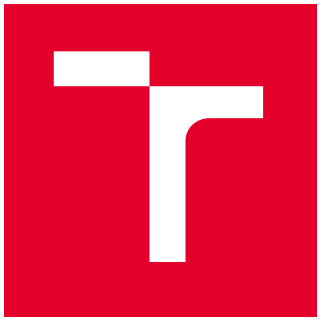


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

VÝVOJ WORKFLOW PRO KONTROLU KVALITY HMOTNOSTNĚ-SPEKTROMETRICKÝCH DAT V PROSTŘEDÍ KNIME

DEVELOPMENT OF WORKFLOW FOR QUALITY CONTROL OF MASS SPECTROMETRY DATA IN KNIME ENVIRONMENT

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. Anna Schneiderová

VEDOUCÍ PRÁCE
SUPERVISOR

Mgr. David Potěšil, Ph.D.

BRNO 2022

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní program **Biomedicínské inženýrství a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Bc. Anna Schneiderová ***ID:*** 203685

Ročník: 2 ***Akademický rok:*** 2021/22

NÁZEV TÉMATU:

Vývoj workflow pro kontrolu kvality hmotnostně-spektrometrických dat v prostředí KNIME

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši na téma kontroly kvality hmotnostně-spektrometrických dat z proteomických experimentů. 2) Zorientujte se v aktuálních možnostech zpracování proteomických dat v prostředí KNIME. 3) Sestavte seznam hlavních hodnotících kriterií kvality hmotnostně-spektrometrických dat z proteomických experimentů na základě literárních údajů i praktických zkušeností. 4) Navrhněte optimální metodiku automatizované kontroly kvality hmotnostně-spektrometrických dat z proteomických experimentů v prostředí KNIME. 5) Sestavte workflow pro automatizovanou kontrolu kvality v prostředí KNIME. 6) Proveďte návrh metodiky pro ověření workflow pro simulovaná a reálná proteomická data a workflow ověřte. 7) Proveďte diskuzi nad výsledky vývoje workflow a zhodnoťte jeho využitelnost a výhody a nevýhody řešení s řešeními z literárních zdrojů.

DOPORUČENÁ LITERATURA:

[1] Aebersold, R.; Mann, M. Mass Spectrometry-Based Proteomics. Nature, 2003, 422, 198–207.

[2] Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. Quantitative Mass Spectrometry in Proteomics: A Critical Review. Anal Bioanal Chem, 2007, 389, 1017–1031.

Termín zadání: 7.2.2022 ***Termín odevzdání:*** 20.5.2022

Vedoucí práce: Mgr. David Potěšil, Ph.D.

Konzultant: Ing. Vojtěch Bartoň

prof. Ing. Ivo Provazník, Ph.D.
předseda rady studijního programu

ABSTRAKT

Proteomický experiment s využitím kapalinové chromatografie a hmotnostní spektrometrie je velmi komplexní technika s celou řadou proměnných, které mohou ovlivnit kvalitu výstupních dat. Průběžná kontrola kvality výstupních dat a stavu používané instrumentace je proto klíčová pro získání kvalitních dat. Pro systematickou a automatizovanou kontrolu kvality dat bylo navrženo workflow, které bylo následně implementováno v prostředí KNIME. Workflow umožňuje získat a zaznamenat vybrané metriky pro kontrolu kvality, které je možné využít ke sledování variability v datech a zachytit počínající technické problémy systému.

KLÍČOVÁ SLOVA

proteomika, hmotnostní spektrometrie, kontrola kvality, identifikace peptidů, identifikace proteinů, kvantifikace peptidů, kvantifikace proteinů, KNIME, OpenMS, Python, SQLite

ABSTRACT

Proteomic experiment using liquid chromatography and mass spectrometry is a complex technique with multiple variables that can affect the quality of output data. Data quality control and instrument status monitoring are therefore essential for high quality data acquisition. Designed workflow, implemented within the KNIME environment, allows systematic and automatic data quality control. The workflow allows to obtain and record selected quality control metrics which could be used to ascertain data variability and prevent technical problems.

KEYWORDS

proteomics, mass spectrometry, quality control, peptide identification, protein identification, peptide quantification, protein quantification, KNIME, OpenMS, Python, SQLite

SCHNEIDEROVÁ, Anna. *Vývoj workflow pro kontrolu kvality hmotnostně-spektrometrických dat z proteomických experimentů v prostředí KNIME*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2022, 74 s. Diplomová práce. Vedoucí práce: Mgr. David Potěšil, Ph.D.

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Bc. Anna Schneiderová
VUT ID autora:	203685
Typ práce:	Diplomová práce
Akademický rok:	2021/22
Téma závěrečné práce:	Vývoj workflow pro kontrolu kvality hmotnostně-spektrometrických dat z proteomických experimentů v prostředí KNIME

Prohlašuji, že svou závěrečnou práci jsem vypracovala samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autorky*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Ráda bych poděkovala vedoucímu diplomové práce panu Mgr. Davidu Potěšilovi, Ph.D. za výborné odborné vedení, cenné konzultace, trpělivost, motivaci a přínosné rady k práci.

Obsah

Seznam obrázků

Úvod

Práce je zaměřená na kontrolu kvality hmotnostně spektrometrických dat. Kontrola kvality dat je důležitou součástí experimentu, aby bylo možné udržovat reprodukovatelnost výsledků, ale i zachytit z dat možné technické problémy vzniklé v průběhu proteomického experimentu. Kontrolu kvality je vhodné provádět systematicky a získané informace uchovávat pro sledování aktuálních experimentů, ale i pro kontrolu kvality v dlouhodobém časovém období. Což umožní zaznamenání vlivu jakýchkoli změn v rámci experimentu na kvalitu dat.

Pro systematickou kontrolu kvality je nutné znát zdroje variability z celého experimentu a možnosti kontroly kvality dat, což umožňují i standardy vkládané mezi měřené vzorky. Tuto problematiku přibližuje první část diplomové práce. Cílem práce je implementovat workflow ke zpracování dat a generování metrik, které umožní automatizaci celého procesu zpracování dat. Zpracování dat má několik kroků, které musí předcházet získání metrik a jejich ukládání. V rámci těchto kroků jsou z hrubých dat získány identifikace a kvantifikace peptidů, ale i proteinů. Možnostem identifikace a kvantifikace se věnuje první kapitola, ze které návrh a implementace workflow vychází.

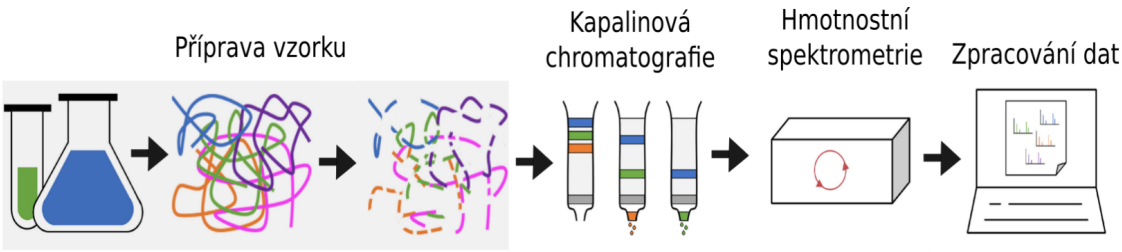
Vytvoření návrhu předchází rešerše možností práce s hmotnostně spektrometrickými daty v KNIME, ve kterém lze používat nejen zabudované funkcionality, ale umožňuje vytvoření vlastních funkcí pomocí skriptovacích jazyků. V KNIME je taktéž možné využít některé externí funkcionality jiných softwarů, například OpenMS, který umožňuje manipulaci s hrubými hmotnostně spektrometrickými daty.

Výstupem diplomové práce by mělo být workflow, které samostatně a automaticky získává metriky pro kvalitu hmotnostně spektrometrických dat a ukládá je do databáze tak, aby byly snadno přístupné uživatelům k hodnocení měření.

1 Kontrola kvality v proteomických experimentech s využitím hmotnostní spektrometrie

Hmotnostní spektrometrie kombinovaná s kapalinovou chromatografií je velmi komplexní technika, která slouží k detekci, identifikaci a kvantifikaci proteinů ve vzorku. Experiment, který začíná přípravou zkoumaného vzorku, je následován hmotnostní spektrometrií a končí analýzou dat, může mít velkou variabilitu, a proto je obtížné získat přesné a reprodukovatelné výsledky. Z tohoto důvodu je důležité, aby všechny experimenty podléhaly detailní a systematické kontrole kvality, která by měla zajistit lepší reprodukovatelnost měření. Mimo variabilitu proteomického experimentu je také vhodné sledovat očekávané charakteristiky zpracovávaných vzorků jako množství a kvalita proteinů. Kontrolu kvality je vhodné navrhovat tak, aby mohla odkrýt odchylky v experimentu v jakékoli části [1].

Proteomický experiment s využitím hmotnostní spektrometrie (MS) se skládá z několika částí. V první fázi je příprava vzorku, která zahrnuje lýzu buněk, denaturaci, redukci, alkylaci proteinů a štěpení proteinů na peptidy. Dalším krokem je kapalinová chromatografie spojená s hmotnostní spektrometrií. Kapalinová chromatografie peptidy separuje a umožňuje tak postupnou eluci peptidů do iontového zdroje hmotnostního spektrometru. V iontovém zdroji (nejčastěji elektrosprej) se peptidy nabíjí a převádí do plynné fáze. Nabité peptidy v závěru vstupují do hmotnostního spektrometru, jehož výstupem jsou hmotnostní spektra zobrazující relativní intenzitu iontů v závislosti na poměru hmotnosti a nábojového čísla (m/z) [2]. Posledním krokem je zpracování dat (obr. 1.1). Peptidy mohou být identifikovány za pomoci vyhledávání v databázích a kvantifikovány, dále jsou peptidy mapovány na proteiny, popřípadě jsou proteiny také kvantifikovány [3].

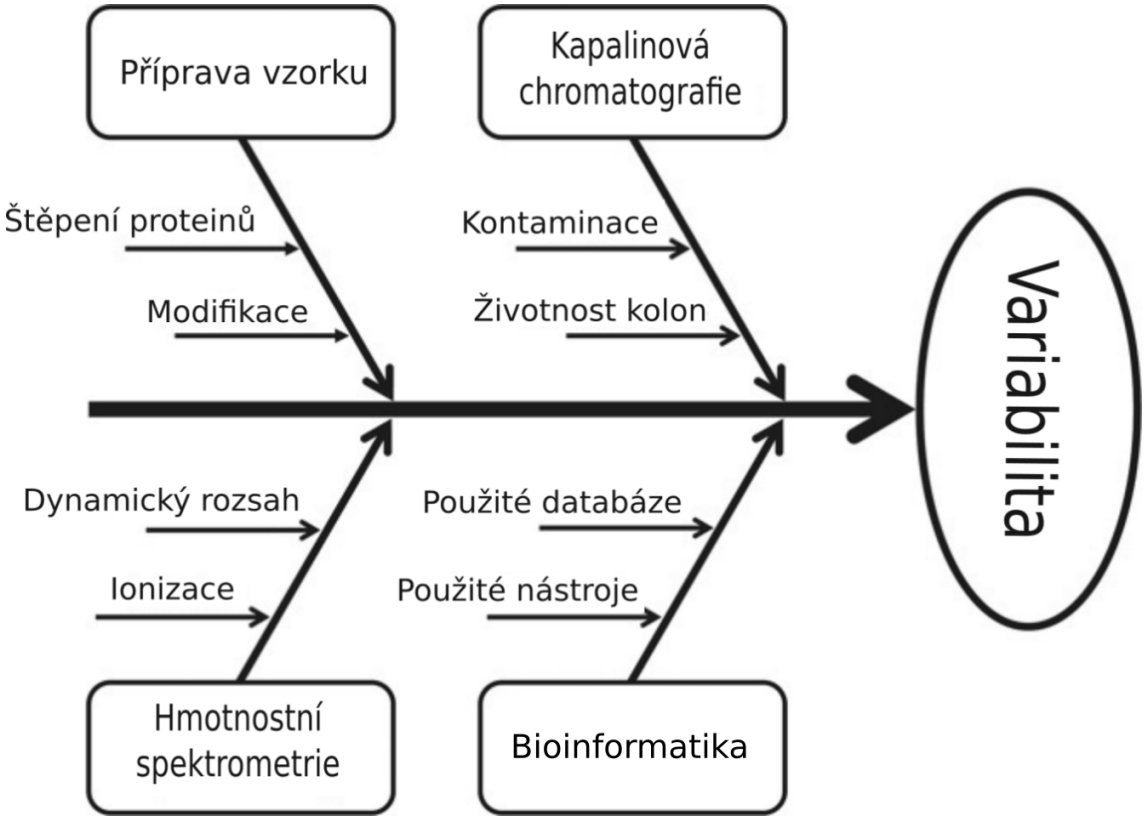


Obr. 1.1: Průběh proteomického experimentu s využitím hmotnostní spektrometrie. Převzato z: [1].

1.1 Zdroj variability v měření

Variabilita dat pochází z jakékoli části výše zmíněného procesu. Ve fázích příprav vzorku a LC-MS analýzy má největší vliv na variabilitu dat extrakce proteinů ze vzorku. Variabilita vzniká např. kontaminací v průběhu zpracování vzorků či nestejnou homogenizací všech zpracovaných vzorků. Na druhém místě ve vlivu na variabilitu je samotná LC-MS analýza. Variabilita je způsobena technickými limity a aspekty LC-MS systémů jako jsou životnosti kolon, přenosem vzorků mezi analýzami, dynamickým rozsahem, ionizací nebo přístrojovou stabilitou. Nejmenší vliv má průběh štěpení proteinů [4]. Ve fázi zpracování dat ovlivňují variabilitu použité databáze a vybrané nástroje a algoritmy (obr. 1.2).

Aby variabilita byla snížena na minimum, je potřeba vytvořit a dodržovat optimalizovaný a přesný pracovní postup, aby bylo dosaženo reprodukovatelných a přesných výsledků. Avšak i tak je potřeba zajištění kvality a systematické kontroly kvality.



Obr. 1.2: Základní zdroje variability v jednotlivých krocích experimentu. Převzato z: [1]

1.1.1 Fáze přípravy vzorku

Fáze přípravy vzorku je velmi důležitá, protože výsledky jsou výhradně závislé na kvalitě vzorku. V této fázi se vyskytuje několik proměnných, které můžou ovlivnit kvalitu měření a jedna z nich jsou neočekávané modifikace proteinu. Modifikační artefakty se můžou objevit po kompletním rozvinutí proteinu. Rozvinutí je nutné pro enzymatické štěpení proteinu na peptidy [5]. Jednou z vyskytujících se modifikací je nežádoucí karbamylace proteinů, která bývá způsobena reverzibilní denaturací proteinu močovinou za zvýšené teploty. Karbamylace negativně ovlivňuje proteolýzu trypsinem, stav náboje, retenční časy, ale i hmotnost peptidů [6]. Dalším zdrojem nežádoucích modifikací může být krok alkylace, kdy se často používá jodacetamid. Při nadměrné alkylaci dochází na N-koncích peptidu k karbamidometylaci. Methioninová oxidace může vzniknout kvůli vystavením vzorku kyslíku (vzduchu). Z těchto důvodů je třeba dbát velké opatrnosti při přípravě vzorků a správně nastavit parametry databázového prohledávání (viz 1.3.1) tak, aby se ověřil možný výskyt očekávaných i neočekávaných modifikací v experimentu. Chybějící či nesprávné identifikace peptidů mohou být způsobeny i semitryptickými a nespecifickými peptidy, které vznikají během proteinového štěpení trypsinem [7]. Ke snížení reprodukovatelnosti mohou vést i ztráty vzorků způsobené adsorpcí peptidů na použité materiály. Pro práci se vzorkem je proto vhodné omezit práci se vzorkem na co nejméně kroků [8]. Během přípravy vzorků je navíc potřeba věnovat pozornost opatřením proti kontaminaci vzorků. Najčastějším typem kontaminace je keratin, který může pocházet z kůže, vlasů či prachu. K identifikaci kontaminace vzorků je nezbytné kontaminanty specifikovat v kroku identifikace peptidů databázovým prohledávání.

1.1.2 Fáze kapalinové chromatografie a hmotnostní spektrometrie

Kapalinová chromatografie

Jelikož kapalinová chromatografie má velký vliv na variabilitu výsledných dat, je nutné monitorovat chromatografická data. Vliv na variabilitu dat můžou mít technické problémy. Mezi ně může patřit chybná kalibrace přístroje nebo zanesená kolona. Jedno z upozornění na výměnu nebo servis kolon kapalinové chromatografie jsou rozšiřující se píky, které udávají zhoršení separace. Píky by měly být v ideálním případě úzké a symetrické. Špatný vliv na tvar píků může mít příliš velké množství analytu v kolonách, ale i změny teploty kolon a prostředí. Dále je nutné se vyhnout kontaminaci přenosem mezi vzorky. Kontaminace může být způsobena analytem z předešlého měření, který se v některé fázi měření objeví v dalším měření. Tento typ kontaminace může být způsoben interakcí vzorků s materiály, se kterými

přijdou do kontaktu, nebo nedostatečným vymytím vzorku či jeho části ze systému před analýzou dalšího vzorku v důsledku přítomnosti tzv. mrtvého objemu. Mrtvé objemy, kde dochází ke smísení vzorků, jsou zdrojem rozšiřování píků a prodloužení času eluce. Podstatný vliv na separaci vzorku mají matriční efekty vzorku, což můžou být například inference separovaného peptidu s jeho matricí. Matriční efekty můžou způsobovat posun retenčních časů peptidů nebo ovlivňovat jejich intenzity či profily píků. [9]

Hmotnostní spektrometrie

Peptidy vystupující z kapalinového chromatografu jsou vstřikovány do hmotnostního spektrometru, kde je analyzována hodnota hmotnost ku náboji (m/z), avšak peptidy musí být prvně před zpracováním v hmotnostním spektrometru ionizovány. Stabilita ionizačního elektrospreje se dá hodnotit dle poklesu v iontovém proudu. Pokud se nachází větší podíl peptidů s nábojem jiným než 2^+, tak to indikuje ionizační problémy a může to mít vliv na poměr identifikací. Fragmentací peptidů kolizně indukovanou disociací mohou vzniknout semitryptické peptidy. Chybnou kalibrací systému mohou vzniknout problémy s přesností měřených m/z hodnot, ale i se sníženou intenzitou iontů v důsledku jejich horší účinností transportu uvnitř hmotnostního spektrometru. Matriční efekty můžou ovlivnit i část analýzy, která využívá hmotnostní spektrometrie. Složky matrice mohou být ionizovány a zaznamenány ve spektru stejně jako peptidy. Mohou tedy snižovat počet ionizovaných peptidů, ale i vnášet do spekter šum. [9]

Přesnost měření hmotnosti je možné kontrolovat pomocí přidání standardů, jejichž složení a hmotnost je známá a je možné je použít ke kalibraci hmotnosti [9]. Dynamický rozsah hmotnostního spektrometru může být kontrolován s využitím rozdílných koncentrací peptidů ve vzorku, naopak senzitivita je kontrolována vstřikováním malých objemů kontrolních standardů (viz 1.2) [10].

1.1.3 Fáze interpretace hrubých dat

Interpretace dat je nedílnou součástí experimentů v proteomice a je nutné, aby do zpracování nebyly zaváděny chyby, které by způsobily nereprodukovatelnost dat. Je zásadní, aby se udržoval pracovní postup fixní, data byla zpracovávána neměnným způsobem z dlouhodobého hlediska. Veškeré zavedené změny v pracovním postupu by měly být zaznamenány a mělo by se na stejných datech určit rozdíl mezi novým a starým postupem. Variabilitu výsledků totiž ovlivňuje nejen výběr algoritmů pro mapování peptidů na spektra, ale i verze nástrojů a kombinace procesních parametrů, jejichž výstup se může velmi lišit.

Kvalitu dat lze hodnotit podle vybraných metrik (viz 1.4). Počet identifikovaných peptidů a proteinů jsou jedny z nejzákladnějších hodnot, ze kterých lze vyvodit celkovou kvalitu experimentu. Avšak je nutné využití více metrik pro komplexnější pohled na data.

1.2 Standardy ke kontrole kvality

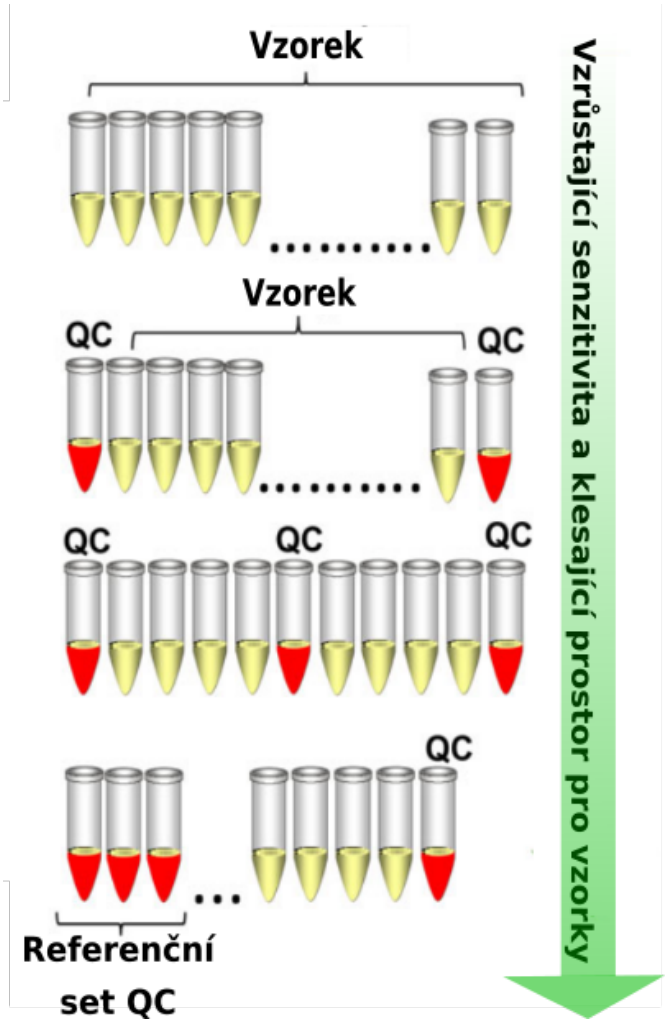
Systematická kontrola funkčnosti LC-MS přístrojů může být provedena vkládáním QC (*quality control*) vzorků – standardů mezi experimentální vzorky a zpracovávány v rámci celého pracovního postupu a díky nim mohou být pozorovány problémy vedoucí ke snížené kvalitě.

