UNIVERZITA PALACKÉHO V OLOMOUCI

Přírodovědecká fakulta
Katedra biochemie

DISERTAČNÍ PRÁCE

Softwarová analýza a predikce v hmotnostní spektrometrii

Autor:	**Mgr. Martin Raus**
Studijní program:	P1416 Biochemie
Studijní obor:	Biochemie
Forma studia:	Prezenční
Vedoucí práce:	**prof. Mgr. Marek Šebela, Dr.**

Prohlašuji, že jsem předloženou diplomovou práci vypracoval samostatně za použití uvedených zdrojů.

V Olomouci dne 15. 7. 2013

Martin Raus

Poděkování

Děkuji prof. Mgr. Marku Šebelovi, Dr. za odborné vedení, cenné rady, připomínky a všechny znalosti, které mi předal. Byl mi vedoucím nejen formálně, ale práce pod jeho vedením měla cíle, směr a postup a na konci byly výsledky, které mají skutečně význam. A to je něco, čeho si na základě řady životních zkušeností velmi vážím.

Můj dík patří i celému kolektivu Katedry biochemie Univerzity Palackého v Olomouci, Centra regionu Haná v Olomouci a Technické Univerzity Vídeň a zvláště všem, se kterými jsem měl možnost spolupracovat. Musím jim poděkovat za vstřícný přístup i rozšíření svých obzorů, stejně tak jako za přátelskou pracovní atmosféru.

V neposlední řadě musím poděkovat i mým rodičům a rodině, kteří za mnou stáli po celou dobu studia a byli mi podporou v mnoha směrech. Ačkoliv nerozuměli tomu, čím jsem se zabýval, přáli mi štěstí a úspěch a věřili, že ve svém studiu uspěji, což bylo zavazující i pro mne.

Věnování

Tuto práci bych chtěl věnovat památce Ing. Petra Orálka (1968 – 2008). V životě člověk potká řada pedagogů, ale s opravdovým učitelem se setká jen jednou. Pokud vůbec má to štěstí takového člověka potkat. Dal mi výborné základy pro obor, který jsem si zvolil. Jeho největší dar však tkví v tom, že mne naučil určitému přístupu a pohledu na svět kolem.

Odešel nečekaně a svět tím ztratil výjimečného člověka.

Čest jeho památce

Bibliografická identifikace

Jméno a příjmení autora	Mgr. Martin Raus
Název práce	Softwarová analýza a predikce v hmotnostní spektrometrii
Typ práce	Disertační
Pracoviště	Katedra biochemie
Vedoucí práce	prof. Mgr. Marek Šebela, Dr.
Rok obhajoby práce	2013
Abstrakt	V teoretické části této práce se pojednává o současné bioinformatice. Popisuje se její význam a studovaná problematika. Jsou zde zmíněny i prostředky, které bioinformatika využívá. Zvláštní důraz je věnován na propojení s oblastí biochemie, proteomiky a hmotnostní spektrometrie v biologii. Praktická část této práce se zabývá softwarovou analýzou dat a predikcí v hmotnostní spektrometrii. Skládá se ze tří kapitol. Každá kapitola se věnuje jednomu z řešených problémů a představuje softwarové aplikace, které byly pro daný účel vytvořeny. První z kapitol se zabývá počítačovou predikcí proteolytického štěpení proteinů. Součástí je i popis aplikace ProteinCutter, která je osvědčeným nástrojem pro predikci štěpení proteinů. Tato aplikace v sobě spojuje schopnosti několika starších aplikací. Navíc přidává další funkce, které nejsou běžně k dispozici. Druhá (nejkratší) kapitola, se věnuje aplikaci Izotop pro výpočet rychlosti biosyntézy cytokininů. Třetí kapitola je nejrozsáhlejší a věnuje se softwarové analýze a porovnání hmotnostních spekter. Zabývá se problematikou rozpoznání píků v hmotnostním spektru. Dále je rozveden postup, jak mezi sebou porovnávat hmotnostní spektra, což je stěžejním tématem této části. Součástí je i popis aplikace Biospean, která byla vyvinuta pro účely hromadného porovnávání hmotnostních spekter. Tato aplikace již byla s úspěchem vyzkoušena pro řešení praktických úkolů. Zvláštní pozornost je věnována třem specifickým případům, kdy byla aplikace Biospean využita způsobem, jaký nebyl na počátku jejího vývoje vůbec plánován.
Klíčová slova	ProteinCutter, Biospean, hmotnostní spektrometrie, spektrum, bioinformatika, biochemie, predikce štěpení, analýza
Počet stran	111
Počet příloh	6
Jazyk	český

Bibliographical identification

Autor's first name and surname	Mgr. Martin Raus
Title	Software analysis and prediction in mass spectrometry
Type of thesis	PhD
Department	Department of biochemistry
Supervisor	prof. Mgr. Marek Šebela, Dr.
The year of presentation	2013
Abstrakt	The theoretical part of this thesis deals with current bioinformatics describing its significance and issues that are studied. There are also bioinformatics tools mentioned. A special emphasis is placed on linking the discipline with the areas of biochemistry, proteomics and biological mass spectrometry. The experimental part concerns software-based data analysis and prediction in mass spectrometry. It consists of three chapters. Each chapter addresses a practical task and introduces a software application that has been developed for the given purpose. The first chapter deals with computer prediction of a protelytic cleavage of proteins. It also includes a description of the application ProteinCutter, which proved to be a reliable tool for the prediction of protein digestion. This application combines the capabilities of several previous applications with some additional features, which are not commonly available. The second (and shortest) chapter introduces the software application Izotop, which allows to calculate the rate of cytokinin biosynthesis. The third chapter is the largest one and it is dedicated to software analysis and comparison of mass spectra. It deals with the issue of peak identification in a mass spectrum. Furthemore, there is a procedure for comparing mass spectra elaborated in detail, which is the central theme of this text section. The chapter also includes a description of the application Biospean that has been developed for extensive comparisons of mass spectra. This application has been successfully tested to solve practical tasks. A special attention is paid to three specific cases where the Biospean was used in a way that had not been expected at all in the beginning of its development.
Keywords	ProteinCutter, Biospean, mass spectrometry, spectra, bioinformatics, biochemistry, prediction of protein cleavage
Number of pages	111
Number of appendices	6
Language	czech

Obsah

1 Úvod

Bioinformatika je mezivědním oborem, který v posledních dvou desetiletích nabývá stále většího výnamu při podpoře výzkumu v biologických oborech. Tato práce se v první části věnuje krátkému shrnutí činností, kterými se bioinformatika zabývá. Z prostorových důvodů nepůjde ani v nejmenším o podrobné, natož o vyčerpávající shrnutí.

Stěžejní částí této práce je totiž praktické využití bioinformatiky, konkrétně vývoj specializovaného softwaru pro potřeby biologického výzkumu na Katedře biochemie PřF UP v Olomouci.

Prvním popisovaný softwarem je webová aplikace ProteinCutter, která byla vytvořena s cílem získat nástroj pro predikci štěpení proteinů. Cílem přitom bylo splnit specifické požadavky KBC PřF UP, kterým nevyhovoval žádný stávající software. Dalším požadavkem bylo spojit do jediného nástroje funkce, které byly dostupné i v jiných nástrojích, ale žádný z nástrojů neobsahoval všechny z požadovaných funkcí. Ve výsledku tak vznikl software, který je v mnoha ohledech unikátní a současně v sobě spojuje schopnosti, které byly doposud dostupné na více nástrojích.

Druhým softwarem je jednoúčelová aplikace Izotop, jejímž úkolem je výpočet rychlosti biosyntézy cytokininů z hmotnostně spektrometrických dat. Tato aplikace je méně významná, je zde ale uvedena, neboť byla součástí publikovaného článku, který je přílohou disertační práce.

Nejvýznamnějším softwarovým nástrojem je však aplikace Biospean, které je zde věnována největší část. Slouží k analýze a porovnávání hmotnostních spekter, přičemž podnětem k jejímu vytvoření byl nedostatek vhodného a dostupného softwaru schopného porovnávat hmotnostní spektra hromadně. Aplikace umožňuje nahradit podobně zaměřené komerční aplikace, nemá některé jejich nevýhody a je více zaměřena na specifické potřeby KBC PřF UP. Svým pojetím se v mnoha ohledech vymyká ostatním aplikacím z této oblasti.

Všechny z uvedených aplikací byly již použity k výzkumným účelům, přičemž získané výsledky jsou součástí publikovaných článků, které jsou přílohou této práce.

2 Cíle práce

Hlavní cíle předložené práce byly následující:

1) Zpracování souhrnného přehledu o roli bioinformatiky v současné vědě, úkolech které plní a prostředcích, které používá. S ohledem na zaměření praktické části práce je zvláštní důraz věnován na využití bioinformatiky pro biologii, zvláště pro proteomiku a hmotnostní spektrometrii.
2) Vytvoření softwaru pro predikci proteomického štěpení. Výsledkem je webová aplikace **ProteinCutter**, které se věnuje druhá část této práce.
3) Vytvoření softwaru pro výpočet rychlosti látkové výměny (syntéza cytokininů). Výsledkem je webová aplikace **Izotop**, které je věnována třetí část práce.
4) Vytvoření softwaru pro analýzu a hromadné porovnání hmotnostních spekter. Tento úkol je nejvýznamnější částí této disertační práce. Výsledkem je webová aplikace **Biospean**, které je věnována čtvrtá část práce.

V době odevzdání této práce jsou uvedené aplikace veřejně přístupné na webových stránkách Katedry biochemie PřF UP (http://biochemie.upol.cz/soft.php).

3 Bioinformatika jako vědní obor

3.1 Bioinformatika a důvody jejího vzniku

Bioinformatika je vědní disciplína s relativně krátkou historií. Z jejího názvu lze usoudit, že se jedná o hraniční (mezioborovou) vědu mezi biologickými obory (biologie - nauka o živých organizmech) a informatikou (vědy zabývající se získáváním, uchováváním, zpracováním, využitím a prezentací dat).

Definice bioinformatiky není jednoznačná, což ukazuje i část rozhovoru *„Bioinformatika - na půl cestě mezi algoritmy a životem“* s Doc. Fatimou Cvrčkovou z února 2006 [BWord 2006]:

> ***Jak bioinformatiku chápete vy sama, respektive v jaké podobě se jí zabýváte a co v rámci tohoto předmětu vyučujete?***
>
> *Napadlo mne onehdy najít si českou definici bioinformatiky na Googlu. Na jednom z prvních míst na mě vyběhl příspěvek z elektronické konference Záhady, podrubrika Psychotronika, podle kterého je bioinformatika obor, který se zabývá jevy, jako je hypnóza, telepatie, dálkové vnímání... Jak vidno, s definicí bioinformatiky to zjevně není tak úplně jednoduché.*
>
> *Mně se líbí dosti rozšířené pojetí, podle něhož je bioinformatika oblast na pomezí biologie a informatiky, která se zabývá zpracováním, prohledáváním a analýzou dat o sekvenci, struktuře a popřípadě i funkci biologických makromolekul, tedy hlavně DNA a proteinů.*
>
> *V literatuře se ale dá najít i vymezení širší – totiž že do bioinformatiky patří jakékoli využití počítačů k hledání odpovědí na biologické otázky. Pak by tam spadlo také statistické zpracování fyziologických, lékařských a dokonce i ekologických dat. Nemyslím ale, že by toto širší vymezení (odpovídající zhruba anglickému termínu "computational biology") bylo k něčemu dobré. To jsou jiná témata, dělají to obvykle jiní lidé a jinými metodami, nanejvýš někdy na stejném hardwaru. Možná se taková definice může hodit třeba k vymezení struktury pracoviště či ke zdůvodnění žádosti o peníze, když se zrovna nosí mezioborovost, ale jinak sotva.*

Takže i z citovaného je zřejmé, že jednoznačné vymezení není a některé definice mohou být víceméně úsměvné. O nejednoznačném vymezení definice bioinformatiky se hovoří i na Wiki stránce portálu Bioinformatics.org [Bioinformatics], kde je uvedeno [BioinformaticsWiki]:

> *Bioinformatics has been defined many different ways, since practitioners do not always agree upon the scope of its use within the biological and computer sciences, but it is always considered a combination of both sciences, along with other contributing disciplines.*

Česky:

> *Bioinformatika je definována mnoha různými způsoby, protože odborníci se ne vždy dohodnou na rozsahu použití v rámci biologie a informatiky, ale je vždy považována za kombinaci obou věd, spolu s dalšími přispívajícími obory.*

Uveďme si pro srovnání jednu z definici v češtině [CzWiki Bi]:

> *Bioinformatika je vědní disciplína, která se zabývá metodami pro shromáždování, analýzu a vizualizaci rozsáhlých souborů biologických dat, zejména dat molekulárně-biologických.*
>
> *Předmětem zájmu a používanými metodami se bioinformatika prolíná s dalšími příbuznými obory, např. molekulární biologií, genomikou, proteomikou, genetikou, výpočetní biologií, matematickou biologií, systémovou biologií, teoretickou biologií, biomedicínskou informatikou, biomedicínským inženýrstvím, výpočetní chemií, informatikou a počítačovou lingvistikou.*

Z této široké definice vyplývají jednak nástroje a jednak oblasti působení bioinformatiky.

