DEREPLIKACE LÁTEK A *DE NOVO* CHARAKTERIZACE MALÝCH MOLEKUL Z HMOTNOSTNÍCH SPEKTER

JIŘÍ NOVÁK a VLADIMÍR HAVLÍČEK

Mikrobiologický ústav AV ČR, Vídeňská 1083, Praha 4
vlhavlic@biomed.cas.cz

Došlo 29.6.21, přijato 24.8.21.

Klíčová slova: CycloBranch, dereplikace, *de novo* charakterizace, hmotnostní spektrometrie, metabolomika, hmotnostně spektrometrické zobrazování, kapalinová chromatografie, izotopová struktura

https://doi.org/10.54779/chl20220011

Obsah

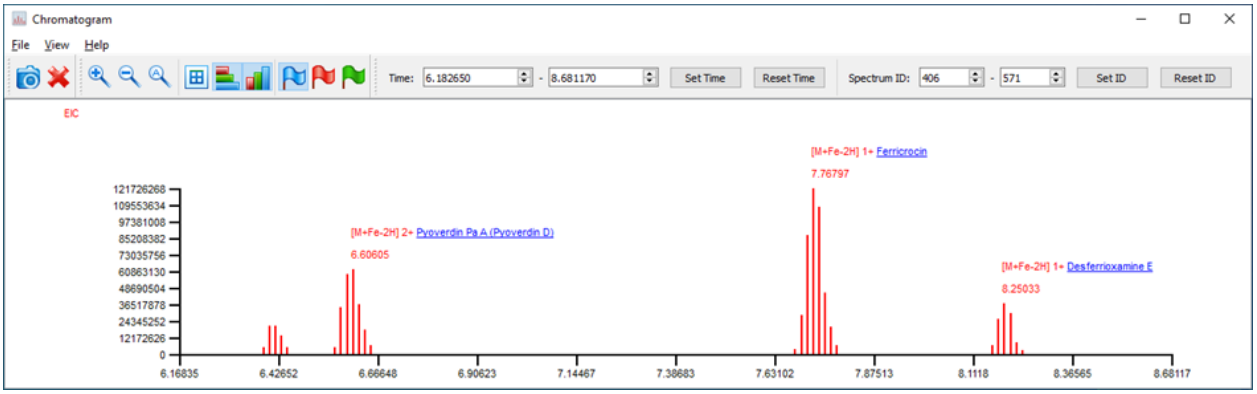
1. Úvod

V tomto příspěvku volně navazujeme na *Základy interpretace hmotnostních spekter*[1] a *Řešené příklady interpretace produktových spekter peptidů*[2] uvedené v dřívějším dvojčíslí *Chemických listů* věnovanému hmotnostní spektrometrii. Ukážeme, jak lze využít open-source aplikaci CycloBranch[3] (https://ms.biomed.cas.cz/cyclobranch/) pro dereplikaci, tedy proces přiřazení již známých chemických látek a *de novo* charakterizaci malých molekul v datových souborech obsahujících hmotnostní spektra. Zaměříme se na spektra získaná kombinací kapalinové chromatografie ve spojení s hmotnostní spektrometrií (LC-MS) a spektra z datových souborů zobrazovací hmotnostní spektrometrie (MSI)[4,5].

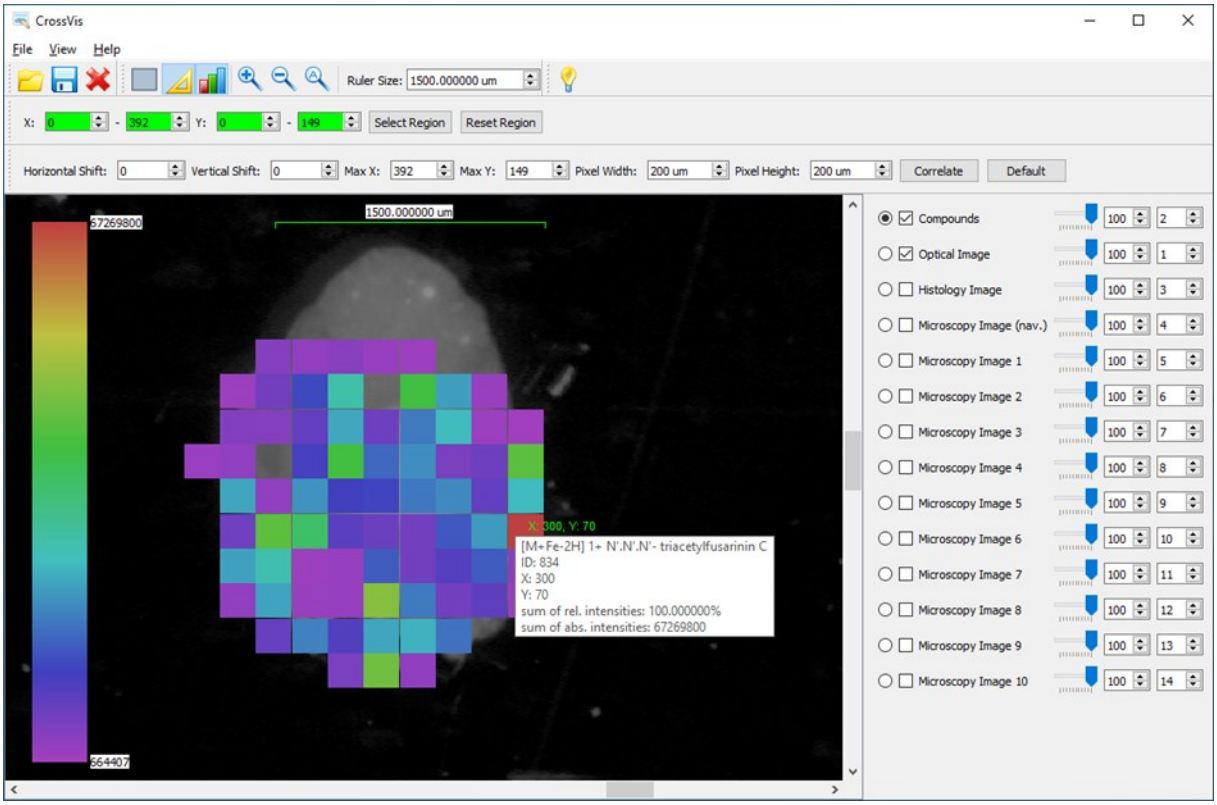
Zatímco LC-MS data můžeme z informatického pohledu chápat jako posloupnost hmotnostních spekter lišících se v retenčním čase, spektra získaná pomocí zobrazovací spektrometrie pak jako matici, kde pro každé spektrum známe souřadnice [x, y]. CycloBranch této analogie využívá a oba typy spekter zpracovává obdobným způsobem v dávkovém módu. Názvy či sumární vzorce látek přiřazených v LC-MS datech můžeme zobrazit v chromatogramu (obr. 1). U látek nalezených v MSI datech provádíme fúzi s optickým obrazem získaným jinou zobrazovací metodou (obr. 2). Naše aplikace se zvláště hodí pro spektra změřená s vysokým rozlišením[6] a látky s menší molekulovou hmotností (≤ 2000 Da), kterými mohou být i mikrobiální sekundární metabolity[7], cyklické peptidy[8], metalofory[9–11], apod. Naším příspěvkem se snažíme adresovat nedostatek studijního materiálu, který je v české odborné literatuře věnován této problematice.

2. Dereplikace

Dereplikace je proces založený na vyhledávání v databázi již známých látek. Pro každou látku potřebujeme znát její sumární vzorec, kterému odpovídá teoretická izotopová obálka[12,13]. Následně provedeme porovnání všech teoretických obálek se všemi experimentálně získanými izotopovými profily. K sestavení teoretických izoto-



Obr. 1. **Dereplikace látek z chromatogramu standardních sideroforů.** Zleva jsou zobrazeny železité formy pyoverdinu E, pyoverdinu D, ferricrocinu a desferrioxaminu E. Ve výstupu *de novo* analýzy se místo názvů látek zobrazují sumární vzorce. (Barevná verze obrázku je dostupná na webových stránkách časopisu Chemické listy).



