Chem. Listy *107*, 44–53 (2013) Referát

PROGRAMOVÁ APLIKACE PRO PŘEDPOVĚĎ VÝSLEDKŮ ŠTĚPENÍ PROTEINŮ PROTEOLYTICKÝMI ENZYMY

MARTIN RAUS, DAVID KOPEČNÝ a MAREK ŠEBELA

Katedra biochemie a Centrum regionu Haná pro biotechnologický a zemědělský výzkum, Přírodovědecká fakulta, Univerzita Palackého, Šlechtitelů 11, 783 71 Olomouc
martin_raus@post.cz; marek.sebela@upol.cz

Došlo 23.11.11, přepracováno 29.3.12, přijato 5.4.12.

Klíčová slova: databáze, hmotnostní spektrometrie, peptidové mapování, protein, proteolytický enzym, sekvence

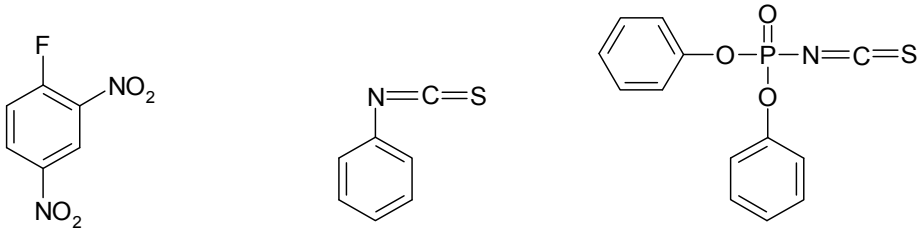
Obsah

1. Úvod

Proteiny jsou biologické makromolekuly s rozmanitou funkcí. Jedná se o lineární polymery v genech kódovaných L-α-aminokyselin propojených vazbami do polypeptidového řetězce[1]. Prvenství v rozpoznání této zákonitosti se připisuje německému chemikovi Franzi Hoffmeisterovi[2]. Pořadí aminokyselin vytváří aminokyselinovou sekvenci a určuje jedinečnou primární strukturu[3]. Jde o formu biologického kódu, který skrývá fyzikálně-chemické předpoklady a zákonitosti pro budování vyšších struktur daného proteinu (struktura sekundární, terciární případně kvarterní)[4]. Toho se využívá při předpovědi (predikci) prostorového uspořádání polypeptidového řetězce (angl. „protein folding“). Kromě strukturních informací je z aminokyselinové sekvence možné vyčíst také informace ve vztahu k biologické funkci proteinu, např. charakteristický motiv pro vazbu koenzymu[5]. Určování proteinových sekvencí a využití jejich znalosti pro identifikaci proteinů ve vzorku patří k předmětům vědecké disciplíny zvané proteomika[1]. V textu jsou v historické posloupnosti shrnuty přístupy pro analýzu aminokyselinových sekvencí a představena internetová programová aplikace pro předpověď výsledků štěpení proteinů proteolytickými enzymy (využívá se při sekvencování hmotnostní spektrometrií) a výpočet biochemických parametrů proteinů z aminokyselinové sekvence.

2. Sekvencování proteinů s použitím chemických činidel

Metodice určování sekvence proteinu (= sekvencování) byla v biochemii věnována značná pozornost. Za prvořadou osobnost v této souvislosti je právem považován britský biochemik Frederick Sanger, který určil úplnou aminokyselinovou sekvenci inzulinu[6]. Použil tzv. Sangerovo činidlo (2,4-dinitro-1-fluorbenzen, DNFB; obr. 1), které reaguje s přístupnými aminoskupinami proteinu, zvláště s aminoskupinou na tzv. *N*-konci (první aminokyselina v pořadí). Peptidy vzniklé hydrolýzou inzulinu Sanger podrobil dvojrozměrné separaci a získal peptidovou mapu (angl. „fingerprint“ – otisk prstu). Peptid obsahující *N*-konec byl rozpoznán podle žlutého zbarvení vzniklého značením s DNFB. Opakováním této procedury při rozdílných podmínkách počáteční hydrolýzy Sanger určil sekvenci mnoha peptidů a jejich skládáním do delších sekvencí dospěl k výsledku[7–10]. Určení sekvence inzulinu bylo klíčové i pro myšlenky a důkazy ohledně kódování proteinů v molekule DNA[11]. Švédský biochemik Pehr Victor Edman v 50. letech 20. stol. publikoval několik prací, kde popsal použití činidla fenylisothiokyanátu (PITC, Edmanovo činidlo; obr. 1)[12,13], které reaguje v mírně bazickém prostředí s *N*-koncovou aminoskupinou proteinu či peptidu. V kyselém prostředí se značená aminokyselina odštěpí jako anilinothiazolinonový derivát, látka je extrahována do organického rozpouštědla a kyselinou převedena na stabilní fenylthiohydantoinový derivát příslušné aminokyseliny identifikovatelný nejlépe kapalinovou chromatografií. Tento postup je možné opakovat i pro další v sekvenci následující aminokyseliny (Edmanova degradace). Jako maximální možný počet proveditelných cyklů se udává číslo 50–60 (cit.[1]). Nevýhodou je chemické blokování *N*-konce některých proteinů a peptidů nebo jeho nepřístupné skrytí v molekule a dále nemožnost určit polohu disulfidových vazeb. Cyklický sled reakcí bylo možné dobře automatizovat, na trhu se tak od konce 60. let 20. století objevily automatické sekvenátory[14,15]. Nutno dodat, že je možné provádět i *C*-koncové sekvencování s chemickým značením *C*-koncové aminokyseliny difenylfosforylisothiokyanátem (DPPITC, obr. 1) a jejím odštěpením ve formě