V první část definice uvádí, co jsou pracovní prostředky bioinformatiky. Jedná se o využití výpočetní techniky, informatiky a příbuzných oborů s cílem získat, vyhodnotit, reprezentovat a vizualizovat biologická data tak, aby byla přijatelným způsobem srozumitelná pro člověka. Bioinformatika tedy nepoužívá laboratoř, ale pracuje *in silico* (na počítači či jiném zařízení tohoto druhu), kde zpracovává data, která pouze reprezentují to, co se odehrává ve skutečnosti, ať už *in vivo* nebo *in vitro* (naživo nebo laboratorně).

Pracovními nástroji bioinformatiky jsou potom výpočetní technika a metody z oblasti informatiky (technologie, nástroje a algoritmy) a s ní blízce příbuzných oborů – matematiky, statistiky, fyziky, inženýrství a dalších vědních, technických a technologických oborů.

Druhá část definice specifikuje jaký druh dat se zpracovává. Není toho málo. Do oblasti působení bioinformatiky snadno zahrneme velice široký záběr činností, počínaje výzkumem molekul přes oblast proteomiky, genetiky, různé druhy počítačového modelování až po software, který řídí moderní lékařské přístroje. Svým způsobem začíná bioinformatika u digitálního lékařského teploměru a končí u zařízení za stovky milionů, která slouží např. pro krystalografii.

Zde je třeba si uvědomit, že v dnešní době výpočetní techniky oblast působení bioinformatiky začíná de facto už v samotné laboratoři (nebo na jiném místě, kde se zkoumaný děj odehrává), kde nějaké zařízení (senzor) získá data reprezentující sledovanou skutečnost a v digitalizované formě je zaznamená. Tato data jsou poté zpracována, uložena a dále využita, vyhodnocena či interpretována dalšími zařízeními. Většinou s cílem, aby na konci celého procesu byl výstup ve formě, která je srozumitelná a použitelná pro člověka.

Z uvedeného tedy vyplývá, že bioinformatika je „jen“ podpůrný obor, jakási servisní služba, která převezme data od specialistů primárních oborů (chemiků, biologů, biochemiků, lékařů), zpracuje je a vyhodnocené výsledky ve srozumitelné formě vrátí těmto specialistům k interpretaci.

Historické důvody pro vznik bioinformatiky (a dalších podpůrných oborů) vyplývají z rozvoje technologií užívaných ve vědeckém výzkumu a z problémů, které tím vyvstaly. Technologický pokrok v posledních desetiletích otevřel většině exaktních oborů dosud nevídané možnosti v podobě laboratorních přístrojů a měřících zařízení, které dokáží poskytnout vědcům obrovské množství údajů.

A právě množství změřených dat je paradoxně zdrojem problému. Přístroje doslova chrlí záplavu syrových dat v podobě čísel, která v jakési formě reprezentují reálné skutečnosti. Přístroje snadno vyprodukují megabyty dat, které se v rámci celého projektu snadno změní v gigabyty či terabyty. A není v lidských silách je zpracovat ručně. V mnoha oborech získáváme data tak rychle, že je nestačíme zpracovávat. Dá se říci, že pro většinu exaktních vědy je v současnosti největší brzdou rychlejšího pokroku právě nedostatek kapacit pro vyhodnocování získaných dat.

Proto se k většině exaktních oborů přidružují podpůrné obory, které se zabývají právě zpracováním dat pomocí výpočetní techniky. Jejich úkol je jediný – přefiltrovat pomocí důmyslných algoritmů data tak, aby lidský prvek umístěný na konci celého zpracování byl ušetřen manuálního zpracování dat a dostal jen takové informace, které potřebuje pro rozumové

vyhodnocení a vyvození závěrů. Přestože stroje za nás nedokáží kreativně myslet (a doufejme že toho nikdy nebudou schopny), je dnes jejich role při automatizovaném vyhodnocení výsledků nezastupitelná.

V případě biochemie můžeme jako příklad uvést hmotnostní spektrometrii. Z jediného zkoumaného vzorku získáme data v řádu jednotek megabytů. Řádově podobné velikosti dat získáme i z metod, které jsou založené na zpracování a analýze obrazu. Například vícerozměrná chromatografie či rentgenová krystalografie. V případě skenu např. z gelového skeneru dosahuje velikost dat snadno hodnot v jednotkách až desítkách megabytů. Pouhých tisíc vzorků pak představuje v součtu obrovské množství dat.

Stejným problémům čelíme např. při analýze sekvencí DNA či studiu proteinů, kdy pracujeme se sekvencemi, které obsahují tisíce prvků. A opět se jedná o situaci, která bez výpočetní techniky a vysoké míry automatizace není řešitelná.

Pro představu o tom, o jak velkém objemu dat se bavíme, uveďme, že kdybychom si chtěli na tiskárně vytisknout data o velikosti jednoho megabytu, neboli 1 048 576 Bytů (každý odpovídající některému z 256 znaků ASCII tabulky), vyšlo by z tiskárny asi 582 normostran[1] textu. Běžné CD (700 MB) tak může obsahovat více než 400 tisíc stran textu a DVD nosič (4,7 GB) již 2,8 milionu stran textu.

Výzkumný projekt tedy může snadno vyprodukovat a vyžadovat zpracování dat, jejichž objem je ekvivalentní milionům stran textu. Takové množství dat opravdu nelze zpracovat „ručně“ a výzkum je v takovém případě na využití výpočetní techniky přímo závislý.

Z výše uvedeného je tedy zřejmé, že bez různých forem bioinformatiky se současné biologické obory neobejdou a její význam dále poroste s potřebou zpracovávat stále větší množství dat. A ačkoliv bioinformatika sama o sobě není klasickou vědou a nepřináší klasický vědecký pokrok tak jak je laicky vnímán (tj. posunování vědeckých poznatků), je podpůrným nástrojem zasahujícím do mnoha oborů, který klasické vědě (tak jak je tradičně chápána) otevírá prostředky právě pro onen klasický vědecký výzkum.

Tato práce se ovšem bude zabývat pouze úzkou částí bioinformatiky a to zpracováním dat týkajících se biochemie, konkrétně biologických makromolekul. Ve skutečnosti je zaměření ještě užší a to na softwarovou analýzu dat získaných hmotnostní spektrometrií.

1 Pojem normostrana definuje stránku textu obsahující 30 řádků a 60 znaků na řádek, což je zhruba ekvivalentní straně ve formátu A4.

3.2 Kdo je bioinformatik

Je také na místě si položit otázku, kdo je bioinformatikem, resp. koho můžeme za bioinformatika považovat. Pokud se zamyslíme nad všemi oblastmi bioinformatiky, zjistíme, že stanovit přesnou hranici není možné. Současně je problém specialistu bioinformatika najít, protože málokdo má současně talent pro biologickou i technickou stránku bioinformatiky. Mezioborová odlišnost je opravdu velká a rozsah obou oborů příliš rozsáhlý.

V podstatě pouhým konzumentem bioinformatiky je zaškolený uživatel. Může to být vědec pracující s počítačovou aplikací stejně tak, jako třeba obsluha např. nemocničního tomografu. Ačkoliv taková osoba příliš nerozumí tomu, na jakém principu zařízení funguje, je zaškolena jak je použít, je schopna na své uživatelské úrovni provést běžné provozní nastavení a je schopna zadat vstupní data a převzít vypočtené výsledky. Takový uživatel vůbec nemusí rozumět informatice a samotné zařízení je pro něj jen černá krabička s tlačítky na ovládání. I když se jedná o využití bioinformatiky, takovou osobu za bioinformatika asi označíme jen stěží.

Za bioinformatika už může být považován pokročilý uživatel, který je odborníkem na biologický obor a současně je obeznámeným s informatikou a jejími technologiemi na takové úrovni, aby rozuměl tomu, jak zařízení nebo nástroj pracuje. Takový člověk dokáže nejen používat např. specializovaný software, ale dokáže jej na pokročilé úrovni i nastavovat, spravovat, modifikovat a v případě potřeby jej použít inovativním způsobem, např. tak, že spojí dohromady dva dosud nezávislé nástroje tak, aby spolupracovaly a vytvořily novou funkčnost. Takové použití už vyžaduje hlubší znalosti výpočetní techniky, znalost fungování počítače, alespoň základní povědomí o typech a struktuře datových souborů a další pokročilé znalosti, vč. základů programování.

Takového člověka už můžeme označit za bioinformatika. Zvládá svůj biologický obor na vysoké odborné úrovni a přitom rozumí výpočetní technice, ví jak jsou skutečnosti z reálného světa reprezentovány v počítači a dokáže s takovými daty pracovat takovým způsobem, že má nad zpracováním dat přehled a kontrolu. Takový bioinformatik „vznikne“ tak, že se biochemik (nebo podobný specialista) naučí něco více o počítačích. Přesto u něho bude převládat původní biologická odbornost a v oblasti informatiky mu bude chybět potřebná intuice a celkový rozhled.

Opačným případem bioinformatika je původně informatik, který se dostane ke zpracování biologických dat. U něj bude silnou stránkou znalost technologií a dostatečný přehled o všech

stránkách výpočetní techniky, ale bude silně chybět intuice a rozhled v biologickém oboru. Bude schopen vytvářet počítačové programy, řešit problémy pomocí matematických a statistických metod, kombinovat různé druhy nástrojů, zpracovávat data pomocí databází, sestavit a zprovoznit výpočetní techniku pro rozsáhlé výpočetní účely (např. clustery) a spoustu dalšího. Nicméně bude závislý na komunikaci s odborníky na biologickou stránku oboru.

Otázkou k diskusi tedy je, jestli je lepší bioinformatika „vytvořit" tak, že biologa/biochemika doškolíme na programování, nebo zda vzít informatika a naučit ho něco z biologie. Oba případy mají stejnou nevýhodu – oběma bude chybět znalost druhého oboru a především potřebná intuice a vztah k němu. Není to něco, co by se dalo brát na lehkou váhu. Každý člověk má přirozeně dané primární zaměření a jiný obor se sice může více či méně mechanicky naučit, ale těžko se v něm bude cítit jistý. Univerzální jedinec zvládající oboje by byla velká výjimka.

Jednoznačná odpověď co je lepší tedy není. Pokud budeme shánět spíše odbornou obsluhu již existujícího systému, je lepší varianta zaškoleného biologa. Pokud ovšem budeme potřebovat vytvořit nový nástroj, programovat, pokud bude nutné zpracovávat různorodá data v surové podobě a pokud bude třeba využívat výpočetní techniku kreativním způsobem a kombinovat různé postupy, je lepším řešením zaškolený informatik.

Určitou výhodou informatika v tomto případě je, že je díky povaze svého oboru zvyklý zcela běžně převádět obecné problémy z reálného světa do počítačové reprezentace. Informatik běžně řeší problémy typu potřebujeme evidenci skladu, potřebujeme analýzu dopravy, potřebujeme řídící software pro výtah, potřebujeme vizualizaci technických výkresů … Informatik je prostě zvyklý dívat se na různorodé situace v reálném světě technickým pohledem a interpretovat realitu do podoby algoritmu. Řešení bioinformatické úlohy vždy začíná převedením biologické reality do počítačové reprezentace. Jen tento první krok vyžaduje úzkou komunikaci s odborníky z biologického oboru. Zbytek už je jen rutinní práce s výpočetní technikou. A zde už má informatik proti biologicky vzdělanému kolegovi navrch.

