Chem. Listy *101*, 279–286 (2007) Referát

# MĚŘENÍ PROTEOMICKÝCH DAT HMOTNOSTNÍ SPEKTROMETRIÍ A JEJICH BIOINFORMATICKÁ INTERPRETACE

**PAVEL ŘEHULKA, HELENA ŘEHULKOVÁ a JOSEF CHMELÍK**

*Ústav analytické chemie Akademie věd ČR, Veveří 97, 602 02 Brno*
*rehulka@iach.cz, rehulkova@iach.cz, chmelik@iach.cz*

Došlo 30.9.05, přepracováno 25.9.06, přijato 26.10.06.

Klíčová slova: hmotnostní spektrometrie, bioinformatika, identifikace bílkovin

## Obsah

1. Úvod
2. Identifikace bílkovin hmotnostní spektrometrií
   2.1. Peptidové mapování
   2.2. Sekvenování peptidů hmotnostní spektrometrií
3. Bioinformatika
   3.1. Databázové vyhledávací programy
   3.2. Skórovací algoritmy pro identifikaci bílkovin peptidovým mapováním

## 1. Úvod

Úspěch při identifikaci bílkovin závisí na kvalitě dat získaných hmotnostní spektrometrií a jejich interpretaci s použitím bioinformatiky. S vývojem jemných ionizačních technik[1–5] se rozvíjely také metody identifikace bílkovin. Směs bílkovin se rozdělí vhodnou separační technikou. Separované bílkoviny se enzymaticky štěpí na směsi peptidů, které se dále analyzují hmotnostní spektrometrií. Nejcitlivější a rychlou metodou je peptidové mapování (peptide mass fingerprinting – PMF)[6–10], ve kterém jsou přesně změřené molekulové hmotnosti peptidů porovnány s teoreticky vypočítanými hmotnostmi peptidů z bílkoviny přítomné v databázi, přičemž toto srovnání je provedeno pro každou bílkovinu v databázi zvlášť. Bílkovina z databáze vykazující největší shodu při použití vhodného skórovacího algoritmu identifikuje analyzovanou bílkovinu. Jinou možností identifikace bílkovin je odvodit krátký sekvenční úsek některého z peptidů přítomných v peptidové směsi na základě jeho fragmentace a použitím jiného vhodného vyhledávacího programu identifikovat bílkovinu v databázi. *De novo* sekvenování celých peptidů se pak používá, pokud selžou předchozí identifikační postupy.

## 2. Identifikace bílkovin hmotnostní spektrometrií

### 2.1. Peptidové mapování

Počátky peptidového mapování spadají do roku 1989, kdy tento přístup poprvé prezentovali Henzel a spol.[11]. Původní Laemmliho myšlenka ukázala, že produkty štěpení bílkoviny jsou charakteristické pro analyzovanou bílkovinu a že peptidové mapování lze provádět gelovou elektroforézou[12]. Henzel a spol.[6] dále rozvedli tuto myšlenku a bílkovinné lyzáty analyzovali hmotnostní spektrometrií. Bílkoviny separované gelovou elektroforézou byly převedeny z gelu na membránu a po jejich redukci a alkylaci byly štěpeny trypsinem nebo bromkyanem. Vzniklé peptidy byly hmotnostně-spektrometricky analyzovány sektorovým hmotnostním analyzátorem s ionizační technikou „fast atom bombardment“ (FAB). Citlivost této metody však byla malá, což zabránilo jejímu využití jako běžného nástroje pro identifikaci bílkovin.

Poté, co se začaly vyrábět první komerční přístroje typu MALDI-TOF MS, nahradila citlivější technika MALDI-TOF MS techniky dříve používané v PMF (cit.[6]), která byla současně vyvinuta i v jiných skupinách[7–10]. Technická zlepšení TOF hmotnostní analýzy (použití zpožděné extrakce[13] a reflektronu[14]) byla příčinou dalšího zvýšení citlivosti a přesnosti analýzy, což mělo za následek rozšíření techniky PMF jako hlavního nástroje pro identifikaci bílkovin na konci minulého století[15].

Dřívější metody PMF prováděly štěpení bílkovin na membráně, na kterou byly přeneseny po jejich separaci gelovou elektroforézou. Později bylo ukázáno, že bílkoviny obarvené stříbrem mohou být štěpeny přímo v gelu bez nutnosti jejich přenosu na membránu a následně pak identifikovány hmotnostní spektrometrií[16]. Současná typická strategie identifikace bílkovin spočívá v separaci bílkovin 2D gelovou elektroforézou, redukci v gelu, alkylaci a štěpení separované bílkoviny, hmotnostně-spektrometrické analýze vzniklých peptidů a bioinformatickém zpracování získaných dat[17].

Další výhodou PMF vedle citlivosti a rychlosti této metody je skutečnost, že pro vlastní analýzu se spotřebuje pouze malá část vzorku (přibližně 5–10 %). To poskytuje možnost využití dalších typů analýzy téhož vzorku (sekvenování peptidů metodou MS/MS, studium posttranslačních modifikací aj.).

