Revista Virtual de Química

Article

http://dx.doi.org/10.21577/1984-6835.20230066

Análise de Dados de Metabolômica em Produtos Naturais: uma Revisão-Tutorial

Analysis of Metabolomics data in Natural Products: a Tutorial Review

Naydja M. Maimone,[a,b,#] Alana K. Pereira,[c,d,e,#] Leila Gimenes,[f] Hocelayne P. Fernandes,[c,e] Taícia P. Fill,[d] Ricardo M. Borges,[g] Simone P. Lira,[a] João B. Fernandes,[c] Anelize Bauermeister[h,i,]*

[a] Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz", Departamento de Ciências Exatas, CEP 13418-900, Piracicaba-SP, Brasil
[b] Simon Fraser University, Department of Chemistry, V5A 1S6, Burnaby-BC, Canadá
[c] Universidade Federal de São Carlos, Departamento de Química, CEP 13565-905, São Carlos-SP, Brasil
[d] Universidade Estadual de Campinas, Instituto de Química, Departamento de Química Orgânica, CEP 13083-887, Campinas-SP, Brasil
[e] Leiden University, Institute of Biology, Plant Sciences, 2333 BE, Leiden-SH, The Netherlands
[f] Centro de Recursos Genéticos Vegetais, Instituto Agronômico, CEP 13075-630, Campinas-SP, Brasil
[g] Universidade Federal do Rio de Janeiro, Instituto de Pesquisas de Produtos Naturais Walter Mors, CEP 21941-599, Rio de Janeiro-RJ, Brasil
[h] University of California, Skaggs School of Pharmacy and Pharmaceutical Science, 92122, San Diego-CA, USA
[i] Universidade de São Paulo, Instituto de Ciências Biomédicas, CEP 05508-000, São Paulo-SP, Brasil.
[#] Both authors contributed equally to this manuscript.

***E-mail:** ane.meister@usp.br*

Recebido: 27 de Abril de 2023

Aceito: 30 de Setembro de 2023

Publicado online: 24 de Novembro de 2023

The incredible natural diversity that we can admire at different ecosystems is a result of genetic combinations along with complex metabolic reactions. For centuries, mankind has learned from nature and used its resources to improve the quality of our life. Therefore, the better understanding of the metabolites hole in living organisms' metabolism and/or in ecological interactions represent a potential strategy to streamline several processes under investigation, such as drug development. In this context, metabolomics has emerged as an indispensable tool to analyze the metabolites in biological systems through spectral information. Metabolomics has been applied to answer biological questions of increasing complexity, which demands robust methodologies to assure reproducibility and reliability in the results presented to the scientific community. Hence, this tutorial-review covers a step-by-step guide for metabolomics analysis, including free available tools. Data acquisition and the main steps of data processing are presented, and we also address commonly applied statistical analysis and data visualization approaches that can improve the comprehension and interpretation of complex datasets. In addition, we discuss some annotation tools and bring up good practices and how to apply them. This work provides background on resources and concepts needed from all the researchers interested in this topic.

Keywords: LC-MS/MS; molecular networking; statistical data analysis; metabolites annotation; natural products; GNPS.

1. Introdução

A Química de Produtos Naturais (QPN) é a ciência que estuda os metabólitos especializados (previamente chamados de "metabólitos secundários")[1] produzidos por organismos vivos como plantas, macro e microrganismos.[2] O estudo desses metabólitos, de suas vias biossintéticas e de seu papel no meio ambiente contribuem para uma melhor compreensão dos processos biológicos, uma vez que são essenciais para o desenvolvimento, equilíbrio, defesa e comunicação entre organismos que compartilham o mesmo habitat.[3] Particularmente, por apresentarem uma ampla diversidade e complexidade estrutural, os metabólitos especializados podem ainda exibir diversas propriedades biológicas de interesse para os setores farmacêutico,[4,5] cosmético,[6] alimentício,[7] agroquímico,[8] dentre outros.

O conhecimento prévio do perfil químico de uma amostra de origem natural é crucial em estudos de prospecção e isolamento de metabólitos de interesse. Tal compreensão permite o direcionamento do estudo e auxilia na prevenção do reisolamento de substâncias já conhecidas e indesejadas, o que se tornou um problema frequentemente observado na QPN.[9,10] Essa estratégia permite redução de tempo, de esforços e de custos gastos em projetos científicos com essa finalidade. Entretanto, o processo de desreplicação (identificação ou anotação dos metabólitos previamente descritos presentes em uma amostra), quando realizado de modo manual, exige muito tempo do pesquisador e está sujeito a diversos erros e vieses. Portanto, o desenvolvimento de ferramentas analíticas e computacionais que visam automatizar e otimizar essa etapa têm sido de extrema relevância para o aperfeiçoamento da QPN.[11,12]

Ademais, a QPN tem se beneficiado com a constante evolução das ciências "ômicas", como a genômica, a proteômica e a metabolômica.[13] Particularmente, esta última busca realizar uma avaliação abrangente qualitativa e quantitativa da dinâmica metabólica envolvida em quaisquer sistemas ou organismos vivos, em decorrência de modificações genéticas, ambientais, de estresse ou estímulo fisiopatológico, visando melhor compreendê-los.[3,14] Dessa forma, a metabolômica trata da análise de metabólitos primários e especializados dos organismos (até 2.000 Da), permitindo realizar comparações entre os perfis metabólicos de diferentes amostras.[15] Resumidamente, os estudos podem ser realizados de duas maneiras: i) quando se

faz uso de um conhecimento prévio sobre os metabólitos de interesse e é aplicado esse conhecimento durante o método de extração e toda a etapa de aquisição dos dados, adotando uma abordagem alvo (do inglês *Targeted*); ou ii) quando se tem por objetivo reunir informações sobre o maior número possível de metabólitos nos sistemas biológicos, levando em consideração todas as informações presentes nos conjuntos de dados, utilizando-se então uma abordagem não-alvo (do inglês *Untargeted*).[14,16]

Podemos citar diferentes técnicas analíticas empregadas em estudos de metabolômica. A ressonância magnética nuclear (RMN ou NMR, do inglês *Nuclear Magnetic Resonance*), por exemplo, possui papel fundamental na elucidação estrutural de metabólitos. No entanto, a espectrometria de massas (EM ou MS, do inglês *Mass Spectrometry*) tem se destacado amplamente neste amplo campo de pesquisa, e principalmente na QPN, pois permite a análise de misturas complexas com substâncias de diferentes propriedades físico-químicas[11,17]. A técnica possui flexibilidade para acoplamento com técnicas cromatográficas[18,19] e permite a detecção de metabólitos presentes em baixas concentrações devido à sua alta sensibilidade.[20,21] Além disso, EM requer quantidades muito pequenas de amostra (menos de 1 mg de extrato), o que demanda relativamente pouco material inicial de trabalho e possibilita a utilização de montantes mínimos de solventes para extração e análise. Devido a esta economia no uso de solventes, a EM é considerada uma técnica analítica "verde". [22,23]

A relativa facilidade na aquisição de dados de EM e maior acessibilidade à técnica fizeram com que ela começasse a ser empregada por diversos grupos de pesquisa empregando diferentes técnicas e diferentes equipamentos, o que levou a geração de grande quantidade de dados. Dessa forma, métodos automatizados de processamento e análise de dados de EM têm sido desenvolvidos de forma a contribuir para viabilizar e acelerar estes processos para garantir uma melhor interpretação dos resultados. No entanto, se faz necessário aos pesquisadores aprender a utilizar estas ferramentas a fim de gerar dados compreensivos que de fato reflitam o que é observado em suas amostras.[24] Por exemplo, sabemos que a utilização de ferramentas estatísticas facilita a interpretação de conjuntos dados de grande complexidade, através de verificação de tendências entre os grupos de amostras[25,26]. Todavia, a obtenção de resultados acurados requer que o processamento dos dados seja realizado de forma coerente com o tipo de dados a fim de se obter informações o mais representativas quanto possível a partir das amostras. Isso inclui, por exemplo, desconsiderar ao máximo sinais de ruído e abranger o maior número de metabólitos possíveis, mesmo metabólitos detectados em baixas intensidades.[27] Embora existam muitas plataformas e ferramentas gratuitas para o processamento de dados, o manuseio e o uso efetivo destas podem ser incipientes devido à ausência de tutoriais disponíveis, principalmente em português.

Dessa forma, o objetivo dessa revisão-tutorial é contribuir para o desenvolvimento de projetos de estudantes de graduação, pós-graduação e de pesquisadores que desejam compreender e realizar estudos utilizando uma abordagem metabolômica (ou suas ferramentas) na área de QPN. Nós discorremos sobre as diferentes formas de aquisição de dados utilizando a EM e apresentamos ferramentas para análise e processamento desses dados, além da exploração das informações por meio de análise estatísticas. Associamos, então, essas informações à construção de redes moleculares e a anotação de metabólitos, discutindo diferentes métodos desde o uso de bibliotecas de espectro como de bancos de dados de produtos naturais, empregando ferramentas *in silico*, árvores de fragmentação, subestruturas, entre outros. Por fim, abordamos também como conciliar essas informações para otimizar a interpretação das informações obtidas.

2. Aquisição dos Dados Espectrais a Partir das Amostras Biológicas

Antes de discutirmos a aquisição dos dados em si, precisamos considerar algumas etapas que são cruciais e podem influenciar diretamente nos resultados de um projeto. A amostragem, o método de extração dos metabólitos e a abordagem de controle de qualidade precisam ser cautelosamente previamente delineadas (Figura 1). Por exemplo, se analisarmos uma planta que foi coletada apenas na estação chuvosa, não obteremos dados representativos do metabolismo da espécie como um todo; nesse caso, várias amostragens precisariam ser realizadas ao longo de diferentes estações do ano, incluindo diferentes indivíduos. Outro fato importante é que cada tipo de solvente pode favorecer a extração de determinadas classes de metabólitos em detrimento de outras. A descrição detalhada dessas etapas não é o objetivo deste trabalho, mas há outras revisões sobre estes tópicos que podem ser encontradas na literatura.[14,28,29] Estes fatores irão refletir na relevância das informações adquiridas e nas possíveis conclusões sobre a importância biológica dos metabólitos no sistema analisado (Quadro 1a e b). O planejamento experimental levando em considerações essas etapas contribui para uma maior reprodutibilidade, robustez, otimização metodológica e, consequentemente, maior confiabilidade nos resultados obtidos.[30,31]

A escolha do espectrômetro de massas deve ser direcionada de acordo com o objetivo da pesquisa. Um espectrômetro de massas é composto principalmente por uma fonte de ionização e um analisador de massas. Dentre as várias fontes de ionização existentes, podemos citar a ionização por impacto eletrônico (IE ou EI, do inglês *Electron Impact*), normalmente acoplada a cromatógrafos e análise de amostras gasosas e termoestáveis – muito empregada para análise de ácidos graxos, terpenos, entre outros; e a ionização por electrospray (IES ou ESI, do inglês *Electrospray Ionization*), normalmente acoplada a

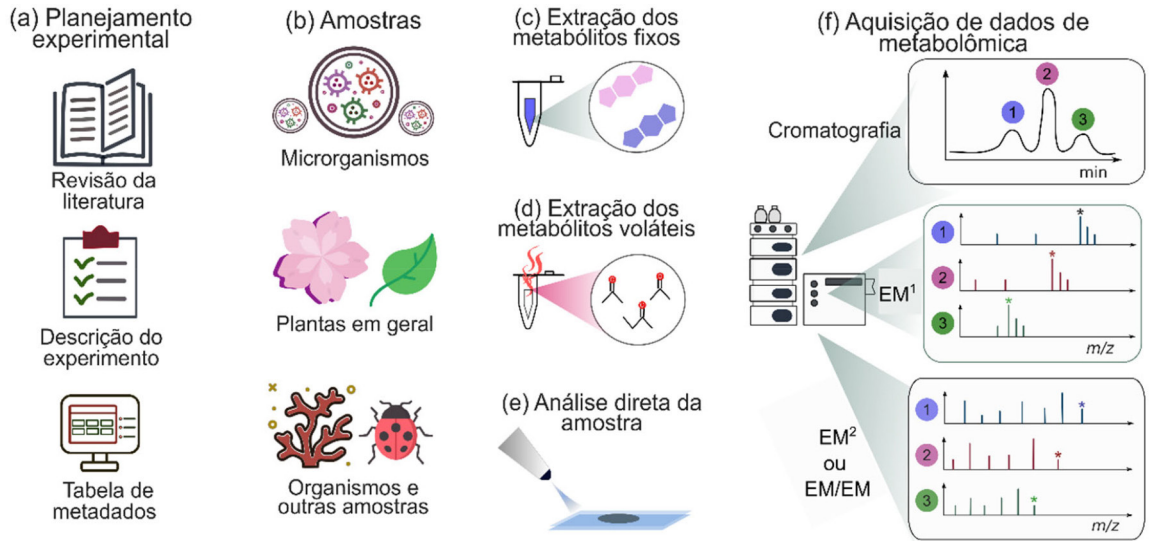

Figura 1. Esquema representativo das principais etapas envolvidas na execução de um projeto de pesquisa de produtos naturais empregando metabolômica. (a) Planejamento experimental: buscas na literatura, descrição e organização do experimento e da tabela de metadados. (b) Amostragem: definição das amostras que serão utilizadas, tais como microrganismos (bactérias, fungos etc.), plantas e outros materiais de origem biológica em geral (macroorganismos, amostras de plasma sanguíneo, urina etc.), sedimento etc. As diversas formas de extração e análise dos metabólitos por (c) extração dos metabólitos fixos, (d) voláteis ou, ainda (e) análise direta da amostra. (f) Aquisição dos dados de EM, envolvendo ou não cromatografia, seja ela líquida (CL) ou gasosa (CG), espectros de EM^1 e espectros de fragmentação ou EM^2

Quadro 1. Informações relevantes em estudos metabolômicos relacionados à amostragem e extração dos metabólitos

(a) Critérios a serem considerados na etapa de amostragem
Plantas: Efeitos sazonais, localização geográfica, umidade, órgãos e idade da planta, armazenamento e estocagem.
Microrganismos: Efeitos dos componentes do meio de cultura, quantidade de microrganismo inoculado (concentração do inóculo), tempo de incubação necessário, temperatura controlada, entre outros.
Macroorganismos e demais amostras biológicas em geral: Localização geográfica, efeitos de temperatura e clima local, fontes de alimentos disponíveis, idade do organismo, sexo, entre outros.
Réplicas biológicas ou autênticas: Consiste em analisar amostras biológicas independentes. Por exemplo, utilizar diferentes folhas de um mesmo exemplar (espécime), ou folhas de diferentes indivíduos da mesma espécie, ou ainda utilizar diversas placas de cultivo de uma mesma linhagem de microrganismo. Nesse caso, verifica-se a variância biológica dentro de um mesmo grupo amostral, realizando-se o mesmo procedimento experimental para todas as amostras.
Número de amostras: Não existe um número padrão de amostras indicado para quaisquer estudos em metabolômica. O tamanho amostral ideal depende em grande parte das variáveis intrínsecas das amostras e do efeito observado experimentalmente. Além de ser limitado pela capacidade prática de cada laboratório ou instrumentação disponível. O indicado é sempre que possível montar uma versão prévia reduzida como piloto do estudo proposto. Os dados experimentais deste piloto poderão, então, serem utilizados para se obter uma estimativa razoável do número ideal de amostras necessárias para o estudo.
(b) Etapas para a otimização da extração de metabólitos
Planejamento experimental: Neste contexto, trata-se da análise das respostas do experimento frente a variação de cada parâmetro testado, com o objetivo de se otimizar a produção, detecção ou extração de um metabólito. Pode-se testar diferentes solventes, o tempo de extração e distintos métodos de extração, por exemplo. Um bom planejamento experimental garantirá a robustez do método proposto.
Extração ou partição líquido-líquido, **extração em fase-sólida** [32–34] **e QuEChERS** (do inglês, *Quick, Easy, Cheap, Effective, Ruged and Safe*)[35–38] são métodos seletivos e eficazes no preparo de amostras, *clean-up* e redução do efeito matriz.

cromatografia líquida e utilizada para análise de amostras líquidas e/ou solúveis em solventes como metanol e acetonitrila – muito empregada em estudos de aminoácidos e metabolitos especializados em geral. Vale ressaltar aqui que o acoplamento da EM a um sistema cromatográfico permite uma melhor separação dos constituintes das amostras, e consequentemente, uma análise mais detalhada.[30] Porém, há opções para amostras ou estudos que requerem análises diretas na matriz de interesse, como estudos *in situ* ou *in vivo* (tecido biológico, placa de cultura etc.), sem a realização de extração prévia dos metabólitos.[17,39] O imageamento e visualização da distribuição espacial dos metabólitos nestes casos podem ser obtidos ao se empregar as técnicas de ionização eletrospray por dessorção (DESI, do inglês *Desorption Electrospray Ionization*),[40] e de ionização por dessorção a laser assistida por matriz (MALDI, do inglês *Matrix-assisted laser desorption/ionization*).[41]

Por ser uma fonte de ionização branda, na IES os metabólitos são geralmente detectados intactos na forma de íons adutos. Neste caso, no modo de ionização positivo

os íons são comumente detectados na forma de agregados iônicos com próton $[M+H]^+$ ou com cátions ($[M+Na]^+$, $[M+K]^+$, $[M+NH_4]^+$, etc.); enquanto que no modo de ionização negativa geralmente é detectado o íon referente a molécula sem um próton $[M-H]^-$, ou na forma de agregados iônicos com carga negativa ($[M+Cl]^-$, $[M-HCOO]^-$, etc.). A adição de ácidos e bases voláteis na fase móvel pode auxiliar na resolução dos picos cromatográficos e na ionização dos analitos. Como exemplo, os ácidos fórmico e trifluoroacético podem ser empregados no modo positivo de ionização, e hidróxido de amônio ou acetato de amônio no modo negativo.[42] O modo de ionização pode ser escolhido de acordo com a classe química dos metabólitos alvos. Alcaloides, por exemplo, ionizam melhor no modo de ionização positivo, enquanto ácidos ionizam melhor no modo negativo.