V praxi nejefektivnějším řešením by zřejmě byl pracovní tým složený jednak z biologicky vzdělaného experta zaškoleného do výpočetní techniky a jednak z informatika zaškoleného do základů biologického oboru. Jejich spolupráce by eliminovala nevýhody, který by měli, pokud by pracovali samostatně.

3.3 Využití bioinformatiky v oblasti biochemie

Biochemie využívá služeb bioinformatiky v několika směrech. Nejčastěji zmiňovanou oblastí je sekvenční analýza, která je většinou dělena na dvě samostatné části. A to konkrétně na sekvenční analýzu proteinů a na sekvenční analýzu nukleových kyselin. Obě analýzy se zabývají studiem a porovnáváním sekvencí řetězců molekul, konkrétně proteinových a nukleových. A zde narážíme na dvojí vnímání této problematiky, na odlišný pohled z úhlu informatika a biochemika. Ačkoliv se po stránce informatiky a technické realizace jedná u obou sekvenčních analýz o prakticky jedno a totéž (zpracování textových řetězců a jejich porovnání), z hlediska biochemie jde o zcela odlišné oblasti.

Další významným polem působení bioinformatiky v oblasti biochemie je strukturní bioinformatika. Zabývá se studiem a modelováním prostorového uspořádání molekul. Zde se už z hlediska technické realizace jedná o poměrně náročnou disciplínu, která vyžaduje rozsáhlé zapojení matematiky a dalších oborů. Např. krystalografie využívá fyzikální principy a zpracování výsledků se neobejde ani bez těchto znalostí.

Nesmíme zapomenout ani na další významnou oblast, kde se bioinformatika také uplatňuje a tou je hmotnostní spektrometrie. Analýza a porovnání hmotnostních spekter má velký význam pro rychlou identifikaci zkoumaných vzorků a jejich více či méně automatizované zpracování výrazně zjednodušuje práci. Zpracování dat z hmotnostní spektrometrie je postaveno na analýze průběhu signálu, což je problém především z oblasti elektrotechniky. Hmotnostní spektrometrie přitom fakticky představuje zcela samostatný obor, který se netýká jen biochemie.

Analýza dat z hmotnostních spekter, jejich srovnání a predikce je i náplní tato práce.

Bioinformatika je samozřejmě využitelná i pro další úlohy. Namátkou můžeme zmínit například analýzu dat z různých separačních metod, např. gelové elektroforézy, kde se výpočetní technika uplatní např. při analýze skenů z gelového skeneru. Tím se dostáváme k další významné disciplíně informatiky, kterou je zpracování obrazu.

Pojďme se podrobněji podívat na jednotlivé oblasti.

3.4 Sekvenční data – záznam proteinů a nukleových kyseliny

Významnou součástí bioinformatiky je analýza sekvencí proteinů a nukleových kyselin, resp. dat, která tyto sekvence reprezentují.

Proteiny jsou biopolymery s relativní molekulová hmotností 10^3 až 10^6 složené z aminokyselin. V biologii a biochemii se pojmem aminokyseliny většinou rozumí alfa-L-aminokyseliny. Dvacet, resp. dvaadvacet, základních aminokyselin tvoří hlavní stavební složku všech proteinů. Tyto aminokyseliny spojené peptidovu vazbou vytváří dlouhé lineární řetězce, které jsou základem veškerého života. Počet existujících proteinů je velmi vysoký – díky dvaceti aminokyselinám je při délce řetězce n možné vytvořit n^{20} kombinací. To je číslo opravdu vysoké. Při pouhých deseti aminokyselinách (tzv. oligopeptid, což ani není považováno za protein) dostáváme 10^{20} (jednička s dvaceti nulami) různých kombinací. Posloupnost aminokyselin v každém proteinu je označovaná jako primární struktura. Znalost primární struktury je základem pro výzkum proteinu.

Nukleové kyseliny jsou makromolekulární látky, které tvoří polynukleotidové řetězce. Tyto řetězce zajišťují uchovávání a přenos genetických informací. Zatímco deoxyribonukleová kyselina (DNA) je v podstatě zdrojový kód programu pro tvorbu živých buněk a jejich činnost, ribonukleová kyseliny (RNA) má řadu podtypů, které plní různé specifické role. Nukleové kyseliny tedy nalezneme ve všech živých buňkách a virech. Tvoří je posloupnost nukleotidů, což jsou jsou fosforylované nukleosidy. Nukleotidy jsou základními stavebními kameny nukleových kyselin podobně jako aminokyseliny u proteinů. Jejich posloupnost (opět označována jako primární struktura) v sobě uchovává genetickou informaci.

Z hlediska počítačového zpracování je nutné oba tyto typy sekvencí nějakým způsobem reprezentovat ve formě, která je použitelná pro počítač. Zde je celá situace naštěstí velice jednoduchá. Jednak existuje mezinárodně uznávaný systém značení pomocí písmen (IUPAC kód) [Tabulka 1], který je snadno čitelný pro člověka a je současně vhodnou datovou formou použitelnou pro počítač. A druhou výhodou je, že proteiny i nukleotidy mají na úrovni primární sekvence jednoduchou lineární strukturu, která je snadno zpracovatelná počítačem.

Symboly jednotlivých nukleotidů a aminokyselin jsou následující:

nukleotidy – nukleové kyseliny

A	adenin
C	cytosin
G	guanin
T	thymin
U	uracil
R	A nebo G
Y	C nebo T
S	G nebo C
W	A nebo T
K	G nebo T
M	A nebo C
B	C nebo G nebo T
D	A nebo G nebo T
H	A nebo C nebo T
V	A nebo C nebo G
N	jakákoliv báze
. nebo -	mezera

aminokyseliny – proteiny

A	Ala	alanine
B	Asx	aspartát nebo asparagin
C	Cys	cystein
D	Asp	aspartát
E	Glu	glutamát
F	Phe	fenylalanin
G	Gly	glycin
H	His	histidin
I	Ile	isoleucin
K	Lys	lysin
L	Leu	leucin
M	Met	methionin
N	Asn	asparagin
P	Pro	prolin
Q	Gln	glutamin
R	Arg	arginin
S	Ser	serin
T	Thr	threonin
V	Val	valin
W	Trp	tryptofan
X	Xaa	jakákoliv aminokyselina
Y	Tyr	tyrosin
Z	Glx	glutamin nebo glutamát

Tabulka 1: Znaková reprezentace nukleotidů a aminokyselin dle IUPAC

zdroj: http://www.bioinformatics.org/sms2/iupac.html [26. 6. 2013]
Cvrčková F. (2006) Úvod do praktické bioinformatiky, str.16 ,Academia, Praha, 2006

Díky těmto dvěma skutečnostem (písmena jako symboly a lineární stuktura) je princip záznamu, uložení a zpracování sekvenčních dat v počítači zredukován na v praxi poměrně „triviální“ práci s textovými řetězci, kde záznam sekvence tvoří řada po sobě následujících písmen reprezentujících jednotlivé stavební prvky. Práce s textovým řetězcem sice není pro samotný počítač zcela triviální a rychlá (po stránce výpočetní režie je práce se znaky náročnější než s čísly), ale je již spolehlivě zvládnutá a prakticky ověřená.

Pro záznam struktury proteinové nebo nukleotidové sekvence se použije prostá posloupnost, která se uloží do souboru nebo databáze. Pokud se sekvence ukládá do souboru, jedná se o uložení v podobě prostého textu (plain text), konkrétně pomocí znakové sady ASCII [ASCII] (základních 128 znaků „počítačové abecedy“). Výhodou takového způsobu záznamu je, že lze otevřít v prakticky jakémkoliv textovém editoru[2].

2 Textový editor je program pro editaci neformátovaného textu. Příkladem takového velmi jednoduchého programu tohoto typu je např. Notepad (Poznámkový blok). Existují samozřejmě i další (mnohem mocnější) editory vhodné pro tento účel a schopné editovat obsah na úrovni jednotlivých znaků. Naprosto nevhodným nástrojem jsou ovšem textové procesory (např. MS Word), které nedovolí upravit samotný obsah a svým specifickým způsobem ukládají nejen textový obsah, ale i jeho formátování.

Samotná sekvence však často nestačí a je žádoucí mít v souboru uložené další informace. Proto bylo zavedeno několik různých formátů a jejich modifikací, které v rámci smluvených pravidel umožňují vložit do textu i další informace o sekvenci, např. název a další poznámky.

Oblíbeným a velmi rozšířeným formátem pro uložení je **formát FASTA** [FASTA 1]. Jde o textový formát [FASTA 2], který používá znaku „>“ k označení začátku obsahu. Přitom samotný označený řádek není sekvencí, ale pouze popisem. Je to fakticky hlavička, které obsahuje jednak název či jiné označení sekvence a může samozřejmě obsahovat i další informace, např. komentáře. Samotná sekvence začíná až na dalším řádku.

Příklad obsahu souboru ve formátu FASTA:

```
>gi|186681228|ref|YP_001864424.1| phycoerythrobilin:ferredoxin oxidoreductase
MNSERSDVTLYQPFLDYAIAYMRSRLDLEPYPIPTGFESNSAVVGKGKNQEEVVTTSYAFQTAKLRQIRA
AHVQGGNSLQVLNFVIFPHLNYDLPFFGADLVTLPGGHLIALDMQPLFRDDSAYQAKYTEPILPIFHAHQ
QHLSWGGDFPEEAQPFFSPAFLWTRPQETAVVETQVFAAFKDYLKAYLDFVEQAEAVTDSQNLVAIKQAQ
LRYLRYRAEKDPARGMFKRFYGAEWTEEYIHGFLFDLERKLTVVK
```

V samotné sekvenci jsou pouze znaky reprezentující jednotlivé nukleotidy nebo aminokyseliny. V souvislosti s tím je třeba si uvědomit, že se z hlediska počítače jedná o textový řetězec, resp. ASCII, tudíž se v něm mohou vyskytnout i různé bílé znaky (konec řádku – enter, mezera, tabulátor a další). Co se týče smluvené konvence, neměla by sekvence obsahovat prázdné řádky. Ty jsou chápány jako začátek nové sekvence. Samotné zalomení sekvence pomocí konce řádku (klávesa Enter) a případné formátování pomocí mezer nejsou zapovězené.

Relativně volná struktura klade samozřejmě větší nároky na software, který dokáže takový soubor načíst. Zde je vhodné zmínit, že program pracující s textovými řetězci, které může editovat i neznalý uživatel, by měl být alespoň částečně robustní. Jinými slovy měl by být schopen se vypořádat s nevhodnými znaky, prázdnými mezerami a dalšími znaky, které může nezasvěcený uživatel do sekvence přidat. Nejzákladnější ochranou by mělo být ignorování znaků, které do sekvence nepaří.

FASTA samozřejmě není jediný formát, který se pro zápis používá. Nicméně je nejzákladnějším způsobem jak zaznamenat sekvenční data. Záznam v podobě textového souboru s písmeny označujícími aminokyseliny nebo nukleotidy je základem. Jednodušší to už být nemůže. Existují samozřejmě další modifikace, které vznikly kvůli snaze zaznamenat další informace.

Svoje vlastní varianty datových formátů tak mají například sekvenční databáze DDJB [DDBJ], Genbank [GenBank] a EMBL-EBI [EMBL]. Tyto formáty jsou také textové, ale už mají svoji strukturu mnohem přísněji stanovenou. Musí mít např. definovanou hlavičku s popisem zdroje

(organizmus), je možné do nic zapsat informace o autorovi, poznámky, do které kategorie patří (viry, bakterie, rostliny, primáti, lidské vzorky), verzi a revizi souboru a řadu dalšího.

Samotné technické řešení využívá vložení smluvených klíčových slov. Například klíčové slovo „source“ označuje zdroj vzorku, ze kterého byla sekvence získána. Vznikám tím jakýsi logicky strukturovaný soubor, který má své logicky související části. Tento způsob zápisu je svým způsobem „samopopisný“, takže zdrojový soubor může snadno editovat i člověk zběžně seznámený se způsobem zápisu.

Například formát souborů pro databázi EMBL [EMBL-formát] má začátek každého řádku označen dvoupísmennou zkratkou, která označuje, co daný řádek znamená. Např. „ID“ je identifikátor, „DT“ datum, „KW“ klíčové slovo, „DE“ je definice, „OS“ znamená druh organizmu atd.