Obr. 2. **Dereplikace triacetylfusarininu C z datového imzML souboru zobrazovací hmotnostní spektrometrie.** Čtverec odpovídá jednomu spektru v imzML souboru definovanému souřadnicemi [x, y] a skutečnému rozměru 200 µm × 200 µm. Barva čtverce odpovídá intenzitě píku nalezené látky v daném spektru. Spodní vrstva zobrazuje fotografii sklíčka, která byla pořízena pro správnou orientaci a korelaci s MS daty (12T Solarix FTICR, Bruker Daltonics, Billerica, MA, USA). Software umožňuje provést fúzi s libovolným histologickým vyšetřením, pokud máme k dispozici obrázek v TIF, JPG, BMP, PNG nebo GIF formátu. Při *de novo* analýze se místo názvů látek zobrazují sumární vzorce. (Barevná verze obrázku je dostupná na webových stránkách časopisu Chemické listy).

pových obálek potřebujeme znát tabulkové hodnoty, které definují relativní atomové hmotnosti nuklidů jednotlivých prvků a jejich procentuální zastoupení v přírodě[14,15]. S jejich využitím vygenerujeme sumární vzorec pro každou kombinaci nuklidů jednotlivých prvků, ze kterých je daná látka složena.

V následujícím příkladu použijeme látku desferrioxamin E (FOXE) s molekulovým vzorcem $^{12}C_{27}{}^{1}H_{48}{}^{14}N_{6}{}^{16}O_{9}$, pro který můžeme vytvořit kombinace s proměnlivým počtem stabilních nuklidů, např. ^{13}C, ^{17}O, a tedy varianty $^{12}C_{25}{}^{13}C_{2}{}^{1}H_{48}{}^{14}N_{6}{}^{16}O_{9}$, $^{12}C_{26}{}^{13}C{}^{1}H_{48}{}^{14}N_{6}{}^{16}O_{8}{}^{17}O$, apod. Pro každý takto získaný sumární vzorec vypočteme hodnotu *m/z* (poměr hmotnosti a náboje) a relativní intenzitu každého odpovídajícího teoretického píku v izotopovém klastru.

Předpokládejme, že máme obecnou látku se sumárním vzorcem $C_aH_bN_cO_dS_e$, kde a, b, c, d, $e \geq 0$, a kterou pro účely výpočtu rozdělíme na jednotlivé chemické prvky C_a, H_b, N_c, O_d a S_e. Počet kombinací izotopů roste s počtem atomů daného prvku podle definice multinomického rozvoje[1,13]. Uvažujeme-li např. dva stabilní izotopy uhlíku ^{12}C, ^{13}C a hodnotu $a = 3$, dostaneme s využitím binomické věty čtyři kombinace izotopů $(^{12}C + {}^{13}C)^3 = {}^{12}C{}^{12}C{}^{12}C + 3 \times {}^{12}C{}^{12}C{}^{13}C + 3 \times {}^{12}C{}^{13}C{}^{13}C + {}^{13}C{}^{13}C{}^{13}C$. Počet kombinací v tomto případě vypočteme jako $a + 1$; tedy $3 + 1 = 4$.

Stejný postup můžeme aplikovat i na nuklidy vodíku (^{1}H, ^{2}H) a dusíku (^{14}N, ^{15}N). V případě nuklidů kyslíku (^{16}O, ^{17}O, ^{18}O) je již potřeba použít trinomický rozvoj. Pro dva atomy kyslíku, tedy $d = 2$, rozvádíme $(^{16}O + {}^{17}O + {}^{18}O)^2 = {}^{16}O{}^{16}O + {}^{17}O{}^{17}O + {}^{18}O{}^{18}O + 2 \times {}^{16}O{}^{17}O + 2 \times {}^{16}O{}^{18}O + 2 \times {}^{17}O{}^{18}O$. Počet kombinací odpovídá vzorci $(d+1) \times (d+2) / 2$. Pro čtyři stabilní izotopy a e atomů síry odpovídá počet kombinací $(e+1) \times (e+2) \times (e+3) / 6$, apod.

Kalkulace kombinací izotopů pro jednotlivé a nejčastěji se vyskytující biogenní prvky nevyžaduje mnoho výpočetního času ani paměti. Problém nastává ve chvíli, kdy se snažíme určit kombinace izotopů pro celou molekulu, například FOXE ($C_{27}H_{48}N_6O_9$). Výše uvedeným postupem dostaneme $(a+1) \times (b+1) \times (c+1) \times ((d+1) \times (d+2) / 2) = (27+1) \times (48+1) \times (6+1) \times ((9+1) \times (9+2) / 2) = 28 \times 49 \times 7 \times 55 = 528\,220$ kombinací stabilních izotopů. Tato metoda není z výpočetního hlediska optimální, a to ani pro relativně malé molekuly, kdy snadno překročíme desítky

i stovky miliónů kombinací. Avšak výhodou je, že násobením polynomů můžeme kromě kombinací jednotlivých izotopů snadno určit i jejich relativní intenzity v rámci izotopové obálky. V dalším textu budeme nejintenzivnější píky nuklidů ^{1}H, ^{12}C, ^{14}N a ^{16}O pro jednoduchost uvádět bez nukleonového čísla.

Pro praktickou aplikaci můžeme využít fakt, že mnoho kombinací izotopů tvoří píky s velmi malou intenzitou, a můžeme tedy omezit „šířku“ teoretické izotopové obálky. K optimalizaci výpočtu zavedeme konstantu n, která definuje maximální počet atomů daného prvku, které budeme nahrazovat izotopy. Např. pro $n = 5$ tak můžeme vypočítat počet atomů uhlíku n_C, které budeme v dané molekule nahrazovat izotopy, jako $n_C = \min(a, n)$. Uvažujeme-li $n_C = 5$, potom generujeme pouze kombinace C_5, $C_4{}^{13}C$, $C_3{}^{13}C_2$, $C_2{}^{13}C_3$, $C^{13}C_4$ a $^{13}C_5$. Pokud platí, že $a > n_C$, doplníme zbývající počet atomů uhlíku (tj. $a - n_C$) nejčastěji se vyskytujícím nuklidem ^{12}C. Pro FOXE tímto zjednodušeným postupem získáme kombinace C_{27}, $C_{26}{}^{13}C$, $C_{25}{}^{13}C_2$, $C_{24}{}^{13}C_3$, $C_{23}{}^{13}C_4$ a $C_{22}{}^{13}C_5$. Analogicky můžeme definovat $n_H = \min(b, n)$, $n_N = \min(c, n)$, $n_O = \min(d, n)$, $n_S = \min(e, n)$, aj. Pokud předpokládáme, že $n_H = n_N = n_O = 5$, pro FOXE dostaneme $(5+1) \times (5+1) \times (5+1) \times ((5+1) \times (5+2) / 2) = 6 \times 6 \times 6 \times 21 = 4536$ sumárních vzorců odpovídajících kombinacím izotopů, tedy počet akceptovatelný pro výkon standardních osobních počítačů.