Vkládané standardy mohou být různé, avšak většinou se používají tři odlišné standardy pro kontrolu kvality. První standard obsahuje jednoduchou peptidovou směs složenou z peptidů BSA (*bovine serum albumin*), enolasy nebo cytochromu c. Druhý standard je více komplexní, obsahující celobuněčný lyzát (např. HeLa buněk) [11].

Dále existují specializované vzorky ke kontrole kvality LC-MS analýzy. Složení vzorku je takové, aby peptidy měly různou hydrofobicitu a jejich gradient byl charakteristický. Standardy obsahující peptidy s indexovaným retenčním časem (iRT) se využívají k normalizaci a korekci posunu retenčního času experimentů nebo zarovnání retenčních časů více experimentů. Lze na nich také sledovat vliv matrice analyzovaného vzorku na jejich separaci (tvar chromatografického píku a retenční čas) či chování v MS (intenzita, fragmentační spektra) a toto pozorování následně extrapolovat i na ostatní analyzované peptidy přítomné v samotném vzorku - např. pokles intenzity peptidů standardu velmi pravděpodobně znamená stejný efekt i u peptidů pocházejících ze vzorku.[12]

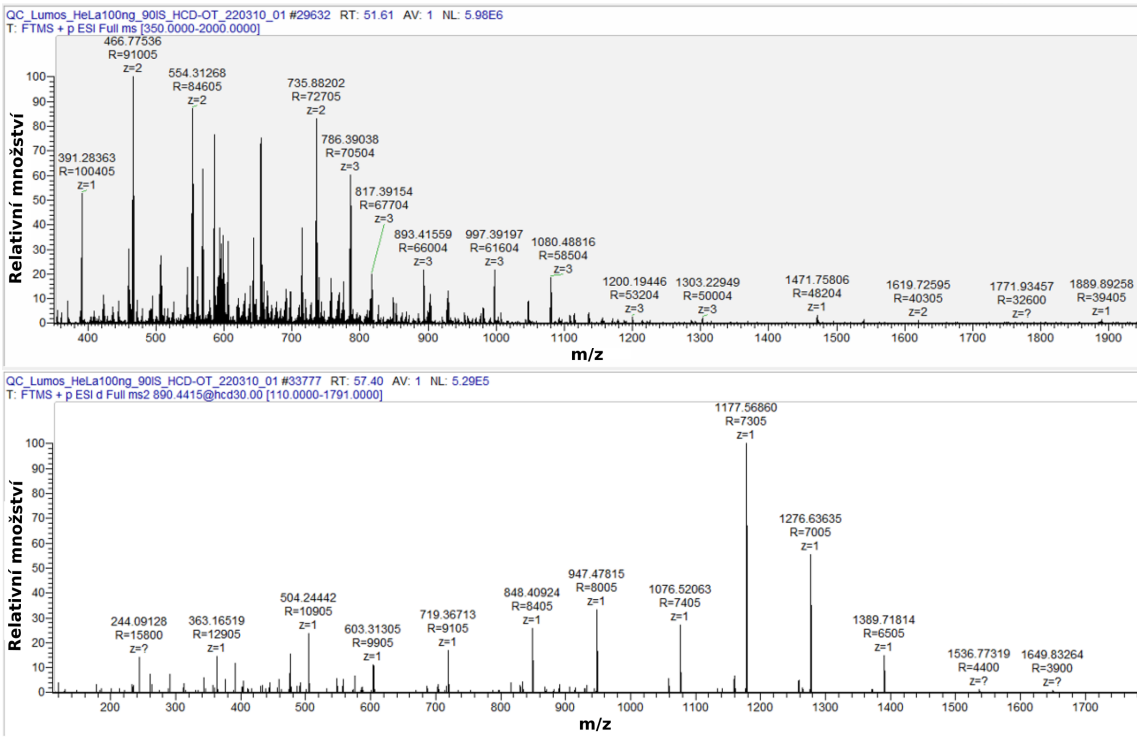
1.2.1 Vkládání standardů mezi experimenty

Standard s jednoduchou peptidovou směsí je zařazován k měření s větší frekvencí - několikrát denně, aby mohla být rychle zjištěna a zaznamenána funkčnost přístrojů, obzvlášť kapalinové chromatografie na základě šířky píků a retenčních časů. Komplexnější standard je pro kontrolu zařazován méně často (většinou jedenkrát denně), i z důvodů časové náročnosti měření. Tento standard je používán k hodnocení hmotnostního spektrometru. Při měření se většinou používají velmi malé množství vzorků, aby byly testovány limity detekovatelnosti. Standard se směsí syntetických peptidů může být buď měřen samostatně nebo je přidán do druhého komplexního standardu nebo do experimentálních vzorků. Z toho důvodu by se neměly syntetické peptidy překrývat svými vlastnostmi s originálními peptidy.



Obr. 1.3: Obrázek řazení kontrolních vzorků (QC) mezi analyzované vzorky. Převzato z: [13]

Způsob zařazení kontrolních standardů mezi experimentální vzorky je silně závislý na designu experimentu. Standardy mohou být například zařazeny na začátku a konci měření skupiny vzorků nebo pravidelně za stanovený počet vzorků, což umožní sledovat změny během měření a vyvarovat se ztrátám vzorků. Druhou možností je vkládání skupiny standardů na začátku měření a dále pokračovat s pravidelných vkládáním standardů, z nichž může být statisticky získán odhad průběhu stability kvality měření přístrojů (obr. 1.3). Tuto skupinu standardů před měřením experimentálních vzorků lze nahradit daty z předešlých měření a statistiky mohou být vyvozeny z nich. Avšak tato předešlá měření nelze použít pokud došlo ve změně použitého standardu. [13]



Obr. 1.4: Ukázka MS spektra na prvním řádku a MS/MS fragmentovaného spektra na druhém řádku ze vzorku HeLa.

1.3 Zpracování dat z měření

Data, která jsou výstupem proteomického experimentu s využitím hmotnostního spektrometru, jsou v podobě hmotnostních spekter zaznamenaných po chromatografické separaci peptidové směsi (obr. 1.4). MS nebo MS/MS spektrum je vyjádřeno závislostí intenzity iontů na poměru m/z hodnoty iontů analyzované látky. MS/MS spektra vznikají změřením fragmentových iontů, které vzniknou rozpadem prekurzorových iontů po jejich excitaci. Data nesou i další informace než hodnoty intenzit a poměrů m/z ze spekter, např.: charakteristiky eluce z kapalinového chromatografu. Tyto metadata jsou následně využívána v poloautomatických a automatických postupech pro zpracování a analýzu spektrálních dat [14]. MS spektra jsou na výstupu v různých datových formátech, a tedy nejednotné, protože formát výstupních dat závisí na dodavateli hmotnostního spektrometru. Pokud je to možné, pracuje se s primárním formátem výstupních dat, případně se data převádí na volně přístupný a standardizovaný formát .mzML.

Spektra jsou používány k identifikaci, charakterizaci nebo absolutní a relativní kvantifikaci peptidů a proteinů, ale i k určení hmotnosti intaktních proteinů či určení druhu a místa posttranslačních modifikací (např. fosforylace, acetylace, oxidace, glykosylace).

Zpracování a analýza dat a metadat z hmotnostního spektrometru má většinou pět základních kroků. První z nich je konverze nezpracovaných MS dat do otevřených xml formátů (např. do .mzML), se kterými umí pracovat analyzující software. Druhým krokem je identifikace spekter pomocí databázového prohledávání a třetí krok je validace získaných identifikací, čtvrtým krokem je kvantifikace, na což navazuje pátý krok, který obsahuje řešení *protein inference* - zpětná rekonstrukce peptidů na proteiny.

1.3.1 Identifikace peptidů

Identifikace peptidů a proteinů za pomocí hmotnostní spektrometrie se dělí na přístup *bottom up* a *top down*.

Top down přístup vynechává proteolytického štěpení proteinů a přímo proteinová směs je separována, proteiny jsou v hmotnostním spektrometru fragmentovány a v závěru jsou identifikovány za pomoci databázového prohledávání.

Bottom up přístup má dva standardní postupy. První přístup využívá specifického proteolytického štěpení proteinů a separace naštěpených peptidů za pomoci např. kapalinové chromatografie a následné nalezení MS/MS spekter peptidů. V závěru se identifikace proteinů provádí za pomoci databázového prohledávání, ve kterém se srovnávají fragmenty peptidů se sekvencemi v databázi. Přiřazením fragmentového spektra k peptidové sekvenci se generuje sada shod peptid - spektrum (PSM). Druhý přístup postupně využívá separace proteinů (gelovou elektroforézou nebo kapalinovou chromatografií), následně proteolytického štěpení, nalezení MS spekter peptidů a nakonec identifikace pomocí srovnání peptidových map s databázemi sekvencí - peptidové mapování. Peptidové mapování je umožněno na základě známých hmotností jednotlivých peptidů a získaných hmotností fragmentů z hmotnostního spektrometru. *Bottom up* metoda je aktuálně více populární, protože se vyskytuje menší absolutní chyba při měřených menších m/z hodnotách, ale i senzitivita hmotnostního spektrometru je mnohem větší na peptidové úrovni než proteinové. Také analýzu celých proteinů komplikuje přítomnost modifikací.

Peptidové mapování

Peptidové mapování funguje na základě shody experimentálních získaných hodnot hmotností peptidů z MS spektra (tedy nedělají se MS/MS fragmentová spektra peptidů) specificky štěpených proteázou tvořících peptidovou mapu s hodnotami hmotností teoretických peptidových map. Teoretické hodnoty se získají in-silico štěpením peptidových sekvencí podle určité proteázy. Experimentální a teoretické hodnoty

jsou přiřazeny k sobě s určitou hodnotou nejistoty a výsledkem prohledávání je žebříček proteinů s nejpodobnějšími peptidovými mapami a jejich skóre. Tato metoda je v dnešní době minimálně používána.

De novo sekvenování

De novo sekvenování je metoda pro identifikování spekter, která je používána v případě chybějících či velmi limitovaných informací v databázích. Často je tedy využívána k obtížným identifikacím a charakterizování peptidových modifikací a proteinových polymorfismů. V této metodě jsou peptidy nebo proteiny interpretovány pomocí vzájemné vzdálenosti píků ze spekter fragmentovaných iontů. K identifikovaným sekvencím se mohou následně vyhledávat podobné sekvence v databázi. K tomu slouží nástroj, který používá vyhledávání pomocí srovnání zarovnaných sekvencí (např. BLAST).

MS/MS databázové prohledávání

Databázové prohledávání MS/MS spekter je založeno na porovnávání naměřených spekter fragmentovaných peptidů (fragmentačních map) a teoretických fragmentačních map odvozených z databáze (např. UniProt či NCBI) známých proteinových sekvencí, které by měly nebo mohly být v analyzovaném vzorku. Kromě předpokládaných proteinů by měla databáze obsahovat proteiny pocházející z možného zdroje kontaminace (trypsin, keratin, BSA). Fragmentační mapy jsou odvozeny z databáze podle použitého enzymu pro štěpení proteinů.

Pro každou shodu spekter je počítáno PSM skóre, které vyjadřuje podobnost spekter. PSM skóre vyjadřuje kolik iontů z teoretického spektra je přítomno v naměřeném spektru. Na výpočet podobnosti naměřených a teoretických spekter jsou používány softwarové nástroje, které jsou komerční i volně přístupné (např: Sequest, Mascot, X!Tandem, MyriMatch, MS-GF+). [15]

Pro shody spekter se vrací seznam PSM seřazených podle jejich skóre. Pravděpodobnost nalezení špatné shody spekter se vyjařuje pomocí FDR (*false discovery rate*). FDR je v kontextu peptidových identifikací podíl falešně pozitivních PSM ze všech PSM nad určitou hranicí. Zjištění pravděpodobnosti špatných shod a všech PSM se zajišťuje pomocí *target - decoy* metody. [16]

Decoy databáze je vytvořena pozměněním originálních proteinových (*target*) sekvencí (např. reverzí, pseudo-reverzí, randomizací). Tyto sekvence by tedy neměly být přítomny ve vzorku, a pokud jsou nalezeny měly by být považovány za falešně pozitivní identifikace. *Target-decoy* metoda by měla poskytnout vyrovnaný poměr *target-decoy* peptidů a ideálně by neměly být velké překryvy mezi *target* a *decoy* peptidy. Špatné PSM mají většinou nižší skóre a správné PSM vyšší skóre. Avšak

je nutné najít ideální hranici, od které jsou PSM uznány za správné a je následně možné vypočítat FDR. FDR o hodnotě 0,01 udává, že 99 % PSM bude správně.

V souvislosti s FDR se používá q-hodnota, která může být považována za minimální FDR při, kterém jsou PSM akceptovány. Q-hodnota se tedy zvýší pokaždé, když se zvýší počet falešně pozitivních výsledků.

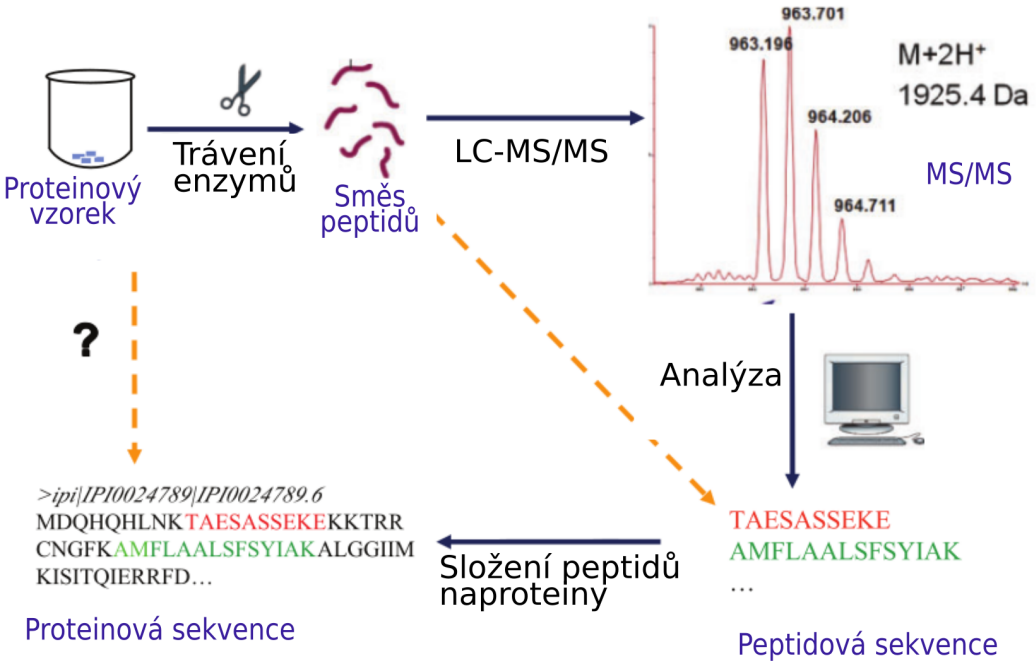
1.3.2 Zpětné sestavení seznamu proteinů

Identifikace proteinů, tedy zpětné skládání proteinů z identifikovaných peptidů – PSM ve vzorku, které jsou seřazeny s daným skóre, se nazývá *protein inference problem* (viz obr. 1.5). V tomto kroku se mimo jiné řeší, že identifikovaný peptid může být sdílen mezi více proteiny z pohledu použité proteinové databáze, nebo z pohledu reportovaného seznamu proteinů. V tomto kontextu se hovoří o tzv. proteotypických peptidech, tedy peptidech, které jsou unikátní pro daný protein v rámci použité proteinové databáze. Druhým typem jsou pak peptidy unikátní pouze v rámci sady reportovaných proteinů, které nemusí být nutně zároveň i proteotypické. V dnešní době je častější náhled na "unikátnost"daného peptidu s ohledem pouze na reportovanou sadu proteinů. Unikátní proteiny jsou pravděpodobněji nalezeny pro delší peptidové sekvence.

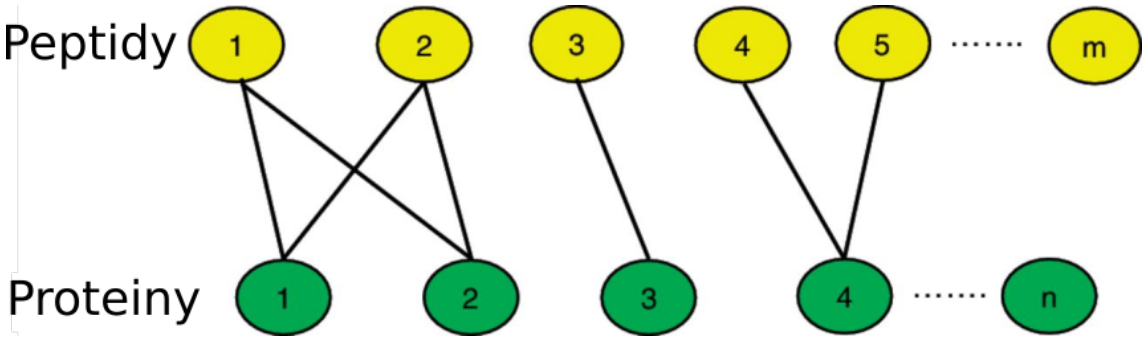
Jako sdílené peptidy jsou nejčastěji míněny ty peptidy, které jsou společné pro dva a více reportovaných proteinů. Tyto sdílené peptidy, také nazývané degenerované sdílené peptidy, vedou často k sadě proteinů nazývané nejednoznačná skupina proteinů. Původem těchto neunikátních proteinů jsou náhodné shody PSM, izoformy proteinů či konzervované regiony sdílené v proteinové rodině. Tyto peptidy mohou být zařazeny do skupin podle možnosti odlišitelnosti proteinů. Proteiny, které jsou odlišitelné, nesdílí spolu žádné peptidy. Naopak rozdílné proteiny musí být rozlišitelné od sebe aspoň jedním peptidem a nerozlišitelné proteiny spolu sdílí všechny peptidy. Podmnožina proteinu obsahuje jen peptidy, které jsou v druhém proteinu. Poslední z možností jsou proteiny, které obsahují jen peptidy, které jsou součástí jiných proteinů. Často není možné určit, který protein je doopravdy ve vzorku, dokud není nalezen unikátní peptid, který se nachází jen v jednom proteinu. [17]

Častý model pro řešení tohoto problému je biparitní graf, který je znázorněn na obr. 1.6. V grafu první a druhý protein má stejnou skupinu přiřazených identifikovaných peptidů (první a druhý peptid), které jsou degenerované a nejde u nich bez žádné další informace rozlišit, který z proteinů se nachází ve vzorku. Naopak třetí protein má přiřazený jen jeden identifikovaný peptid a není snadné určit jeho existenci oproti čtvrtému proteinu, který má dva identifikované peptidy. [18]

Tento problém řeší různé algoritmy (např. Epifany, Fido, Percolator, PIA). Algoritmy mají kombinatorické nebo pravděpodobnostní přístupy. Jeden ze způsobů



Obr. 1.5: Znázornění postupu získání proteinů z peptidových identifikací. Proteiny ve vzorku jsou neznámé a pro jejich identifikaci a určení množství, je použit experiment s využitím hmotnostní spektrometrie. Zpracováním získaných dat z hmotnostního spektrometru jsou získány identifikace peptidů, které jsou zpětně sestaveny na proteiny. Převzato z: [18].



Obr. 1.6: Biparitní graf používaný k sestavení peptidů na proteiny. Vrchní řada vrcholů představuje identifikované peptidy a spodní řada kandidáty proteinů, které se mohou nalézat ve vzorku (proteiny alespoň s jedním přiřazeným peptidem). Převzato z: [18].

používá princip parsimonie ke sestavení peptidů na protein s využitím biparitního grafu. Princip parsimonie uplatňuje teorii Occamovy břitvy, podle které by měla být skupina proteinů minimální (nejmenší možné) velikosti, která pokryje veškeré identifikované peptidy. Pokud tedy všechny peptidy patřící do jedné proteinové rodiny mohou být obsaženy v jednom proteinu, tak je velmi pravděpodobné, že tento protein je přítomný ve vzorku. K nalezení řešení je často použit hladový algoritmus, který iterativně prochází vrcholy proteinů v biparitním grafu a hledá podskupinu proteinů, které pokrývají všechny vrcholy peptidů v grafu (obr. 1.6). Když jsou všechny peptidy pokryty, nevyužité proteiny jsou eliminovány.

1.3.3 Kvantifikace proteinů

Kvantifikace proteinů slouží ke zjištění absolutního nebo relativního množství proteinů ve vzorku. Absolutní kvantifikace zjišťuje přesné množství proteinu ve vzorku a slouží k tomu např. metoda AQUA používající izotopově značené peptidy kvantifikovaných proteinů. Relativní kvantifikace umožňuje sledovat snížení či zvýšení exprese proteinu v porovnávaných vzorcích. Kvantifikační metody se dělí na metody využívající značení (metabolické, chemické, enzymatické) a metody bez značení – *label-free*.