Pro počítače je čtení takového souboru samozřejmě náročnější. Počítač musí celý řetězec rozložit na jednotlivé prvky a vyseparovat z nich klíčové informace. Odborně mluvíme o tzv. parsování nebo též o syntaktické analýze [Molnár et al., 1987]. Tento proces je již složitější než čtení samotného FASTA souboru. V případě počítače vyžaduje více provozní režie a na čtecí algoritmus klade větší požadavky.

V případě složitějších či velmi volných pravidel pro zápis dat (tzv. gramatiky) může být parsovací algoritmus poměrně rozsáhlý. Problém není ani tak v řešitelnosti, ale především v otrocké pracnosti při sestavení analyzátoru. Pro tyto účely naštěstí existují i specializované programy schopné část gramatických pravidel vygenerovat (např. yacc, flex, bison). Z hlediska informatiky tedy jde o úlohu, která je sice pracná, ale v podstatě rutinní. Používá se již dlouho k mnoha účelům a po teoretické (resp. obecné) stránce je velmi dobře zvládnuta.

Záznam, datová reprezentace a zpracování proteinových a nukleotidových sekvencí tedy nepředstavuje zásadnější problém. Z hlediska informatiky se využívá zcela rutinních nástrojů.

3.5 Zpracování sekvenčních dat – hledání podobností

Jedním z nejvýznamnějších úkolů, které současná bioinformatika plní je porovnávání podobnosti sekvenčních dat. V obou případech sekvenčních dat (proteinů i nukleových kyselin) je snaha najít co nejdelší shodné subsekvence porovnávaných řetězců. Ze shodných podčástí DNA lze odvodit například příbuznost různých druhů organizmů.

Nalezení shodných úseků a vzájemné podobnosti se označuje jako **zarovnání** či **přiřazení** („alignment“) [Cvrčková 2006a], [Xiong 2006a], [Pietrokovski 1999], [Altschul 2011], [Zuker 2004]. Z hlediska informatiky se jedná o úkol nalézt nejdelší společné sekvence ve dvou porovnávaných textových podřetězcích. Řešení této úlohy není zcela triviální. Je třeba si uvědomit, že počítač nedokáže porovnávat různě dlouhé řetězce, ale pouze jednotlivé znaky (i když není problém vytvořit funkčnost, která toto zapouzdří). Porovnávání řetězců se provádí tak, že oba řetězce čteme současně znak po znaku a kontrolujeme, za se jednotlivé znaky shodují.

Porovnání proteinových a nukleotidových sekvencí má ovšem i další zádrhele. V první řadě nelze zaručit, že podobné části sekvence začnou na stejném místě. Jeden z řetězců může být např. jen fragment. Navíc se některé části mohou opakovat na více místech.

A nesmíme samozřejmě zapomenout ani na skutečnost, že pro potřeby biochemie musíme být schopni detekovat i řetězce, kde došlo ke změně na některé z pozic. Dvě dostatečně dlouhé sekvence lišící se jen v několika málo znacích jsou z hlediska biochemie také příbuzné. Typickým příkladem může být mutace u nukleové kyseliny, kde došlo k záměně některého z nukleotidů. Takovéto drobné mutace např. odlišují nejen jedince, ale např. i organizmy různých druhů, které ovšem patří ke stejnému rodu. Stupeň podobnosti a shody je v takovém případě klíčový ke stanovení příbuznosti.

Z uvedených důvodů má tedy porovnání biochemických sekvencí svoje specifika a vyžaduje poněkud odlišný přístup než při porovnávání běžných řetězců.

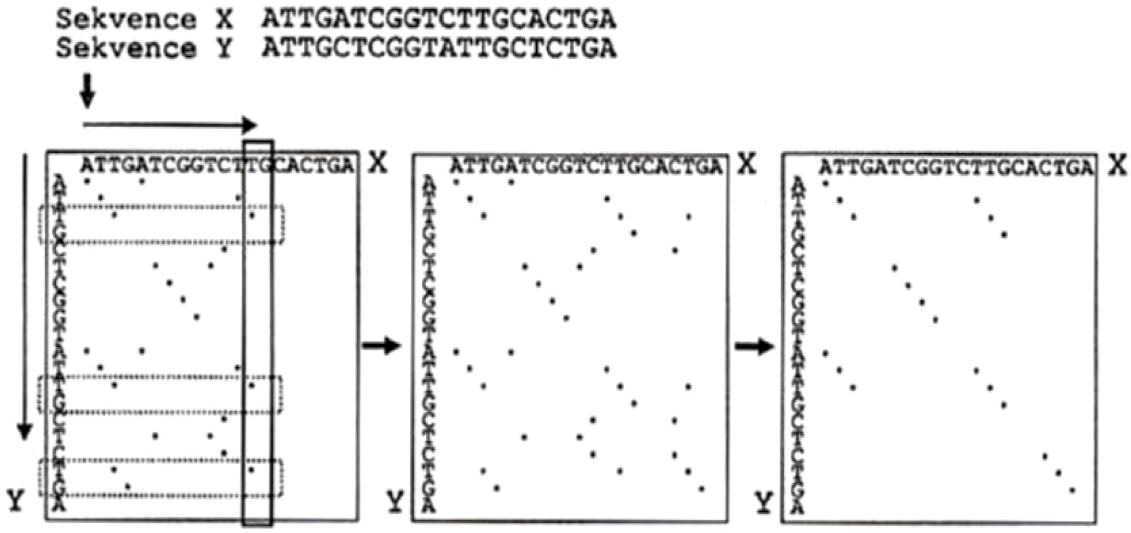
Prvním krokem pro porovnání dvou řetězců je nalezení shodných znaků. To se provede pomocí bodové matice (resp. bodového diagramu) [Obr 1]. V paměti počítače se použije matice, resp. dvourozměrné pole, z hlediska běžného uživatele se reprezentace provádí bodovým grafem/diagramem.

Princip spočívá v tom, že se porovnávané řetězce X a Y napíší na okraje tabulky a do políček tabulky se udělá záznam na to místo, kde dochází k průsečíku stejných znaků (viz. obrázek).

Z hlediska počítače se jedná o dvojitý cyklus nad polem [m,n], kde m a n jsou délky řetězců X a Y. V kódu programovacího jazyka „typu C“:

```
for (i=0; i<m; i++) {
       for (j=0; j<n; j++) {
              if (X[i] == Y[j]) M[i,j] = true;
       }
}
```

… kde X a Y jsou řetězce, m a n jejich délka, M je matice s výsledkem a proměnné i a j jsou indexy jednotlivých pozic v matici resp. sekvenci (indexováno od 0).



Obr 1: hledání shody ve dvou řetězcích pomocí matice
zdroj: Cvrčková F. (2006) Úvod do praktické bioinformatiky, str.41, Academia, Praha, 2006

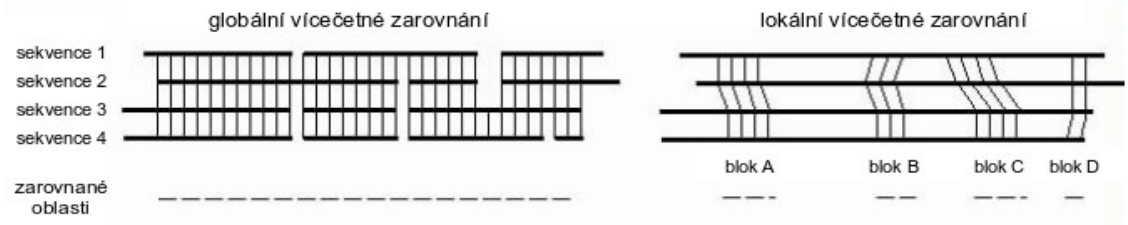
Postup značení zobrazuje první část obrázku [Obr 1], výsledek hledání je uprostřed. Diagonály značí sekvence, které jsou shodné v obou řetězcích.

Dalším krokem je vyčištění výstupu od pozic, které kvůli nedostatečné délce za shodné řetězce nepovažujeme. Pokud si například zvolíme, že délka nalezeného řetězce musí být alespoň tři znaky, dostaneme výstup v podobě pravé části obrázku.

Po stránce algoritmizace v počítači nejde o nijak náročnou úlohu – hledání začne od levého horního rohu přes všechny pozice matice a hledáme začátek řetězce. Když ho nalezneme, tak postupujeme od nalezené pozice směrem doprava a dolů a podle počtu nalezených pozic počítáme délku podřetězce. Tam kde není dostatečná délka, tam políčka vymažeme. Z hlediska programovacího jazyka jde o trojitý cyklus.

Ve chvíli, kdy máme nalezené pozice shodných řetězců, zkonstruujeme samotné zarovnání. To může být **lokální** nebo **globální** [Obr 2]. V případě globálního se snažíme nalézt nejdelší úsek napříč celou délkou řetězců za cenu toho, že budeme ignorovat neshodné úseky a budeme do řetězce vnášet mezery. V případě lokálního se snažíme nalézat jen kratší lokální úseky, ovšem s větším důrazem na přesnost. Globální přiřazení má větší smysl jen u podobně dlouhých a velmi podobných řetězců a využívá se pro mnohačetná přiřazení. V případě dobré shody se lokální zarovnání blíží globálnímu.

V případě porovnávání více řetězců se provádí zarovnání tak, že se snažíme najít společnou podmnožinu úseků, která je víceméně shodná pro všechny prohledávané sekvence. A stejně jako při porovnání dvou řetězců lze i pro více sekvencí lze vytvářet jak globální, tak lokální zarovnání.



Obr 2: Globální a lokální zarovnání
zdroj: http://bioinfo.weizmann.ac.il/~pietro/Making_and_using_protein_MA/Global.vs.Local_MA.jpg [26. 6. 2013]

Obrázek [Obr 2] celou situaci demonstruje zcela názorně. Máme několik řetězců, každý s jinou délkou a podobné sekvence jsou umístěny, resp. začínají, na různých místech. Globální zarovnání se snaží o největší shodu po celé délce sekvencí bez posunutí po sobě jdoucích bloků, zatímco lokální se zaměřuje na podobnost jen v rámci lokálních bloků, které mohou být o určitý úsek posunuty.

Tato metoda ovšem vyhovuje jen v případě, že chceme použít možnosti shoda-neshoda. Existují i situace, kdy musíme zohledňovat nejen čistou shodu (tj. písmena na obou pozicích řetězce jsou stejná), ale i možnosti, že i když se znaky neshodují, tak určitá kombinace je zajímavější než jiné. Např. některé nukleotidy nebo aminokyseliny si jsou charakterově bližší nebo je jejich kombinace vzácnější a tudíž více charakteristická.

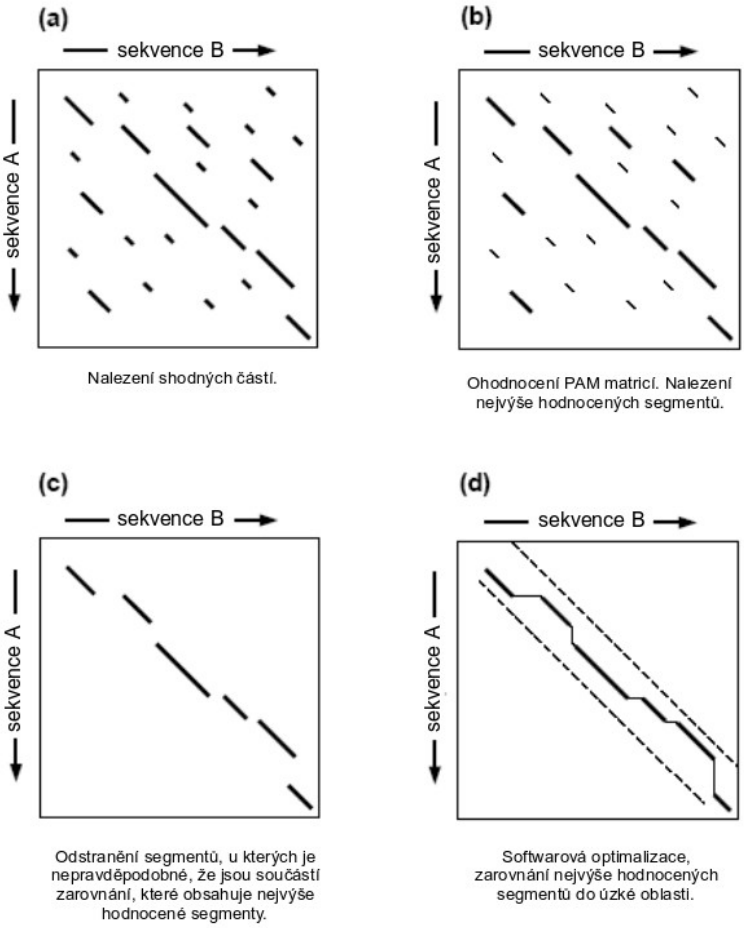
V tom případě je zajímavé je nějakým způsobem zvýhodnit, použít vyhodnocení „sice se úplně neshodují, ale jsou si hodně podobné“. Narážíme zde na problém vyhodnotit podobnost, která není popsatelná zcela exaktně. V takovém případě se opět použije matice, nicméně značení jednotlivých pozic se neprování boolovskou hodnotou ano/ne, ale použije se hodnota (označovaná jako skóre). V takovém případě se použije např. hodnota 0 pro neshodu a 10 pro plnou shodu. A různé stupně méně zajímavých kombinací se označí hodnotami od 1 do 9. Lze samozřejmě použít i jiné hodnoty.

