonstrating its potential in metabolomics studies. Despite its disadvantages, MALDI is a promising technique for metabolomics studies because of its high sensitivity, high analysis scan speed (≈ 90 seconds) and lower contaminant influence on ionisation⁵⁵.

3.2.3 The use of fragmentation spectra (MS/MS and MSⁿ) in metabolomics

Another area that is growing within metabolomics research is the use of fragmentation spectra to aid structural identification. A series of databases has been established, and with each day, the information increases. As discussed before, liquid chromatography hyphenated methods usually use ESI for ionisation, in addition to APCI and others. In the studies employing direct injection, they normally employed ionisation sources that involve a soft ionisation process, which usually does not form fragments³⁹. An additional step is necessary for MS/MS analyses, namely, the isolation of an ion, which has an increased internal energy after receiving some type of activation that allows the beginning of the fragmentation process. Collision induced dissociation (CID) is without a doubt the most frequently used process¹⁷⁵⁻¹⁷⁶. The use of equipment such as the Orbitrap has also distributed other similar processes for use in metabolomics studies (for major details, we recommend the review article published by Perry and collaborators¹⁷⁷). In the pioneering article by Perry¹⁷⁷, the author describes the dissociation technique induced by collision in detail, which is based on the activation of the ion by collision with an inert gas (generally, Ar, CO₂, N₂). Kinetic energy transference occurs through collision with the gas, and the ion is transformed by internal energy, inducing the fragmentation. The mechanism of this fragmentation can be elucidated with a certain precision, and a sequence of fragmentation steps originates similar fragmentation ions for distinct substituent groups¹⁷⁸. However, the difficulty of elucidating the mechanism is not only related to discovering the classical neutral eliminations but also to characterising the dynamic transformations, which occur during the diffusion of the species in the direction of the analyser (formation of metastable ions, rearrangements, etc.)¹⁷⁹. In this sense, Vessecchi and co-workers showed that computational methods (to describe the reaction site and energetic profile) in combination with experimental parameters (energy resolved plot, solvent system and labelled experiments) can be useful for understanding the fragmentation mechanisms involved in mass spectrometry analysis to obtain more comprehensive data^{146-148, 180-181}. These studies clearly show that the quantity of energy provided can lead to different spectra, and thus, the lack of homogeneity in the quantity of the transferred energy (similar values) and the lack of homogeneity between the used gases can be a barrier for a more general comparative metabolomics analysis with data obtained from different laboratories and with different equipment.

The application of this associated technique using homologous substances has demonstrated that different activations of natural products analysed by different analysers can lead to similar or completely different spectra. Recently, our group compared the fragmentation spectra of two sesquiterpene lactone stereoisomers

of the furanoheliangolide type (budlein A and centratherin), which have β - and α -orientations, respectively, for the side chain at C-8, as obtained by QqQ, QTOF and Orbitrap-HCD analysers. All analyses were supported by computational methods, and it was possible to suggest the kinetic control of the two different fragmentation pathways. A significant difference between the spectra acquired on QqQ and Orbitrap-HCD and the spectra acquired with TOF was clearly demonstrated¹⁸². These data were important in assisting structure elucidation, but they can also be an error source in the analysis of sesquiterpene lactones by MS for inexperienced researchers in the spectrometry area. Another important concern is related to acquiring the first critical energy to yield fragments. To induce this first fragmentation with the desired ion energy, the energy can be dispersed, and when the barrier is overcome, it can be too high for subsequent reactions and, in this case, other fragmentations do not occur. This phenomenon can be clarified when analysing the fragmentation of flavonoids, for example, pinobanksin and 3-O-acetylpinobanksin. As expected, pinobanksin provides the whole series of ions used for the structural determination of flavonoids by MS/MS analysis. In contrast, the acetylation of the oxygen in position 3 leads to an initial neutral elimination of the acetate portion, and the resulting ion no longer possesses internal energy for further fragmentation, which leads to a totally different profile from that expected for non-glycosylated flavonoids¹⁸³. These examples show that the use of MS/MS or MSⁿ is extremely important for metabolomics studies, but for the interpretation of data, a detailed observation of the ions that show abnormal behaviour is necessary because the reason might merely be the effects of the internal energy quantity of the analysed ion.

3.2.4 Hyphenated mass spectrometry methods: GC/LC-MS

Hyphenated mass spectrometry methods are so-called separation-based methods coupled with mass spectrometry. Prior to MS analysis, metabolites are separated by a separation-based method, namely, gas chromatography (GC), liquid chromatography (LC) or capillary electrophoresis (CE). In GC and LC, compounds are separated by exploiting different interactions of the substances with the stationary phase. However, in CE, the separation is based on the size-to-charge ratio of the ionic molecules⁴³. CE-MS will not be further discussed here because it is beyond the scope of this review. Reviews on the subject can be found in Soga¹⁸⁴, Yan and collaborators¹⁸⁵, and Ramautar and collaborators¹⁸⁶. The choice of separation technique is made according to the type of molecules present in the target sample. GC is suitable for hydrophobic, low molecular weight compounds, such as essential oils, hydrocarbons, esters, and metabolite derivatives with reduced polarity⁴³. Furthermore, these compounds must be heat-stable and volatile. Previous derivatisation can be performed in case the compounds are not sufficiently volatile^{7, 43}. Different ionisation methods can be applied depending on the type of molecules targeted for analysis. The ionisation method that is most commonly used in combination with GC is electron ionisation (EI). EI is a hard ionisation method, where an excess amount of energy is transferred, causing strong molecular fragmentation⁴. For LC-MS, soft ionisation technologies are more commonly used. Soft ionisation technologies transfer less energy, and thus, less molecular fragmentation is induced^{4, 9}. For this

reason, it is necessary to induce fragmentation through, for example, the use of CID to produce fragments. Atmospheric pressure chemical ionisation MS coupled to LC (APCI-LC-MS) is applied in the analyses of non-polar metabolites, such as sterols, fatty acids, phospholipids, steroids, triterpenes and others¹¹⁰. However, as previously mentioned, electrospray ionisation LC-MS (LC-ESI-MS) is commonly applied to plant metabolomics studies, where it is used to analyse a wide array of metabolites, from sugar to fatty acids^{4, 9, 43, 110}. The separation-based MS methods are highly sensitive (detection limit at 10⁻¹²-10⁻¹⁵ moles) and the costs are relatively low compared to, for example, NMR methods^{9, 43, 49}. In addition, this method has good availability, relative quantitation is possible, the chemical information is increased with data obtained from the chemical nature (by the separation technique) and mass (by MS) of the compounds, and combinations with other detectors are also possible, such as a diode array detector (DAD) that provides information about the UV spectra^{9, 43, 110}. However, the separation-based MS methods require significant time^{43, 110}, and the absolute quantitation of metabolites is not possible because calibration curves for each individual compound would be required, which is unrealistic, and many compounds have not been identified⁹. In Table 2, the main advantages and disadvantages of hyphenated mass spectrometry methods (GC-MS and LC-MS) in plant metabolomics research are briefly described.

The high time consumption in separation-based MS methods is one relevant problem, as previously mentioned. In this context, new stationary phases and modifications to the instruments were developed to reduce this time. For GC-MS, new columns, such as narrow bore column, were developed to reduce the analysis times to approximately 1 and 10 min, namely, the Ultra Fast-GC and Fast-GC, respectively¹⁸⁷⁻¹⁸⁸. In addition, comprehensive two-dimensional gas chromatography (GC x GC) has been employed to improve the peak resolution, separation capacity and reproducibility of complex mixtures¹⁸⁹. In LC-MS, monolithic columns and fused core particle (2.7 μ m) packing columns have been used to reduce the time and increase the resolution¹⁹⁰⁻¹⁹². The size particle sub-2 μ m packing columns are highly efficient for this purpose, and they have high sensitivity and reproducibility. However, they operate at very high pressures, so ultra-high performance liquid chromatography (UHPLC) systems must be used¹⁹³. The stationary phase of C18 (reverse phase) is most frequently used in plant metabolomics, despite new phases such as HILIC^{191, 194}.

Table 2: Advantages and disadvantages of hyphenated mass spectrometry methods

	Advantages	Disadvantages
GC-MS 9, 43	Suitable for analysis of low molecular weight hydrophobic compounds (essential oils, hydrocarbons, esters). Volatile compounds can be directly analysed. Databases available*, due to the reproducibility of mass spectra (fragmentation) Suitable for analysis of mixtures	Not suitable for non-volatile and thermo-unstable compounds. Requires derivatisation depending on the type of molecules that are analysed. Derivatisation can mask the result.
LC-MS 43, 46, 110, 173	Suitable for analysis of relatively polar compounds with low, moderate or high molecular weights, depending on which ionisation method is applied. Derivatisation not required. Thermo-unstable compounds can be analysed.	Fragmentation patterns are poorly reproducible, so databases are therefore problematic. Not all compounds in a plant extract will ionise under the same conditions; the technique is therefore more suited for targeted analysis. Different types of adducts can be produced, depending the nature of the compounds.

*Golm Metabolome Database, <http://csbdb.mpimp-golm.mpg.de> and Fiehn Laboratory Database, <http://fiehnlab.ucdavis.edu/>