Pro predikci teoretických intenzit jednotlivých izotopických píků odpovídajících vygenerovaným sumárním vzorcům použijeme opět metodu založenou na multinomickém rozvoji[12,13]. Při výpočtu použijeme následující pravděpodobnosti výskytu stabilních izotopů uhlíku, vodíku, dusíku a kyslíku: $p(C) = 0{,}9893$, $p(^{13}C) = 0{,}0107$, $p(H) = 0{,}999885$, $p(^{2}H) = 0{,}000115$, $p(N) = 0{,}99632$, $p(^{15}N) = 0{,}00368$, $p(O) = 0{,}99757$, $p(^{17}O) = 0{,}00038$ a $p(^{18}O) = 0{,}00205$. Ve FOXE pravděpodobnost výskytu iontu $[C_{25}{}^{13}C_2H_{48}N_6O_9+H]^+$ vypočteme jako součin pravděpodobností výskytů jednotlivých prvků, tedy $p(C_{25}{}^{13}C_2H_{49}N_6O_9) = p(C_{25}{}^{13}C_2) \times p(H_{49}) \times p(N_6) \times p(O_9) = 0{,}02922$, kde $p(C_{25}{}^{13}C_2) = 27! / (25! \times 2!) \times p(C)^{25} \times p(^{13}C)^2 = 0{,}03071$; $p(H_{49}) = p(H_{49}{}^{2}H_0) = 49! / (49! \times 0!) \times p(H)^{49} \times p(^{2}H)^0 = 0{,}99438$; $p(N_6) = p(N_6{}^{15}N_0) = 6! / (6! \times 0!) \times p(N)^6 \times p(^{15}N)^0 = 0{,}97812$; a konečně $p(O_9) = p(O_9{}^{17}O_0{}^{18}O_0) = 9! / (9! \times 0! \times 0!) \times p(O)^9 \times p(^{17}O)^0 \times p(^{18}O)^0 = 0{,}97834$.

Podobně vypočteme pravděpodobnost výskytu monoizotopického iontu $[C_{27}H_{48}N_6O_9+H]^+$ jako $p(C_{27}H_{49}N_6O_9) = p(C_{27}) \times p(H_{49}) \times p(N_6) \times p(O_9) = 0{,}74792 \times 0{,}99438 \times 0{,}97812 \times 0{,}97834 = 0{,}71169$. Uvedeným postupem určíme i pravděpodobnosti zbývajících izotopických kombinací. Po vygenerování teoretické obálky musíme ještě provést normalizaci intenzit. Pokud je relativní intenzita teoretického iontu $[C_{27}H_{48}N_6O_9+H]^+$ rovna 100 %, pak intenzita $[C_{25}{}^{13}C_2H_{48}N_6O_9+H]^+$ je 4,1 %. Protože při výpočtu používáme faktoriály, které mohou způsobit přetečení nebo zaokrouhlovací chyby i při použití 64bitového počítače, je vhodné naprogramovat algebraickou optimalizaci uvedených výrazů před jejich výpočtem. Například výraz $27! / (25! \times 2!)$ můžeme vypočítat jako $27 \times 26 / 2$. Nakonec odstraníme teoretické píky s velmi malou intenzitou ($\leq 0{,}1$ %).

Pro všechny teoretické píky vypočteme odpovídající hodnoty *m/z* jako součet monoizotopických hmotností příslušných nuklidů. Následně procházíme seznam píků setříděný v klesajícím pořadí podle intenzity a provádíme jejich shlukování podle předem definované hodnoty rozlišení FWHM (Full Width at Half Maximum). Tedy podle šířky píku v polovině jeho výšky uvažujeme-li, že spektrum je v profilovém nikoliv čárovém módu. Hodnotu FWHM uvádíme v Daltonech [Da]. Je-li rozdíl *m/z* hodnot dvou píků menší nebo roven hodnotě FWHM, oba píky sloučíme do jednoho. Novou hodnotu *m/z* vypočteme jako vážený průměr původních hodnot, kde váhy jsou definovány intenzitami původních teoretických píků. Intenzitu nového píku pak vypočteme jako součet intenzit původních píků.

Při porovnávání teoretického a experimentálního spektra ke každému teoretickému píku přiřadíme experimentální pík s minimální odchylkou hodnoty *m/z*. Pokud přiřadíme experimentální pík k nejvyššímu teoretickému píku v dané obálce, nastavíme teoretickou hodnotu intenzity nejvyššího píku na intenzitu změřeného píku. Intenzity ostatních píků v teoretické obálce pak proporcionálně redukujeme a odstraníme píky, jejichž intenzita je nižší než minimální hodnota zadaná na vstupu.

Abychom eliminovali co nejvíce falešně pozitivních výsledků, provedeme ještě filtrování s využitím vlastností izotopových obálek. Na vstupu můžeme definovat minimální počet píků, které musí být v každé izotopové obálce anotovány. Pokud hledáme $[M+H]^+$ ionty, použijeme např. hodnotu 2 (hledáme monoizotopický ^{12}C pík a jeho ^{13}C analog). Pro železité formy sideroforů $[M+Fe-2H]^+$ lze nastavit např. hodnotu 3 (sledujeme monoizotopický pík M a partnerské píky v obálce odpovídající ^{13}C a ^{54}Fe). Filtrovací kritérium můžeme posílit požadavkem, aby daná látka s daným minimálním počtem anotovaných izotopických píků byla nalezena ve více spektrech. Pro MSI data můžeme navolit, že se má vyskytovat minimálně v 50 spektrech (větší oblast zájmu zahrnující padesát pixelů, region of interest, ROI), pro LC-MS data pak, že má být ve 2 nebo 3 spektrech navazujících přesně za sebou, apod.

Experimentální a teoretickou izotopovou obálku můžeme dále porovnat např. pomocí úhlové vzdálenosti (kosinové podobnosti) $\delta(A, B) = \cos^{-1}((\Sigma_i a_i b_i) / (\sqrt{(\Sigma_i a_i^2)}\sqrt{(\Sigma_i b_i^2)}))$, kde A a B jsou vektory intenzit a kde $a_i \in A$ a $b_i \in B$ jsou teoretické a experimentální intenzity píku i. Vzdálenost δ můžeme dále normalizovat do intervalu $<0, 1>$ dělením hodnotou $\pi/2$, přičemž hodnota 0 znamená nejlepší skóre. S takto vypočteným skóre můžeme provést statistické hodnocení výsledků.

Při analýze produktových spekter[16] například v proteomických experimentech se k tomuto účelu běžně používá metoda založená na poměru počtu falešně pozitivních výsledků vůči počtu všech nalezených výsledků (tzv. false-discovery rate, FDR)[17]. Před vyhledáváním v databázi proteinových sekvencí do této databáze přidáme neexistující sekvence sloužící jako „návnada“ (tzv. decoy database). Často se používá metoda, při které se sekvence aminokyselin ve vstupní databázi proteinových sekvencí

zapíší v obráceném pořadí a z nich se následně generují krátké peptidové sekvence stejným způsobem jako z originálních proteinových sekvencí. Tím je dosaženo efektu, že struktura dat je podobná a zároveň, že počet validních sekvencí přibližně odpovídá počtu falešných peptidových sekvencí. Pokud se k experimentálnímu spektru daného peptidu přiřadí teoretické spektrum návnady, víme, že se jedná o falešně pozitivní výsledek. Při stanovení FDR pro danou prahovou hodnotu skóre t zahrneme do výpočtu pouze položky, které mají vzdálenost $\delta \leq t$. Hodnotu *FDR* pak vypočteme podle vzorce $FDR = FP / (FP + TP)$, kde *FP* (false positives) je počet falešně pozitivních výsledků a *TP* (true positives) je počet správně přiřazených výsledků.

Podobný postup lze aplikovat i na metabolomická data analyzovaná bez využití MS/MS. V tomto případě generujeme návnady ve formě neexistujících (zcela nepravděpodobných) teoretických izotopových obálek. Pro data získaná zobrazovací hmotnostní spektrometrií můžeme v literatuře najít postup, který k sumárním vzorcům jednotlivých látek přidává adukty, které se nemohou běžně vyskytovat (většina prvků z periodické tabulky kromě H, Na, K a některých prvků s velkým protonovým číslem)[18]. V jiné práci autoři přidávají k sumárním vzorcům v databázi atom vodíku, čímž zajistí, že součet valenčních stavů všech atomů ve vzorci nebude odpovídat reálné látce, a že návnady nebudou ovlivněny výraznou změnou hmotnosti, která vznikne po přidání prvku s velkou atomovou hmotností[19]. Z naší zkušenosti jsou však hodnoty *FDR* vypočtené těmito postupy velmi citlivé na vstupní data na rozdíl od metody, která je běžně používaná v proteomice. Proto je vždy nutné výsledky manuálně zkontrolovat a konfrontovat s příslušnou literaturou, tedy ověřit, zdali zkoumaný organismus může produkovat nalezenou látku nebo ne.