Z hlediska programování algoritmu se pro záznam skóre použije tzv. **substituční matice** [Cvrčková 2006b], kde řádky a sloupce odpovídají jednotlivým znakům a jejich průsečík obsahuje příslušné skóre. Tyto matice jsou symetrické dle hlavní diagonály, kombinace AB má stejnou váhu jako BA.

Vyhledávání shody poté probíhá tak, že se vybírají pozice podle hodnoty skóre. Stanovení hodnot skóre samozřejmě nemůže být libovolné, ale musí vycházet ze statistiky výskytu v reálných řetězcích. Zde je informatik zpracovávající úlohu odkázán na poznatky biochemiků získaných v laboratoři. Způsobů jak stanovit skóre je více. Nejčastěji se používají matice IUPAC, PAM a jejich varianty JTT, Gonnet či BLOSUM.

Pokud chceme vytvořit zarovnání pro více sekvencí, musíme porovnat všechny sekvence metodou „každý s každým“ a v získaných mezivýsledcích najít společné části (průnik). Tento postup je ovšem velmi náročný na výpočetní výkon.

Z tohoto důvodu se používají heuristické metody. Heuristická metoda obecně je metoda, která, zjednodušeně řečeno, nejdříve (rychle) zkontroluje jen několik kritérií, které by mohly být příznakem shody. A teprve v případě, že je výsledek této kontroly nadějný (vypadá to, že by shoda mohla být), provede se důkladnější (pomalé) porovnání.

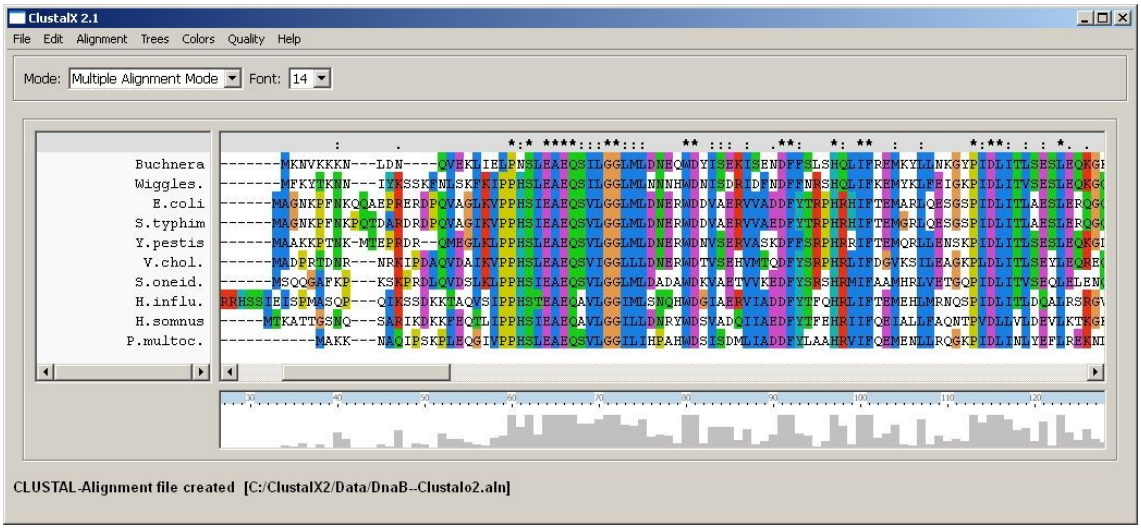


Obr 3: Ukázka algoritmu FASTA

zdroj: http://www.ctu.edu.vn/~dvxe/Bioinformatic%20course/mod4/mod4_2_files/fastafig2.gif [26. 6. 2013]

Z heuristických metod pro zarovnání sekvencí jsou nejznámější metody **FASTA** [Xiong 2006b], [Pearson & Lipman 1987], [Fasta homepage], která je demonstrována na obrázku [Obr 3] a **BLAST** [Xiong 2006c], [BLAST homapage], případně jejich modifikované varianty. Současně jde také o název nejznámějších softwarových nástrojů pro zarovnání sekvencí.

Samozřejmě kromě samotných programů FASTA a BLAST existuje řada dalších odvozených či funkčně podobných programů [Tools 4 MP], [SAS], [OS4SA], [LoSAS]. Zatímco FASTA i BLAST jsou schopné porovnávat jak proteinové, tak nukleotidové sekvence, některé nástroje jsou pouze pro nukleotidové nebo jen pro proteinové (např. CS-BLAST – „context specific BLAST“) sekvence. Mezi další poměrně známé programy pro vícečetní zarovnání patří například **Clustal** [Clustal homepage], [Thompson et al.,1997] [Obr 4], který je opět k dispozici ve více variantách (ClustalX, ClustalW, Clustal Omega).



Obr 4: Pracovní prostředí programu ClustalX
zdroj: http://akira.ruc.dk/~olesk/sekvens/Clustal_alignment.jpg

I v případě samotných nástrojů FASTA a BLAST se nejedná o jednotlivé programy, ale o celý balík nástrojů. Např. nástroj BLAST obsahuje nástroje blastn (nukleotidy), blastp (proteiny), blastpgp (PSI-BLAST), blastx, tblastx, tblastn, megablast (hromadné zpracování).

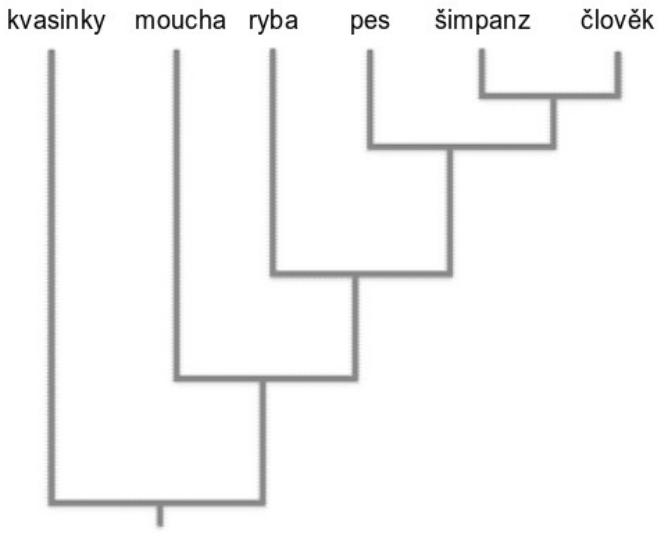
S rozvojem technologií se objevují i varianty optimalizované pro efektivnější využití moderního hardwaru, konkrétně funkcí pro paralelní zpracování. Jednak pro efektivní využití vícevláknového zpracování (vícejádrové procesory a víceprocesorové počítače) ale například i pro využití v současnosti rychle se rozvíjejících technologií jako je např. CUDA, která využívá pro výpočet procesory grafických karet (GPU) firmy nVidia [NVIDIA]. Příkladem takové varianty je např. program CUDA-BLASTP [Liu 2011].

V současné době tedy tato oblast zvládnutá po teoretické i praktické stránce a vývoj směřuje spíše k vylepšování výkonu, k výpočetním metodám a technologiím založeným na paralelizaci a zvyšování rychlosti výpočtů.

Kromě samotného výpočtu podobnosti se software výrazně uplatňuje i v další části zpracování dat a to v jejich **vizualizaci**. Je nepochybě zajímavé, že software dokáže kvantitativně popsat podobnost dvou sekvencí a vyjádřit je jako nějakou formu metriky (procenta, skóre atd.), ale tato čísla jsou pro lidské vnímání obtížněji pochopitelná. Je hezké, že software dokáže pro *n* sekvencí vytvořit tabulku n×n se vzájemnými podobnostmi, ale lidská bytost pohledem do tabulky plné čísel podobnosti nevidí.

Pro názornější vizualizaci podobností se proto v biochemii, ale i v dalších oborech, využívají dendrogramy a fylogenetické stromy.

Dendrogram [Cvrčková 2006f] [DEN1], [DEN2], [DEN3] je typ stromového grafu, který využívá shlukování prvků dle jejich vlastností, konkrétně hierarchické clusterizace [Obr 5]. Jeho výhodou je, že je na první pohled zřejmé, které prvky si vzájemně nejvíce odpovídají. Ty, které jsou nejvíce podobné jsou spojeny větvemi stromu do jednoho uzlu. Z tyto uzly se následně opět spojují do dalších podle podobnosti.



Obr 5: Dendrogram podobnosti živočišných druhů
zdroj: http://www.aarongolas.com/media/2009/08/simple-dendrogram.gif [26. 6. 2013]

Existuje několik postupů, jak dendrogram vytvořit, např. algoritmus UPGMA [Opperdoes 1997]. Konstrukce dendrogramu je v principu velice jednoduchá. Jde ji tvořit oběma směry (od kořene k listům nebo od listu ke kořenu). Na počátku máme několik prvků a každý z nich má

určité vlastnosti, některé jsou společné s ostatními prvky, některé specifické. Průniky vlastností dvou různých prvků nám dají různé podmnožiny společných vlastností, na základě kterých se popisuje podobnost. Podmínkou je, aby existovala podmnožina vlastností která je společná pro všechny prvky (jinak bychom dostali ne jeden, ale dva či více dendrogramů). Popišme si zjednodušeně konstrukci od listů ke kořenu.

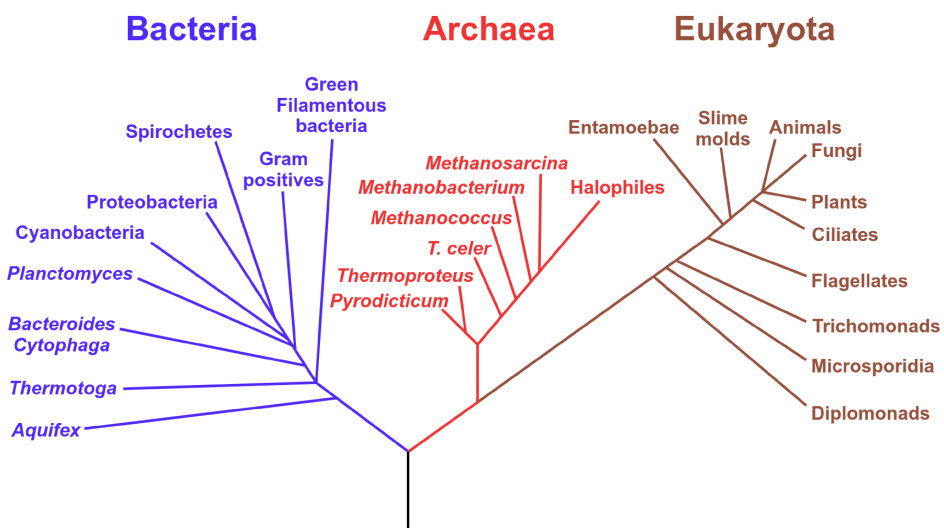
U prvků nejprve provedeme výpočet podobnosti (spočítáme shodu na základě shodných vlastností) mezi všemi prvky stylem „každý s každým“. Prvky, které si jsou nejvíce podobné sloučíme do společného uzlu. Tomuto uzlu přiřadíme vlastnosti, které jsou průnikem vlastností obou sloučených prvků a prohlásíme ho za nový prvek. Původní sloučené prvky vyjmeme a vytvoříme z nich první stupeň stromu.

Následně zopakujeme celu operaci znovu, tj. vypočteme podobnosti a opět sloučíme nejvíce podobné prvky do nového uzlu. Takto postupujeme dokud nesloučíme poslední dva prvky. Tím dostaneme kořen a výpočet končí.