A EM detecta e analisa íons em fase gasosa, e os espectros de massas são resultados do registro da intensidade desses íons pela razão massa/carga (*m/z*). Um experimento comumente empregado em metabolômica é o experimento de fragmentação, no qual é aumentada a energia interna de um determinado íon isolado em fase gasosa (íon precursor), de forma que essa energia seja dissipada pela quebra de ligações químicas. O espectro obtido pela fragmentação (espectro de fragmentação, espectro de EM/EM, EM^2 ou tandem-EM) contém valiosas informações estruturais que são utilizadas para auxiliar na identificação dos metabólitos ou, mesmo que os dados obtidos não permitam a identificação, ainda se pode obter indícios sobre as classes químicas as quais eles pertencem, o que é útil para direcionar as etapas seguintes do trabalho.

Ainda com relação aos experimentos de fragmentação, podemos citar dois métodos amplamente utilizados em metabolômica: i) a aquisição independente de dados (DIA, do inglês *Data Independent Acquisition*) na qual o analisador de massas fragmenta todos os íons detectados no EM^1 [43,44] e ii) a aquisição dependente de dados (DDA, do inglês *Data Dependent Acquisition*), que seleciona íons específicos do EM^1 para fragmentação conforme configuração nos parâmetros do equipamento, sendo geralmente utilizado um valor base de intensidade ou número de íons por *scan*.[45–48]

Na etapa de aquisição, alguns cuidados podem ser tomados para facilitar o posterior processamento dos dados para melhorar sua robustez e qualidade, visando evitar interpretações equivocadas. No Quadro 2 apresentamos alguns desses detalhes, como por exemplo a ordem de análise das amostras, utilização de controles e réplicas, e quais controles de qualidade podem ser considerados para minimizar os erros de interpretação dos dados.

3. Como Extrair as Informações Químicas dos Dados Espectrais?

Para extrair as informações químicas presentes nos dados espectrais são necessárias várias etapas de processamento destes dados, dentre as quais algumas são opcionais (e assim sendo, sua utilização irá depender principalmente da pergunta a ser respondida e da qualidade das informações obtidas). Por exemplo, se o objetivo do estudo é realizar apenas uma análise exploratória rápida a nível de conteúdo químico para um conjunto de amostras, deverá ser suficiente empregar uma análise clássica de redes moleculares (ou *molecular networking*) em conjunto com buscas automatizadas em bibliotecas espectrais. No entanto, caso se pretenda elaborar uma comparação entre diferentes grupos

Quadro 2. Sumário de quesitos relevantes relativos ao controle de qualidade para aquisição de dados de EM

Controles de qualidade para aquisição de dados de EM[49]
Ordem de aquisição dos dados: A aquisição dos dados deve ser feita em ordem aleatória, para evitar tendências e facilitar a identificação de possíveis erros experimentais. Aleatoriedade deve ser considerada desde a etapa de cultivo e/ou coleta do material biológico. Erros inseridos por falta de aleatoriedade são conhecidos como efeito de bloco (*batch-effect*). Uma boa revisão sobre este tema pode ser encontrada na referência 50. [50]
Amostras de controle de qualidade (CQ ou QC, do inglês *Quality Control*): O CQ auxilia na verificação da robustez e reprodutibilidade do método analítico, bem como no controle da instrumentação utilizada. O CQ pode ser criado a partir de uma mistura de padrões comerciais que podem possuir variadas propriedades físico-químicas à escolha do pesquisador. Alguns trabalhos utilizam uma mistura de iguais quantidades de todas as amostras a serem analisadas com essa mesma finalidade.[51]
Amostras-branco ou controle: Informações irrelevantes e que potencialmente podem gerar interpretações equivocadas sobre os resultados, como por exemplo os sinais presentes nas amostras brancos e controles, devem ser removidos previamente à análise dos dados. Por isso se faz necessária a aquisição de dados relativos aos solventes extratores e à fase móvel empregada na cromatografia, bem como do meio de cultivo ou da matriz utilizada (em caso de amostras biológicas).
Réplicas analíticas: Consistem em analisar-se a mesma amostra mais de uma vez nas mesmas condições (mesmo equipamento analítico e mesma metodologia). Nesse caso, monitora-se a robustez analítica instrumental.
Uso de padrão interno: Em abordagens não-alvo, um ou mais compostos conhecidos (e de preferência que não estejam presentes originalmente no material estudado) podem ser adicionados às amostras em concentrações padronizadas para facilitar as etapas de normalização dos dados e aumentar a precisão da quantificação de metabólitos de interesse. Essa abordagem pode ser expandida inclusive para permitir correções de sinais entre diferentes grupos de amostras analisadas (*inter-batch correction*). Além disso, padrões isotopicamente marcados (dessa vez, de compostos que se espera que estejam presentes no material estudado) podem ser utilizados em estudos com abordagem alvo também para verificação da robustez do método de extração e análise.
Calibração: A calibração periódica do equipamento é necessária para ajustar espectrômetros de massas de alta-resolução. Isso pode ser feito com uso de um padrão de referência geralmente obtido de forma comercial. Os espectros adquiridos das amostras em estudo devem ser calibrados, e isso pode ser feito de forma manual ou de modo automático por alguns equipamentos.

biológicos, então análises mais detalhadas com suporte estatístico se farão necessárias e muitos controles podem ser requeridos.

Na literatura, alguns autores costumam dividir o processamento de dados em pré-processamento (termo utilizado para referir-se às etapas de detecção dos picos, construção do cromatograma etc.) e processamento (referindo-se às análises dos dados, como análises estatísticas, criação de redes moleculares etc.). No entanto, nós entendemos que não há a necessidade de dividir os termos empregados neste contexto e, sendo assim, somente utilizaremos o termo "processamento" neste trabalho. O fluxograma a seguir (Figura 2) apresenta as principais etapas de processamento em um experimento de metabolômica, as quais serão mais detalhadas nos próximos tópicos.

3.1. Como tornar seus dados mais acessíveis (conversão dos dados brutos)

Para o processamento e análise dos dados em plataformas públicas, os arquivos precisam estar em formato aberto e compatível para que os dados possam ser acessados. Os *softwares* de equipamentos de EM geralmente geram os dados em um formato específico de cada empresa (.wiff, .d, .raw, .lcd etc.) e, consequentemente, muitas vezes esses dados são lidos única e exclusivamente por seus respectivos *softwares*. Por isso, há a necessidade de conversão dos arquivos originais para formato aberto, ou seja, um formato que permita sua leitura por diferentes plataformas. *Softwares* como o MSConvert, da ProteoWizard, podem ser utilizados para essa finalidade. Mais detalhes para conversões de dados são encontrados na documentação da plataforma *Global Natural Products Social Molecular Networking* (GNPS).[52–54]

3.2. Como organizar as informações a respeito das amostras (metadados)

A organização das informações referentes às amostras é imprescindível para a otimização das análises e interpretação dos resultados. A tabela de metadados contribui para a descrição detalhada das informações dos dados, podendo incluir os nomes dos arquivos, identificação das amostras, nome das espécies dos organismos, local de coleta, metodologia e solventes de extração, equipamentos e métodos usados para a aquisição de dados, atividade biológica associada, doença associada, dieta do organismo

Figura 2. Fluxograma com descrição geral do processo de análise de dados de metabolômica na QPN

estudado etc. Organizar uma tabela de metadados é uma tarefa que demanda tempo, atenção e padronização. No entanto, seu emprego apresenta vários benefícios e, portanto, recomendamos sua aplicação. Os metadados facilitam o trabalho dos colaboradores do projeto, aceleram a análise dos dados na maioria das plataformas públicas, são compatíveis com diversas ferramentas de metabolômica, auxiliam no direcionamento das análises estatísticas, entre outros fatores.

Para ressaltar a importância da padronização da descrição dos metadados, vamos considerar aqui um exemplo hipotético onde, em um único laboratório, um aluno #1 descreve 'acetato de etila' para o solvente utilizado na extração dos metabólitos, enquanto um segundo aluno #2 descreve 'AcOEt', um terceiro aluno #3 descreve 'AE' e um quarto aluno #4 simplesmente não anota em lugar algum qual foi o solvente utilizado em suas extrações. Nesse caso, como é possível saber concretamente se os dados são comparáveis ou não? Vamos agora ampliar a escala de comparação e pensar em dados obtidos por diferentes grupos de pesquisa, ou ainda, grupos de diferentes lugares do mundo: como seria possível comparar dados? Com o crescente aumento da interdisciplinaridade e internacionalização dos projetos de pesquisa, cada vez mais faz-se necessário uma padronização na descrição dos dados. Neste contexto, vamos aproveitar para mencionar o crescente aumento de bancos de dados públicos,[55] onde os dados podem ser acessados e reanalisados por diferentes pesquisadores ao redor do mundo. A proposta é promissora, mas como comparar dados se não há metadados correspondentes, ou se estes não são compreensíveis? Este desafio fez surgir a iniciativa para criação do ReDU (do inglês *Reanalysis of Data User Interface)*, um modelo geral para descrição dos metadados que pode ser empregado em qualquer laboratório,[56] visando facilitar o compartilhamento e comparação de dados gerados por diferentes grupos de pesquisa. Cabe ressaltar, no entanto, que o formato atual do ReDU ainda é relativamente complexo por não ser exclusivamente voltado a QPNs, mas ainda assim é a melhor iniciativa que temos neste sentido no momento.

3.3. Como gerar uma tabela de íons a partir dos dados espectrais

O processamento de dados de EM obtidos em experimentos de metabolômica constitui um grande desafio, porém é crucial para permitir interpretações biológicas.[57] Esse grande desafio está relacionado à alta complexidade dos dados obtidos. Várias plataformas têm sido desenvolvidas e implementadas com o intuito de melhorar e facilitar esta etapa, como OpenMS[58,59], MZMine[60,61], MSDial[62], XCMS[63] e mzTab-M[64], todas de acesso público e com alguma interface gráfica. Ferramentas baseadas nas linguagens R ou Python também são públicas, porém demandam certo conhecimento de informática, o que pode ser impeditivo para alguns estudantes e pesquisadores. Outros *softwares* comerciais de empresas podem ser utilizados com a mesma finalidade, porém não são de acesso gratuito. Neste tutorial, discutiremos brevemente as principais etapas empregadas pelo *software* MZMine 2.53,[61] e mais detalhes sobre o processamento podem ser encontrados em Borges *et al.*, 2022.[65] Optamos por detalhar especificamente sobre o MZmine devido a este ser um dos *softwares* mais utilizados, entre os de uso aberto, pela comunidade brasileira de QPN. Além disso, o MZmine em particular é uma ferramenta que possibilita o acompanhamento dinâmico do processamento através da pré-visualização de como os parâmetros selecionados atuarão diretamente nos dados brutos, como por exemplo mostrando o que será considerado como ruído (e assim descartado) a partir da determinação da intensidade mínima dos íons para um conjunto de dados. A versão 3 do software recentemente foi disponibilizada pelo grupo desenvolvedor, trazendo novas funções interativas. A Figura 3 ilustra as principais etapas empregadas no processamento de dados de EM no MZmine.

Antes de iniciar o processamento, recomenda-se visualizar o cromatograma de íons totais (CIT ou TIC, do inglês *Total Ion Chromatogram*) dos arquivos originais e verificar algumas características gerais dos dados, sendo elas: (i) o nível do ruído dos sinais, identificando a intensidade dos menores sinais do cromatograma para evitar a exclusão de sinais relevantes de intensidade baixa; (ii) verificar a quantidade de pontos (*scans*) que formam os picos cromatográficos; e (iii) o tempo de duração dos sinais cromatográficos (de maior e menor duração). Estes parâmetros serão importantes para direcionar várias etapas do processamento dos dados (Figura 3a). Sugerimos também a utilização da ferramenta *Show Preview* para conferência do que está sendo selecionado de acordo com os parâmetros estabelecidos.

3.3.1. Detecção dos sinais dos íons e filtro do sinal-ruído

A detecção dos íons é baseada principalmente na definição do nível de ruído (*noise level*). É nessa etapa que é definido o que será considerado como sinal e o que será excluído das etapas posteriores da análise por ser considerado ruído (Figura 3b). Se for do interesse do pesquisador, aqui também é possível fazer a detecção dos espectros de EM^2 que devem ser considerados na análise. Esses dados (de EM^2) podem ser posteriormente utilizados em análises específicas como para construção de redes moleculares (sobre as quais falaremos mais adiante nesse tutorial) ou para anotação das *features*. Os dados de EM podem estar organizados no modo centróide ou perfil. O modo centróide reduz os pontos adquiridos de *m/z* à uma única representação do ponto máximo de um espectro, enquanto o modo perfil mostra todos os pontos registrados pelo EM. Os dados no modo centróide são mais concisos que os do modo perfil, e são utilizados para a realização dessa etapa do processamento dos dados.[66]

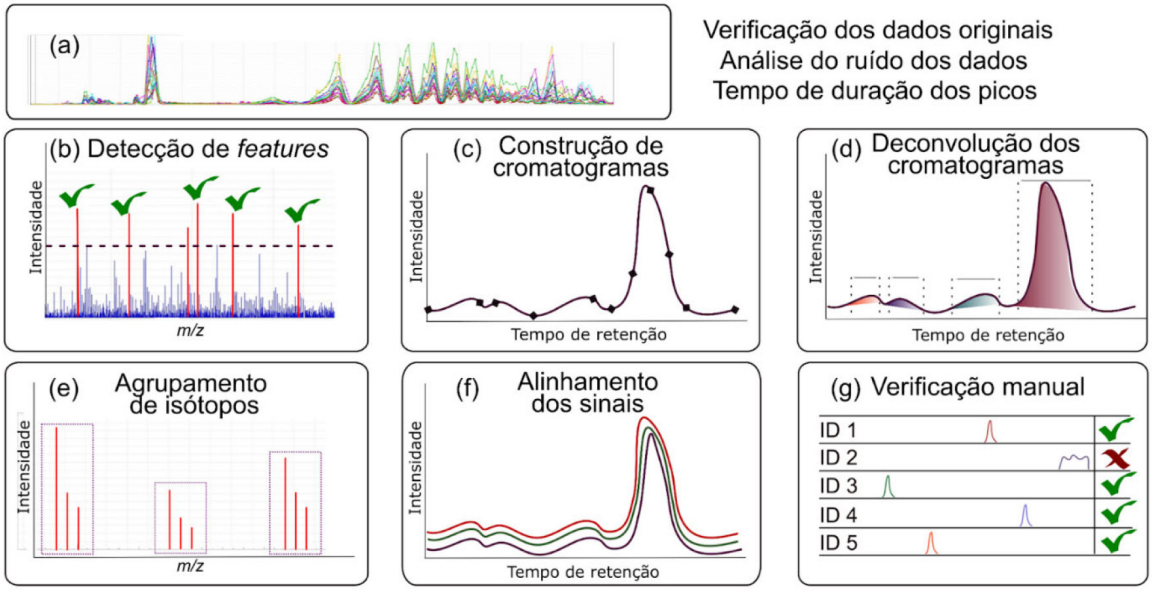

Figura 3. Principais etapas do processamento dos dados de CL-EM e/ou CL-EM/EM. (a) Análise inicial do dado original, com foco para informações relevantes como o nível do ruído, altura do pico e tempo de retenção dos picos. (b) Detecção dos *features* (íons de EM^1 e EM^2). (c) Agrupamento dos *features* detectados com base nos dados de EM^1. (d) Deconvolução dos cromatogramas. (e) Agrupamento dos isótopos. (f) Alinhamento dos picos. (g) Verificação manual dos dados obtidos através do alinhamento

3.3.2. Construção dos cromatogramas de íons detectados

Nessa etapa os íons de uma mesma razão massa carga, detectados na etapa anterior, serão agrupados em um cromatograma (Figura 3c). Esta etapa determina sinais verdadeiros presentes nos dados através da definição de intensidade e altura mínima que caracteriza um pico cromatográfico. Além disso, o número de pontos caracterizará o formato e a resolução do pico, permitindo assim a construção dos picos reais no cromatograma e descartando-se sinais-ruído (que podem passar pelo filtro discutido no tópico anterior).