3. *De novo* charakterizace

De novo charakterizaci sumárních vzorců lze v LC-MS a MSI datech provádět analogicky jako při dereplikaci. Klíčovým rozdílem je fakt, že nemáme k dispozici databázi látek, ve které bychom mohli vyhledávat, a proto musíme vytvořit virtuální databázi sumárních vzorců. Budeme tedy generovat kombinace s opakováním ze vstupního seznamu chemických prvků (např. H, C, O, N). Takovým přístupem však získáme mnoho falešně pozitivních sumárních vzorců, kterým nebudou odpovídat žádné existující chemické látky. Touto problematikou se ve své práci zabývali již autoři Kind a Fiehn[20], kteří definovali 7 heuristických pravidel pro filtrování falešně pozitivních sumárních vzorců založených na:

(1) omezení počtu atomů ve vzorci,
(2) pravidlech podle Lewise a Seniora využívajících znalosti valenčních stavů jednotlivých prvků,
(3) izotopových obálkách,
(4) poměru počtu atomů vodíku a uhlíku,
(5) poměrech počtu atomů dalších prvků (dusík, kyslík, fosfor a síra) vůči počtu atomů uhlíku,
(6) statistickém pozorování četnosti výskytu jednotlivých prvků v existujících databázích a
(7) výskytu trimetylsilylovaných látek[1,20].

V našem případě (FOXE) nejprve omezíme celkový počet atomů ve vzorci a volitelně i maximální počty atomů jednotlivých prvků ve vstupním seznamu. Ze všech vygenerovaných kombinací prvků pak ponecháme jen ty, které splňují tři základní pravidla podle Seniora[21]. První pravidlo říká, že součet valenčních stavů všech atomů ve vzorci musí být sudý. Druhé pravidlo, že součet valenčních stavů musí být větší nebo roven dvojnásobku maximálního valenčního stavu a konečně třetí pravidlo, že součet valenčních stavů musí být větší nebo roven $2 \times (\alpha - 1)$, kde α je počet atomů ve vzorci. Při výpočtu se však z výkonového hlediska nevyplatí testovat druhé pravidlo, protože eliminuje jen velmi málo vzorců s velmi malou molekulovou hmotností (např. CH_2).

V dalším kroku otestujeme, jestli pro zbývající sumární vzorce převedené na typy iontů hledané uživatelem (např. $[M+H]^+$ a $[M+Na]^+$) platí, že jejich teoretické hodnoty *m/z* jsou v intervalu mezi minimální a maximální experimentální hodnotou *m/z*, které jsou rovněž zadány uživatelem. Pokud ano, aplikujeme dále výše uvedená pravidla (4), (5) a (6). Detaily lze nalézt v tabulkách I, II a III v literatuře[20]. V našem případě pravidla (4) a (5) rozšiřujeme ještě o volitelné pravidlo, které říká, že poměr počtu atomů dusíku a kyslíku má být menší nebo roven 1, což je výhodné zejména při *de novo* analýze peptidů. Pro představu ještě uveďme, že šesté heuristické pravidlo podle Kinda & Fiehna například říká, že pokud vzorec obsahuje skupinu prvků N, O, P, a S více než jedenkrát, pak pro počet výskytů jednotlivých prvků platí, že $N < 10$, $O < 20$, $P < 4$ a $S < 3$.

Vzhledem k tomu, že generování velkého množství teoretických izotopových obálek je výpočetně náročné, provedeme ještě předtím redukci zbývajících sumárních vzorců pomocí analyzovaného souboru dat. Vstupní soubor obsahující LC-MS nebo MSI data prohledáme a zjistíme, zda obsahuje monoizotopické píky odpovídající zkoumaným vzorcům. Pokud neobsahuje, můžeme daný vzorec z výsledků odstranit, protože víme, že i kdybychom vygenerovali izotopovou obálku, ve výsledku by se obálka nenašla. Podobně jako v předchozím případě můžeme vyžadovat, aby daná látka byla nalezena ve více spektrech. Pro látky, které projdou všemi filtrovacími kritérii, vygenerujeme izotopické obálky pomocí multinomiální expanze a dále pokračujeme stejným postupem jako u dereplikace.

Protože pomocí *de novo* přístupu anotujeme mnohem více experimentálních píků než při dereplikaci, je vhodné detailně analyzovat i chyby měření hodnot *m/z* a odchylky teoretických a experimentálních intenzit pro jednotlivé izotopické píky[3]. Pro daný izotopický pík a nejvyšší pík v obálce vypočteme absolutní hodnotu rozdílu hodnot *m/z*. Toleranci chyby označíme $\tau_{m/z}$ a otestujeme pomocí vzorce $\max_i \| TMZ_i - TMMZ| - |EMZ_i - EMMZ\| / TMMZ \times 10^6 \leq \tau_{m/z}$, kde TMZ_i je teoretická *m/z* hodnota izotopického píku, EMZ_i je experimentální *m/z* hodnota odpovídajícího píku, *TMMZ* je *m/z* hodnota nejintenzivnějšího teoretického píku a *EMMZ* je *m/z* hodnota nejintenzivnějšího experimentálního píku. Pro danou izotopovou obálku tedy vy-

počteme rozdíl mezi teoretickou hodnotou *m/z* určitého izotopu a nejvyššího píku $|TMZ_i - TMMZ|$. Stejný postup aplikujeme i na odpovídající experimentální píky $|EMZ_i - EMMZ|$. Dále vypočteme rozdíl obou hodnot. Protože postup aplikujeme na všechny izotopické píky v obálce, vybereme maximální hodnotu. Následně provedeme ještě převod na jednotky ppm vydělením hodnotou *TMMZ* a vynásobením koeficientem 10^6. Pokud je výsledná hodnota $\leq \tau_{m/z}$, danou látku zachováme, v opačném případě ji z výsledků vyhledávání odstraníme.

Parametr $\tau_{m/z}$ je vhodný i pro data, která byla naměřena s vysokým rozlišením, ale nebyla správně nakalibrována. Můžeme tedy prohledávat spektra s tolerancí chyby měření hodnoty *m/z* například 10 ppm a dále nastavit, že chyba rozdílů hodnot *m/z* mezi nejintenzívnějším píkem v obálce a libovolným jiným izotopickým partnerem může být maximálně 3 ppm. Pro úplnost zdůrazněme, že zatímco chybu měření v tomto případě uvažujeme v intervalu <–10, 10> ppm, hodnota $\tau_{m/z}$ se pohybuje v intervalu <0, 3> ppm.