Metody bez značení vychází z velkého množství kvantitativních dat, kde je zaznamenáno až desítky spekter pro jednu sekundu. Spektra jsou často složeny do LC-MS map. Pomocí metod bez značení lze zjistit absolutní kvantifikaci například při využití metody APEX – metoda absolutní proteinová exprese nebo metody Detekce LCMS píků – metoda iterativně v dvoudimenzionálním prostoru (hodnoty m/z, retenční čas) hledá LC-MS píky (peptidy, které mají charakteristickou isotopovou distribuci). Tyto techniky poskytují odhad absolutního množství bez nutnosti

použití značených peptidů, ale většinou poskytují méně přesné a správné výsledky. K relativním metodám kvantifikace bez značení se řadí *spectral counting* přístup, který sleduje počty MS/MS spekter proteinu v pozorovaných vzorcích. Tento parametr by měl lineárně odpovídat relativnímu množství proteinu ve vzorku. Metoda tedy předpokládá, že při větším množství proteinu ve vzorku se vyskytuje i více spekter identifikovaných k danému proteinu. Tato metoda je však často nepřesná a ne vždy počet spekter lineárně koreluje k množství peptidů ve vzorku. Metodu TOP3 je možné použít pro absolutní i relativní kvantifikaci. Metoda ke kvantifikaci používá MS signál tří tryptických peptidů s největší intenzitou. [19]

1.4 Metriky kontroly kvality

Pro sledování průběhu experimentů se používají metriky pro kontrolu kvality. Metriky umožňují kvantifikovat variabilitu ve výsledcích experimentu a časové trendy. Experiment může být podle nich následně optimalizován, případně může být odhalen začínající technický problém. Metriky jsou zaměřeny hlavně na hodnocení spolehlivosti kapalinové chromatografie a hmotnostní spektrometrie. Metriky jsou extrahovány z hrubých dat a je možné je ukládat v standardním formátu .qcML, avšak který je jeho tvůrci postupně nahrazován novým formátem .mzQC. [20]

Metriky kontroly kvality se dělí na intra-experimentální a inter-experimentální. Intra-experimentální metriky jsou získány z jednoho průběhu experimentu, a tedy jsou na úrovni jednotlivých skenů a identifikací (např. chromatogram celkového iontového proudu závislého na retenční čase, přesnost hmotnosti identifikovaných spekter). Inter-experimentální metriky posuzují kvalitu měření přístroje v dlouhodobém časovém horizontu. Porovnávají tedy hodnoty z více měření navzájem. K tomu je nutné hodnoty vhodně uchovávat – například v databázi.

Další rozdělení metrik vyplývá z fází zpracování hrubých dat, ve kterých jsou metriky získány. Můžeme je rozdělit do tří kategorií – instrumentální, metriky s identifikací (ID metriky) a získané bez identifikací (*ID-free* metriky). Hodnoty jsou získávány a počítány ze spekter kromě instrumentálních metrik.

Metriky bez ID jsou výhradně ze spekter a zachytávají informace z celého průběhu LC-MS (např. tvar píků vnitřních standardů, průběh TIC popisující chromatografii, počet MS a MS/MS spekter, *scan rate* – popisující získání spekter, distribuce náboje změřených MS/MS spekter popisující ionizaci).

Metriky založené na ID jsou získány ze spekter a následující identifikace a kvantifikace. Do těchto metrik jsou řazeny např. počet identifikací v podobě počtu PSM, peptidů a proteinů, které jsou ovlivněny procesními parametry identifikačního prohledávání. Dalšími metrikami jsou například pokrytí sekvence známého vzorku,

který obsahuje jen jeden protein nebo metriky porovnávající retenční časy stejných peptidů, které slouží pro zjištění chromatografické stability.

Instrumentální metriky jsou získány přímo z přístroje a ne z měřených spekter. Většinou se jedná o velmi citlivé metriky týkající se např. stavu iontového zdroje nebo vakua. Tyto metriky jsou vhodné ke kontrole přístroje a plánování údržby, ale nemůžou být přímo spojovány s výsledky experimentu a většinou nejsou součástí .mzML souboru. Velmi se liší mezi jednotlivými přístroji a dodavateli přístrojů. [21]

2 Zpracování proteomických dat v KNIME

Software KNIME je velmi flexibilní v možnostech způsobů práce s daty, protože podporuje instalaci programovacích jazyků (R, Python), ale i rozšíření pro některé programy. Jeden z velmi propracovaných programů je OpenMS, který je volně přístupný víceplatformní a speciálně navržený pro tvorbu reprodukovatelné analýzy MS dat. Nabízí nástroje vytvořené v jazyce C++ a Python pro klasické zpracování hmotnostně spektrometrických dat s využitím volně dostupnými standardizovanými datovými formáty. [22]

2.1 OpenMS

OpenMS pokrývá celou škálu požadavků na vytvoření pracovního postupu zpracování MS dat, které jsou velmi variabilní a vyžadují různé přístupy analýzy. Vliv na tyto požadavky má v prvé řadě typ dat (např. proteomická, metabolomická, lipidomická data), další vliv mají separační a fragmentační metody, akvizice dat (DDA, DIA). OpenMS pokrývá různé kvantifikační metody (bez značení, isobarické a isotopické značení), ale má i širokou nabídku databázového vyhledávání k identifikaci peptidů, kde software nabízí řadu vyhledávačů (Mascot, MS-GF+ adapter, Myrimatch, OMSSA, X!Tandem).

Knihovna OpenMS je koncipována tak, aby jednotlivé algoritmy či nástroje mohly být seskládány do komplexního pracovního postupu, což usnadňuje rozsáhlá dokumentace na stránkách OpenMS. Dokumentace kromě popisu funkcionality obsahuje popis vstupních proměnných, ale i kompatibility algoritmů mezi sebou. Vizuálně lze postup z OpenMS algoritmů sestavit v grafických programech KNIME, Galaxy a dalších. V KNIME je možnost importu OpenMS funkcí a každý algoritmus je k dispozici v podobě jedné pracovní jednotky - nodu. Nody jsou propojitelné a často je u nich možnost nastavení vstupních parametrů i výstupních datových formátů. OpenMS poskytuje propracovanou dokumentaci jednotlivých funkcí, možnosti nastavení vstupních parametrů, ale i doporučení použití předcházajících a následujících funkcí.

OpenMS je ideální pro kontrolu kvality dat, protože lze získat nespočet metrik (např. počet spekter, peptidů a proteinů, statistiky přesnosti hmotnosti, rozsahu retenčních časů a poměru hmotnosti k náboji). Lze vygenerovat grafy nebo podkladová data pro grafy (např. TIC chromatogram či histogram rozdělení náboje detekovaných iontů). Informace o metrikách mohou být exportovány v souboru .qcML či v novějším formátu .mzQC, avšak pro tento novější formát nejsou prozatím přizpůsobeny další algoritmy a grafy z tohoto formátu nelze generovat. Soubory je

možné zobrazit a prohlížet ve webovém prohlížeči nebo je exportovat v PDF. Využití OpenMS v KNIME má velké výhody v možnosti rozšíření jeho funkcionality vlastními vytvořenými nody za pomoci některého ze skriptovacích jazyků dostupných v KNIME jako Java, R či Python. Vytvořený pracovní postup z nodů je možný jednoduše ukládat a sdílet.

2.2 Balíčky skriptovacích jazyků Python a R

Nástroje v KNIME podporují práci se sktriptovacími jazyky, a proto je možné využít jazyky R či Python k vytvoření workflow zpracování MS dat a kontrolu kvality. Python obsahuje několik knihoven, které se zaměřují na proteomiku a hmotnostní spektrometrii. Mezi ně patří např. Spectrum utils [23], pymzML [24], Pyteomics [25] či pyOpenMS, který je rozšiřujícím nástrojem pro práci se nízkoúrovňovými funkcemi OpenMS. Umožňuje vytváření nových algoritmů pro práci s MS daty. Je to balíček jazyka Python, díky kterému je umožněno propojit funkcionalitu pyOpenMS i s dalšími bioinformatickými balíčky a snadno data zobrazit.

Mezi knihovny jazyka R pro práci s MS daty patří např. MSnbase [26], MSstats [27], MALDIquant [28], MSstatsQC [29], pmartR [30] nebo baliček proteoQC [31], který umožňuje vygenerovat HTML report o kvalitě experimentu.

2.3 Datové formáty hmotnostně spektrometrických dat v KNIME

Proteomická data je možné zpracovávat v KNIME s využitím výše zmíněné knihovny OpenMS, která umožňuje práci s daty v některých standardních formátech definovaných PSI (*Proteomics Standard Initiative* – viz https://www.hupo.org/Proteomics-Standards-Initiative). PSI je součástí *Human Proteome Organization*, jejíž cílem je definovat proteomické standardy pro usnadnění porovnání, výměny a ověřování dat.

2.3.1 PSI formátové standardy

Standardní formáty definované PSI jsou založené na XML a liší se dle využití (např. skladování MS dat, identifikací peptidů nebo proteinů a kvantifikací). Tomu odpovídají formáty .mzML, .mzIndentML, .mzQuantML, ale jsou i formáty .mzTab (ukládání finálních výsledků), .TraML (pro monitorování vybraných reakcí), .GelML (pro ukládání gelů) nebo .mzQC (pro ukládání metrik kvality – *Quality control*, QC). Mezi výhody těchto formátů patří jejich otevřenost, dokumentace, čitelnost a interoperabilita mezi softwary. Avšak mají i své nevýhody, např. častá nutnost počáteční

konverze hrubých dat přímo uloženým hmotnostním spektrometrem z formátů, které se liší podle výrobce hmotnostních spektrometrů.

Formát .mzML

Formát .mzML je založený na otevřenosti, vychází z XML a je určený k uchovávání výstupních dat z hmotnostního spektrometru. Formát je možné následně využít k archivování, sdílení a zpracování dat. Formát má jasně definovanou strukturu, hodnoty jsou uchovávány pomocí kontrolovaného slovníku pro termíny a definice hmotnostně spektrometrických metadat.

Formát historicky vychází z formátů mzData a mzXML. které byly vyvinuty HUPO PSI a ISB (*Institute for System Biology*) v letech 2003-2005. Následně v roce 2009 vznikl formát .mzML 1.0.0, který předchozí dva nahradil. Hlavním cílem bylo sjednotit formáty, udržet jejich flexibilitu, ale i schéma a využití kontrolovaného slovníku. V návaznosti na .mzML se udržuje kontrolovaný slovník, sémantický validátor a dokumentace.

Formát je navržený tak, aby obsahoval informace z jednoho MS běhu, tedy metadata o spektrech a spektra samotná. V hlavičce nese element *cvList*, který obsahuje informace o kontrolovaném slovníku, který je použit ve zbytku souboru. Hlavička obsahuje další elementy – *fileDescription* (upřesňujicí základní informace o spektrech), *referenceParamGroupList* (obsahující seznam skupin termínů z kontrolovaného slovníku), *sampleList* (nesoucí informaci o vzorku), *instrumentConfiguration* (obsahující informace o přístrojích použitých k měření), *softwareList*, *dataProccessingList* (poskytující historii zpracování dat v RAW formátu), *acquisitionSettingList* (informující o speciálních vstupních parametrech hmotnostního spektrometru). V další části jsou už samotná spektra a za nimi můžou být zaznamenány chromatogramy. [14]

Většina metadat je uložena v elementu *cvParam*, který poskytuje referenci k termínu PSI MS kontrolovaném slovníku. Každý termín má detailní definici a může mít definovaný datový typ nebo jednotku. Kontrolovaný slovník je v OBO formátu, což je otevřený formát pro ukládání ontologií.

Formát .mzQC

Formát .mzQC je vytvářen a podporován HUPO PSI organizací. Tomuto formátu se přímo věnuje *The Quality Control Working Group*, která vytváří standardy a doporučení, jak popisovat hmotnostně spektrometrická data a jejich analytické výsledky.

Historicky formátu předchází .qcML formát [20], který byl založený na XML, avšak .mzQC je definován jako JSON (*JavaScript Object Notation*) syntaxí z důvodů malých paměťových nároků a má širokou podporu v programovacích jazycích.

Účelem .mzQC formátu je reportovat metriky vypočítané QC nástroji (Quameter, PTX-QC, QCloud), umožnit reportovat kvalitu pro posouzení funkčnosti nástrojů, poskytovat metriky jako vstup do vizualizačních a reportovacíh nástrojů, zaznamenávat dlouhodobé QC metriky k monitorování stárnutí nástrojů v čase či kontrolovat kvalitu sady MS experimentů skrz jednotlivé běhy, různé biologické nebo technické podmínky, studie a laboratoře.

Standardní formát .mzQC slouží k výměně, přesunu a archivování metrik kvality generovaných hmotnostní spektrometrií. Tento formát používá část (QC CV) kontrolovaného slovníku (PSI-MS QC CV) pro reprezentování pojmů definující metriky kvality a hodnoty spjaté s metrikami. QC CV hodnoty jsou reprezentovány kódy *MS:4000000 – MS:4999999* a kontrolovaný slovník je definován jako OBO soubor. Každá metrika obsahuje jméno, definici, typ hodnoty (hodnota, seznam, tabulka, matice) a může obsahovat komentář, jednotku nebo kategorizaci. Nové termíny mohou být přidány do kontrolovaného slovníku po podání žádosti na stránkách PSI-MS CV Githubu, kde je připraven modul, který provede uživatele v podání všech potřebných informací. Pro ukládání obrázků do .mzQC formátu je doporučeno použít vlastní metriku s dostatečným popisem a hodnotu, tedy obrázek, vložit jako *base64* kódovaný řetězec.

Soubor .mzQC vždy obsahuje několik objektů (verzi, datum vzniku .mzQC souboru, kontakt, popis, pole metrik – *runQuality*), nastavení metrik – *setQuality* a informaci o kontrolovaném slovníku) viz obr. 2.1. Pole a nastavení metrik obsahují vždy elementy *metadata* a *qualityMetrics* pro jeden běh. *QualityMetric* má strukturu základního elementu .mzQC formátu *cvParameter*, který obsahuje identifikátor, jméno, popis, hodnotu a jednotku.

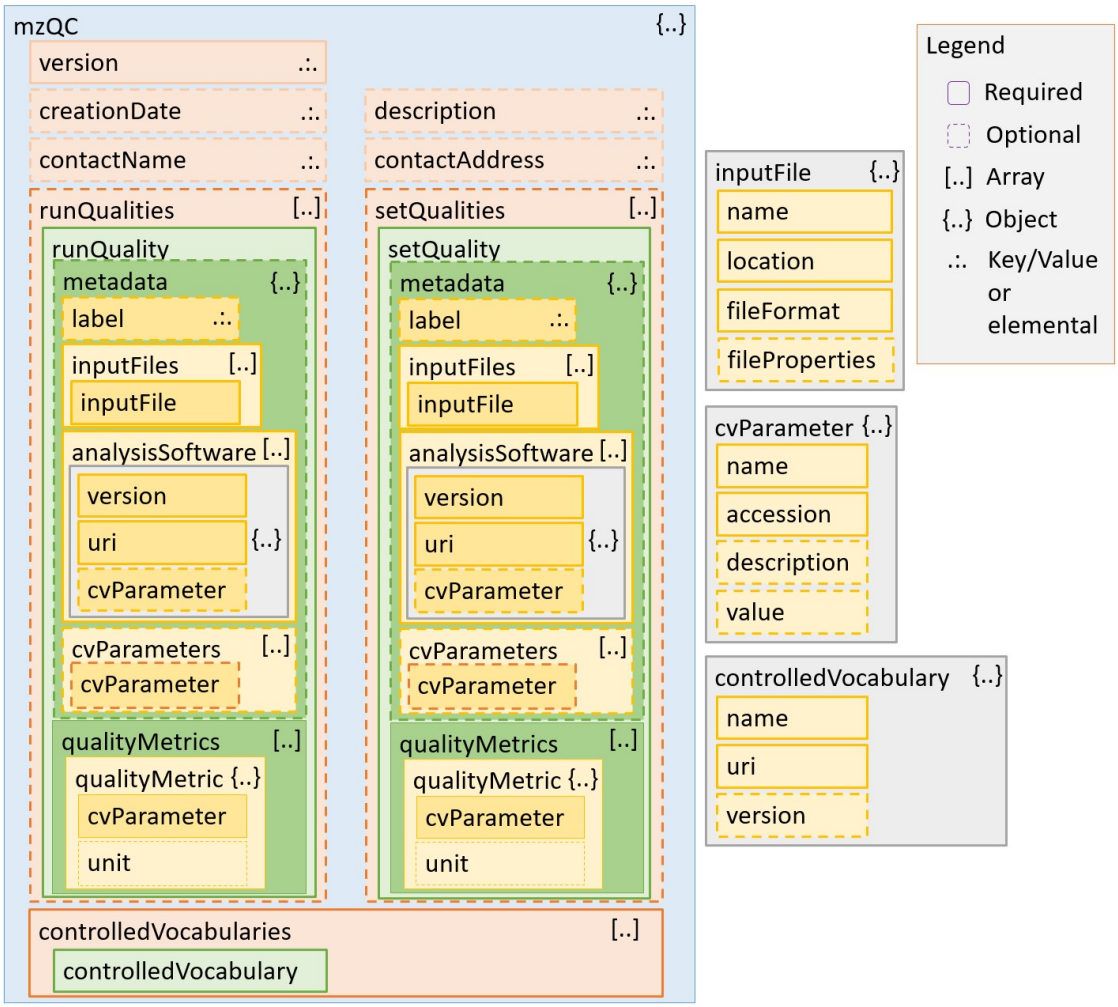
Pro manipulaci s .mzQC formátem existuje knihovna *mzqc-pylib*, avšak soubor tohoto formátu lze vygenerovat OpenMS nástrojem *QCCalculator*.

2.3.2 OpenMS formáty

Software OpenMS definuje své formáty, které jsou používány moduly softwaru a liší se podle využití. Mezi tyto formáty patří .featureXML, .idXML, .consensusXML.

Formát .idXML

Formát .idXML je další z formátů definovaný a používaný v OpenMS. Je to formát XML, který obsahuje informace o peptidových a proteinových identifikacích z proteinového databázového prohledávání a je využívaný v hmotnostně spektrometrické analýze dat. Soubor může nést informace o identifikacích z více běhů. Alternativní variantou mimo OpenMS .idXML je .pepXML formát, který se používá v *Trans-Proteomic Pipeline* [32], nebo .mzIndentML standardní formát PSI. Pro konverzi



Obr. 2.1: Schéma .mzQC souboru, který slouží pro ukládání metrik pro kontrolu kvality. Z tohoto schématu vychází skript pro získání metrik a jejich uložení do databáze v rámci workflow pro automatickou kontrolu kvality dat. Názvy ve schématu jsou reálně použité názvy v .mzQC formátu. Schéma je převzato z github stránek HUPO-PSI organizace.

do OpenMS formátu .idXML je možné použít nástroj *IDFileConverter*. S formátem dále mohou pracovat OpenMS nástroje *IDFilter*, *PeptideIndexer*, *IDPosteriorErrorProbability*, *IDMerger*, *ConsensusID* a další. [22]

Formát .featureXML

OpenMS ukládá data o LC-MS pících, které mohou být anotovány peptidovými identifikacemi, do formátu .featureXML. LC-MS pík je detekovaný dvou či více rozměrný vzor, které leží v rovině, která je definována osami m/z a retenčním časem. LC-MS pík představuje chromatografický profil v časové rovině a isotopový vzor v rovině m/z. Do tohoto formátu lze z tabulkové podoby konvertovat za pomoci OpenMS nástroje *FileConverter* a zpětnou konverzi umožňuje *TextExporter*. Mezi další nástroje, které podporují a používají formát .featureXML, patří např. *FeatureFinder*, *IDmapper*, *IDConflictResolver*.

Formát .consensusXML

Formát .consensusXML je používaný pro ukládání konsensuálních map vytvořených z více vstupních souborů. Konsensuální mapy vznikají porovnáním a sjednocením informací z více běhů, tedy koriguje se osa retenčního času, která může být mezi běhy posunutá nebo škálovaná. Formát je také používán k uchování informací z vícenásobné peptidové identifikace databázovým prohledáváním a výsledné konsensuální identifikace.

OpenMS používá tento formát například s nástroji *MapAlignerPoseClustering*, *ConsensusMapNormalizer*, *FeatureLinkerUnlabeled* nebo *ProteinQuantifier*.

3 Návrh metodiky automatizované kontroly kvality hmotnostně spektrometrických dat

Automatizovaná kontrola kvality hmotnostně spektrometrických dat by měla být v programu KNIME v podobě workflow, tedy neměnného sledu pracovních jednotek (nodů a metanodů), které jasně definovaným způsobem pracují s daty. Každý nod má svůj definovaný vstup, výstup a svoji funkcionalitu.

Samotným vstupem do workflow budou raw data ve formátech, které jsou závislé na typu hmotnostního spektrometru, kterým jsou při měření generovány. Tyto data bude nutné převádět do formátu .mzML, protože to je vstupní formát dat pro nástroje OpenMS. Formát .mzML je standardizovaný, volně přístupný a používaný v hmotnostní spektrometrii. Vstupní soubory budou postupně načítány z určené složky, která bude samotným workflow sledována, zda v ní nepřibyly nové nezpracované soubory, čímž bude zajištěna automatičnost nahrání dat do workflow.