A tolerância de massa (*m/z*) define o erro de massa (em Da ou ppm) para considerar diferentes *features* como correspondendo ou não ao mesmo sinal. Aqui, isóbaros (metabólitos de mesma *m/z*) de diferentes tempos de retenção, ou seja, correspondentes a substâncias diferentes, serão agrupados em um mesmo cromatograma, e então se faz necessária uma próxima etapa do processamento, a deconvolução.

3.3.3. Deconvolução dos cromatogramas

A etapa de deconvolução é importante para separar os *features* que constituem os cromatogramas (Figura 3d).[60] Em dados de CL-EM, a deconvolução classifica como diferentes *features* íons com mesma massa (isóbaros) que apresentam diferentes tempos de retenção. Já em dados de CG-EM a deconvolução, que deve ser composta por uma etapa de deconvolução cromatográfica e outra espectral, consiste em separar computacionalmente os componentes co-eluídos, fornecendo um espectro puro com a contribuição de cada componente para o cromatograma de íons extraídos (CIE ou EIC, do inglês *Extracted Ion Chromatogram*). Diferentes algoritmos podem ser usados para deconvoluir os picos cromatográficos, e como exemplo podemos mencionar alguns dos mais utilizados: o *local minimum search*, ou uma opção um pouco mais simples: o *baseline cut-off*. Nesta etapa é realizada a associação dos dados de EM^1 com os de EM^2, utilizando a função *scan pairing* do MZMine. A descrição destes métodos pode ser encontrada em Borges *et al.*, 2022.[65]

3.3.4. Agrupamento dos isótopos

A identificação de isótopos é relevante para a remoção de informações redundantes, permitindo o agrupamento desses íons como um único *feature* (Figura 3e). Após essa etapa, recomenda-se verificar a tabela de dados e buscar por íons específicos e de interesse. Essa verificação confere a possibilidade de retornar às etapas anteriores e mudar os parâmetros aplicados caso esses íons tenham sido perdidos durante o processamento.

3.3.5. Alinhamento dos sinais

Durante a aquisição de dados, pequenas oscilações na fase móvel, temperatura ou pressão podem acarretar variações nas corridas cromatográficas. Para corrigir essa limitação, é possível utilizar algoritmos capazes de alinhar os *features* detectados entre diferentes amostras (Figura 3f), como o *join aligner*, *ADAP aligner* e *RANSAC*.[65]. Um dos parâmetros para o alinhamento engloba o peso de importância utilizado para *m/z* e para o tempo de retenção (TR ou RT, do inglês *Retention Time*) que pode variar de acordo com o método e equipamento utilizado. Por exemplo, se os dados forem adquiridos em um equipamento de alta resolução, a *m/z* deverá ter o maior peso. Nessa parte, uma tabela com todos os *features* alinhados das amostras será construída e poderá ser manualmente inspecionada (Figura 3g).

Após a verificação da tabela alinhada, é possível avaliar se há a necessidade de aplicação de etapas adicionais de processamento, por exemplo: i) a criação de filtros de exclusão de sinais detectados nos brancos dos solventes e/ou meio de cultivo (que não contém sinais cromatográficos relevantes para a análise e sim interferentes) ou mesmo a criação de um filtro que selecione apenas os *features* com alta reprodutibilidade entre as amostras, ii) *gap-filling* para recuperar sinais perdidos durante alguma etapa do processamento devido ao não cumprimento de algum dos parâmetros utilizados, iii) a normalização dos sinais visando uma análise quantitativa relativa ao padrão interno ou CQ e, iv) é ainda possível realizar a identificação automática de adutos comuns em EM e íons produtos advindos da fragmentação *in-source*, como perdas de água, que podem acontecer na fonte de ionização.[67,68] O *software* MZMine permite que as informações usadas no processamento sejam salvas como projetos, podendo revisitá-las sempre que necessário.

4. Análises Estatísticas

Diferentes métodos estatísticos podem ser empregados para auxiliar na análise dos dados de metabolômica e direcionar futuras etapas das pesquisas. Análises uni e multivariadas, por exemplo, são amplamente empregadas para indicar a presença de *outliers* que podem ser removidos do conjunto de dados, ou ainda para indicar a presença de subgrupos em um grupo de amostras pelo emprego de métodos que permitam avaliar a similaridade e/ou dissimilaridade entre elas. Neste caso, temos como exemplos a análise multivariada de dados com estudos de análise da componentes principais (ACP ou PCA, do inglês *Principal Component Analysis*) e mínimos quadrados parciais (PLS, do inglês *Partial Least Squares*), ou ainda análise discriminante por mínimos quadrados parciais (PLS-DA, do inglês *Partial Least Squares - Discriminant Analysis*).[26,69–71] No entanto, análises estatísticas geralmente abrangem métodos conceitualmente complexos e que demandam conhecimento em computação, e por isso plataformas de análise voltadas à metabolômica, com interface gráfica e que propiciem o emprego desses métodos de modo mais simples podem facilitar e agilizar a análise de dados por pesquisadores da área. Sendo assim, neste tutorial nós optamos por apresentar algumas das funções mais comumente utilizadas da plataforma aberta MetaboAnalyst (https://www.metaboanalyst.ca/),[26] amplamente empregada tanto por iniciantes quanto por pesquisadores experientes em QPN.

Além das análises estatísticas, a plataforma MetaboAnalyst oferece diferentes opções para análise de dados obtidos em estudos de metabolômica, como análises de biomarcadores, processamento de dados de CL-EM, entre outros.[72,73] Os dados precisam estar organizados com um número mínimo de três réplicas por grupo amostral, previamente organizadas no formato adequado (tabela com valores separados por vírgula em formato .csv), correlacionando cada grupo avaliado com as amostras as quais eles abrangem. O MZmine possui um módulo de exportação de dados para O MetaboAnalyst, que gera uma tabela formatada especificamente para a plataforma. Aqui é importante ressaltar que fatores relativos ao controle de qualidade para aquisição de dados de EM (citados no Quadro 2) e características intrínsecas a cada experimento devem ser considerados. Por exemplo, se suas amostras são plantas coletadas em diferentes épocas do ano e, portanto, possuem tempos de armazenamento diferentes, é indicado realizar uma comparação de presença e ausência de metabólitos ao invés de se considerar as variações de seus valores quantitativos entre as amostras (uma vez que neste caso estes valores podem apresentar variações não naturais devido a possível degradação). De modo similar, caso se esteja trabalhando com metabólitos extraídos de meios de cultivo de microrganismos, é essencial obter dados de brancos dos meios de cultivo para que estes possam ser excluídos da tabela de dados antes da realização das análises estatísticas.

Para gerar as figuras apresentadas neste tutorial, nós utilizamos dados de dois diferentes microrganismos (A e B), contendo seis réplicas cada. Os dados foram previamente processados no MZmine, e então a tabela gerada foi importada ao MetaboAnalyst para análise, e cada uma das etapas empregadas será detalhada a seguir (os resultados estão apresentados na Figura 4).

4.1. Checagens iniciais e filtragem dos dados

Os dados exportados do MZMine para O MetaboAnalyst têm uma organização que compreende cada amostra como uma coluna da tabela, às quais são inseridos os parâmetros de agrupamento definidos pelo pesquisador, e os valores de quantificação dos *features* em cada uma das linhas. O MetaboAnalyst exige ao menos três réplicas para cada grupo pré-definido pois esse critério assegura robustez das análises experimentais sobre o sistema biológico. Caso em determinado projeto não seja possível se obter réplicas autênticas das amostras, uma sugestão seria criar pequenos grupos a partir de amostras que possuam perfis químicos semelhantes entre si, o que deve feito com extremo cuidado.[26]

Os valores correspondentes à área ou intensidade do pico de certos íons podem ser iguais a zero ou caracterizar uma célula em branco (*missing values*) para algumas amostras, indicando que o íon não foi detectado naquela amostra. Entretanto, valores nulos podem dificultar a normalização matemática e favorecer equívocos na interpretação dos dados. Por padrão, O MetaboAnalyst substitui os zeros ou valores em branco por ⅕ do valor mínimo de cada *feature* dentro do conjunto de dados. Em caso de baixa repetibilidade dos *features* por réplicas de amostras, essas informações podem ser deletadas da tabela de dados ou se pode utilizar filtros de análises que excluam *features* não

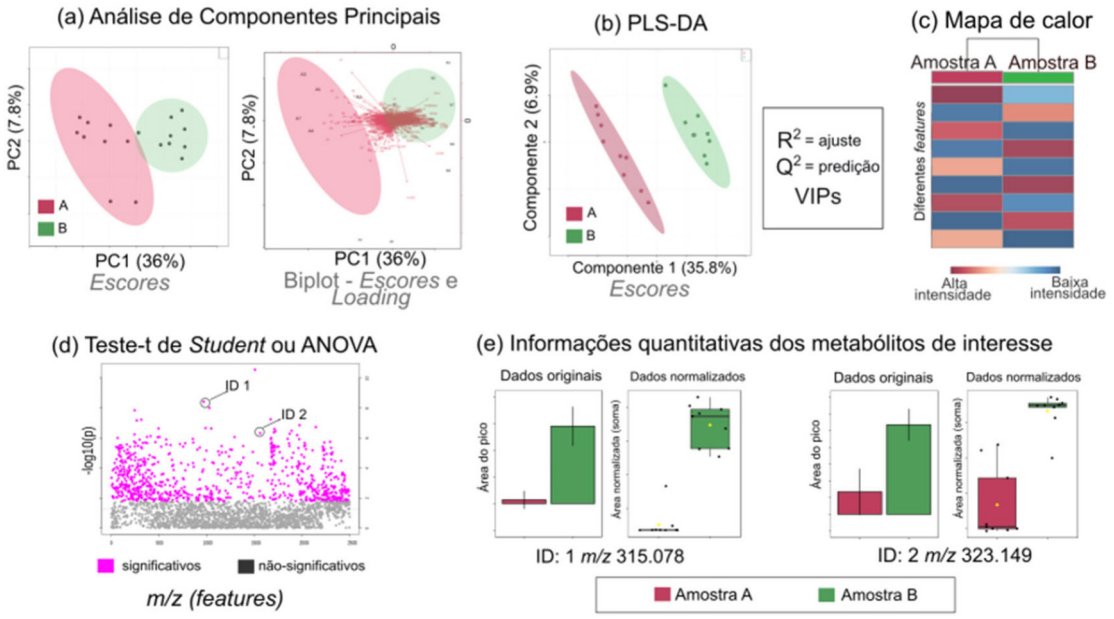

Figura 4. Resultados das análises realizadas na plataforma *MetaboAnalyst* com os grupos de amostras referente a dois diferentes microrganismos: (a) análise de componentes principais (ACP) com o gráfico dos *escores* (amostras) e *loadings*, referente aos *features*. (b) Análise discriminante de mínimos quadrados parciais (PLS-DA) que devem ser verificadas pelos valores de R^2 (capacidade de ajuste do modelo) e Q^2 (capacidade de previsão do modelo), além de verificar a lista de *features* significativos pelas variáveis importantes na projeção (VIPs). (c) *Mapa de calor* com as informações da análise de agrupamento hierárquico (AAH ou HCA, do inglês *Hierarchical Cluster Analysis*) e a distribuição dos *features* correlacionados às amostras. (d) Análises univariadas como teste-t *Student* ou ANOVA com os íons significativos da análise, caracterizados por IDs (que caracterizam os *features*). (e) Informações quantitativas de *features* significativos ou selecionados mostrando por exemplo a distribuição entre os grupos de amostras

representativos, contribuindo para que somente íons com reprodutibilidade sejam considerados (conforme discutido no tópico 3.3.5). O MetaboAnalyst apresenta a opção de filtro de dados (*data filtering*) para facilitar a exclusão de dados não representativos, baseando-se nos CQ amostrais ou em filtros estatísticos.[74]

4.2. Normalização dos dados

A normalização dos dados permite a correção de possíveis alterações na quantificação das *features* provenientes da análise de amostras com concentrações diferentes, erros experimentais ou analíticos. Os procedimentos de normalização podem ser divididos em três diferentes categorias: normalização por amostra, transformação dos dados e escalonamento dos dados.[75] As decisões sobre a aplicação destas funções (sendo que mais de um procedimento pode ser realizado sobre o mesmo conjunto de dados) dependem das características das amostras a serem analisadas, assim como todas as etapas prévias de preparo e da aquisição dos dados. Geralmente este tipo de procedimento é realizado para melhor adequar os dados à posteriores análises de inferência e para melhorar a visualização e facilitar a interpretação dos gráficos a serem gerados.

De maneira resumida, a normalização por amostra se baseia na normalização de todo o conjunto de dados por parâmetros como a média do conjunto, pela soma, com base em um padrão interno ou amostra de referência (CQ), entre outros, de forma que os dados obtenham uma distribuição Gaussiana.[76] A transformação de dados é realizada através da aplicação de uma função matemática a todos valores em um conjunto de dados, reescalando-os de modo não linear e de forma a deixar sua distribuição mais simétrica,[77] o que é recomendado para reduzir o efeito de análises por grupos (*batch-effects)*. No MetaboAnalyst os dados podem ser transformados por transformação logarítmica (base 10), raiz quadrada ou raiz cúbica. Já o escalamento dos dados contribui para ajustar as diferenças na quantificação dos metabólitos, considerando um fator de escalonamento para relativizar estes valores.[77] Se as variáveis possuem importâncias diferentes, geralmente se realiza a centragem na média dos dados (*mean-centering*). Quando todos os *features* têm a mesma importância, a normalização não pode alterar os valores de forma significativa, e dessa forma auto escalamento (*auto-scaling*) é recomendado, pois assim os dados centrados na média são divididos pelo desvio padrão de cada variável. Ainda, pode ser realizado o escalonamento de Pareto (*Pareto scaling*), que divide os dados centrados na média pela raiz quadrada do desvio padrão de cada variável, sendo assim uma normalização com pouca alteração dos dados originais.[69,77]

A plataforma MetaboAnalyst permite que as diferentes alternativas de normalização oferecidas sejam testadas de modo prático, facilitando a verificação de qual(is) melhor se adequa(m) aos dados. Depois da normalização, a plataforma realiza as análises estatísticas uni- e multivariadas que podem ser acessadas *online* e, caso o usuário deseje, é possível ainda exportar as linhas de código executadas (linguagem R) para que alterações manuais sejam realizadas e as análises refeitas localmente.

4.3. Análises multivariadas

Quando duas ou mais variáveis são medidas durante um experimento, os dados resultantes são chamados de multivariados.[78,79] No caso de metabolômica, o número de variáveis pode facilmente chegar a dezenas de milhares. As ferramentas de análise para esses dados têm por objetivo a redução da dimensionalidade dos dados sem perder informações importantes das amostras. Em geral, as análises são aplicadas para resolver problemas em estudos com grandes matrizes de dados no qual ocorrem diversos tipos de efeitos metabólicos, como amostras de sistemas biológicos de origens naturais.[80] Essas análises podem ser não-supervisionadas, com objetivo exploratório, ou supervisionadas, quando já se conhece grupos específicos entre as amostras (Figura 4a).

4.3.1. Análises não supervisionadas

Análises não supervisionadas de dados são utilizadas quando se deseja verificar a similaridade ou dissimilaridade entre as amostras, sendo as análises de coordenadas principais (ACoP ou PCoA, do inglês *Principal Coordinate Analysis*), um dos métodos amplamente empregados para esse fim.[81,82] A ACoP permite uma visualização gráfica dos dados em 2D ou 3D utilizando novos eixos de projeção onde as variâncias são maximizadas, correlacionando as amostras às variáveis analisadas (*features* ou *loadings*).[81] Embora não esteja disponível na plataforma MetaboAnalyst, tal a análise pode ser acessada na plataforma GNPS após a utilização das ferramentas de redes moleculares.[83] Outra alternativa bastante utilizada em metabolômica é a ACP que busca comparar os dados multivariados através de uma representação visual. Na Figura 4a podemos ver uma ACP onde claramente ocorre a formação de dois grupos distintos formados pelos dois microrganismos que usamos, com as réplicas de cada grupo apresentando alta similaridade entre si. A análise de agrupamento hierárquico, também bastante empregada em metabolômica, é um método para reconhecimento de padrões onde os resultados são apresentados na forma de árvore hierárquica (dendrograma) ou grupos (*clusters*) de similaridades.[84,85] Tais métodos são bastante úteis para a detecção de *outliers*, amostras com perfil químico muito discrepante diante as demais réplicas (que podem provir de possíveis erros de organização ou aquisição do material analisado, ou de contaminações em meios de cultivo de microrganismos, por exemplo). Essas amostras consideradas *outliers* podem ser então excluídas das análises, caso o pesquisador julgue necessário.