Toleranci chyby rozdílu teoretické a experimentální intenzity označíme τ_{int} a otestujeme pomocí vzorce

$$\max_i |TI_i / TMI - EI_i / EMI| \times 100 \leq \tau_{int}$$

kde TI_i je teoretická relativní intenzita izotopického píku, EI_i je odpovídající experimentální relativní intenzita, *TMI* je relativní intenzita nejvyššího teoretického píku a *EMI* je relativní intenzita nejvyššího experimentálního píku. Protože po přiřazení teoretických a experimentálních píků nastavujeme hodnotu intenzity nejvyššího teoretického píku na hodnotu intenzity přiřazeného experimentálního píku, přičemž intenzity ostatních teoretických píků proporcionálně redukujeme, můžeme vzorec přepsat do tvaru:

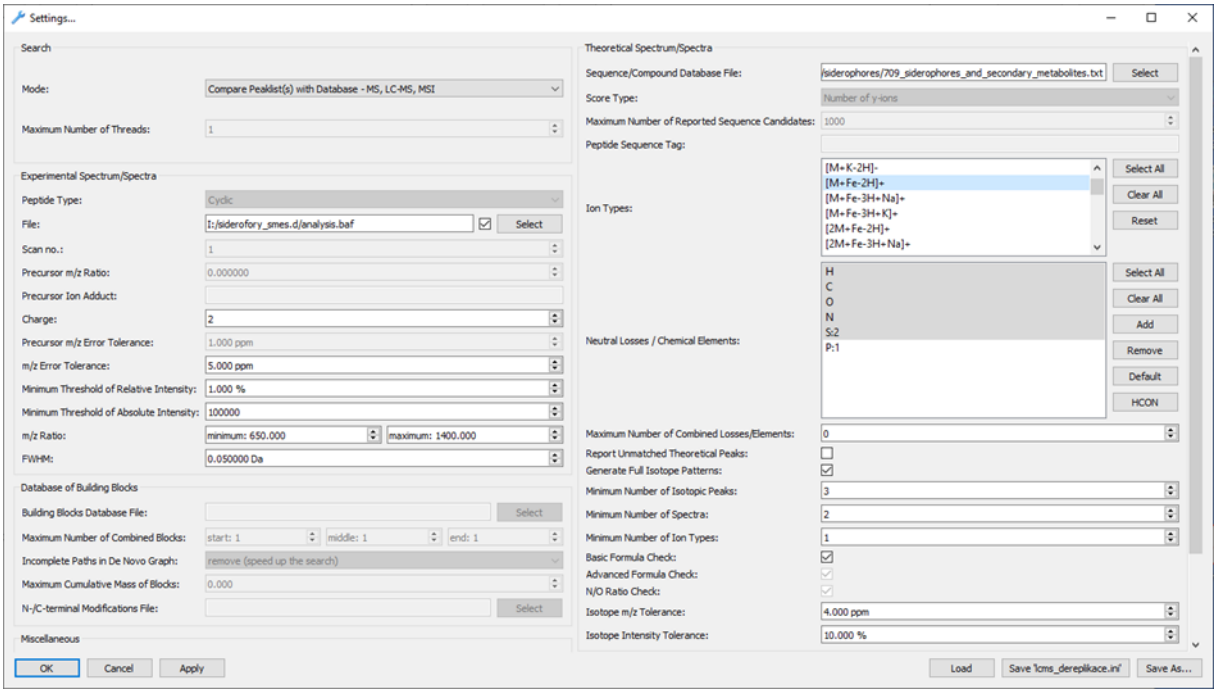
$$\max_i |(TI'_i - EI_i) / EMI| \times 100 \leq \tau_{int}$$

kde $TI'_i = TI_i \times EMI / TMI$ je normalizovaná teoretická intenzita. Pro vysvětlení uveďme, že pokud je *EMI* = 100 % a τ_{int} = 10 %, tolerance chyby relativní intenzity jednotlivých izotopů je 10 %. Pro *EMI* = 50 % je tolerance chyby relativní intenzity izotopů jen 5 %, pro EMI = 20 % jen 2 %, atd.

4. Bioinformatická podpora

Při analýze dat pomocí aplikace CycloBranch nejprve otevřeme dialog nastavení (v hlavním okně vybereme „Search → Settings" (obr. 3). Pro dereplikaci vybereme z nabídky „Compare Peaklist(s) with Database – MS, LC-MS, MSI", pro *de novo* analýzu sumárních vzorců pak režim „Compound Search – MS, LC-MS, MSI". Ostatní módy slouží pro analýzu MS/MS spekter[2]. Vstupní soubor s hmotnostními spektry může být ve formátech založených na jazyce XML (eXtensible Markup Language). Pro LC-MS data doporučujeme formát mzML (cit.[22]), při zobrazovací analýze (MSI, mass spectrometry imaging) je nutné ze softwaru výrobce vyexportovat data ve formátu imzML (cit.[23]).

Pro zpracování imzML souboru je potřeba nainstalovat knihovnu OpenMS ve verzi 2.3 a vyšší[24]. Aplikace je



Obr. 3. **Příklad nastavení aplikace pro dereplikaci v LC-MS datech.** Pro *de novo* analýzu změníme mód programu na hodnotu „Compound Search – MS, LC-MS, MSI" a podle velikosti zkoumaných látek nastavíme hodnotu „Maximum Number of Combined Elements" např. na 150 nebo 200. Při analýze MSI dat vybereme vstupní soubor ve formátu imzML a volitelně zvýšíme hodnotu „Minimum Number of Spectra" např. na 50.

navržená i pro zpracování objemných dat, které bývají často výstupem MSI analýzy (řádově desetitisíce spekter, desítky gigabytů). Profilová data z imzML souboru jsou rozdělena po 100 spektrech a s využitím nástrojů z knihovny OpenMS následně v dávce zpracována do výstupního imzML souboru, který je řádově menší než původní soubor a obsahuje čárová spektra.

Kromě formátu XML je možné zadat vstupní data i v nativních formátech některých výrobců, v současné době jsou podporovány formáty baf (Bruker), raw (Thermo) a raw (Waters). Alternativně můžeme použít i formát txt (na každém řádku uvedeme hodnotu *m/z* a intenzitu oddělenou tabulátorem, jednotlivá spektra oddělíme prázdným řádkem).

Dále nastavíme maximální hodnotu náboje generovaných iontů „Charge“. Chceme-li generovat teoretické píky odpovídající iontům $[M+H]^+$ a $[M+2H]^{2+}$, použijeme hodnotu 2. Pro píky odpovídající iontům $[M–H]^–$ zadáme hodnotu –1. Dále nastavíme toleranci přesnosti hodnot *m/z* (např. ±5 ppm). Volitelně nastavíme i minimální práh relativní a absolutní intenzity analyzovaných píků, přičemž obě hodnoty se aplikují současně. Zejména při *de novo* analýze je důležité vhodně nastavit interval, ve kterém budou zkoumané hodnoty *m/z* (minimum a maximum „*m/z* ratio“). Dostatečným zúžením tohoto intervalu můžeme výrazně omezit množství vygenerovaných sumárních vzorců[3]. Dosáhneme tak nejen urychlení výpočtu, ale ušetříme i hlavní paměť počítače. Dále nastavíme hodnotu FWHM, která by měla odpovídat zkoumaným spektrům. Doporučujeme nejprve zobrazit zkoumaná profilová spektra (např. v akvizičním software) a odhadnout typickou šířku píku v polovině jeho výšky.

V pravé části dialogu zvolíme databázi látek, ve které budeme vyhledávat. V současnosti jsou k dispozici databáze neribozomálních peptidů[25], sideroforů[9], lipidů[26] a mikrobiálních metabolitů rodů *Alternaria*, *Aspergillus*, *Candida*, *Cladosporium*, *Eurotium*, *Fennellia*, *Metarhizium*, *Paecilomyces*, *Phoma* a *Trichothecium*. Vzhledem k tomu, že databáze je v textovém formátu, je možné si jednoduše vytvořit vlastní databázi s využitím vestavěného editoru, vlastního skriptu nebo s využitím šablony pro Microsoft Excel (šablona je volně ke stažení na stránkách aplikace CycloBranch). Dále vybereme typ generovaných iontů, např. $[M+H]^+$ nebo $[M+Fe–2H]^+$.

Pokud provádíme dereplikaci, je možné navolit seznam neutrálních ztrát (např. H_2O a NH_3) spolu s maximálním počtem jejich opakování. Například hodnota 2 znamená, že pro každý teoretický pík budou generovány i píky odpovídající ztrátám H_2O, NH_3, H_2OH_2O, NH_3NH_3 a H_2ONH_3. Tato funkce má však častější využití při analýze MS/MS spekter.