## 2.2. Sekvenování peptidů hmotnostní spektrometrií

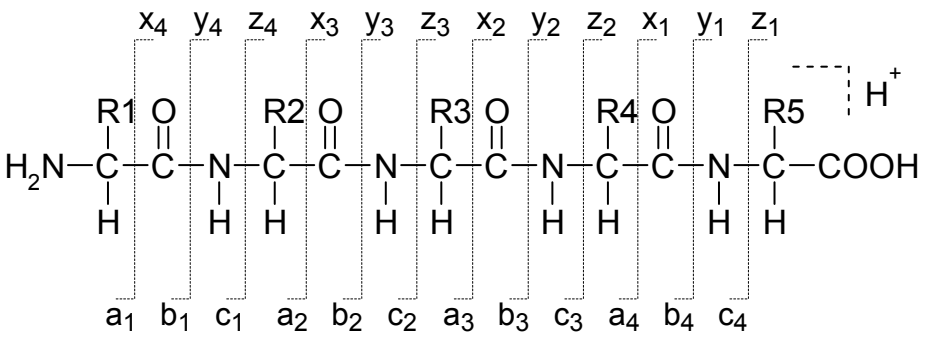
Již v 60. letech minulého století byly rozpoznány možnosti hmotnostní spektrometrie co se týče citlivosti a informačního obsahu ve výsledných datech[18]. První pokusy sekvenování peptidů hmotnostní spektrometrií s elektronovou ionizací (EI) byly provedeny již v roce 1965 (cit.[19,20]). Tento způsob využíval informace ze štěpení malých derivatizovaných peptidů v iontovém zdroji. Zavedení ionizační techniky FAB v roce 1981 (cit.[21,22]) zlepšilo efektivitu ionizace peptidů s vyšší molekulovou hmotností a tato technika pak byla spojena s MS/MS (cit.[23–25]). Největší vliv na rozvoj sekvenování peptidů prostřednictvím MS/MS měla komercializace ionizačních technik ESI a MALDI. Ty jsou používány v kombinaci s mnoha konfiguracemi hmotnostních analyzátorů pro MS/MS (trojitý kvadrupól[26], iontová past[27], kvadrupól/průletový analyzátor (time-of-flight, TOF)[28], TOF/TOF[29] atd.). Tyto techniky jsou velmi citlivé a mnohé z nich mohou být automatizovány pro analýzu velkého počtu vzorků.

Zjištění sekvence peptidů může být provedeno nepřímo metodou PMF. Po úspěšné identifikaci bílkoviny jsou experimentální molekulové hmotnosti peptidů přiřazeny k sekvencím peptidů a tím je určena primární struktura analyzovaných peptidů. V dalším textu se pod pojmem „sekvenování peptidů hmotnostní spektrometrií“ rozumí získání informace o primární struktuře peptidu užitím informace získané z fragmentačního hmotnostního spektra.

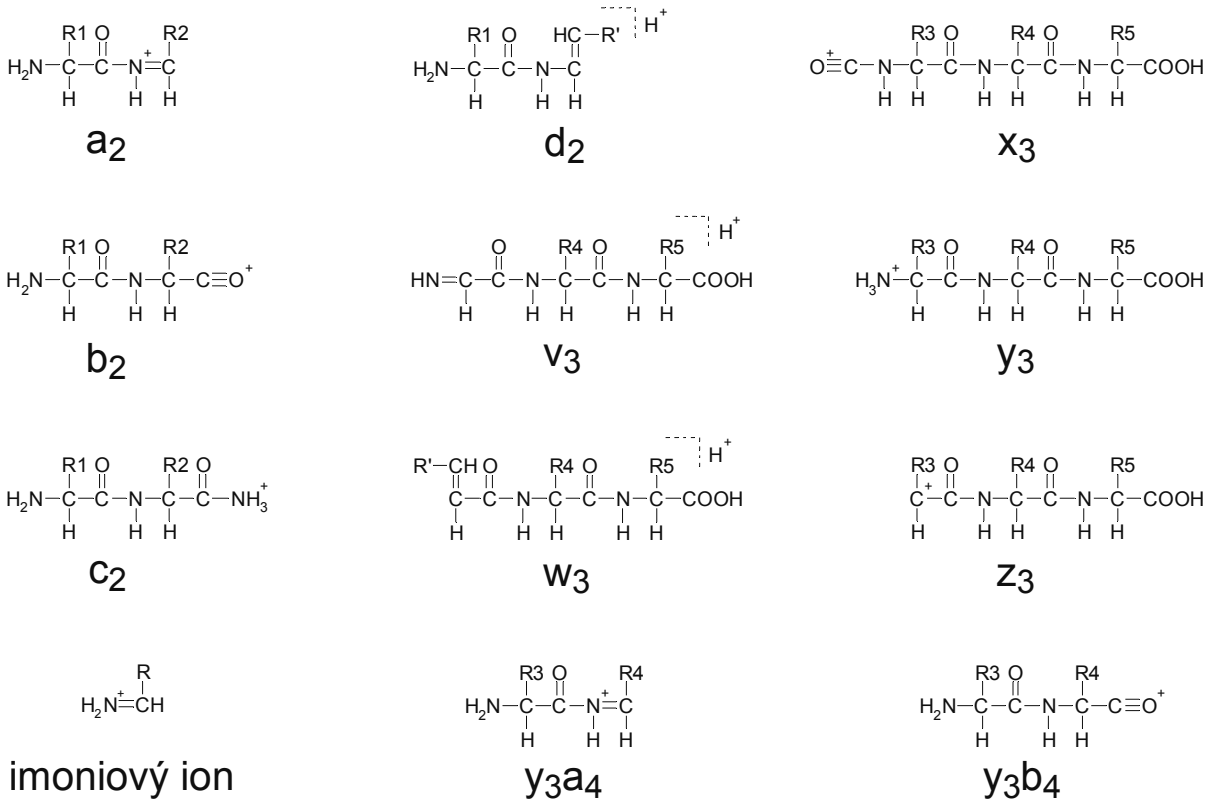
Hlavní myšlenkou sekvenování peptidů hmotnostní spektrometrií je užití fragmentačního spektra pro získání alespoň částečné informace o primární struktuře peptidu. Ta dále s použitím bioinformatických nástrojů slouží k identifikaci bílkoviny. Důležitou skutečností je poměrně dobře definované fragmentační chování peptidů v MS/MS experimentu, kdy nejčastější typy fragmentace probíhají na peptidovém řetězci. Obzvláště při experimentech s nízkoenergetickou kolizně indukovanou disociací (CID – collision-induced dissociation) jsou nejčastějšími typy fragmentových iontů tzv. *a*-, *b*- a *y*-ionty (viz obr. 1 a obr. 2), spojené někdy se ztrátou malé neutrální molekuly ($NH_3$, $H_2O$).