4.3.2. Análises supervisionadas

A PLS-DA contribui para a distinção entre os grupos já conhecidos, pois ele associa aos dados originais à uma nova matriz de informações, relacionando classes às amostras e variáveis, que passam a se chamar fatores ou variáveis latentes (Figura 4b).[69,86] A análise contribui para a seleção das variáveis responsáveis pela separação entre os grupos e ajuda em processos de classificação.[87,88]

A análise de PLS-DA permite a avaliação do modelo de previsão, oferecendo valores de Q^2 relacionado a capacidade de prever a qual grupo a amostra pertence e R^2 relacionado ao ajuste do modelo, ou de como os pontos distribuídos no gráfico tem a tendência a formar uma reta. Esses parâmetros também englobam as informações sobre as variáveis importantes para a construção do modelo, ou seja, aquelas que mais contribuem com a separação dos grupos. Essas informações podem ser correlacionadas através da importância da variável na projeção (VIP, do inglês *Variable Importance in Projection*) que descreve uma estimativa quantitativa do poder discriminante de cada *feature,* elencando quais os possíveis metabólitos diferenciais entre os grupos.[25]

Ainda na PLS-DA, para conjuntos de dados relativos a grandes quantidades de amostras é possível realizar o teste de permutação dos dados, que verifica se o modelo possui uma boa capacidade preditiva ou classificatória mesmo aplicado de forma aleatória. Este teste pode ser realizado com centenas ou milhares de permutas, sendo que para cada uma delas um modelo PLS-DA é construído com os dados permutados. Um valor de p é estimado e indica quantas vezes os modelos permutados foram melhores que aquele com as informações originais, assim espera-se que seja menor ou igual a 5%.[79]

Diversas outras abordagens de ferramentas de análises supervisionadas podem ser realizadas na plataforma MetaboAnalyst, como por exemplo: i) mapas de calor (no inglês, *Heat-maps*) que associam dados da AAH com as informações dos *features* correspondentes às amostras (Figura 4c); ii) mapas de correlação (no inglês, *Correlation Maps*) entre amostras e variáveis, que visam verificar a similaridade entre as réplicas por exemplo em formas de mapas de calor, ou ainda iii) florestas de decisão aleatória (no inglês, *Random Forests)* que consistem em uma método de aprendizado de máquina (*machine learning*) amplamente utilizado para identificar *features* que melhor classificam os dados em diferentes grupos, fornecendo os erros de classificação através de tabelas de contingência.[89] A plataforma também possibilita a alteração simples da escolha de parâmetros e algoritmos, e fornece informações adicionais sobre as diferentes análises e referências necessárias para entender melhor as diversas ferramentas oferecidas.[72,90] Nesse caso, o mais recomendado é testar e verificar a melhor abordagem de análise para os seus dados.

4.4. Análises univariadas

As análises univariadas são voltadas para a avaliação dos *features* (metabólitos) individualmente. Elas podem ser aplicadas diretamente para determinar a presença de biomarcadores para certos grupos amostrais ou utilizadas para testar a significância de *features* previamente selecionadas em análises multivariadas.[91] No contexto das

análises univariadas, primeiramente é necessário verificar a homogeneidade dos dados. No MetaboAnalyst é possível acessar essa informação na função *generate report* na seção *download*. Assim, pode-se optar por aplicar testes estatísticos específicos para casos de distribuição paramétrica (curva de distribuição dos dados normal) ou não-paramétrica (curva de distribuição assimétrica). De modo geral os testes paramétricos são mais robustos e, portanto, recomenda-se que os dados sejam normalizados antes das análises (ver tópico 4.2). Em todo caso, a plataforma oferece opções de testes para ambos os casos, a escolha do pesquisador. No âmbito das análises paramétricas, O MetaboAnalyst oferece por exemplo o teste-T de *Student,* a ser empregado apenas para comparação entre dois grupos, e a análise de variância ANOVA, utilizada para comparação da variação das *features* entre diversos grupos de amostras (Figura 4d).[92] As opções não paramétricas a estas análises são, respectivamente, teste de Kruskal-Wallis e teste de Wicoxon. De modo bastante resumido, estes testes resultam em um valor numérico (chamado *F* para ANOVA, *t* para t-Student, *H* para Kruskal-Wallis etc.) que é confrontado com valores tabelados (específicos de cada teste) que associam parâmetros de nível de probabilidade e graus de liberdade. Caso o valor obtido na análise seja menor que o listado na tabela para os parâmetros do estudo, considera-se a hipótese de não-significância estatística (H_0), ou seja, o valor não pode ser considerado diferente do que se observaria caso as amostras avaliadas fossem iguais. Ao contrário, se o valor for superior ao tabelado, considera-se que há significância estatística (H_1) e que as amostras não podem ser consideradas iguais. Para uma discussão mais aprofundada sobre as particularidades das análises aplicadas em estatística univariada e suas aplicações, recomendamos Sokal e Rohlf, 2012.[93] Com base nestes métodos, os íons significativos podem ser avaliados individualmente de acordo com sua quantificação, baseada na área do pico, entre as amostras (Figura 4e).

É importante citar que análises univariadas aplicadas em conjuntos de dados grandes, como os que usualmente observamos em metabolômica, podem gerar um elevado número de falsos positivos. Isso ocorre pois cada *feature* é submetida a um teste estatístico próprio, e quanto maior o número de testes realizados, maior a chance randômica da ocorrência de erros. Este fato é conhecido como problema dos testes múltiplos (do inglês, *Multiple Testing Problem*). Para lidar com isso, determinados métodos podem ser aplicados para a correção dos *p*-valores inicialmente obtidos, como os testes de Bonferroni e de correção de taxas de falsas descobertas (FDR, do inglês *False Discovery Rate*)[94,95], sendo este último integrado ao MetaboAnalyst.

5. Redes Moleculares Facilitam a Visualização dos Dados

A aquisição de dados por EM pode chegar a gerar milhares de espectros em minutos,[17,96] resultando em conjuntos de dados grandes demais para serem analisados somente de modo manual. Um outro grande desafio na área de metabolômica é encontrar o maior número de informações a respeito dos dados coletados pela EM (que podem abranger análogos desconhecidos de compostos já descritos, compostos totalmente desconhecidos, adutos de ionização e íons oriundos de fragmentação *in-source*), o que limita a interpretação global dos dados coletados.[97,98] Para facilitar a organização e principalmente a visualização dos dados de EM, uma estratégia de criação de redes moleculares foi introduzida em 2012, em um estudo que demonstrou a utilidade desta abordagem para compreensão do perfil metabólico de diversos microrganismos.[99] Desde então, este tipo de análise vem recebendo destaque em diversas áreas de conhecimento, como por exemplo na pesquisa de produtos naturais e busca por novos potenciais fármacos, no estudo de interações ecológicas, na avaliação do metabolismo de medicamentos, entre muitos outros.[100–105]

A construção de redes moleculares caracteriza uma abordagem de organização de dados de EM com base na similaridade estrutural de compostos; uma vez que substâncias com estruturas semelhantes compartilham padrões de fragmentação (EM^2) semelhantes, estes espectros serão agrupados nas redes formando famílias moleculares, o que facilita a exploração, inspeção e anotação dos dados.[106] Esta técnica se fundamenta em um algoritmo de alinhamento de espectros baseado no cosseno. Os íons fragmentos, assim como suas respectivas intensidades, são transformados em escala vetorial, e o cosseno do ângulo formado do vetor resultante é utilizado para fim de comparação e agrupamento dos espectros. Como resultado dessas comparações, temos nas redes moleculares a representação de cada *feature* como um nodo, e as ligações entre nodos são indicativas de compostos estruturalmente relacionados com valor de cosseno iguais ou superiores ao patamar previamente definido.[107]

Atualmente, a plataforma GNPS permite a criação de redes moleculares a partir dos *workflows* "clássico" e "*feature-based molecular networking*" (FBMN).[67] No primeiro caso, arquivos em formatos abertos (como .mzML, .mzXML ou .mgf) devem ser previamente adicionados ao servidor da plataforma e os espectros de EM^2 são automaticamente extraídos e agrupados através do algoritmo MS-Cluster, que gera espectros de consenso para *features* de mesmo íon precursor e íons fragmento similares (Figura 5a).[108] Uma das vantagens deste método é que ele não requer que o pesquisador utilize outras ferramentas para processamento prévio dos dados. Além disso, geralmente íons minoritários também são detectados nas amostras, uma vez que a análise considera todos os espectros de fragmentação presentes nos dados de $CL-EM^2$. No entanto, como desvantagem, o MS-Cluster não considera o TR para gerar os espectros de consenso e, portanto, pode agrupar espectros referentes a diferentes compostos em determinadas ocasiões (como quando há

certos tipos de isômeros nas amostras) , ou ainda apresentar na rede molecular mais de um nodo referente a espectros de um mesmo íon.

Para gerar o FBMN são utilizados dados previamente processados (ver tópico 3), o que permite caracterizar íons em diferentes tempos de retenção como diferentes *features* (o que propicia a diferenciação de isômeros, por exemplo).[67] Uma outra vantagem desse método é a possibilidade da incorporação de dados quantitativos (área do sinal cromatográfico), o que facilita a correlação deste tipo de análise com análises estatísticas uni e multivariadas. No entanto, como o processamento inclui várias etapas, é muito comum haver perda de íons minoritários no conjunto de dados.

Recentemente, outros métodos para comparar a similaridade espectral de íons em grandes conjuntos de dados vem sendo criados, como por exemplo o Spec2Vec[107], que também se encontra disponível na plataforma GNPS. Esta ferramenta computa os *scores* de similaridade com base em uma abordagem de aprendizagem de máquina não supervisionada, caracterizando uma nova abordagem de comparação entre espectros de EM^2. O Spec2Vec *score* tem demonstrado eficácia superior ao *score* de cosseno em determinados casos, levando a uma maior acurácia nos resultados.[111] Para uma melhor visualização e edição das redes moleculares, independente do *workflow* utilizado para sua criação, o uso do *software* Cytoscape é recomendado.[112]

6. Anotação dos Metabólitos

Enquanto a elucidação estrutural de substâncias novas presentes em amostras biológicas comumente exige a utilização de múltiplas técnicas analíticas, sem contar as inúmeras etapas de fracionamento para o isolamento, a anotação de metabólitos conhecidos nos extratos brutos pode ser realizada através da comparação dos dados espectrais experimentais com os dados de referência disponíveis.[113] Em metabolômica, em especial na área de produtos naturais, detectar metabólitos previamente descritos logo nas etapas iniciais de um trabalho é algo que pode diminuir significativamente os custos, tempo e esforço necessários para os estudos.[114] Atualmente, o processo de anotação pode ser realizado de forma automatizada em certas plataformas, no entanto uma inspeção minuciosa dos resultados deve ser realizada para cada metabólito anotado. Além disso, devemos recordar que a busca em bibliotecas espectrais está limitada às informações nelas disponíveis, e por isso utilizar mais de uma plataforma é sempre recomendado. Infelizmente, a automatização e a relativa facilidade de utilização de ferramentas de anotação têm levado à publicação de muitos trabalhos com listas de metabólitos anotados automaticamente e não devidamente inspecionados. Este processo não deve de modo algum ser empregado de modo superficial, toda anotação deve ser cuidadosamente avaliada por um pesquisador com

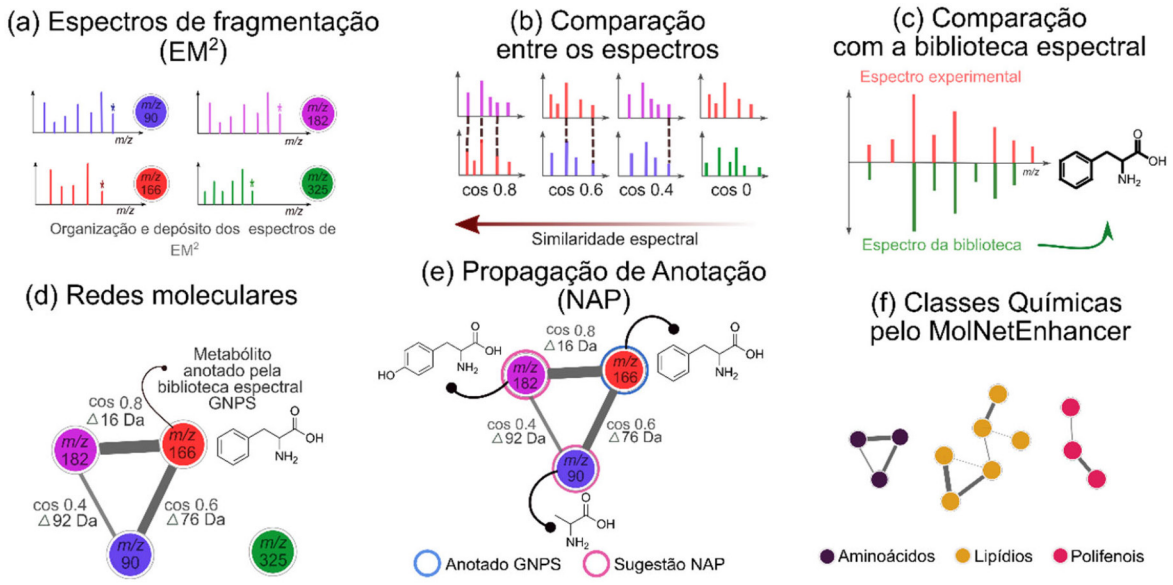

Figura 5. Visualização geral de um exemplo da construção de rede molecular geradas com a plataforma GNPS. (a) Inicialmente os espectros de fragmentação de massas são organizados de acordo com sua similaridade. (b) Depois comparados e classificados pelo índice de similaridade baseado no cosseno. (c) Os espectros são comparados com a biblioteca de espectros da plataforma GNPS, e quando esses espectros se combinam com um índice de cosseno igual ou maior àquele escolhido nas definições dos parâmetros, tem-se um *match* espectral. (d) As redes moleculares são construídas através da similaridade entre os espectros em uma disposição visual em formato de redes, na qual os nodos representam um ou mais espectros de massas semelhantes ou idênticos, conectados por ligações *(edges)* de acordo com a similaridade entre os espectros. Espectros muito distintos não são conectados em redes. (e) A ferramenta de redes moleculares pode colaborar para propagar a anotação de metabólitos correlacionados, a partir das perdas ou incrementos de partes estruturais, através do NAP. (f) De acordo com os padrões de fragmentação, também é possível sugerir a classe química que os metabólitos daquela rede pertencem, a ferramenta MolNetEnhancer realizada essa organização de forma automática[109,110]

conhecimento em química orgânica para atestar sua plausabilidade.

Em busca da padronização de anotações em estudos de metabolômica, a iniciativa para padronização da metabolômica (MSI, do inglês *Metabolomic Standards Initiative*)[115] apresentou uma abordagem sistemática para a classificação da identificação de metabólitos em quatro níveis de anotação.[115–117] Anotação nível 1 (identificação) é obtida apenas quando se realiza a co-eluição do metabólito de interesse com um padrão comercial disponível (o que pode ser realizado para substâncias conhecidas) e/ou pelo isolamento do composto seguido de sua caracterização estrutural a partir de técnicas como RMN (geralmente utilizado para substancias desconhecidas ou não disponíveis comercialmente). Anotação nível 2 é alcançada quando se possui dois ou mais tipos dados experimentais comparados com dados de uma biblioteca espectral ou da literatura, considerando por exemplo os valores de EM^1 e EM^2 (dados geralmente empregados na busca automática em bibliotecas espectrais), ou ainda quando é possível comparar dados de ultravioleta (UV), informações de fonte biológica ou bioatividade entre um composto analisado e a literatura. O nível 3 de anotação se dá quando um metabólito é anotado com apenas um dado experimental (exemplo EM^1 ou EM^2), e consequentemente, a anotação deve ser considerada apenas a nível de classe química. A anotação obtida por emprego de ferramentas *in silico* são geralmente classificadas como de nível 3. O nível 4 de anotação é caracterizado quando apenas é possível assumir que uma *feature* detectada corresponde a um metabólito, porém este é desconhecido. Nesse nível de anotação, é possível realizar o cálculo da fórmula molecular pela *m/z* observada, considerando-se o erro com base nas massas teórica e experimental da fórmula molecular predita. Essa informação é ao menos válida para nortear estudos alvo para isolamento desses metabólitos, por exemplo. [115,116] A descrição do nível de anotação obtida para os metabólitos analisados tem sido cada vez mais exigida em publicações, seja pelas normas das revistas ou pelos revisores. Essa informação deixa claro aos leitores de um trabalho o nível de confiança relativo a estas anotações. Além disso, é comum se observar, principalmente entre pesquisadores iniciantes na área, dúvidas ou confusão quanto ao emprego dos termos "anotação" e "identificação". O termo "identificação" somente pode ser utilizado ao se referir a uma anotação nível 1. Não se pode considerar uma anotação obtida a partir de uma biblioteca espectral ou ferramenta *in silico* como a identificação de um metabólito. É necessário aos pesquisadores atentarem-se aos termos utilizados em seus trabalhos para evitar interpretações equivocadas de seus resultados.