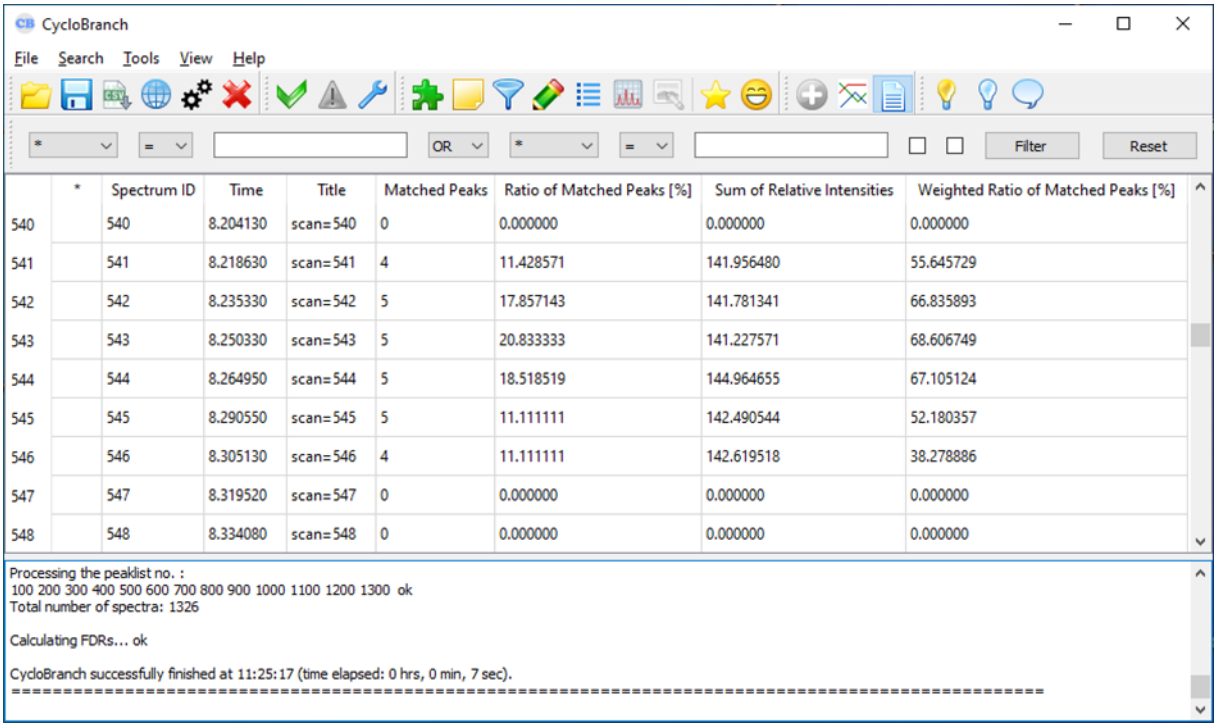
Pokud provádíme *de novo* analýzu, nastavíme ve stejném okně seznam chemických prvků, ze kterých budeme skládat jednotlivé sumární vzorce (např. H, C, O, N) a omezíme maximální počet prvků ve vzorci (např. 150 nebo 200). Chceme-li omezit maximální počet výskytů jednoho prvku, učiníme tak přímo v seznamu prvků např. pomocí zápisu „S:2“. V takovém případě tedy povolíme výskyt maximálně dvou atomů síry. Aplikace umožňuje reportovat i všechny vygenerované, ale nepřiřazené teoretické píky, což může být někdy výhodné pro malé databáze látek. V případě velkých databází nebo velkého množství generovaných sumárních vzorců doporučujeme kvůli zvýšení výkonu aplikace a přehlednosti výsledků ponechat tuto možnost vypnutou.

Pomocí volby „Generate Full Isotope Patterns“ povolíme generování teoretických izotopových obálek. Pokud bychom tuto funkci neaktivovali, budou se generovat pouze teoretické píky odpovídající monoizotopickým iontům. To může být vhodné pro rychlé prohledání spekter, obvykle však dostaneme velké množství falešně pozitivních výsledků. Parametrem „Minimum Number of Isotopic Peaks“ nastavíme minimální vyžadovaný počet izotopických píků v teoretické izotopové obálce, které musí být přiřazeny experimentálním píkům, aby daná látka byla hlášena jako nalezená. V případě, že nastavíme minimální počet izotopických píků na hodnotu vyšší než 1 a hledáme ionty typu $[M+Fe–2H]^+$, automaticky se z výsledků vyhledávání odstraní látky, pro které nebyl nalezen pík odpovídající izotopu ^{54}Fe. Dále se odstraní látky, pro které je poměr intenzit píků odpovídajících $^{54}Fe/^{56}Fe$ větší než empiricky stanovená hodnota 0,1. Pokud nastavíme vysokou hodnotu hmotnostního rozlišení formou FWHM ≤ 0,001, aplikují se podobná pravidla i na poměry intenzit píků v rámci jemných izotopových struktur. Takto můžeme postihnout a rozlišit i nuklidy $^{34}S/^{32}S$, $^{41}K/^{39}K$, $^{65}Cu/^{63}Cu$, $^{60}Ni/^{58}Ni$, $^{62}Ni/^{58}Ni$, $^{66}Zn/^{64}Zn$, $^{67}Zn/^{64}Zn$, $^{68}Zn/^{64}Zn$, apod.

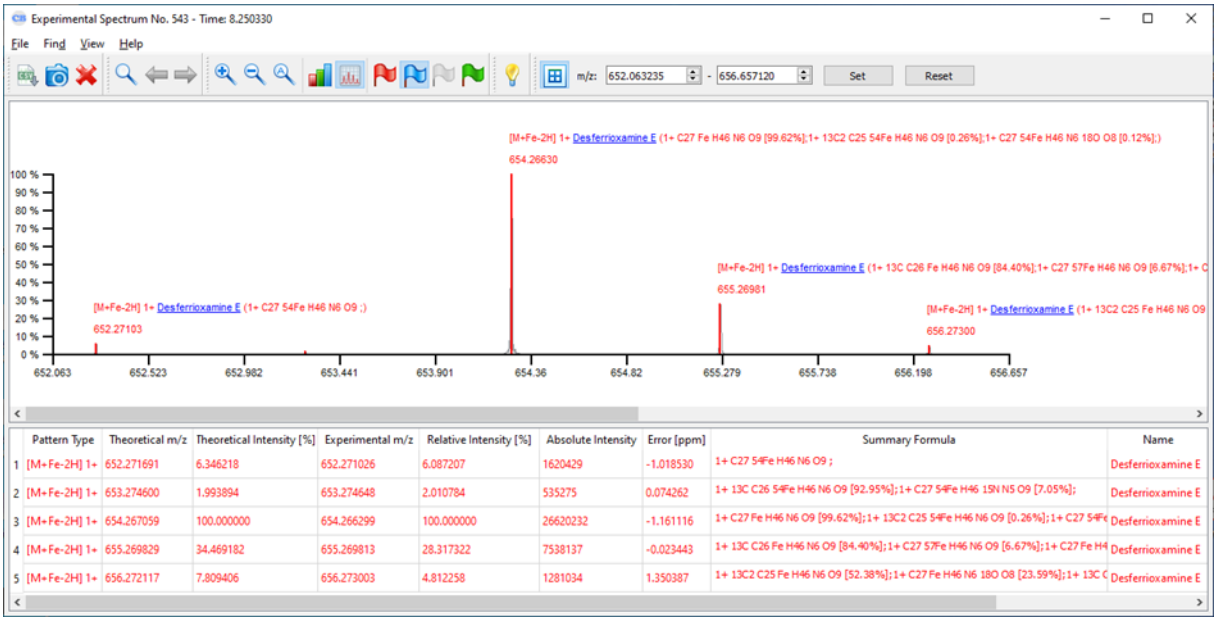
Podobně lze využít i další parametr, který definuje minimální počet spekter, ve kterých musí být daná látka nalezena. Jak již bylo uvedeno v části věnované dereplikaci, pro LC-MS data můžeme použít např. hodnotu 2 nebo 3. Zde navíc platí podmínka, že spektra, ve kterých byla látka nalezena, musí být v chromatogramu hned za sebou. Pro MSI data můžeme nastavit například hodnotu 50, tedy látka musí být nalezena minimálně v 50 „pixelech“. Přestože existují i práce, které se zabývají rozložením těchto pixelů v ploše obrázku[18], v současné verzi naší aplikace tato funkce není implementována. Jedním z důvodů je i fakt, že v závislosti na nastavení (např. prahu minimální intenzity), nemusíme při zobrazení hledané látky nutně pozorovat souvislou plochu a mohli bychom tak přijít o informace, které hledáme.