Jedná se o postup, který začíná na mnoha skupinách s jediným prvkem, který má mnoho vlastností. Postupně se počet skupin snižuje, ale roste počet prvků ve skupině a současně se snižuje počet vlastností, které má skupina společné pro všechny prvky v ní obsažené.

Pro konstrukci dendrogramů lze použít řadu nástrojů, např. MEGA [MEGA homepage] nebo yEd [yEd homepage].

Druhým používaným diagramem je **fylogenetický strom** neboli „strom života“ [Baum 2008], [Phylo1], [Phylo2], [Phylo3]. Jedná se opět o stromový graf, viz. [Obr 6], který ovšem nevzniká slučováním párů jako u dendrogramu, ale jeho jednotlivé uzly značí, kde došlo k oddělení nového druhu. Od dendrogramu se liší tím, že z jednoho uzlu může vycházet více větví a původní větev může pokračovat (původní druh přežije i když z něj vznikl odvozený druh). U dendrogramu sloučením větví původní větve zanikají.



Obr 6: Fylogenetický strom
zdroj: http://upload.wikimedia.org/wikipedia/commons/7/70/Phylogenetic_tree.svg [26. 6. 2013]

Využití je jak již lze usuzovat především v oblasti genetiky a nukleotidových sekvencí. Ke konstrukci fylogenetických stromů lze použít řadu nástrojů, jmenujme pro příklad třeba program Archaeopteryx [Zmasek].

3.6 Biologické databáze

Biologické databáze [Cvrčková 2006c], [ItBD], [König 2010], [BW-ov], [LoBD] jsou ve své podstatě (elektronická) úložiště biologických dat, která pocházejí z vědeckých experimentů nebo z počítačových analýz. Nemusí jít nutně jen o biologická data ve smyslu klasických dat získaných z měřících přístrojů, ale do databáze lze uložit např. i publikace (články, knihy, časopisy).

Z technického hlediska a z pohledu informatika je **databáze** [Šeda 2002] soubor dat a k nim nezbytné softwarové nástroje pro manipulaci s daty. Softwarové nástroje někdy označujeme krkolomným označením SŘBD (systém řízení báze dat) nebo DBMS (database management system). Umožňují v datech vyhledávat, organizovat je, provádět nad nimi tzv. dotazy (např. statistické úlohy typu „kolik záznamů splňuje podmínku x“) a provádět s nimi řadu dalších činností. Co se týče druhu dat, je možné uchovávat prakticky všechny současné datové typy, od textových řetězců až po datově velké objekty, například bitmapové obrázky, zvuk, video.

Drtivá většina dnes používaných databázových systémů využívá tzv. relační databázový model (pracuje s relacemi nad elementárními tabulkami s daty), přestože existují i další modely (síťový, hierarchický, objektový...). Z konkrétních systémů jmenujme např. velmi jednoduchý

SQLite [SQLite] (který ani není databázovým systémem v pravém slova smyslu), rozšířený a populární MySQL [MySQL], kvalitně zpracovaný PosgreSQL [PgSQL] disponující velkým množstvím funkcí, kancelářský MS Access (který je vhodný opravdu jen do kanceláře), komerční MS SQL, případně produkty firem Oracle [Odb] nebo IBM [IBMdb], které ovšem míří především do prostředí velkých firem.

Uvedené databázové systémy patří k těm, které jsou univerzální a lze je obecně použít k vytvoření jakékoliv struktury dat (teprve data a jejich struktura vytvoří specializaci databáze), tj. lze je použít i k bioinformatickým účelům. Liší se jen zaměřením a optimalizací na určité rozsahy dat – některé jsou lepší na menší množství dat a jejich předností je rychlost, jiné se hodí více pro velké objemy dat a jejich přednosti jsou ve funkcích a podpoře. Kromě univerzálních databází existují samozřejmě i úzce zaměřené systémy (třeba produkty firmy SAP). Ty se ovšem na účely bioinformatiky příliš nehodí.

Technologická úroveň většiny současných databázových systémů poskytuje dostatečnou základnu pro hromadné zpracování prakticky libovolného druhu dat, nicméně samotný návrh databáze není zcela triviální úkol. Efektivnost a použitelnost samotné databáze pro konkrétní úkol stále silně závisí na návrhu struktury a reprezentace samotných dat, stejně jako na způsobu přístupu k nim.

Pokud se podíváme na databáze z pohledu bioinformatiky, budeme se zabývat **biologickými databázemi**. Jedná se především databáze pro genetiku, proteomiku, chemické struktury, hmotnostní spektrometrii, medicínské účely a další disciplíny. V biologických databázích najdeme data o proteinových řetězcích, o funkcích genů, o struktuře molekul o podobnosti biologických sekvencí a struktur a spoustu dalšího.

Pokud bychom chtěli charakterizovat současné nejpoužívanější veřejné biologické databáze, tak je můžeme rozdělit na moderované a nemoderované [Cvrčková 2006d]. **Nemoderované** jsou v podstatě skladištěm dat, kam kdokoliv může nahrát svoje data. Zda jsou tato data důvěryhodná a věrohodná už zajištěno není. U **moderovaných** je nově vložený záznam nejprve prověřen a musí projít určitým recenzním procesem či splnit jiná kritéria kvality, např. musí jít o data již zveřejněná v recenzované publikaci.

Většina současných hromadně používaných biologických databází je dnes přístupná přes Internet, formou webového rozhraní. Data jsou v nich uspořádána tak, že uživatelé je mohou snadno prohlížet online, vyhledávat v nich a stahovat si je pro offline použití v různých formátech.

Z databází nukleotidových sekvencí (např. DNA) tvoří „velkou trojku“ americká databáze **GenBank** [GenBank], evropská **EMBL** (Nucleotide Sequence database) [EMBL] a japonská **DDBJ** (DNA Data Bank of Japan) [DDBJ]. Společně jsou součástí konsorcia **International Nucleotide Sequence Database Collaboration** (INSDC) [INSDC]. Jejich obsah je víceméně identický, protože tyto tři databáze velice úzce spolupracují a sdílí (zrcadlí) navzájem svůj obsah [Cvrčková 2006e]. Při vložení sekvence do jedné z databází dojde poměrně rychle (v řádu hodin) k její distribuci i do ostatních dvou. Databáze se tedy od sebe liší v podstatě jen svým technickým řešením, po stránce datové jsou víceméně ekvivalentní.

Z databází proteinů jsou nejvýznamnějšími hráči americký **PIR** (Protein Information Resource) [PIR] a švýcarský **Swiss-Prot** [ExPASy]. Na mezinárodní úrovni bylo spojením jednotlivých databází vytvořeno konsorcium **UniProt** [UniProt], které spojuje evropský EBI (European Bioinformatics Institute), švýcarský SIB (Swiss Institute of Bioinformatics, provozovatel Swiss-Protu) a již zmíněný americký PIR (Protein Information Resource). Tím vzniká proteinový ekvivalent k nukleotidovému INSDC.

Tyto databáze samozřejmě nejsou jen skladištěm dat. Každému záznamu je přidělena jednoznačná identifikace, může být upravován, doplňován o poznámky, revidován, je evidována jeho historie změn (verzování) a řada dalších informací.

Aby bylo možné pracovat s vloženými daty, jsou vytvořena webová rozhraní (v případě např. EMBL-EBI označované jako SRS, u GenBank zase NCBI Entrez), které uživatelům umožňují využívat služeb databází on-line (přes webový prohlížeč).

K celému systému je připojena i řada nástrojů, které jsou běžně dostupné i jako samostatné programy (BLAST, CLUSTAL). Samotné databáze pak poskytují i externí API (aplikační rozhraní s nabízenými funkcemi, v tomto případě pro vzdálený přístup), které umožňuje využívat jejich služeb i vzdáleně a výsledky funkcí využívat i pro aplikace, které vytvořil někdo jiný. Např. v případě EMBL-EBI je používáno rozhraní Atlas REST APIs [RAPIs], které komunikuje prostřednictvím technologií Ajax (XML, JSON).

3.7 Strukturní bioinformatika a strukturní databáze

Dalším druhem biologických databází jsou **strukturní databáze**. Ty, jak už název napovídá, obsahují informace o struktuře molekul, v biochemii především proteinů a nukleových kyselin (RNA i DNA). Strukturou se rozumí prostorové (3D) uspořádání molekul, které má vliv na vlastnosti dané molekuly. Máme tím na mysli sekundární, terciální či kvarterní strukturu.

Strukturální informatika také usiluje o využití výpočetní techniky ke stanovení, predikci či vyhodnocení co nejpřesnějšího uspořádání struktur, které byly experimentálně zkoumány. Zde se se jedná například o vyhodnocení dat z rentgenové krystalografie či magnetické rezonance, které jakožto detekční metody primárně produkují pouze dílčí data (difrakční obrazce z rentgenu, echo z magnetické rezonance), která je třeba teprve vyhodnotit pomocí matematických metod. Problém se většinou převede na řešení soustav rovnic, které reprezentují vlnové funkce.

Z hlediska počítačového zpracování je 3D struktura molekuly mnohem komplikovanější než lineární sekvence. Jednak je třeba si uvědomit, že molekula je složená z jednotlivých atomů, které nemají při interakci s okolím pevnou pozici. Mezi jednotlivými atomy jsou různé vazby. A samozřejmě shluky atomů tvoří molekuly aminokyselin či nukleotidů, které se chovají jako ucelené stavební jednotky. Ze skupin molekul potom vznikají v rámci terciální a kvarterní struktury celé segmenty (shluky, klubka, listy, helixy), které se navazují na další segmenty.

Když si tyto skutečnosti uvědomíme, tak samotná reprezentace molekuly v prostoru pomocí počítače vyžaduje poměrně promyšlený přístup a využití pokročilejší matematiky. Zamysleme se nad tím, co všechno se musí uložit do souboru s informacemi o struktuře molekuly.

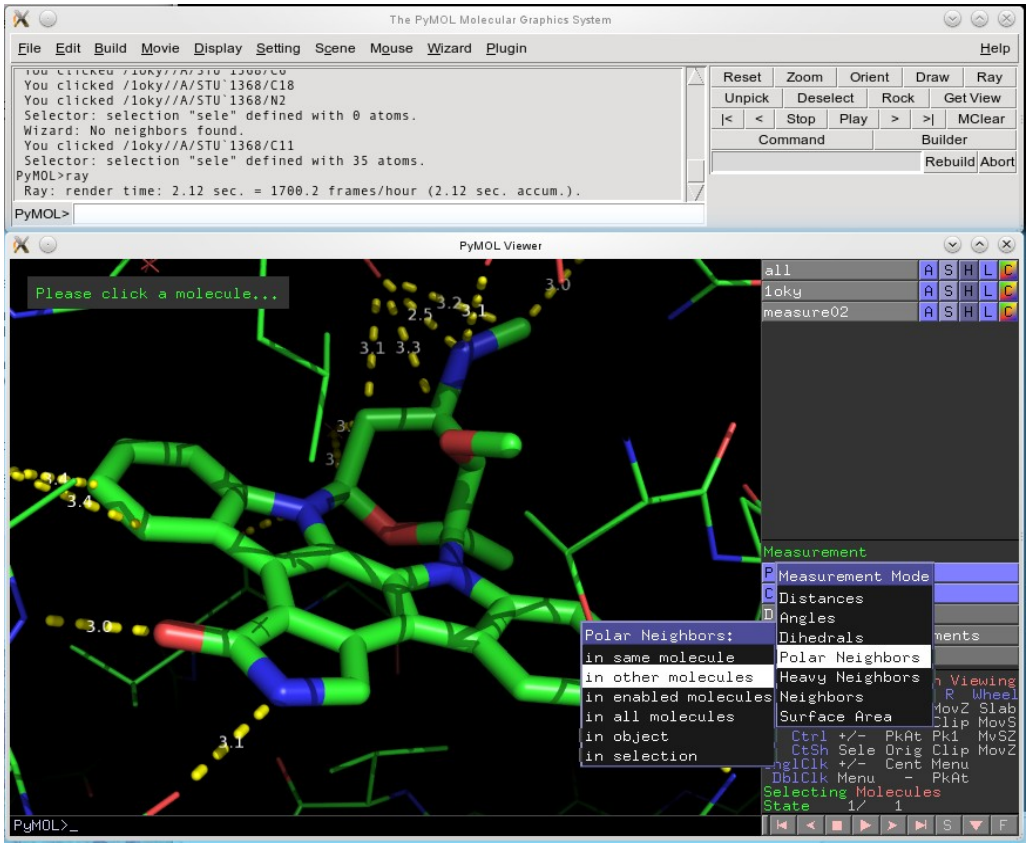
Ani základní stavební kameny (aminokyseliny či nukleotidy) nemají jednoznačný tvar. Je tady nutné evidovat každý atom samostatně a to tak, že je definován chemickým prvkem (např. kyslík či uhlík) a trojrozměrnou polohu v prostoru. Dále je třeba evidovat chemické vazby mezi atomy a to tak, že každou vazbu definují identifikátory dvou atomů a druh vazby. Dále je samozřejmě žádoucí mít možnost seskupovat atomy do molekul a molekuly do segmentů (např. helixů a listů), které tvoří části terciální a kvarterní struktury.