Além disso, é importante ressaltar que as estruturas anotadas de maneira automática (neste contexto, este termo pode ser entendido como sinônimo de "sugeridas"), por uso de ferramentas computacionais, devem ser todas conferidas manualmente (Figura 5d). O processo automatizado serve para dar um direcionamento e para acelerar a análise, mas isso não significa que a avaliação do pesquisador deixa de ser requerida. Os espectros depositados em bibliotecas espectrais são referentes a uma dada estrutura, porém o pesquisador deve ter em mente, por exemplo, que outras estruturas podem apresentar o mesmo perfil (espectro) de fragmentação. Os flavonoides isômeros quercetina e morina, por exemplo, não podem ser diferenciados apenas por dados de EM^1 e EM^2, portanto quando se observa a anotação de rutina ou isoquercetina com uso de ferramentas computacionais não se pode considerar a presença destes compostos pelo simples fato de não ser possível afirmar que se tem uma quercetina e não uma morina. Outro exemplo muito importante de ser mencionado é a presença de monossacarídeos ligados a uma estrutura principal por uma ligação *O*-glicosídica. Nesse caso, a ligação *O*-glicosídica é quebrada durante a fragmentação e o glicosídeo sai da estrutura de forma neutra e nenhuma informação estrutural é obtida, portanto, quando se observa a perda neutra de 160 u.m.a. pode-se sugerir a presença de um hexosídeo, mas não podemos afirmar se é uma glucose ou uma galactose, mesmo que o espectro depositado na biblioteca de espectros seja referente a uma estrutura com uma galactose. Se não podemos afirmar quais grupos químicos fazem parte da ligação, também não podemos fazer afirmações a respeito da estereoquímica, se o composto é *R* ou *S*, se é *cis* ou *trans*, se a ligação é alfa (α) ou beta (β)-glicosídica, salvo algumas raríssimas exceções descritas na literatura e, neste caso, a maneira como a diferenciação pôde ser feita deve estar bem detalhada no corpo do manuscrito. Dessa forma, o pesquisador deve usar as anotações obtidas automaticamente das plataformas apenas como um guia e criar suas próprias tabelas de compostos que realmente são possíveis de anotação dentro do seu conjunto de dados. Para avaliar a plausabilidade de uma anotação, alguns passos podem ser adotados. Primeiramente, pode ser realizada a avaliação do erro observado entre o *m/z* experimental e o calculado para o metabólito anotado. Nesta etapa, deve-se ter conhecimento do erro máximo tolerável a partir do equipamento empregado para aquisição dos dados de EM. Por exemplo, se estes foram gerados em um espectrômetro de massas do tipo quadrupolo-tempo de voo (QTOF, do inglês Q*uadrupole Time-Of-Flight*) devidamente calibrado, o erro observado geralmente não deve ser superior a 20 ppm. Em seguida, pode ser avaliado o *mirror plot*, que é uma comparação gráfica dos espectros de fragmentação experimentais e da biblioteca. Para se poder concluir que um bom *match* espectral foi obtido, deve ser observado o número de íons fragmentos existentes e quantos destes aparecem em ambos os espectros, sendo que quanto maior o número de íons compartilhados, mais plausível será a anotação. Por fim, pode ser realizada ainda uma busca na literatura para checar se há dados que corroborem as anotações obtidas no contexto experimental, como por exemplo fonte biológica, identificação taxonômica, região geográfica de ocorrência etc.

6.1. Anotações com base em bibliotecas espectrais e estruturais

A plataforma GNPS, além de propiciar análises de redes moleculares, também submete as *features* detectadas a um processo de anotação automatizado, o qual busca por espectros similares em bibliotecas espectrais integradas na própria plataforma (Figuras 5b e 5c).[54,118] Quando um espectro da biblioteca é encontrado pela ferramenta como uma possível estrutura para um dado espectro experimental, diz-se que um possível *hit* foi encontrado, ou ainda o termo espectro combinado ou *match* espectral costuma ser usado. Atualmente, tais bibliotecas fornecem juntas espectros relativos a mais de 2 milhões de metabólitos analisados por CG-EM e CL-EM/EM. Ademais, é possível aos usuários adicionarem seus próprios espectros de compostos identificados e compartilhá-los com a comunidade em geral.

Para ampliar a investigação das anotações, outras plataformas *online* estão disponíveis e podem ser utilizadas na busca de estruturas e de espectros de fragmentação para a comparação com dados experimentais.[119] Podemos citar aqui o COCONUT (COlleCtion of Open NatUral producTs) (disponível em https://coconut.naturalproducts.net/), um banco bastante amplo e abrangente. Entretanto, apesar de conter informações sobre substâncias obtidas de fontes naturais, este também inclui muitas substâncias obtidas de alimentos e oriundas da combustão de derivados de petróleo, por exemplo.[119] A Tabela 1 apresenta alguns exemplos de bibliotecas espectrais e seu principal conteúdo.

6.2. Ferramentas *in silico* de anotação

Apesar das bibliotecas espectrais facilitarem a rápida identificação de metabólitos conhecidos, em média apenas 2-5% dos espectros detectados em uma amostra biológica costumam ser anotados a partir desta abordagem, embora essa taxa possa chegar a até ~15% em matrizes mais estudadas, como por exemplo *Escherichia coli* e amostras de plasma sanguíneo e urina.[98,128] Uma alternativa para expandir as taxas de anotações realizadas em estudos de metabolômica é a utilização de ferramentas baseadas em fragmentação *in silico*, que geralmente envolvem a utilização de métodos fundamentados em química quântica, aprendizado de máquinas, entre outros.[129,130] Esta abordagem permite explorar o potencial representado por diversos bancos de estruturas como PubChem[130] e Chemspider[131], que a título de exemplo possuem juntos informações estruturais de cerca de 100 milhões de compostos, mas deste total menos de 1% apresentam dados de espectrometria de massas.[132] A plataforma GNPS, além de favorecer a desreplicação de compostos com base em sua biblioteca espectral, que estima-se conter espectros de referência correspondentes a cerca de 2,5% dos produtos naturais atualmente conhecidos,[109] também integra opções de análise com as ferramentas de previsão estrutural *in silico* Network Annotation Propagation (NAP),[129] Dereplicator[133] e Dereplicator+[134], que são bastante úteis principalmente para a anotação das classes químicas de compostos agrupados dentro de uma mesma família molecular.

A ferramenta NAP[129] se baseia na topologia das redes moleculares para ranquear as estruturas candidatas indicadas pelo algoritmo *in silico* MetFrag[135] a cada *feature* de interesse, automatizando o processo de propagação de anotações fundamentado na similaridade estrutural que é observada entre compostos relacionados entre si dentro de uma mesma família molecular (Figura 5e). Atualmente, os bancos de

Tabela 1. Descrição das principais bibliotecas de espectros de massas e estruturas públicas e *online* disponíveis para a comparação dos espectros de EM^2

Base de dados	Dados presentes	Referência
GNPS	Biblioteca pública de espectros de massas (EM/EM), com associação com diferentes bibliotecas	https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp[54,120]
NISTwebbook (Online) **NIST (Comercial)** **NIST/EPA/NIH Mass Spectral Library- Wiley (Comercial)**	Bibliotecas que fornecem espectros de massas para CG-EM. Utilizadas para identificação de substâncias naturais e sintéticas	https://webbook.nist.gov/chemistry/# https://www.wiley.com/en-ai/NIST+EPA+NIH+Mass+Spectral+Library+2020-p-9781119750291
WILEY	O maior banco com dados de IE-EM. Também apresenta dados de CID-EM/EM	https://onlinelibrary.wiley.com/[121]
METLIN	Biblioteca pública com metabólitos em geral	https://metlin.scripps.edu/[122]
MassBank of Japan, EU e North America	Metabólitos utilizados em ciências da vida em geral. O mais longo repositório de dados de EM.	http://massbank.jp/[123,124] https://massbank.eu/MassBank/[123,124] http://mona.fiehnlab.ucdavis.edu/[123,124]
MassBank	Dados de IE, EM^2e EM^n.	https://massbank.eu/MassBank/[123]
mzCloud	Metabólitos em geral	https://www.mzcloud.org/[125]
HMDB	Para metabólitos majoritariamente humanos e de mamíferos	https://hmdb.ca/[126]
RIKEN tandem mass spectral database (ReSpect)	Dados de EM^2 e tempo de retenção de alguns metabólitos	http://www.csrs.riken.jp/en/database/index.html/[62]
$NuBBE_{DB}$	Dados sobre produtos naturais isolados no Brasil.	https://nubbe.iq.unesp.br/portal/nubbe-search.html[127]

dados disponíveis para pronta utilização em análises com o NAP são a biblioteca GNPS, Human Metabolome Database (HMDB)[126], SuperNatural II (SupNat), ChEBI, DrugBank, FooDB e Natural Products Atlas (NPAtlas),[136,137] mas os usuários podem adicionar suas próprias bibliotecas *in house* caso desejem. O NAP não requer *matches* espectrais em uma família molecular para anotar as prováveis estruturas que a compõem; mesmo nestes casos a ferramenta consegue anotar corretamente cerca de 63% das subestruturas que constituem o composto ranqueado com maior *score,* no entanto, esta taxa chega a 81% em famílias moleculares onde há *matches*. Devido à possibilidade de aumentar o número de metabólitos com anotação, o NAP tem sido utilizado com sucesso na investigação de diferentes amostras de origem natural, incluindo plantas e microrganismos.[138–141] O NAP utilizado em conjunto com algoritmos de classificação química hierárquica como o *Classyfire* também pode favorecer a anotação das famílias moleculares quanto aos diferentes níveis da taxonomia química (superclasse, classe, subclasse).[142–144]

Existem também ferramentas *in silico* para anotação de classes químicas específicas, como por exemplo o *Dereplicator*[133], uma ferramenta para anotação de produtos naturais da classe dos peptídeos (PNPs, do inglês *Peptidic Natural Products*), que inclui diversos antibióticos e demais compostos bioativos de importância médica e ambiental.[145,146] Os PNPs, ao contrário do que geralmente se observa em peptídeos proteinogênicos, não se limitam a conformações estruturais majoritariamente lineares, podendo apresentar configurações muito mais complexas. Além disso, enquanto os peptídeos proteinogênicos estudados pela proteômica tradicional são compostos pelos 20 aminoácidos essenciais, os PNPs podem possuir centenas de aminoácidos distintos em sua composição, o que reduz a eficiência de *softwares* desenvolvidos especificamente para proteômica em sua análise.[147] Como estratégia para favorecer a desreplicação de PNPs, o *Dereplicator* computa os *matches* espectrais obtidos para um conjunto de espectros experimentais a partir tanto de seu banco de dados interno de PNPs conhecidos, quanto de uma *database decoy* constituída por peptídeos teóricos (não-existentes), cujas composições de aminoácidos são similares às dos peptídeos conhecidos. A ferramenta considera a taxa de falsas descobertas como a razão entre o número de PNPs anotados a partir de um banco de dados *decoy* (que contém peptídeos inexistentes, porém com composições de aminoácidos similares a peptídeos existentes) e o número de anotações obtidas pelos bancos de dados padrão. A significância estatística (*p*-valores) dos *matches* é calculada, e com esta avaliação a ferramenta gera uma lista de candidatos para cada espectro considerado, com o PNP de maior *score* indicado como anotação provável.[148] O *Dereplicator*, de modo similar ao NAP, também utiliza as redes moleculares para propagar anotações, considerando modificações estruturais inferidas para peptídeos agrupados em mesmas famílias moleculares. Além disso, o algoritmo Varquest[149] é integrado ao *Dereplicator* e pode ser utilizado para buscar estudadas variantes modificadas de PNPs conhecidos, inclusive indicando a localização da modificação proposta na estrutura do composto.[149,150] O *Dereplicator* +[134] é uma ferramenta que expande a abordagem de desreplicação de PNPs do *Dereplicator* para aplicação em outras classes de produtos naturais, como policetídeos, alcalóides, flavonóides, benzenóides, entre outros. Ambas as ferramentas têm sido amplamente aplicadas especialmente em estudos de PNs microbianos, incluindo a busca por metabólitos ativos em ensaios biológicos.[151–153]

Quando a anotação de um metabólito não é atingida, uma alternativa é a anotação de subestruturas (partes da estrutura) que podem ser úteis para caracterização de funções orgânicas e classes químicas presentes em uma dada amostra. MS2LDA[154,155] é uma ferramenta baseada em um algoritmo não-supervisionado utilizada para detectar e mapear grupos comuns de íons fragmento e perdas neutras (chamadas *Mass2Motifs*) a partir de conjuntos de espectros de fragmentação, que podem ser representativos de sub estruturas moleculares e possibilitar a identificação de subfamílias e modificações compartilhadas entre diferentes compostos.[154–156] Para facilitar a anotação de *Mass2Motifs*, a *database* MotifDB é integrada à ferramenta, porém é necessário se atentar ao fato de que padrões de fragmentação podem abranger sub estruturas isoméricas, e, além disso, validações manuais devem ser realizadas principalmente se as estruturas anotadas são provenientes de amostras de origem diferente das analisadas.[109] O MS2LDA pode ser acessado a partir de sua plataforma *online* (http://ms2lda.org/) ou pela plataforma do GNPS.[155] Esta ferramenta tem possibilitado a caracterização da diversidade metabólica de microrganismos e plantas em diversos estudos e, em abordagens integradas às redes moleculares, têm facilitado a detecção de potenciais novos derivados de compostos já conhecidos.[156–159]

Para simplificar a integração dos resultados obtidos a partir das redes moleculares, MS2LDA e de ferramentas *in silico*, desenvolveu-se o MolNetEnhancer.[144] Esta ferramenta possibilita a combinação das informações geradas pelos demais *workflows* disponíveis na plataforma GNPS e promove de modo automatizado a classificação química hierárquica das estruturas representantes de cada família molecular através da ontologia *Classyfire*[142], o que em termos práticos possibilita a visualização das classes químicas dos metabólitos observados na rede molecular (Figura 5f). O MolNetEnhancer tem sido utilizado em diversos trabalhos para facilitar a compreensão geral de perfis metabólicos e favorecer a assimilação de características estruturais para os espectros detectados em amostras de origem natural.[160–162]

6.3. Outras ferramentas in silico de anotação vinculadas ao GNPS

Uma outra abordagem promissora utilizada para a desreplicação de compostos desconhecidos em estudos

de metabolômica é a utilização de ferramentas baseadas em algoritmos que computam árvores de fragmentação a partir espectros de EM^2 ou EM^n que têm sido desenvolvidas pelo grupo do professor Sebastian Böcker em Jena na Alemanha.[163–165] O SIRIUS 4[166,167], por exemplo, permite análises de padrão isotópico e de fragmentação a partir de dados de espectrometria de massas de alta resolução (EMAR ou HRMS, do inglês *High Resolution Mass Spectrometry*) e espectros de EM^2. O algoritmo ZODIAC[167], integrado ao SIRIUS, é efetivo para determinação de fórmulas moleculares mesmo para compostos de massa molecular superior a 500 Da (a taxa de fórmulas moleculares corretamente designadas para *features* de maior massa costuma ser relativamente baixa).[166] Além disso, o método CSI:FingerID,[132,168] que combina a computação das árvores e aprendizado de máquina para a previsão *in silico* de propriedades moleculares, é utilizado pelo SIRIUS para procurar compostos candidatos para as *features* desconhecidas de uma amostra em bancos de dados como o PubChem, Human Metabolome Database, CHEBI, entre outros.[144] O SIRIUS também dispõe de uma ferramenta computacional para anotação sistemática de classes químicas chamado CANOPUS,[169] que performa a classificação automatizada dos compostos com base na ontologia Classyfire.[142] Atualmente, o SIRIUS pode performar as análises a partir de dados de EM^2 processados em *softwares* como o MZmine2 (*peak lists* em formato .mgf) ou dados brutos em formato aberto (.mzXML, .mzML), sendo que neste último caso é necessário realizar o processamento dentro da própria ferramenta como etapa inicial da análise.

7. Considerações Finais

Nessa revisão-tutorial, discutimos de modo geral e sumarizado as principais etapas de aquisição e de processamento de dados de metabolômica, com detalhamento sobre sua aplicação para análises estatísticas, construção de redes moleculares e anotação dos metabólitos. Os procedimentos aqui discutidos baseiam-se em *softwares* e plataformas gratuitas e *open-source* a fim de auxiliar os usuários em suas análises, tornando assim a ciência mais democrática, aberta e acessível a grupos de pesquisa que possuem recursos financeiros limitados para a obtenção de licenças de *softwares* pagos.