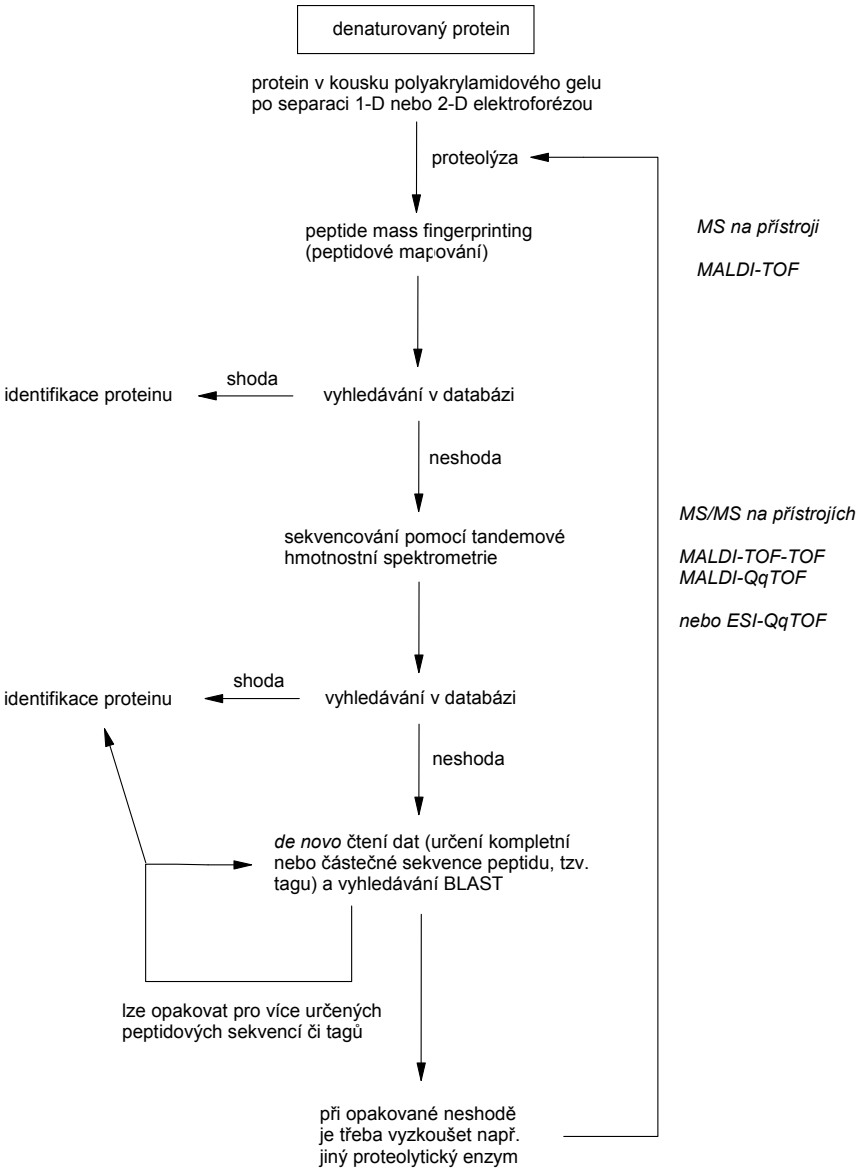
Obr. 1. **Chemická činidla pro sekvenční analýzu proteinů: 2,4-dinitro-1-fluorbenzen (DNFB), fenylisothiokyanát (PITC) a difenylfosforylisothiokyanát (DPPITC)**

thiohydantoinového derivátu. Použití je však méně časté neboť vyžaduje větší množství vzorku[16].

3. Sekvencování proteinů hmotnostní spektrometrií

Použití Edmanových sekvenátorů bylo běžné do poloviny 90. let 20. století. To však již probíhal nástup hmotnostní spektrometrie (MS) peptidů a proteinů, která se nakonec v oblasti analýzy aminokyselinové sekvence prosadila. V roce 2008 sekvenátory zmizely z komerčního trhu přístrojů[17]. Klíčovým předpokladem pokroku byl objev tzv. měkkých (angl. „soft“) ionizačních technik pro MS, které umožňují šetrnou ionizaci složitějších organických molekul a makromolekul při zachování struktury – tedy bez fragmentace. Jde o ionizaci elektrosprejem (angl. „electrospray ionization“; zkratka ESI) a laserovou desorpci a ionizaci s účastí matrice (angl. „matrix-assisted laser desorption/ionization“; zkratka MALDI). Princip ESI byl rozpracován již v 60. letech 20. století[18]. Systematická práce na jejím použití pro velké biologické molekuly (např. proteiny) byla oceněna Nobelovou cenou za chemii pro amerického vědce Johna Bennetta Fenna v roce 2002. Profesor Fenn a spol.[19] v průběhu 80. let 20. století vyvinuli postup, ve kterém je zředěný roztok makromolekuly sprejován z kovové jehly mikrostříkačky nebo pokovené kapiláry, na kterou je přivedeno vysoké elektrické napětí v řádu kilovoltů. Nabité kapičky vzniklé za atmosférického tlaku jsou vysušovány proudícím inertním plynem. Rozpouštědlo se postupně odpařuje a uvolňují se ionty, které jsou tlakovým rozdílem vtaženy do hmotnostního analyzátoru přístroje[20]. Objev MALDI byl oceněn Nobelovou cenou za chemii pro japonského vědce Koiči Tanaku, a to taktéž v roce 2002. Jeho tým zjistil, že jemný práškový kobalt s glycerolem usnadňuje ionizaci analytu[21]. Nesporný je však i příspěvek německých vědců (Michael Karas a Franz Hillenkamp), kteří již předtím provedli laserovou desorpci a ionizaci molekul v přítomnosti malé organické sloučeniny jako matrice[22] a později popsali použití této ionizační techniky i pro proteiny[23]. Při technice MALDI je vzorek krystalován spolu s velkým nadbytkem matrice (původní matricí byla kyselina nikotinová). Na vzorek s matricí dopadají krátké pulsy laserového světla a dojde k odpaření zasažené části. Nadbytek matrice absorbuje energii laseru, rychle se šířící oblak matrice v plynné fázi unáší molekuly analytu do vakua v hmotnostním analyzátoru a usnadňuje přitom proces jejich ionizace[20].

Pro určování aminokyselinové sekvence pomocí MS jsou k dispozici následující možnosti: *1*) peptidové mapování; *2*) sekvencování peptidů tandemovou hmotnostní spektrometrií (MS/MS; využívá se fragmentace v okolí peptidové vazby) a *3*) „top-down“ sekvencování intaktních proteinů (jednoduchý český ekvivalent anglického termínu není, vysvětlení níže). Peptidové mapování (angl. „peptide mass fingerprinting“, PMF) je založeno na štěpení proteinu specifickými proteolytickými enzymy (např. trypsinem nebo peptidasami Glu-C, Arg-C aj.) a měření molekulové hmotnosti peptidů na MALDI přístrojích (obr. 2). Sada hmotnostních čísel slouží jako peptidová mapa pro identifikaci proteinu vyhledáváním v databázi sekvencí na základě srovnání s předpovědí štěpení[24,25]. Skutečné sekvencování peptidů vzniklých chemickým či enzymovým štěpením proteinů využívá čtení posloupnosti aminokyselin z odpovídacích hmotnostních rozdílů v rámci určité série fragmentů, které vznikají během MS/MS měření v kolizní cele MALDI i ESI přístroje při srážkách prekurzorového peptidu s částicemi kolizního plynu[26]. Podobnou avšak principiálně odlišnou variantou je samovolný rozpad za iontovým zdrojem (angl. „post-source decay“, PSD) na MALDI přístrojích[27]. Není-li sekvence studovaného proteinu uložena v databázi, jde o tzv. *de novo* sekvencování (obr. 2). Pro zcela neznámé proteiny se *de novo* sekvencování kombinuje s identifikací na základě podobnosti se známými proteiny pomocí algoritmu BLAST (angl. „basic local alignment search tool“ – nástroj pro běžné vyhledávání podobných sekvencí v databázích, který výchází z identifikace krátkých homologních subsekvencí bez mezer s následným rozšiřováním vyhledávání v okolí subsekvencí s cílem získat lokálně seřazené sekvence, do nichž mo-