3.2.5 Direct injection MS (DIMS)

Direct injection MS (DIMS) is an analysis performed by directly injecting a sample into the ionisation source. Atmospheric pressure ionisation techniques are usually applied, with ESI being most frequently used⁵⁶. Soft ionisation techniques are preferably applied in metabolomics studies because they cause less fragmentation than hard ionisation techniques, resulting in a smaller number of signals and, consequently, less complex spectra⁴³. The ions are generated in the ionisation source, and they are subsequently separated in the analyser for detection, so a mass spectrum is produced. There is a large variety of mass analysers applied in metabolomics studies by DIMS and LC-MS, such as single-stage quadrupole (Q), triple quadrupole (QqQ), quadrupole ion trap (QIT), Orbitrap, time of flight (TOF) and Fourier transform ion cyclotron mass spectrometers (FT-ICR-MS)^{39, 56}, and a combination of some of these analysers can be found in hybrid spectrometers, such as Q-TOF. The analysers can be of two types, namely, low or high resolution, and the latter enables high mass accuracy. Despite the high cost of the high resolution equipment, it is possible to obtain high repeatability, peak annotation and more precise data alignment. To distinguish between isobars (compounds with the same nominal mass), Dettmer and collaborators⁵⁶ and Wolfender and collaborators³⁹ recommended the use of high-resolution mass spectrometers such as TOF and FT-ICR in combination with DIMS. The quadrupoles, which are frequently used in metabolomics studies, have been applied in GC-MS and LC-MS¹¹⁰. They are composed of four parallel metal rods that are submitted to direct current voltage and radiofrequency in opposite pairs, and they alternate. The m/z range is limited, in addition to having low

resolution. They can be used alone in a single quadrupole (Q), in combination with three quadrupoles (QqQ) or in hybrid systems. Quadrupoles have several advantages, such as good speed and sensitivity, low cost, a relatively good dynamic range, and good quantification capability, and they are well suited to continuous ion sources such as ESI^{56, 110, 195}. In QqQ, the quadrupoles are arranged in a linear sequence in which the second acts as a collision cell, and it allows the acquisition of MS/MS data. QqQ can be operated using different methods, such as full scan MS, single ion monitoring (SIM), neutral loss scan, product scan (daughter scan) and selected reaction monitoring (SRM)¹⁹⁵. In the product scan, the fragmentation is yielded, and all the fragment ions are scanned. Thus, it is commonly used for the identification of compounds in qualitative analyses, but the neutral loss may also be applied to compounds of interest to identify the ions that have lost neutral moieties⁵⁶. The other highlighted acquisition mode is SRM, which monitors a specific precursor ion and its product ion or different product ions in the multiple reaction monitor (MRM), improving the sensitivity and confidence of the target identity. This acquisition mode is widely used in quantification studies^{56, 196-197}. The cylindrical quadrupole ion trap (QIT) is based on the same principle of Q, but with a different geometry. Another type is the linear quadrupole ion trap, which has a linear and non-cylindrical geometry. Ions may be isolated and maintained in the trap for the fragmentation, even ions of low abundance, after the ions are ejected and destroyed in the detector to generate a mass spectrum. It allows the acquisition of MSⁿ (consecutive steps) representing more information for structural identification, but it is not good for quantification. In addition to the other limitations described for QIT, it also has low resolution, low scan speed and poor applications in UHPLC^{110, 195, 198}. The importance of MSⁿ data in the structural identification of compounds can be exemplified in isoflavones, in which QIT was employed to confirm the relationship between precursor and product ions¹⁹⁹. It was also employed in direct injection as a fast method to obtain MS and MSⁿ data for the compounds from plant extracts²⁰⁰⁻²⁰¹. Recently, other trap-based mass spectrometers were developed for FT-ICR-MS and Orbitrap, but with high resolution and mass precision. FT-ICR-MS in particular has a very high resolution and mass precision (1-2 ppm), and it is also cited as one method of choice in some reviews on plant metabolomics^{4, 38-39, 43}. This method employs a strong magnetic field in which the ions are trapped, where they show circular motion and the frequency depends on their m/z . The calibration of these machines is very stable, and the detection is non-destructive¹⁹⁵. However, the acquisition time is too long, and thus, it is better for DIMS analyses than for LC-MS; the dynamic range is medium, and the typical m/z range of work is 30 to 10000, which is not a problem in plant metabolomics²⁰². FT-ICR-MS was first applied by Aharoni and collaborators²⁰³ in 2002; however, since then, its application has been limited, mainly because of high instrument costs^{43, 56}. In contrast, the Orbitrap, which is operated using an electrostatic field, has a higher scan speed than the FT-ICR-MS¹⁹⁵ and is more applicable to HPLC and UHPLC, making it a method of choice in metabolomics studies by DIMS and LC²⁰⁴⁻²⁰⁵. It is also a non-destructive detection technique that shows high resolution, high mass accuracy (1-2 ppm) and good dynamic range. In addition, it

is possible to obtain MSⁿ data with high accuracy that aids the structural identification of metabolites^{177, 195, 206}. TOF is another widely employed analyser in metabolomics^{39, 110}. Ions are separated according their arrival time in the detector, and there are linear and/or reflectron instruments. In the first, the resolution is lower than in the reflectron, so the second is the most frequently used. A variety of reflectron instruments can be found, such as a reflectron in V or W mode that has an increased flight path, so it has the increased mass resolving power of a TOF analyser^{195, 207-208}. The resolving power is mass dependent and is also related to the length of the flight path. However, TOF equipment allows only MS analyses, and hybrid systems can employ MS/MS analyses, such as the Q-TOF that is widely used¹¹⁰. In addition, its scan speed is high, so it is versatile and applicable to fast chromatographic methods, and it has high levels of resolution and mass accuracy²⁰⁷⁻²⁰⁸. The main disadvantages of DIMS are the impossibility of distinguishing between chemical isomers and the lower quantity of chemical information that is obtained relative to hyphenated MS methods⁴³. A possible way to circumvent the chemical isomer problem is to perform MSⁿ analyses⁴³. By performing MSⁿ analyses, previous chromatographic separations can be avoided if information obtained by the fragmentation patterns can be used to differentiate the isomers²⁰⁹. Furthermore, chemical information can also be increased by performing analyses in negative as well as in positive ion modes³⁹. The type of mass analyser and the ionisation method have a major influence on the resulting mass accuracy, resolution, and detection limit, and advantages and disadvantages depend highly upon the chosen instrument type.

4 Preprocessing and Statistical Analysis of Plant Metabolomics Data

4.1 Introduction

Metabolomics experiments produce an enormous amount of data. Statistical and computational methods have been applied to evaluate these large data sets^{47, 210}. Data acquisition often occurs at such a high speed that it is no longer seen as the limiting factor; instead, data assessment is considered to be the time limiting factor⁹⁶. Mathematical and statistical skills are therefore an essential prerequisite for extracting as much information as possible from metabolomics data^{35, 211}. Data generated by high-throughput metabolomics experiments require their own sophisticated computational and statistical tools, which can be routinely applied in metabolomics experiments³⁶. These standard methods are developed by adapting analytical methods from other *-omics* sciences, which have been developed in the fields of chemometrics, chemoinformatics, and bioinformatics^{36, 211}. Some sources refer to the data processing tools employed in metabolomics as chemometric methods^{34, 211, 218}, and others refer to them as methods combined from bioinformatics and chemoinformatics by arguing that metabolomics is not only about the identification and quantification of metabolites but also about putting the data into context within the biology³⁶. All three terms are overlapping, and brief definitions are given in Table 3.

Table 3: Definitions of chemometrics, chemoinformatics and bioinformatics

Term	Explanation
Chemometrics ²¹²⁻²¹⁵	The term chemometrics was coined by Svante Wold ²¹⁴⁻²¹⁵ and appeared in the early 1970s. It is defined as follows: "Chemical discipline that uses mathematics, statistics and formal logic (a) to design or select optimal experimental procedures; (b) to provide maximum relevant chemical information by chemical data; and (c) to obtain knowledge about chemical systems." ²¹² The chemical structure of a distinct compound is not of interest, in contrast with chemoinformatics. However, applied statistical methods (machine learning) are overlapping in both disciplines ²¹³ .
Chemoinformatics ^{213, 216}	The term chemoinformatics first appeared in 1998 ²¹⁶ and is defined as the application of informatics methods to solve chemical problems ²¹⁶ . It includes chemical structure modelling as well as the application of statistical methods (machine learning) ²¹³ . More detailed descriptions can be found in Varnek ²¹³ .
Bioinformatics ²¹⁷	The term bioinformatics was introduced by Hesper and Hogeweg in 1970 ²¹⁷ . It is defined as the study of informatic processes in biotic systems, and it provides a systematic approach to statistical and computational tools for the treatment of experimental data from biological samples.

The field of computational metabolomics is still evolving, and it is thought that developments in this field will be a key factor in making metabolomics a standard analytical procedure in different areas of research³⁶. Wishart³⁶ states that computational metabolomics will not only have a major impact on data analysis but also on (i) metabolomics databases, (ii) metabolomics laboratory information management systems (LIMS) and data standards, (iii) spectral analysis tools for metabolomics and (iv) metabolic modelling.

It is suggested that whenever a metabolomics experiment is carried out, familiarisation with those aspects of computational informatics should be sought. The storage of raw data plays a particularly important role, as will be discussed in Section 4.3, but it is also essential because metabolomics is still a very young developing research field and better data processing methods might be developed in the future²¹⁰.

Since several extensive reviews have been written on metabolomics data analysis and preprocessing (*e. g.* ^{34, 36, 219-222}), including scripts for the statistical program Matlab in a ready-to-use format on the webpage <http://bioinformatics.mpimp-golm.mpg.de/> in Steuer and collaborators²¹⁹, we merely provide some working examples of our routine techniques written in R on <http://labpib.fmrp.usp.br/~rsilva/metabolomics/review.html>. In the following data, preprocessing is briefly introduced, and the advances of different aspects of statistical data analysis in plant metabolomics that have enabled metabolic network modelling are discussed.

4.2 Preprocessing Mass Spectrometric Metabolomics Data

Variance and bias is introduced during the analysis of samples by

mass spectrometric methods, which cannot be attributed to real differences. Data preprocessing includes all steps that try to eliminate this variance^{218, 223} by aiming to reduce the complexity and extract the most important features of the raw data²²⁴. Data preprocessing is an essential step in metabolomics data analysis, and as described by Van den Berg and coworkers²²⁵, it can greatly influence data interpretation.

According to Castillo and collaborators²²⁴, preprocessing includes noise filtering, normalisation, peak detection, deisotoping, alignment and identification. Noise filtering is designed to separate compound signals from background signals originating from chemical or instrumental interference. Normalisation is applied to correct the systematic variation and for direct comparison of different samples. Peak detection accounts for the correct identification of a compound signal and its intensity, and deisotoping algorithms group isotopic peaks that originate from the same compound. Isotopes would otherwise be considered different compounds during data analysis and would consequently falsify data interpretation. In different samples, slightly different m/z values may be obtained for the same compound. To compare different samples, alignment must be performed, and depending on whether individual metabolites are identified or not, finding out the chemical identity of a compound may also be necessary²²⁴.

Normalisation, as well as alignment, is possible with or without the addition of an internal standard²²⁴⁻²²³. A frequent topic of discussion in metabolomics research is whether to add internal standards or not. Smith and collaborators²²³ do not recommend the addition of internal standards for the alignment of metabolomics data. They name various drawbacks, such as the wrong assumption that deviations in retention time are linear, the additional time-consuming step in sample preparation, and the fact that internal standards may mask other experimentally relevant analytes. Instead, the authors suggest the use of an algorithm implemented in XCMS^{223, 232}, a free software program for LC-MS preprocessing. This algorithm considers the nonlinearity of retention time deviation without the addition of any internal standards.

Data preprocessing software is usually proprietary or depends on data produced by a vendor's specific instrument. To favour standardisation, reproducibility, and exchangeability of metabolomics data, efforts have been made towards the development of freely accessible, open-source software^{223-224, 226}.