Esperamos possibilitar que estudantes e pesquisadores iniciantes da área possam obter uma boa base para fundamentar seus próximos trabalhos, otimizando suas análises e tornando-se hábeis a trabalhar com as mesmas ferramentas de metabolômica utilizadas pelos mais prolíficos grupos de pesquisa em QPNs, muitas das quais discutidas neste trabalho. Assim, contribuímos para que os pesquisadores do Brasil e demais países lusófonos consigam contribuir com trabalhos de alto impacto para nosso campo de pesquisa.

Agradecimentos

Os autores NMM, AKP e JBF agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro do processo 0001. A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelos apoios financeiros: NMM [#2022/01529-4], AKP [#2016/12304-2 e #2018/21936-8], JBF [#2012/25299-6], LG [#2018/26066-1], HPF [# 2015/09208-9 e #2018/23144-1], e AB [#2017/17648-4 e #2018/24865-4].

Referências

1. Beniddir, M. A.; Kang, K. Bin; Genta-Jouve, G.; Huber, F.; Rogers, S.; Van Der Hooft, J. J. J.; Advances in decomposing complex metabolite mixtures using substructure- And network-based computational metabolomics approaches. *Natural Product Reports* **2021**, *38*, 1967. [Crossref] [PubMed]
2. Khan, R. A.; Natural products chemistry: The emerging trends and prospective goals. *Saudi Pharmaceutical Journal* **2018**, *26*, 739. [Crossref]
3. Funari, C. S.; Castro-Gamboa, I.; Cavalheiro, A. J.; Da Silva Bolzani, V.; Metabolômica, uma abordagem otimizada para exploração da biodiversidade brasileira: Estado da arte, perspectivas e desafios. *Química Nova* **2013**, *36*, 1605. [Crossref]
4. Kharwar, R. N.; Mishra, A.; Gond, S. K.; Stierle, A.; Stierle, D.; Anticancer compounds derived from fungal endophytes: Their importance and future challenges. *Natural Product Reports* **2011**, *28*, 1208. [Crossref] [PubMed]
5. Hassan, M. Z.; Osman, H.; Ali, M. A.; Ahsan, M. J.; Therapeutic potential of coumarins as antiviral agents. *European Journal of Medicinal Chemistry* **2016**, *123*, 236. [Crossref] [PubMed]
6. Traversier, M.; Gaslondes, T.; Milesi, S.; Michel, S.; Delannay, E.; Polar lipids in cosmetics: recent trends in extraction, separation, analysis and main applications. *Phytochemistry Reviews* **2018**, 17, 1179. [Crossref]
7. Velioglu, Y. S.; Mazza, G.; Gao, L.; Oomah, B. D.; Antioxidant Activity and Total Phenolics in Selected Fruits, Vegetables, and Grain Products. *Journal of Agricultural and Food Chemistry* **1998**, *46*, 4113. [Crossref]
8. Céspedes, C. L.; Salazar, J. R.; Ariza-Castolo, A.; Yamaguchi, L.; Ávila, J. G.; Aqueveque, P.; Kubo, I.; Alarcón, J.; Biopesticides from plants: *Calceolaria integrifolia* s.l. *Environmental Research* **2014**, *132*, 391. [Crossref] [PubMed]
9. Krell, T.; Matilla, M. A.; Antimicrobial resistance: progress and challenges in antibiotic discovery and anti-infective therapy. *Microbial Biotechnology* **2022**, *15*, 70. [Crossref] [PubMed]
10. El-Elimat, T.; Figueroa, M.; Ehrmann, B. M.; Cech, N. B.; Pearce, C. J.; Oberlies, N. H.; High-resolution MS, MS/MS, and UV database of fungal secondary metabolites as a dereplication protocol for bioactive natural products. *Journal of Natural Products* **2013**, *76*, 1709. [Crossref] [PubMed]
11. Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; De Felicio, R.; Fenner,

A.; Wong, W. R.; Linington, R. G.; Zhang, L.; Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C.; Molecular networking as a dereplication strategy. *Journal of Natural Products* **2013**, *76*, 1686. [Crossref] [PubMed]

12. Amaral, J. C.; Da Silva, M. M.; Da Silva, M. F. G. F.; Alves, T. C.; Ferreira, A. G.; Forim, M. R.; Fernandes, J. B.; Pina, E. S.; Lopes, A. A.; Pereira, A. M. S.; Novelli, V. M.; Advances in the Biosynthesis of Pyranocoumarins: Isolation and 13C-Incorporation Analysis by High-Performance Liquid Chromatography-Ultraviolet-Solid-Phase Extraction-Nuclear Magnetic Resonance Data. *Journal of Natural Products* **2020**, *83*, 1409. [Crossref] [PubMed]
13. Berlinck, R. G. S.; De Borges, W. S.; Scotti, M. T.; Vieira, P. C.; A Química de Produtos Naturais do Brasil do Século XXI. *Química Nova* **2017**, *40*, 706. [Crossref]
14. Pilon, A. C.; Selegato, D. M.; Fernandes, R. P.; Bueno, P. C. P.; Pinho, D. R.; Neto, F. C.; Freire, R. T.; Castro-Gamboa, I.; Bolzani, V. S.; Lopes, N. P.; Metabolômica de plantas: Métodos e desafios. *Química Nova* **2020**, *43*, 329. [Crossref]
15. Canuto, G. A. B.; Da Costa, J. L.; Da Cruz, P. L. R.; De Souza, A. R. L.; Faccio, A. T.; Klassen, A.; Rodrigues, K. T.; Tavares, M. F. M.; Metabolômica: definições, estado-da-arte e aplicações representativas. *Química Nova* **2018**, *41*, 75. [Crossref]
16. Bauermeister, A.; Mannochio-Russo, H.; Costa-Lotufo, L. V.; Jarmusch, A. K.; Dorrestein, P. C.; Mass spectrometry-based metabolomics in microbiome investigations. *Nature Reviews Microbiology* **2022,** *20*, 143. [Crossref] [PubMed]
17. Soares, M. S.; Da Silva, D. F.; Forim, M. R.; Das Graças Fernandes Da Silva, M. F.; Fernandes, J. B.; Vieira, P. C.; Silva, D. B.; Lopes, N. P.; De Carvalho, S. A.; De Souza, A. A.; Machado, M. A.; Quantification and localization of hesperidin and rutin in *Citrus sinensis* grafted on *C. limonia* after *Xylella fastidiosa* infection by HPLC-UV and MALDI imaging mass spectrometry. *Phytochemistry* **2015**, *115*, 161. [Crossref] [PubMed]
18. Baidoo, E. E. K.; Benke, P. I.; Keasling, J. D.; Em *Microbial Systems Biology*; Navid, A. 1a. ed.; Humana Press: Totowa, 2012, cap. 9.
19. Lisec, J.; Schauer, N.; Kopka, J.; Willmitzer, L.; Fernie, A. R.; Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nature Protocols* **2006**, *1*, 387. [Crossref] [PubMed]
20. Zhang, A.; Sun, H.; Wang, X.; Mass spectrometry-driven drug discovery for development of herbal medicine. *Mass Spectrometry Reviews* **2018**, *37*, 307. [Crossref] [PubMed]
21. Wolfender, J. L.; Nuzillard, J. M.; Van Der Hooft, J. J. J.; Renault, J. H.; Bertrand, S.; Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography-High-Resolution Tandem Mass Spectrometry and NMR Profiling, in Silico Databases, and Chemometrics. *Analytical Chemistry* **2019**, *91*, 704. [Crossref] [PubMed]
22. Molina-Díaz, A.; Beneito-Cambra, M.; Moreno-González, D.; Gilbert-López, B.; Ambient mass spectrometry from the point of view of Green Analytical Chemistry. *Current Opinion in Green and Sustainable Chemistry* **2019**, *19*, 50. [Crossref]
23. Liu, P.; Forni, A.; Chen, H.; Development of solvent-free ambient mass spectrometry for green chemistry applications. *Analytical Chemistry* **2014**, *86*, 4024. [Crossref] [PubMed]
24. Borges, R. M.; Resende, J. V. M. Deconstructing Metabolomics Within Natural Products: an Invitation for a Discussion. *Química Nova* **2021**, *44*, 1392. [Crossref]
25. Cho, H. W.; Kim, S. B.; Jeong, M. K.; Park, Y.; Miller, N. G.; Ziegler, T. R.; Jones, D. P.; Discovery of metabolite features for the modelling and analysis of high-resolution NMR spectra. *International Journal of Data Mining and Bioinformatics* **2008**, *2*, 176. [Crossref] [PubMed]
26. Cambiaghi, A.; Ferrario, M.; Masseroli, M.; Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. *Briefings in Bioinformatics* **2017**, *18*, 498. [Crossref] [PubMed]
27. Houriet, J.; Vidar, W. S.; Manwill, P. K.; Todd, D. A.; Cech, N. B.; How Low Can You Go? Selecting Intensity Thresholds for Untargeted Metabolomics Data Preprocessing. *Analytical Chemistry* **2022**, *94*, 17964. [Crossref] [PubMed]
28. Belinato, J. R.; Bazioli, J. M.; Sussulini, A.; Augusto, F.; Filla, T. P.; Microbial metabolomics: Innovations and applications. *Química Nova* **2019**, *42*, 546. [Crossref]
29. Salem, M. A.; De Souza, L. P.; Serag, A.; Fernie, A. R.; Farag, M. A.; Ezzat, S. M.; Alseekh, S.; Metabolomics in the context of plant natural products research: From sample preparation to metabolite analysis. *Metabolites* **2020**, *10*, 1. [Crossref]
30. Zhou, B.; Xiao, J. F.; Tuli, L.; Ressom, H. W.; LC-MS-based metabolomics. *Molecular BioSystems* **2012**, *8*, 470. [Crossref] [PubMed]
31. Kim, H. K.; Verpoorte, R.; Sample preparation for plant metabolomics. *Phytochemical Analysis* **2010**, *21*, 4. [Crossref] [PubMed]
32. Ahmad, S.; Siddiqui, M. R.; Ali, M. S.; Wabaidur, S. M.; Alam, M. S.; Alam, N.; Alothman, Z. A.; Khan, M. A.; Khan, M. R.; Solid phase extraction and LC-MS/MS method for quantification of venlafaxine and its active metabolite O-desmethyl venlafaxine in rat plasma. *Journal of the Chilean Chemical Society* **2016**, *61*, 3130. [Crossref]
33. Wu, Q.; Xu, Y.; Ji, H.; Wang, Y.; Zhang, Z.; Lu, H.; Enhancing coverage in LC–MS-based untargeted metabolomics by a new sample preparation procedure using mixed-mode solid-phase extraction and two derivatizations. *Analytical and Bioanalytical Chemistry* **2019**, *411*, 6189. [Crossref] [PubMed]
34. Trenholm, R. A.; Vanderford, B. J.; Snyder, S. A.; On-line solid phase extraction LC-MS/MS analysis of pharmaceutical indicators in water: A green alternative to conventional methods. *Talanta* **2009**, *79*, 1425. [Crossref] [PubMed]
35. Izcara, S.; Casado, N.; Morante-Zarcero, S.; Sierra, I.; A miniaturized QuEChERS method combined with ultrahigh liquid chromatography coupled to tandem mass spectrometry for the analysis of pyrrolizidine alkaloids in oregano samples. *Foods* **2020**, *9*, 1319. [Crossref] [PubMed]
36. Kaczyński, P.; Łozowicka, B.; A novel approach for fast and simple determination pyrrolizidine alkaloids in herbs by ultrasound-assisted dispersive solid phase extraction method coupled to liquid chromatography-tandem mass spectrometry.

Journal of pharmaceutical and biomedical analysis **2020**, *187*, 113351. [Crossref] [PubMed]

37. Walker, K.; Duringer, J.; Craig, A. M.; Determination of the ergot alkaloid ergovaline in tall fescue seed and straw using a QuEChERS extraction method with high-performance liquid chromatography-fluorescence detection. *Journal of Agricultural and Food Chemistry* **2015**, *63*, 4236. [Crossref] [PubMed]
38. Casado, N.; Perestrelo, R.; Silva, C. L.; Sierra, I.; Câmara, J. S.;
39. An improved and miniaturized analytical strategy based on μ-QuEChERS for isolation of polyphenols. A powerful approach for quality control of baby foods. *Microchemical Journal* **2018**, *139*, 110. [Crossref]
40. Li, B.; Ge, J.; Liu, W.; Hu, D.; Li, P.; Unveiling spatial metabolome of Paeonia suffruticosa and Paeonia lactiflora roots using MALDI MS imaging. *New Phytologist* **2021**, *231*, 892. [Crossref] [PubMed]
41. Claude, E.; Jones, E. A.; Pringle, S. D. Em *Methods in Molecular Biology;* Cole, L. M., 1a. ed. Humana Press: New York, 2017, cap. 7.
42. Ernst, M.; Silva, D. B.; Silva, R. R.; Vêncio, R. Z. N.; Lopes, N. P.; Mass spectrometry in plant metabolomics strategies: From analytical platforms to data acquisition and processing. *Natural Product Reports* **2014**, *31*, 784. [Crossref] [PubMed]
43. Chiaradia, M. C.; Collins, C. H.; Jardim, I. C. S. F.; Estado da arte da cromatografia associada à espectrometria de massas acoplada à espectrometria de massas na análise de compostos tóxicos em alimentos. *Química Nova* **2008**, *31*, 623. [Crossref]
44. Ludwig, C.; Gillet, L.; Rosenberger, G.; Amon, S.; Collins, B. C.; Aebersold, R.; Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular Systems Biology* **2018**, *14*, e8126. [Crossref] [PubMed]
45. Li, K. W.; Gonzalez-lozano, M. A.; Koopmans, F.; Smit, A. B.; Recent Developments in Data Independent Acquisition (DIA) Mass Spectrometry: Application of Quantitative Analysis of the Brain Proteome. *Frontiers in Molecular Neuroscience* **2020**, *13*, 564446. [Crossref]
46. Cavalcanti, G. D. A.; Borges, R. M.; Carneiro, G. R. A.; Padilha, M. C.; Pereira, H. M. G.; Variable Data Independent Acquisition and Data Mining Exploring Feature-Based Molecular Networking Analysis for Untargeted Screening of Synthetic Cannabinoids in Oral Fluid. *Journal of the American Society for Mass Spectrometry* **2021**, *32*, 2417. [Crossref] [PubMed]
47. Davies, V.; Wandy, J.; Weidt, S.; Van Der Hooft, J. J. J.; Miller, A.; Daly, R.; Rogers, S.; Rapid Development of Improved Data-Dependent Acquisition Strategies. *Analytical Chemistry* **2021**, *93*, 5676. [Crossref]
48. Kalli, A.; Smith, G. T.; Sweredoski, M. J.; Hess, S.; Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: Focus on LTQ-orbitrap mass analyzers. *Journal of Proteome Research* **2013**, *12*, 3071. [Crossref] [PubMed]
49. Olivon, F.; Roussi, F.; Litaudon, M.; Touboul, D.; Optimized experimental workflow for tandem mass spectrometry molecular networking in metabolomics. *Analytical and Bioanalytical Chemistry* **2017**, *409*, 5767. [Crossref] [PubMed]
50. Phapale, P.; Rai, V.; Mohanty, A. K.; Srivastava, S.; Untargeted Metabolomics Workshop Report: Quality Control Considerations from Sample Preparation to Data Analysis. *Journal of the American Society for Mass Spectrometry* **2020**, *31*, 2006. [Crossref] [PubMed]
51. Liu, Q.; Walker, D.; Uppal, K.; Liu, Z.; Ma, C.; Tran, V. L.; Li, S.; Jones, D. P.; Yu, T.; Addressing the batch effect issue for LC/MS metabolomics data in data preprocessing. *Scientific Reports* **2020**, *10*, 1. [Crossref] [PubMed]
52. Torres, C. L.; Sardela, V. F.; Scalco, F. B.; de Aquino Neto, F. R.; Garrett, R.; Development and Application of a Test Mixture for Untargeted Liquid Chromatography-Mass Spectrometry Analysis of Urine Samples. *Química Nova* **2022**, *45*, 89. [Crossref]
53. ProteoWizard: Download. Disponível em: <https://proteowizard.sourceforge.io/download.html>. Acesso em: 26 março 2023.
54. Wang, M.; Nothias, L.-F.; Kang, K. Bin; Protsyuk, I; Mass Spectrometry File Conversion - GNPS Documentation. Disponível em: <https://ccms-ucsd.github.io/GNPSDocumentation/fileconversion/>. Acesso em: 26 março 2023.
55. Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W. T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C. C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C. C.; Yang, Y. L.; Humpf, H. U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya, C. A. P.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P. M.; Phapale, P.; Nothias, L. F.; Alexandrov, T.; Litaudon, M.; Wolfender, J. L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D. T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N.; Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016**, *34*, 828. [Crossref] [PubMed]
56. Jarmusch, S. A.; Van Der Hooft, J. J. J.; Dorrestein, P. C.; Jarmusch, A. K.; Advancements in capturing and mining mass spectrometry data are transforming natural products research. *Natural Product Reports* **2021**, *38*, 2066. [Crossref] [PubMed]