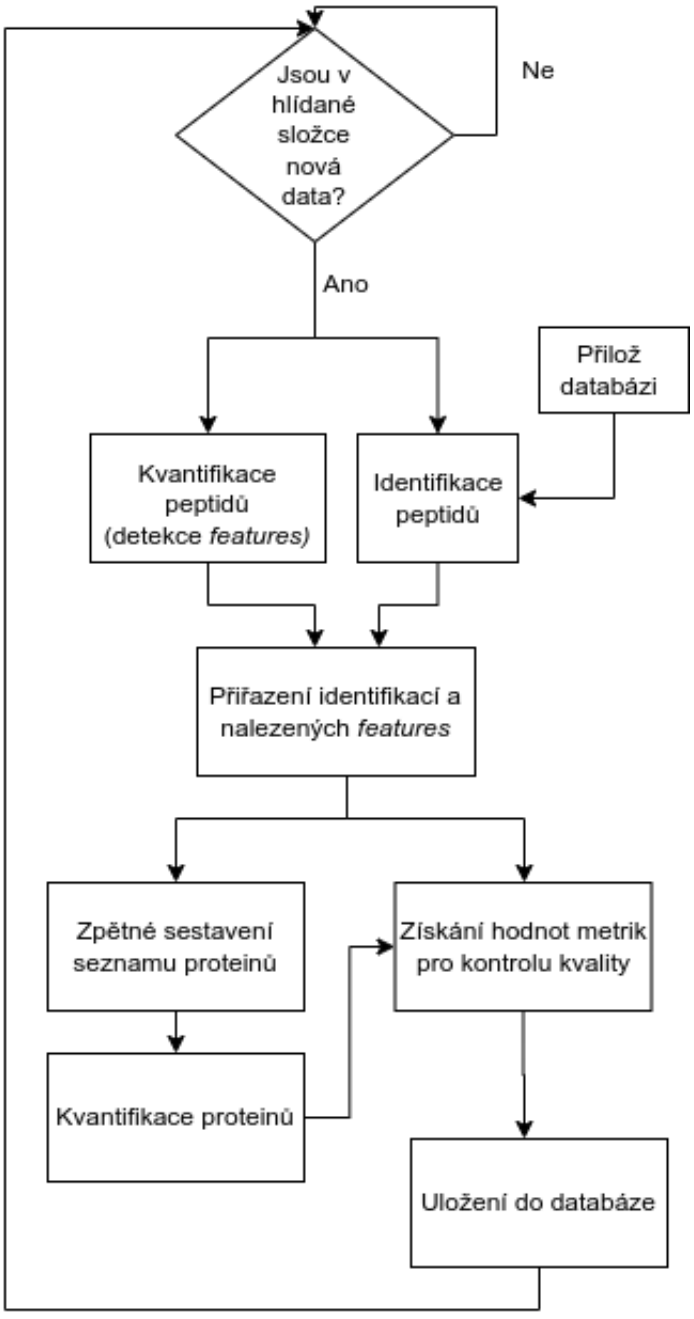
Workflow bude z velké části využívat funkcionalit OpenMS, které jsou k dispozici v podobě pracovních jednotek. Workflow automaticky najde pík v LC-MS mapě, identifikuje a kvantifikuje je a následně bude vytvořen soubor formátu .mzQC, který shrnuje informace - metriky pro kontrolu kvality dat. Z tohoto souboru budou extrahované vybrané parametry pro kontrolu kvality dat a uloženy do databáze (viz obr. 3.1).

Pro identifikování peptidů bude vybrán jeden z dostupných nástrojů pro prohledávání databáze s ohledem na rychlost a kvalitu identifikace (např. MSGFPlusAdapter). Databáze proteinů pro vyhledávání bude poskytnuta v podobě dedikovaného a neměnného FASTA souboru, ze kterého budou vytvořeny i *decoy* proteiny, které budou součástí databáze pro identifikace. Vytváření decoy proteinů je zařazeno pro zjištění hodnoty FDR (*false discovery rate* – poměr falešně pozitivních výsledků).

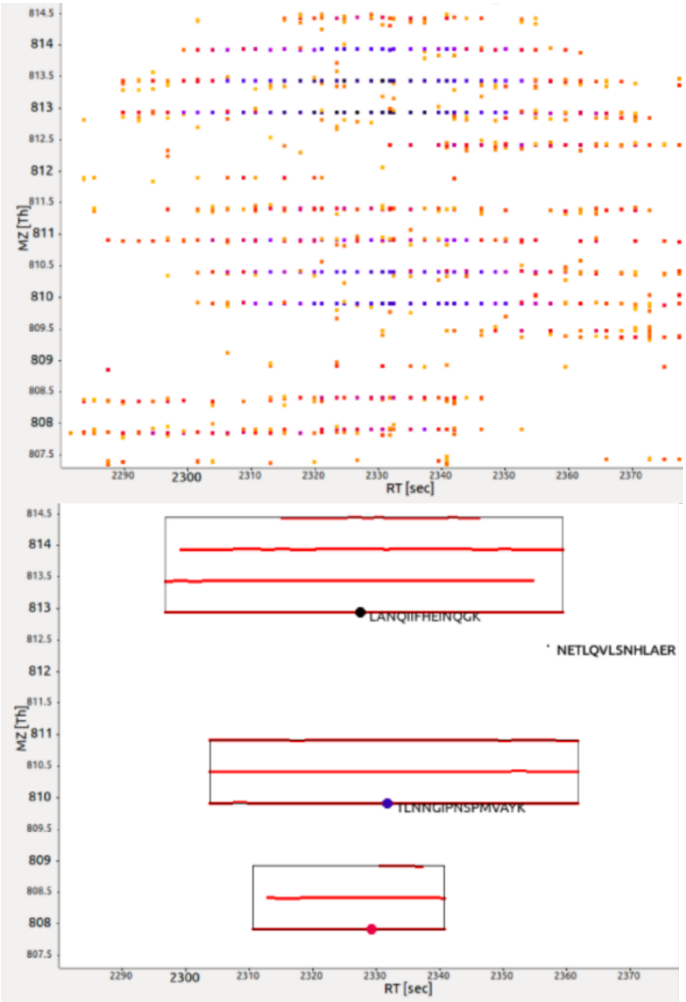
Nezávisle na identifikacích proběhne detekce a kvantifikace LC-MS signálů (píků, anglicky *features*) všech analytů včetně peptidů pomocí nástroje OpenMS. Algoritmus pro detekci signálů hledá pozice všech analytů ve třírozměrném poli, jehož dimenze jsou retenční čas, m/z hodnota a intenzita. Vyhledávání probíhá iterativně, dokud nenajde model isotopového profilu a retenčního času, který je složen z bodů.

Informace z identifikace a kvantifikace budou následně propojeny. K nalezeným a kvantifikovaným signálům budou tedy přiřazeny identifikace (obr. 3.2).

V další části bude sestaven proteinový seznam pomocí pracovních jednotek věnujících se inferenci proteinů a jejich výstupem by měl být seznam identifikovaných proteinů. Na to by měla navazovat pracovní jednotka kvantifikující proteiny, jejíž výstupem bude množství daných proteinů ve vzorku.



Obr. 3.1: Diagram zjednodušeného návrhu automatického zpracování hmotnostně spektrometrických dat a získání metrik pro kontrolu kvality a jejich ukládání do databáze.



Obr. 3.2: Ukázka spojení hodnot identifikovaných a kvantifikovaných peptidů. Nahoře je snímek LC-MS mapy, kde na ose x jsou hodnoty retenčního času a na ose y hodnoty m/z. Dole je stejná pozice v mapě, avšak jsou už nalezeny LC-MS píky, které jsou anotovány přiřazenými identifikacemi. Převzato z: [33].

Výstupní data budou v přehledném formátu prohlídnutelná a průběžně ukládána do databáze. Do databáze se bude rovněž ukládat informace o vstupních datech a procesním workflow pro reprodukovatelnost a jednoznačné určení způsobu, jakým výstupní data vznikla a ostatní soubory (např. grafy či soubor .mzQC shrnující metriky kontroly kvality).

4 Workflow pro automatizovanou kontrolu hmotnostně spektrometrických dat

V rámci práce bylo vytvořeno workflow, které slouží k automatizované kontrole hmotnostně spektrometrických dat – převážně kontrolních (QC) vzorků pro kontrolu stavu LC-MS systému v laboratoři. Cílem je automaticky získávat metriky pro kontrolu kvality dat z kontrolních standardních vzorků a metriky ukládat tak, aby bylo možné kontrolovat jednotlivé běhy a jejich kvalitu, ale také je porovnávat mezi sebou a sledovat dlouhodobě kvalitu hmotnostně spektrometrických dat.

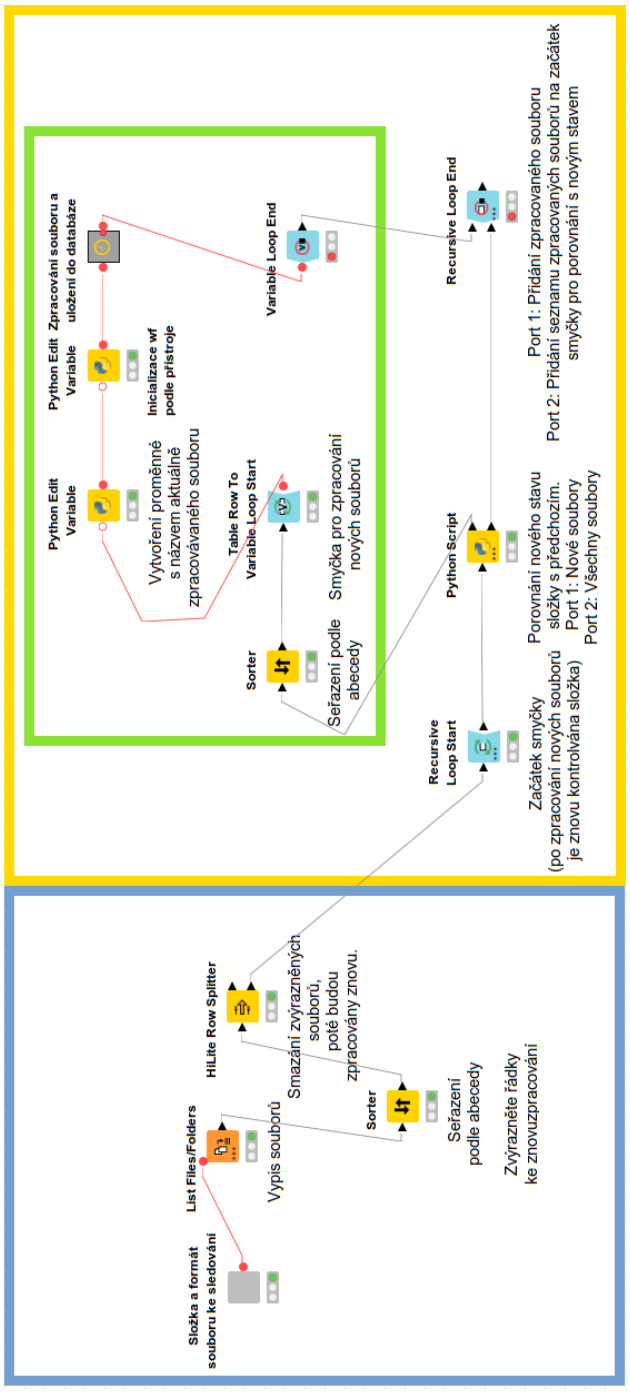
Workflow je navrženo tak, aby mohlo být spuštěno kdekoli, kde je nainstalován software KNIME a jeho závislosti. Workflow má přístup ke složce, do které jsou ukládána data z hmotnostního spektrometru a sleduje stav souborů, které má zpracovat.

Workflow bylo vytvořeno v softwaru KNIME s využitím importovaných funkcí knihovny OpenMS, skriptovacích jazyků Python a R, ale i jazyka SQL pro vytvoření a práci s databází SQLite. Předpokladem pro práci s workflow je nutné umět pracovat s KNIME, alespoň pro základní nastavení workflow a jeho spuštění.

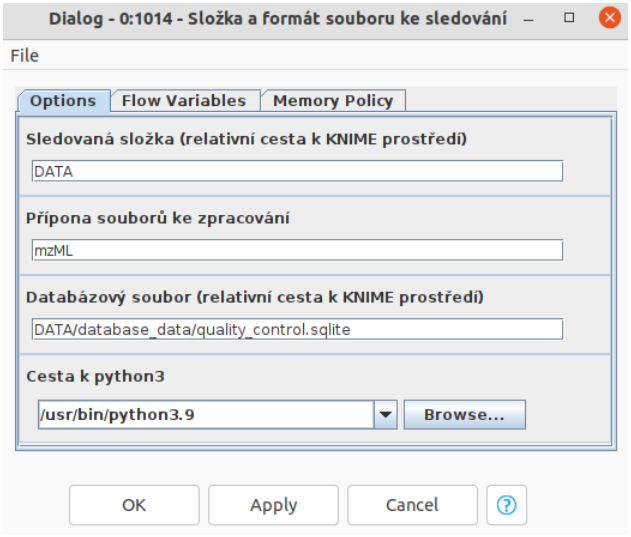
Workflow je sestaveno z nodů, které mají vlastní funkcionalitu a předávají si získaná data pro další zpracování nebo uložení. V KNIME jsou dvě možnosti předání dat v rámci workflow. Jednou možností je uložení informace do proměnné – *flow_variable*, které KNIME shrnuje do jedné proměnné, která odpovídá slovníku všech proměnných (*flow_variables*) v Pythonu. Slovník je předán při jakémkoli propojení dvou nodů (růžový spoj značí přímo předání slovníku, šedé spoje předávají výstupní data i se slovníkem). Druhá možnost předání dat je přes šedé spoje, které předávají výstupní data o určitém formátu (např. tabulkový) dalšímu nodu, který přijímá na vstupu stejný formát dat (viz obr. 4.1).

4.1 Části workflow

Workflow podle návrhu obsahuje části zajišťující automatické načítání nových dat, ale i možnost znovu zpracování starších dat (obr. 4.1). Po načítání dat následuje část, která se liší podle přístroje, na kterém probíhalo měření. V této části probíhá identifikace a kvantifikace peptidů nalezením LC-MS píků. Identifikace jsou mapovány k píkům, následně jsou hodnoty filtrovány pomocí FDR. Dalším krokem je proteinová úroveň. Z obou úrovní jsou získány metriky pro kontrolu kvality dat a v poslední části jsou uloženy do databáze.



Obr. 4.1: Workflow pro automatickou kontrolu kvality dat. Modrá část workflow obsahuje nody, které slouží k nastavení cesty ke sledované složce a umožnění zásahu do práce workflow označením souborů, které mají být zpracovány. Žlutá část workflow obsahuje smyčku, která kontroluje složku a ve chvíli, kdy se objeví nové soubory ve složce, spustí se zelená část workflow. V zelené části je smyčka, která postupně zpracovává nové soubory (samotné získání metrik a zapsání do databáze). Ve chvíli, kdy jsou všechny soubory ze seznamu nových souborů zpracovány, vrátí se běh workflow do žluté části.



Obr. 4.2: Vstupní formulář, který slouží k definování cesty sledované složky, přípony souborů ke zpracování, cestu k databázi a Pythonu.

4.1.1 Automatické načítání dat

Workflow, které v jádru zpracovává proteomická data, je zabaleno do rekurzivních smyček pro automatické načítání dat, kterým předchází několik nodů sloužících k nastavení základních informací a opětovné zpracování souborů.

Vstupní formulář k workflow

Na začátku workflow pro automatickou kontrolu dat je komponenta, která umožňuje pomocí formuláře (obr. 4.2) definování hodnot, které jsou uloženy do proměnných, které jsou dále používány ve workflow (např. načítání dat, ukládání dat do souborového systému a databáze). Ve formuláři je tedy specifikována relativní cesta ke složce, do které jsou ukládána data určená ke zpracování. Další pole formuláře je pro definování souborové přípony, pro snadné filtrování souborů ve sledované složce. A v závěru jsou pole pro vložení cesty k databázi a cesta k Python prostředí, které je ve workflow používáno.

Znovu zpracování souborů

Hned po vstupní komponentě s formulářem jsou tři nody, které umožňují uživatelský zásah do automatické práce workflow. Jelikož workflow sleduje stav zpracovaných a nezpracovaných souborů, tento zásah umožňuje znovu zprocesovat soubory, které jsou už ve stavu zpracovaných souborů.

Soubory je možné ke znovu zpracování určit zvýrazněním v nodu *Sorter* (viz. modrá část workflow na obr. 4.1. Zvýrazněné soubory jsou v nodu *HiLite Row Splitter* vyřazeny ze seznamu zpracovaných souborů, což je v dalším kroku zahrne do opětovného zpracování.

Sledování stavu kontrolované složky

Automatičnost workflow a spuštění procesu zpracování dat a získání metrik je obstaráno dvěma vnořenými KNIME smyčkami. Vstupem první smyčky (žlutá část workflow na obr. 4.1) je tabulka zpracovaných souborů, která je získána buď z části pro nastavení opětovného zpracování souboru, a nebo z ukončení předchozího cyklu, který vrací aktualizovanou tabulku zpracovaných souborů.

Vstupní tabulka už zpracovaných souborů je přeposlána do nodu *Python script*, který ve *while* smyčce pravidelně získává z kontrolované složky seznam všech .mzML souborů a porovnává je se vstupní tabulkou. Pokud se ve složce nacházejí nové soubory je *while* smyčka ukončena a tabulka nových souborů je ve workflow posunuta k dalšímu nodu, kde jsou soubory seřazeny abecedně.

Po seřazení tabulka souborů vstupuje do druhé KNIME smyčky, která už je v zelené části workflow na obr. 4.1. Smyčka slouží k procházení jednotlivých nových souborů, které jsou dále zpracovány a získané metriky uloženy. Jakmile jsou data zapsána v databázi z posledního souboru, smyčka je ukončena a seznam zpracovaných souborů je aktualizován a poslán na začátek první smyčky.

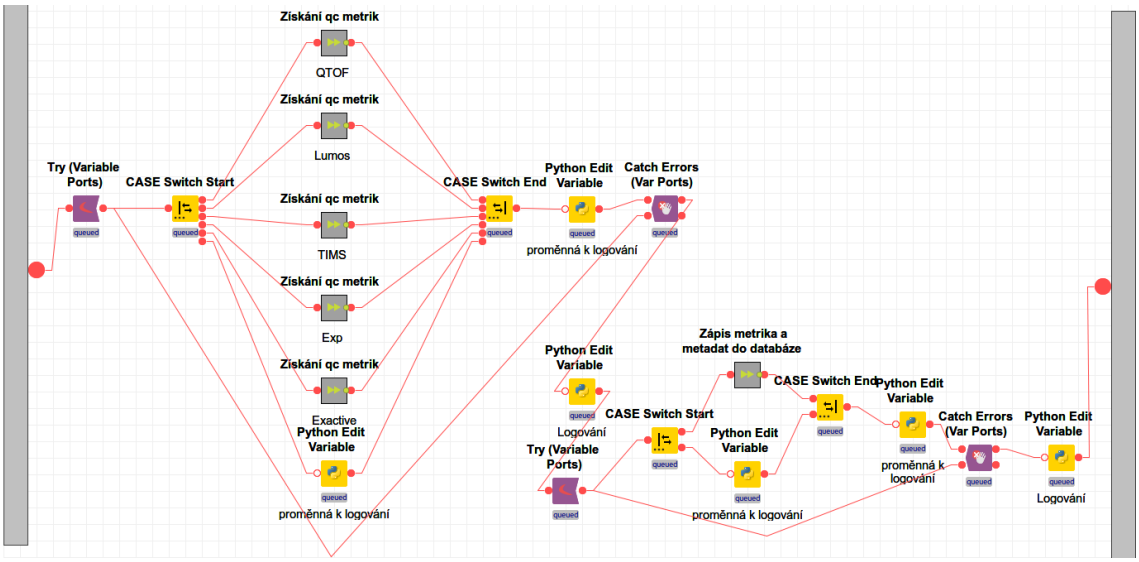
4.1.2 Zpracování souboru .mzML

Hlavní částí workflow je samotné zpracování vstupních souborů .mzML. Za účelem získání metrik pro kontrolu kvality je potřeba prvně provést identifikaci a kvantifikaci peptidů a proteinů. Tato část zpracování se liší podle přístroje, na kterém proběhlo měření. Workflow se z toho důvodu dělí do pěti větví (viz obr. 4.3), které se liší v různém nastavení. Větve jsou nastaveny přímo podle přístrojů v laboratoři, pro kterou je workflow vytvářen. Pokud by byl používán nový přístroj, je nutné přidat novou větev a nastavit ji podle vlastností přístroje, zadat identifikátor nového přístroje podle jeho zkratky do proměnné, která určuje větev v nodu *Case Switch Start*, a také je nutné zadat záznam o přístroji do databáze (viz. 4.14).

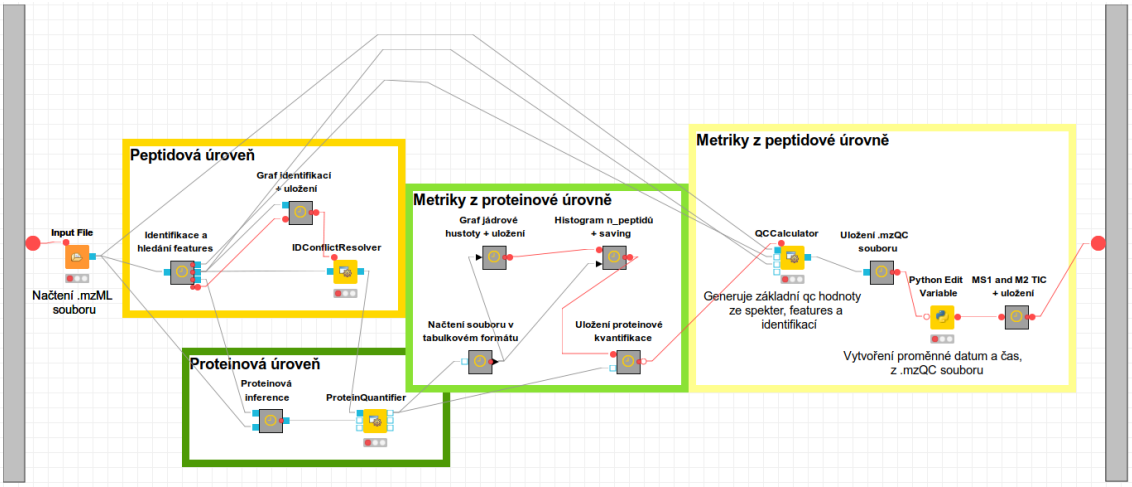
V rámci každé větve je metanod, který obsahuje veškeré kroky pro získání metrik kvality (obr. 4.4).