The advantages of such open-source software programs are that they may be easily customised if some basic programming skills are available with no need to develop new algorithms. Furthermore, the exchange of such algorithms and data within the community is facilitated²²³⁻²²⁴. Several open-source software programs are in use, and some of them have been developed for metabolomics research, whereas others have been adapted from software used in proteomics²²⁴.

Reviews on data preprocessing, as well as commonly applied open-source and vendor-specific software for mass spectrometry-based metabolomics (focused especially on LC-MS), can be found in works by Castillo and coworkers²²⁴ and Katajamaa and Orešić²²⁷, for example. When open-source software is chosen, raw data usually must be converted to an open format. Different tools are available and described in Castillo and collaborators²²³.

The metabolomics community is continuously growing, and new and exciting analytical strategies are steadily being developed to increase the amount of information extracted from mass spectrometric data sets. Such strategies include the development of software for the facilitation of data exchange²²⁸ as well as the development of databases, as discussed in Section 4.3. The more robust GC-MS experiments can be reproduced in different laboratories, and spectral databases have been successfully employed in routine analysis²²⁹. Popular free software, such as TagFinder²³⁰ and TargetSearch²³¹, are available for spectral preprocessing, which results in the typical data table of peaks and corresponding intensities that is used as input for multivariate data analysis. The preprocessing of LC-MS data is an area of constant research. A large part of the research community is working with XCMS^{223, 232} and MZmine²³³, but new software programs are being steadily developed²³⁴ to increase the quality and efficiency during preprocessing. For example, a package for preprocessing mass spectrometric data in R has recently been developed²³⁵.

Detailed tutorials are available for both GC-MS²³⁶ and LC-MS²³⁷. A major bottleneck in LC-MS data is its reproducibility. Several chromatographic parameters influence the retention time, and a similar approach to retention index calculation, as used for GC-MS, has not yet been established for LC-MS²³⁸⁻²³⁹. The most commonly used ionisation type in LC-MS is ESI, which produces a complex mixture of ions and makes reproducibility, as well as interpretation, more problematic²⁴⁰. Many methods have been developed to overcome these difficulties, such as the use of the internal data structure to find adducts and isotopic patterns and either the use of this information as a complexity filter or of additional information for modelling²⁴¹⁻²⁴³. Furthermore, the problem of retention time deviation²⁴⁴ has also been addressed.

As discussed in Section 4.3, the use of MSⁿ in compound identification has been hindered by the lack of databases in which spectra from LC-MS experiments²⁴⁵ are stored.

4.3 Metabolic Network Modelling

As previously discussed, metabolomics has a wide variety of applications, and each of these applications demands a different set of statistical analysis tools to extract the necessary information. There is generally no single technique that is able to extract all information contained in a large multidimensional dataset generated by metabolomics experiments, and instead, a combination of different tools is required. However, one of the main goals of metabolomics experiments is to position the analysed metabolites in an organism-centred biochemical network to investigate their influence on a specific condition of a cell or organism. In the following section, we introduce advances in different aspects of statistical analysis that have enabled metabolic network modelling.

The main challenges of metabolic network modelling in metabolomics are as follows: 1) the necessity of different analytical platforms to analyse a large quantity of compounds, 2) the lack of reference databases, and 3) the lack of integration methods (methods that can integrate data from different sources). In the following section, recent initiatives for the establishment of central databases, new integration methods, and procedures to formalise the storage and exchange of the growing information

collected from specific organisms are introduced.

Metabolism is driven by an organism-specific catalytic potential, meaning the capacity a distinct organism has to produce certain metabolites. This catalytic potential is determined by specific enzymes that are present in the organism. Defining a set of reactions that an organism can perform, including its reactants and products, is therefore the first step in defining an organism's metabolome. To achieve this goal, a so-called organism specific Genome-scale Metabolic (GEM) Model is constructed. These GEMs use information from genome databases and other information of any kind found in the literature. This model involves genome annotations, the identification of associated reactions and their stoichiometry, localisation assignments, biomass composition determinations, energy requirement estimations and reaction directionality²⁴⁵⁻²⁴⁶. GEMs of individual organisms are available in a standard format called Systems Biology Markup Language (SBML). Although GEMs are more common for microorganisms, primarily because of their simpler cellular organisation, plant cellular compartmentalisation and specialisation have also been addressed by related models²⁴⁷⁻²⁴⁹. An increasing number of plant GEMs has been proposed, which shows a capacity to predict the observed phenotypes²⁵⁰. Populating pathway databases would enable even faster production of different plant GEMs.

In traditional methods for constructing GEMs, the first step is to use a function inference, in which enzyme functions are inferred by comparing the sequence of a gene that codifies a known enzyme with the sequence of a gene that codifies an unknown enzyme (structure similarity)²⁴⁶. Newer methods have recently been applied by using additional context-specific information to support metabolism network reconstruction and network topology (arrangement of nodes, *e.g.*, metabolites, and edges, *e.g.*, reactions, which represent a specific metabolic process or state of a specific organism). These context-specific networks (cells from different parts of the organism show different metabolic networks, such as cells under different environmental conditions) were assessed by a metabolic Context-specificity Assessed by Deterministic Reaction Evaluation (mCADRE)²⁴⁸. This model allowed for the inference of a tissue-specific network based on gene expression data and metabolic network topology, together with an evaluation of functional capabilities during model building²⁴⁸. Another approach is the recently developed global biochemical reconstruction using sampling (GLOBUS)²⁵¹, which integrates sequence homology and context-based correlations under a single principled framework.

Flux balance analysis (FBA) uses GEMs to calculate the flow of metabolites through the metabolic network, which then enables the prediction of the growth rate of an organism or the production rate of a specific metabolite²⁵². In complement to FBA, the experimental ¹³C-Metabolic Flux Analysis has been used to study dynamic metabolic modifications in plants²⁵³⁻²⁵⁴, and the interplay of *in silico* simulations and experimental measurements has been reported²⁵⁵.

As the research community keeps informing organism-specific models (GEMs) with new network and pathways elements, knowledge of an organism-specific metabolism is steadily increased, and experimental analyses can be continuously enriched by more specific and detailed models.

As seen in previous sections, the use of different analytical techniques is necessary to analyse compounds with different physicochemical properties. Projects such as the Human Metabolome Database (hmdb)²⁵⁶ make available an extraordinary database including data acquired by different analytical techniques²⁵⁷, which provides insight into the metabolism of different cell types. Because of the increasing amount of information obtained by network models, similar projects are also observed in the plant metabolomics community, for example, in the project PlantMetabolomics.org²⁵⁸.

However, the lack of standard protocols for metabolomics as used in other *-omics* fields, for example, in transcriptomics, inhibits advances made by the metabolomics community and is therefore regarded as one of the main bottlenecks in the field²⁵⁹. A recent initiative for a database with detailed experimental information and accessibility of metadata, called MetaboLights²⁶⁰, could eventually overcome these barriers. In this way, sufficient information for modelling spectral information through the comparative analysis of similar experiments could be provided and, furthermore, could provide a means for disseminating metabolomics laboratory information management systems (LIMS) and data standards²⁶¹.

Despite the numerous limitations discussed above, promising studies have been carried out in metabolic network modelling.

An analysis of network weighted correlations has been successfully used to support network topology interpretations²⁶²⁻²⁶⁵. When associated with mass difference mapping, these networks may be used to infer reactions in the sample or uncharacterised compound transformations^{264, 266}. The same idea can also be extended to MS/MS spectra, which share common fragmentation patterns, eventually leading to the inference of a compound that has been transformed by a biochemical reaction²⁶⁷⁻²⁶⁸. Similar fragmentation patterns may infer similar chemical structures, and thus, there is a possibility that two substances with similar fragmentation patterns result from a compound that has been biochemically transformed. The storage of this information and of particular compound identities can help to close the feedback loop of *in silico* fragmentation^{245, 265}. Simulated spectra based on these databases can lead to common sub-structures and thus reveal new and possibly unknown compound structures²⁶⁹.

Correlation and fragmentation networks may not be well represented by classical ways of describing metabolism as dissociated pathways and may lose the cellular context in which metabolic experiments were performed²⁷⁰. Network exploratory tools are essential if metabolism alterations under specific conditions are to be explored. Cytoscape software (<http://www.cytoscape.org/>)²⁷¹⁻²⁷² provides a large collection of tools for network visualisation and inference. Cytoscape is a dynamic visualisation tool with which the user can navigate large correlation networks and potentially discover important relationships.

Because experimental information is limited by the diverse reasons discussed earlier, compounds in the network must be linked without having sufficient information due to the fact that, for example, certain metabolites could not be detected with the chosen analytical method²⁷³. Inference methods are gap filling methods, which can position incomplete information in the

network and reconstruct the network topology of a specific metabolic condition²⁷⁴⁻²⁷⁶. However, this type of utopic inference network depends heavily on advances in compound annotation/identification²⁷⁷. One of the most promising lines of research on compound annotation takes into account data and metadata and turns them into probabilistic claims. The mathematical tool used to perform this uncertainty quantification is known as Bayesian Inference. Computational implementations are available (<http://www.dcs.gla.ac.uk/inference/metsamp/> or <http://labpib.fmrp.usp.br/methods/probmetab/>), although widespread adoption by the community still depends on usability issues²⁷⁸.

Integrating metabolism information from different experimental and theoretical sources is one of the current challenges for metabolic function interpretation and dynamic metabolic modelling^{221, 279}, and the integration of this information with other *-omics* fields is predicted by systems biology²⁸⁰. Exciting achievements have been reported with platforms such as COVAIN²⁸¹. Such platforms are essential because they merge information obtained from different analytical techniques, integrate information from different sources, such as correlation structures and stoichiometric relations, and extend metabolite coverage to enable stochastic metabolism modelling²⁸². Figure 5 is a graphical interpretation of the modelling and experimentation feedback loop, and it shows how different metabolism research fields can be integrated to improve metabolic network modelling.

5 Conclusions and Future Perspectives

Despite the recent development of the plant metabolomics field, its expansion is notable, and different applications have been described. However, it is necessary to exercise caution about the acquisition, processing and interpretation of data because there are many limitations related to plant metabolomics, from the harvesting of plant material to the acquisition and processing of data. The integration of other *-omics* fields, such as proteomics, transcriptomics and genomics, is a way to study and understand a whole biological system, but it may cause network discordance when integrating information about metabolism²⁸³. One important point is that the extraction methods and analytical tools must be able to obtain all the metabolites, but even with those recent developments in analytical and bioinformatics tools, this is not yet possible or may never be possible, certainly not in a manner that is compatible with high-throughput methods. Additionally, there are many steps to harvesting the material, of which there are several parameters that influence the extraction profile, as well as steps for sample preparation and analyses, which can also have consequences for the data quality.