57. Reanalysis of Data User Interface for MS2 (ReDU). Disponível em: <https://redu.ucsd.edu/>. Acesso em: 26 março 2023.
58. Katajamaa, M.; Oresic, M.; Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A* **2007**, *1158*, 318. [Crossref]
59. Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H. C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O.; OpenMS: A flexible open-source software platform for mass spectrometry data analysis. *Nature Methods* **2016**, *13*, 741. [Crossref] [PubMed]
60. Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O.; OpenMS - An open-source software framework for mass spectrometry. *BMC Bioinformatics* **2008**, *9*, 1. [Crossref] [PubMed]
61. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M.; MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**, *11*, 1. [Crossref] [PubMed]
62. MZmine. Disponível em: <http://mzmine.github.io/>. Acesso em: 26 março 2023.
63. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; Vandergheynst, J.; Fiehn, O.; Arita, M.; MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods* **2015**, *12*, 523. [Crossref] [PubMed]
64. Wang, M.; Jarmusch, A. K.; Vargas, F.; Aksenov, A. A.; Gauglitz, J. M.; Weldon, K.; Petras, D.; da Silva, R.; Quinn, R.; Melnik, A. V.; van der Hooft, J. J. J.; Caraballo-Rodríguez, A. M.; Nothias, L. F.; Aceves, C. M.; Panitchpakdi, M.; Brown, E.; Di Ottavio, F.; Sikora, N.; Elijah, E. O.; Labarta-Bajo, L.; Gentry, E. C.; Shalapour, S.; Kyle, K. E.; Puckett, S. P.; Watrous, J. D.; Carpenter, C. S.; Bouslimani, A.; Ernst, M.; Swafford, A. D.; Zúñiga, E. I.; Balunas, M. J.; Klassen, J. L.; Loomba, R.; Knight, R.; Bandeira, N.; Dorrestein, P. C.; Mass spectrometry searches using MASST. *Nature Biotechnology* **2020**, *38*, 23. [Crossref] [PubMed]
65. Hoffmann, N.; Rein, J.; Sachsenberg, T.; Hartler, J.; Haug, K.; Mayer, G.; Alka, O.; Dayalan, S.; Pearce, J. T. M.; Rocca-Serra, P.; Qi, D.; Eisenacher, M.; Perez-Riverol, Y.; Vizcaíno, J. A.; Salek, R. M.; Neumann, S.; Jones, A. R.; MzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics. *Analytical Chemistry* **2019**, *91*, 3302. [Crossref] [PubMed]
66. Borges, R. M.; Resende, J. V. M.; Moraes, A. O. de; Pereira, A. K.; Garrett, R.; Bauermeister, A.; Silva, A. J. R. da; Guia para processamento de dados de cromatografia acoplada a espectrometria de massas. *Química Nova* **2022**, *45*, 608. [Crossref]
67. Smith, R.; Mathis, A. D.; Ventura, D.; Prince, J. T.; Proteomics, lipidomics, metabolomics: A mass spectrometry tutorial from a computer scientist's point of view. *BMC Bioinformatics* **2014**, *15*, 1. [Crossref] [PubMed]
68. Nothias, L. F.; Petras, D.; Schmid, R.; Dührkop, K.; Rainer, J.; Sarvepalli, A.; Protsyuk, I.; Ernst, M.; Tsugawa, H.; Fleischauer, M.; Aicheler, F.; Aksenov, A.; Alka, O.; Allard, P.-M.; Barsch, A.; Cachet, X.; Caraballo, M.; Da Silva, R.; Dang, T.; Garg, N.; Gauglitz, J.; Gurevich, A.; Isaac, G.; Jarmusch, A.; Kameník, Z.; Kang, K. Bin; Kessler, N.; Koester, I.; Korf, A.; Gouellec, A. Le; Ludwig, M.; Christian, M.; McCall, L.-I.; McSayles, J.; Meyer, S.; Mohimani, H.; Morsy, M.; Moyne, O.; Neumann, S.; Neuweger, H.; Nguyen, N. H.; Nothias-Esposito, M.; Paolini, J.; Phelan, V.; Pluskal, T.; Quinn, R.; Rogers, S.; Shrestha, B.; Tripathi, A.; van der Hooft, J.; Vargas, F.; Weldon, K.; Witting, M.; Yang, H.; Zhang, Z.; Zubeil, F.; Kohlbacher, O.; Böcker, S.; Alexandrov, T.; Bandeira, N.; Wang, M.; Dorrestein, P.; Feature-based Molecular Networking in the GNPS Analysis Environment. *Nature Methods* **2020**, *17*, 905. [Crossref] [PubMed]
69. Schmid, R.; Petras, D.; Nothias, L. F.; Wang, M.; Aron, A. T.; Jagels, A.; Tsugawa, H.; Rainer, J.; Garcia-Aloy, M.; Dührkop, K.; Korf, A.; Pluskal, T.; Kameník, Z.; Jarmusch, A. K.; Caraballo-Rodríguez, A. M.; Weldon, K. C.; Nothias-Esposito, M.; Aksenov, A. A.; Bauermeister, A.; Albarracin Orio, A.; Grundmann, C. O.; Vargas, F.; Koester, I.; Gauglitz, J. M.; Gentry, E. C.; Hövelmann, Y.; Kalinina, S. A.; Pendergraft, M. A.; Panitchpakdi, M.; Tehan, R.; Le Gouellec, A.; Aleti, G.; Mannochio Russo, H.; Arndt, B.; Hübner, F.; Hayen, H.; Zhi, H.; Raffatellu, M.; Prather, K. A.; Aluwihare, L. I.; Böcker, S.; McPhail, K. L.; Humpf, H. U.; Karst, U.; Dorrestein, P. C.; Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nature Communications* **2021**, *12*, [Crossref] [PubMed]
70. Worley, B.; Powers, R.; Multivariate Analysis in Metabolomics. *Current Metabolomics* **2013**, *1*, 92. [Crossref] [PubMed]
71. Blaise, B. J.; Correia, G. D. S.; Haggart, G. A.; Surowiec, I.; Sands, C.; Lewis, M. R.; Pearce, J. T. M.; Trygg, J.; Nicholson, J. K.; Holmes, E.; Ebbels, T. M. D.; Statistical analysis in metabolic phenotyping. *Nature Protocols* **2021**, *16*, 4299. [Crossref] [PubMed]
72. Hijaz, F.; Nehela, Y.; Killiny, N.; Possible role of plant volatiles in tolerance against huanglongbing in citrus. *Plant Signaling and Behavior* **2016**, *11*, 1. [Crossref] [PubMed]
73. Pang, Z.; Chong, J.; Zhou, G.; De Lima Morais, D. A.; Chang, L.; Barrette, M.; Gauthier, C.; Jacques, P. É.; Li, S.; Xia, J.; MetaboAnalyst 5.0: Narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research* **2021**, *49*, W388. [Crossref] [PubMed]
74. MetaboAnalyst. Disponível em: <https://www.metaboanalyst.ca/>. Acesso em: 26 março 2023.
75. Schiffman, C.; Petrick, L.; Perttula, K.; Yano, Y.; Carlsson, H.; Whitehead, T.; Metayer, C.; Hayes, J.; Rappaport, S.; Dudoit, S.; Filtering procedures for untargeted LC-MS metabolomics data. *BMC Bioinformatics* **2019**, *20*, 1. [Crossref] [PubMed]
76. Misra, B. B.; Data normalization strategies in metabolomics: Current challenges, approaches, and tools. *European Journal of Mass Spectrometry* **2020**, *26*, 165. [Crossref] [PubMed]
77. Mizuno, H.; Ueda, K.; Kobayashi, Y.; Tsuyama, N.; Todoroki, K.; Min, J. Z.; Toyo'oka, T.; The great importance of normalization

of LC–MS data for highly-accurate non-targeted metabolomics. *Biomedical Chromatography* **2017**, *31*, e3864. [Crossref] [PubMed]

78. van den Berg, R. A.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J.; Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics* **2006**, *7*, 1. [Crossref] [PubMed]
79. Saccenti, E.; Hoefsloot, H. C. J.; Smilde, A. K.; Westerhuis, J. A.; Hendriks, M. M. W. B.; Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics* **2014**, *10*, 361. [Crossref]
80. Szymańska, E.; Saccenti, E.; Smilde, A. K.; Westerhuis, J. A.; Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* **2012**, *8*, 3. [Crossref] [PubMed]
81. Fonville, J. M.; Richards, S. E.; Barton, R. H.; Boulange, C. L.; Ebbels, T. M. D.; Nicholson, J. K.; Holmes, E.; Dumas, M. E.; The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *Journal of Chemometrics* **2010**, *24*, 636. [Crossref]
82. Jollife, I. T.; Cadima, J.; Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2016**, *374*, 20150202. [Crossref] [PubMed]
83. Bridges Junior, C. C.; Hierarchical cluster analysis. *Psychological Reports* **1966**, *18*, 851. [Crossref]
84. Zuur, A. F.; Ieno, E. N.; Smith, G. M. Em *Analysing Ecological Data*; Zuur, A. F.; Ieno, E. N.; Smith, G. M., eds.; Springer: New York, 2007, cap. 15.
85. Steinbach, M.; Ertöz, L.; Kumar, V. Em *New Directions in Statistical Physics*; Wille, L.C., ed.; Springer: Berlin, 2004, cap. 16.
86. Ferreira, M. M. C.; *Quimiometria: Conceitos, métodos e aplicações*, 1a. ed., Editora da Unicamp: Campinas, 2015.
87. Pereira, A. K.; *Tese de Doutorado*, Universidade Federal de São Carlos, 2020. [Link]
88. Wold, S.; Sjöström, M.; Eriksson, L.; PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2001**, *58*, 109. [Crossref]
89. Ruiz-Perez, D.; Guan, H.; Madhivanan, P.; Mathee, K.; Narasimhan, G.; So you think you can PLS-DA? *BMC Bioinformatics* **2020**, *21*, 1. [Crossref] [PubMed]
90. Oza, V. H.; Aicher, J. K.; Reed, L. K.; Random forest analysis of untargeted metabolomics data suggests increased use of omega fatty acid oxidation pathway in *Drosophila melanogaster* larvae fed a medium chain fatty acid rich high-fat diet. *Metabolites* **2019**, *9*, 1. [Crossref] [PubMed]
91. Chong, J.; Wishart, D. S.; Xia, J.; Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Current Protocols in Bioinformatics* **2019**, *68*, e86. [Crossref] [PubMed]
92. Considine, E. C.; Thomas, G.; Boulesteix, A. L.; Khashan, A. S.; Kenny, L. C.; Critical review of reporting of the data analysis step in metabolomics. *Metabolomics* **2018**, *14*, 1. [Crossref]
93. De Melo Cazal, C.; Forim, M. R.; Da Silva, M. F. G. F.; Vieira, P. C.; Fernandes, J. B.; Development and validation of a RP-HPLC method to determine the xanthyletin content in biodegradable polymeric nanoparticles. *Química Nova* **2014**, *37*, 1145. [Crossref]
94. Sokal, R. R.; Rohlf, F. J.; *Biometry: the principles and practice of statistics in biological research,* 4a. ed. W.H. Freeman and Company: New York, 2012.
95. Vinaixa, M.; Samino, S.; Saez, I.; Duran, J.; Guinovart, J. J.; Yanes, O.; A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. *Metabolites* **2012**, *2*, 775. [Crossref] [PubMed]
96. Chen, Y.; Li, E. M.; Xu, L. Y.; Guide to Metabolomics Analysis: A Bioinformatics Workflow. *Metabolites* **2022**, *12*, 7. [Crossref] [PubMed]
97. Bouslimani, A.; Sanchez, L. M.; Garg, N.; Dorrestein, P. C.; Mass spectrometry of natural products: Current, emerging and future technologies. *Natural Product Reports* **2014**, *31*, 718. [Crossref] [PubMed]
98. Quinn, R. A.; Nothias, L. F.; Vining, O.; Meehan, M.; Esquenazi, E.; Dorrestein, P. C.; Molecular Networking As a Drug Discovery, Drug Metabolism, and Precision Medicine Strategy. *Trends in Pharmacological Sciences* **2017**, *38*, 143. [Crossref] [PubMed]
99. Da Silva, R. R.; Dorrestein, P. C.; Quinn, R. A.; Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112*, 12549. [Crossref] [PubMed]
100. Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B.; Yang, J.; Mass spectral molecular networking of living microbial colonies. *PNAS* **2012**, *109*, 1743. [Crossref] [PubMed]
101. Marner, M.; Patras, M. A.; Kurz, M.; Zubeil, F.; Förster, F.; Schuler, S.; Bauer, A.; Hammann, P.; Vilcinskas, A.; Schaberle, T. F.; Glaeser, J.; Molecular networking-guided discovery and characterization of stechlisins, a group of cyclic lipopeptides from a *Pseudomonas* sp. *Journal of Natural Products* **2020**, *83*, 2607. [Crossref] [PubMed]
102. Kang, K. Bin; Park, E. J.; Da Silva, R. R.; Kim, H. W.; Dorrestein, P. C.; Sung, S. H.; Targeted Isolation of Neuroprotective Dicoumaroyl Neolignans and Lignans from *Sageretia theezans* Using in Silico Molecular Network Annotation Propagation-Based Dereplication. *Journal of Natural Products* **2018**, *81*, 1819. [Crossref] [PubMed]
103. Oppong-Danquah, E.; Parrot, D.; Blümel, M.; Labes, A.; Tasdemir, D.; Molecular networking-based metabolome and bioactivity analyses of marine-adapted fungi co-cultivated with phytopathogens. *Frontiers in Microbiology* **2018**, *9*, 1. [Crossref] [PubMed]
104. Proença, D. N.; Heine, T.; Senges, C. H. R.; Bandow, J. E.; Morais, P. V.; Tischler, D.; Bacterial Metabolites Produced Under Iron Limitation Kill Pinewood Nematode and Attract *Caenorhabditis elegans*. *Frontiers in Microbiology* **2019**, *10*, 1. [Crossref] [PubMed]
105. Le Daré, B.; Ferron, P. J.; Allard, P. M.; Clément, B.; Morel, I.; Gicquel, T.; New insights into quetiapine metabolism using molecular networking. *Scientific Reports* **2020**, *10*, 1. [Crossref] [PubMed]