Poslední volbou je minimální počet různých iontů, které musí být nalezeny pro danou látku „Minimum Number of Ion Types“. Nastavením hodnoty na 2 a výběrem iontů $[M+Na]^+$ a $[M+K]^+$ tak můžeme zajistit, že ve výsledku uvidíme pouze látky, pro které se najdou oba typy iontů zároveň. U LC-MS analýzy se zvolené typy iontů mohou objevit v odlišném retenčním čase. Při zpracování MSI dat musí být oba typy iontů nalezeny ve stejných spektrech. Všechny tři uvedené parametry lze libovolně kombinovat. Můžeme tak hledat látky, pro které chceme ve výstupu vidět minimálně dva teoretické izotopické píky přiřazené experimentálním píkům minimálně ve třech spektrech za sebou apod.

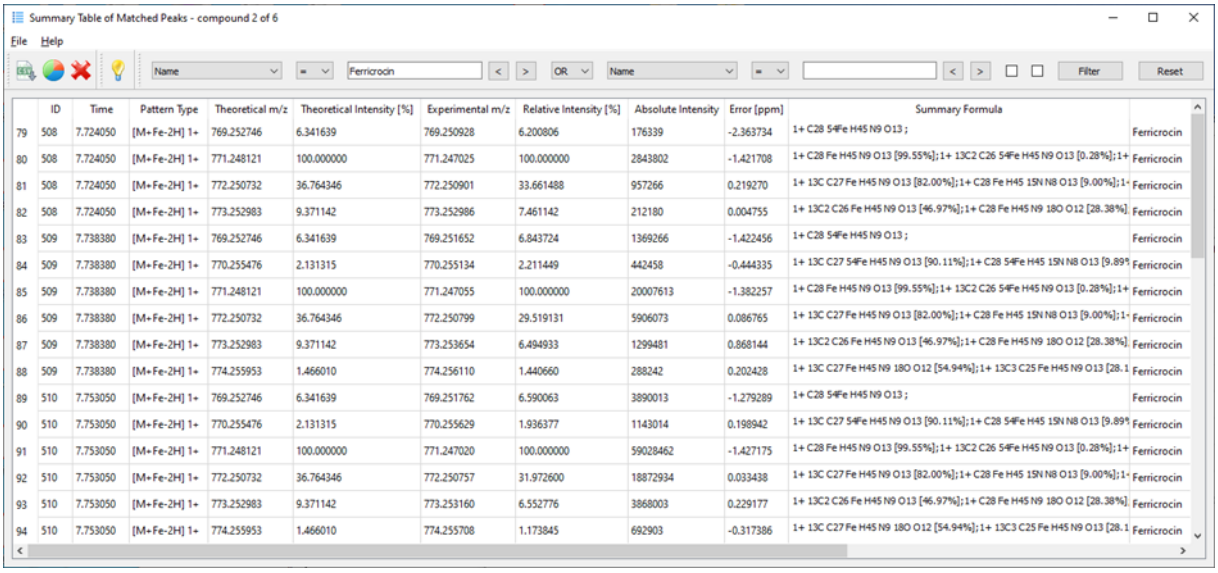
Pokud provádíme *de novo* analýzu, doporučujeme zapnout kontrolu vygenerovaných sumárních vzorců pomocí funkce „Basic Formula Check“, která aktivuje použi-



Obr. 4. **Hlavní okno a výstup aplikace zobrazující seznam jednotlivých spekter v LC-MS analýze.** Pro každé spektrum je zobrazena statistika počtu anotovaných píků, součet jejich intenzit, aj. Dvojklikem lze otevřít detail daného spektra. Z nabídky „Tools“ lze rovněž otevřít chromatogram nebo okno pro obrazovou fúzi. Při analýze MSI dat se místo sloupce udávajícího retenční čas zobrazí sloupce se souřadnicemi [x, y] daného spektra. (Barevná verze obrázku je dostupná na webových stránkách časopisu Chemické listy).



Obr. 5. **Detail hmotnostního spektra.** Spektrum lze otevřít dvojklikem na příslušný řádek v hlavním okně programu, na chromatografický pík v okně zobrazujícím chromatogram, na čtverec odpovídající danému spektru v okně s obrazovou fúzí nebo na řádek v tabulce zobrazující seznam všech anotovaných píků v celém souboru. (Barevná verze obrázku je dostupná na webových stránkách časopisu Chemické listy).



	ID	Time	Pattern Type	Theoretical m/z	Theoretical Intensity [%]	Experimental m/z	Relative Intensity [%]	Absolute Intensity	Error [ppm]	Summary Formula	
79	508	7.724050	[M+Fe-2H] 1+	769.252746	6.341639	769.250928	6.200806	176339	-2.363734	1+ C28 54Fe H45 N9 O13 ;	Ferricrocin
80	508	7.724050	[M+Fe-2H] 1+	771.248121	100.000000	771.247025	100.000000	2843802	-1.421708	1+ C28 Fe H45 N9 O13 [99.55%];1+ 13C2 C26 54Fe H45 N9 O13 [0.28%];1+	Ferricrocin
81	508	7.724050	[M+Fe-2H] 1+	772.250732	36.764346	772.250901	33.661488	957266	0.219270	1+ 13C C27 Fe H45 N9 O13 [82.00%];1+ C28 Fe H45 15N N8 O13 [9.00%];1+	Ferricrocin
82	508	7.724050	[M+Fe-2H] 1+	773.252983	9.371142	773.252986	7.461142	212180	0.004755	1+ 13C2 C26 Fe H45 N9 O13 [46.97%];1+ C28 Fe H45 N9 18O O12 [28.38%]	Ferricrocin
83	509	7.738380	[M+Fe-2H] 1+	769.252746	6.341639	769.251652	6.843724	1369266	-1.422456	1+ C28 54Fe H45 N9 O13 ;	Ferricrocin
84	509	7.738380	[M+Fe-2H] 1+	770.255476	2.131315	770.255134	2.211449	442458	-0.444335	1+ 13C C27 54Fe H45 N9 O13 [90.11%];1+ C28 54Fe H45 15N N8 O13 [9.89%	Ferricrocin
85	509	7.738380	[M+Fe-2H] 1+	771.248121	100.000000	771.247055	100.000000	20007613	-1.382257	1+ C28 Fe H45 N9 O13 [99.55%];1+ 13C2 C26 54Fe H45 N9 O13 [0.28%];1+	Ferricrocin
86	509	7.738380	[M+Fe-2H] 1+	772.250732	36.764346	772.250799	29.519131	5906073	0.086765	1+ 13C C27 Fe H45 N9 O13 [82.00%];1+ C28 Fe H45 15N N8 O13 [9.00%];1+	Ferricrocin
87	509	7.738380	[M+Fe-2H] 1+	773.252983	9.371142	773.253654	6.494933	1299481	0.868144	1+ 13C2 C26 Fe H45 N9 O13 [46.97%];1+ C28 Fe H45 N9 18O O12 [28.38%]	Ferricrocin
88	509	7.738380	[M+Fe-2H] 1+	774.255953	1.466010	774.256110	1.440660	288242	0.202428	1+ 13C C27 Fe H45 N9 18O O12 [54.94%];1+ 13C3 C25 Fe H45 N9 O13 [28.1	Ferricrocin
89	510	7.753050	[M+Fe-2H] 1+	769.252746	6.341639	769.251762	6.590063	3890013	-1.279289	1+ C28 54Fe H45 N9 O13 ;	Ferricrocin
90	510	7.753050	[M+Fe-2H] 1+	770.255476	2.131315	770.255629	1.936377	1143014	0.198942	1+ 13C C27 54Fe H45 N9 O13 [90.11%];1+ C28 54Fe H45 15N N8 O13 [9.89%	Ferricrocin
91	510	7.753050	[M+Fe-2H] 1+	771.248121	100.000000	771.247020	100.000000	59028462	-1.427175	1+ C28 Fe H45 N9 O13 [99.55%];1+ 13C2 C26 54Fe H45 N9 O13 [0.28%];1+	Ferricrocin
92	510	7.753050	[M+Fe-2H] 1+	772.250732	36.764346	772.250757	31.972600	18872934	0.033438	1+ 13C C27 Fe H45 N9 O13 [82.00%];1+ C28 Fe H45 15N N8 O13 [9.00%];1+	Ferricrocin
93	510	7.753050	[M+Fe-2H] 1+	773.252983	9.371142	773.253160	6.552776	3868003	0.229177	1+ 13C2 C26 Fe H45 N9 O13 [46.97%];1+ C28 Fe H45 N9 18O O12 [28.38%]	Ferricrocin
94	510	7.753050	[M+Fe-2H] 1+	774.255953	1.466010	774.255708	1.173845	692903	-0.317386	1+ 13C C27 Fe H45 N9 18O O12 [54.94%];1+ 13C3 C25 Fe H45 N9 O13 [28.1	Ferricrocin