Pro ukládání prostorové struktury se používají především soubory ve formátu **PDB** [PDBdoc], jehož struktura veřejně zdokumentovaná a standardizována. Podobně jako např. formát FASTA, je i PDB souborem textovým, nicméně je mnohem striktněji strukturován. Umožňuje uložení nejen samotné 3D struktury, která je podstatně více složitá než lineární sekvence, ale i dalších pomocných informací, komentářů atd. A protože se potřeby strukturní bioinformatiky stále vyvíjejí, ani tento formát nezůstává stabilní, ale existuje v několika verzích, které postupně reflektovaly nové požadavky. Nejnovější verze tohoto formátu má označení 3.30 a pochází z listopadu 2012.

Kromě nejrozšířenějšího formátu PDB však existují i další strukturní datové formáty. Můžeme uvést třeba **CIF** [CIF] (Crystallographic Information File), skupina formátů označovaných jako **CTFiles** (Chemical Table files) [CTF] a další.

Podobně jako existují veřejné databáze pro sekvenční data, existují i strukturní databáze fungující na podobném principu. Mezi nejvýznamnější patří **PDB** [PDB] (Protein Data Bank) po které je pojmenován i datový formát PDB, dále například databáze **NDB** [NDB] (Nucleic Acid Database), **CSD** [CSD] (Cambridge Structural Database) nebo již zmiňovaný **UniProt** [UniProt]. A opět bychom zde nalezli podobné schéma jako u sekvenčních databází – webové rozhraní pro uživatele, centrální správa databáze, API pro poskytování služeb.

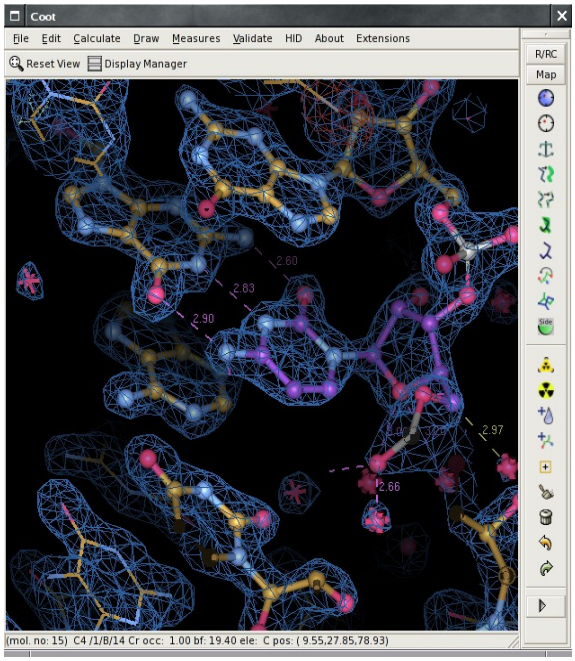
V oblasti samotného softwaru pro zpracování strukturních dat nalezneme také řadu vyspělých programů. Mezi nejvýznamnější patří velmi široce používaný **PyMOL** [PyMOL] [Obr 7], který je populární i díky ještě nedávno přívětivé licenční politice.



Obr 7: Pracovní prostředí programu PyMOL

zdroj: http://www.pymol.org/sites/default/files/pymol_snap.png [26. 6. 2013]

V současnosti je perspektivním softwarovým nástrojem například **COOT** (Crystallographic Object-Oriented Toolkit) [COOT] [Obr 8], který je do budoucna nadějný i díky svojí licenci, která neumožňuje komercializaci a poskytuje prostor pro otevřený vývoj.



Obr 8: Pracovní prostředí programu COOT

zdroj: http://upload.wikimedia.org/wikipedia/en/7/72/Coot-crystallography-software.png [26. 6. 2013]

Z dalších softwarových nástrojů stojí za pozornost **Jmol** [Jmol]. Je vytvořen v jazyce Java, díky tomu je multiplatformní a na rozdíl od většiny podobných nástrojů nevyžaduje ze strany hostujícího systému grafickou akceleraci. Je tedy široce využíván webovými portály, kde slouží k zobrazování struktury molekul – ve webových prohlížečích uživatelů potom Jmol běží jako plugin, který graficky a interaktivně zobrazuje strukturu molekuly.

Mohlo by se tedy zdát, že strukturní bioinformatika má většinu problémů vyřešených. Zdaleka to tak není. V současné době řada problémů stále čeká na vyřešení. Například za současného stavu technologií stále není možné obecně stanovit prostorovou strukturu molekuly pouze na základě znalosti jejich strukturního vzorce, což je schopnost o kterou by nepochybně byl velký zájem. Současné nástroje umožňují jen určitou míru aproximace, která je založená na znalostech určitých zákonitostí, např. faktu, že molekuly se snaží zaujímat energeticky nejvhodnější polohu tak, aby se síly jednotlivých atomů co nejméně ovlivňovaly.

Matematické modely pro odhad prostorového uspořádání v principu pracují tak, že se na počátku stanoví nějaký počáteční stav, který je jen přibližný. Následně se provádí (někdy i metodou pokus-omyl) menší úpravy, které postupně stav zlepšují a snižují chyby. Každý krok

úprav (iterace) by měl snížit chybu. Výpočet končí buď snížením chyby na přijatelnou úroveň, nebo zastavením výpočtu buď po určitém počtu iterací, nebo v okamžiku kdy další iterace již nesnižují chybu a výpočet se tzv. „zacyklí“. Jedná se tedy opět o určitý heuristický model. Samozřejmě už vyžaduje poměrně rozsáhlý matematický aparát, znalosti z fyzikální chemie a také pokročilé algoritmy či samoučící se systémy (např. neuronové sítě) z oblasti informatiky. Navíc získaný výsledek většinou vyžaduje i dodatečnou manuální korekci (například vyloučení nemožných variant) ze strany uživatele.

Na poli strukturní bioinformatiky je tedy stále mnoho neprobádaného.

3.8 Bioinformatika a separační metody

Bioinformatika v současné době hraje významnou roli také při zpracování výsledků získaných různými separačními metodami. **Separační metody** [SepM] slouží k identifikaci jednotlivých složek směsí na základě chemických nebo fyzikálních vlastností jednotlivých složek. V biochemii se nejčastěji používají různé druhy chromatografie, separace pomocí membrán (osmóza, dialýza), elektromagnetických polí (elektroforéza, hmotnostní spektrometrie), gravitace (sedimentace, centrifugace) a dalších.

Informatika se dostává i do těchto oblastí a to jak primárně v podobě řídícího softwaru pro jednotlivá zařízení, tak softwaru pro následnou analýzu a interpretaci získaných dat. Například v případě gelové elektroforézy se uplatňují specializované skennery pro získání digitálního obrazu gelu, který se následně analyzuje dalším softwarem.

Použití softwaru k analýze skenovaného gelu přitom dokazuje, jak blízko k sobě mají exaktní vědy. Pro potřeby biochemie jsou prostřednictvím výpočetní techniky zpracovávána data pomocí metod z počítačové grafiky, jako je například hranová detekce [Lindeberg 2001] či prahování [Tresh], což jsou postupy, které vznikly pro účely zpracování obrazu a jsou založeny na matematických, resp. statistických principech.

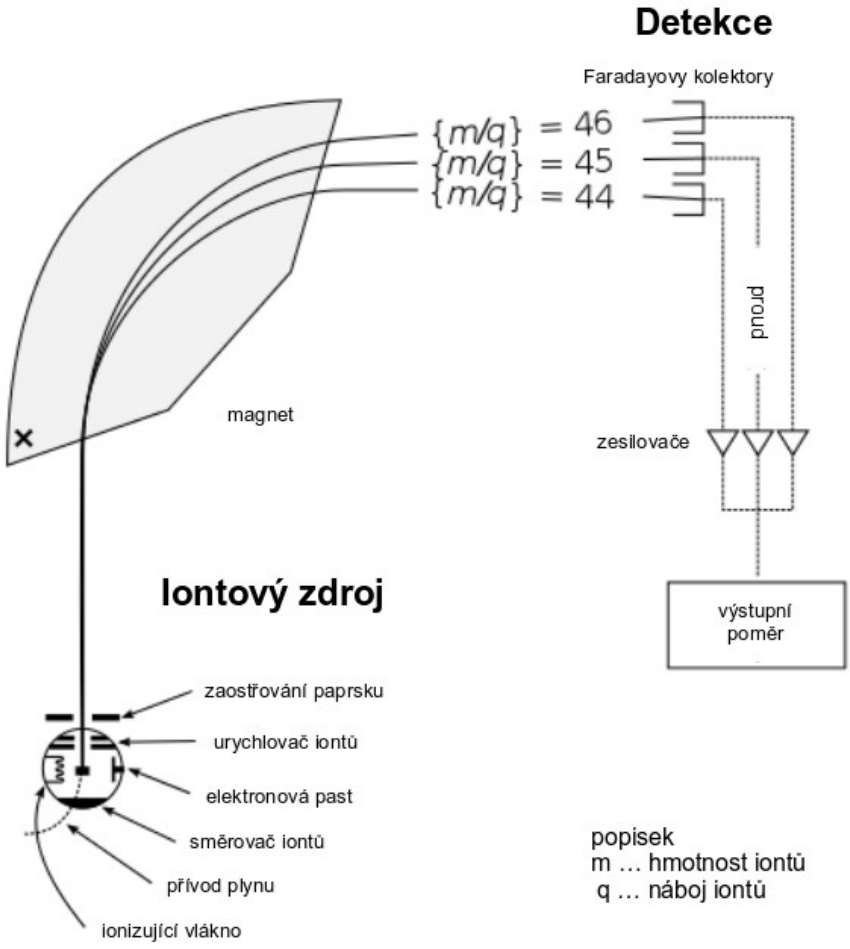
Další oblastí masivního využití bioinformatiky k analýze dat je **hmotnostní spektrometrie** [Poustka 2007], [Reusch 2013], [Polášek 2005], [Holčapek 2013]. Jedná se o separační metodu založenou na rozdělení částic na základě poměru m/z (m – hmotnost, z – počet nábojů vzniklého iontu).

Měření se provádí pomocí hmotnostního spektrometru[3]. Spektrometr [Obr 9] se skládá ze tří částí – iontového zdroje, hmotnostního analyzátoru a detektoru.

3 Nezaměňovat s *optickým spektrometrem*, který pracuje na principu rozkladu světla procházejícího látkou (spektroskopie).

Úkolem iontového zdroje je převést zkoumaný vzorek do ionizovaného stavu. Princip spočívá v odpaření vzorku a jeho ionizaci. To lze provést elektronovou ionizací, chemickou ionizací, elektrosprejem, MALDI („matrix assisted laser desorption ionization“), indukčně vázanou plazmou a dalšími způsoby…

Následně dojde k separaci iontů dle m/z poměru. To lze provést několika způsoby – za pomoci magnetického či elektromagnetického pole (magnetický, resp. elektrostatický analyzátor zakřivující dráhu letu iontů, viz. [Obr 9]), pomocí elektrického napětí (kvadrupól, iontová past), iontovou cyklotronovou rezonancí či pomocí doby letu (TOF – „time of flight“).