Obr. 2. **Běžná strategie sekvenční analýzy proteinů pomocí hmotnostní spektrometrie.** Schéma nezahrnuje tzv. „top-down“ strategii sekvencování, kdy je čistý intaktní protein fragmentován v hmotnostním spektrometru. Zkratky pro MS přístroje vysvětleny v oddíle 3, dále TOF – „time-of-flight“ tj. analyzátor doby letu, Q – kvadrupólový analyzátor, q – kvadrupólová kolizní cela

hou být vloženy mezery)[28]. Nejnovější záležitostí je „top-down“ sekvencování, které poskytuje sekvenční informaci na základě fragmentace čistého intaktního proteinu nikoli peptidů. Existují varianty v závislosti na použitém přístroji, kdy se principiálně liší proces fragmentace (ECD – disociace záchytem elektronu, angl. „electron capture dissociation“, na přístrojích s iontovým cyklotronem a Fourierovou transformací; ISD – rozpad v iontovém zdroji, angl. „in-source decay“, na MALDI přístrojích a ETD – disociace přenosem elektronu, angl. „electron transfer dissociation“, na přístrojích s iontovou pastí)[29–31]. Dnes běžné sekvencování pomocí hmotnostní spektrometrie má svoje nevýhody, z nichž některé lze překonat pokročilou technologií na špičkových přístrojích. U tandemové hmotnostní spektrometrie jsou to např. nejednoznačnost daná výskytem izobarických aminokyselin (I/L, K/Q) a problematické výsledky při získání neúplných fragmentačních spekter vlivem určitých aminokyselinových modifikací[32].

4. Nepřímé určování aminokyselinové sekvence a bioinformatika

Aminokyselinovou sekvenci proteinu lze odvodit také nepřímo na základě kódující genové sekvence. K amplifikaci určitého genu potřebujeme polymerasovou řetězovou reakci (PCR) s oligonukleotidovými primery[33,34], dále klonování amplikonu do vhodného plasmidu[35] a DNA sekvencování[36]. Výsledek se nakonec získá překladem zjištěné nukleotidové sekvence genu pomocí abecedy genetického kódu[37,38]. Rozmanité projekty sekvencování genů a genomů (genom = úplná genetická informace organismu, tj. soubor všech genů) začaly od přelomu 70. a 80. let 20. století rychle plnit nejen databáze genových sekvencí, ale i proteinové databáze. Genomika využívá pro čtení malých genomů (do 7000 bp) tzv. „shotgun" sekvencování (český ekvivalent se neužívá), kdy rozbitím genomové DNA na náhodné fragmenty získáme velké množství materiálu pro analýzu. Sekvencování se provádí opakovaně, aby přečtené fragmenty (angl. „reads") v dostatečném množství přesahovaly a z těchto přesahů bylo možné sestavit sekvenci celého genomu[39]. Pro rozsáhlé genomy se využívá umělých bakteriálních chromosomů (angl. „bacterial artificial chromosome", BAC), což jsou plazmidy obsahující fragmenty genomové DNA studovaného organismu o obvyklé velikosti 150–350 kbp (cit.[40]). Postupná sekvenční analýza BAC klonů (angl. „clone-by-clone", CBC) je souborem dílčích „shotgun" projektů. Pro vlastní určení sekvence se užívá tzv. pyrosekvencování[41] (reakční směs obsahuje řadu enzymů pro katalýzu reakcí navazujících na reakci DNA polymerasy a produkujících v konečné fázi fluorescenční světlo jako důsledek připojení nukleotidu v narůstající sekvenci), nyní běžné v tzv. 454 pikolitrové variantě zavedené firmou 454 Life Sciences[42].

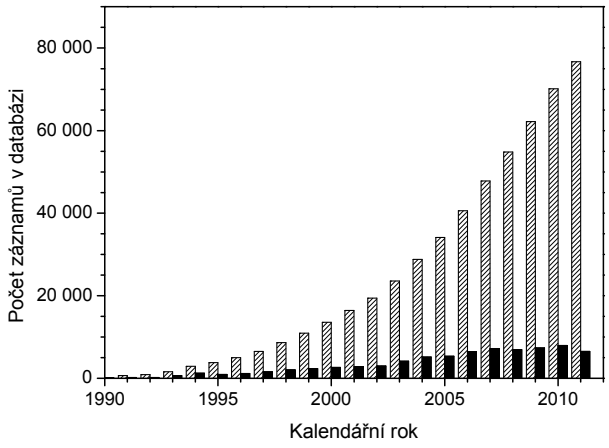
S rostoucím množstvím dat v centralizovaných databázích se zrodila i nová vědecká disciplína. Bioinformatika byla původně úzce spojena s genetikou a genomikou, a to díky genomovým projektům a výsledkům automatizovaného čtení a skládání částečných sekvencí komplementárních DNA (angl. „expressed sequence tags";ESTs)[43]. S exponenciálním přívalem nových sekvencí bylo třeba databáze nejen udržovat a budovat uživatelská rozhraní pro vkládání, sdílení a poskytování sekvenčních dat (např. internetová aplikace Entrez organizace National Center for Biotechnology Information, NCBI, při National Institutes of Health, NIH, v Bethesdě, Maryland, USA[44]), ale začít data opatřovat anotacemi, analyzovat a interpretovat. Dnes je úkolem bioinformatiky nejen vyvíjet vhodné nástroje a služby pro přístup, používání a správu databází biologických informací (sekvence, struktura, funkce), ale zejména konstrukce nových algoritmů, výpočetních a statistických procedur, programů a teorií pro vysvětlování vztahů mezi jednotlivými záznamy v databázi. V případě proteinů se zájem soustředí na problematiku sekvenční homologie, sdružování sekvencí do proteinových rodin a nadrodin (angl. „superfamily"), předpověď prostorové struktury, posttranslačních modifikací a funkce, porovnávání strukturních domén, strukturní modelování nebo popis protein-proteinových interakcí[45].

5. Databáze sekvencí

Prostřednictvím Internetu je dostupná řada databází nukleotidových nebo aminokyselinových sekvencí, které se obsahově dělí na primární, kompozitní, sekundární a ostatní (specializované)[1,46]. Pokud jde o proteiny, objevuje se také dělení na databáze s vysokou úrovní péče a ověřování (angl. „curated"), souborné (angl. „omnibus"), počítačově přeložené (angl. „translated") a ostatní. Primární databáze obsahují pouze sekvence s anotací. Hlavními primárními zdroji pro nukleové kyseliny jsou GenBank (americká, spravuje NCBI), ENA („European Nucleotide Archive")/EMBL-Bank („European Molecular Biology Laboratory Nucleotide Sequence Database") a DDBJ („DNA Data Bank of Japan")[44], pro proteiny pak např. PIR, SWISS-PROT, TrEMBL, PSD-Kyoto, PRF a NRL-3D. V kompozitních databázích jsou spojena data z více zdrojů, což umožňuje provádět efektivní vyhledávání, zvláště pokud jsou dozorem eliminována nadbytečná opakování sekvencí (redundance)[1]. Jako typické příklady kompozitních databází lze uvést pro proteiny NCBInr[47] (spojuje data z GenBank – překlady, GenPept, PDB, PIR, PRF, SWISS-PROT, TrEMBL), MSDB[48] tj „Mass Spectrometry protein sequence DataBase" (data z GenBank – překlady, PIR, SWISS-PROT, TrEMBL, NRL-3D; aktualizována do r. 2006) a OWL[49] (PIR, SWISS-PROT, GenBank-překlady, NRL-3D).