Po větvení workflow na přístroje jsou data uložena do databáze, to zabezpečuje poslední metanod na obrázku 4.3.

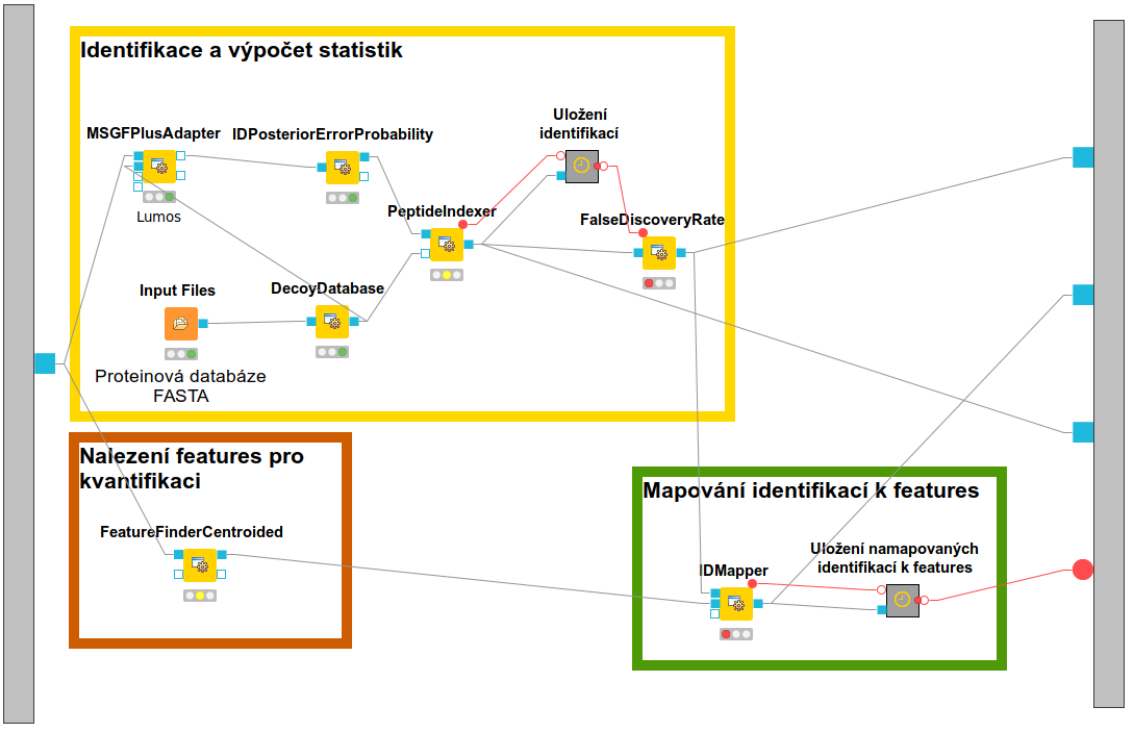
Nody pro získání dat i zapsání do databáze jsou zaobaleny nody *Try – Catch Errors*. Nody slouží k tomu, aby při jakékoli chybě v nodech, které jsou mezi nimi,



Obr. 4.3: Část workflow, která se stará o samotné identifikace, kvantifikace a získání metrik. Worflow se dělí podle přístroje, na kterém proběhlo měření a byl z něj získaný .mzML soubor. V druhé části je obsažen metanod, který se stará o uložení do databáze. Části zpracování dat i zápis do databáze jsou zabaleny mezi nody *Try – Catch Errors*, které by měli v případě chyby ve zpracování či zápisu do databáze umožnit chod workflow s dalším souborem. Na to jsou navázány skripty, které pomáhají zalogovat úspěšnost zpracování a zápisu dat. Tato část workflow je uvnitř metanodu „Zpracování souboru a uložení do databáze“ na obr. 4.1 vpravo nahoře.



Obr. 4.4: V rámci každé větve workflow pro určitý přístroj jsou vždy kroky identifikací a kvantifikací na peptidové úrovni (vyznačené tmavě žlutým obdélníkem), na proteinové úrovni (tmavě zelený obdélník), poté jsou získány metriky z proteinové úrovně (světle zelený obdélník) a metriky z peptidové úrovně (světle žlutý obdélník).



Obr. 4.5: Část workflow, která obsahuje nody pro identifikaci (žlutá část workflow) peptidů a nalezení LC-MS píků (hnědá část), které jsou na sebe následně mapovány (zelená část). Tato část workflow je součástí metanodu „Identifikace a hledání *features*“ v tmavě žluté části workflow na obr. 4.4.

se workflow nezastavilo, ale pokračovalo dál s uvedením chyby do záznamu. Tedy pokud nastane chyba v metanodu pro získání QC metrik, workflow skočí na nod *Catch Errors* a pokračuje ve spouštění dalších nodů. Z toho důvodu je ve workflow použité logování – zápis do souboru .log, zda daný soubor byl zpracován, a poté zda byl úspěšný zápis do databáze.

4.1.3 Získání identifikací a kvantifikací peptidů

Identifikaci a kvantifikaci peptidů a proteinů má na starost série nodů, které jsou k dispozici na obrázku 4.5.

Identifikace peptidů

Identifikace peptidů je zajištěna OpenMS nodem *MSFGPlusAdapter*, který má na vstupu .mzML soubor se spektry a FASTA soubor s proteinovou databází. Databáze je rozšířena i o *decoy* sekvence vytvořením reverzních sekvencí z originálních sekvencí nodem *DecoyDatabase*, v jehož nastavení je nutné uvést používaný enzym ke

štěpení dle typu měřeného vzorku (zde trypsin). Nod *MSFGPlusAdapter* získá identifikace pomocí skórování MS/MS spekter podle peptidů odvozených z proteinové databáze a na výstup generuje .idXML soubor. Adapter podporuje nastavení pro různé přístroje a postupy měření, proto se jeho nastavení liší skrz větve workflow. Adapter pro databázové prohledávání byl vybrán z více nabízejících se adapterů podle množství identifikovaných peptidů.

K získaným identifikacím je pomocí nodu *IDPosteriorErrorProbability* přidána hodnota pravděpodobnosti, že peptidové identifikace jsou nesprávně přiřazeny. Dalším krokem je aktualizace peptidových identifikací o informaci, zda se jedná o shodu se správnou nebo *decoy* sekvencí, což obstarává nod *PeptideIndexer*, který má na vstupu soubor s identifikacemi a proteinovou databázi. Informace, o kterou shodu se jedná, vyžaduje výpočet FDR nodem *FalseDiscoveryRate*. V nodu je nastaven výpočet FDR jen na peptidové úrovni a zároveň je zde provedena filtrace peptidových identifikací tak, aby globální FDR byla 1 % (q-hodnota $< 0{,}01$).

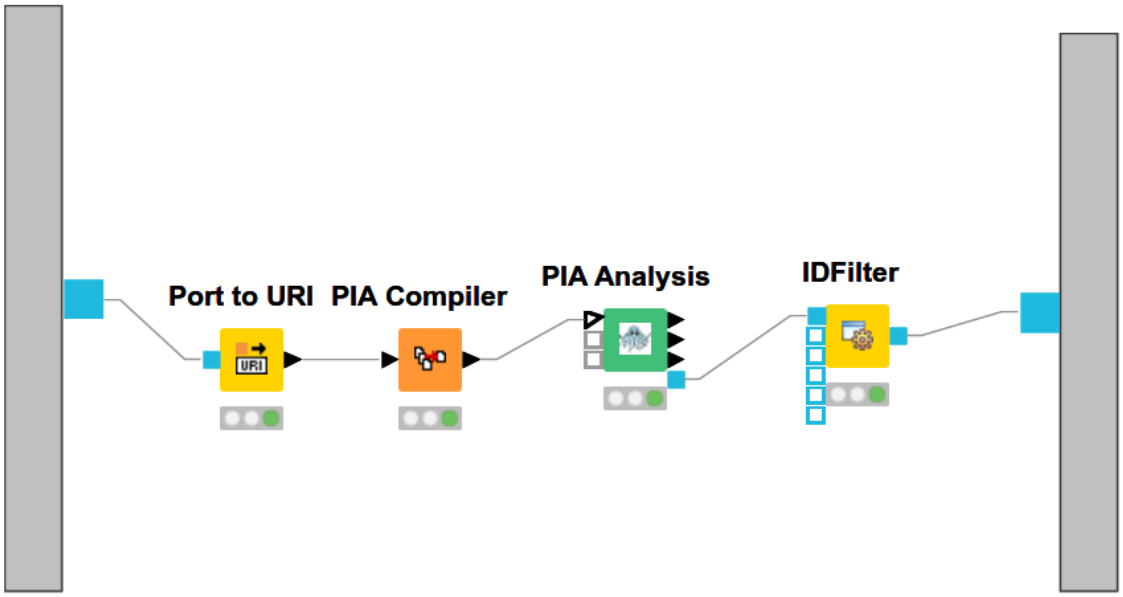
Nalezení LC-MS píků a mapování

Paralelně s identifikací probíhá detekce LC-MS píků za účelem kvantifikace s využitím nodu *FeatureFinderCentroided*. Vstupem je soubor se spektry .mzML a na výstupu je soubor .featureXML.

K propojení informací z identifikace peptidů (.idXML soubor) a kvantifikace (.featureXML soubor) je použit nod *IDMapper*. Mapování je založeno na retenčním čase a hodnotách m/z, kdy je identifikace přiřazena k LC-MS píku, pokud se nachází uvnitř hranice píku nebo je blízko centroidu píku. Identifikace jsou ponechány, i když nejsou přiřazeny k žádnému píku, ale jsou označeny v datech, že nebyly přiřazeny. Výstupem je soubor formátu .featureXML. Z těchto dat je v metanodu „Graf identifikací“ vytvořen graf (v části peptidová úroveň na obr. 4.4), který zobrazuje kumulativní intenzity identifikovaných a neidentifikovaných peptidů, který je uložen do databáze s metrikami.

Filtrace identifikovaných peptidů

Identifikované peptidy jsou následně profiltrovány nodem *IDConlictResolver* 4.4 tak, aby ke každému píku byla přiřazena jen jedna identifikace s nejlepším skóre. Tento krok je důležitý před proteinovou kvantifikací, protože píky s vícenásobnou identifikací nejsou použity k proteinové kvantifikaci.



Obr. 4.6: Část workflow pro identifikaci proteinů, který je součástí metanodu „Proteinová inference“.

4.1.4 Získání identifikací a kvantifikací proteinů

Identifikace proteinů

Identifikace proteinů má na starosti metanod „Proteinová inference“ (na obr. 4.4 na proteinové úrovni), který zabaluje nody aplikace *PIA Analysis* 4.6. PIA [35] je univerzální aplikace, která umožňuje prozkoumat výsledky peptidových identifikací, ale hlavní využitá funkcionalita je algoritmus proteinové inference. Před samotnou PIA analýzou je použit nod *PIA Compiler*, který upraví data do požadovaného formátu, do kterého vstupují data z nodu *PeptideIndexer* ve formátu .idXML. *PIA Analysis* na vstupu přebírá zkompilovaná data a může být poskytnut soubor .mzML se spektry (pro vytvoření interaktivního náhledu PSM). PIA nabízí tří různé metody pro proteinovou inferenci (metodu Occamovy břitvy, Spektrum extraktor, reportování všech proteinových skupin). Pro workflow byla vybrána metoda Occamovy břitvy, protože na datech dosahovala nejlepších výsledků v počtu identifikovaných proteinů. PIA nabízí i filtrování proteinových identifikací podle FDR, avšak filtrace není nastavena, protože data jsou filtrována až v dalším kroku nodem *IDFilter*.

IDFilter je nastaven, aby filtroval podle skóre pro proteinové skupiny na hranici 0.01 a aby vyloučil nepřiřazené peptidové identifikace.

Kvantifikace proteinů

Kvantifikace proteinů – zjištění množství proteinů je vypočteno z intenzit píků z peptidové úrovně pomocí nodu *ProteinQuantifier*. Intenzity píků jsou prvně akumulo-

vány do peptidového množství podle namapovaných peptidových identifikací k píkům a následně jsou množství peptidů zprůměrovány pro získání množství proteinů. Krok z peptidové úrovně na proteinovou používá generalizovanou metodu „Top 3“, která použije N proteotypických peptidů s největším množství pro každý protein k výpočtu množstvím proteinu. [36]

ProteinQuantifier má na vstupu data z nodu *IDConflictResolver* a metanodu „Proteinová inference“. A na jeho výstupu jsou data v tabulkovém formátu. Z proteinové kvantifikace jsou následně získány metriky pro vytvoření grafu jádrové hustoty a histogramu (viz. obr. 4.4), které jsou zapsány s metrikami do databáze.

4.1.5 Získání metrik

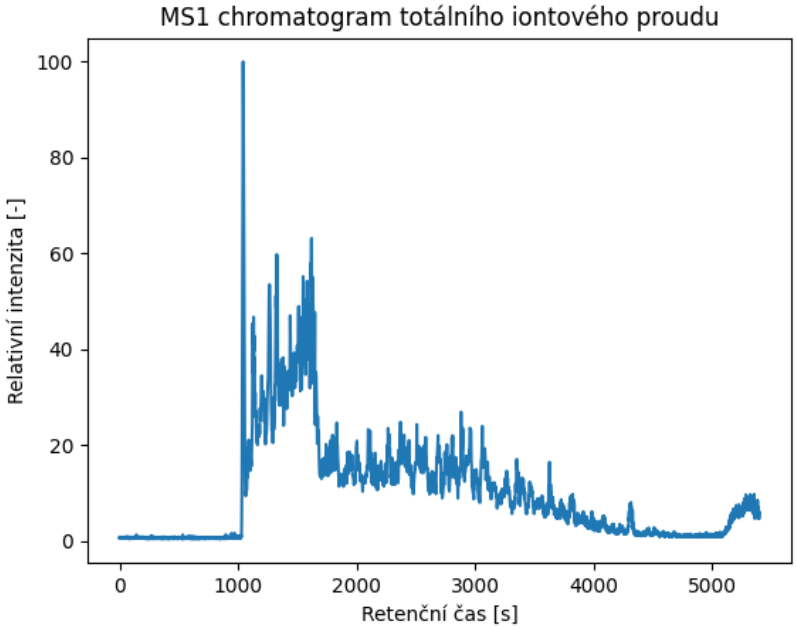
Metriky pro popis kvality hmotnostně spektrometrických dat jsou sbírány z peptidové i proteinové úrovně, z toho důvodu do zpracování dat jsou zařazeny identifikace a kvantifikace peptidů a proteinů.

Metriky peptidové úrovně

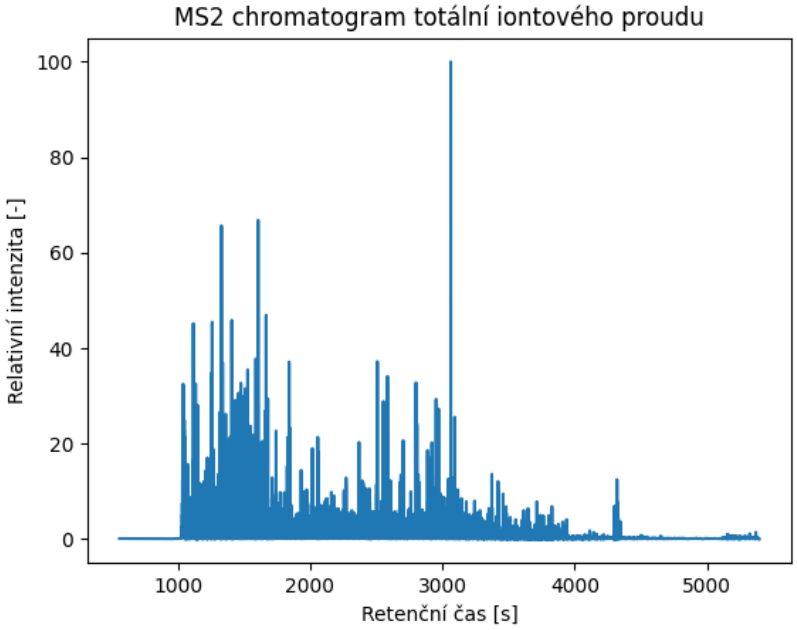
O sesbírání metrik z peptidové úrovně se stará nod *QCCalculator* (na obr. 4.4), který umožňuje vygenerovat starší (.qcML) i novější (.mzQC) formát souboru, který je následně ve workflow používán. Z .mzQC jsou v Python skriptu získány vybrané metriky. Skript předpokládá zavedenou JSON strukturu souboru .mzQC (viz obr. 2.1), a tak pomocí Python knihovny Json jsou převedeny z JSON formátu do Pythonu (objekty na slovník a pole na seznam). Následně jsou ze slovníku pod klíčem *QualityMetrics* vyindexovány metriky a podle jejich identifikátoru (*accession*) jsou přiděleny do databáze. Metriky, které jsou jako jedna hodnota, jsou přímo zapsány do databáze. Mezi vybrané metriky patří počet MS a MS/MS spekter, počet chromatogramů, doba trvání retenčního času, rozsah m/z hodnot, plocha pod TIC, počet MS skoků a poklesů, počet detekovaných sloučenin, celkový počet PSM, počet identifikovaných peptidů, průměrná délka identifikovaných peptidů, průměrná hodnota vynechaných štěpení peptidů, počet identifikovaných proteinů a průměrná hodnota identifikačního skóre.

Z metrik, které jsou zaznamenány jako vektor hodnot jsou vytvořeny v rámci workflow grafy. Jednou z takových metrik je MS chromatogram celkového iontového proudu – MS TIC 4.7. Tato metrika je zobrazena v podobě grafu, podobně je vypracován graf MS/MS TIC 4.8. Chromatogram iontového proudu zobrazuje sumu intenzit píků všech hmotnostních spekter vzhledem k retenčnímu času.

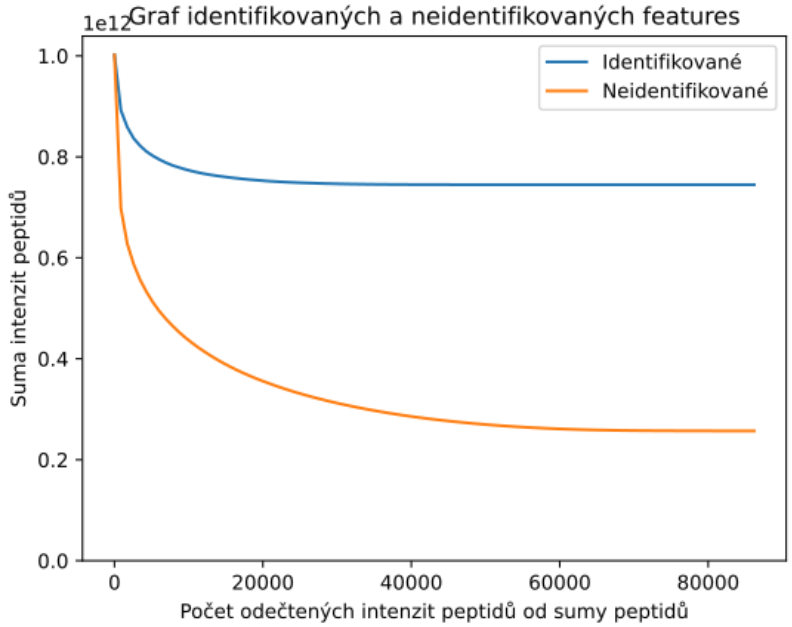
Na peptidové úrovni je vytvořen další graf ze souboru .featureXML po mapování identifikací k píkům, který zobrazuje kumulativní intenzity identifikovaných a neidentifikovaných píků 4.9. V grafu na ose x je počet odečtených intenzit píků od



Obr. 4.7: Chromatogram MS celkového iontového proudu, na ose x je retenční čas a na ose y je relativní intenzita.



Obr. 4.8: Chromatogram MS/MS celkového iontového proudu, na ose x je retenční čas a na ose y je relativní intenzita.



Obr. 4.9: Graf zobrazuje sumu intenzit identifikovaných a neidentifikovaných píků. Na ose x je počet odečtených intenzit peptidů od celkové sumy peptidů (intenzity peptidů jsou seřazeny sestupně). Na ose y je suma intenzit peptidů.

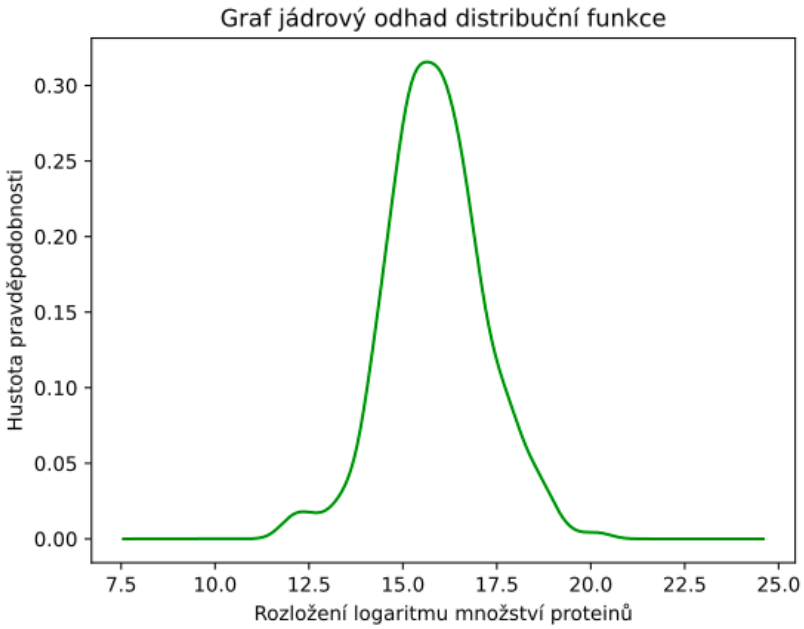
sumy všech intenzit (identifikovaných i neidentifikovaných) píků. Na ose y je tedy suma intenzit píků a jednotlivé sumy jsou vypočítány tak, že jsou vždy postupně od sumy všech intenzit odečítány sestupně seřazené intenzity identifikovaných nebo neidentifikovaných peptidů.