Existuje více způsobů, jak využít informaci získanou z MS/MS analýzy peptidů pro identifikaci bílkovin. Přímý způsob je interpretace fragmentačního spektra ručně a použití získané sekvenční informace k prohledávání databáze bílkovin buď jednoduchým vyhledávacím programem (např. MS-Pattern[30]) nebo programem založeným na vyhledávání podobných sekvencí v databázi bílkovin používajícím vhodné algoritmy (např. FASTA algoritmus[31] na adrese[32] nebo BLAST algoritmus[33] na adrese[34]). Interpretace fragmentačních spekter není jednoduchá, ale použití těchto nástrojů je výhodné zejména v situaci, kdy sekvence analyzované bílkoviny není přítomna v databázi, ale sekvence podobné bílkoviny (např. z jiného organismu) v databázi je. Tím může být analyzovaná bílkovina identifikována na základě vysoké podobnosti s bílkovinou přítomnou v databázi.

Jiný způsob identifikace bílkovin s použitím dat z fragmentačního spektra zavedli Mann a Wilm[35], kteří využili skutečnosti, že většina fragmentačních spekter obsahuje krátkou a snadno stanovitelnou sekvenci fragmentových iontů identifikující krátkou sekvenci primární struktury analyzovaného peptidu. Tato částečná sekvence rozděluje peptid na tři části, kde druhou částí je částečná interpretovaná sekvence peptidu a první a třetí část jsou zbývající peptidové fragmenty. Konstrukce $m_1$TAG$m_3$ se nazývá peptidový sekvenční úsek (peptide sequence tag), kde $m_1$ (resp. $m_3$) je molekulová hmotnost prvního (resp. posledního) fragmentového iontu, na kterém začíná (resp. končí) interpretovaná krátká sekvence označená písmeny „TAG“. Peptidový sekvenční úsek spolu s molekulovou hmotností rodičovského iontu poskytuje specifickou informaci o analyzovaném peptidu. Bílkovinná databáze je pak prohledávána s použitím této specifické informace. Každý peptid odpovídající specificitě enzymu a molekulové hmotnosti analyzovaného peptidu ve specifikované toleranci je testován na přítomnost identifikovaného úseku v sekvenci peptidu na pozici dané údaji $m_1$ a $m_3$ (přičemž je obvykle testována možnost jak pro sérii *b*-iontů, tak *y*-iontů). Pro jednoznačnou identifikaci bílkoviny obvykle stačí, když je interpretovaný sekvenční úsek delší než dvě aminokyseliny.




Obr. 1. **Struktura lineárního protonovaného peptidu a nomenklatura jeho fragmentových iontů**



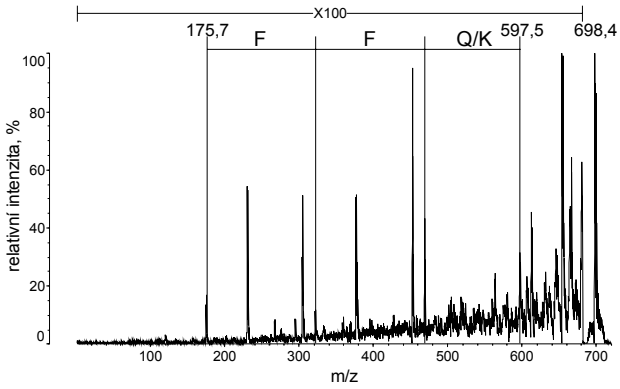

Obr. 2. **Struktura fragmentových iontů lineárního protonovaného peptidu**

Příklad fragmentačního spektra s krátkou interpretovanou sekvencí analyzovaného peptidu je ukázán na obr. 3. Peptidový sekvenční úsek získaný vyhodnocením tohoto spektra byl zadán do vyhledávacího databázového programu PeptideSearch[36]. Výsledkem byla jednoznačná identifikace β-amylasy z ječmene.

Jedním z proteomických nástrojů v souboru programů Protein Prospector[37] je program nazývaný MS-Tag (cit.[38]), který byl původně vyvinut Clauserem a Bakerem[39] pro interpretaci PSD spekter, jež obvykle neobsahují dostatečné množství informace pro kompletní *de novo* sekvenaci peptidu, ale stále je v nich dost specifické informace pro identifikaci bílkoviny databázovým vyhledáváním. Program vybírá z databáze peptidy vzniklé teoretickým štěpením zvoleným enzymem a současně odpovídající teoretickou molekulovou hmotností ve zvoleném rozsahu experimentální hodnotě molekulové hmotnosti peptidu. Peptid, který toto splňuje, je teoreticky fragmentován a experimentálně získané hmotnosti fragmentových iontů jsou porovnány s takto vytvořeným seznamem teoretických hodnot. Výsledkem databázového hledání je seznam bílkovin, jež prošly všemi omezujícími kritérii (včetně počtu odpovídajících identifikovaných fragmentových iontů). Tento seznam je seřazen podle nejmenšího počtu neidentifikovaných fragmentových iontů a za identifikovanou bílkovinu je považována bílkovina umístěná na prvním místě. Je-li kvalita fragmentačního spektra dostatečná, bílkovina je přítomna v databázi a vyhledávací parametry jsou správně nastaveny, pak je bílkovina identifikována jednoznačně. Aplikací programu MS-Tag na data získaná z fragmentačního spektra na obr. 3 byla opět identifikována β-amylasa.