Sekundární databáze obsahují informace získané bioinformatickou analýzou sekvencí v primárních zdrojích. Z nich je populární databáze PROSITE[50] popisující funkční místa v proteinech (např. vazebná místa, aktivní místa enzymů), strukturní domény nebo proteinové rodiny. Ze specializovaných databází je možné zmínit např. databázi ENZYME[51] na serveru Expasy („Expert Protein Analysis System"), která je s databázemi SWISS-PROT, PROSITE a SWISS-2DPAGE mimo jiné informační zdroje spravována organizací SIB („Swiss Institute of Bioinformatics"), nebo databázi PDB[52] („Protein Data Bank") spravovanou konsorciem RCSB („Research Collaboratory for Structural Bioinformatics") a zaměřující se především na trojrozměrné struktury proteinů. Na počátku roku 2012 bylo v PDB téměř 80 000 struktur (v roce 2004 to bylo 24 000 struktur, v roce 2000 pouhých 5000), což je sice ohromné číslo, ale v kontrastu s více než stovkou miliónů sekvencí v GenBank stále zanedbatelné (obr. 3). V databázi NDB[53] („Nucleic Acid Database"; The State University of New Jersey, USA) bylo počátkem roku 2012 asi 5800 struktur nukleových kyselin. Zmínku si jistě zaslouží i databáze MEROPS[54], která je věnována proteolytickým enzymům (peptidasám), jejich substrátům a inhibitorům. Databázi spravuje The Wellcome Trust Sanger Institute v Hinxtonu, Velká Británie.

Proteinová databáze SWISS-PROT[55] bývá často považována za nejlepší přinejmenším pokud jde o kvalitu



Obr. 3. **Exponenciální nárůst počtu záznamů v databázi PDB v posledních dvaceti letech**

anotací (přístupová čísla, taxonomie, literatura, funkce, vlastnosti, odvozená data aj.). Má dnes více než 500 000 záznamů s minimální redundancí. Její sekvence pocházejí z databáze PIR[56] („Protein Information Resource“; Georgetown Medical Center, USA), která je komplexnější a aktuálnější[1]. SWISS-PROT databáze vznikla v roce 1986 a původně její fungování zajišťovala univerzita v Ženevě, Švýcarsko, spolu s tehdejším předchůdcem dnešní organizace EBI („The European Bioinformatics Institute“). Je to databáze s vysokou úrovní péče a manuálně anotovanými sekvencemi. Spolu s automaticky anotovanými překlady nukleotidových sekvencí v ENA (EMBL-Bank), které tvoří proteinovou databázi TrEMBL (zavedena 1996), je SWISS-PROT součástí velké databáze UniProt Knowledge Base (UniProt KB)[57] a je v péči konsorcia institucí EBI, SIB a PIR. Souborná databáze NCBInr je ceněna pro svou aktuálnost, neboť se v ní objevují i překlady nejnovějších kódujících nukleotidových sekvencí (CDS) z databáze GenBank. V souvislosti s GenBank je třeba zmínit ještě odvozenou referenční databázi RefSeq[58] („Reference Sequence Collection“) spravovanou stejnou organizací (NCBI). RefSeq je neredundantní soubor sekvencí DNA, RNA a proteinů s vysokou úrovní péče. Obsahuje vždy jeden příklad biologické molekuly pro vybrané modelové organismy (na počátku roku 2012 jich bylo zhruba 17 000 – v GenBank je zastoupeno více než 380 000 organismů) [44,59] s odděleným přístupem pro genomovou DNA, transkripty a proteiny z těchto transkriptů. Přístupová čísla v RefSeq mají prefix, který ukazuje, zda již byla sekvence ověřena příslušným oddělením NCBI. Zkontrolované záznamy mají prefix NP, ostatní mají prefix XP.