Metriky proteinové úrovně

Na proteinové úrovni ze získaných dat jsou vytvořeny grafy (obr. 4.4), které jsou ukládány s ostatními metrikami do databáze. Grafy, které nejsou vytvořeny z dat .mzQC souboru neodpovídá žádná hodnota v kontrolovaném slovníku, a tedy ani jim nepřipadá žádná QC CV hodnota. Proto je jim přidělena v databázi *accession*, která formátem odpovídá QC CV hodnotám, ale začíná číslem jedna (QC CV hodnoty začínají číslem čtyři).

Graf jádrového odhadu distribuční funkce znázorňuje hustotu pravděpodobnosti množství proteinů 4.10. Graf je vytvořen z hodnot množství proteinů, které byly kvantifikovány nodem *ProteinQuantifier*. Graf má na ose x rozložení logaritmované (ln) intenzity proteinů a na ose y hustotu pravděpodobnosti.

Druhým grafem je histogram zobrazující počet peptidů přiřazených k proteinu 4.11, který z dat kvantifikovaných proteinů využívá sloupce s počty peptidů na



Obr. 4.10: Graf jádrový odhad distribuční funkce zobrazuje hustotu pravděpodobnosti množství proteinů. Na ose y je hustota pravděpodobnosti a na ose x je rozložení logaritmu množství proteinů.

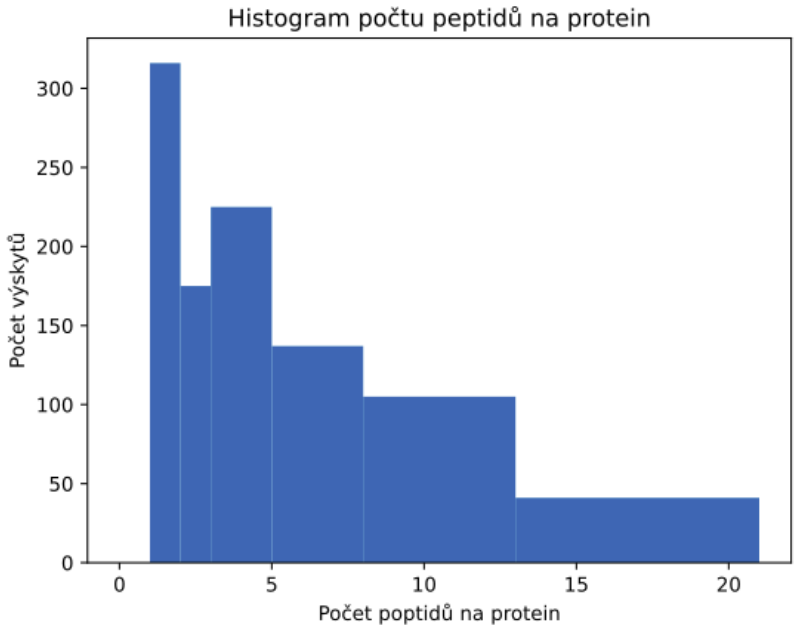
protein. Histogram je přizpůsobený tomu, že nejvíce proteinů má často přiřazeny jen nižší jednotky peptidů.

4.1.6 Uložení metrik do databáze

V předchozích částech workflow byla získána data, která je potřeba ukládat ve vhodné podobě pro další využití, zpracování nebo vizualizaci. Prvním krokem bylo uložení metrik kvality dat do jednoduchého tabulkového formátu, avšak pro lepší možnost manipulace s daty, přehlednost a větší rychlost je navržena databáze, která se dělí do několika propojených tabulek.

Databáze je vytvořena v relačním databázovém systému SQLite, který je napsaný v jazyce C. SQLite je založena na jednoduchosti, malé velikosti a databáze se ukládá do souboru .sqlite. SQLite umožňuje paralelní čtení ze souborů, avšak paralelní zápis možný není. Mezi nevýhody SQLite patří nižší výkon pro velké množství zápisů do databáze, ale je možné rychlost zvýšit využitím příkazy *begin transaction* a *commit* [37].

Databáze v rámci workflow je používaná nejen k zápisu metrik kvality, ale i k zapsání metadat o měření (např. datum měření, název použitého přístroje, metody, jméno měřeného standardu). Pro ukládání potřebných informací je vytvořeno sedm

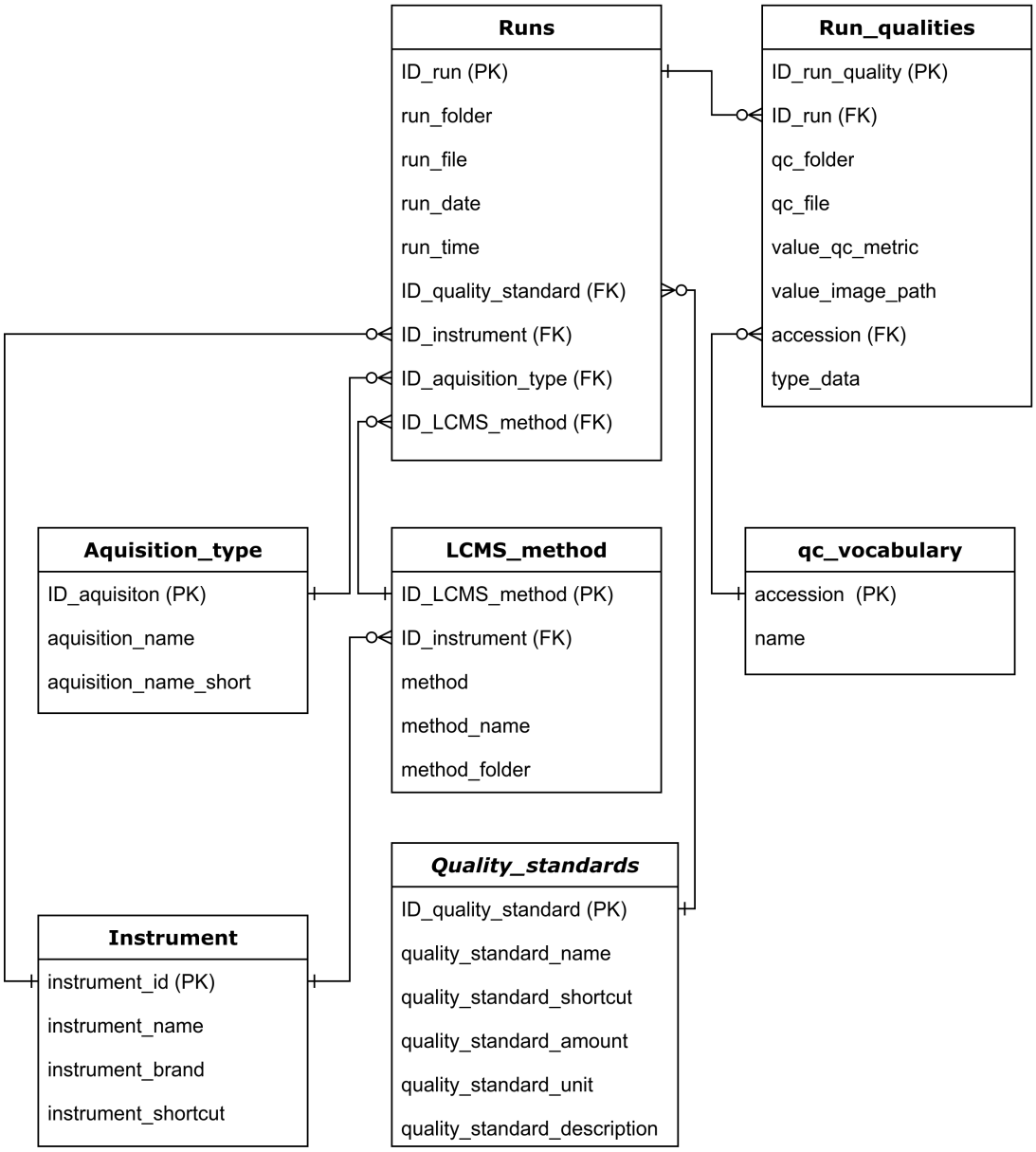


Obr. 4.11: Histogram, který zobrazuje hodnoty počtu peptidů na protein.

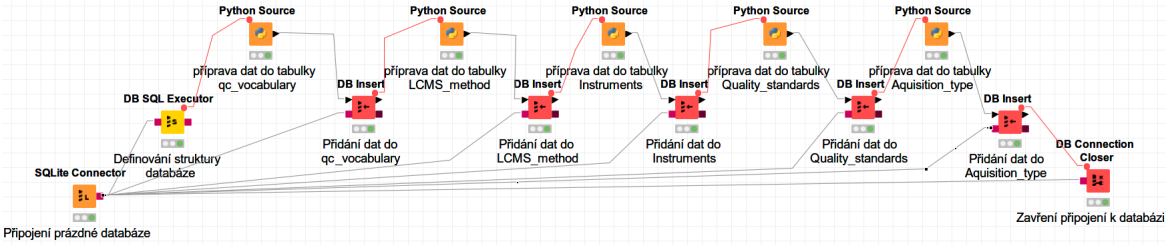
provázaných tabulek (obr. 4.12).

Tabulky by mohly být teoreticky rozděleny na část tabulek, které jsou předefinované a workflow pro automatickou kontrolu kvality dat z nich jen čte hodnoty, které jsou pomocí identifikátoru zaznamenány do jedné z tabulek *Run_qualities* 4.1.6 nebo *Runs* 4.1.6. Tento první typ tabulek nese informaci o dostupných přístrojích, standardech pro kontrolu kvality, akvizičních typech, LC-MS metodách a vybraných metrikách z kontrolovaného slovníku. Do druhého typu tabulek se právě zapisuje ve workflow pro automatickou kontrolu kvality dat, tedy je do nich zapisováno pokaždé, když je zpracováván nový soubor. K druhému typu tabulek patří *Run_qualities* a *Runs*.

Pro vytvoření struktury databáze je navrhnuto menší samostatné workflow (obr. 4.13), které prvně nadefinuje všech sedm tabulek s jejich atributy, jejich datové typy, primární a cizí klíče. Tato struktura je napsána v SQL a v KNIME implementována v nodu *DB SQL Executor*. Pro zápis do databáze je na začátku potřeba připojení k databázi, u které se předpokládá, že je prázdná, a protože je použitý SQLite databázový systém, připojení je vytvořeno nodem *SQLite Connector*. Po vytvoření struktury databáze ve workflow jsou připojena data (*Python source* nod) a jsou vkládána do databázových tabulek (*DB Insert*) jako první záznamy. Data jsou vkládána jen do tabulek, které slouží pro čtení ve workflow pro automatickou kontrolu kvality dat. Po naplnění tabulek je práce s databází ukončena nodem *DB Connection Closer*. Připojení k databázi je nutné ukončit z důvodu, aby nedocházelo



Obr. 4.12: Databáze je složená ze sedmi tabulek, které jsou vždy provázány vztahem jedna k mnoha. Tabulky *Runs* a *Run qualities* jsou určeny k zapisování metrik a metadat z každého souboru (měření), ostatní tabulky uchovávají informace o přístrojích, akvizicích, vybrané metriky z kontrolovaného slovníku, LC-MS metodách, standardech pro měření kvality.



Obr. 4.13: Workflow pro vytvoření struktury databáze. Workflow je složeno z připojení k prázdné databázi, definici struktury databáze pomocí *DB SQL Executor*, naplnění tabulek záznamy pomocí série dvojic nodů *Python source*, kde je definovaná tabulka s daty, a *DB Insert*, kde jsou data jako záznam vloženy do dané tabulky v databázi. V závěru je nod *DB Connection Closer*, který závírá připojení k databázi.

později k žádostem o paralelní zápis do databáze, které u SQLite není možné.

Do databázových tabulek je možné velmi snadno přidat nové záznamy (např. za účelem přidání nového používaného přístroje nebo nové metriky kvality). V software KNIME je k tomu potřeba jen nod k připojení k databázi, nod pro vytvoření tabulky se záznamem, nod pro zápis do databáze a nod k uzavření připojení k databázi (viz obr. 4.14)

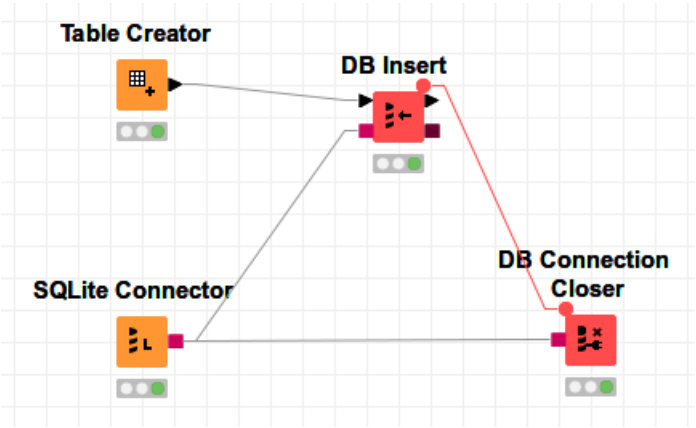
Tabulka *qc vocabulary*

Tabulka, která obsahuje vybrané metriky z kontrolovaného slovníku používaném v souboru formátu .mzQC se jmenuje *qc_vocabulary* a vždy obsahuje číslo metriky a jméno metriky obr. 4.15. Číslo metriky je použité jako primární klíč tabulky a v tabulce *Run_qualities* je cizím klíčem. Metriky byly vybrány podle využití v rámci proteomické laboratoře.

Číslo metriky odpovídá identifikátoru metriky v kontrolovaném slovníku, ale je v databázi používán bez prvních tří znaků, který jsou v kontrolovaném slovníku vždy dvě písmena a dvojtečka. Tedy pokud je v kontrolovaném slovníku identifikátor *QC:4000065*, v databázi je jako primární klíč u atributu *Accession* identifikátor *4000065*. Primární klíč má tedy datový typ číselný. Jména metriky jsou taktéž přímo převzané z kontrolovaného slovníku a jejich atribut předpokládá datový typ textový. Metrik bylo vybráno celkem 21 (viz obr. 4.16).

Tabulka *Runs*

Tabulka *Runs* obsahuje informace o měření (viz. obr. 4.17). Je v ní zaznamenáno umístění a jméno souboru .mzML (obě hodnoty jsou datového typu text), ze kterého vychází následná analýza, datum a čas měření (textový datový typ), identifikátoru



Obr. 4.14: Příklad přidání nového záznamu do tabulky. *SQLite Connector* je nutný pro připojeni k databázi, do které má být záznam přidán. V nodu *Table Creator* je předpřipraven záznam do databáze, kde názvy sloupců odpovídají atributům v databázové tabulce a řádky jsou už samotné záznamy, které se do tabulky vkládají. U atributu nesoucí primární klíč není nutné definovat záznam, pokud atribut předpokládá, že primární klíč se automaticky inkrementuje s novým záznamem. V nodu *DB Insert* je nutné vybrat tabulku, do které je záznam přidáván. Následně stačí sérii nodů spustit.

qc_vocabulary
accession (PK)
name

Obr. 4.15: Tabulka uchovávající metriky z kontrolovaného slovníku. *Accession* je použitý jako primární klíč.

accession	name
4000059	Number of MS1 spectra
4000060	Number of MS2 spectra
4000135	Number of chromatograms
4000053	RT duration
4000138	MZ acquisition range
4000067	Total ion current chromatogram
4000077	Area under TIC
4000069	MS1 Total ion current chromatogram
4000172	MS1 signal jump (10x) count
4000173	MS1 signal fall (10x) count
4000070	MS2 Total ion current chromatogram
4000257	Detected Compounds
4000186	Total number of PSM
4000187	Number of identified peptides
4000214	Identified peptide lengths - mean
4000209	Missed cleavages - mean
4000185	Number of identified proteins
4000204	Identification score - mean
1000001	Protein abundance
1000002	Number of peptides per protein
1000003	Gtraph identified and non identified features

Obr. 4.16: Tabulka vkládaných metrik z kontrolovaného slovníku. *Accession* je použitý jako primární klíč.

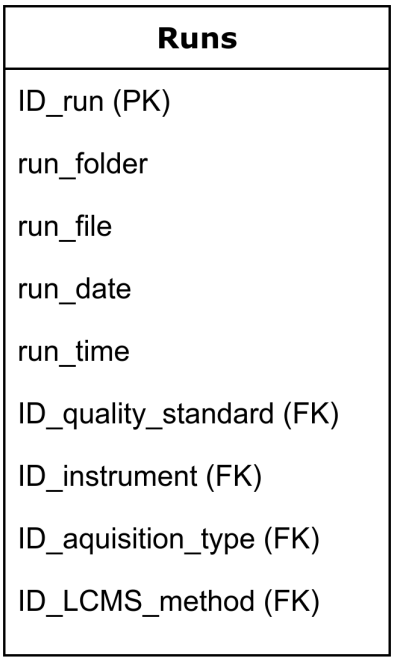
jména měřeného standardu. Následně jsou zaznamenány identifikátory použitého přístroje, akvizice a LC-MS metody použité k měření (veškeré hodnoty identifikátorů jsou v této tabulce číselné).

Jméno složky, kde je .mzML soubor uložen, je podle nastavené kontrolované složky pro automatické zpracování dat a jméno souboru, je zapsáno podle aktuálně zpracovávaného vstupního souboru do workflow. Datum a čas je získán z .mzQC souboru, kde je uložen v rámci sesbíraných metrik.

Z názvu .mzML souboru, který musí mít laboratoří zavedený název podle dané konvence, je získána informace o měřeném standardu pro kontrolu kvality, použitém přístroji a množství standardu. Konvence názvu .mzML souboru je určená do podoby *QC_přístroj_QC standard_množství v ng_další informace podle potřeby.mzML*. V názvu souboru by měl přístroj odpovídat zkratce (*instrument_shortcut*) z některých přístrojů v tabulce *Instrument*, QC standard by měl odpovídat některé zkratce standardu (*quality_standard_shortcut*) v tabulce *Quality_standards*. Všechny hodnoty jsou následně použity v rámci ukládání dat do databáze a je tedy velmi důležité, aby soubor byl uložen a nazván správně.

Tabulka *LCMS method*

Tabulka *LCMS method* uchovává informaci jaká metoda byla použita pro měření vzorku (obr. 4.18). Metoda se liší podle měřeného přístroje, proto je v tabulce přes cizí klíč zaznamenán identifikátor přístroje. Dále tabulka nese informaci, zda se jedná

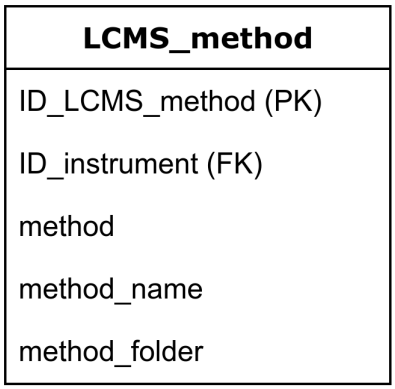


Obr. 4.17: Tabulka nesoucí informace o jednom měření, kde je primární klíč *ID_run*, který se automaticky zvyšuje s novým záznamem a má čtyři cizí klíče značených FK. Tabulka nese tedy hlavně metadata o měření.

o LC, MS nebo LC-MS metodu, jméno metody a složku (hodnoty jsou textového typu), kde je metoda uložena. Primárním klíčem tabulky je *ID_LCMS_method* (identifikátor se automaticky zvyšuje s novým záznamem) a tento identifikátor je použitý jako cizí klíč v tabulce *Runs*. Informace v této tabulce jsou vyplněny podle zavedených metod a jejich umístění v laboratoři 4.19.

Tabulka *Instrument*

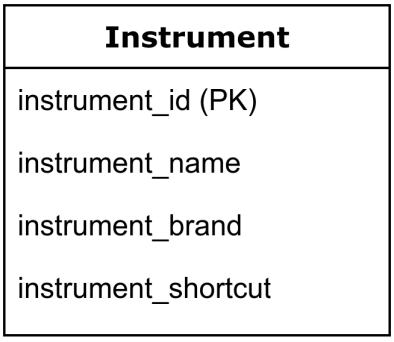
Tabulka *Instrument* 4.20 je určena k zaznamenání základních informací o přístrojích použitých k měření. Tabulka obsahuje číselný identifikátor přístroje, který je použitý jako primární klíč a automaticky se zvyšuje při novém záznamu. Dále je obsaženo jméno přístroje, jeho používaná zkratka a výrobce přístroje (hodnoty jsou textové). Tato tabulka je propojena s tabulkou *Runs*, z důvodů přiřazení přístroje použitého u daného měření. Také je propojena s tabulkou *LCMS_method*, protože každá metoda je používána u určitého přístroje. Přístroje, které jsou vloženy do databáze jsou na obrázku 4.21.



Obr. 4.18: Tabulka slouží k zaznamenání dostupných LC, MS nebo LC-MS metod pro hmotnostní spektrometrii. Primárním klíčem tabulky je *ID_LCMS_method* a přes cizí klíč *ID_instrument* je uloženo id přístroje, pro který je metoda použita.