Compartmentalisation is another problem; there are experimental approaches to analysing spatial transcript and protein sites; however, this approach to metabolites is too recent and needs to be developed for use in high-throughput techniques, such as MALDI imaging. Thus, all these points demonstrate the necessity for a robust yet functional and pragmatic methodology and protocol standardisation.

Some different methods have been developed for extractions to reduce the problems already reported for this step, for example, the comprehensive extraction method¹⁰¹. In fact, the new methods have not yet been adequately evaluated for use in metabolomics,

which corresponds to the perpetuation of the use of classical methods in this area. In addition, as metabolite analysis in specific tissues and cells is further complicated, developments are needed in terms of analytical tools, including automation for routine use⁵³. In terms of separation-based methods, the analysis time is one of the main disadvantages, but this time has been significantly reduced by variations in classical techniques, such as UHPLC¹⁹³ and Ultra Fast-GC¹⁸⁷, as well as new stationary phases^{39, 194}. Detection MS based-methods still have limitations and must be used with care in the generation and interpretation of data to avoid untrue statements, as described in section 3.2. ESI is one of the most frequently used techniques³⁹. Moreover, MALDI is beginning to be applied to plant metabolomics^{Error! Reference source not found-55, 112}, demonstrating its promise in the field; however, it still requires significant studies regarding the ionisation of low molecular weight metabolites.

Taken together, all these data indicate that metabolomics studies require a more universal protocol, but they are developing as a way for investigators to understand living organisms.

6 Acknowledgments

The authors thank FAPESP (Process: 2009/54098-6, 2012/18031-7, 2011/05215-0, 2010/14926-4), CNPq, INCT_if, and CAPES for financial support.

Notes

^a Núcleo de Pesquisa em Produtos Naturais e Sintéticos, Department of Physics and Chemistry, Faculty of Pharmaceutical Sciences, University of São Paulo, Av. Do Café, s/n, 14040-903 Ribeirão Preto - SP, Brazil. Fax: 0055 16 3602 4178; Tel: 0055 16 3602 4158; E-mail: npelopes@fcfrp.usp.br

^b Department of Computing and Mathematics, FFCLRP, University of São Paulo, Av. Bandeirantes, 3900, 14040-901 Ribeirão Preto - SP, Brazil

† comment added by the authors

7 References

- 1 Rochfort, J. *Nat. Prod.*, 2005, **68**, 1813.
- 2 O. Fiehn, *Comp. Funct. Genom.*, 2001, **2**, 155.
- 3 R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, D. B. Kell, *Trends in Biotechnol.*, 2004, **22**, 245.
- 4 J. Kopka, A. Fernie, W. Weckwerth, Y. Gibon, M. Sitt, *Genom. Biol.*, 2004, **5**, 109.
- 5 M. Tomita, in *Metabolomics: The Frontier of Systems Biology*, ed. M. Tomita and T. Nishioka, Springer Verlag, Tokyo, 2005, ch. 1, pp. 1-6.
- 6 R. Goodacre, *Metabolomics*, 2005, **1**, 1.
- 7 T. Moritz, A. I. Johansson, in *Metabolomics, Metabonomics and Metabolite Profiling*, ed. W. J. Griffiths, RSC Publishing, Cambridge, 2008, ch. 10, pp. 254-272.
- 8 U. Roessner, A. Luedemann, D. Brust, O. Fiehn, T. Linke, L. Willmitzer, A.R. Fernie, *The Plant Cell*, 2001, **13**, 11.
- 9 R. Verpoorte, Y. H. Choi, N. R. Mustafa, H. K. Kim, *Phytochem. Rev.*, 2008, **7**, 525.
- 10 S. G. Villas-Bôas, A. Koulman, G.A. Lane, in *Metabolomics A Powerful Tool in Systems Biology*, ed. J. Nielsen, M. C. Jewett, Springer-Verlag, Berlin Heidelberg, 2007, pp. 11-52.
- 11 S. G. Oliver, M. K. Winson, D. B. Kell, F. Baganz, *Trends Biotechnol.*, 1998, **16**, 373.
- 12 J. Allen, H. M. Davey, D. Broadhurst, J. K. Heald, J. J. Rowland, S. G. Oliver, D. B. Kell, *Nat Biotechnol.*, 2003, **21**, 692.

- 13 L. M. Raamsdonk, B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M. C. Walsh, J. A. Berden, K. M. Brindle, D. B. Kell, J. J. Rowland, H. V. Westerhoff, K. van Dam, S. G. Oliver, *Nat. Biotechnol.*, 2001, **19**, 45.
- 14 J. C. Lindon, J. K. Nicholson, E. Holmes, H. Antti, M. E. Bollard, H. Keun, O. Beckonert, T. M. Ebbels, M. D. Reilly, D. Robertson, G. J. Stevens, P. Luke, A. P. Breau, G. H. Cantor, R. H. Bible, U. Niederhauser, H. Senn, G. Schlatterbeck, U. G. Sidelmann, S. M. Laursen, A. Tymiak, B. D. Car, L. Lehman-McKeeman, J. Colet, A. Loukaci, C. Thomas, *Toxicol. Appl. Pharmacol.*, 2003, **187**, 137.
- 15 J. K. Nicholson and I. D. Wilson, *Nat. Rev. Drug Discov.*, 2003, **2**, 668.
- 16 J. K. Nicholson, J. C. Lindon, E. Holmes, *Xenobiotica*, 1999, **29**, 1181.
- 17 J. L. Griffin and M. E. Bollard, *Curr Drug Metab.*, 2004, **5**, 389.
- 18 J. C. Lindon, E. Holmes, J. K. Nicholson, *Expert Rev Mol Diagn.*, 2004, **4**, 189.
- 19 J. K. Nicholson, J. Connelly, J. C. Lindon, E. Holmes, *Nat Rev Drug Discov.*, 2002, **1**, 153.
- 20 D. G. Robertson, M. D. Reilly, J. D. Baker, *Expert Opin. Drug Metab. Toxicol.*, 2005, **1**, 363.
- 21 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, D. J. Grainger, *Nat. Med.*, 2002, **8**, 1439.
- 22 T. Kuhara, in *Metabolomics: The Frontier of Systems Biology*, ed. M. Tomita, T. Nishioka, Springer Verlag, Tokyo, 2005, ch. 12, pp. 167-192.
- 23 D. S. Wishart, L. M. M. Querengesser, B. A. Lefebvre, N. A. Epstein, R. Greiner, J. B. Newton, *Clin. Chem.*, 2001, **47**, 1918.
- 24 M. Oostendorp, U. F. Engelke, M. A. Willemsen, R. A. Wevers, *Clin. Chem.*, 2006, **52**, 1395.
- 25 S. H. Moolenaar, U. F. Engelke, R. A. Wevers, *Ann. Clin. Biochem.*, 2003, **40**, 16.
- 26 D. B. Kell, D. B. *Drug Discov. Today*, 2006, **11**, 1085.
- 27 S. M. Watkins, J. B. German, *Curr. Opin. Mol. Ther.*, 2002, **4**, 224.
- 28 S. M. Watkins, P. R. Reifsnnyder, H. J. Pan, J. B. German, E. H. Leiter, *J. Lipid Res.*, 2002, **43**, 1809.
- 29 J. R. Griffiths, M. Stubbs, *Adv. Enzyme Regul.*, 2003, **43**, 67.
- 30 S. M. Watkins, J. B. German, *Curr. Opin. Biotechnol.*, 2002, **13**, 512.
- 31 J. Taylor, R. D. King, T. Altmann, O. Fiehn, *Bioinformatics*, 2002, **18**(Suppl. 2), S241.
- 32 G. S. Catchpole, M. Beckmann, D. P. Enot, M. Mondhe, B. Zywicki, J. Taylor, N. Hardy, A. Smith, R. D. King, D. B. Kell, O. Fiehn, J. Draper, *PNAS*, 2005, **102**, 14458.
- 33 L. Rizhsky, H. Liang, J. Shuman, V. Shualey, S. Davletova, R. Mittler, *Plant Physiol.*, 2004, **134**, 1683.
- 34 M. Brown, W. B. Dunn, D. I. Ellis, R. Goodacre, J. Handl, J. D. Knowles, S. O'Hagan, I. Spasic, D. B. Kell, *Metabolomics*, 2005, **1**, 39.
- 35 V. Shulaev, *Brief. Bioinform.*, 2006, **7**, 128.
- 36 D. S. Wishart, *Brief. Bioinform.*, 2007, **8**, 279.
- 37 K. Oksman-Caldentey, K. Saito, *Curr. Opin. Biotechnol.*, 2005, **16**, 174.
- 38 R. D. Hall, *New Phytologist*, 2006, **169**, 453.
- 39 J. L. Wolfender, S. Rudaz, Y. H. Choi, H. K. Kim, *Curr. Med. Chem.*, 2013, **20**, 1056.
- 40 W. Weckwerth, G. Kahl, *The Handbook of Plant Metabolomics*, Wiley-VCH Verlag GmbH & Co KGaA, Weinheim, Germany, 2013.
- 41 R. J. Bino, R. D. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B.J. Nikolau, P. Mendes, U. Roessner-Tunali, M.H. Beale, R.N. Trethewey, B.M. Lange, E.S. Wurtele, L.W. Sumner, *Trends Plant Sci.*, 2004, **9**, 418.
- 42 O. Fiehn, *Plant Mol. Biol.*, 2002, **48**, 155.
- 43 J. M. Hagel, P. J. Facchini, *Phytochem. Rev.* 2008, **7**, 479.
- 44 P. Krishnan, N. J. Kruger, R. G. Ratcliffe, *J. Exp. Bot.*, 2005, **56**, 255.
- 45 R. L. Last, A. D. Jones, Y. Shachar-Hill, *Nat. Rev. Mol. Cell Biol.*, 2007, **8**, 167.
- 46 U. Roessner, in *Metabolome Analysis: An Introduction*, ed. S. G. Villas-Bôas, U. Roessner, M. A. E. Hansen, J. Smedsgaard, J. Nielsen, 2007, John Wiley & Sons, New Jersey, 2007, ch. 8, pp. 215-238.
- 47 U. Roessner, L. Willmitzer, A. R. Fernie, *Plant Cell Rep.*, 2002, **21**, 189.
- 48 K. Saito, F. Matsuda, *Annu. Rev. Plant Biol.*, 2010, **61**, 463.
- 49 L. W. Sumner, P. Mendes, R. A. Dixon, *Phytochemistry*, 2003, **62**, 817.
- 50 R. Verpoorte, Y. H. Choi, H. K. Kim, *Phytochem. Rev.*, 2007, **6**, 3.
- 51 H. K. Kim, Y. H. Choi, R. Verpoorte, *Trends Biotechnol.*, 2011, **29**, 267.
- 52 D. Ryan, K. Robards, *Anal. Chem.*, 2006, **78**, 7954.
- 53 S. Moco, B. Schneider, J. Vervoort, *J. Proteome Res.*, 2009, **8**, 1694.
- 54 Y. Li, B. Shrestha, A. Vertes, *Anal. Chem.*, 2008, **80**, 407.
- 55 D. Miura, Y. Fujimura, H. Tachibana, H. Wariishi, *Anal. Chem.*, 2010, **82**, 498.
- 56 K. Dettmer, P. A. Aronov, B. D. Hammock, *Mass Spectrom. Rev.*, 2007, **26**, 51.
- 57 M. Oldiges, S. Lütz, S. Pflug, K. Schroer, N. Stein, C. Wiendahl, *Appl. Microbiol. Biotechnol.*, 2007, **76**, 495.
- 58 S. G. Villas-Bôas, S. Mas, M. Åkesson, J. Smedsgaard, J. Nielsen, *Mass Spectrom. Rev.*, 2005, **24**, 613.
- 59 J. Nielsen, M.C. Jewett, in *Metabolomics A Powerful Tool in Systems Biology*, ed. J. Nielsen, M.C. Jewett, Springer-Verlag, Heidelberg, 2007, pp. 1-10.
- 60 M. R. Viant, C. Ludwig, U. L. Günther, in *Metabolomics, Metabonomics and Metabolite Profiling*, ed. W.J. Griffiths, RSC Publishing, Cambridge, 2008, ch. 2, pp. 44-70.
- 61 K. Dettmer, B. D. Hammock, *Environ Health Perspect.*, 2004, **112**, A396.
- 62 K. Hollywood, D.R. Brison, R. Goodacre, *Proteomics*, 2006, **6**, 4716.
- 63 C. W. W. Beecher, in *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, ed. G. G. Harrigan, R. Goodacre, Kluwer Academic Publishers, Dordrecht, 2003, ch. 17, pp. 311-319.
- 64 H. Jenkins, N. Hardy, M. Beckmann, J. Draper, A. R. Smith, J. Taylor, O. Fiehn, R. Goodacre, R. J. Bino, R. Hall, J. Kopka, G. A. Lane, B. M. Lange, J. R. Liu, P. Mendes, B. J. Nikolau, S. G. Oliver, N. W. Paton, S. Rhee, U. Roessner-Tunali, K. Saito, J. Smedsgaard, L. W. Sumner, T. Wang, S. Walsh, E. S. Wurtele, D. B. Kell, *Nat. Biotechnol.*, 2004, **22**, 1601.
- 65 R. N. Bennett, R. M. Wallsgrove, *New Phytol.*, 1994, **127**, 617.
- 66 M. Wink, *Phytochemistry*, 2003, **64**, 3.
- 67 L. Gobbo-Neto, N. P. Lopes, *Quim. Nova*, 2007, **30**, 374.
- 68 U. Roessner, F. Pettolino, in *Metabolomics A Powerful Tool in Systems Biology*, ed. J. Nielsen, M.C. Jewett, Springer-Verlag, Berlin Heidelberg, 2007, pp. 253-278.
- 69 K. R. Markham, G. J. Tanner, M. Caasi-Lit, M. I. Whitecross, M. Nayudu, K. A. Mitchell, *Phytochemistry*, 1998, **49**, 1913.
- 70 M. Kusano, T. Tohge, A. Fukushima, M. Kobayashi, N. Hayashi, H. Otsuki, Y. Kondou, H. Goto, M. Kawashima, F. Matsuda, R. Niida, M. Matsui, K. Saito, A. R. Fernie, *Plant J.*, 2011, **67**, 354.
- 71 K. Hirata, M. Asada, E. Yatani, K. Miyamoto, Y. Miura, *Planta Med.*, 1993, **59**, 46.
- 72 H. K. Kim, R. Verpoorte, *Phytochem. Anal.*, 2010, **21**, 4.
- 73 E. Urbanczyk, C. Baxter, A. Kolbe, J. Kopka, L. J. Sweetlove, A. R. Fernie, *Planta*, 2005, **221**, 891.
- 74 M. G. V. Silva, A. A. Craveiro, F. J. A. Matos, M. I. L. Machado, J. W. Alencar, *Fitoterapia*, 1999, **70**, 32.
- 75 K. Itenov, P. Mølgaard, U. Nyman, *Phytochemistry*, 1999, **52**, 1229.
- 76 J. W. Fairbairn, P. N. Suwal, *Phytochemistry*, 1961, **1**, 38.
- 77 M. Veit, W. Bilger, T. Mühlbauer, W. Brummet, K. Winter, *Plant Physiol.*, 1996, **148**, 478.
- 78 D. A. Herms, W. J. Mattson, *Q. Rev. Biol.*, 1992, **67**, 283.
- 79 D. R. Crankshaw, J. H. Langenheim, *Biochem. Syst. Ecol.*, 1981, **9**, 115.
- 80 K. A. Leiss, Y. H. Choi, I. B. Abdel-Farid, R. Verpoorte, P. G. L. Klinkhamer, *J. Chem. Ecol.*, 2009, **35**, 219.

