http://dx.doi.org/10.21577/0100-4042.20170838 *Quim. Nova,* Vol. 45, No. 5, 608-620, 2022

Nota Técnica

GUIA PARA PROCESSAMENTO DE DADOS DE CROMATOGRAFIA ACOPLADA A ESPECTROMETRIA DE MASSAS

Ricardo Moreira Borges[a,*], João Victor Mendes Resende[a], Aldebaran Oliveira de Moraes[a], Alana Kelyene Pereira[b], Rafael Garrett[c], Anelize Bauermeister[d,e] e Antonio Jorge Ribeiro da Silva[a]

[a]Instituto de Pesquisas de Produtos Naturais Walter Mors, Universidade Federal do Rio de Janeiro, 21941-902 Rio de Janeiro – RJ, Brasil

[b]Instituto de Química, Universidade Estadual de Campinas, 13083-970 Campinas – SP, Brasil

[c]Instituto de Química, Universidade Federal do Rio de Janeiro, 21941-909 Rio de Janeiro – RJ, Brasil

[d]Instituto de Ciências Biomédicas, Universidade de São Paulo, 05508-000, São Paulo – SP, Brasil

[e]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 92093, San Diego – CA, Estados Unidos

Recebido em 28/07/2021; aceito em 05/10/2021; publicado na web em 11/11/2021

GUIDE FOR CHROMATOGRAPHY COUPLED TO MASS SPECTROMETRY DATA PROCESSING. In this work, a discussed and step-wise tutorial for LC-MS and GC-MS data processing using the open-access software MZMine2 is presented and discussed. The rationale behind each step was demonstrated to enable the readers to go through their own data and process it accordingly. The main lesson to be learned is that each parameter must be chosen in light of the raw data and no guidelines should suggest a predetermined value. Still, it is worth mentioning that ideal values for each parameter do not exist, and that the user might end up investing too much time futilely optimizing values. Our suggestion is to process your data in light of the raw data (and the study design) following the *preview figure* result and the resulting *feature list* generated in each processing step, interpret your data, and go back to process it again to tune the detection of important *features*.

Keywords: mass spectrometry; data processing; MZMine2; GC-MS; LC-MS.

INTRODUÇÃO

A espectrometria de massas (MS) é uma das ferramentas de detecção e produção de informações estruturais mais utilizada em química orgânica e de produtos naturais (PN). Juntamente com a ressonância magnética nuclear (RMN), forma um grupo importante de técnicas instrumentais para identificação estrutural inequívoca de moléculas.[1] No contexto da relevância das análises por MS, deve-se incluir as técnicas acopladas que utilizam a cromatografia como forma de separação dos compostos: cromatografia em fase gasosa acoplada à MS (GC-MS, do inglês *gas chromatography-mass spectrometry*) e cromatografia em fase líquida acoplada à MS (LC-MS, do inglês *liquid chromatography-mass spectrometry*). Dentre as diversas vantagens e abrangência desses sistemas de separação e detecção, destacam-se, principalmente, as reduções de efeitos de matriz (p. ex. supressão iônica) e a separação de compostos isóbaros.[2] Abrangência de métodos, se refere à possibilidade de separar compostos de diferentes características físico-químicas usando colunas cromatográficas com fases estacionárias diversas,[3] como apolares, polares, de troca iônica e mecanismos mistos. Adicionalmente, é possível analisar os compostos separados pela cromatografia usando diferentes formas de ionização (p. ex. ionização por elétrons na GC-MS e ionização por *electrospray* nos modos positivo e/ou negativo na LC-MS),[4] diferentes modos de aquisição de dados (p. ex. aquisição dependente de dados – DDA e independente de dados – DIA em espectrômetros de massas híbridos)[5] e com medida de massas em baixa ou alta resolução.

Dados de cromatografia acoplada à MS de amostras complexas, como as de origem natural, geralmente apresentam cromatogramas também complexos com sobreposição de sinais cromatográficos. O processo de desconvolução pode auxiliar na separação dos mesmos. Desconvolução se refere ao processo de resolução de sinais sobrepostos em seus constituintes únicos ou *features*; esse termo será usado para se referir a uma informação produzida pela combinação de um dado de MS e um dado da dimensão cromatográfica. Deve-se reconhecer que é crescente o uso dessas ferramentas em estudos envolvendo PN e metabolômica em suas diversas vertentes.[1,6-8] Hoje, diversos grupos de pesquisa no mundo que atuam na área de PN usam a LC-MS e/ou GC-MS em seus estudos, inclusive para protocolos de desreplicação[6,8] de extratos brutos e frações enriquecidas em determinadas classes de compostos. No entanto, percebe-se que muitos pesquisadores ainda deixam de processar (e analisar em profundidade) os dados de suas próprias amostras, trabalhando até com arquivos impressos de espectros obtidos de sinais cromatográficos escolhidos manualmente. Portanto, é clara a necessidade de uma discussão maior e divulgação de metodologias de processamento de dados em detalhe e de forma prática.

O objetivo deste trabalho é explicitar as etapas envolvidas e fornecer referências para ajudar os grupos de pesquisa nacionais, principalmente os alunos de graduação e pós-graduação, a escolherem os melhores parâmetros para processar seus dados de LC-MS e GC-MS. Para isso, o *software* MZMine2[9] será usado como ferramenta de processamento de dados pelos seguintes motivos: (1) ele é um programa gratuito e de código aberto; (2) permite a visualização prévia dos resultados a cada etapa do processamento; (3) permite a inclusão de novos *scripts* com métodos variados para processamento com diferentes objetivos; (4) é uma das ferramentas mais utilizadas mundialmente (junto com o XCMS[10] e MS-Dial[11]). Não será excluída a possibilidade de os usuários escolherem outras ferramentas de preferência, desde que a escolha dos parâmetros seja feita de forma racional.

*e-mail: ricardo_mborges@ufrj.br

Tabela 1. Dados e origem dos dados do repositório Metabolomics Workbench utilizados nesta demonstração

LC-MS			
Tipo de amostra	Origem (código)	Referência/DOI	Comentários
Planta	ST000240 (Material Suplementar 2.A)	10.21228/M8FK5P	ACE Excel 2 C18-PFP (100 x 2,1mm, 2um).
Coral	PR000747 (ST001163)[15] (Material Suplementar 2.B)	10.21228/M8469M	ACE Excel 2 C18-PFP (100 x 2,1mm, 2um)
Microrganismo	ST001199 (Material Suplementar 2.C)	10.21228/M8CD73	Acquity CSH Phenyl Hexyl (1,7 uM, 1,0 x 100 mm)
Urina humana	ST001122 (Material Suplementar 2.D)	10.21228/M8GD6N	Waters Acquity BEH Amide (150 x 2,1mm, 1,7um)
Plasma humano	ST000601 (Material Suplementar 2.E)	10.21228/M8FC7C	Agilent Zorbax RRHD SB-C18 (100 x 2,1mm, 1,8um)
GC-MS			
Tipo de amostra	Origem (código)	Referência/DOI	Comentários
Planta	ST001056[16] (Material Suplementar 2.F)	10.21228/M81M4X	TG-5MS
Inseto	ST000025[17] (Material Suplementar 2.G)	10.21228/M85P4V	Restek Rtx-5Sil MS (30 x 0,25mm, 0,25um)

*A discussão será feita com foco em cada etapa do processamento utilizando os dados, mas não dando sequência completa ao processamento de cada dado.

PARTE EXPERIMENTAL

O *software* MZMine2 pode ser obtido gratuitamente em seu site oficial (http://mzmine.github.io/, data de acesso 26/07/2021), através da aba *download*. Para este guia, foi utilizada a versão 2.53. Para utilizar o *software*, é necessária a instalação prévia do Java SE runtime (JRE) versão 1,7 ou posterior. Não é necessário instalar o MZMine2 em seu computador, basta descompactá-lo após seu *download* e executar o arquivo referente ao seu sistema operacional (p. ex. Microsoft Windows, Mac OS X ou Linux).

Para demonstrar cada etapa do processamento, foram utilizados dados de GC-MS e LC-MS obtidos no repositório aberto Metabolomics Workbench (https://www.metabolomicsworkbench.org/). Foram selecionados grupos de dados de diferentes naturezas, incluindo de origem humana, como plasma e urina, insetos, organismos marinhos, microrganismos e plantas (Tabela 1). Este trabalho é um exemplo da importância de repositórios de dados públicos, tais como o Metabolomics Workbench[12] que têm papel importante nos princípios de *FAIR*.[13] Com objetivo de expandir o uso de dados coletados em repositórios, um tutorial mostrando as etapas para *download* de dados para os repositórios Metabolomics Workbench e MassIVE foi montado (Material Suplementar 1).

Os dados de LC-MS utilizados foram previamente convertidos ao formato .mzXML com o *software* MSConvert[14] antes de serem submetidos ao processamento de dados. Os dados de GC-MS utilizados foram obtidos no formato .cdf, produzidos pelo *software* nativo do próprio equipamento de aquisição. Esses arquivos .mzXML e .cdf foram então importados para o *software* MZMine2. Este aceita como dados de entrada os formatos .NetCDF, .mzXML, .mzData, .mzXML, dentre outros formatos, tais como Thermo RAW files, Waters RAW folders, Agilent CSV files e Compressed files. Os parâmetros utilizados em cada etapa dos processamentos para essa demonstração são dispostos em cada momento de uso na discussão deste trabalho.

RESULTADOS E DISCUSSÃO

Os autores declaram que os resultados apresentados neste trabalho não implicam em nenhum juízo de valor sobre as informações químicas geradas em cada projeto ou as interpretações biológicas resultantes dos dados químicos. Nosso objetivo se limita a criar um guia comentado em língua portuguesa para tornar a etapa de processamento de dados mais acessível e disseminada entre todos os usuários das ferramentas de LC-MS e GC-MS.

Etapas do processamento

O objetivo deste trabalho é a discussão dos parâmetros a serem utilizados em diferentes etapas de processamento de dados de LC-MS e GC-MS. Logo, não será descrito um procedimento para processamento completo de um estudo. Aqui, alguns projetos disponíveis em repositório *online* foram selecionados e processados por alunos em estágio inicial de treinamento na área de metabolômica. Informações sobre esses processamentos estão descritas no material suplementar.

Serão discutidos os efeitos de diferentes parâmetros em cada etapa do processamento para os diferentes projetos escolhidos no repositório *online* (Quadro 1): (1) Detecção e listagem dos sinais de MS; (2) Construção dos cromatogramas de íons extraídos dos sinais detectados; (3) Desconvolução dos cromatogramas de íons extraídos; (4) Desconvolução de espectros de MS (específico para GC-MS); (5) Reconhecimento do perfil isotópico; (6) Alinhamento de *features* entre amostras; e (7) Aplicação de filtros.