S method_name	S method_folder	I ID_instrument	S method
74min_07	D:\Methods\AUR2-25075C18A-CSI\	3	LC
DDA PASEF-standard_201102	D:\MethodsimsControl_methods\	3	MS
74min-IS_2sec_MaxIT-050_AGC050_210714	D:\Methods_EASY_PepMap_25cmX75um_2um\	4	LCMS

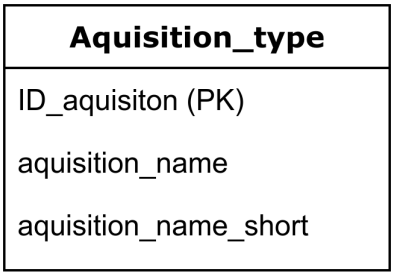
Obr. 4.19: Vkládané hodnoty do databázové tabulky *LCMS_method*, které odpovídají používaným metodám v laboratoři.



Obr. 4.20: Tabulka je používána k zaznamenání informací o přístrojích použitých k měření. Jejím primárním klíčem je *instrument_id*.

instrument_name	instrument_brand	instrument_shortcut
Impact II	Bruker	QTOF
Orbitrap Fusion Lumos	Thermo	Lumos
timsTOF Pro	Bruker	TIMS
Orbitrap Exploris 480	Thermo	Exp
Q_Exactive	Thermo	Exactive

Obr. 4.21: Tabulka přístrojů (jejich jméno, výrobce a zkratka), které jsou vloženy do databáze, aby byly k dispozici pro zápis do tabulky *Runs* při zpracování souboru s jedním měřením.



Obr. 4.22: Tabulka nese jméno a zkratku možných akvizic (DDA, DIA) pro měření. *ID_aquisition* je použito jako primární klíč.

Tabulka *Aquisition type*

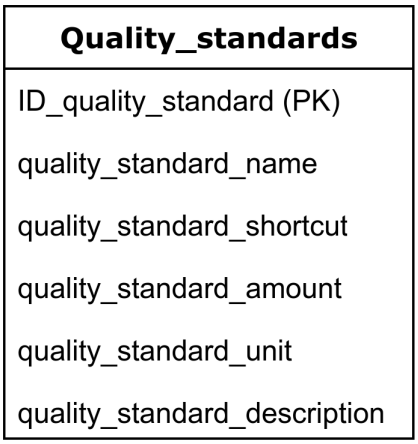
Typy akvizice jsou uloženy v tabulce *Aquisition_type* 4.22, ve které je uloženo jméno a zkratka akvizice (hodnoty jsou textové). Tabulka má primární klíč *ID_Aquisition_type*, který je číselný a automaticky se zvyšuje při novém záznamu. Akvizice, které jsou používány pro měření se dělí na DDA – *data dependend analysis* (analýzu datově závislou) a DIA – *data independent analysis* (analýzu datově nezávislou) (vložená data do databáze jsou na obr. 4.23). Identifikátor akvizice je použitý v tabulce *Runs*, aby mohla být k měření použitá informace o akvizici.

Tabulka *Quality standards*

Informace o použitém standardu jsou zaznamenány v tabulce *Quality_standards* 4.24, ve které je obsaženo jméno a zkratka standardu pro kontrolu kvality (tex-

aquisition_name	aquisition_name_short
data dependent acquisition	DDA
data independent acquisition	DIA

Obr. 4.23: Vložené hodnoty do databázové tabulky *Aquisition type*.



Obr. 4.24: Tabulka je určena k uložení jména standardu pro kontrolu kvality, jeho množství a popis. Primárním klíčem tabulky je *ID_quality_standard*.

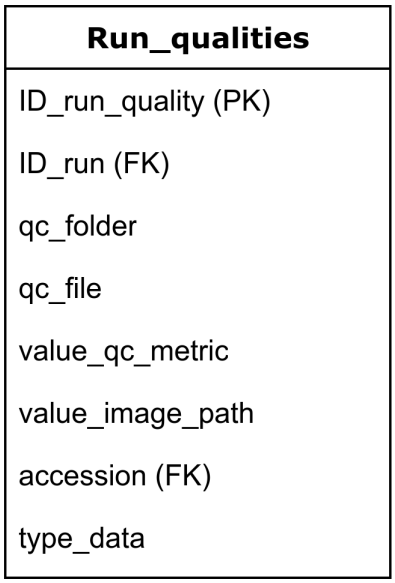
S quality_standard_name	S quality_standard_shortcut	D quality_standard_amount	S quality_standard_unit	S quality_standard_description
iRT	blank	0.01		100x diluted stock solution
HeLa cell line digest	HeLa	100	ng	100ng
MEC cell line digest	MEC	100	ng	100ng

Obr. 4.25: Hodnoty vložené do tabulky *Quality standards* pro potřeby workflow automatické kontroly kvality dat.

tový datový typ), ale i jeho používané množství pro měření (číselný datový typ) a jeho jednotka (textový datový typ). Navíc tabulka umožňuje přidat informaci o standardu do popisu (textový datový typ). Primárním klíčem tabulky je číselný *ID_quality_standard*, který se automaticky zvyšuje s novým záznamem, a identifikátor použitého standardu pro kontrolu kvality je zaznamenán v tabulce *Runs* přes cizí klíč. Vložené hodnoty do tabulky *Quality standards*, které jsou používány následně ve workflow pro automatickou kontrolu kvality dat jsou na obrázku 4.25.

Tabulka *Run qualities*

Metriky kvality, které jsou získány ve workflow zpracováním vstupních dat a uloženy v souboru .mzQC, jsou do databáze zaznamenány v tabulce *Run_qualities* (obr. 4.26). Každá metrika má svůj číselný primární klíč, který je uložen do tabulky pod atributem *ID_run_quality* a jeho hodnota se automaticky zvyšuje s novým záznamem. Také je každá metrika přiřazena k jednomu měření, ze kterého je získána, což je zařízeno propojením s tabulkou *Runs* číselným cizím klíčem *ID_run*, a tedy jedno měření může mít více metrik. Do tabulky je možné uložit cestu a název .mzQC souboru (atributy v databázové tabulce mají textový datový typ), ze kterého jsou metriky získány. Tabulka umožňuje uložit metriku jako číselnou hodnotu. Pokud se



Obr. 4.26: Tabulka je určena pro uložení naměřených a získaných hodnot metrik. Zapisované metriky jsou vždy přiřazeny k určitému měření přes cizí klíč *ID_run* a je definován typ metriky přiřazením přes cizí klíč *accession*.

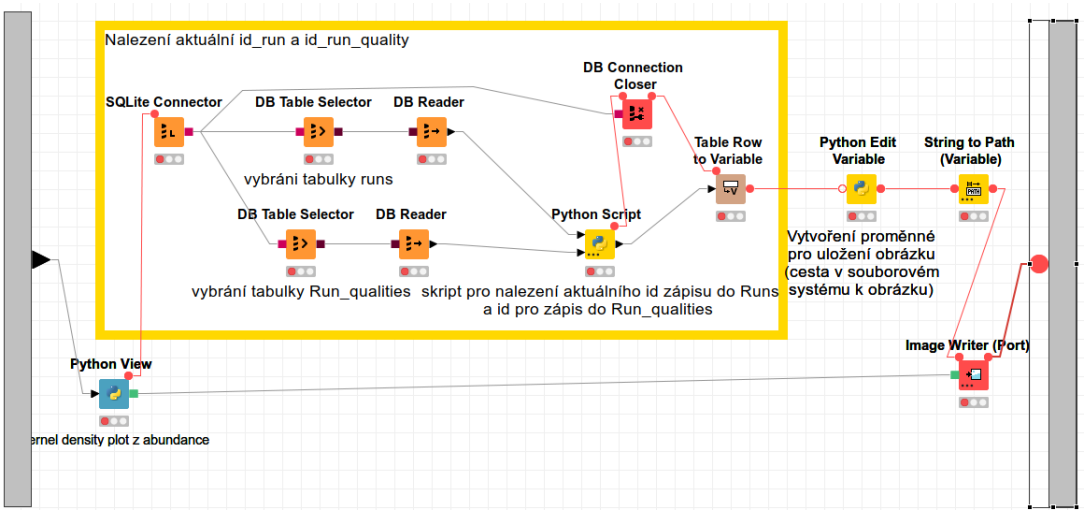
jedná o metriku, ze které byl vytvořen graf, tak je možné uložit cestu ke grafu do tabulky.

Workflow z určitých metrik vytváří obrázky automaticky a aby grafy měly nadefinovaný unikátní název, je cesta složena z jednotlivých identifikátorů pro měření. Přesněji, obrázek je pojmenován „*identifikátor metriky_identifikator aktualniho měření_ identifikátor aktuální zapisované metriky.svg*“. Identifikátor metriky je získán z tabulky *qc_vocabulary*, která je i v tabulce *Run_qualities* jako cizí klíč. Identifikátor aktuálního měření je získán z tabulky *Runs*, který je též předán jako cizí klíč a identifikátor aktuální získané metriky je určen právě identifikátor pro aktuálně ukládanou metriku v tabulce *Run_qualities*. Ukázka vygenerování grafu ze zpracovaných dat ve workflow a získání hodnot z databáze pro vytvoření názvu souboru pro uložení grafu je na obr. 4.27.

Dále obsahuje tabulka *Run_qualities* číselný identifikátor metriky z tabulky *qc_vocabulary* a určení typu dat, které jsou do tabulky ukládány, aby bylo umožněno snadné filtrování metrik při práci s databází.

Zápis do tabulek *Runs* a *Run qualities*

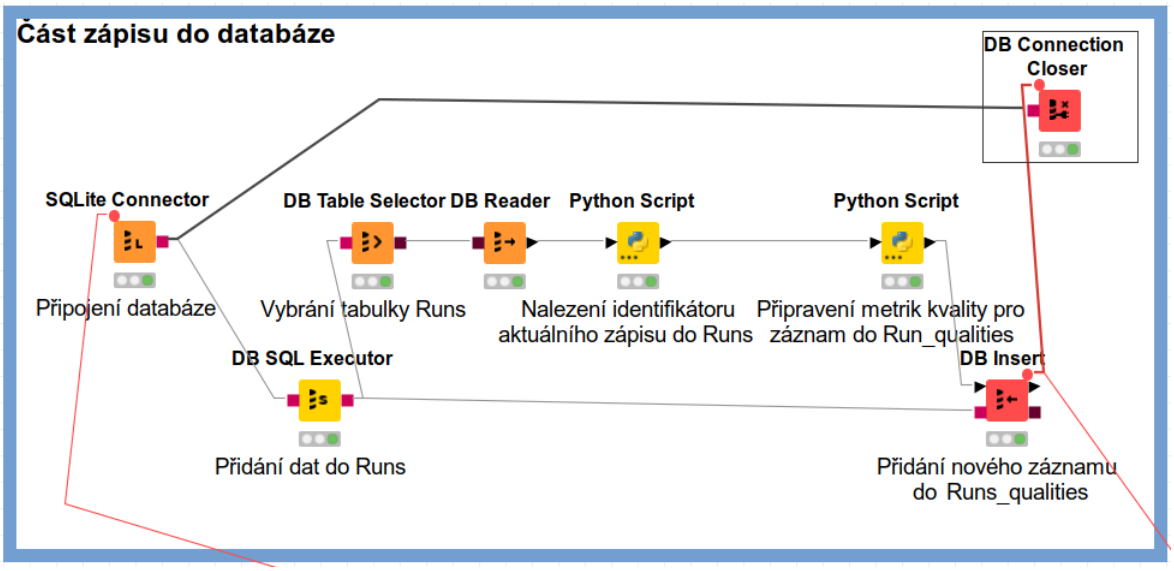
Do databázové tabulky *Run_qualities* a *Runs* jsou hodnoty ukládány ve workflow pro automatickou kontrolu kvality dat po zpracování dat v první části workflow (obr. 4.28). Ve workflow je prvně nodem *SQLite Connector* připojena databáze, na který



Obr. 4.27: Ukázka části z workflow pro automatickou kontrolu kvality dat, která ze zpracovaných dat generuje graf v nodu *Python View* a následně z databáze získá hodnotu identifikátorů pro aktuální zápis do databázové tabulky *Runs* a *Run_qualities* (nody ve žluté oblasti). V závěru je vytvořena cesta pro uložení grafu do souborového systému, převedení cesty do KNIME proměnné, které je použita k zapsání grafu do souborového systému a proměnná je ve workflow zapamatována pro pozdější zápis cesty do databázové tabulky *Run_qualities*.

navazuje nod *DB SQL Executor*, ve kterým je implementováno jazykem SQL přidání hodnot do tabulky *Runs*. Dále jsou přidány získané hodnoty z workflow – datum a čas měřeného vzorku, cesta ke složce a název .mzML souboru, poté jsou uloženy identifikátory hodnot z propojených tabulek přes cizí klíč. Identifikátory hodnot jsou nalezeny tak, že získaná hodnota z workflow je porovnána se záznamy v tabulce a je vrácen odpovídající identifikátor. Například ve workflow je zjištěna hodnota standardu pro kontrolu kvality HeLa a v databázové tabulce *Quality standards* má HeLa identifiktor 2, je vrácena hodnota 2 a uložena k atributu *ID_quality_standard*.

Na zápis do tabulky *Runs* navazuje série nodů, který získají hodnotu identifikátoru aktuálního zápisu do tabulky *Runs*, který je následně použit pro zápis záznamů do tabulky *Run Qualities*. Zapisované záznamy jsou předpřipraveny v nodu *Python Script*. Hodnoty jsou buď vyčteny z .mzQC souboru, a nebo už byly získány dříve ve workflow a byly uloženy v proměnné. Příklad připravené tabulky pro zápis do *Run_Qualities* je na obrázku 4.29.



Obr. 4.28: Část workflow pro automatickou kontrolu kvality dat, který zajišťuje zapsání do databáze při zpracování nového souboru. Tato část workflow zajišťuje zápis do tabulek *Runs* a *Run_qualities*.

accession	type_data	value_image_path	ID_run	value_qc_metric	qc_folder	qc_file
1000003	path	/home/anna/proteo...	2	NaN	/home/anna/...	QC_Lumos...
1000001	path	/home/anna/proteo...	2	NaN	/home/anna/...	QC_Lumos...
1000002	path	/home/anna/proteo...	2	NaN	/home/anna/...	QC_Lumos...
4000069	path	/home/anna/proteo...	2	NaN	/home/anna/...	QC_Lumos...
4000070	path	/home/anna/proteo...	2	NaN	/home/anna/...	QC_Lumos...
4000059	single_value	?	2	7,234	/home/anna/...	QC_Lumos...
4000060	single_value	?	2	35,906	/home/anna/...	QC_Lumos...
4000135	single_value	?	2	1	/home/anna/...	QC_Lumos...
4000053	single_value	?	2	5,399	/home/anna/...	QC_Lumos...
4000077	single_value	?	2	1,204,010,205	/home/anna/...	QC_Lumos...
4000172	single_value	?	2	16	/home/anna/...	QC_Lumos...
4000173	single_value	?	2	0	/home/anna/...	QC_Lumos...
4000257	single_value	?	2	948	/home/anna/...	QC_Lumos...
4000186	single_value	?	2	727	/home/anna/...	QC_Lumos...
4000214	single_value	?	2	10.264	/home/anna/...	QC_Lumos...
4000209	single_value	?	2	0.05	/home/anna/...	QC_Lumos...
4000204	single_value	?	2	0	/home/anna/...	QC_Lumos...
4000185	single_value	?	2	553	/home/anna/...	QC_Lumos...
4000187	single_value	?	2	611	/home/anna/...	QC_Lumos...

Obr. 4.29: Ukázka připravených hodnot pro zápis do tabulky *Run_qualities*.

4.2 Metodika kontroly workflow

Metodiku kontroly workflow byla vytvořena podle návrhu workflow a tak, aby bylo zkontrolováno zda jsou splněny základní požadavky na workflow a jeho použitelnost v praxi. Pro otestování byla vytvořena sada simulovaných dat a byl použitý jeden soubor s reálnými daty.

4.2.1 Kontrolované požadavky na workflow

Základních požadavků na workflow pro automatickou kontrolu hmotnostně spektrometrických dat je několik. Mezi ně patří automatičnost workflow, která je nutná pro plynulost a samostatnost zpracování dat. Uživatel by měl mít co nejméně práce s obsluhou workflow.

Dalším požadavkem je správné zpracování dat, uložení dat a metrik pro možnou zpětnou kontrolu kvality nebo i sledování vývoje kvality v čase podle jakýchkoli zvolených metrik. Z toho důvodu by mělo workflow splňovat i možnost kontroly, zda data byla v pořádku zpracována a uložena.

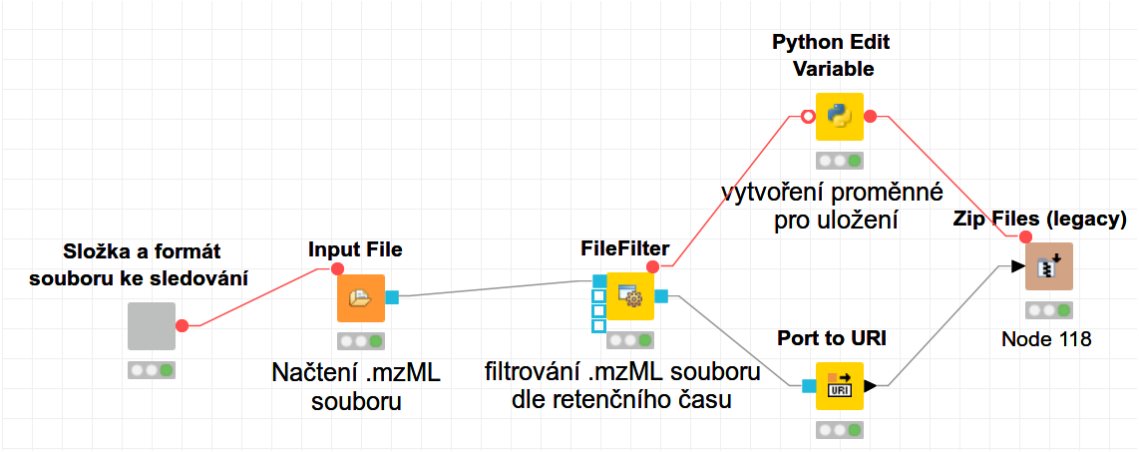
Workflow by mělo být schopné získat srovnatelné hodnoty identifikací a zpracování by mělo být srovnatelně časově náročné jako aktuálně používaný software, který workflow bude nahrazovat v proteomické laboratoři, pro kterou bylo workflow navrhnuto.

Workflow by mělo zpracovat data podle správného přístroje, na kterém byla data měřena.

4.2.2 Testovací data

K ověření workflow bylo připraveno několik souborů simulovaných dat a jeden soubor reálných dat, protože zpracování reálného souboru je časově náročnější. Časová náročnost je z důvodu velkého objemu reálných dat (zhruba od 500 MB), ale také i z důvodu většího množství peptidových sekvencí (delší trvání identifikace s databázovým prohledáváním).

Simulovaná data jsem vytvořila z .mzML souboru z reálného měření HeLa buněčné linie na přístroji Orbitrap Fusion Lumos, ze kterého jsem vyseparovala naměřená spektra podle retenčního času. Vybrané spektra jsou v rozsahu retenčních časů: 0-120 s, 1000-1120 s, 2000-2120 s, 3000-3120 s, 5200-5320 s ze souboru, ze kterého byl vytvořen TIC chromatogram 4.7 (zmíněno pro porovnání, kolik mohlo být identifikací a kvantifikací ve vybraných časech). Pro vyseparování dat byl použitý nod v KNIME *FileFilter*, ve kterém jsem vybrala různé časové úseky, tak aby soubory obsahovaly různé množství naměřených spekter 4.30.



Obr. 4.30: Způsob přípravy simulovaných dat v KNIME. Simulovaná data byla vytvořena z .mzML souboru, který byl vyfiltrován pomocí nodu *FileFilter* podle retenčního času a data byla následně uložena.

Soubory jsem pojmenovala podle předepsaného formátu (viz. 4.1.6) a dala jsem jim do názvu zkratky přístrojů pro všechny přístrojové větve ve workflow (větvení dle přístroje určuje zkratka přístroje v názvu souboru). Vytvořila jsem tedy 5 souborů s názvy, které jsou podle předpisu názvu souboru bezchybně a k tomu jsem vytvořila soubory, které mají chyby v názvu přístroje, v názvu standardu pro kontrolu kvality i v množství. V celku bylo tedy vytvořeno 8 souborů se simulovanými daty.

Reálná data, která byla použita pro testování pochází ze stejného přístroje, na kterém byl měřen stejný typ vzorku pro kontrolu kvality HeLa. Ze souboru s reálnými daty („QC_Exactive_HeLa_100ng_74min_DDA_220124_01.mzML“) je vytvořen menší soubor obsahující jen část spekter, který je součástí přílohy (viz popis k příloze A).

4.2.3 Testování workflow

Testování se simulovanými daty

Worfklow bylo prvně otestováno pomocí simulovaných dat, kterými bylo zkontrolováno zda workflow zvládne automaticky načíst a zpracovat data, která ve složce už existují. Poté byly postupně do složky přidávány nové soubory, aby bylo ověřeno, zda workflow zvládá zjistit nový stav složky se soubory a zařadit je ke zpracování.

Mezi simulovanými daty, byl i soubor, který neobsahoval velké množství spekter, ze souboru byly identifikovány jen 4 peptidy a žádný protein a to vedlo k vygenerování prázdné tabulky z proteinové úrovně identifikace a kvantifikace workflow. Tvorba grafů z prázdné tabulky vedlo k chybám. Z toho důvodu jsem přidala v nodech generující grafy kontrolu, zda vstupní data nejsou prázdná a pokud ano, je

vytvořen prázdný graf. Díky tomu workflow může bezchybně pokračovat a uložit získané metriky do databáze, a poté zpracovávat nové soubory.

Soubor se špatnou zkratkou názvu přístroje způsobil, že nebyla vybrána žádná z větví workflow s přístrojem, ale větev, která jen zaznamená do .log souboru, že daný soubor nebyl zpracován, což patří mezi očekávané chování. Následně data nejsou zaznamenána do databáze, protože nebyla ani žádná data získána. Workflow se poté vrací na začátek ke zjištění stavu sledované složky.