6. Programová aplikace pro předpověď výsledků štěpení proteinů

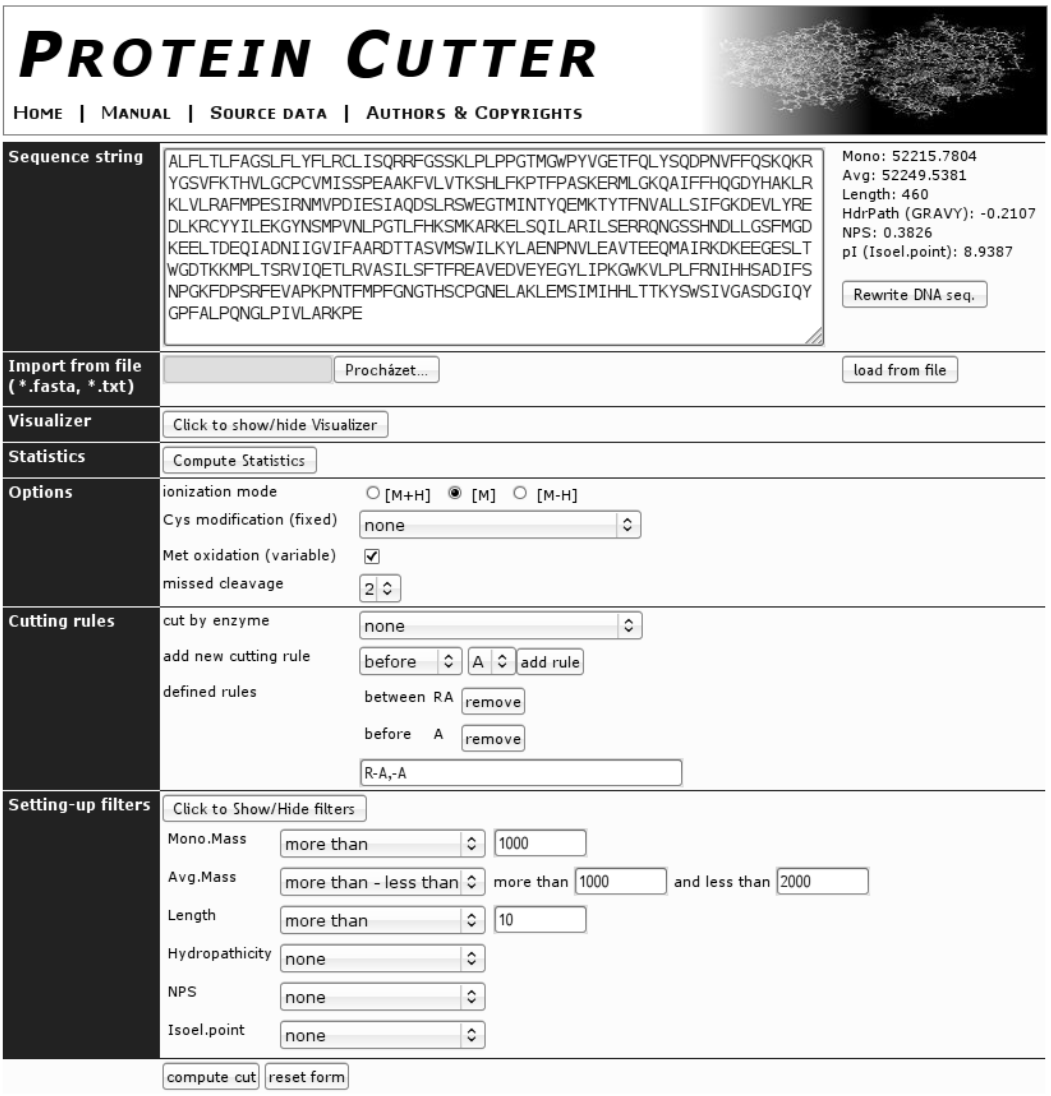
6.1. Představení aplikace

Programová aplikace Protein Cutter (název podle angl. slovesa „cut“ tj. krájet, štípat; http://biochemie.upol.cz/software/proteincutter), kterou bychom chtěli představit, byla vyvinuta pro předpověď primární struktury peptidů vzniklých štěpením proteinů působením proteolytických enzymů s předpokládaným použitím v biochemii a proteomice. Umožňuje zadat vstupní data (aminokyselinovou sekvenci proteinu nebo nukleotidovou sekvenci odpovídajícího genu), kde chceme provést štěpení a stanovit pravidla, podle kterých budou generovány teoretické peptidy. Výstupem jsou aminokyselinové sekvence peptidů, které dle zadaných pravidel teoreticky mohou vzniknout. Získaná i vkládaná data jsou vizuálně doplněna hodnotami fyzikálně-chemických parametrů počítanými na základě aminokyselinových sekvencí (molekulová hmotnost, isoelektrický bod aj.). Ze starších srovnatelných aplikací je možné zmínit např. programy portálu ExPASy[60]: PeptideMass (http://web.expasy.org/peptide_mass/) a PeptideCutter (http://web.expasy.org/peptide_cutter/), dále MS-Digest (http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msdigest) nebo Protein Calculator (http://www.scripps.edu/~cdputnam/protcalc.html). Všechny tyto aplikace umožňují předpověď (predikci) výsledků štěpení a výpis teoretických peptidů (včetně molekulových hmotností) po uživatelském výběru proteolytického enzymu a zadání předpokládaných modifikací v aminokyselinové sekvenci, avšak postrádají některé možnosti, které jsme jako přidanou hodnotu vložili do aplikace Protein Cutter. Běžně chybí informace o isoelektrickém bodu (výjimkou je program Protein Calculator), není možné vkládat nukleotidové sekvence DNA či mRNA pro predikci peptidů příslušného translačního produktu, nejsou poskytovány informace o zastoupení jednotlivých aminokyselin, indexu GRAVY (číslo počítané na základě zastoupení hydrofilních a hydrofobních aminokyselin, angl. „grand average of hydropathicity“) aj. Kromě toho není běžné definovat vlastní pravidla štěpení a omezené je i třídění a filtrování výsledků. Nová aplikace umožňuje i vizualizaci zastoupení jednotlivých aminokyselin a přidává možnost pokračovat v teoretickém „štěpení“ jednotlivých generovaných peptidů s novými uživatelskými pravidly simulující činnost směsi proteolytických enzymů. Účelem nebylo konkurovat licencovaným programům dodávaným k hmotnostním spektrometrům pro proteomické aplikace nebo jiným lokálně instalovaným programům, jako je např. volně distribuovaný mMass[61,62]. Program mMass je primárně určený jako podpora pro MS proteinů a peptidů a v aktuální verzi 5.0 umožňuje kromě práce s hmotnostními spektry mj. editování sekvencí (včetně predikce štěpení), vkládání aminokyselinových modifikací nebo fragmentaci peptidů simulující výsledky MS/MS analýzy. Je však specializovaný a nedovoluje tak získání informací využitelných v jiných oblastech studia proteinů.

6.2. Práce s aplikací

Zpracování dat probíhá ve čtyřech krocích – vložení dat, nastavení možností výpočtů, definování pravidel štěpení a nastavení filtrů výstupu. Vložení dat je možné provést buď manuálně, zapsáním proteinové sekvence do formuláře, nebo importem ze vstupního souboru. Formát vstupu odpovídá běžným zvyklostem, tj. jde o řetězec jednopísmenných zkratek aminokyselin. Při načítání dat zvládá aplikace soubory ve formátu FASTA a TXT (prostý text, angl.„plain-text“). Při zadání pořadí aminokyselin (nukleotidové sekvence jsou automaticky přeloženy kliknutím na tlačítko „Rewrite DNA seq.“ – přepsat DNA sekvenci) jsou vedle okna formuláře (obr. 4) zobrazovány hodnoty parametrů vypočítaných z vložené sekvence: monoisotopová a průměrná molekulová hmotnost, celkový počet aminokyselin, indexy GRAVY (pozitivní hodnota – hydrofobní protein/peptid, negativní hodnota – hydrofilní protein/peptid) a NPS (zastoupení hydrofobních aminokyselin v rozmezí 0–1), hodnota isoelektrického bodu. Údaje se aktualizují automaticky a v případě vložení chybného symbolu je uživatel na tuto chybu upozorněn. K vloženému řetězci lze volitelně zobrazit i statistiku (část „Statistics“), která uvádí kvantitativní zastoupení jednotlivých aminokyselin (v tabulce i graficky) a také si nechat zobrazit pozice jednotlivých uživatelem zvolených aminokyselin (část „Visualizer“).

Druhým krokem je nastavení možností výpočtů. Aplikace poskytuje možnost nastavit, zda bude vypočítávána ionizovaná či neutrální forma peptidu ($[M+H]^+$, M, $[M-H]^-$),



Obr. 4. **Příklad vstupního okna programu Protein Cutter s vloženou aminokyselinovou sekvencí proteinu a zadáním uživatelem zvolených pravidel štěpení**

dále fixní modifikace cysteinu (bez modifikace, karbamidomethylace CysCAM či karboxymethylace CysCM) a také variabilní oxidace methioninu na methioninsulfoxid (hmotnostní nárůst o 16 Da). Významnou volbou je potom možnost nechat vypočítat také vynechaná štěpení („missed cleavages“), kdy aplikace vypočte všechny možné kombinace peptidů, které mohou vzniknout při opomenutí jednotlivých štěpných míst proteolytickým enzymem (např. z důvodu prostorové nepřístupnosti). V grafickém rozhraní je možné nastavit až čtyři vynechaná štěpení a vnitřní funkce umí neomezený počet, nicméně pro praxi mají význam hodnoty 0, 1 a 2.

Ve třetím kroku je nutné definovat pravidla pro štěpení (specifičnost enzymu vůči aminokyselinovým zbytkům v proteinech). Z připravené nabídky se vybere konkrétní proteolytický enzym, kde je známa specifičnost štěpení (aplikace si pravidla nastaví automaticky), nebo je možné si zvolit vlastní pravidla, tedy vybrat aminokyseliny, kde lze štěpení předpokládat. K rychlému výběru je připraveno devatenáct nejčastěji používaných enzymů, pro uživatelské zadání jiných pravidel štěpení je k dispozici jednoduché rozhraní. V posledním kroku se provádí nastavení filtrů výstupu. Použije se rozhraní, které umožní zadat, jaké peptidy (aminokyselinové sekvence) budou zobrazeny. Pro každý z šesti parametrů počítaných ze sekvence je možné nastavit intervaly hodnot „větší než“, „menší než“ nebo rozsah „od – do“. Ve výpisu se potom objeví jen ty aminokyselinové sekvence peptidů, které splní všechny podmínky definované ve filtrech (obr. 4).