Quadro 1. Etapas do procedimento de processamento de dados de GC-MS e LC-MS usando o *software* MZMine2

Etapas para GC-MS	Etapas para LC-MS
Importação de dados	Importação de dados
Listagem de sinais de *m/z**	Listagem de sinais de *m/z**
Construção de EICs (dados de MS1 apenas)	Construção de EICs (dados de MS1 apenas)
Desconvolução cromatográfica	Desconvolução cromatográfica
Desconvolução espectral**	Agrupamento de perfil isotópico***
Alinhamento entre amostras	Alinhamento entre amostras
Aplicação de filtros****	Aplicação de filtros****

*incluir MS e MS/MS (apenas quando há dados de MS/MS). **etapa específica para GC-MS com ionização por elétrons. *** etapa dependente da resolução usada no espectrômetro de massas. ****etapa opcional.

Detecção e listagem dos sinais de MS (etapa comum para LC-MS e GC-MS)

[MZMine2| Menu: Raw data methods >> Feature Detection >> Mass detection] (Figura 1S)

Nessa etapa, o objetivo é indicar ao *software* quais sinais em MS1 (dados adquiridos em modo de varredura total de íons) devem ser considerados para as próximas etapas. É importante definir um limite máximo permitido para o nível de ruído instrumental (parâmetro *Noise level*). Em geral, cada tipo de equipamento/analisador de massas apresenta um nível diferente de ruído. Por exemplo, em analisadores de massas do tipo Orbitrap o valor geralmente utilizado para o ruído é 10^5 (entrada no *software* como 1E5). Enquanto que para analisadores do tipo TOF esse valor é de 10^3 (1E3). Porém, esses valores podem variar de acordo com a amostra a ser analisada, parâmetros de

aquisição de dados e fabricantes dos equipamentos. Sugere-se a análise dos dados brutos das amostras para esse tipo de avaliação.

O procedimento indicado para essa etapa de detecção e listagem dos sinais de MS é: (1) avaliar diferentes regiões da linha de base (incluindo o período do cromatograma relativo ao volume morto da análise cromatográfica) e anotar os valores de *m/z* e intensidades dos sinais; (2) avaliar diferentes sinais cromatográficos de intensidade baixa e anotar os valores de *m/z* e intensidades para comparar com os sinais anotados anteriormente; e (3) definir e informar um valor seguro, que não se sobreponha à intensidade de sinais referentes à metabólitos (Figura 1).

Uma sugestão para essa etapa seria utilizar valores mais baixos e incluir algum ruído nesse momento para considerar flutuações da linha de base, uma vez que outras etapas do processamento poderão excluir sinais indesejáveis. A escolha do valor usado para filtrar o ruído deve sempre ser feita utilizando a ferramenta *Show preview*, observando, quais sinais estão sendo excluídos. Na Figura 1, é demonstrada a seleção de pontos da linha de base em diferentes tempos de retenção para visualização das intensidades dos sinais no espectro de MS (Figura 1; setas vermelhas) e a seleção de sinais cromatográficos relativamente pequenos para visualização das intensidades mínimas a serem consideradas como sinais (Figura 1; setas azuis). Deve-se desconsiderar os sinais de MS constantes ao longo da linha de base como responsáveis por sinais cromatográficos. Vale também mencionar que dificilmente haverá um valor perfeito para listar e desconvoluir com sucesso todos os sinais detectados pelo instrumento, mas esse processo só será reprodutível com um processamento computacional devidamente documentado. É necessário informar também a janela de tempo de retenção (*Retention time*) e nível de MS (*MS level*, com o *Browser* em *Mass detector*) do qual a listagem será feita.

Como resultado dessa etapa de listagem dos sinais feita para

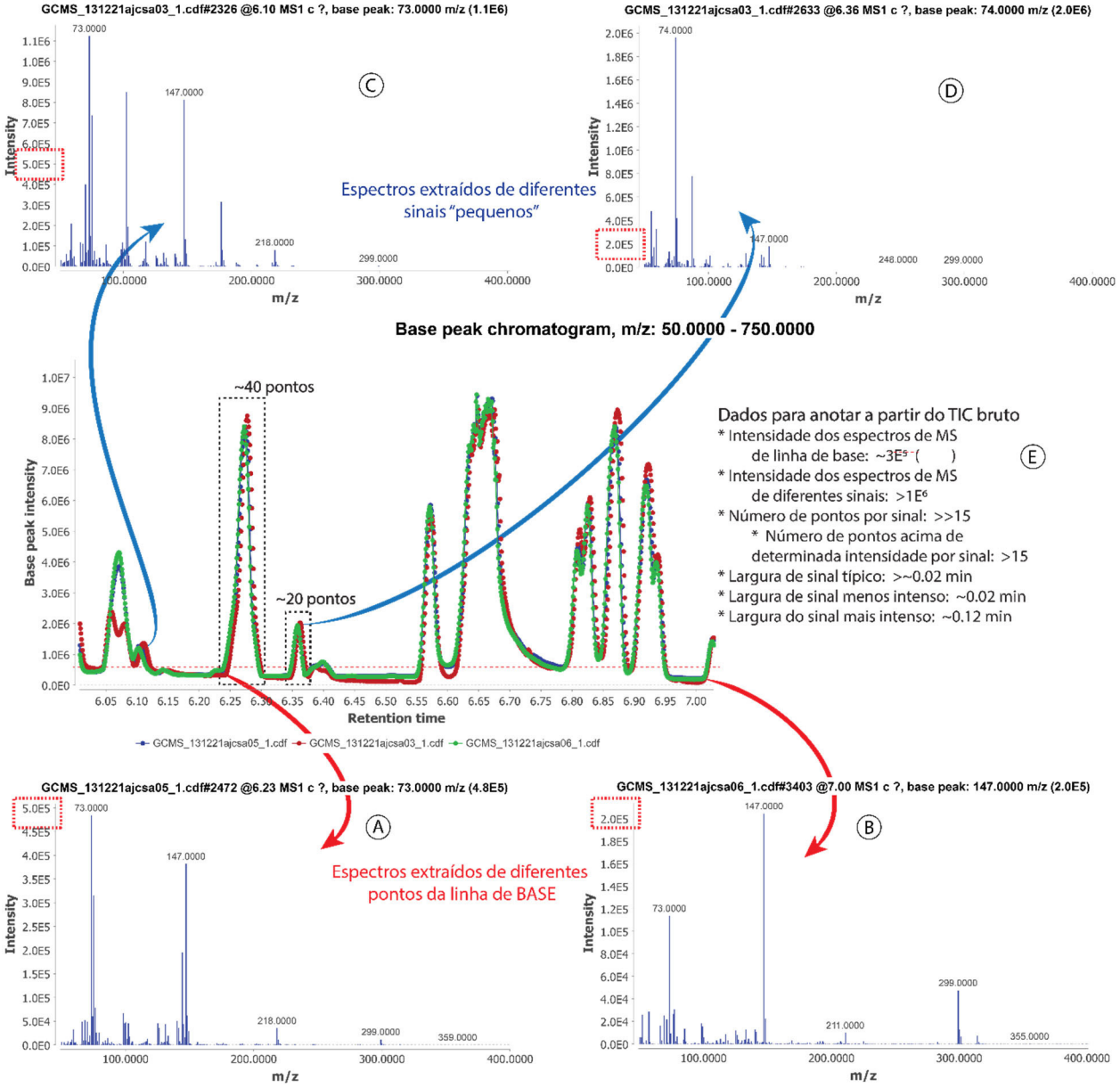

Figura 1. *Procedimento para análise dos dados brutos (TIC – total ion current chromatogram) e obtenção dos parâmetros fundamentais para o processamento (Dado: ST000025, GC-MS). A: Espectro de MS obtido de um ponto de linha de base (6,23 min); B: Espectro de MS obtido de outro ponto de linha de base (7,08 min); C: Espectro de MS obtido de um sinal cromatográfico de baixa intensidade (6,10 min); D: Espectro de MS obtido de um sinal cromatográfico de alta intensidade (6,36 min); E: Informações a serem observadas e coletadas para dar sequência ao processamento*

cada *scan* do(s) arquivo(s) importado(s), serão criadas variáveis (listas de massas; *Mass list*) ainda na janela *raw data files* para cada *scan*. Essa informação pode ser conferida abrindo cada uma dessas listas do *Mass list,* ou visualmente, abrindo uma figura com os dados coloridos conforme estão ou não listados (Figura 2). Observem na Figura 2 que, mesmo nos espectros de MS de pontos da linha de base, aparecem alguns sinais marcados em vermelho. Em se tratando de dados adquiridos por experimento de DDA na espectrometria de massas sequencial (*tandem* MS ou MS/MS), por exemplo, esses irão incluir espectros de fragmentação que também devem ter seus sinais listados da mesma forma que foi feito para MS acima. Também é possível (e indicado) a visualização para os espectros de MS/MS para seleção do *Noise level.* A intensidade desses sinais é geralmente duas ordens de magnitude menor que o MS1. Nesse caso, deve-se selecionar a opção *MS level* como sendo 2 em *Set filters*.

Construção dos cromatogramas de íons extraídos (etapa comum para LC-MS e GC-MS)

[MZMine2| Menu: Raw data methods >> Feature Detection >> ADAP chromatogram builder] (Figura 2S)

Nessa etapa, aqueles dados listados nas *Mass list* são usados para a criação do cromatograma de íon extraído (EIC – *extracted ion chromatogram*) para cada sinal listado. Uma série de parâmetros são utilizados para determinar valores de *m/z* que são detectados continuamente e o que pode ser aceito como um sinal cromatográfico. O *script Automated Data Analysis Pipeline* (ADAP)[18] é o método mais indicado e ele atua com base nos parâmetros: (1) nome da lista de massas definida pela etapa anterior (*Mass list name*); (2) número de pontos/*scan* que podem ser usados para caracterizar um sinal cromatográfico (*Min group size in # of scans*); (3) intensidade mínima que um ponto deve ter para ser considerado como parte de um sinal cromatográfico (*Group intensity threshold*); (4) intensidade mínima da altura máxima de um sinal cromatográfico para que esse seja considerado (*Min highest intensity*) e; (5) tolerância máxima permitida para que valores de *m/z* de diferentes *scans* possam ser alinhados em um *feature* (*m/z tolerance* no MZmine2). Geralmente aqui é utilizado o valor do erro experimental em massa do equipamento usado para a aquisição dos dados. Por exemplo, *m/z tolerance* de 0,02-0,005 (*m/z*) para dados de alta resolução/exatidão (Orbitraps e TOFs).[19] Quando o parâmetro *ppm* em *m/z tolerance* estiver 0,0, ele será desconsiderado, e vice-versa. Por outro lado, para dados adquiridos em instrumentos de baixa resolução/exatidão (comum em GC-MS com analisadores de massas quadrupolo ou armadilha de íons), uma variação unitária (ou nominal) pode ser admitida nesse parâmetro (atenção a situações em que a exportação dos dados foi feita de modo a reportar valores unitários apenas, p. ex. o dado ST000025).