- 81 T. J. Schmidt, U. Bomme, A. W. Alfermann, *Planta Med.*, 1998, **64**, 268.
- 82 U. Roessner, C. Wagner, J. Kopka, R. N. Trethewey, L. Willmitzer, *Plant J.*, 2000, **23**, 131.
- 83 M. A. Farag, M. G. F. Eldin, H. Kassem, M. A. Fetouh, *Phytochem. Anal.*, 2013, **24**, 277.
- 84 R. Shroff, F. Vergara, A. Muck, A. Svatos, J. Gershenzon, *J. Proc. Natl. Acad. Sci. USA*, 2008, **105**, 6196.
- 85 B. Li, S. H. Hansen, C. Janfelt, *Int. J. Mass Spectrom.*, 2013, **348**, 15.
- 86 N. Harbourne, E. Marete, J. C. Jacquier, D. O'Riordan, *Lwt-Food Sci. Technol.*, 2009, **42**, 1468.
- 87 J. L. Schmitke, C. R. Wescott, A. M. Klibanov, *J. Am. Chem. Soc.*, 1996, **118**, 3360.
- 88 A. Oikawa, K. Saito, *Plant J.*, 2012, **70**, 30.
- 89 H. Y. Zhou, C. Z. Liu, *J. Chromatogr. A*, 2006, **1129**, 135.
- 90 M. Schad, R. Mungur, O. Fiehn, J. Kehr, *Plant Methods*, 2005, **1**, 2 (doi: 10.1186/1746-4811-1-2).
- 91 T. Nelson, N. Gandotra, S. L. Tausta, *Curr. Opin. Plant Biol.*, 2008, **11**, 567.
- 92 E. Abbott, D. Hall, B. Hamberger, J. Bohlmann, *BMC Plant Biology*, 2010, **10**, 106.
- 93 B. Ebert, D. Zöller, A. Erban, I. Fehrlé, J. Hartmann, A. Niehl, J. Kopka, J. Fishan, *J. Exp. Bot.*, 2010, **61**, 1321.
- 94 D. Hölscher, R. Shroff, K. Knop, M. Gottschaldt, A. Crecelius, B. Schneider, D. G. Heckel, U. S. Schubert, A. Svatos, *Plant J.*, 2009, **60**, 907.
- 95 B. St-Pierre, F. A. Vazquez-Flota, V. De Luca, *Plant Cell*, 1999, **11**, 887.
- 96 K. Morgenthal, S. Wienkoop, F. Wolschin, W. Weckwerth, in *Methods in Molecular Biology, vol. 358: Metabolomics: Methods and Protocols*, ed. W. Weckwerth, Totowa, 2007, ch. 4, pp. 57-75.
- 97 E. A. Schmelz, J. Engelberth, J. H. Tumlinson, A. Block, H. T. Alborn, *Plant J.*, 2004, **39**, 790.
- 98 C. Deng, N. Liu, M. Gao, X. Zhang, *J. Chromatogr. A*, 2007, **1153**, 90.
- 99 F. Maltese, F. Van der Kooy, R. Verpoorte, *Nat. Prod. Commun.*, 2009, **4**, 447.
- 100 N. D. Yuliana, M. Jahangir, R. Verpoorte, H. Y. Choi, *Phytochem. Rev.*, 2013, **12**, 293.
- 101 N. D. Yuliana, A. Khatib, R. Verpoorte, Y. H. Choi, *Anal. Chem.*, 2011, **83**, 6902.
- 102 E. Martino, I. Ramaiola, M. Urbano, F. Bracco, S. Collina, *J. Chromatogr. A*, 2006, **1125**, 147.
- 103 A. Mustafa, C. Turner, *Anal. Chim. Acta*, 2011, **703**, 8.
- 104 M. H. Lee, C. C. Lin, *Food Chem.*, 2007, **105**, 223.
- 105 Y. Liu, T. Chen, Y. Qiu, Y. Cheng, Y. Cao, A. Zhao, W. Jia, *Anal. Bioanal. Chem.*, 2011, **400**, 1405.
- 106 H. K. Kim, Y. H. Choi, T. J. C. Luijck, R. A. V. Rocha, R. Verpoorte, *Phytochem. Anal.*, 2004, **15**, 257.
- 107 V. Mandal, Y. Mohan, S. Hemalatha, *Pharmacog. Rev.*, 2007, **1**, 7.
- 108 M. Vinatoru, *Ultrason. Sonochem.*, 2001, **8**, 303.
- 109 J. Gullberg, P. Jonsson, A. Nordström, M. Sjöström, T. Moritz, *Anal. Biochem.*, 2004, **331**, 283.
- 110 J. W. Allwood, R. Goodacre, *Phytochem. Anal.*, 2010, **21**, 33.
- 111 G. Romanik, E. Gilgenast, A. Przyjazny, M. Kamiński, *J. Biochem. Biophys. Methods*, 2007, **70**, 253.
- 112 B. Kleidus, D. Vitamvasová, V. Kubán, *J. Chromatogr. A*, 1999, **839**, 261.
- 113 C. W. Huie, *Anal. Bioanal. Chem.*, 2002, **373**, 23.
- 114 K. Hoyerová, A. Gaudinová, J. Malbeck, P. I. Dobrev, T. Kocábek, B. Šolcová, A. Trávníčková, M. Kamínek, *Phytochemistry*, 2006, **67**, 1151.
- 115 M. C. Hennion, *J. Chromatogr. A*, 1999, **856**, 3.
- 116 A. Zwir-Ferenc, M. Biziuk, *Polish. J. of Environ. Stud.*, 2006, **15**, 677.
- 117 K. Glowinski, G. Zgórk, M. Kozyra, *J. Chromatogr. A*, 1996, **730**, 25.
- 118 M. Stobiecki, P. Wojtaszek, K. Gulewicz, *Phytochem. Anal.*, 1997, **8**, 153.
- 119 D. Vuckovic, *Anal. Bioanal. Chem.*, 2012, **403**, 1523.
- 120 M. Fraccaroli, S. Nicoletti, F. Maltese, Y. H. Choi, F. Guzzo, M. Levi, R. Verpoorte, *Biotechnol. Lett.*, 2008, **30**, 2031.
- 121 R. Nakabayashi, K. Saito, *Anal. Bioanal. Chem.*, 2013, **405**, 5005.
- 122 R. J. Luján, F. P. Capote, M. D. L. Castro, *Phytochem. Anal.*, 2009, **20**, 221.
- 123 A. Khatib, E. G. Wilson, H. K. Kim, A. W. M. Lefeber, C. Erkelens, Y. H. Choi, R. Verpoorte, *Anal. Chim. Acta*, 2006, **559**, 264.
- 124 G. Karasová, J. Lehotay, J. Sádecká, I. Skacáni, M. Lachová, *J. Sep. Sci.*, 2005, **28**, 2468.
- 125 J. Fenn, *Angew. Chem. Int. Ed.*, 2003, **42**, 3871.
- 126 P. Jayaweera, A. T. Blades, M. G. Ikonomou, P. Kebarle, *J. Am. Chem. Soc.*, 1990, **112**, 2452.
- 127 M. G. Ikonomou, A. T. Blades, P. Kebarle, *Anal. Chem.*, 1991, **63**, 1989.
- 128 A. E. M. Crotti, R. L. Vessecchi, J. L. C. Lopes, N. P. Lopes, *Quim. Nova*, 2006, **29**, 287.
- 129 P. Kebarle, U. H. Verkerk, in *Electrospray and MALDI Mass Spectrometry*, ed. R. B. Cole, John and Wiley & Sons, 2nd edn., 2010, ch. 1, pp. 3-48.
- 130 J. B. Fenn, *Int. J. Mass Spectrom.*, 2000, **200**, 459.
- 131 M. Yamashita, J. B. Fenn, *J. Phys. Chem.*, 1984, **88**, 4451.
- 132 N. P. Lopes, C. B. W. Stark, P. J. Gates, J. Staunton, *Analyst*, 2002, **127**, 503.
- 133 N. P. Lopes, C. B. W. Stark, H. Hong, P. J. Gates, J. Staunton, *Analyst*, 2001, **126**, 1630.
- 134 N. P. Lopes, P. J. Gates, J. P. G. Wilkins, J. Staunton, *Analyst*, 2002, **127**, 1224.
- 135 N. P. Lopes, F. A. A. Paz, P. J. Gates, *Rev. Bras. Cienc. Farm.*, 2006, **42**, 363.
- 136 F. A. Dorr, J. C. Tomaz, N. P. Lopes, E. Pinto, *Rapid Commun. Mass Spectrom.*, 2008, **22**, 2015.
- 137 H. G. J. Mol, R. C. J. Van Dam, P. Zomer, P. P. J. Mulder, *Food Addit. Contam. Part A Chem. Anal. Control Expo. Risk Assess*, 2011, **28**, 1405.
- 138 G. J. Van Berkel, F. Zhou, *Anal. Chem.*, 1995, **67**, 2916.
- 139 G. J. Van Berkel, G. E. Giles, J. S. Bullock, L. J. Gray, *Anal. Chem.*, 1999, **71**, 5288.
- 140 R. Vessecchi, A. E. M. Crotti, T. Guaratini, P. Colepiccolo, S. E. Galembeck, N. P. Lopes, *Mini-Rev. Org. Chem.*, 2007, **4**, 75.
- 141 T. Guaratini, R. L. Vessecchi, F. C. Lavarda, P. M. B. G. Maia Campos, Z. Naal, P. J. Gates, N. P. Lopes, *Analyst*, 2004, **129**, 1223.
- 142 T. Guaratini, R. Vessecchi, E. Pinto, P. Colepiccolo, N. P. Lopes, *J. Mass Spectrom.*, 2005, **40**, 963.
- 143 T. Guaratini, P. J. Gates, K. H. M. Cardozo, P. M. B. G. M. Campos, P. Colepiccolo, N. P. Lopes, *Eur. J. Mass Spectrom.*, 2006, **12**, 71.
- 144 T. Guaratini, P. J. Gates, E. Pinto, P. Colepiccolo, N. P. Lopes, *Rapid Commun. Mass Spectrom.*, 2007, **21**, 3842.
- 145 R. L. Vessecchi, S. E. Galembeck, N. P. Lopes, P. G. B. D. Nascimento, A. E. M. Crotti, *Quim. Nova*, 2008, **31**, 840.
- 146 R. Vessecchi, Z. Naal, J. N. C. Lopes, S. E. Galembeck, N. P. Lopes, *J. Phys. Chem. A*, 2011, **115**, 5453.
- 147 R. Vessecchi, J. N. C. Lopes, N. P. Lopes, S. E. Galembeck, *J. Phys. Chem. A*, 2011, **115**, 12780.
- 148 R. Vessecchi, F. S. Emery, S. E. Galembeck, N. P. Lopes, *Rapid. Comm. Mass. Spectrom.*, 2010, **24**, 2101.
- 149 M. Karas, D. Bachmann, F. Hillenkamp, *Anal. Chem.*, 1985, **57**, 2935.