Současná verze programu MS-Tag může interpretovat nejen data z PSD experimentu, ale i výsledky získané jinými fragmentačními technikami. MS-Tag lze též použít v případě, kdy spektrum obsahuje dostatečné množství informace pro interpretaci *de novo*; je však třeba, aby homologický peptid byl přítomen v databázi.

Sofistikovaný program SEQUEST využívá pro databázové vyhledávání vzájemné korelace mezi experimentálním a teoretickým fragmentačním spektrem vytvořeným ze sekvence peptidu přítomného v databázi[40,41]. Tento program nejprve zjednoduší experimentální fragmentační spektrum, aby nedocházelo k nesprávným identifikacím bílkovin. Maximální počet píků v experimentálním fragmentačním spektru je 200 a tento seznam je v prvním vyhledávacím kole porovnáván s teoretickými seznamy fragmentových iontů peptidů v databázi, které navíc mají odpovídající enzymovou specificitu a molekulovou hmot-




Obr. 3. **PSD fragmentační spektrum trypsinového peptidu bílkoviny extrahované vodou z obilek ječmene (*Hordeum vulgare L.*) a separované za použití 1D gelové elektroforézy;** v horní části obrázku je znázorněn interpretovaný sekvenční úsek analyzovaného peptidu

nost. Peptidy jsou ohodnoceny zjednodušeným skórovacím algoritmem, který závisí na součtu intenzit iontů, souvislosti iontů v identifikované sekvenci a podílu identifikovaných fragmentových iontů. Pro každou z 500 sekvencí peptidů s nejvyšším skóre po prvním kole vyhledávání je vytvořeno teoretické fragmentační spektrum (odvozené podle speciálních pravidel), které je pak vzájemně korelováno s upraveným experimentálním fragmentačním spektrem rychlou Fourierovou transformací. Výsledné skóre pak popisuje, jak je experimentální spektrum podobné spektru odvozenému teoreticky pro určitou peptidovou sekvenci.

Program SEQUEST byl původně navržen pro analýzu fragmentačních spekter nízkoenergetických CID experimentů[40,42,43]. Program byl dále upraven též pro analýzu spekter z analýzy PSD (cit.[44]), vysokoenergetických CID analýz[45] a pro identifikaci modifikovaných peptidů[42,44,46]. Nevýhodou je, že skórovací algoritmus neposkytuje žádnou absolutní míru spolehlivosti výsledku identifikace, což nakonec vyžaduje přítomnost pracovníka posuzujícího relevantnost výsledků. Řešením této situace může být např. aplikace skórovací rutiny (SEQUEST-NORM) normalizující XCorr hodnoty pro jejich nezávislost na délce sekvence peptidu a velikosti použité databáze[47] nebo použití neuronových sítí spolu s vhodným statistickým modelem[48].

## 3. Bioinformatika

Obrovská množství experimentálních dat vyžadují sofistikované metody pro jejich interpretaci. Současný rychlý rozvoj v oblasti informatiky umožňuje provádět analýzy, které by ještě před několika lety nebyly možné.

Pro identifikaci bílkovin jsou nejdůležitějším zdrojem databáze DNA a bílkovinných sekvencí, které jsou využívány různými vyhledávacími programy. Existují tři druhy databází: primární nukleotidové sekvenční databáze, komplexní databáze sekvencí bílkovin a administrované databáze sekvencí bílkovin. Příkladem nukleotidové sekvenční databáze je GenBank[49]. Tato databáze je z větší části výsledkem přímého vložení nukleotidových sekvencí jednotlivými experimentátory. Další sekvence v databázi jsou získány z vědecké literatury. Přímé vkládání sekvencí do databáze způsobuje rychlý růst její velikosti, ale cenou za to je poměrně vysoká četnost chyb. Primárně zodpovědnými osobami za editaci a opravu sekvenčních dat jsou sami experimentátoři vkládající odpovídající data do databáze. TrEMBL (cit.[50]) je příklad komplexní databáze bílkovin sestavené překladem nukleotidové sekvenční databáze. Databáze TrEMBL je částečně odlišná od jiných databází téhož typu, neboť slouží jako dočasný archiv před vložením dané položky do bílkovinné databáze SwissProt[50]. SwissProt je dobrým příkladem administrované databáze sekvencí bílkovin. Jednotlivé položky pro databázi SwissProt jsou vyhodnocovány skupinou vědců předtím, než jsou do této databáze vloženy. To má sice za následek nízký celkový počet sekvencí v databázi, ale výhodou zůstává nízká chybovost a nadbytečnost dat a vysoký stupeň anotace jednotlivých položek.

### 3.1. Databázové vyhledávací programy

Databázové vyhledávací programy jsou různé programy, které akceptují na vstupu experimentální data (event. zpracovaná) a při průchodu vybranou databází specificky porovnávají tato vstupní data s každou položkou v databázi. Položky, které splňují určitá kritéria, jsou vypsány na výstup jako výsledek programu.