Observem que não há uma etapa de exclusão de ruído de linha de base (*background subtraction*). Isso é justificável, pois a definição dos sinais cromatográficos feita pelo *script* ADAP tende a ser bem sucedida, já que, idealmente, ruído não tem formato de sinal cromatográfico.[20] De forma mais clara, essa etapa vai considerar um EIC válido quando ele tiver, pelo menos, um certo número de pontos para formar um sinal cromatográfico (*Min group size in # of scans*) acima do valor informado para *Group intensity threshold* e produzindo um sinal com valor máximo acima do *Min highest intensity* também informado pelo usuário.

A escolha dos parâmetros nessa etapa de construção dos cromatogramas não pode ser acompanhada usando a visualização

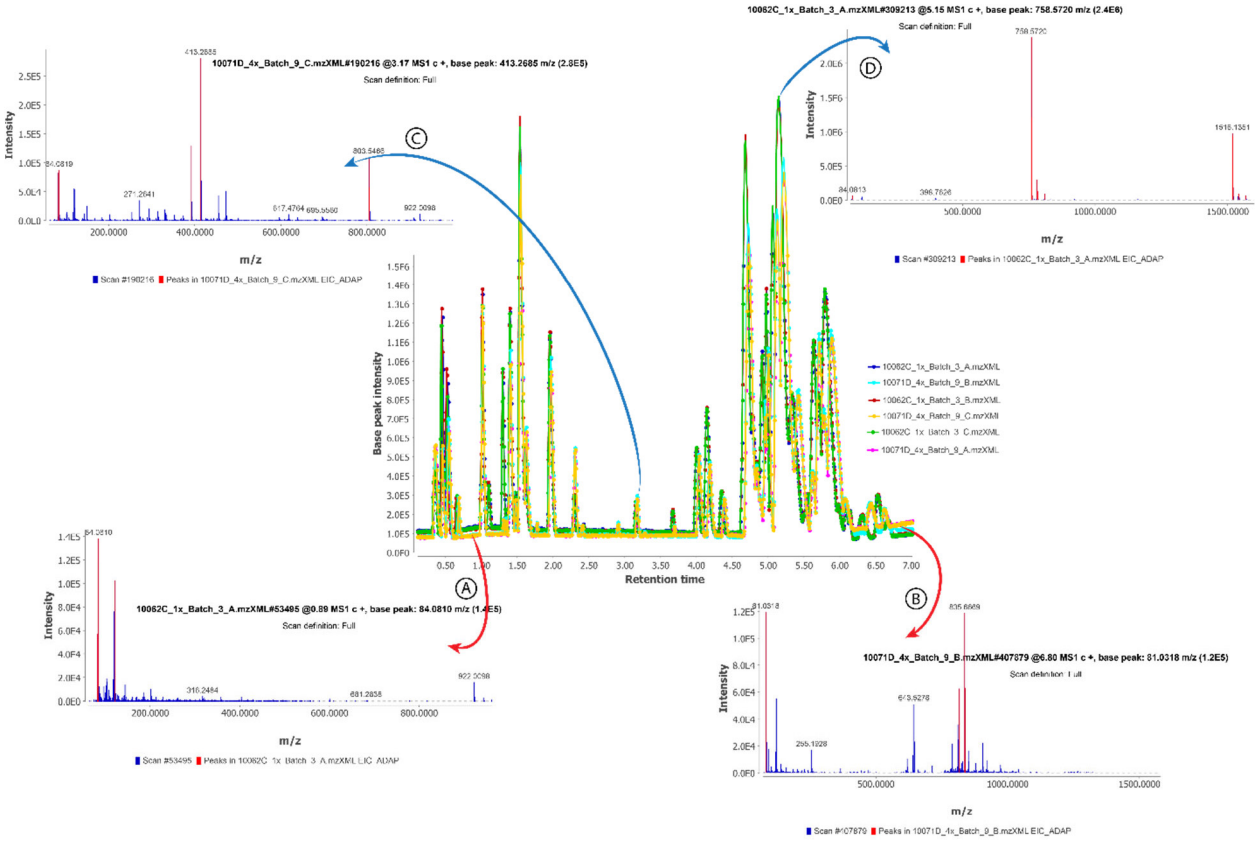

Figura 2. *Demonstração da visualização dos sinais nos espectros de MS em cada scan (Dado: ST000601, LC-MS). Os sinais marcados em vermelho são aqueles listados em Mass lists. Os sinais nos espectros de MS não marcados em vermelho (em azul) foram excluídos pelo Noise level (neste caso, estabelecido como sendo 6E4). A: Espectro de MS obtido de um ponto de linha de base (0,89 min); B: Espectro de MS obtido de outro ponto de linha de base (6,80 min); C: Espectro de MS obtido de um sinal cromatográfico de baixa intensidade (3,17 min); D: Espectro de MS obtido de um sinal cromatográfico de alta intensidade (5,15 min)*

prévia (não existe a ferramenta *Show preview*), mas é indicado que sejam avaliados diretamente a partir dos cromatogramas de íons totais (TIC) obtidos dos dados brutos. O procedimento indicado é: (1) abrir os TIC para os dados brutos importados (função oferecida com o uso do 'botão direito' do *mouse* sobre cada *raw file*), incluindo as amostras branco (que compõem parte dos dados de controle de qualidade; sempre indicado) utilizando a ferramenta *Show TIC* no MZMine2 com os parâmetros desejados; (2) selecionar sinais cromatográficos de tamanhos variados para anotação dos parâmetros requeridos pelo método ADAP *Chromatogram builder* no MZMine2 (Figura 1).

É esperado que experimentos de aquisição de dados no MS que produzam mais pontos (*scans*) por sinal cromatográfico acabem por permitir uma definição de sinal mais bem sucedida. Para isso, o conceito de tempo do ciclo de varredura (*scan cycle time*) deve ser bem compreendido pelo analista/operador do instrumento. Por exemplo, deve-se considerar que experimentos que envolvam aquisição de dados em modo positivo e negativo alternadamente (*fast switch*) na ionização por *electrospray* em LC-MS podem resultar em um *scan cycle time* elevado e limitar a capacidade do *script* ADAP *chromatogram builder* já que cada sinal cromatográfico será definido por menos pontos. Existirá uma chance maior do ruído ser erroneamente definido como uma *feature*. Da mesma forma, é notável que na maioria dos experimentos adquiridos com um analisador de alta velocidade de transmissão (p.ex.: quadrupolo simples), o *scan cycle time é* mais curto do que aqueles experimentos adquiridos em analisadores de mais baixa velocidade de transmissão (p.ex.: por transformada de Fourier e ressonância ciclotrônica de íons). Logo os sinais cromatográficos em GC-MS são normalmente construídos por mais de 10 pontos (Figura 3).

Como resultado dessa etapa, observa-se a construção de uma série de EICs para cada dado em processamento em listas de *features* e essas variáveis serão expostas na janela *Feature list* ao lado direito da janela principal do MZMine2. A partir desse ponto, quando um *feature* possuir dados de MS/MS associados, esses estarão presentes em cada *feature.* Porém, as próximas etapas são feitas apenas com os EICs produzidos utilizando apenas dados de MS em nível 1 (*MS level 1*), ou de íons precursores. É muito importante que o usuário entenda que é possível acompanhar cada etapa de processamento visualizando cada sinal cromatográfico e espectro de MS/MS associado, inclusive monitorando aqueles sinais de maior interesse.

Desconvolução dos cromatogramas de íons extraídos (etapa comum para LC-MS e GC-MS)

[MZMine2| Menu: Feature list methods >> Feature Detection >> Chromatogram deconvolution] (Figura 3S)

Em diversas ocasiões, nos EICs produzidos pela etapa anterior, serão observados mais do que 1 sinal cromatográfico por *feature* (Figura 4). Isso é resultado da ocorrência de sinais de *m/z* de mesmo valor (dentro do *m/z threshold*) em diferentes tempos de retenção e, em muitos casos, são isóbaros detectados em tempos de retenção diferentes. É necessário separar esses *features* distintos para que se possa analisá-los à luz da expectativa de que cada *feature* deve caracterizar apenas um componente de uma amostra. É sabido que na ionização por *electrospray* na LC-MS um único componente pode ser detectado em diferentes formas, caracterizando diferentes *features*, como no caso dos adutos $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, dentre muitos outros.[21] A esse procedimento, de separação de dois ou mais sinais cromatográficos em *features* únicos, dá-se o nome de desconvolução cromatográfica (*Chromatogram deconvolution*). Ela é aplicada diretamente nas listas de *features* produzida pela etapa anterior, ou seja, nos EICs.

Entre os métodos disponíveis para a desconvolução dos sinais cromatográficos, os disponíveis no MZMine2 são: (1) Baseline cut-off; (2) Noise amplitude; (3) Savitzky-Golay; (4) Local minimum search; (5) Wavelets (XCMS); e (6) Wavelets (ADAP).

Os métodos *Baseline cut-off* e *Noise amplitude* são mais simples e indicados para um pequeno conjunto de dados com sinais cromatográficos bem definidos. De forma geral, ambos requerem apenas altura mínima (*Min peak height*), a largura mínima de um sinal cromatográfico (*Min peak duration*) e o nível do ruído, que é o que diferencia os dois métodos. *Baseline cut-off* considera apenas o nível da linha de base (*Baseline level*) e por isso é indicado para dados que apresentem baixo valor de ruído e linha de base continua. Já o *Noise amplitute* considera a amplitude/variação do ruído, e portanto, é mais indicado para dados com variações do nível do ruído durante uma mesma análise cromatográfica e entre os cromatogramas. A partir dos valores escolhidos para esses parâmetros, os sinais cromatográficos que aparecem acima do nível de ruído serão integrados.

Os parâmetros a serem informados são: altura mínima de um sinal cromatográfico (*Min peak height*), largura mínima de um sinal cromatográfico (*Min peak duration*) e a intensidade mínima aceitável para a segunda derivada (*Derivative threshold level*).