- 150 M. Karas, D. Bachmann, U. Bahr, F. Hillenkamp, *Int. J. Mass Spectrom. Ion Processes*, 1987, **78**, 53.
- 151 R. C. Beavis, B. T. Chait, *Rapid Commun. Mass Spectrom.*, 1989, **3**, 432.
- 152 R. C. Beavis, T. Chaudhary, B. T. Chait, *Org. Mass Spectrom.*, 1992, **27**, 156.
- 153 M. Karas, H. Ehring, E. Nordhoff, B. Stahl, K. Strupat, F. Hillenkamp, M. Grehl, B. Krebs, *Org. Mass Spectrom.*, 1993, **28**, 1476.
- 154 K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida, T. Matsuo, *Rapid Commun. Mass Spectrom.*, 1988, **2**, 151.
- 155 K. Tanaka, *Angew. Chem. Int. Ed.*, 2003, **42**, 3861.
- 156 M. Karas, U. Bahr, U. Giessmann, *Mass Spectrom. Rev.*, 1991, **10**, 335.
- 157 R. C. Beavis, *Org. Mass Spectrom.*, 1992, **27**, 653.
- 158 U. Bahr, M. Karas, F. Hillenkamp, *Fresenius' J. Anal. Chem.*, 1994, **348**, 783.
- 159 D. P. Pavarini, D. B. Silva, C. A. Carollo, A. P. F. Portella, S. R. Latansio-Aidar, P. O. Cavalin, V. C. Oliveira, B. H. P. Rosado, M. P. M. Aidar, V. S. Bolzani, N. P. Lopes, C. A. Joly, *J. Mass Spectrom.*, 2012, **47**, 1482.
- 160 A. Tholey, E. Heinze, *Anal. Bioanal. Chem.*, 2006, **386**, 24.
- 161 M. Lu, Y. Lai, G. Chen, Z. Cai, *Anal. Chem.*, 2011, **83**, 3161.
- 162 J. A. Liu, L. Xiong, S. Zhang, J. C. Wei, S. X. Xiong, *Metabolomics*, 2012, **2**, S1.
- 163 L. H. Cohen, A. I. Gusev, *Anal. Bioanal. Chem.*, 2002, **373**, 571.
- 164 R. Knochenmuss, in *Electrospray and MALDI Mass Spectrometry*, ed. R. B. Cole, John and Wiley & Sons, 2nd edn., 2010, ch. 5, pp. 149-183.
- 165 A. A. Parent, T. M. Anderson, D. J. Michaelis, G. L. Jiang, P. B. Savage, M. R. Linford, *Appl. Surf. Sci.*, 2006, **252**, 6746.
- 166 A. Orinak, R. Orinakova, H. F. Arlinghaus, G. Vering, S. Hellweg, *Surf. Interface Anal.*, 2006, **38**, 599.
- 167 T. Knubovets, L. Song, E. Golenko, Y. Wang, A. Mikulskis, V. Panchagnulu, P. Jackson, E. Scrivener, W. F. Patton, *Mol. Cell Proteomics*, 2006, **5**, S91.
- 168 K. Nakamura, Y. Suzuki, N. Goto-Inoue, C. Yoshida-Noro, A. Suzuki, *Anal. Chem.*, 2006, **78**, 5736.
- 169 Y. L. Yang, Y. Xu, P. Straight, P. C. Dorris, *Nat. Chem. Biol.*, 2009, **5**, 885.
- 170 M. L. Reyzer, R. M. Caprioli, *J. Proteome Res.*, 2005, **4**, 1138.
- 171 Y. J. Lee, D. C. Perdian, Z. H. Song, E. S. Yeung, B. J. Nikolau, *Plant J.*, 2012, **70**, 81.
- 172 Y. Li, B. Shrestha, A. Vertes, *Anal. Chem.*, 2007, **79**, 523.
- 173 P. D. Fraser, E. M. A. Enfissi, M. Goodfellow, T. Eguchi, P. M. Bramley, *Plant J.*, 2007, **49**, 552.
- 174 L. C. Keles, N. I. de Melo, G. P. Aguiar, K. A. L. Wakabayashi, C. E. de Carvalho, W. R. Cunha, A.E.M. Crotti, J. L. C. Lopes, N. P. Lopes, *Quim. Nova*, 2010, **33**, 2245.
- 175 S. A. McLuckey, *J. Am. Mass Spectrom.*, 1992, **3**, 599.
- 176 M. Baker, W. Garbyelski, *Int. J. Mass Spectrom.*, 2007, **262**, 128.
- 177 R. H. Perry, R. G. Cooks, R. J. Noll, *Mass Spectrom. Rev.*, 2008, **27**, 661.
- 178 A. E. M. Crotti, J. L. C. Lopes, N. P. Lopes, *J. Mass Spectrom.*, 2005, **40**, 1030.
- 179 H. I. Kentamaa, R. G. Cooks, *J. Am. Soc.*, 1985, **107**, 1881.
- 180 A. E. M. Crotti, E. S. Bronze-Uhle, P. G. B. D. Nascimento, P. M. Donate, S. E. Galembeck, R. Vessecchi, N. P. Lopes, *J. Mass Spectrom.*, 2009, **44**, 1733.
- 181 R. J. Vessecchi, C. A. Carollo, J. N. C. Lopes, A. E. M. Crotti, N. P. Lopes, S. E. Galembeck, *J. Mass Spectrom.*, 2009, **44**, 1224.
- 182 L. R. Sartori, R. Vessecchi, H. Humpf, F. B. Da Costa, N. P. Lopes, *Rapid Commun. Mass Spectrom.*, submitted.
- 183 L. Gobbo-Neto, P. J. Gates, N. P. Lopes, *Rapid Commun. Mass Spectrom.*, 2008, **22**, 3802.
- 184 T. Soga, in *Methods in Molecular Biology*, vol. 358: *Metabolomics: Methods and Protocols*, ed. W. Weckwerth, Totowa, 2007, ch. 8, pp. 129-137.
- 185 K. Yan, J. Brumbaugh, J. Barrett, S. Jung, S. Hu, *Curr. Metab.*, 2013, **1**, 121.
- 186 R. Ramautar, O. A. Mayboroda, G. W. Somsen, G. J. De Jong, *Electrophoresis*, 2011, **32**, 52.
- 187 J. W. Allwood, A. Erban, S. Koning, S. K. B. Dunn, A. Luedemann, A. Lommen, L. Kay, R. Löscher, J. Kopka, R. Goodacre, *Metabolomics*, 2009, **5**, 479.
- 188 C. Bicchì, C. Brunelli, C. Cordero, P. Rubiolo, M. Galli, A. Sironi, *J. Chromatogr. A*, 2004, **1024**, 195.
- 189 M. F. Almstetter, P. J. Oefner, K. Dettmer, *Anal. Bioanal. Chem.*, 2012, **402**, 1993.
- 190 M. Rigobello-Masini, J. C. P. Penteado, J. C. Masini, *Anal. Bioanal. Chem.*, 2013, **405**, 2107.
- 191 N. L. Kuehnbaum, P. Britz-McKibbin, *Chem. Rev.*, 2013, **113**, 2437.
- 192 V. V. Tolstikov, A. Lommen, K. Nakanishi, N. Tanaka, O. Fiehn, *Anal. Chem.*, 2003, **75**, 6737.
- 193 X. Wang, H. Sun, A. Zhang, P. Wang, Y. Han, *J. Sep. Sci.*, 2011, **34**, 3451.
- 194 K. Spagou, H. Tsoukali, N. Raikos, H. Gika, I. D. Wilson, G. Theodoridis, *J. Sep. Sci.*, 2010, **33**, 716.
- 195 M. E. Bier, in *Electrospray and MALDI Mass Spectrometry*, ed. R. B. Cole, John and Wiley & Sons, 2nd edn., 2010, ch. 9, pp. 265-344.
- 196 Y. Sawada, K. Akiyama, A. Sakata, A. Kuwahara, H. Otsuki, T. Sakurai, K. Saito, M. Y. Hirai, *Plant Cell Physiol.*, 2009, **50**, 37.
- 197 A. R. S. Ross, S. J. Ambrose, A. J. Cutler, J. A. Feurtado, A. R. Kermode, K. Nelson, R. Zhou, S. R. Abrams, *Anal. Biochem.*, 2004, **329**, 324.
- 198 T. Kind, O. Fiehn, *Bioanal. Rev.*, 2010, **2**, 23.
- 199 J. Kang, L. A. Hick, W. E. Price, *Rapid. Commun. Mass Spectrom.*, 2007, **21**, 857.
- 200 L. Mattoli, F. Cangi, A. Maidecchi, C. Ghiara, E. Ragazzi, M. Tubaro, L. Stella, F. Tisato, P. Traldi, *J. Mass Spectrom.*, 2006, **41**, 1534.
- 201 A. Koulman, B. A. Tapper, K. Fraser, M. Cao, G. A. Lane, S. Rasmussen, *Rapid Commun. Mass Spectrom.*, 2007, **21**, 421.
- 202 S. C. Brown, G. Kruppa, J. L. Dasseux, *Mass Spectrom. Rev.*, 2005, **24**, 223.
- 203 A. Aharoni, R. C. H. De Vos, H. A. Verhoeven, C. A. Maliapaard, G. Kruppa, R. Bino, D. B. Goodenowe, *OMICS*, 2017, **6**, 217.
- 204 G. Madalinski, E. Godat, S. Alves, D. Lesage, E. Genin, P. Levi, J. Labarre, J. C. Tabet, E. Ezan, C. Junot, *Anal. Chem.*, 2008, **80**, 3291.
- 205 W. B. Dunn, D. Broadhurst, M. Brown, P. N. Baker, C. W.G. Redman, L. C. Kenny, D. B. Kell, *J. Chromatogr. B*, 2008, **871**, 288.
- 206 Y. Xu, J. F. Heilner, G. Madalinski, E. Genin, E. Ezan, J. C. Tabet, C. Junot, *Anal. Chem.*, 2010, **82**, 5490.
- 207 M. Guilhaus, V. Mlynski, D. Selby, *Rapid Commun. Mass Spectrom.*, 1997, **11**, 951.
- 208 T. J. Cornish, R. J. Cotter, *Anal. Chem.*, 1993, **65**, 1043.
- 209 F. W. McLafferty, *Science*, 1981, **214**, 280.
- 210 B. Mehrotra, Mendes P., in *Biotechnology in Agriculture and Forestry 57: Plant Metabolomics*, ed. K. Saito, R.A. Dixon, L. Willmitzer, Springer Verlag, Berlin Heidelberg, 2006, section 2, ch. 1, pp. 105-115.
- 211 J. Trygg, E. Holmes, T. Lundstedt, *J. Proteome Res.*, 2007, **6**, 469.
- 212 D. L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufmann, *Chemometrics: A Textbook*. Elsevier, Amsterdam, 1988.
- 213 A. Varnek, I. I. Baskin, *Mol. Inf.*, 2011, **30**, 20.
- 214 S. Wold, M. Sjöström, 1998. Chemometrics, present and future success. *Chemom. Intell. Lab. Syst.* **44**, 3.