Výstupem z výpočtu je tabulka s aminokyselinovými sekvencemi teoretických štěpných peptidů (obr. 5). U každého peptidu jsou dále zobrazeny hodnoty (případně filtrované hodnoty) vypočítaných parametrů, jmenovitě monoizotopová hmotnost (angl. „monoisotopic mass“), průměrná hmotnost (angl. „average mass“), délka řetězce, index GRAVY, index NPS a isoelektrický bod. U hmotnostních parametrů se na základě nastavení možností výpočtů (viz výše) započítávají modifikace methioninu a cysteinu. Pro potřeby uživatele je možné nechat výstup seřadit vzestupně nebo sestupně podle hodnot jednotlivých parametrů. K dispozici je i zobrazení dat vhodné pro tisk s možností skrytí zadávacího formuláře. Každý z teoretických peptidů lze podrobit dalšímu teoretickému štěpení, je tedy možné simulovat současné působení více proteolytických enzymů.

6.3. Technologie aplikace

Uživateli stačí jakýkoliv moderní prohlížeč a funkční připojení k Internetu. Aplikaci není možné provozovat bez přístupu k příslušnému serveru, ten však může být instalován lokálně. Samotná aplikace má z technického hlediska dvě části – část na straně uživatele (to, co vidíme v prohlížeči) a část na straně serveru, která realizuje samotné výpočty.

Část na straně uživatele je grafickým výstupem ze serverové části. Toto tzv. uživatelské rozhraní je validováno dle standardů W3C, což zajišťuje správné zobrazení ve

Sort by: position | alphabetically | monoisotopic mass | average mass | length | hydropathicity | NPS | isoel.point

Position	Fragment string	Mono.mass	Avg.mass	Length	Hydropathicity	NPS	Isoel.point	
61	AEGVETALKL	1029.5706	1030.1868	10	0.3400	0.3000	4.2475	cut this peptide
146	AFILEPIQGE	1115.5862	1116.2797	10	0.4900	0.5000	3.6155	cut this peptide
213	ALGGGILPVS AVL	1165.7070	1166.4269	13	1.8692	0.5385	5.9250	cut this peptide
145	AAFILEPIQG E	1186.6234	1187.3585	11	0.6091	0.4545	3.6155	cut this peptide
156	AGVVIPPDGY LK	1227.6863	1228.4544	12	0.4833	0.5833	6.1456	cut this peptide
32	AFYNDRFPVF	1274.6084	1275.4294	10	0.0000	0.6000	6.1465	cut this peptide
5	AVNQGHCHPK ILK	1443.7768	1444.7246	13	-0.4923	0.3077	9.4175	cut this peptide
47	ALFGYDMVLP MNTG	1527.7101 1Mo: 1543.7050 2Mo: 1559.6999	1528.8138 1Mo: 1544.8132 2Mo: 1560.8126	14	0.6286	0.4286	3.7750	cut this peptide
132	AIERIFKEKG DRV	1559.8783	1560.8175	13	-0.7615	0.3077	9.2064	cut this peptide
18	ALHDQADRLT VSSR	1567.8066	1568.7107	14	-0.6857	0.2143	7.1247	cut this peptide
182	ADEIQTGLAR TGKML	1602.8399 1Mo: 1618.8348	1603.8609 1Mo: 1619.8603	15	-0.2333	0.2000	6.2922	cut this peptide
168	AVRDLCSKYN VLMI	1623.8476 1Mo: 1639.8425	1624.9898 1Mo: 1640.9892	14	0.6571	0.4286	8.2746	cut this peptide
132	AIERIFKEKG DRVA	1630.9154	1631.8964	14	-0.5786	0.2857	9.2064	cut this peptide
130	AEAIERIFKE KGDRV	1759.9580	1761.0119	15	-0.7733	0.2667	6.4399	cut this peptide
197	ACDWEDVRPD VVILGK	1813.9032	1815.0809	16	-0.0063	0.4375	4.0647	cut this peptide
130	AEAIERIFKE KGDRVA	1830.9951	1832.0907	16	-0.6125	0.2500	6.4399	cut this peptide
1	SGYSAVNQGH CHPKILK	1837.9257	1839.1091	17	-0.5706	0.2941	9.3084	cut this peptide
32	AFYNDRFPVF AEYLT	1851.8832	1853.0644	15	0.0067	0.5333	4.1860	cut this peptide
71	ARKWGYEKKK IPNDE	1860.9846	1862.1194	15	-2.1333	0.2667	9.9296	cut this peptide

Obr. 5. **Příklad výstupního okna programu Protein Cutter s výsledky seřazenými vzestupně podle monoizotopové molekulové hmotnosti peptidů**

všech moderních internetových prohlížečích. Grafická část je vytvořena v jazycích XHTML, CSS a JavaScript. Dynamické funkce potom využívají knihovnu JQuery (MIT/GPL licence, jazyk JavaScript), která aplikaci umožňuje použít technologie AJAX. Pro kreslení grafu se využívá knihovna phpMyGraph (svobodná licence, jazyk PHP5).

V rozhraní jsou použity technologie označované jako Web2. Ty opouštějí klasickou koncepci webu, kdy kliknutí na aktivní prvek vyžaduje opětovné načtení stránky. Aplikace typu Web2 komunikují s webovým serverem tzv. „na pozadí“ a provádějí změny v obsahu stránky bez opakovaného načítání. Uživatel tak má pocit, že pracuje s běžnou (nikoliv webovou) aplikací. Takto fungují např. našeptávače (angl. „autocomplete“) nebo internetové mapy. V samotném Protein Cutteru se toto projevuje např. při automatickém počítání veličin vstupního řetězce, při signalizaci chybného vstupního znaku, nebo při nastavování pravidel štěpení a filtrů. Technologie Webu2 tak dávají aplikaci v řadě směrů zajímavé vlastnosti.

Serverová část aplikace je vytvořena ve skriptovacím jazyku PHP5. Pro její fungování je nutné použít webový server podporující jazyk PHP5. Nejvhodnějším typem takového webového serveru je samozřejmě Apache (který je také využíván), nicméně je možné použít i jiný. Pro samotný běh není třeba na serveru ani v konfiguraci PHP5 provádět žádná nestandardní nastavení nebo úpravy. Při tvorbě aplikace bylo pro výpočetní jádro použito objektové programování a došlo při tom k oddělení vzhledu od výpočetního jádra (logiky) aplikace. To do budoucna otevírá možnost použít již vytvořené funkce i pro jiné aplikace a to buď bez úprav, nebo jen s malými úpravami v kódu.

6.4. Čím je aplikace zajímavá

Aplikace v sobě spojuje množství funkcí, které dosud byly umístěny v několika nezávislých aplikacích. To umožňuje získat potřebné informace jediným výpočtem a na jednom místě. Výhodou je i správnost vypočítávaných parametrů. Různé už existující aplikace dávají při řešení stejných výpočtů rozdílné výsledky. Autoři proto dbali na to, aby výpočty byly co nejsprávnější. Konstanty, vzorce a algoritmy použité při výpočtech parametrů teoretických štěpných peptidů byly proto ověřeny z více zdrojů. Významným prvkem je univerzálnost. Uživateli se nabízí možnost nadefinovat si vlastní pravidla štěpení proteinů, což obdobné aplikace neumožňují. Takto bylo například zařazeno štěpení proteinů působením prolylendoproteasy z *Aspergillus niger*, přičemž predikce ukázala výbornou shodu s experimentálními výsledky[63]. Mezi dalšími přednostmi aplikace lze uvést i uživatelské prostředí. Vzhled aplikace samozřejmě není pro získané výsledky podstatný, ale vůči uživateli je ohleduplné, aby rozhraní bylo snadno použitelné a umožňovalo rychlé získání žádaných informací. Webové rozhraní aplikace bylo proto navrženo i s ohledem na přehlednost a uživatelskou přívětivost.