O método *Local minumum search* é mais indicado para casos onde há a relação sinal/ruído alta e com sinais cromatográficos bem definidos. É um método ideal para análises-alvo (ou quando há sinais limpos, intensos e com linha de base mínima). Ele busca por pontos mais baixos entre dois sinais cromatográficos para traçar os limites de integração respeitando parâmetros de intensidade e duração dos sinais. Um cromatograma com intensidade alta de ruído ou sinais cromatográficos muito serrilhados podem representar uma limitação,

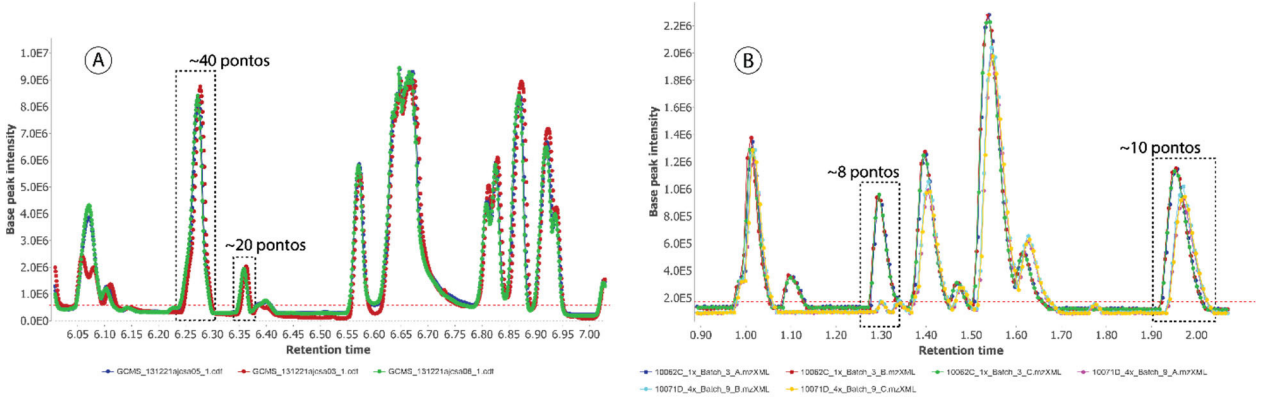

Figura 3. *Diferença do número de pontos que compõem um sinal cromatográfico entre uma análise feita em GC-MS (A) (dado: ST000025) e em LC-MS usando uma coluna cromatográfica de interação hidrofílica-HILIC (B) (dado: ST000601). Esse parâmetro, assim como todos os outros, devem ser escolhidos à luz dos dados brutos de cada estudo, incluindo diferentes amostras dentro do estudo*

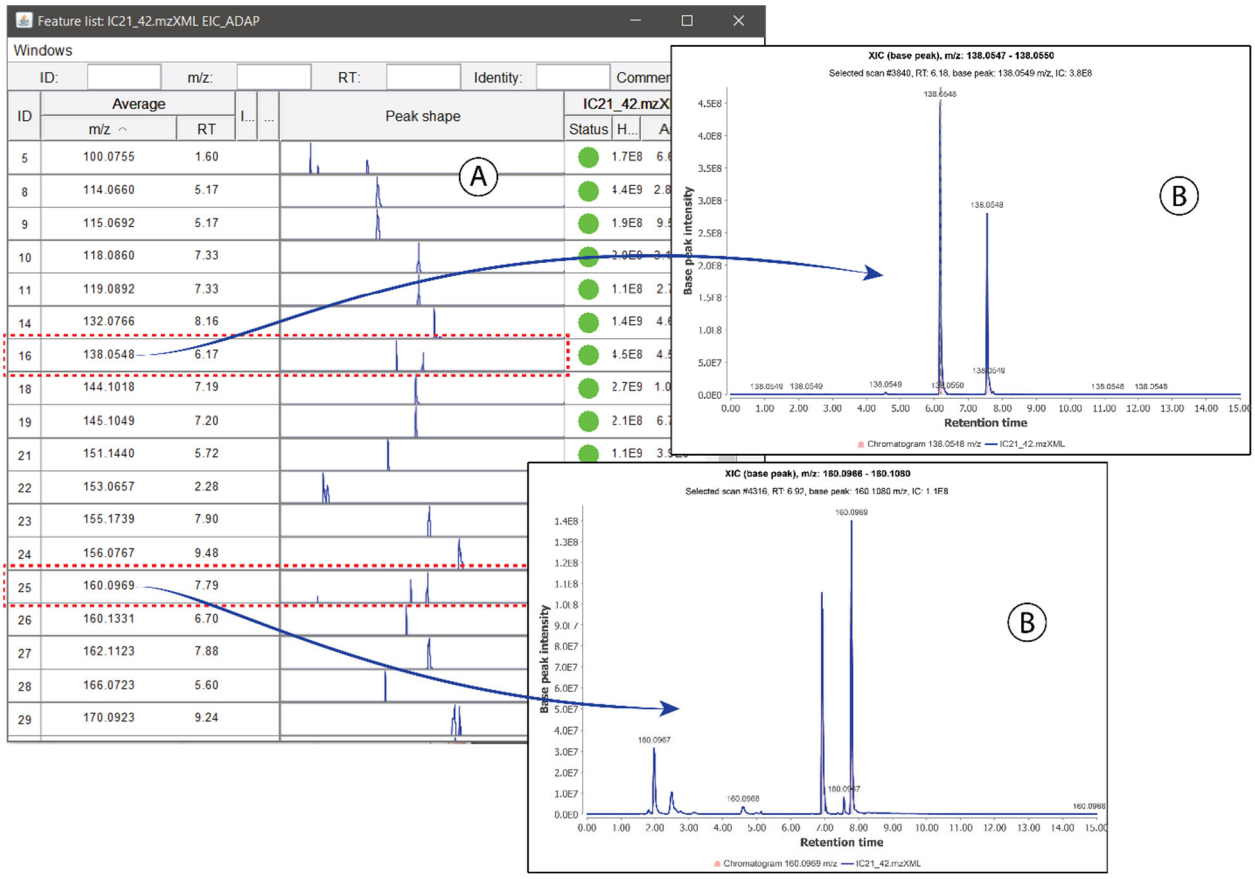

Figura 4. *Demonstração da necessidade de desconvolução de sinais cromatográficos (Dado: ST001122, LC-MS). Idealmente cada feature deverá compor um dado de MS e um dado de tempo de retenção apenas. A: Lista de features obtida logo após a etapa de Chormatogram builder; B: Exemplos de EICs de dois m/z diferentes que precisam ser desconvoluídos para fornecer apenas um sinal cromatográfico para cada m/z*

já que esse método irá reconhecer sinais em demasia e dividi-los em muitos *features* diferentes. Os parâmetros a serem otimizados são: (1) limite mínimo para pontos detectados para serem considerados um sinal cromatográfico em porcentagem (*Chromatographic threshold*); (2) período de tempo mínimo considerado para a detecção de um ponto mínimo local que caracteriza a separação mínima entre dois sinais cromatográficos para se buscar por um 'vale' (*Search minimum in RT range (min)*; parâmetro sensível em situações onde há sinais "serrilhados"); (3) intensidade mínima de um sinal cromatográfico em relação ao ponto mais alto do cromatograma (*Minimum relative height*); (4) intensidade absoluta mínima para que um sinal cromatográfico seja considerado (*Minimum absolute height*); (5) razão mínima entre o ponto mais alto de um sinal cromatográfico e os pontos adjacentes (*Min ratio of peak top/edge*; parâmetro sensível em situações onde há sinais "serrilhados"); (6) limite inferior de duração de um sinal cromatográfico (*Peak duration range*).

O método *Wavelets (XCMS)* necessita que o usuário do MZMine2 instale um programa paralelo gratuito, chamado de R, com um conjunto de funções que estão disponíveis no sítio http://bioconductor.org/biocLite.R. Como se trata de um método bastante similar ao mais moderno *Wavelets (ADAP)*, que será priorizado em seguida, não será detalhado aqui, mas pode ser encontrado em uma referência para os leitores interessados.[22]

O método *Wavelets (ADAP)*[23-26] promove a detecção de sinais cromatográficos dentre os EICs usando a ferramenta transformada *wavelet* contínua (CWT)[27] que tem se mostrado útil em diversos segmentos de pesquisas que envolvem detecção de sinais não estacionários, como é, aproximadamente, o caso para cromatogramas. Os coeficientes *wavelet* são calculados em diferentes escalas e localizações (parâmetros esses informados pelos usuários) e a determinação de localização e fronteiras dos sinais cromatográficos são estabelecidos. Não cabe aqui discutir em detalhes o uso de CWT na detecção de sinais, mas algumas referências podem ser indicadas para o leitor mais interessado.[19,20,23-29] De forma mais clara, quanto maiores os coeficientes *wavelet*, maior a confiança no sinal cromatográfico detectado. Os parâmetros a serem otimizados são: (1) limite mínimo para razão sinal-ruído (*S/N Threshold*); (2) método para estimativa de ruído (*S/N estimator*); (3) limite inferior de intensidade para um sinal cromatográfico (*min feature height*); (4) razão entre o maior coeficiente *wavelet* e a área sob a curva de cada sinal cromatográfico que é usado para excluir sinais menos consistentes, já que a tendência é de que sinais com áreas muito grandes produzam coeficientes maiores (normalizando os coeficientes *wavelet* pelas áreas; *coefficient/area threshold*); (5) variação aceitável para largura dos sinais cromatográficos (*Peak duration range*); (6) variação limite de escalas *wavelet* a serem utilizadas na construção das matrizes dos coeficientes expressos em tempo de retenção em minutos que serão aplicadas ao cromatograma e que devem ser escolhidos conforme as larguras dos sinais cromatográficos observados no cromatograma (*RT wavelet range*) (Figura 5). Na Figura 5 observa-se uma situação de escolha do parâmetro mais crítico nessa etapa, o *RT wavelet range*. O objetivo dessa etapa é tentar separar sinais cromatográficos produzidos por constituintes isóbaros, mas deve-se visualizar todas as situações que possam ser críticas para integrar o sinal por completo (Figura 5; caso negativo em A e positivo em B e D) ou não dividir um sinal no meio (Figura 5 C). Infelizmente, a otimização dos parâmetros para um conjunto de sinais pode levar a impossibilidade de integração completa de um sinal cromatográfico

(Figura 5; seta vermelha em B) e isso deve ser considerado à luz dos resultados esperados para o planejamento do estudo.

A escolha do método mais indicado, dentre todos os apresentados anteriormente, para a desconvolução de sinais cromatográficos dependerá dos dados adquiridos e o usuário poderá testar os diferentes métodos e avaliar os resultados usando a ferramenta *Show preview*. Espera-se que as discussões apresentadas aqui referentes aos prós e contras de cada método possam ajudar o usuário a definir qual o melhor método para aplicar ao seu conjunto de dados.

Quando há dados de MS/MS adquiridos no experimento, é importante marcar as opções que pareiam os espectros de MS/MS com os *features* que serão desconvoluidos, segundo parâmetros de desvios de valores de *m/z* e tempo de retenção (Opção *scan pairing*).

Desconvolução de espectros de MS (etapa específica para GC-MS) [MZMine2| Menu: Feature list methods >> Spectral deconvolution >> Hierarchical Clustering | Multivariate Curve Resolution] (Figura 4S)

Em análises por LC-MS é muito comum a aquisição de dados de MS/MS utilizando o experimento DDA. Nesse modo, espectros de MS/MS são adquiridos após uma etapa de seleção de um íon precursor usando uma janela pequena de isolamento do íon, e consequentemente, cada espectro de íons produtos pode ser relacionado diretamente com seu íon precursor. Já para GC-MS, onde é muito comum o uso da fonte de ionização por elétrons (*Electron Ionization*), as fragmentações ocorrem na fonte, ou seja, sem uma etapa prévia de seleção dos íons. Consequentemente, cada *scan* (espectro) detectado apresenta todos os íons precursores juntamente com íons produtos. Ou seja, esses espectros de massas obtidos por GC-MS podem conter mistura de íons fragmentos de diferentes metabólitos. Nesse caso, é necessário que se adicione uma etapa de processamento para separação dos íons dos espectros de fragmentação para cada componente da mistura, a desconvolução de espectros de MS (*Spectral deconvolution*).