- 215 S. Wold, *Chemom. Intell. Lab. Syst.*, 1995, **30**, 109.
- 216 J. Gasteiger, *Anal Bioanal Chem*, 2006, **384**, 57.
- 217 P. Hogeweg, *Plos Comp. Biol.*, 2011, **7**, 1.
- 218 J. Trygg, J. Gullberg, A.I. Johansson, P. Jonsson, T. Moritz, in *Biotechnology in Agriculture and Forestry 57: Plant Metabolomics*, ed. K. Saito, R.A. Dixon, L. Willmitzer, Springer Verlag, Berlin Heidelberg, 2006, section 2, ch. 2, pp. 117-128.
- 219 R. Steuer, K. Morgenthal, W. Weckwerth, J. Selbig, in *Methods in Molecular Biology*, vol. 358: *Metabolomics: Methods and Protocols*, ed. W. Weckwerth, Totowa, 2007, ch. 7, pp. 105-126.
- 220 L. W. Sumner, E. Urbanczyk-Wochniak, C. D. Broeckling, *Methods Mol. Biol.*, 2007, **406**, 409.
- 221 A. K. Smilde, J. A. Westerhuis, H. C. J. Hoefsloot, S. Bijlsma, C. M. Rubingh, D. J. Vis, R. H. Jellema, H. Pijl, F. Roelfsema, J. Van der Greef, *Metabolomics*, 2010, **6**, 3.
- 222 M. A. E. Hansen, in *Metabolome Analysis: An Introduction*, ed. S. G. Villas-Boas, U. Roessner, M. A. E. Hansen, J. Smedsgaard, J. Nielsen, John Wiley & Sons, New Jersey, 2007, ch. 5, pp. 146-187.
- 223 C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, *Anal. Chem.*, 2006, **78**, 779.
- 224 S. Castillo, P. Gopalacharyulu, L. Yetukuri, M. Orešič, *Chemom. Intell. Lab. Syst.*, 2011, **108**, 23.
- 225 R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, M. J. van der Werf, *BMC Genomics*, 2006, **7**, 142.
- 226 P. G. A. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, Jr, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, R. Aebersold, *Nat. Biotechnol.*, 2004, **22**, 1459.
- 227 M. Katajamaa, M. Orešič, *J. Chromatogr. A*, 2007, **1158**, 318.
- 228 M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Rudermann, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessler, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M. Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, P. Mallick, *Nat. Biotechnol.*, 2012, **30**, 918.
- 229 J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmüller, P. Dörmann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, D. Steinhauser, *Bioinformatics*, 2005, **21**, 1635.
- 230 A. Luedemann, K. Strassburg, A. Erban, J. Kopka, *Bioinformatics*, 2008, **24**, 732.
- 231 A. Cuadros-Inostroza, C. Caldana, H. Redestig, M. Kusano, J. Lisec, H. Peña-Cortés, L. Willmitzer, M. A. Hannah, *Bioinformatics*, 2009, **10**, 428.
- 232 R. Tautenhahn, G. J. Patti, D. Rinehart, G. E. Siuzdak, *Anal. Chem.*, 2012, **84**, 5035.
- 233 T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, *BMC Bioinformatics*, 2010, **11**, 395.
- 234 X. Wei, W. Song, X. Shi, I. Koo, B. Wang, J. Zhang, X. Yin, Y. Tang, B. Bogdanov, S. Kim, Z. Zhou, C. McClain, X. Zhang, *Anal. Chem.*, 2011, **83**, 7668.
- 235 S. Gibb, K. Strimmer, *Bioinformatics*, 2012, **28**, 2270.
- 236 J. Hummel, N. Strehmel, C. Bölling, S. Schmidt, D. Walther, J. Kopka in *The Handbook of Plant Metabolomics*, ed. W. Weckwerth, G. Kahl, Wiley-VCH Verlag GmbH & Co KGaA, Weinheim, Germany, 2013, ch. 18, pp. 321-343.
- 237 G. J. Patti, R. Tautenhahn, G. Siuzdak, *Nat Protoc.*, 2012, **7**, 508.
- 238 N. Strehmel, J. Hummel, A. Erban, K. Strassburg, J. Kopka, *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.*, 2008, **871**, 182.
- 239 R. Kaliszan, *Chem Rev.*, 2007, **107**, 3212.
- 240 A. Jankevics, M. E. Merlo, M. De Vries, R. J. Vonk, E. Takano, R. Breitling, *Metabolomics*, 2012, **8**, 29.
- 241 C. Kuhl, R. Tautenhahn, C. Bottcher, T. R. Larson, S. Neumann, *Anal. Chem.*, 2012, **84**, 283.
- 242 R. A. Scheltema, A. Jankevics, R. C. Jansen, M. A. Schwertz, R. Breitling, *Anal. Chem.*, 2012, **83**, 2786.
- 243 T. Pluskal, T. Uehara, M. Yanagida, *Anal. Chem.*, 2012, **84**, 4396.
- 244 D. J. Creek, A. Jankevics, K. E. V. Burgess, R. Breitling, M. P. Barrett, *Bioinformatics*, 2012, **28**, 1048.
- 245 S. Neumann, F. Rasche, S. Wolf, S. Böcker in *The Handbook of Plant Metabolomics*, ed. W. Weckwerth, G. Kahl, Wiley-VCH Verlag GmbH & Co KGaA, Weinheim, Germany, 2013, ch. 16, pp. 289-303.
- 246 I. Thiele B. Ø. Palsson, *Nat. Prot.*, 2010, **5**, 93.
- 247 T. Y. Kim, S. B. Sohn, Y. B. Kim, W. J. Kim, S. Y. Lee, *Curr Opin Biotechnol.*, 2012, **23**, 617.
- 248 Y. Wang, J. A. Eddy, N. D. Price, *BMC Syst. Biol.*, 2012, **6**, 153.
- 249 C. G. de O. Dal'Molin, L. E. Quek, R. W. Palfreyman, S. M. Brumbley, L. K. Nielsen, *Plant Physiol.*, 2010, **154**, 1871.
- 250 C. G. de O. Dal'Molin, L. K. Nielsen, *Curr. Opin. Biotechnol.*, 2013, **24**, 271.
- 251 G. Plata, T. Fuhrer, T. L. Hsiao, U. Sauer, D. Vitkup, *Nat. Chem. Biol.*, 2012, **8**, 848.
- 252 J. D. Orth, I. Thiele, B. Ø. Palsson, *Nat. Biotechnol.*, 2010, **28**, 245.
- 253 N. J. Kruger, R. G. Ratcliffe, *Biochimie*, 2009, **91**, 697.
- 254 J. Schwender, *Front Plant Sci.*, 2011, **2**, 63.
- 255 J. Hay, J. Schwender, *Plant J.*, 2011, **67**, 513.
- 256 I. Nobeli, J. M. Thornton, *Bioessays*, 2006, **28**, 534.
- 257 N. Psychogios, D. D. Hau, J. Peng, A. C. Guo, R. Mandal, S. Bouatra, I. Sinelnikov, R. Krishnamurthy, R. Eisner, B. Gautam, N. Young, J. Xia, C. Knox, E. Dong, P. Huang, Z. Hollander, T. L. Pedersen, S. R. Smith, F. Bamforth, R. Greiner, B. McManus, J. W. Newman, T. Goodfriend, D. S. Wishart, *PLoS One*, 2011, **16**, e16957.
- 258 P. Bais, S. M. Moon, K. He, R. Leitao, K. Dreher, T. Walk, Y. Sucaet, L. Barkan, G. Wohlgenuth, M. R. Roth, E. S. Wurtele, P. Dixon, O. Fiehn, B. M. Lange, V. Shuaev, L. W. Sumner, R. Welti, B. J. Nikolau, S. Y. Rhee, J. A. Dickerson, *Plant Physiol.*, 2010, **152**, 1807.
- 259 A. R. Fernie, M. Stitt, *Plant Physiol.*, 2012, **158**, 1139.
- 260 C. Steinbeck, P. Conesa, K. Haug, T. Mahendrakar, M. Williams, E. Maguire, P. Rocca-Serra, S. A. Sansone, R. M. Salek, J. L. Griffin, *Metabolomics*, 2012, **8**, 757.
- 261 O. Fiehn, D. Robertson, J. Griffin, M. Werf, B. Nikolau, N. Morrison, L. W. Sumner, R. Goodacre, N. W. Hardy, C. Taylor, J. Fostel, B. Kristal, R. Kaddurah-Daouk, P. Mendes, B. Van Ommen, J. C. Lindon, S. A. Sansone, *Metabolomics*, 2007, **3**, 175.
- 262 R. Steuer, *Brief Bioinform.*, 2006, **7**, 151.
- 263 R. Breitling, S. Ritchie, D. Goodenowe, M. L. Stewart, M. P. Barrett, *Metabolomics*, 2006, **2**, 155.
- 264 F. Jourdan, R. Breitling, M. P. Barrett, D. Gilbert, *Bioinformatics*, 2008, **24**, 143.
- 265 E. Gaquerel, C. Kuhl, S. Neumann, *Metabolomics*, 2013, **9**, 904.
- 266 R. J. M. Weber, M. R. Viant, *Chemom. Intell. Lab. Syst.*, 2010, **104**, 75.
- 267 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, *PNAS*, 2012, **109**, E1743.
- 268 D. K. Barupal, P. K. Haldiya, G. Wolgemuth, T. Kind, S. L. Kothari, K. E. Pinkerton, O. Fiehn, *BMC Bioinformatics*, 2012, **13**, 99.
- 269 S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann, *BMC Bioinformatics*, 2010, **11**, 148.