106. Lyu, Q.; Hsu, C.-C.; Can Diet Influence Our Health by Altering Intestinal Microbiota-Derived Fecal Metabolites? *mSystems* **2018**, *3*, e00187. [Crossref] [PubMed]
107. Nguyen, D. D.; Wu, C. H.; Moree, W. J.; Lamsa, A.; Medema, M. H.; Zhao, X.; Gavilan, R. G.; Aparicio, M.; Atencio, L.; Jackson, C.; Ballesteros, J.; Sanchez, J.; Watrous, J. D.; Phelan, V. V.; Van De Wiel, C.; Kersten, R. D.; Mehnaz, S.; De Mot, R.; Shank, E. A.; Charusanti, P.; Nagarajan, H.; Duggan, B. M.; Moore, B. S.; Bandeira, N.; Palsson, B.; Pogliano, K.; Gutiečrez, M.; Dorrestein, P. C.; MS/MS networking guided analysis of molecule and gene cluster families. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, E2611. [Crossref] [PubMed]
108. Huber, F.; Ridder, L.; Verhoeven, S.; Spaaks, J. H.; Diblen, F.; Rogers, S.; Van Der Hooft, J. J. J.; Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Computational Biology* **2021**, *17*, e1008724. [Crossref] [PubMed]
109. Frank, A. M.; Monroe, M. E.; Shah, A. R.; Carver, J. J.; Bandeira, N.; Moore, R. J.; Anderson, G. A.; Smith, R. D.; Pevzner, P. A.; Spectral archives: Extending spectral libraries to analyze both identified and unidentified spectra. *Nature Methods* **2011**, *8*, 587. [Crossref] [PubMed]
110. Hooft, J. J. J. Van Der; Mohimani, H.; Dorrestein, P. C.; Duncan, K. R.; Bauermeister, A.; Linking genomics and metabolomics to chart specialized metabolic diversity. *Chemical Society Reviews* **2020**, *49*, 3297. [Crossref] [PubMed]
111. Aron, A. T.; Gentry, E. C.; McPhail, K. L.; Nothias, L. F.; Nothias-Esposito, M.; Bouslimani, A.; Petras, D.; Gauglitz, J. M.; Sikora, N.; Vargas, F.; van der Hooft, J. J. J.; Ernst, M.; Kang, K. Bin; Aceves, C. M.; Caraballo-Rodríguez, A. M.; Koester, I.; Weldon, K. C.; Bertrand, S.; Roullier, C.; Sun, K.; Tehan, R. M.; Boya P, C. A.; Christian, M. H.; Gutiérrez, M.; Ulloa, A. M.; Tejeda Mora, J. A.; Mojica-Flores, R.; Lakey-Beitia, J.; Vásquez-Chaves, V.; Zhang, Y.; Calderón, A. I.; Tayler, N.; Keyzers, R. A.; Tugizimana, F.; Ndlovu, N.; Aksenov, A. A.; Jarmusch, A. K.; Schmid, R.; Truman, A. W.; Bandeira, N.; Wang, M.; Dorrestein, P. C.; Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nature Protocols* **2020**, *15*, 1954. [Crossref] [PubMed]
112. Sheng, Y.; Xue, Y.; Wang, J.; Liu, S.; Jiang, Y.; Fast screening and identification of illegal adulteration in dietary supplements and herbal medicines using molecular networking with deep-learning-based similarity algorithms. *Analytical and Bioanalytical Chemistry* **2023**, *415*, 3285. [Crossref] [PubMed]
113. Otasek, D.; Morris, J. H.; Bouças, J.; Pico, A. R.; Demchak, B.; Cytoscape Automation: Empowering workflow-based network analysis. *Genome Biology* **2019**, *20*, 1. [Crossref] [PubMed]
114. Pilon, A. C.; Vieira, N. C.; Amaral, J. G.; Monteiro, A. F.; Da Silva, R. R.; Spíndola, L. S.; Castro-Gamboa, I.; Lopes, N. P.; Molecular networks: An analysis on annotations and discovery of new assets. *Química Nova* **2021**, *44*, 1168. [Crossref]
115. Fox Ramos, A. E.; Evanno, L.; Poupon, E.; Champy, P.; Beniddir, M. A.; Natural products targeting strategies involving molecular networking: different manners, one goal. *Natural Product Reports* **2019**, *36*, 960. [Crossref] [PubMed]
116. Fiehn, O.; Robertson, D.; Griffin, J.; vab der Werf, M.; Nikolau, B.; Morrison, N.; Sumner, L. W.; Goodacre, R.; Hardy, N. W.; Taylor, C.; Fostel, J.; Kristal, B.; Kaddurah-Daouk, R.; Mendes, P.; van Ommen, B.; Lindon, J. C.; Sansone, S. A.; The metabolomics standards initiative (MSI). *Metabolomics* **2007**, *3*, 175. [Crossref] [PubMed]
117. Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W. M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R.; Proposed minimum reporting standards for chemical analysis. *Metabolomics* **2007**, *3*, 211. [Crossref] [PubMed]
118. Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J.; Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science and Technology* **2014**, *48*, 2097. [Crossref] [PubMed]
119. de Souza Moura, M.; Bellete, B. S.; Vieira, L. C. C.; Sampaio, O. M.; Uso de Redes Moleculares para Anotações de Compostos em Metabolômica. *Revista Virtual de Química* **2022**, *14*, 214. [Crossref]
120. Sorokina, M.; Steinbeck, C.; Review on natural products databases: where to find data in 2020. *Journal of Cheminformatics* **2020**, *12*, 1. [Crossref]
121. GNPS - Analyze, Connect, and Network with your Mass Spectrometry Data. Disponível em: <https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp>. Acesso em: 26 março 2023.
122. Wiley Online Library | Scientific research articles, journals, books, and reference works. Disponível em: <https://onlinelibrary.wiley.com/>. Acesso em: 26 março 2023.
123. Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G.; METLIN: A metabolite mass spectral database. *Therapeutic Drug Monitoring* **2005**, *27*, 747. [Crossref] [PubMed]
124. MassBank | Personal MassBank Mass Spectral DataBase. Disponível em: <http://massbank.jp/>, <https://massbank.eu/MassBank/>, <http://mona.fiehnlab.ucdavis.edu/>. Acesso em: 26 março 2023.
125. Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T.; MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* **2010**, *45*, 703. [Crossref] [PubMed]
126. mzCloud – Advanced Mass Spectral Database. Diponível em: <https://www.mzcloud.org/>. Acesso em: 26 março 2023.
127. Human Metabolome Database. Disponível em: <https://hmdb.ca/>. Acesso em: 26 março 2023.
128. Valli, M.; Dos Santos, R. N.; Figueira, L. D.; Nakajima, C. H.; Castro-Gamboa, I.; Andricopulo, A. D.; Bolzani, V. S.; Development of a natural products database from the biodiversity of Brazil. *Journal of Natural Products* **2013**, *76*, 439. [Crossref] [PubMed]

129. Bowen, B. P.; Northen, T. R.; Dealing with the unknown: Metabolomics and metabolite atlases. *Journal of the American Society for Mass Spectrometry* **2010**, *21*, 1471. [Crossref] [PubMed]
130. da Silva, R. R.; Wang, M.; Nothias, L. F.; van der Hooft, J. J. J.; Caraballo-Rodríguez, A. M.; Fox, E.; Balunas, M. J.; Klassen, J. L.; Lopes, N. P.; Dorrestein, P. C.; Propagating annotations of molecular networks using in silico fragmentation. *PLoS Computational Biology* **2018**, *14*, 1. [Crossref]
131. PubChem. Disponível em: <https://pubchem.ncbi.nlm.nih.gov/>. Acesso em: 26 março 2023.
132. ChemSpider | Search and share chemistry. Disponível em: <https://www.chemspider.com/>. Acesso em: 26 março 2023.
133. Blaženović, I.; Kind, T.; Ji, J.; Fiehn, O.; Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites* **2018**, *8*, 1. [Crossref] [PubMed]
134. Mohimani, H.; Gurevich, A.; Mikheenko, A.; Garg, N.; Nothias, L. F.; Ninomiya, A.; Takada, K.; Dorrestein, P. C.; Pevzner, P. A.; Dereplication of peptidic natural products through database search of mass spectra. *Nature Chemical Biology* **2017**, *13*, 30. [Crossref] [PubMed]
135. Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.; Shcherbin, E.; Nothias, L. F.; Dorrestein, P. C.; Pevzner, P. A.; Dereplication of microbial metabolites through database search of mass spectra. *Nature Communications* **2018**, *9*, 1. [Crossref] [PubMed]
136. Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S.; MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics* **2016**, *8*, 1. [Crossref] [PubMed]
137. Van Santen, J. A.; Poynton, E. F.; Iskakova, D.; Mcmann, E.; Alsup, T. A.; Clark, T. N.; Fergusson, C. H.; Fewer, D. P.; Hughes, A. H.; Mccadden, C. A.; Parra, J.; Soldatou, S.; Rudolf, J. D.; Janssen, E. M. L.; Duncan, K. R.; Linington, R. G.; The Natural Products Atlas 2.0: A database of microbially-derived natural products. *Nucleic Acids Research* **2022**, *50*, D1317. [Crossref] [PubMed]
138. The Natural Products Atlas. Disponível em: <https://www.npatlas.org/>. Acesso em: 26 março 2023.
139. Di Ottavio, F.; Gauglitz, J. M.; Ernst, M.; Panitchpakdi, M. W.; Fanti, F.; Compagnone, D.; Dorrestein, P. C.; Sergi, M.; A UHPLC-HRMS based metabolomics and chemoinformatics approach to chemically distinguish 'super foods' from a variety of plant-based foods. *Food Chemistry* **2020**, *313*, 126071. [Crossref] [PubMed]
140. Remy, S.; Solis, D.; Silland, P.; Neyts, J.; Roussi, F.; Touboul, D.; Litaudon, M.; Isolation of phenanthrenes and identification of phorbol ester derivatives as potential anti-CHIKV agents using FBMN and NAP from *Sagotia racemosa*. *Phytochemistry* **2019**, *167*, 112101. [Crossref] [PubMed]
141. Pham, H. T.; Lee, K. H.; Jeong, E.; Woo, S.; Yu, J.; Kim, W. Y.; Lim, Y. W.; Kim, K. H.; Kang, K. B.; Species Prioritization Based on Spectral Dissimilarity: A Case Study of Polyporoid Fungal Species. *Journal of Natural Products* **2021**, *84*, 298. [Crossref] [PubMed]
142. Maimone, N. M.; de Oliveira, L. F. P.; Santos, S. N.; de Lira, S. P. Elicitation of *Streptomyces lunalinharesii* secondary metabolism through co-cultivation with *Rhizoctonia solani*. *Microbiological Research* **2021**, *251*, 126836. [Crossref] [PubMed]
143. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S.; ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **2016**, *8*, 1. [Crossref] [PubMed]
144. Lee, J.; da Silva, R. R.; Jang, H. S.; Kim, H. W.; Kwon, Y. S.; Kim, J. H.; Yang, H.; In silico annotation of discriminative markers of three *Zanthoxylum* species using molecular network derived annotation propagation. *Food Chemistry* **2019**, *295*, 368. [Crossref] [PubMed]
145. Ernst, M.; Kang, K. Bin; Caraballo-Rodríguez, A. M.; Nothias, L. F.; Wandy, J.; Chen, C.; Wang, M.; Rogers, S.; Medema, M. H.; Dorrestein, P. C.; van der Hooft, J. J. J.; Molnetenhancer: Enhanced molecular networks by integrating metabolome mining and annotation tools. *Metabolites* **2019**, *9*, 1. [Crossref] [PubMed]
146. Ling, L. L.; Schneider, T.; Peoples, A. J.; Spoering, A. L.; Engels, I.; Conlon, B. P.; Mueller, A.; Schäberle, T. F.; Hughes, D. E.; Epstein, S.; Jones, M.; Lazarides, L.; Steadman, V. A.; Cohen, D. R.; Felix, C. R.; Fetterman, K. A.; Millett, W. P.; Nitti, A. G.; Zullo, A. M.; Chen, C.; Lewis, K.; A new antibiotic kills pathogens without detectable resistance. *Nature* **2015**, *517*, 455. [Crossref] [PubMed]
147. Nakashima, H.; Ichiyama, K.; Inazawa, K.; Ito, M.; Hayashi, H.; Nishihara, Y.; Tsujii, E.; Kino, T.; FR901724, a novel anti-human immunodeficiency virus (HIV) peptide produced by Streptomyces, shows synergistic antiviral activities with HIV protease inhibitor and 2',3'-dideoxynucleosides. *Chemical Pharmaceutical Bulletin* **1996**, *19*, 405. [Crossref] [PubMed]
148. Ricart, E.; Pupin, M.; Müller, M.; Lisacek, F.; Automatic Annotation and Dereplication of Tandem Mass Spectra of Peptidic Natural Products. *Analytical Chemistry* **2020**, *92*, 15862. [Crossref] [PubMed]
149. Mohimani, H.; Kim, S.; Pevzner, P. A.; A new approach to evaluating statistical significance of spectral identifications. *Journal of Proteome Research* **2013**, *12*, 1560. [Crossref] [PubMed]
150. Gurevich, A.; Mikheenko, A.; Shlemov, A.; Korobeynikov, A.; Mohimani, H.; Pevzner, P. A.; Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nature Microbiology* **2018**, *3*, 319. [Crossref] [PubMed]
151. VarQuest – Center for Algorithmic Biotechnology. Disopnível em: <https://cab.spbu.ru/software/varquest/>. Acesso em: 26 março 2023.
152. Atencio, L. A.; Boya P., C. A.; Martin H., C.; Mejía, L. C.; Dorrestein, P. C.; Gutiérrez, M.; Genome Mining, Microbial Interactions, and Molecular Networking Reveals New Dibromoalterochromides from Strains of *Pseudoalteromonas* of Coiba National Park-Panama. *Marine Drugs* **2020**, *18*, 456. [Crossref] [PubMed]

153. Amiri Moghaddam, J.; Crüsemann, M.; Alanjary, M.; Harms, H.; Dávila-Céspedes, A.; Blom, J.; Poehlein, A.; Ziemert, N.; König, G. M.; Schäberle, T. F.; Analysis of the Genome and Metabolome of Marine Myxobacteria Reveals High Potential for Biosynthesis of Novel Specialized Metabolites. *Scientific Reports* **2018**, *8*, 1. [Crossref] [PubMed]
154. Velasco-Alzate, K. Y.; Bauermeister, A.; Tangerina, M. M. P.; Lotufo, T. M. C.; Ferreira, M. J. P.; Jimenez, P. C.; Padilla, G.; Lopes, N. P.; Costa-Lotufo, L. V.; Marine Bacteria from Rocas Atoll as a rich source of pharmacologically active compounds. *Marine Drugs* **2019**, *17*, 1. [Crossref] [PubMed]
155. Van Der Hooft, J. J. J.; Wandy, J.; Young, F.; Padmanabhan, S.; Gerasimidis, K.; Burgess, K. E. V.; Barrett, M. P.; Rogers, S.; Unsupervised Discovery and Comparison of Structural Families Across Multiple Samples in Untargeted Metabolomics. *Analytical Chemistry* **2017**, *89*, 7569. [Crossref] [PubMed]
156. Wang, M.; Nothias, L.-F.; van der Hooft, J. J. J.; MS2LDA and MotifDB Substructure Discovery - GNPS Documentation. Disponível em: <https://ccms-ucsd.github.io/GNPSDocumentation/ms2lda/>. Acesso em: 26 março 2023.
157. Wandy, J.; Zhu, Y.; Van Der Hooft, J. J. J.; Daly, R.; Barrett, M. P.; Rogers, S.; Ms2lda.org: Web-based topic modelling for substructure discovery in mass spectrometry. *Bioinformatics* **2018**, *34*, 317. [Crossref] [PubMed]
158. Jarmusch, S. A.; Lagos-Susaeta, D.; Diab, E.; Salazar, O.; Asenjo, J. A.; Ebel, R.; Jaspars, M.; Iron-meditated fungal starvation by lupine rhizosphere-associated and extremotolerant *Streptomyces* sp. S29 desferrioxamine production. *Molecular Omics* **2021**, *17*, 95. [Crossref] [PubMed]
159. Nothias-Esposito, M.; Nothias, L. F.; Da Silva, R. R.; Retailleau, P.; Zhang, Z.; Leyssen, P.; Roussi, F.; Touboul, D.; Paolini, J.; Dorrestein, P. C.; Litaudon, M.; Investigation of Premyrsinane and Myrsinane Esters in *Euphorbia cupanii* and *Euphobia pithyusa* with MS2LDA and Combinatorial Molecular Network Annotation Propagation. *Journal of Natural Products* **2019**, *82*, 1459. [Crossref] [PubMed]
160. Zdouc, M. M.; Iorio, M.; Maffioli, S. I.; Crüsemann, M.; Donadio, S.; Sosio, M.; *Planomonospora*: A Metabolomics Perspective on an Underexplored Actinobacteria Genus. *Journal of Natural Products* **2021**, *84*, 204. [Crossref] [PubMed]
161. Peñaloza, E.; Holandino, C.; Scherr, C.; Araujo, P. I. P. de; Borges, R. M.; Urech, K.; Baumgartner, S.; Garrett, R.; Comprehensive Metabolome Analysis of Fermented Aqueous Extracts of *Viscum album* L. by Liquid Chromatography–High Resolution Tandem Mass Spectrometry. *Molecules* **2020**, *25*, 4006. [Crossref] [PubMed]
162. Soldatou, S.; Eldjárn, G. H.; Ramsay, A.; van der Hooft, J. J. J.; Hughes, A. H.; Rogers, S.; Duncan, K. R.; Comparative Metabologenomics Analysis of Polar Actinomycetes. *Marine drugs* **2021**, *19*, 1. [Crossref] [PubMed]
163. Kang, K. Bin; Woo, S.; Ernst, M.; van der Hooft, J. J. J.; Nothias, L. F.; da Silva, R. R.; Dorrestein, P. C.; Sung, S. H.; Lee, M.; Assessing specialized metabolite diversity of *Alnus* species by a digitized LC–MS/MS data analysis workflow. *Phytochemistry* **2020**, *173*, 112292. [Crossref] [PubMed]
164. Böcker, S.; Rasche, F.; Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics* **2008**, *24*, 49. [Crossref] [PubMed]
165. Vaniya, A.; Fiehn, O.; Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *Trends in Analytical Chemistry* **2015**, *69*, 52. [Crossref] [PubMed]
166. Rasche, F.; Scheubert, K.; Hufsky, F.; Zichner, T.; Kai, M.; Svatoš, A.; Böcker, S.; Identifying the unknowns by aligning fragmentation trees. *Analytical Chemistry* **2012**, *84*, 3417. [Crossref] [PubMed]
167. Böcker, S.; Dührkop, K.; Fragmentation trees reloaded. *Journal of Cheminformatics* **2016**, *8*, 1. [Crossref] [PubMed]
168. Ludwig, M.; Nothias, L. F.; Dührkop, K.; Koester, I.; Fleischauer, M.; Hoffmann, M. A.; Petras, D.; Vargas, F.; Morsy, M.; Aluwihare, L.; Dorrestein, P. C.; Böcker, S.; Publisher Correction: Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nature Machine Intelligence* **2020**, *2*, 727. [Crossref]
169. Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S.; Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112*, 12580. [Crossref] [PubMed]
170. Dührkop, K.; Nothias, L. F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M. A.; Petras, D.; Gerwick, W. H.; Rousu, J.; Dorrestein, P. C.; Böcker, S.; Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology* **2021**, *39*, 462. [Crossref] [PubMed]