Dentre os métodos disponíveis, estão: (1) *Hierarchical Clustering* e (2) *Multivariate Curve Resolution*. Esse último será especificado aqui e é o método de preferência dos autores deste guia. Uma de suas vantagens é possuir menos parâmetros a serem informados pelo usuário.

Dentre os parâmetros requeridos pelo método *Multivariate Curve Resolution* estão: (1) largura para janela de desconvolução em minutos (*Deconvolution window width (min)*); (2) tolerância em minutos para tempo de retenção (*Retention time tolerance (min)*); (3) número mínimo de sinais que, será quase sempre igual a 1 (*Minimum Number of Peaks*); (4) ajuste do tempo de retenção do sinal desconvoluído (*Adjust Apex Ret Times*) (Figura 6). É importante destacar que o processamento é realizado nas listas de *features* dos EIC brutos e dos EIC desconvoluídos e, por isso, eles devem ser escolhidos nas opções *Chromatograms* e *Peaks*, como *Specific feature lists*. Nesse método de desconvolução, o cromatograma é dividido em seções de intervalos de tempo (*clusters*), o que irá permitir maior eficiência computacional para os cálculos. Um método multivariado divide cada *cluster* em diferentes grupos conforme o perfil de sinais nos espectros de MS resultando em espectros

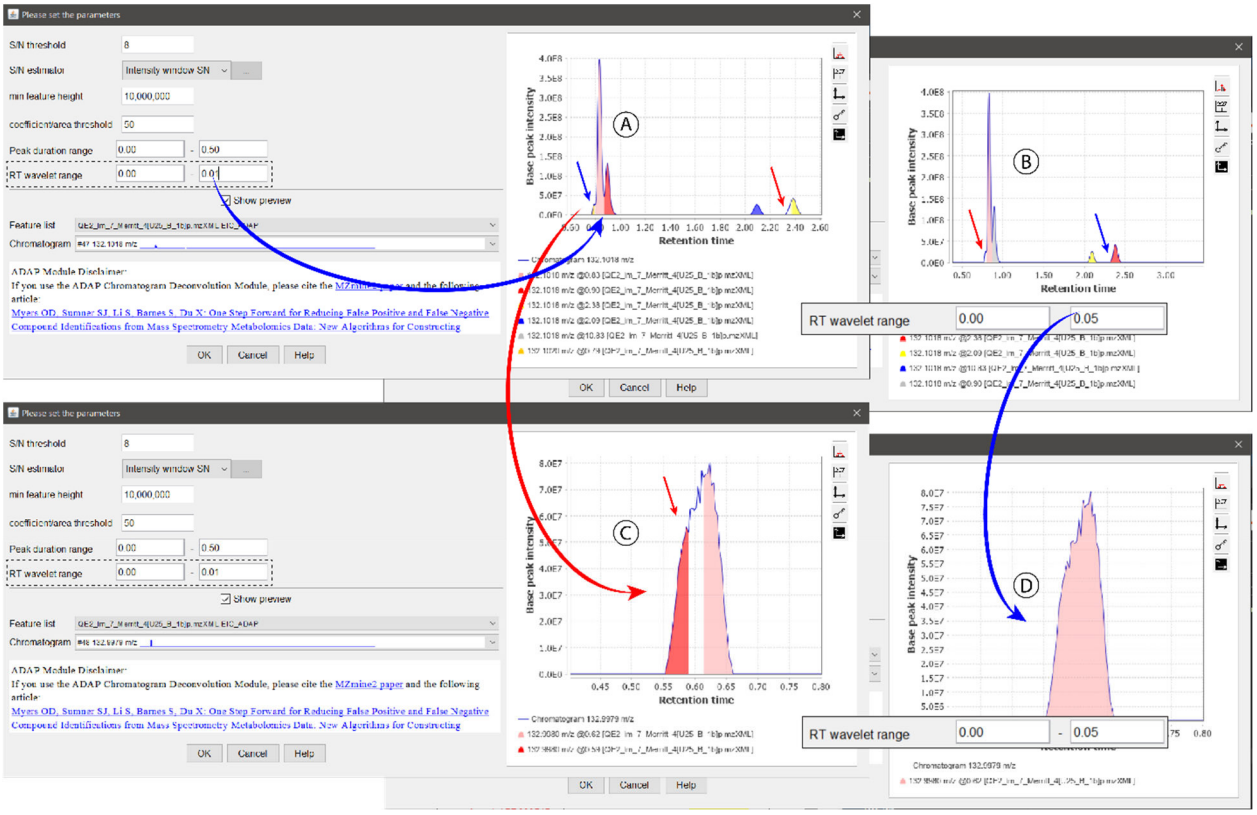

Figura 5. *Demonstração da visualização dos sinais cromatográficos para o procedimento de desconvolução cromatográfica utilizando o método Wavelets (ADAP) (Dado: ST001163, LC-MS). Observa-se aqui 4 situações diferentes: (1) para o intervalo entre 0,60-2,60 minutos (A e B) e com escolha por parâmetros 0,01 (A) e 0,05 (B) para o parâmetro RT wavelet range; (2) para o intervalo entre 0,45-0,80 minutos (C e D) e com escolha por parâmetros 0,01 (C) e 0,05 (D) para o parâmetro RT wavelet range. A: Show preview com parâmetro RT wavelet range em 0,01 evidenciando uma divisão correta dos sinais em 0,8-1 min (seta azul) e uma integração incompleta do sinal em 2,4 min (seta vermelha); B: Show preview com parâmetro RT wavelet range em 0,05 evidenciando a integração satisfatória do sinal 2,4 min (seta azul) e a não integração do 'ombro' em 0,8 min (seta vermelha); C: Show preview com parâmetro RT wavelet range em 0,01 evidenciando divisão indevida de um sinal mais intenso (seta vermelha); D: Show preview com parâmetro RT wavelet range em 0,05 evidenciando a integração satisfatória do sinal 0,6 min. É sempre importante avaliar diferentes chromatograms em diferentes Feature Lists*

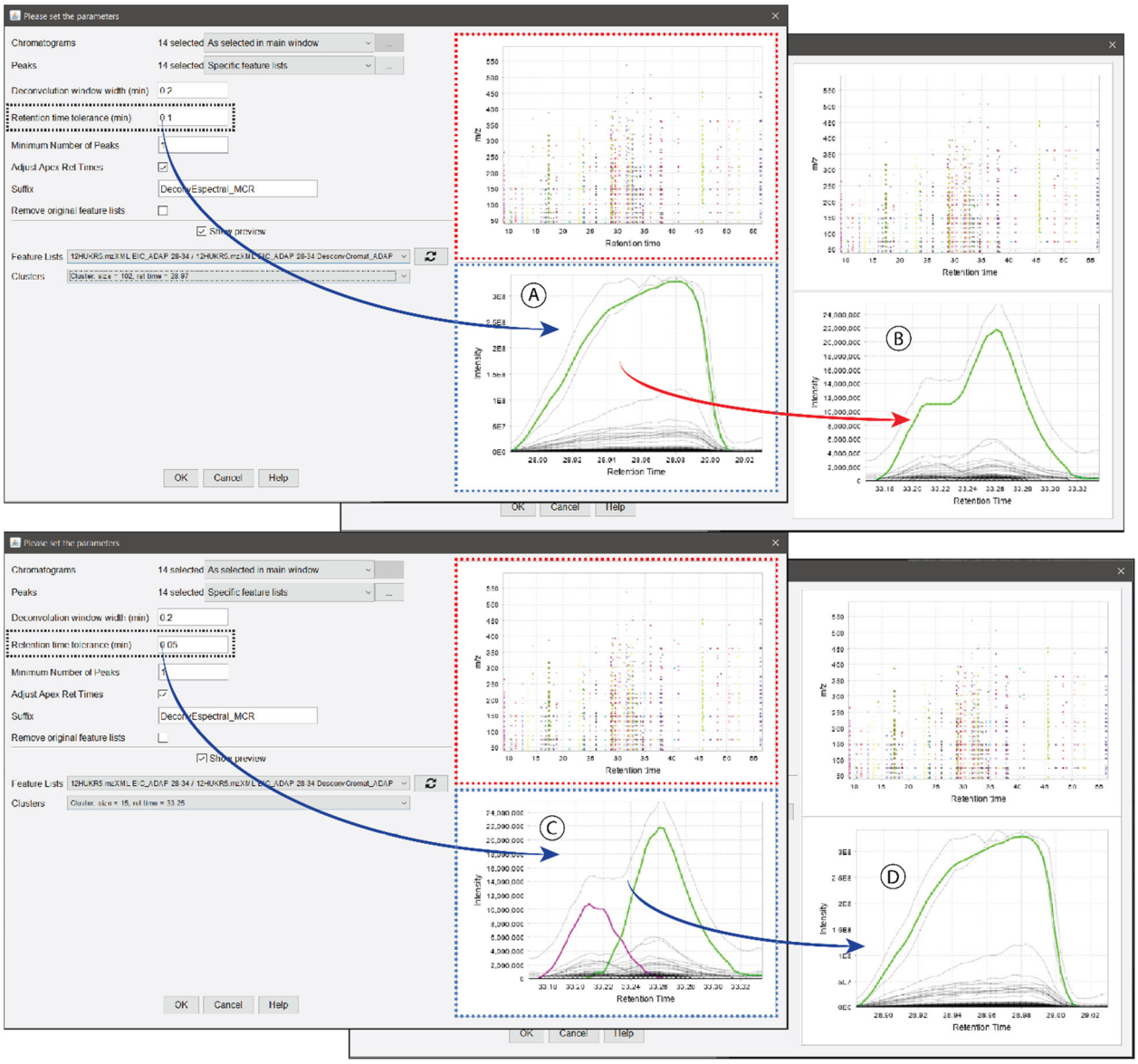

Figura 6. *Demonstração do procedimento de desconvolução espectral usando o método Multivariate Curve Resolution (Dado: ST001056, GC-MS). O tamanho dos clusters foram definidos pelo parâmetro Deconvolution Window Width (min) para englobar 2-3 sinais cromatográficos. Observem que os clusters são identificados e diferenciados por cores na janela marcada em vermelho. Na janela marcada em azul, observa-se o resultado previsto da desconvolução a partir da aplicação dos valores informados para os demais parâmetros. A: Show preview com parâmetro Retention time tolerance (min) de 0,1 evidenciando uma distinção correta do sinal em 28,86 min, mas incorreta dos sinais presentes entre 33,18 e 33,32 min (B). C: Show preview com parâmetro Retention time tolerance (min) de 0,05 evidenciando uma distinção correta dos sinais presentes entre 33,18 e 33,32 min (C) e também do sinal em 28,86 min (D). É sempre importante avaliar diferentes clusters em diferentes Feature Lists*

de fragmentação puros, quando bem sucedido. Então, um balanço entre os parâmetros *Deconvolution window width (min)* e *Retention time tolerance (min)* deve ser explorado utilizando o *Show preview* observando as janelas mais críticas (ou de interesse) no cromatograma. De forma prática, o objetivo é observar a desconvolução de dois ou mais sinais cromatográficos com cores diferentes na janela inferior do *Show preview* (Figura 6 C). Na janela superior do *Show preview* é mostrado o resultado da divisão dos *clusters* no cromatograma selecionado. Essa janela deve ser explorada com a ferramenta *zoom* e analisada em conjunto com os cromatogramas dos dados brutos. O sugerido aqui é encontrar um valor de *Deconvolution window width (min)* que envolva 2 ou 3 sinais cromatográficos para a determinação dos *clusters* e um valor de *Retention time tolerance (min)* conforme a largura mínima dos sinais cromatográficos, Valores altos desse último levarão ao mau desempenho do processo de desconvolução espectral. Um desafio claro em muitas análises será integrar por completo um sinal cromatográfico muito intenso (Figura 6; A e D) usando os mesmos parâmetros escolhidos para a separação de dois sinais parcialmente sobrepostos (Figura 6; B e C). Observem que em um primeiro momento, o usuário pode ser levado a escolher parâmetros para integrar um sinal muito intenso (Figura 6 A) com parâmetros que não iriam separar outros sinais cromatográficos (Figura 6 B). A indicação é sempre verificar pela visualização dos *features* quando modificar algum parâmetro.