270 M. M. W. B. Hendriks, F. A. Van Eeuwijk, R. H. Jellema, J. A. Westerhuis, T. H. Reijmers, H. C. J. Hoefsloot, A. K. Smilde, *Trends Anal. Chem.*, 2011, **30**, 1685. 70

271 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, *Genome Res.*, 2003, **13**, 2498. 5

272 S. Lotia, J. Montojo, Y. Dong, G. D. Bader, A. R. Pico, *Bioinformatics*, 2013, **29**, 1350. 75

273 F. Jourdan, L. Cottret, L. Huc, D. Wildridge, R. Scheltema, A. Hillenweck, M. P. Barrett, D. Zalko, D. G. Watson, L. Debrauwer, *Metabolomics*, 2010, **6**, 312. 10

274 A. Morgat, E. Coissac, E. Coudert, K. B. Axelsen, G. Keller, A. Bairoch, A. Bridge, L. Bougueleret, I. Xenarios, A. Viari, *Nucleic Acids Res.*, 2012, **40**, D761. 80

275 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, *Nucleic Acids Res.*, 2011, **40**, D742. 85

276 L. Cottret, D. Wildridge, F. Vinson, M. P. Barrett, H. Charles, M. F. Sagot, F. Jourdan, *Nucleic Acids Res.*, 2010, **38**, W132. 90

277 W. B. Dunn, A. Erban, R. J. M. Weber, D. J. Creek, M. Brown, R. Breitling, T. Hankemeier, R. Goodacre, S. Neumann, J. Kopka, M. R. Viant, *Metabolomics*, 2012, **9**, 44.

278 S. Veretnik, J. L. Fink, P. E. Bourne, *PLoS Comput. Bio.*, 2008, **4**, e1000136. 95

279 C. F. Quo, C. Kaddi, J. H. Phan, A. Zollanvari, M. Xu, M. D. Wang, G. Alterovitz, *Brief. Bioinform.*, 2012, **13**, 430.

280 J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. Bolival, N. Assad-Garcia, J. I. Glass, M. W. Covert, *Cell*, 2012, **150**, 389. 100

281 X. Sun, W. Weckwerth, *Metabolomics*, 2012, **8**, S81.

282 H. Doerfler, D. Lyon, T. Nägele, X. Sun, L. Fragner, F. Hadacek, V. Egelhofer, W. Weckwerth, *Metabolomics*, 2013, **9**, 564. 105

283 A. R. Fernie, M. Stitt, *Plant Physiol.*, 2012, **158**, 113. 35

40 110

45 115

50 120

55 125

60 130

65 135

Fig. 1: Bibliographic search in Chemical Abstracts Service (CAS) by SciFinder Scholar (August 9th, 2013). *Metabolomics* was used as a keyword in the first step (blue bars), and the search was then refined by using *Plant* in a second step (green bars).

Figure 4: Spectrum of *L. tomentosa* in negative ion mode without a matrix.

Figure 2: Scheme for carotene, xanthophyll and retinoid ionisation.

Figure 3: Spectrum of *L. tomentosa* acquired in negative ion mode with 4-NA at a concentration of 2.5 mg/ml as the matrix

Figure 5: Sketch of how different fields of metabolic research can be integrated to improve metabolic network modelling. Models based on basic metabolism knowledge codified in the genome can be used as a starting point for metabolic experiments. Experimental data should then be shared, and new inferences about the metabolic network can be made, creating a feedback loop. The cycle of modelling and experimentation enables continuous improvement of data integration and phenotype prediction.

Histórico Janus

Universidade de São Paulo
Faculdade de Medicina de Ribeirão Preto
Documento sem validade oficial
FICHA DO ALUNO

17135 - 6293113/1 - Ricardo Roberto da Silva

Sigla	Nome da Disciplina	Início	Término	Carga Horária	Cred.	Freq.	Conc.	Exc.	Situação
5935863-7/1	Elucidação Estrutural de Compostos Orgânicos (Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - Universidade de São Paulo)	08/03/2010	21/06/2010	120	0	0	-	N	Matrícula cancelada
SEL5763-1/2	Reconhecimento de Padrões (Escola de Engenharia de São Carlos - Universidade de São Paulo)	09/03/2010	21/06/2010	180	12	100	A	N	Concluída
IBI5014-2/2	Estatística Básica em Bioinformática (Curso Interunidades: Bioinformática - Universidade de São Paulo)	22/03/2010	13/06/2010	120	8	99	A	N	Concluída
IBI5025-1/1	Introdução à Computação Evolutiva (Curso Interunidades: Bioinformática - Universidade de São Paulo)	11/08/2010	27/10/2010	120	8	90	A	N	Concluída
RGM5806-8/1	Genética Quantitativa I	03/09/2010	16/12/2010	90	6	100	A	N	Concluída
IBI5016-1/1	Métodos Estatísticos em Bioinformática (Curso Interunidades: Bioinformática - Universidade de São Paulo)	10/10/2010	01/01/2011	120	8	99	A	N	Concluída
RAD5001-3/5	Didática (Curso Interunidades: Administração de Organizações - Universidade de São Paulo)	06/06/2011	29/08/2011	90	6	100	A	N	Concluída

	Créditos mínimos exigidos		Créditos obtidos
	Para exame de qualificação	Para depósito de tese	
Disciplinas:	10	20	48
Estágios:			
Total:	10	20	48

Créditos Atribuídos à Tese: 152

Observações:

1) Curso com validade nacional, de acordo com o disposto na Portaria nº 73, de 17.01.2007.

Conceito a partir de 02/01/1997:

A - Excelente, com direito a crédito; B - Bom, com direito a crédito; C - Regular, com direito a crédito; R - Reprovado; T - Transferência.

Um(1) crédito equivale a 15 horas de atividade programada.

Última ocorrência: Matrícula de Acompanhamento em 18/07/2013

Impresso em: 17/01/14 11:23:43