7. Závěr a možnosti dalšího vývoje aplikace

Aplikace Protein Cutter je osvědčeným nástrojem pro výpočty výsledků proteolytického štěpení proteinů a vyhodnocování hmotnostních spekter peptidů z proteolytických digestů[63]. Je tak vhodným pomocníkem pro výzkumnou práci v oblasti biochemie proteinů a proteomiky. V rámci pokračujícího vývoje je plánováno doplnění dalších funkcí. Může se jednat např. o přidání dalších uživatelsky volených výpočtů nebo o rozšíření počtu předvolených proteolytických enzymů. Zajímavým zdokonalením by bylo napojení aplikace na externí databáze proteinů. V technické části aplikace jsou také možnosti pro vylepšení grafického rozhraní a optimalizací zdrojových kódů s cílem zvýšit výkon. Uživatelé by určitě do budoucna uvítali i lepší výstup pro tisk a možnost uložit si vypočtená data ve formě souboru (hodnoty oddělené čárkou, angl. „Comma-Separated Values“, CSV). Pro širší využití je zvažováno i zpřístupnění funkcí výpočetního jádra pro jiné aplikace (vzdálené volání procedur; angl. „Remote Procedure Call“, RPC).

Autoři tímto děkují MŠMT za podporu projektu ED0007/01/01 Centrum regionu Haná pro biotechnologický a zemědělský výzkum.

LITERATURA

1. Smith A. D., Datta S. P., Smith G. H., Campbell P. N., Bentley R., McKenzie H. A. (ed.): *Oxford Dictionary of Biochemistry and Molecular Biology*. Oxford University Press, New York 2000.
2. Hofmeister F.: Naturwiss. Rundschau *17*, 529 (1902).
3. Sanger F.: Adv. Protein Chem. *7*, 1 (1952).
4. Zhang Y.: Curr. Opin. Struct. Biol. *18*, 342 (2008).
5. Chu B. C. H., Lee H.: Curr. Microbiol. *53*, 118 (2006).
6. Sanger F.: Annu. Rev. Biochem. *57*, 1 (1988).
7. Sanger F., Tuppy H.: Biochem. J. *49*, 463 (1951).
8. Sanger F., Tuppy, H.: Biochem. J. *49*, 481 (1951).
9. Sanger F., Thompson E. O. P.: Biochem. J. *53*, 353 (1953).
10. Sanger F., Thompson E. O. P.: Biochem. J. *53*, 366 (1953).
11. Crick F. H., Barnett L., Brenner S., Watts-Tobin R. J.: Nature *192*, 1227 (1961).
12. Edman P.: Acta Chem. Scand. *4*, 283 (1950).
13. Edman, P.: Acta Chem. Scand. *10*, 761 (1956).
14. Edman P., Begg G.: Eur. J. Biochem. *1*, 80 (1967).
15. Niall H. D.: Meth. Enzymol. *27*, 942 (1973).
16. Graham K., Shively J. E.: Anal. Biochem. *307*, 202 (2002).
17. Suckau D., Resemann A.: J. Biomol. Tech. *20*, 258 (2009).
18. Dole M., Mack L. L., Hines R. L., Mobley R. C., Ferguson L. D., Alice M. B.: J. Chem. Phys. *49*, 2240 (1968).
19. Fenn J. B., Mann M., Meng C. K., Wong S. F., Whitehouse C. M.: Science *246*, 64 (1989).