Reconhecimento de perfil isotópico

[MZMine2| Menu: Feature list methods >> Isotopes >> Isotope peak grouper] (Figura 5S)

Até o momento, todos os *features* detectados e desconvoluídos foram construídos dentro de um valor limite referente ao *m/z tolerance* (em *ADAP Chromatogram builder*) através do qual se faz

um alinhamento dos valores de *m/z* entre os diferentes *scans* para constituir um *feature* listado como EIC. Sendo esse *m/z tolerance* sempre menor do que 1,0, é esperado que os sinais que compõem um perfil isotópico estejam separados em *features* diferentes. O perfil isotópico é o nome que se dá ao conjunto de sinais de MS com suas intensidades relativas que englobam o íon molecular (ou as moléculas protonadas/desprotonadas em ESI) e os íons referentes aos seus isótopos mais pesados.[2] Em geral, observa-se um perfil do tipo 'A', 'A+1', 'A+2' (p. ex.: $^{12}C_9H_{10}NO_2$, $^{13}C^{12}C_8H_{10}NO_2$, $^{13}C_2{}^{12}C_7H_{10}NO_2$) conforme a diferença de massa e contribuição de cada isótopo, principalmente em referência à abundância relativa de ^{13}C e ^{12}C. Sabendo que uma substância certamente produzirá um perfil isotópico (desde que o instrumento utilizado produza resolução suficiente para tal distinção (Figura 7), é necessário que se combine os diferentes *features* que compõem um perfil isotópico e esse procedimento é denominado agrupamento de sinais isótopicos (*Isotope peak grouper*).

Esse procedimento é realizado utilizando os parâmetros (1) tolerância máxima de desvio de valor de *m/z* (*m/z tolerance*), (2) tolerância máxima permitida de desvio de valor de tempo de retenção (*Retention time tolerance*) e (3) maior quantidade de cargas permitidas (*Maximum charge*). Trata-se de um procedimento simples e mais similar a um alinhamento de *features* dentro de cada dado e deve-se usar os valores anotados analisando os dados brutos. De forma prática, esse procedimento combina EIC *features* que tenham comportamento cromatográfico similares e forma um *pseudo*-espectro de MS com os sinais dos mesmos.

Alinhamento de features entre amostras (etapa comum para LC-MS e GC-MS)

[MZMine2| Menu: Feature list methods >> Alignment >> ADAP Aligner (GC); Join aligner (LC) ou RANSAC aligner (LC)] (Figura 6S)

[MZMine2| Menu: Feature list methods >> Alignment >> ADAP Aligner (GC)] (Figura 7S)

Finalmente, após processamentos dos sinais cromatográficos e obtenção de espectros de MS consenso para cada *feature* desconvoluído, pode-se seguir com o alinhamento das diferentes listas de *features* de cada dados (na maioria dos casos, de cada amostra).

É muito importante, inclusive como um procedimento de controle de qualidade dos dados, incluir amostras de branco de extração nesse alinhamento para que seja possível apontar sinais de contaminações comuns. Outras amostras de controle de qualidade também são importantes, mas essas podem variar conforme o planejamento do estudo proposto. Como resultado, é produzido uma lista de *features* alinhados, com seus dados de MS combinados e a intensidade de cada *feature* integrado listada para cada dado (ou amostra). Um resultado positivo desse processo todo é a integração e cálculo das áreas sob a curva de cada sinal cromatográfico devidamente desconvoluído. Esse dado poderá ser utilizado em estudos de quantificação absoluta ou quantificação relativa, por exemplo.

Duas opções podem ser utilizadas para o alinhamento dos *features* em dados de LC-MS: (1) *Join aligner*,[9] que utiliza similaridade entre os tempos de retenção e os valores de *m/z* que definem cada *feature*; e (2) *RANSAC aligner*,[30] que é uma adaptação do *Join aligner* que inclui a possibilidade de alinhamento de *features* cujos desvios fogem de um comportamento linear. Esse último é uma opção menos preferida pelos autores deste guia uma vez que não é uma ferramenta determinística e apresenta diversos parâmetros para serem definidos.

O método *Join aligner* apresenta poucos parâmetros para serem definidos, sendo eles principalmente relacionados à tolerância de *m/z* (*m/z tolerance*) e de tempo de retenção (*Retention time tolerance*), juntamente com a informação dos pesos utilizados para esses dois parâmetros (*Weight for m/z* e *Weight for RT*). Nesse caso, dados de MS de alta resolução são mais bem alinhados pelos dados de *m/z*, sendo configurados com um peso maior, enquanto dados adquiridos em equipamentos de baixa resolução pode ser alinhados com um peso maior para o tempo de retenção. Entretanto, há casos onde o *Join aligner* não funciona bem por requerer desvios lineares apenas, daí

Figura 7. *Demonstração da necessidade de se aplicar o método Isotopic peak grouper para combinar os EIC produzidos a partir de sinais de MS que compõem um único perfil isotópico. Dado: ST000025 (GC-MS)*

a indicação do *RANSAC*.[9] É indicado, por final, reanalisar os dados à luz dos resultados de anotação de substâncias (etapa posterior que não é alvo deste trabalho) verificando *features* de interesse e, principalmente, essa etapa de alinhamento.

Especificamente para dados de GC-MS, o método de escolha é o *ADAP Aligner*.[23,31] Esse método requer a construção prévia dos espectros de fragmentação desconvoluídos, pois também utiliza um resultado de cálculos de similaridade entre esses espectros de fragmentação, além da similaridade entre os valores de tempo de retenção. O usuário deve informar um limite mínimo para um *score* de similaridade e o peso a ser dado a esse parâmetro. Também é necessário informar: (1) uma fração mínima das amostras que contenham um *feature* (*Min confidence*); (2) um limite de tolerância para tempo de retenção (*Retention time tolerance*); (3 e 4) *Score threshold* e *Score weight*, que já foram mencionados; (5) o método a ser usado para medir a similaridade entre os tempos de retenção (*Retention time similarity*). Diferente do *Join aligner*, esse método usa tanto os espectros de MS desconvoluídos quanto os tempos de retenção para produzir um parâmetro de similaridade em cada amostra e, depois, alinha esses *features*.

A aplicação do *Gap filling* nessa etapa do processamento é opcional e indicada quando há, após alinhamento, *features* ausentes da lista. Brevemente, o script *Gap filling* (que significa "preenchimento de lacunas") permite recuperar dados de intensidade de *features*, que por algum motivo (p. ex.: na etapa de alinhamento), não foram detectados. Se por um lado, essa etapa pode corrigir possíveis erros, por outro, pode incluir *features* incorretos ou irrelevantes. Não há uma regra e o usuário deve tentar aplicar o que for possível à luz de seus dados. Particularmente para grandes conjuntos de dados, essa ferramenta pode auxiliar em futuras análises estatísticas multivariadas onde os algoritmos não funcionam perfeitamente quando uma quantidade muito grande de valores zero estão presentes. Ao usar o *Gap filling,* esses valores nulos de intensidade (zero) resultantes do alinhamento podem ser preenchidos com valores de intensidade baixos, perto do sinal ruído, substituindo esses valores nulos.

Filtros (etapa opcional, porém comum para LC-MS e GC-MS)
[MZMine2| Menu: Feature list methods >> Filtering]

Nessa última (e opcional) etapa do processamento, o objetivo é filtrar aqueles *features* que não serão usados na interpretação dos dados, por qualquer que seja o motivo. Para isso, alguns filtros estão disponíveis, mas a aplicação de cada um deles vai depender dos objetivos e planejamentos de cada estudo. Por esse motivo, serão listados aqui alguns casos mais comuns.

Em estudos que envolvam uma série de amostras de controle de qualidade, é comum observar a aplicação de um filtro de controle de qualidade (*QC filter*) no qual, por exemplo, apenas serão mantidos aqueles *features* que estejam presentes nas amostras combinadas do controle de qualidade. Essas podem ser (1) uma amostra combinada formada por alíquotas de todas as amostras do estudo (*QC pool* Geral) ou (2) amostras combinadas por grupos formadas por alíquotas de amostras de cada grupo estudado (*QC pool* Grupo). Esse último é o mais indicado quando quantidade de amostra disponível não é uma limitação. Um filtro de prevalência (*prevalence filter*) é usado para excluir aqueles *features* que não são consistentes em todas as réplicas analisadas em um estudo, ou qualquer quantidade relativa dessas réplicas. Por exemplo, o usuário pode manter apenas aqueles *features* que aparecem em, pelo menos, 70 % das réplicas.

Em outros casos, o usuário pode usar um filtro que seja baseado em formato de sinais cromatográficos (*peak shape filter*). Um exemplo disso ocorre no procedimento descrito para identificar *features* de adutos, onde é necessário observar, inclusive, a semelhança dos sinais cromatográficos para afirmar que determinados *features* sejam, de fato, das mesmas substâncias, mas em formas de adutos diferentes.[21]

Um filtro baseado nos resultados estatísticos (*statistically based filter*) é pouco comum de ser encontrado na literatura. Nesse, o usuário exclui aqueles *features* que não obtiveram representatividade nos cálculos de estatística exploratório não-alvo (p. ex.: análise de componentes principais (PCA) ou análise discriminante por mínimos quadrados parciais (PLS-DA)), o que resulta em uma matriz de dados mais próxima de uma situação estatisticamente mais relevante em relação a diferença entre número de amostras e número de variáveis/ *features*. Assim, o usuário repetiria a última etapa do estudo, mas considerando apenas aqueles *features* mais importantes para a projeção (*variables important in projection*-VIP, no caso do PLS-DA) dos cálculos exploratórios. De qualquer forma, é sempre importante que o usuário retorne ao dado bruto para verificação de *feature* que se mostrou de interesse, por qualquer que seja o motivo, para se certificar de que se trata realmente de um sinal cromatográfico.