Existují tři hlavní kategorie databázových vyhledávacích programů pro identifikaci bílkovin. Jsou to programy využívající jakožto vstup pro databázové vyhledávání:

- molekulové hmotnosti peptidů,
- hmotnosti fragmentových iontů z fragmentačního spektra,
- interpretované sekvence aminokyselin.

Tabulka I ukazuje přehled některých webových adres s různými volně dostupnými programy pro identifikaci bílkovin.

### 3.2. Skórovací algoritmy pro identifikaci bílkovin peptidovým mapováním

Nedílnou součástí databázových vyhledávacích programů je skórovací schéma, které funguje jako kritérium pro správné seřazení databázových výsledků podle míry spolehlivosti identifikace bílkoviny. Obzvláště důležité je použití vhodného skórovacího algoritmu v programu identifikujícím bílkoviny PMF. Nejjednodušším kritériem při identifikaci bílkovin pomocí PMF je počet identifikovaných peptidů z experimentálních hodnot ve srovnávané primární struktuře bílkoviny v databázi (čím vyšší počet

Tabulka I
Databázové vyhledávací programy dostupné na internetu

| Název programu | Typ programu | Internetová adresa |
|---|---|---|
| | | **Protein Prospector** (http://prospector.ucsf.edu) |
| MS-Fit | PMF | http://prospector.ucsf.edu/ucsfhtml4.0/msfit.htm |
| MS-Tag | MS/MS | http://prospector.ucsf.edu/ucsfhtml4.0/mstagfd.htm |
| MS-Seq | TAG | http://prospector.ucsf.edu/ucsfhtml4.0/msseq.htm |
| MS-Pattern | TAG | http://prospector.ucsf.edu/ucsfhtml4.0/mspattern.htm |
| | | **Mascot** (http://www.matrixscience.com) |
| PMF | PMF | http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=PMF |
| MS/MS Ion Search | MS/MS | http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=MIS |
| Sequence Query | TAG | http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=SQ |
| | | **PeptideSearch** (http://www.narrador.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html) |
| Peptide Search | PMF | http://www.narrador.embl-heidelberg.de/GroupPages/PageLink/peptidesearch/Services/PeptideSearch/FR_PeptideSearchFormG4.html |
| Peptide Pattern | TAG | http://www.narrador.embl-heidelberg.de/GroupPages/PageLink/peptidesearch/Services/PeptideSearch/FR_PeptidePatternFormG4.html |
| Sequence Only | TAG | http://www.narrador.embl-heidelberg.de/GroupPages/PageLink/peptidesearch/Services/PeptideSearch/FR_SequenceOnlyFormG4.html |
| | | **PROWL** (http://prowl.rockefeller.edu) |
| ProFound | PMF | http://prowl.rockefeller.edu/profound_bin/WebProFound.exe |
| PepFrag | MS/MS | http://prowl.rockefeller.edu/prowl/pepfragch.html |
| | | **ExPASy** (http://www.expasy.org/tools) |
| Aldente | PMF | http://www.expasy.org/tools/aldente |
| Popitam | MS/MS | http://www.expasy.org/tools/popitam |
| TagIdent | TAG | http://www.expasy.org/tools/tagident.html |
| MultiIdent | TAG PMF | http://www.expasy.org/tools/multiident |
| | | **Některé další programy** |
| Phenyx | MS/MS | http://phenyx.vital-it.ch/pwi |
| OMSSA | MS/MS | http://pubchem.ncbi.nlm.nih.gov/omssa |
| GFS | PMF MS/MS | http://gfs.unc.edu |
| PepMAPPER | PMF | http://wolf.bms.umist.ac.uk/mapper |
| FASTA | TAG | http://fasta.bioch.virginia.edu |
| BLAST | TAG | http://www.ncbi.nlm.nih.gov/blast |

PMF – program využívající molekulové hmotnosti peptidů, MS/MS – program využívající hmotnosti fragmentových iontů z fragmentačního spektra, TAG – program využívající interpretované sekvence aminokyselin

nalezených peptidů, tím spolehlivější je identifikace bílkoviny)[6,7,9,10]. Tento typ skórovacího algoritmu je však v mnoha případech nedostačující. Pro spolehlivou identifikaci bílkoviny je v tomto případě nutné, aby se v hmotnostním spektru vyskytovaly signály mnoha peptidů z dané bílkoviny, což v případě nízkomolekulárních bílkovin často není ani teoreticky možné. Často pak dochází k situacím, kdy výsledek databázového vyhledávání obsahuje pouze spoustu náhodných identifikací bílkovin poskytujících velké množství teoretických peptidů, což pak

značně ztěžuje proces identifikace.