Obr. 6. **Přehledná tabulka píků anotovaných v celém souboru dat.** Pomocí šipek lze postupně procházet a zobrazovat názvy všech nalezených látek (resp. sumární vzorce při *de novo* analýze). Po nastavení filtrovacích kritérií se automaticky aktualizují i grafické informace zobrazené v chromatogramu nebo okně s obrazovou fúzí.

tí pravidel pro kontrolu valenčních stavů jednotlivých prvků podle Seniora, viz výše. Funkce „Advanced Formula Check“ pak přidává pravidla (4), (5) a (6) podle Kinda a Fiehna. Poslední možnost „N/O Ratio Check“ slouží pro odstranění látek, ve kterých je poměr počtu atomů dusíku vůči počtu atomů kyslíku vyšší než 1. Tento filtr lze s výhodou použít k zúžení seznamu elementárních složení při analýze peptidů. V případě dereplikace i *de novo* analýzy můžeme nastavit hodnotu $\tau_{m/z}$ parametrem „Isotope *m/z* Tolerance“ a τ_{int} parametrem „Isotope Intensity Tolerance“.

Po nastavení parametrů algoritmus zahájíme příkazem „Search → Run” v hlavním menu aplikace. Po ukončení vyhledávání se v hlavním okně zobrazí seznam jednotlivých spekter (obr. 4), která můžeme dvojitým kliknutím myší otevřít a dále analyzovat (obr. 5). Příkazem „Tools → Summary Table of Matched Peaks“ můžeme zobrazit tabulku všech píků anotovaných ve všech spektrech (obr. 6). Při filtrování řádků v tabulce se automaticky aktualizuje zobrazení chromatografických píků (obr. 1) v případě LC-MS dat nebo zobrazení fúze s optickým obrazem v případě MSI dat (obr. 2). Můžeme takto snadno procházet a zobrazovat nalezené látky. Dvojitým kliknutím myší na daný chromatografický pík nebo na daný „pixel“ v okně s fúzí můžeme rovněž zobrazit detail příslušného spektra. Podrobnosti k obrazové fúzi a další informace lze najít v pokynech pro uživatele na webu aplikace (https://ms.biomed.cas.cz/cyclobranch/docs/html/tutorials.html).

Práce byla podpořena Grantovou agenturou České republiky (21-17044S).

LITERATURA

1. Škultéty L., Novák J., Havlíček V.: Chem. Listy *114*, 145 (2020).
2. Novák J., Havlíček V.: Chem. Listy *114*, 200 (2020).
3. Novák J., Škríba A., Havlíček V.: Anal. Chem. *92*, 6844 (2020).
4. Luptáková D., Havlíček V.: Chem. Listy *114*, 216 (2020).
5. McDonnell L. A., Heeren R. M. A.: Mass Spectrom. Rev. *26*, 606 (2007).
6. Škríba A., Houšť J., Havlíček V.: Chem. Listy *114*, 119 (2020).
7. Krásný L., Strohalm M., Bouchara J.-P., Šulc M., Lemr K., Barreto-Bergter E., Havlíček V.: Mycoses *54*, 37 (2011).
8. Novák J., Lemr K., Schug K. A., Havlíček V.: J. Am. Soc. Mass Spectrom. *26*, 1780 (2015).
9. Pluháček T., Lemr K., Ghosh D., Milde D., Novák J., Havlíček V.: Mass Spectrom. Rev. *35*, 35 (2016).
10. Hider R. C., Kong X. L.: Nat. Prod. Rep. *27*, 637 (2010).
11. Novák J., Sokolová L., Lemr K., Pluháček T., Palyzová A., Havlíček V.: BBA-Proteins Proteomics *1865*, 768 (2017).
12. Novák J., Škríba A., Zápal J., Kuzma M., Havlíček V.: J. Mass Spectrom. *53*, 1097 (2018).
13. Valkenborg D., Mertens I., Lemiere F., Witters E., Burzykowski T.: Mass Spectrom. Rev. *31*, 96 (2012).
14. Audi G., Wapstra A. H.: Nucl. Phys. A *565*, 1 (1993).
15. Rosman K. J. R., Taylor P. D. P.: Pure Appl. Chem. *70*, 217 (1998).
16. Sadílek M.: Chem. Listy *114*, 133 (2020).

17. Käll L., Storey J. D., MacCoss M. J., Noble W. S.: J. Proteome Res. *7*, 29 (2008).
18. Palmer A. a 11 spoluautorů: Nat. Methods *14*, 57 (2017).
19. Wang X. S., Jones D. R., Shaw T. I., Cho J. H., Wang Y. Y., Tan H. Y., Xie B., Zhou S. P., Li Y. X., Peng J. M.: J. Proteome Res. *17*, 2328 (2018).
20. Kind T., Fiehn O.: BMC Bioinformatics *8*, (2007).
21. Senior J. K.: Am. J. Math. *73*, 663 (1951).
22. Martens L. a 17 spoluautorů: Mol. Cell. Proteomics *10*, (2011).
23. Schramm T. a 11 spoluautorů: J. Proteomics *75*, 5106 (2012).
24. Röst H. L. a 27 spoluautorů: Nat. Methods *13*, 741 (2016).
25. Caboche S., Pupin M., Leclere V., Fontaine A., Jacques P., Kucherov G.: Nucleic Acids Res. *36*, D326 (2008).
26. Sud M. a 10 spoluautorů: Nucleic Acids Res. *35*, D527 (2007).

J. Novák and V. Havlíček *(Institute of Microbiology of the Czech Academy of Sciences, Prague, Czech Republic):* **Compound Dereplication and *De Novo* Characterization of Small Molecules by Mass Spectrometry**

We describe the molecular dereplication principles and *de novo* characterization of small molecules obtained from liquid-chromatography mass spectrometry and imaging mass spectrometry data sets. Our methodology aims at supporting chemists and computer programmers to understand the hidden computing algorithms used for metabolomics mass spectrometry data processing. The approaches have been made available in the open-source tool CycloBranch. The presented tutorial extends the interpretation of mass spectra portfolia described in a series of papers published in Chemicke listy, issues 2/2020 and 3/2020.

Keywords: CycloBranch, dereplication, *de novo* characterization, mass spectrometry, metabolomics, mass spectrometry imaging, liquid chromatography, isotopic structure

• Novák J., Havlíček V.: Chem. Listy *116*, 11–19 (2022).
• https://doi.org/10.54779/chl20220011

Acknowledgements
This work was supported by the Czech Science Foundation (21-17044S).