Finalmente, é válido indicar que os usuários produzam, em paralelo com todas essas etapas discutidas, um arquivo de automação da sequência de processamento (*Batch mode*), que pode ser salvo no formato .xml. Com esse arquivo, outros usuários podem processar automaticamente os dados utilizando os mesmos parâmetros pré-selecionados. Incluir essa etapa como rotina durante o processamento e disponibilizar esse arquivo juntamente com os dados brutos garantem reprodutibilidade e parte do que é indicado dentro dos princípios de *FAIR science*.[32]

Validação do guia com dados de repositório

De forma a validar a aplicação deste trabalho como um guia realmente útil, 3 alunos em estágios de Iniciação Científica e Mestrado foram convidados para processar alguns dados obtidos do repositório Metabolomics Workbench (Tabela 1). Foram utilizados apenas os dados brutos com as informações contidas no *Metadata* de cada estudo para processar os dados. Nenhum juízo sobre a qualidade dos dados ou de interpretação biológica foi considerado.

A dúvida mais comum encontrada pelos alunos nos diferentes conjuntos de dados foi sobre como escolher corretamente o valor de *Noise level* logo na etapa inicial do processamento. Em alguns casos (mais evidente em ST001056), diferentes amostras apresentavam diferentes intensidades para linha de base. A sugestão é que o usuário entenda cada etapa do processamento como sendo um filtro e que talvez seja uma boa ideia aumentar o grau de restrição nos valores de cada parâmetro progressivamente a cada etapa. Dessa forma, é mais seguro escolher valores mais baixos de *Noise level* nessa primeira etapa de *Mass detection* esperando que a próxima etapa de *Chromatogram builder* exclua aqueles sinais que não formam um perfil de sinais nos EICs (voltar na descrição do método ADAP *Chromatogram builder*).

Uma grande vantagem do MZMine2 relatada pelos alunos, em relação a outros *softwares* disponíveis para processamento de dados de cromatografia-espectrometria de massas, é a facilidade de acompanhar cada etapa do processamento de dados com os modos de visualização. Assim que um lista de *features* é produzida, o usuário pode abri-la e visualizar as informações de tempo de retenção, *m/z*, área sob a curva, intensidade dos EICs e o formato dos EICs. É ainda possível produzir uma visualização da integração de cada *feature* em meio ao cromatograma TIC inteiro (Figura 8). Dessa forma, cada etapa do processamento pode ser validada conforme os objetivos traçados no estudo.

Comentários gerais e de integração do MZMine2 com outras ferramentas

Certamente, o avanço que o GNPS (*The Global Natural Product*

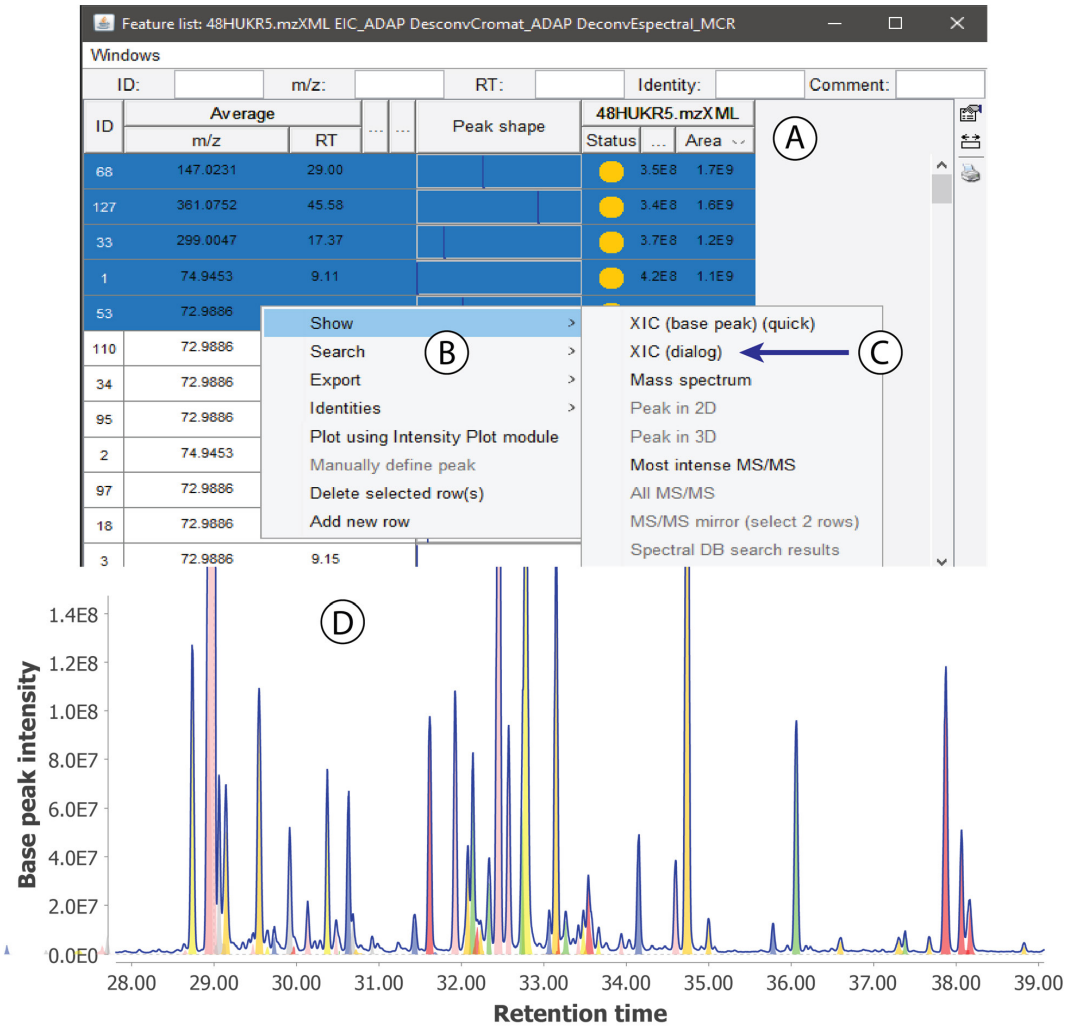

Figura 8. *Demonstração do processo de verificação das integrações obtidas com a aplicação de cada método a partir da lista de features. A: uma lista de features qualquer; B e C: caminho a partir do uso do atalho para visualização (Show) dos features sobrepostos nos EICs (usar o comando XIC (dialog)). D: um exemplo de um cromatograma de TIC obtido com a visualização das integrações obtidas. Essa etapa deve ser seguida para a visualização de todas as etapas do processamento para acompanhamento do que está sendo integrado e o que está sendo excluído da lista de features*

Social Molecular Networking)[33,34] trouxe para a área de PN pode ser apontado como um dos responsáveis pelo maior interesse no uso de ferramentas de processamento de dados de LC-MS, mas essa não é a sua única aplicação. Aliás, a possibilidade de organizar todos os dados de um estudo em uma tabela de sinais integrados com suas áreas ou intensidades traz vantagens óbvias de visualização.

Considerando o resultado do processamento obtido, tabela de integração dos *features* junto com um arquivo único com os dados de MS/MS (em formato .mgf), pode-se utilizar esses dados em análises posteriores com maior segurança, excluir *features* de amostras branco e/ou aberrantes, priorizar *features* consistentes com as amostras de controle de qualidade, identificar sinais em comparação com uma amostra externa de padrões autênticos, etc. E o fato dos *features* estarem indexados sequencialmente de forma organizada, traz outras vantagens já que essa identificação de cada *feature* será mantida em diferentes plataformas utilizadas (p. ex.: GNPS, MetaboAnalyst[35] e Sirius[36,37]). Em uma aplicação muito comum em PN, os usuários podem estar interessados em realizar análises multivariadas a um estudo de desreplicação utilizando redes moleculares. Usando os arquivos resultantes do processamento dos dados desconvoluídos e alinhados (e filtrados quando for oportuno), o usuário garante a indexação dos *features* nas diferentes plataformas, como foi mostrado por Resende *et al.*[38] em que os *features* apontados como VIP foram anotados nas redes moleculares para tentativa de identificação dos marcadores químicos. Nesse contexto, vale a pena mencionar ainda a possibilidade de anotação e reconhecimento de diferentes *features* detectados para um mesmo metabólito na forma de adutos ou fragmentos gerados *in-source* com a utilização do Ion Identity Molecular Networking.[21] O reconhecimento desses *features* pode contribuir para evitar interpretação errônea dos dados. Essa ferramenta pode ser aplicada no processamento dos dados no MZmine2 e a visualização das anotações pode ser conferida nas redes moleculares, onde todos os *features* de um mesmo metabólito serão conectados em uma mesma família molecular.

Independentemente deste trabalho, que pode ser utilizado como um tutorial comentado, é indicado que se faça uso dos tutoriais oficiais de cada *software* e, no caso do MZMine2, usar a opção *Help* disponível em cada etapa e método de processamento. O MZMine2 também mostra uma explicação mais simplificada de cada parâmetro quando o cursor é posicionado sobre ele.

CONCLUSÕES

Com este trabalho, estão resumidas as informações mais importantes para auxiliar grupos de pesquisa no Brasil, assim como outros países de língua nativa portuguesa, (desde alunos de iniciação

científica até pesquisadores mais experientes) a processar seus dados de LC-MS e GC-MS utilizando o MZMine2, uma ferramenta de acesso gratuito e código aberto. Este trabalho em língua portuguesa será útil e dará suporte a inúmeros estudos envolvendo a química de produtos naturais, a metabolômica e as ciências da vida de modo geral.

Finalmente, como uma última etapa de validação deste trabalho como um guia prático, alunos em nível de iniciação científica e mestrado pouco experientes nos tópicos discutidos processaram os dados obtidos em repositório *online* com o sucesso esperado.

MATERIAL SUPLEMENTAR

As etapas completas dos processamentos de alguns dos conjuntos de dados da Tabela 1 estão expostas como Material Suplementar para acompanhamento (Material Suplementar 2), assim como um rápido tutorial para *download* de dados do ambiente MassIVE e do MetabolomicsWorkbench (Material Suplementar 1). Os seguintes conjuntos de dados estão apresentados: ST000240 (LC-MS); ST001163 (LC-MS); ST001199 (LC-MS); ST001122 (LC-MS/MS); e ST001056 (GC-MS). Esse material com figuras usando o MZMine2 está disponível em http://quimicanova.sbq.org.br, na forma de arquivo PDF, com acesso livre.