Dobrý skórovací algoritmus, který bere do úvahy přirozené vlastnosti bílkovin, byl poprvé použit v programu MOWSE (cit.[8]). Základem skórovací funkce je matice ($f_{i,j}$) přiřazená zvlášť pro určitý enzym a zvolenou databázi. Řádky této matice představují 10 kDa intervaly pro molekulové hmotnosti intaktních bílkovin a sloupce pak 100 Da intervaly pro molekulové hmotnosti peptidů vzniklých štěpením bílkovin v databázi zvoleným enzymem. Hlavním důvodem pro seskupení molekulových hmotností peptidů tímto způsobem je různá četnost těchto hmotností pro jejich různé hodnoty (nízkomolekulární peptidy jsou četnější než vysokomolekulární), která je také navíc závislá na molekulové hmotnosti bílkoviny. Frekvenční hodnota prvku matice ($F_{i,j}$) je spočítána jako podíl počtu peptidů v odpovídajícím prvku matice ($N_{i,j}$) a součtu všech peptidů pro celý 10 kDa interval molekulové hmotnosti bílkoviny (suma všech hodnot v řádku matice, $\Sigma_j N_{i,j}$):

$$\left(F_{i,j}\right) = \frac{N_{i,j}}{\sum_j N_{i,j}}$$

Tyto frekvenční hodnoty prvků matice jsou pak pro každý 10 kDa interval (řádek matice) normalizovány vzhledem k největší frekvenční hodnotě prvků matice $F_{i,\max} = \max_j F_{i,j}$ a výsledek pro každý prvek matice je $f_{i,j} = F_{i,j} / F_{i,\max}$, což je reálné číslo mezi 0 a 1. Vlastní databázové vyhledávání probíhá následovně: každá položka v databázi (tj. molekulové hmotnosti peptidů teoreticky naštípané sekvence bílkoviny) je porovnána s naměřenými daty. Experimentální hmotnosti peptidů neodpovídající žádné teoretické hodnotě jsou ignorovány. Pro každou experimentální molekulovou hmotnost odpovídající některé z teoretických molekulových hmotností peptidů porovnávané bílkoviny je přiřazena normalizovaná frekvenční hodnota odpovídajícího prvku v matici ($f_{i,j}$) a tyto hodnoty jsou navzájem vynásobeny, což poskytuje součin $P_n$. Konečná hodnota výsledného skóre je získána jako převrácená hodnota součinu $P_n$ s následnou normalizací tohoto výsledku na bílkovinu o molekulové hmotnosti 50 kDa za účelem redukce náhodné akumulace skóre při porovnávání se sekvencemi vysokomolekulárních bílkovin ($M$ označuje molekulovou hmotnost v kDa aktuálně porovnávané bílkoviny v databázi):

$$\text{skóre} = \frac{50}{P_n \times M}$$

Hlavní předností tohoto skórovacího algoritmu je, že dává peptidům s vyšší molekulovou hmotností větší skórovací váhu a kompenzuje nenáhodné distribuce molekulových hmotností peptidů v bílkovinách o různé molekulové hmotnosti.

Skórovací schéma MOWSE pro PMF metodu je poměrně rozšířené. Program MS-Fit (cit.[51]) v Protein Prospectoru[37] používá toto schéma v částečně pozměněné formě[52]. Skórovací matice je vypočtena poté, kdy jsou již na databázi aplikována kritéria pro druh organismu a rozsah možné molekulové hmotnosti identifikované bílkoviny. Matice má 11 intervalů pro molekulovou hmotnost bílkoviny (10 intervalů po 10 kDa až do molekulové hmotnosti 100 kDa, 11. interval je určen pro všechny molekulové hmotnosti bílkovin nad 100 kDa) a 30 intervalů pro molekulové hmotnosti peptidů (pro 100 Da intervaly až do 3000 Da; peptidy s vyšší molekulovou hmotností se nezapočítávají). Peptidy bez vynechaného štěpného místa ve struktuře přispívají faktorem 1, zatímco peptidy mající nějaká vynechaná štěpná místa přispívají faktorem nazývaným „pfactor“, který zadává uživatel. Nalezené peptidy se pak dělí do dvou skupin: skórující peptidy (např. nemodifikované peptidy) a neskórující peptidy (např. peptid s oxidovaným methioninem a tentýž peptid bez oxidovaného methioninu nebyl nalezen). Výsledné skóre se pak již počítá stejným způsobem jako v případě algoritmu MOWSE.

Jiným široce používaným programem PMF je Mascot Peptide Mass Fingerprinting[53] v souboru programů nazývaném Mascot[54]. Cílem skórovacího algoritmu v Mascot PMF je poskytnout jednoduché pravidlo, které umožní rozhodnout, zda je výsledek identifikace relevantní či nikoliv. Dále by měl poskytnout hodnotu skóre, jež by se dala srovnat s výsledky aplikace jiných databázových vyhledávacích nástrojů a která by též sloužila k optimalizaci parametrů databázového vyhledávání. Program PMF Mascot používá pravděpodobnostní skórovací systém MOWSE. Algoritmus je založen na výpočtu pravděpodobnosti $P$, že pozorovaná shoda mezi experimentálními hodnotami a teoretickými hodnotami posuzované bílkoviny v databázi je náhodná událost[55]. Jako významné se považují hity s pravděpodobností $P$ nižší než 0,05. Z praktických důvodů (např. pro lepší shodu mít vyšší skóre) je celkové skóre udáváno jako $-10 \log_{10} P$. Detailnější popis skórovacího algoritmu nebyl publikován.