20. Veenstra T. D., Yates J. R.: *Proteomics for Biological Discovery*. J. Wiley, Hoboken, New Jersey 2006.
21. Tanaka K., Waki H., Ido Y., Akita S., Yoshida Y., Yoshida T.: Rapid Commun. Mass Spectrom. *2*, 151 (1988).
22. Karas M., Bachmann D., Hillenkamp F.: Anal. Chem. *57*, 2935 (1985).
23. Karas M., Hillenkamp F.: Anal. Chem. *60*, 2299 (1988).
24. Mann M., Højrup P., Roeppstorf P.: Biol. Mass Spectrom. *22*, 338 (1993).
25. Yates J. R. III, Speicher S., Griffin P. R., Hunkapiller T.: Anal. Biochem. *214*, 397 (1993).
26. Hunt D. F., Yates J. R. III, Shabanowitz J., Winston S., Hauer C. R.: Proc. Natl. Acad. Sci. U.S.A. *83*, 6233 (1986).
27. Spengler B.: J. Mass Spectrom. *32*, 1019 (1997).
28. Shevchenko A., Sunyaev S., Loboda A., Shevchenko A., Bork P., Ens W., Standing K. G.: Anal. Chem. *73*, 1917 (2001).
29. Zubarev R. A.; Kelleher N. L.; McLafferty F. W.: J. Am. Chem. Soc. *120*, 3265 (1998).
30. Suckau D., Resemann A.: Anal. Chem. *75*, 5817 (2003).
31. Bunger M. K., Cargile B. J., Ngunjiri A., Bundy J. L., Stephenson J. L. Jr.: Anal. Chem. *80*, 1459 (2008).
32. Kinter M., Sherman N. E.: *Protein Sequencing and Identification Using Tandem Mass Spectrometry*. J. Wiley, New York 2000.
33. Mullis K. B., Faloona F. A.: Methods Enzymol. *155*, 335 (1987).
34. Rabinow F.: *Making PCR: A Story of Biotechnology*, University of Chicago Press, Chicago 1996.
35. Balbas P., Lorence A. (ed.): *Recombinant Gene Expression: Reviews and Protocols (Methods in Molecular Biology, Vol. 267)*. Humana Press, Totowa 2004.
36. Sanger F., Nicklen S., Coulson A. R.: Proc. Natl. Acad. Sci. U.S.A. *74*, 5463 (1977).
37. Crick F. H. C., Barnett L., Brenner S.,Watts-Tobin R. J.: Nature *192*, 1227 (1961).
38. Osawa S., Jukes T. H., Watanabe K., Muto A.: Microbiol. Rev. *56*, 229, 1992.
39. Staden R.: Nucleic Acids Res. *6*, 2601, 1979.
40. Shizuya H., Birren B., Kim U. J., Valeria M., Slepak T., Tachiiri Y., Simon M.: Proc. Natl. Acad. Sci. U.S.A. *89*, 8794, 1992.
41. Ronaghi M., Uhlén M., Nyrén P.: Science *281*, 363 (1998).
42. Margulies M., Egholm M., Altman W. E., Attiya S., Bader J. S., Bemben L. A., Berka J., Braverman M. S., Chen Y. J., Chen Z., Dewell S. B., Du L., Fierro J. M., Gomes X. V., Godwin B. C., He W., Helgesen S., Ho C. H., Irzyk G. P., Jando S. C., Alenquer M. L., Jarvie T. P., Jirage K. B., Kim J. B., Knight J. R., Lanza J. R., Leamon J. H., Lefkowitz S. M., Lei M., Li J., Lohman K. L., Lu H., Makhijani V. B., McDade K. E., McKenna M. P., Myers E. W., Nickerson E., Nobile J. R., Plant R., Puc B. P., Ronan M. T., Roth G. T., Sarkis G. J., Simons J. F., Simpson J. W., Srinivasan M., Tartaro K. R., Tomasz A., Vogt K. A., Volkmer G. A., Wang S. H., Wang Y., Weiner M. P., Yu P., Begley R. F., Rothberg J. M.: Nature *437*, 376 (2005).
43. Boguski M. S.: Curr. Opin. Genet. Dev. *4*, 383 (1994).
44. Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J., Sayers E. W.: Nucleic Acids Res. *39*, D32 (2011).
45. Baxevanis A. D., Ouellette B. F. F. (ed.): *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. J. Wiley, New York 2005.
46. Attwood T., Parry-Smith D.: *Introduction to Bioinformatics*. Prentice Hall, Harlow 1999.
47. Wheeler D. L., Church D. M., Lash A. E., Leipe D. D., Madden T. L., Pontius J. U., Schuler G. D., Schriml L. M., Tatusova T. A., Wagner L., Rapp B. A.: Nucleic Acids Res. *29*, 11 (2001).
48. http://www.proteomics.leeds.ac.uk/bioinf/msdb.html (staženo 25.3.2012)
49. Bleasby A. J., Akrigg D., Attwood T. K.: Nucleic Acids Res. *22*, 3574 (1994).
50. Sigrist C. J. A., Cerutti L. de Castro1 E., Langendijk-Genevaux P. S., Bulliard V., Bairoch A., Hulo N.: Nucleic Acids Res. *38*, D161 (2010).
51. Bairoch A.: Nucleic Acids Res. *28*, 304 (2000).
52. Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E.: Nucleic Acids Res. *28*, 235 (2000).
53. Berman H. M., Olson W. K., Beveridge D. L., Westbrook J., Gelbin A., Demeny T., Hsieh S. H., Srinivasan A. R., Schneider B.: Biophys. J. *63*, 751 (1992).
54. Rawlings N. D., Barrett A. J., Bateman A.: Nucleic Acids Res. *38*, D227 (2010).
55. Bairoch A., Apweiler R.: Nucleic Acids Res. *31*, 360 (2000).
56. George D. G., Barker W. C., Mewes H. W., Pfeiffer F., Tsugita A.: Nucleic Acids Res. *24*, 17 (1996).
57. Wu C. H., Apweiler R., Bairoch A., Natale D. A., Barker W. C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M. J., Mazumder R., O'Donovan C., Redaschi N., Suzek B.: Nucleic Acids Res. *34*, D187 (2006).
58. Pruitt K. D., Tatusova T., Maglott D. R.: Nucleic Acids Res. *35*, D61 (2007).
59. http://www.ncbi.nlm.nih.gov/RefSeq/ , staženo 25.3.2012.
60. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M. R., Appel R. D., Bairoch A., v knize: *The Proteomics Protocols Handbook* (Walker, J. M., ed), str. 571. Humana Press, New York 2005.
61. Strohalm M., Hassman M., Košata B., Kodíček M.: Rapid Commun. Mass. Spec. *22*, 905 (2008).
62. Strohalm M., Kavan D., Novák P., Volný M., Havlíček V.: Anal. Chem. *82*, 4648 (2010).
63. Šebela M., Řehulka P., Kábrt J., Řehulková H., Ožd'an T., Raus M., Franc V., Chmelík J.: J. Mass Spectrom. *44*, 1587 (2009).

M. Raus, D. Kopečný, and M. Šebela *(Department of Biochemistry and Centre of Biotechnological and Agricultural Research, Faculty of Sciences, Palacký University, Olomouc):* **Program Application for the Prediction of Results of Protein Digestion by Proteolytic Enzymes**

Here we introduce the Protein Cutter (http://biochemie.upol.cz/software/proteincutter), a web application for the prediction of results of protein digestion by proteolytic enzymes, which is accessible over the Internet network. In the beginning, previous and current approaches for protein sequencing are summarized. This includes the use of dinitrofluorobenzene and substituted isothiocyanate reagents as well as mass-spectrometry-based strategies and translation of genomic sequences. The following text characterizes bioinformatics as a modern scientific discipline, which solves problems arising from the management and analysis of biological data. The most important nucleotide and amino acid sequence databases are described together with the databases of DNA and protein structures. The program Protein Cutter, which is described in detail with respect to its design and technology, allows predicting peptide sequences generated by proteolytic digestion of a protein (represented by a user-entered amino acid or coding nucleotide sequence). In addition to other comparable applications, Protein Cutter offers more complex information calculated from amino acid sequences (i.e. molecular mass, amino acid composition, isoelectric point, hydropathicity index etc.), it works with nucleotide sequences upon automatic translation, it is open and friendly for user-entered cutting rules and provides more options for the filtration and sorting of results.

OPPK OPERAČNÍ PROGRAM PRAHA KONKURENCESCHOPNOST

PRAHA PRAGUE PRAGA PRAG

EVROPSKÁ UNIE

Mikrobiologický ústav AV ČR, v.v.i. v rámci 4. výzvy Operačního programu Praha – Konkurenceschopnost realizoval projekt

Pražská infrastruktura pro strukturní biologii a metabolomiku (PISBM) CZ.2.16/3.1.00/24023.

Vytvořením nového výzkumného centra jsou do stávající infrastruktury biologického areálu Akademie věd v Praze 4-Krči implementovány špičkové technologie instrumentální analýzy, nezbytné pro udržení strukturně-biologických a biomedicinálních vědeckých skupin na úrovni srovnatelné s vyspělými státy. V rámci realizace projektu byl instalován NMR spektrometr s protonovou pozorovací frekvencí 700 MHz a hmotnostní spektrometr s iontově cyklotronovým hmotnostním analyzátorem (FT-ICR-MS) s magnetickým polem 12 T. Rovněž byl rozšířen stávající NMR spektrometr 600 MHz o spojení s kapalinovou chromatografií, extrakcí na pevné fázi a hmotnostní spektrometrií (HPLC-SPE-NMR-MS). Nedílnou součástí realizace projektu byla nezbytná rekonstrukce a modernizace objektu L v areálu Akademie věd v Praze 4-Krči.
Partnery projektu, jehož finanční objem přesáhl 94 milionů korun, jsou Vysoká škola chemicko-technologická v Praze, Univerzita Palackého v Olomouci a Univerzita Karlova v Praze. Průběh realizace projektu je možno sledovat na stránkách: http://ms.biomed.cas.cz/oppk.php nebo http://www.biomed.cas.cz/mbu/cz/oppk.php. Zde naleznete i aktuální informace o programu slavnostního otevření infrastruktury, které proběhne v polovině ledna 2013.

Evropský fond pro regionální rozvoj
Praha a EU – Investujeme do vaší budoucnosti