AGRADECIMENTOS

R.M.B gostaria de agradecer a Dr. X. Du (University of North Carolina at Charlotte) pelas discussões e elucidações de alguns pontos referentes ao melhor entendimento dos métodos ADAP e a todos os alunos que assistiram as versões não-oficiais do curso de processamento de dados de MS ministrados na UFRJ que incentivaram a criação deste trabalho. Os autores agradecem ainda a todos que rotineiramente disponibilizam seus dados em repositórios abertos para reprocessamento por terceiros e àqueles que trabalham no desenvolvimento de ferramentas gratuitas e de código aberto de bio(quimio)-informática. A.B. agradece à FAPESP (processo 2018/24865-4) pela bolsa disponibilizada para desenvolvimento do estágio no exterior.

REFERÊNCIAS

1. Pilon, A.; Vieira, N.; Amaral, J.; Monteiro, A.; Silva, R.; Spíndola, L.; Castro-Gamboa, I.; Lopes, N.; *Quim. Nova* **2021**, no prelo.
2. Vessecchi, R.; Lopes, N. P.; Gozzo, F. C.; Dörr, F. A.; Murgu, M.; Lebre, D. T.; Abreu, R.; Bustillos, O. V.; Riveiros, J. M.; *Quim. Nova* **2011**, *34*, 1887.
3. Peñaloza, E.; Holandino, C.; Scherr, C.; de Araujo, P. I. P.; Borges, R. M.; Urech, K.; Baumgartner, S.; Garrett, R.; *Molecules* **2020**, *25*, 1.
4. Zhou, Y.; Qin, Q.; Zhang, P. W.; Chen, X. T.; Liu, B. J.; Cheng, D. M.; Zhang, Z. X.; *Sci. Rep.* **2020**, *10*, 2306.
5. de Albuquerque Cavalcanti, G.; Moreira Borges, R.; Reis Alves Carneiro, G.; Costa Padilha, M.; Gualberto Pereira, H. M.; *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 2417.
6. Pilon, A.; Selegato, D.; Fernandes, R.; Bueno, P.; Pinho, D.; Carnevale Neto, F.; Freire, R.; Castro-Gamboa, I.; Bolzani, V.; Lopes, N.; *Quim. Nova* **2020**, *43*, 329.
7. Belinato, J.; Bazioli, J.; Sussulini, A.; Augusto, F.; Fill, T.; *Quim. Nova* **2019**, *42*, 546.
8. Canuto, G.; Costa, J. L.; Cruz, P.; Souza, A.; Faccio, A.; Klassen, A.; Rodrigues, K.; Tavares, M.; *Quim. Nova* **2018**, *41*, 75.
9. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M.; *BMC Bioinformatics* **2010**, *11*, 395.
10. Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G.; *Anal. Chem.* **2006**, *78*, 779.
11. Tsugawa, H.; Ikeda, K.; Takahashi, M.; Satoh, A.; Mori, Y.; Uchino, H.; Okahashi, N.; Yamada, Y.; Tada, I.; Bonini, P.; Higashi, Y.; Okazaki, Y.; Zhou, Z.; Zhu, Z.-J.; Koelmel, J.; Cajka, T.; Fiehn, O.; Saito, K.; Arita, M.; Arita, M.; *Nat. Biotechnol.* **2020**, *38*, 1159.
12. Kale, N. S.; Haug, K.; Conesa, P.; Jayseelan, K.; Moreno, P.; Rocca-Serra, P.; Nainala, V. C.; Spicer, R. A.; Williams, M.; Li, X.; Salek, R. M.; Griffin, J. L.; Steinbeck, C.; *Curr. Protoc. Bioinformatics* **2016**, *53*, 14.
13. McAlpine, J. B.; Chen, S. N.; Kutateladze, A.; MacMillan, J. B.; Appendino, G.; Barison, A.; Beniddir, M. A.; Biavatti, M. W.; Bluml, S.; Boufridi, A.; Butler, M. S.; Capon, R. J.; Choi, Y. H.; Coppage, D.; Crews, P.; Crimmins, M. T.; Csete, M.; Dewapriya, P.; Egan, J. M.; Garson, M. J.; Genta-Jouve, G.; Gerwick, W. H.; Gross, H.; Harper, M. K.; Hermanto, P.; Hook, J. M.; Hunter, L.; Jeannerat, D.; Ji, N. Y.; Johnson, T. A.; Kingston, D. G. I.; Koshino, H.; Lee, H. W.; Lewin, G.; Li, J.; Linington, R. G.; Liu, M.; McPhail, K. L.; Molinski, T. F.; Moore, B. S.; Nam, J. W.; Neupane, R. P.; Niemitz, M.; Nuzillard, J. M.; Oberlies, N. H.; Ocampos, F. M. M.; Pan, G.; Quinn, R. J.; Reddy, D. S.; Renault, J. H.; Rivera-Chavez, J.; Robien, W.; Saunders, C. M.; Schmidt, T. J.; Seger, C.; Shen, B.; Steinbeck, C.; Stuppner, H.; Sturm, S.; Taglialatela-Scafati, O.; Tantillo, D. J.; Verpoorte, R.; Wang, B. G.; Williams, C. M.; Williams, P. G.; Wist, J.; Yue, J. M.; Zhang, C.; Xu, Z.; Simmler, C.; Lankin, D. C.; Bisson, J.; Pauli, G. F.; *Nat. Prod. Rep.* **2019**, *36*, 35.
14. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P.; *Nat. Biotechnol.* **2012**, *30*, 918.
15. Lohr, K. E.; Khattri, R. B.; Guingab-Cagmat, J.; Camp, E. F.; Merritt, M. E.; Garrett, T. J.; Patterson, J. T.; *Sci. Rep.* **2019**, *9*, 6067.
16. Misra, B. B.; Das, V.; Landi, M.; Abenavoli, M. R.; Araniti, F.; *Plant. Sci.* **2020**, *298*, 110548.
17. Chou, H.; Pathmasiri, W.; Deese-Spruill, J.; Sumner, S.; Buchwalter, D. B.; *J. Insect. Physiol.* **2017**, *101*, 107.
18. Jiang, W.; Qiu, Y.; Ni, Y.; Su, M.; Jia, W.; Du, X.; *J. Proteome Res.* **2010**, *9*, 594.
19. Stettin, D.; Poulin, R. X.; Pohnert, G.; *Metabolomics* **2020**, *10*, 143.
20. Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X.; *Anal. Chem.* **2017**, *89*, 8689.
21. Schmid, R.; Petras, D.; Nothias, L. F.; Wang, M.; Aron, A. T.; Jagels, A.; Tsugawa, H.; Rainer, J.; Garcia-Aloy, M.; Duhrkop, K.; Korf, A.; Pluskal, T.; Kamenik, Z.; Jarmusch, A. K.; Caraballo-Rodriguez, 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.; Hovelmann, Y.; Kalinina, S. A.; Pendergraft, M. A.; Panitchpakdi, M.; Tehan, R.; Le Gouellec, A.; Aleti, G.; Mannochio Russo, H.; Arndt, B.; Hubner, F.; Hayen, H.; Zhi, H.; Raffatellu, M.; Prather, K. A.; Aluwihare, L. I.; Bocker, S.; McPhail, K. L.; Humpf, H. U.; Karst, U.; Dorrestein, P. C.; *Nat. Commun.* **2021**, *12*, 3832.
22. Tautenhahn, R.; Bottcher, C.; Neumann, S.; *BMC Bioinformatics* **2008**, *9*, 504.
23. Ni, Y.; Qiu, Y.; Jiang, W.; Suttlemyre, K.; Su, M.; Zhang, W.; Jia, W.; Du, X.; *Anal. Chem.* **2012**, *84*, 6619.
24. Ni, Y.; Su, M.; Qiu, Y.; Jia, W.; Du, X.; *Anal. Chem.* **2016**, *88*, 8802.
25. Smirnov, A.; Jia, W.; Walker, D. I.; Jones, D. P.; Du, X.; *J. Proteome Res.* **2018**, *17*, 470.
26. Smirnov, A.; Qiu, Y.; Jia, W.; Walker, D. I.; Jones, D. P.; Du, X.; *Anal. Chem.* **2019**, *91*, 9069.

27. Wee, A.; Grayden, D. B.; Zhu, Y.; Petkovic-Duran, K.; Smith, D.; *Electrophoresis* **2008**, *29*, 4215.
28. Du, X. Zeisel, S. H.; *Comput. Struct. Biotechnol. J.* **2013**, *4*, e201301013.
29. Du, P.; Kibbe, W. A.; Lin, S. M.; *Bioinformatics* **2006**, *22*, 2059.
30. Fischler, M. Bolles, R.; *Commun. ACM* **1981**, *24*, 381.
31. Jiang, W.; Qiu, Y.; Ni, Y.; Su, N.; Jia, W.; Du, X.; *J. Proteome Res.* **2010**, *9*, 5974.
32. Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J. W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; t Hoen, P. A.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S. A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B.; *Sci. Data* **2016**, *3*, 160018.
33. Nothias, L. F.; Petras, D.; Schmid, R.; Duhrkop, K.; Rainer, J.; Sarvepalli, A.; Protsyuk, I.; Ernst, M.; Tsugawa, H.; Fleischauer, M.; Aicheler, F.; Aksenov, A. A.; Alka, O.; Allard, P. M.; Barsch, A.; Cachet, X.; Caraballo-Rodriguez, A. M.; Da Silva, R. R.; Dang, T.; Garg, N.; Gauglitz, J. M.; Gurevich, A.; Isaac, G.; Jarmusch, A. K.; Kamenik, Z.; Kang, K. B.; Kessler, N.; Koester, I.; Korf, A.; Le Gouellec, A.; Ludwig, M.; Martin, H. C.; McCall, L. I.; McSayles, J.; Meyer, S. W.; Mohimani, H.; Morsy, M.; Moyne, O.; Neumann, S.; Neuweger, H.; Nguyen, N. H.; Nothias-Esposito, M.; Paolini, J.; Phelan, V. V.; Pluskal, T.; Quinn, R. A.; Rogers, S.; Shrestha, B.; Tripathi, A.; van der Hooft, J. J. J.; Vargas, F.; Weldon, K. C.; Witting, M.; Yang, H.; Zhang, Z.; Zubeil, F.; Kohlbacher, O.; Bocker, S.; Alexandrov, T.; Bandeira, N.; Wang, M.; Dorrestein, P. C.; *Nat. Methods* **2020**, *17*, 905.
34. 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.; Crusemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderon, 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. P. C. A. B.; 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.; Rodriguez, 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.; Muller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. O.; Pogliano, K.; Linington, R. G.; Gutierrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N.; *Nat Biotechnol* **2016**, *34*, 828.
35. Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D. S.; Xia, J.; *Nucleic Acids Res.* **2018**, *46*, W486.
36. Bocker, S.; Rasche, F.; *Bioinformatics* **2008**, *24*, i49.
37. Ludwig, M.; Duhrkop, K.; Bocker, S.; *Bioinformatics* **2018**, *34*, i333.
38. Mendes Resende, J. V.; de Sá, N. M. D.; de Oliveira, M. T. L.; Lopes, R. C.; Garrett, R.; Moreira Borges, R.; *Phytochem. Lett.* **2020**, *36*, 99.