Dalším z řady PMF identifikačních programů je ProFound (cit.[56]) na adrese programového komplexu PROWL (cit.[57]). Skórovací algoritmus ProFoundu používá výpočet bayesovské pravděpodobnosti, že porovnávaná bílkovina z databáze je analyzovanou bílkovinou v experimentu[58]. Tato pravděpodobnost je počítána za následujících předpokladů: *1*) analyzovaná bílkovina je přítomná v databázi; *2*) všechny detegované ionty ve spektru pocházejí z enzymatického štěpení analyzované bílkoviny; *3*) náhodná shoda experimentální a teoretické molekulové hmotnosti peptidu se neuvažuje, všechny shody jsou považovány jako shody s teoreticky předpovězenými peptidy. Pravděpodobnost hypotézy, že „bílkovina $k$ je analyzovanou bílkovinou“ za podmínky daných experimentálních dat $D$ a okolních informací $I$ (volby databázového vyhledávání), je dána následující rovnicí[58]:

$$P(k \mid DI) \propto P(k \mid I) \frac{(N-r)!}{N!} \prod_{i=1}^{r} \left\{ \sqrt{\frac{2}{\pi}} \frac{m_{\max} - m_{\min}}{\sigma_i} \times \sum_{j=1}^{g_i} \exp\left[ -\frac{\left(m_i - m_{ij0}\right)^2}{2\sigma_i^2} \right] \right\} F_{\text{vzorec}}$$

s normalizační podmínkou

$$\sum_{k \in \text{databáze}} P(k \mid DI) = 1$$

kde $P(k|I)$ je pravděpodobnost hypotézy $k$ za dané okolní informace $I$; $N$ je teoretický počet peptidů generovaných z bílkoviny $k$ enzymem použitým v experimentu; $r$ je počet nalezených peptidů; $m_{max}-m_{min}$ je rozsah naměřených molekulových hmotností peptidů; $m_i$ je naměřená molekulová hmotnost $i$-té shody s násobností $g_i$ (počet teoretických peptidů odpovídajících $m_i$); $m_{ij0}$ je vypočtená molekulová hmotnost $j$-tého peptidu v $i$-té shodě; $\sigma_i$ je standardní odchylka měření molekulové hmotnosti o hodnotě $m_i$ a $F_{vzorec}$ empirický člen zvyšující pravděpodobnost v případě, že byly pozorovány překrývající se nebo po sobě jdoucí peptidy. Na základě pravděpodobnostní hodnoty pro každou bílkovinu v databázi ProFound spočítá tzv. skóre Z, které určuje, jak významně se identifikovaná bílkovina liší od identifikace prostřednictvím náhodných shod molekulových hmotností v databázi[59].

*Tato práce byla podpořena prostředky výzkumného centra MŠMT 1M06030 „Funkční genomika a proteomika rostlin“.*

LITERATURA

1. Whitehouse C. M., Dreyer R. N., Yamashita M., Fenn J. B.: Anal. Chem. *57*, 675 (1985).
2. Karas M., Hillenkamp F.: Anal. Chem. *60*, 2299 (1988).
3. Fenn J. B., Mann M., Meng C. K., Wong S. F., Whitehouse C. M.: Science *246*, 64 (1989).
4. Karas M., Bachmann D., Bahr U., Hillenkamp F.: Int. J. Mass Spectrom. Ion Processes *78*, 53 (1987).
5. Tanaka K., Waki H., Ido Y., Akita S., Yoshida Y., Yoshida T.: Rapid Commun. Mass Spectrom. *2*, 151 (1988).
6. Henzel W. J., Billeci T. M., Stults J. T., Wong S. C., Grimley C., Watanabe C.: Proc. Natl. Acad. Sci. U.S.A. *90*, 5011 (1993).
7. Mann M., Hojrup P., Roepstorff P.: Biol. Mass Spectrom. *22*, 338 (1993).
8. Pappin D. J. C., Hojrup P., Bleasby A. J.: Curr. Biol. *3*, 327 (1993).
9. James P., Quadroni M., Carafoli E., Gonnet G.: Biochem. Biophys. Res. Commun. *195*, 58 (1993).
10. Yates J. R., Speicher S., Griffin P. R., Hunkapiller T.: Anal. Biochem. *214*, 397 (1993).
11. Henzel W. J., Stults J. T., Watanabe C.: *Proceedings of the Third Symposium of the Protein Society*. Seattle 1989.
12. Cleveland D. W., Fischer S. G., Kirschner M. W., Laemmli U. K.: J. Biol. Chem. *252*, 1102 (1977).
13. Vestal M. L., Juhasz P., Martin S. A.: Rapid Commun. Mass Spectrom. *9*, 1044 (1995).
14. Mamyrin B. A.: Int. J. Mass Spectrom. Ion Processes *131*, 1 (1994).
15. Henzel W. J., Watanabe C., Stults J. T.: J. Am. Soc. Mass Spectrom. *14*, 931 (2003).
16. Shevchenko A., Wilm M., Vorm O., Mann M.: Anal. Chem. *68*, 850 (1996).
17. Jensen O. N., Wilm M., Shevchenko A., Mann M., v knize: *Methods in Molecular Biology: 2-D Proteome Analysis Protocols* (Link A. J., ed.), kap. 52. Humana Press, Totowa 1999.
18. Bieman K.: Ann. Rev. Biochem. *32*, 755 (1963).
19. Barber M., Jolles P., Vilkas E., Lederer E.: Biochem. Biophys. Res. Commun. *18*, 469 (1965).
20. Barber M., Wolstenholme W. A., Guinand M., Michel G., Das B. C., Lederer E.: Tetrahedron Lett. *6*, 1331 (1965).
21. Barber M., Bordoli R. S., Sedgwick R. D., Tyler A. N.: J. Chem. Soc., Chem. Commun. *7*, 325 (1981).
22. Morris H. R., Panico M., Barber M., Bordoli R. S., Sedgwick R. D., Tyler A.: Biochem. Biophys. Res. Commun. *101*, 623 (1981).
23. Desiderio D. M., Katakuse I.: Anal. Biochem. *129*, 425 (1983).
24. Tomer K. B., Crow F. W., Gross M. L., Kopple K. D.: Anal. Chem. *56*, 880 (1984).
25. Hunt D. F., Yates J. R., Shabanowitz J., Winston S., Hauer C. R.: Proc. Natl. Acad. Sci. U.S.A. *83*, 6233 (1986).
26. Haller I., Mirza U. A., Chait B. T.: J. Am. Soc. Mass Spectrom. *7*, 677 (1996).
27. Jonscher K. R., Yates J. R.: Anal. Biochem. *244*, 1 (1997).
28. Shevchenko A., Loboda A., Shevchenko A., Ens W., Standing K. G.: Anal. Chem. *72*, 2132 (2000).
29. Medzihradszky K. F., Campbell J. M., Baldwin M. A., Falick A. M., Juhasz P., Vestal M. L., Burlingame A. L.: Anal. Chem. *72*, 552 (2000).
30. http://prospector.ucsf.edu/ucsfhtml4.0/mspattern.htm, staženo 16. září 2006.
31. Pearson W. R., Lipman D. J.: Proc. Natl. Acad. Sci. U.S.A. *85*, 2444 (1988).
32. http://fasta.bioch.virginia.edu, staženo 16. září 2006.
33. Altschul S. F., Gish W., Miller W., Myers E.W., Lipman D. J.: J. Mol. Biol. *215*, 403 (1990).
34. http://www.ncbi.nlm.nih.gov/blast, staženo 16. září 2006.
35. Mann M., Wilm M.: Anal. Chem. *66*, 4390 (1994).
36. http://www.narrador.embl-heidelberg.de/GroupPages/PageLink/peptidesearch/Services/PeptideSearch/FR_PeptidePatternFormG4.html, staženo 16. září 2006.
37. http://prospector.ucsf.edu, staženo 16. září 2006.
38. http://prospector.ucsf.edu/ucsfhtml4.0/mstagfd.htm, staženo 16. září 2006.
39. Clauser K. R., Baker P., Burlingame A. L.: Anal. Chem. *71*, 2871 (1999).
40. Eng J. K., McCormack A. L., Yates J. R.: J. Am. Soc. Mass Spectrom. *5*, 976 (1994).
41. http://fields.scripps.edu/sequest, staženo 16. září 2006.
42. Yates J. R., Eng J. K., McCormack A. L., Schieltz D.: Anal. Chem. *67*, 1426 (1995).
43. Yates J. R., Eng J. K., McCormack A. L.: Anal. Chem. *67*, 3202 (1995).
44. Griffin P. R., MacCoss M. J., Eng J. K., Blevins R.