Chyby v názvu souboru na pozicích zkratky standardu pro kontrolu kvality a jeho množství nezpůsobilo žádné problémy se zpracováním dat a ani s celkovým zápisem do databáze. Chybné hodnoty z názvu do databáze zaznamenány nejsou.

Workflow po malé úpravě kontrolních podmínek zvládlo automaticky zpracovat simulovaná data, získat metriky a uložit do databáze. Při chybě v části zpracování dat nebo zápisu do databáze zaznamenalo nezpracování dat v .log souboru.

Testování s reálnými daty

Soubor s reálnými daty byl bez problémů zpracován a získané metriky byly uložený do databáze. Workflow dokázalo identifikovat 7333 PSM a 2677 proteinů při 1% FDR na PSM úrovni za 13 minut. Software Mascot, který je aktuálně používán pro semiautomatické zpracování QC analýz identifikoval 5449 PSM a 1619 proteinů při 1% FDR na PSM úrovni. Mascot zvládl samotné databázové prohledání během přibližně 30 s, ovšem spolu s přípravou MS dat pro databázové hledání a zpracování výsledku trval proces přibližně 8 minut. Nutno podotknout, že výstupem byla pouze informace o počtu a kvalitě identifikovaných proteinů, bez kvantitativního zhodnocení.

4.3 Zhodnocení workflow

Navržené workflow, které bude využíváno v rámci Centrální laboratoře Proteomika ke kontrole kvality hmotnostně spektrometrických dat, bylo vytvořeno v prostředí KNIME s použitím knihovny OpenMS, která obsahuje funkce pro práci s hmotnostně spektrometrickými daty.

4.3.1 Hodnocení výběru knihoven a programu KNIME

Navržení workflow v KNIME je výhodné, protože je uživatelsky velmi přívětivý. Umožňuje snadnou práci s daty a je velmi flexibilní z pohledu rozšiřování workflow, díky velkému množství zabudovaných funkcí, ale i možnosti doinstalování mnoha

funkcí. Práci s KNIME usnadňuje i dostupná dokumentace a diskuzní fórum s poměrně velkým uživatelským zázemím. Z toho důvodu, může podstatnou část workflow upravovat kdokoli, i bez znalostí z informatického pozadí. K tomu přispívá možnost využití v KNIME knihovny OpenMS, které má z pohledu hmotnostní spektrometrie taktéž zdokumentovanou rozsáhlou a komplexní funkcionalitu s možností jejího nastavení s pomocí parametrů prostřednictvím formulářů u nabízených funkcionalit.

Workflow je díky této široké nabídce navrhnuto podle potřeb Centrální laboratoře Proteomika, aniž by k celému popsanému procesu byl KNIME potřeba kombinovat s jinými softwary a navíc jejich použití je volně dostupné. Z tohoto pohledu je spojení KNIME a OpenMS poměrně výhodné oproti jiným softwarům a knihovnám, které většinou nenabízí celou škálu funkcí a algoritmů a nebo jsou komerční. Ostatní knihovny (v Python nebo R) většinou poskytují algoritmy jen zvlášť pro identifikaci, kvantifikaci nebo práci s metrikami, ale většinou nabídka není široká (např. v algoritmech pro identifikaci peptidů pomocí databázového prohledávání). Také by bylo nutné kombinovat více knihoven, které by mohly dohromady zaštítit celé zpracování dat a získání metrik.

Z druhé strany pohledu použití KNIME má i několik nevýhod. Mezi ně patří složitější předávání dat skrz celé workflow, které je omezené na vstupní, výstupní porty z nodů a na jednu proměnnou *flow_variables* (odpovídající v Python slovníku), ve které uložená data mohou být nepřehledná. Z pohledu programátora může být KNIME omezující – například funkce vytvořené v jednom skriptovacím nodu nelze volat v druhém a je nutné znova napsání funkce. V případě rozsáhlejších funkcí je však možnost napsání např. vlastních python modulů a jejich import v jednotlivých nodech v rámci celého workflow. Další nevýhodou je, že zpracování velkých dat může být poměrně paměťově náročné, protože KNIME si ukládá u některých nodů soubory po dobu běhu workflow.

4.3.2 Hodnocení z pohledu využitelnosti

Workflow pro automatickou kontrolu kvality hmotnostně spektrometrických dat je navržené tak, aby nahradilo komerční program používaný v Centrální laboratoři Proteomika a aby celou část zpracování provedlo samostatně a bez nutnosti uživatelského vstupu a uživatel měl na konci přístup k uloženým datům v databázi z měřených standardů pro kontrolu kvality. Workflow bude používané ve skupině na denní bázi.

Workflow je navržené tak, aby bylo volně dostupné i pro další uživatele. Umožňuje to použití otevřených softwarů a jejich dostupné dokumentace, ale i jednoduchost použití KNIME, ve kterým si uživatel workflow může upravit dle vlastních

potřeb. K workflow pro automatickou kontrolu dat je k dispozici i malé worflow pro vytvoření databáze, kterou uživatel může použít v navržené podobě, ale i ji upravit podle vlastní představy.

S databází lze snadno pracovat v KNIME pomocí databázových nodů. Je tedy možné v KNIME získat data, která jsou v databází uložena, ale i samotnou databázi upravovat, aniž by byl potřeba jiný software.

Oproti existujícím řešením (např. OpenMS workflow [33]) workflow nabízí celkovou automatičnost, nástavba nad získáním metrik v ukládání dat do souborové databáze, získání metrik z proteinové úrovně, ale i zpracování dat z nového formátu .mzQC.

Závěr

Teoretická část diplomové práce se věnuje teorii kontroly kvality proteomického experimentu s využitím hmotnostní spektrometrie. První část práce popisuje experiment od přípravy vzorku přes hmotnostní spektrometrii až po zpracování dat a přibližuje problematiku zdrojů variability v měření v každé části experimentu a přibližuje možnosti kontroly kvality. Důraz je kladen na zpracování dat z měření, protože je to podstatou pro získání informací z hrubých dat nejen o zkoumaném vzorku, ale i o průběhu experimentu. Tyto informace je možné následně zpracovat jako metriky pro kontrolu kvality dat. Metriky umožňují kvantifikovat variabilitu ve výsledcích experimentu a časové trendy, což může být následně použito k optimalizaci experimentu a odhalení počínajících technických problémů.

Pro praktickou část byla provedena rešerše možností zpracování proteomických dat v softwaru KNIME a návrh metodiky kontroly kvality hmotnostně spektrometrických dat.

Podle tohoto návrhu bylo implementováno workflow, které bude používáno v Centrální laboratoři Proteomika. Workflow vytvořené v softwaru KNIME s naimportovanou knihovnou OpenMS automaticky zpracovává data z hmotnostního spektrometru. Na peptidové a proteinové úrovni jsou ze vstupního .mzML souboru získány identifikace (s použitím *MSFGPlusAdapter* a *PIA Analysis* nodů) a kvantifikace (pomocí nodů *FeatureFinderCentroided* a *ProteinQuantifier*). Z obou úrovní jsou získány vybrané metriky pro hodnocení kvality dat. Metriky jsou vybrány ze souboru .mzQC pomocí QC CV identifikátorů na peptidové úrovni a na proteinové jsou vybrány z výstupních dat nodu *ProteinQuantifier*. Z některých metrik jsou vytvořeny grafy, které jsou i s metrikami uloženy do databáze. Pro zápis metrik do databáze je jako primární klíč použitý právě odpovídající identifikátor z QC CV. S metrikami jsou do databáze uložený veškerá potřebná metadata o měření.

Workflow tedy slouží k tomu, aby uživatel pro množství změřených vzorků kontroly kvality na hmotnostním spektrometru měl bezpracně k dispozici vybrané metriky v databázi. S daty v databázi následně může volně pracovat a přizpůsobit si výběr metrik a metadat do tabulky dle vlastních potřeb. Výběr může být třeba pro kontrolu kvality z celé sady metrik z posledního měření pro aktuální kontrolu, ale výběr třeba jedné metriky z posledního roku může ukázat průběh kvality v dlouhodobém hledisku.

Literatura

1. BITTREMIEUX, Wout; TABB, David L; IMPENS, Francis; STAES, An; TIMMERMAN, Evy; MARTENS, Lennart; LAUKENS, Kris. Quality control in mass spectrometry-based proteomics. *Mass Spectrometry Reviews*. 2018, roč. 37, č. 5, s. 697–711.
2. ZHU, Wenhong; SMITH, Jeffrey W; HUANG, Chun-Ming. Mass spectrometry-based label-free quantitative proteomics. *Journal of Biomedicine and Biotechnology*. 2009, roč. 2010.
3. BANTSCHEFF, Marcus; LEMEER, Simone; SAVITSKI, Mikhail M; KUSTER, Bernhard. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and bioanalytical chemistry*. 2012, roč. 404, č. 4, s. 939–965.
4. PIEHOWSKI, Paul D; PETYUK, Vladislav A; ORTON, Daniel J; XIE, Fang; MOORE, Ronald J; RAMIREZ-RESTREPO, Manuel; ENGEL, Anzhelika; LIEBERMAN, Andrew P; ALBIN, Roger L; CAMP, David G et al. Sources of technical variability in quantitative LC–MS proteomics: human brain tissue sample analysis. *Journal of proteome research*. 2013, roč. 12, č. 5, s. 2128–2137.
5. KARTY, Jonathan A; IRELAND, Marcia ME; BRUN, Yves V; REILLY, James P. Artifacts and unassigned masses encountered in peptide mass mapping. *Journal of Chromatography B*. 2002, roč. 782, č. 1-2, s. 363–383.
6. SUN, Shisheng; ZHOU, Jian-Ying; YANG, Weiming; ZHANG, Hui. Inhibition of protein carbamylation in urea solution using ammonium-containing buffers. *Analytical biochemistry*. 2014, roč. 446, s. 76–81.
7. MA, Ze-Qiang; DASARI, Surendra; CHAMBERS, Matthew C; LITTON, Michael D; SOBECKI, Scott M; ZIMMERMAN, Lisa J; HALVEY, Patrick J; SCHILLING, Birgit; DRAKE, Penelope M; GIBSON, Bradford W et al. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *Journal of proteome research*. 2009, roč. 8, č. 8, s. 3872–3881.
8. MAGDELDIN, Sameh; MORESCO, James J; YAMAMOTO, Tadashi; YATES III, John R. Off-line multidimensional liquid chromatography and auto sampling result in sample loss in LC/LC–MS/MS. *Journal of proteome research*. 2014, roč. 13, č. 8, s. 3826–3836.

9. RUDNICK, Paul A; CLAUSER, Karl R; KILPATRICK, Lisa E; TCHEKHOVSKOI, Dmitrii V; NETA, Pedatsur; BLONDER, Nikša; BILLHEIMER, Dean D; BLACKMAN, Ronald K; BUNK, David M; CARDASIS, Helene L et al. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Molecular & Cellular Proteomics*. 2010, roč. 9, č. 2, s. 225–241.

10. ZUBAREV, Roman A. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics*. 2013, roč. 13, č. 5, s. 723–726.

11. KÖCHER, Thomas; PICHLER, Peter; SWART, Remco; MECHTLER, Karl. Quality control in LC-MS/MS. *Proteomics*. 2011, roč. 11, č. 6, s. 1026–1030.

12. BURKHART, Julia Maria; PREMSLER, Thomas; SICKMANN, Albert. Quality control of nano-LC-MS systems using stable isotope-coded peptides. *Proteomics*. 2011, roč. 11, č. 6, s. 1049–1057.

13. BEREMAN, Michael S. Tools for monitoring system suitability in LC MS/MS centric proteomic experiments. *Proteomics*. 2015, roč. 15, č. 5-6, s. 891–902.

14. DEUTSCH, Eric W. Mass spectrometer output file format mzML. In: *Proteome bioinformatics*. Springer, 2010, s. 319–331.

15. NOOR, Zainab; AHN, Seong Beom; BAKER, Mark S; RANGANATHAN, Shoba; MOHAMEDALI, Abidali. Mass spectrometry–based protein identification in proteomics—a review. *Briefings in bioinformatics*. 2021, roč. 22, č. 2, s. 1620–1638.

16. AGGARWAL, Suruchi; YADAV, Amit Kumar. False discovery rate estimation in proteomics. In: *Statistical Analysis in Proteomics*. Springer, 2016, s. 119–128.

17. NESVIZHSKII, Alexey I; AEBERSOLD, Ruedi. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics*. 2005, roč. 4, č. 10, s. 1419–1440.

18. HUANG, Ting; WANG, Jingjing; YU, Weichuan; HE, Zengyou. Protein inference: a review. *Briefings in bioinformatics*. 2012, roč. 13, č. 5, s. 586–614.

19. BANTSCHEFF, Marcus; SCHIRLE, Markus; SWEETMAN, Gavain; RICK, Jens; KUSTER, Bernhard. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*. 2007, roč. 389, č. 4, s. 1017–1031.

20. WALZER, Mathias; PERNAS, Lucia Espona; NASSO, Sara; BITTREMIEUX, Wout; NAHNSEN, Sven; KELCHTERMANS, Pieter; PICHLER, Peter; TOORN, Henk WP van den; STAES, An; VANDENBUSSCHE, Jonathan et al. qcML: an exchange format for quality control metrics from mass spectrometry experiments. *Molecular & Cellular Proteomics*. 2014, roč. 13, č. 8, s. 1905–1913.

21. BITTREMIEUX, Wout; VALKENBORG, Dirk; MARTENS, Lennart; LAUKENS, Kris. Computational quality control tools for mass spectrometry proteomics. *Proteomics*. 2017, roč. 17, č. 3-4, s. 1600159.

22. PFEUFFER, Julianus; SACHSENBERG, Timo; ALKA, Oliver; WALZER, Mathias; FILLBRUNN, Alexander; NILSE, Lars; SCHILLING, Oliver; REINERT, Knut; KOHLBACHER, Oliver. OpenMS–A platform for reproducible analysis of mass spectrometry data. *Journal of biotechnology*. 2017, roč. 261, s. 142–148.

23. BITTREMIEUX, Wout. spectrum_utils: A Python package for mass spectrometry data processing and visualization. *Analytical chemistry*. 2019, roč. 92, č. 1, s. 659–661.

24. BALD, Till; BARTH, Johannes; NIEHUES, Anna; SPECHT, Michael; HIPPLER, Michael; FUFEZAN, Christian. pymzML—Python module for high-throughput bioinformatics on mass spectrometry data. *Bioinformatics*. 2012, roč. 28, č. 7, s. 1052–1053.

25. LEVITSKY, Lev I; KLEIN, Joshua A; IVANOV, Mark V; GORSHKOV, Mikhail V. Pyteomics 4.0: five years of development of a Python proteomics framework. *Journal of proteome research*. 2018, roč. 18, č. 2, s. 709–714.

26. GATTO, Laurent; GIBB, Sebastian; RAINER, Johannes. MSnbase, Efficient and Elegant R-Based Processing and Visualization of Raw Mass Spectrometry Data. *Journal of Proteome Research*. 2020, roč. 20, č. 1, s. 1063–1069.

27. CHOI, Meena; CHANG, Ching-Yun; CLOUGH, Timothy; BROUDY, Daniel; KILLEEN, Trevor; MACLEAN, Brendan; VITEK, Olga. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*. 2014, roč. 30, č. 17, s. 2524–2526.

28. GIBB, Sebastian; STRIMMER, Korbinian. MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*. 2012, roč. 28, č. 17, s. 2270–2271.

29. DOGU, Eralp; TAHERI, Sara Mohammad; OLIVELLA, Roger; MARTY, Florian; LIENERT, Ian; REITER, Lukas; SABIDO, Eduard; VITEK, Olga. MSstatsQC 2.0: R/Bioconductor package for statistical quality control of mass spectrometry-based proteomics experiments. *Journal of proteome research*. 2018, roč. 18, č. 2, s. 678–686.

30. STRATTON, Kelly G; WEBB-ROBERTSON, Bobbie-Jo M; MCCUE, Lee Ann; STANFILL, Bryan; CLABORNE, Daniel; GODINEZ, Iobani; JOHANSEN, Thomas; THOMPSON, Allison M; BURNUM-JOHNSON, Kristin E; WATERS, Katrina M et al. Pmartr: Quality control and statistics for mass spectrometry-based biological data. *Journal of proteome research*. 2019, roč. 18, č. 3, s. 1418–1425.

31. WEN, B; GATTO, L. proteoQC: An R package for proteomics data quality control. *R package version 1.3*. 2014, roč. 2.

32. DEUTSCH, Eric W; MENDOZA, Luis; SHTEYNBERG, David; FARRAH, Terry; LAM, Henry; TASMAN, Natalie; SUN, Zhi; NILSSON, Erik; PRATT, Brian; PRAZEN, Bryan et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics*. 2010, roč. 10, č. 6, s. 1150–1159.

33. WEISSER, Hendrik; NAHNSEN, Sven; GROSSMANN, Jonas; NILSE, Lars; QUANDT, Andreas; BRAUER, Hendrik; STURM, Marc; KENAR, Erhan; KOHLBACHER, Oliver; AEBERSOLD, Ruedi et al. An automated pipeline for high-throughput label-free quantitative proteomics. *Journal of proteome research*. 2013, roč. 12, č. 4, s. 1628–1644.

34. GATTO, Laurent; BRECKELS, Lisa M; NAAKE, Thomas; GIBB, Sebastian. Visualization of proteomics data using R and bioconductor. *Proteomics*. 2015, roč. 15, č. 8, s. 1375–1389.

35. USZKOREIT, Julian; PEREZ-RIVEROL, Yasset; EGGERS, Britta; MARCUS, Katrin; EISENACHER, Martin. Protein inference using PIA workflows and PSI standard file formats. *Journal of proteome research*. 2018, roč. 18, č. 2, s. 741–747.

36. SILVA, Jeffrey C; GORENSTEIN, Marc V; LI, Guo-Zhong; VISSERS, Johannes PC; GEROMANOS, Scott J. Absolute Quantification of Proteins by LCMSE: A Virtue of Parallel ms Acquisition* S. *Molecular & Cellular Proteomics*. 2006, roč. 5, č. 1, s. 144–156.

37. HIPP, Richard D. *SQLite*. 2020. Ver. 3.31.1. Dostupné také z: `https://www.sqlite.org/index.html`.

Seznam symbolů a zkratek

LC-MS	kapalinová chromatografie spojená s hmotnostní spektrometrií
MS	hmotnostní spektrometrie
MS/MS	tandemová hmotnostní spektrometrie
LC	kapalinová chromatografie
m/z	hodnota hmotnost ku náboji
QC	kontrola kvality
TIC	totální iontový chromatogram
PSM	shoda spektra k peptidu
QC CV	kontrolovaný slovník kontroly kvality
PSI-MS	Proteomická standardová iniciativa hmotnostní spektrometrie
OBO	Otevřená Biologická a Biomedicínská Ontologie
JSON	JavaScriptový objektový zápis
PK	primární klíč
FK	cizí klíč
DB	databáze
SQL	standardizovaný strukturovaný dotazovací jazyk pro databáze
HeLA	buněčná linie lidských epiteliálních buněk

A Návod ke spuštění workflow

Tato příloha slouží jako krátký návod ke spuštění workflow v programu KNIME.

A.1 Potřebné soubory a software

Ke spuštění workflow je potřebný samotný program KNIME (workflow byl vytvořen ve verzi 4.5.2), který lze stáhnout z oficiálních stránek KNIME: `https://www.knime.com/downloads/download-knime`.

V KNIME je nutné mít nastavenou cestu k Pythonu (návod je k dispozici na stránce: *Configure the KNIME Python Integration* a instalované *KNIME Extensions – Bioinformatics & NGS* pro možnost použití OpenMS funkcí (návod je zde: *KNIME Integreations*). Workflow používá verzi Pythonu 3.8 a jeho knihovny Pandas, os, time, matplotlib, numpy, json a datetime a verzi OpenMS 2.6.0.

Další soubory jsou v příloze diplomové práce. V příloze je workflow „definice_databaze.knwf“ (nutné k definování struktury databáze), workflow „kontrola_kvality .knwf“ (samotný workflow pro automatickou kontrolu kvality hmotnostně spektrometrických dat) a složku *DATA* obsahující složku „mzml_data“ s .mzML soubory („QC_Lumos_HeLa_100ng.mzML“), .fasta soubory pro identifikaci databázovým prohledáváním („cRAP_universal_181122.fasta“, „UniProtKB_Human_can_20180912.fasta“) a složka „database_data“ obsahující databázový soubor „quality_control.sqlite“ s definovanou databází.

A.2 Příprava workflow ke spuštění

Workflow je do KNIME možné importovat postupem: *File – Import KNIME Workflow... – Select file – jmeno_workflow.knwf*. Do KNIME je nutné naimportovat workflow „kontrola_kvality.knwf“ a je možné naimportovat „definice_databaze.knwf“.

Workflow je následně potřeba otevřít ze záložky *KNIME Explorer*. Pokud by nebyl k dispozici soubor už s definovanou strukturou databáze, bylo by nutné prvně otevřít a spustit workflow „definice_databaze“. Takto postačí otevřít workflow „kontrola_kvality“ a v prvním metanodu „Složka a formát souboru ke sledování“ nastavit cesty k datům, databázi a Pythonu. Poté je ještě nutné u všech metanodů rozdělených pro přístroje nastavit cestu k .fasta souborům. Ve workflow je možné se k tomutu nastavení dostat tímto postupem: „Zpracování souboru a uložení do databáze“ – „Získání qc metrik“ – „Identifikace a hledání features“ – „Proteinová databáze FASTA“ – vložit do nodu cestu k oběma .fasta souborům poskytnutých v příloze.

A.3 Spuštění workflow

Na závěr stačí workflow spustit vybráním posledního nodu *Recursice Loop End* a stisknutím klávesy F7 se celé workflow spustí. Pro zpracování souboru .mzML je nutné ho do sledované složky po spuštění přidat.