A., Aaronson J. S., Yates J. R.: Rapid Commun. Mass Spectrom. *9*, 1546 (1995).
45. Yates J. R., Eng J. K., Clauser K. R., Burlingame A. L.: J. Am. Soc. Mass Spectrom. *7*, 1089 (1996).
46. Yates J. R., Morgan S. F., Gatlin C. L., Griffin P. R., Eng J. K.: Anal. Chem. *70*, 3557 (1998).
47. MacCoss M. J., Wu C. C., Yates J. R.: Anal. Chem. *74*, 5593 (2002).
48. Razumovskaya J., Olman V., Xu D., Uberbacher E. C., VerBerkmoes N. C., Hettich R. L., Xu Y.: Proteomics *4*, 961 (2004).
49. http://www.ncbi.nlm.nih.gov/Genbank/index.html, staženo 16. září 2006.
50. http://www.expasy.org/sprot/, staženo 16. září 2006.
51. http://prospector.ucsf.edu/ucsfhtml4.0/msfit.htm, staženo 16. září 2006.
52. http://prospector.ucsf.edu/ucsfhtml4.0/instruct/fitman.htm#scoring, staženo 16. září 2006.
53. http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=PMF, staženo 16. září 2006.
54. http://www.matrixscience.com/search_form_select.html, staženo 16. září 2006.
55. Perkins D. N., Pappin D. J. C., Creasy D. M., Cottrell J. S.: Electrophoresis *20*, 3551 (1999).
56. http://prowl.rockefeller.edu/profound_bin/WebProFound.exe, staženo 16. září 2006.
57. http://prowl.rockefeller.edu, staženo 16. září 2006.
58. Zhang W. Z., Chait B. T.: Anal. Chem. *72*, 2482 (2000).
59. http://129.85.19.192/profound/zscore.pdf, staženo 16. září 2006.

**P. Řehulka, H. Řehulková, and J. Chmelík** *(Institute of Analytical Chemistry, Academy of Sciences of the Czech Republic, Brno):* **Acquisition of Proteomic Data Using Mass Spectrometry and Their Bioinformatics Interpretation**

**Purpose of Review**

The review provides a brief overview of techniques used for interpretation of mass spectrometric data acquired in proteomic experiments.

**Findings**

Processing of the data by bioinformatics is an important part of protein identification in proteomics. Analyzed samples can be measured using different mass spectrometric techniques and the obtained data require different methods of processing. Genomic databases rapidly grow and a suitable way of manipulation with data becomes a crucial step for their correct interpretation.

**Conclusions**

Basic methods of interpretation of mass spectrometric data are summarized and some examples show how the corresponding programs work.

**VÝZVA**

k podávání žádostí o podporu
česko-americké vědeckotechnické spolupráce

Bližší informace a formuláře na www.amvis.cz

Uzavírka **přijímání návrhů je 31. května 2007.**

*Kontakt:*
AMVIS, o.p.s.
Senovážné nám. 24
116 47 Praha 1

e-mail: amvis@amvis.cz
tel.: 234 621 222 nebo 441
fax: 234 621 552