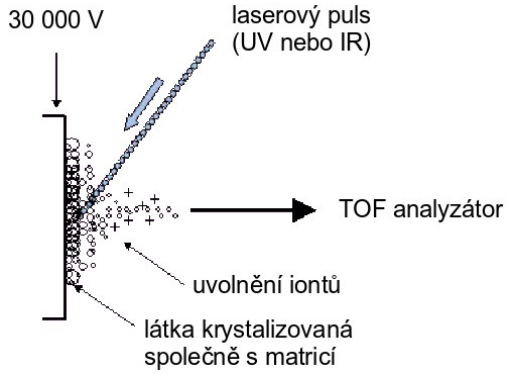


Obr 9: Schéma hmotnostního spektrometru

zdroj: http://upload.wikimedia.org/wikipedia/commons/0/0d/Mass_Spectrometer_Schematic.svg [26. 6. 2013]

Posledním krokem je detekce iontů. To se provádí detektorem, na jehož povrchu dochází vlivem dopadu iontů k indukci proudu. Detektorů je opět více druhů, uveďme např. elektronový násobič nebo Faradayova klec [Fklec].

Jedním z nejvýznamnějších druhů hmotnostní spektrometrie je použití laserové ionizace vzorku umístěného na matrici, tzv. MALDI („matrix assisted laser desorption ionization“) [Obr 10], ve spojení se separací TOF („time of flight“). V tomto případě se mluví o metodě MALDI-TOF [MTOF], [Havliš 1999] případně dvourozměrné MALDI-TOF/TOF.



Obr 10: MALDI-TOF - ionizace pomocí laseru
zdroj: http://qbab.aber.ac.uk/roy/mss/maldi.gif

Výstupem ze spektrometru je **hmotnostní spektrum**, což je soubor dat, který se zobrazuje jako graf, resp. spojitá křivka ve spojnicovém grafu, kde je na horizontální ose (osa x) je uveden poměr hmotnosti a náboje [*m/z*] a na vertikální ose (ose y) je vynesena intenzita.

Z hlediska počítačového zpracování se jedná o soubor diskrétních bodů ve tvaru [*m/z*, intenzita]. Většinou se data ukládají v podobě strukturovaného textového souboru (ASCII), ale lze použít i jiné formy, např. CSV soubor (což je opět varianta textu).

V tuto chvíli nastupuje bioinformatika k analýze získaných spekter. Každé hmotnostní spektrum obsahuje desítky až stovky tisíc bodů a velikost takového souboru je v řádech megabyte. Ruční zpracování takového množství dat není v praxi možné. Pro potřeby běžného uživatele je třeba data převést do podoby, která je lépe srozumitelná lidskému vnímání.

V praxi se tedy použije počítačová vizualizace (graf) se zvýrazněním extremálních hodnot intenzity. Tyto extremální hodnoty se označují jako **píky**. Projevují se lokálním skokovým zvýšením intenzity signálu a znamenají přítomnost většího množství iontů v dané hodnotě [*m/z*].

Úkolem softwaru pro zpracování těchto dat je detekovat píky a v případě potřeby nalézat podobnosti v různých spektrech. To s ohledem na množství dat v každém spektru klade požadavek na jejich co nejvíce automatizované zpracování. A to jednak ve formě nalezení podstatných informací ve hmotnostních spektrech (píky), ale také při jejich vzájemném porovnání. A v neposlední řadě i při predikci, kdy je cílem předpovědět na základě teoretických znalostí výsledek a poté ho ověřit proti reálně získaným výsledkům.

O významu hmotnostní spektrometrie hovoří i velký počet programů [Xu & Ma 2006], [MSS], který je k jejich zpracování dostupný. Z nejznámějších můžeme uvést jako příklad programy **Mascot** [MASCOT], **Phenix** [Phenix], **mMass** [mMass], **MALDI Biotyper** [Biotyper], **Saramis** [Sar].

Hmotnostní spektrometrií se zabývají i aplikace, které jsou součástí této disertační práce.

3.9 Další úkoly a nástroje bioinformatiky

Uvedené úlohu samozřejmě nejsou jedinou náplní bioinformatiky. V řadě zmíněných oblastí by bylo možné jím mnohem více do hloubky. Rovněž je řada oblastí, ke kterým jsem se v této teoretické části ani nepřiblížil.

Dalo by se např. rozsáhle psát o počítačové simulaci genové exprese a syntézy proteinů. Zde je velice zajímavé, že genetika a informatika (minimálně její části) mají společného mnohem více než se na první pohled zdá. Mnohé postupy probíhající v genetice, např. při přepisu a kopírování DNA, jsou principiálně velmi podobné postupům, které používá informatika při práci s daty.

V informatice jsou např. pro ochranu konzistence dat používány tzv. paritní kódy a kontrolní součty, které mají mnoho společného s ochrannými mechanizmy, které používá genetika k opravám chyb v DNA. Dvojkový kód používaný počítačem a překlad zdrojového kódu programovacího jazyka na binární program složený ze strojových instrukcí je v řadě rysů podobný čtyřkovému paritními kódu DNA a jejímu přepisu na proteiny, na nichž stojí živé organizmy.

Nemalou část by zabral detailnější popis nástrojů výpočetní techniky, které bioinformatika využívá. Zmiňme jen to nejpodstatnější. V předchozí části byl zmíněn **textový editor** pro úpravu textových souborů[4]. Pro hromadné výpočty a vizualizaci dat pak lze použít **tabulkový kalkulátor** (MS Excel, LO Calc, Gnumeric). Schopnost použít tyto nástroje patří k základům počítačové gramotnosti.

Na další nástroje již musí být uživatel pokročilejší. Uživatelské využívání **databázových systémů** (zmíněných výše), pokročilých statistických a matematických softwarů typu **CAS** (computer algebra system, např. hojně rozšířené nástroje Matlab [MATLAB], Maple [Maple], nebo jejich svobodná alternativa GNU Octave [Octave]) nebo úzce specializovaných programů

4 Příkladem textového editoru je třeba Notepad (Poznámkový blok). Nezaměňovat s textovým procesorem, např. MS Wordem.

(viz. uváděné výše) již vyžaduje zaškoleného uživatele, který je dobře obeznámen s používanými nástroji a často vyžaduje zvládnutí i základů nějakého programovacího jazyka.

Pro informatika pracujícího v oblasti bioinformatiky je samozřejmě nutná i znalost dalších nástrojů. Především se jedná o znalosti **algoritmizace** (obecná úroveň řešení problému, teoretický základ pro programování) a **programování** (zvládnutí konkrétních programovacích jazyků) a samozřejmě také schopnost administrovat počítačové systémy.

Pokud se zastavíme u samotné algoritmizace, bioinformatika samozřejmě vyžaduje znalosti běžných postupů algoritmizačních/programovacích postupů. Nicméně v rámci oblasti působení bioinformatiky jsou některé schopnosti zdůrazněny. V první řadě je to práce s textovými řetězci a čtením souborů. Dále metody spojené se statistikou a vyhodnocováním numerických dat (interpolace, výpočty funkčních závislostí). Nelze zapomenout ani na grafické prostředky (vizualizace dat) nebo v poslední době se mohutně rozšiřující nástroje umělé inteligence, samoučící se systémy (neuronové sítě) apod.

Detailní popis mnoha algoritmů a postupů využívaných či přímo specifických pro oblast bioinformatiky lze ve velmi dobře zpracované podobě nalézt např. ve výborné publikaci *A. Polanski, M. Kimmel: Bioinformatics, Springer 2007* nebo *Xiong J., Essential Bioinformatics, Cambridge University Press 2006.*

Pří detailním pohledu na tuto problematiku se tedy ukazuje, že práce bioinformatika má s biochemií v laboratoři jen velmi málo společného a vyžaduje především specializaci z kategorie analytik-matematik-vývojář-programátor, kde jsou pracovním nástrojem počítače, matematické metody a programovací jazyky.

Z programovacích jazyků jsou pro potřeby bioinformatiky nejvýznamnější **jazyky interpretované** [intJ]. Vyznačují se multiplatformností (použitelné kdekoliv) a flexibilitou i když jsou pomalejší než jazyky kompilované. Lze je velmi snadno použít k rychlé tvorbě jednoúčelových skriptů (především na transformaci dat v textových souborech).

Z interpretovaných jazyků je velmi populární (a to nejen v bioinformatice) především jazyk Python [Python], který je oblíbený pro snadné použití, rychlost a kompatibilitu s mnoha externími nástroji. Jako skriptovací jazyk se často užívá např. i Perl [Perl] (lze použít i jeho zjednodušený webový dialekt PHP [PHP]), nicméně ten je se svojí složitější syntaxí poněkud náročnější. Použitelné jsou samozřejmě i další interpretované jazyky, např. Ruby [Ruby]. Pro tyto účely je do jisté míry použitelný i bash/shell [shell] (interpretovaný jazyk unixového příkazového řádku).

Zde je zajímavé, že řada z uvedených jazyků má k dispozici rozšíření (ať ve formě knihoven, pluginů nebo nástrojů) pro vědecké účely. Pro potřeby biologie uveďme sady nástrojů jako je například Biopython [BioPython], BioPerl [BioPerl] nebo BioRuby [BioRuby].

Pokud dojde na potřebu vytvořit rozsáhlejší aplikace či je požadavek na rychlost, je vhodné využít spíše **kompilované jazyky** (především velice rychlý C a výrazně rozšířený C++ [C&CPP]). Ty je nutné vždy přeložit (zkompilovat) do binární podoby pro konkrétní operační systém a nejsou tedy tak flexibilní. Nicméně mají svůj význam kvůli vyšší rychlosti. Řada větších aplikací je často tvořena v Javě [Java], což je jazyk hybridní (je předkompilovaný do rychlejšího bytekódu, který je ale stále interpretovaný) a velmi univerzální s rozsáhlým množstvím funkcí.

Specifickou kapitolou jsou potom programovací jazyky pro **webové aplikace** [WApp1], [WApp2] (např. webové portály proteinových databází), kde je nutné využívat kombinací více technologií (databáze, serverová část rozhraní, klientská část rozhraní a jejich propojení).

V řadě případů jsou vyžadovány i administrátorské schopnosti. Většina softwaru pracujícího s velkým množstvím dat využívá pro jejich skladování **databázový systém** (viz. výše). Je ovšem velký rozdíl mezi používáním databází z uživatelského hlediska (rozhraní) a jejich konfigurací, údržbou a provozní správou ze strany administrátora či vývojáře. Pokud tedy nastane nutnost použít pro uložení dat databázi, měl by bioinformatik-vývojář být schopen nainstalovat a nakonfigurovat databázový systém a vytvořit v systému samotnou databázi (datovou strukturu reprezentující data, která budou ukládána).

S administrátorskými úkoly souvisí i používání **jiných operačních systémů** než je řada Windows, především systémů na bázi Unixu. I když se většině lidí vybaví pod pojmem „počítač“ pouze klasické PC s OS Windows, je monopol tohoto systému na osobních počítačích anomálie. V jiných segmentech (např. mobilní zařízení, síťová zařízení, servery, superpočítače) nalezneme zcela jiné typy operačních systémů a Windows je zde jen okrajový hráč. Například v červnu 2013 z 500 nejvýkonnějších počítačů světa [TOP500] 476 používalo Linux, 16 Unix, 4 smíšené řešení, jeden BSD a pouze tři používaly OS Windows.

Pro většinu náročných úkolů (systémy provozované v režimu 24/7[5] – servery, webové portály, rozsáhlé databáze, superpočítače, clustery[6]) se masivně používají operační systémy typu **Unix**,

5 Zkratka pro nepřetržitý provoz: 24 hodin denně, 7 dní v týdnu.

6 Počítačový cluster je systém složený s mnoha menších počítačů, které společně vykonávají určitou činnost. Navenek se tváří jako jeden počítač. Cluster je řízen centrálním uzlem, který rozděluje řešenou úlohu na menší části, které předává na vyřešení jednotlivým počítačům, které jsou spolu navzájem propojené (většinou počítačovou sítí). Silnou stránkou clusteru je schopnost efektivně řešit problémy, které lze paralelizovat, tj.rozdělit na menší části, které lze řešit samostatně.

jejichž výhodou je především efektivnější fungování při velké zátěži, především při zpracování mnoha úloh současně, stabilita a v neposlední řadě i bezpečnost. Takové vlastnosti dělají z těchto systémů zajímavé nástroje i pro vědu. Pro Unix existuje řada vědeckých programů [42FLSS],[UbSci], které na systémech Windows nejsou podporovány, nebo nelze plně využít jejich potenciálu (např. vícevláknové zpracování úloh).

Mezi unixové systémy patří nejen klasický Unix, ale i Solaris, BSD, populární Mac OS X, mobilní iOS a především početné systémy řady **GNU Linux** (včetně Androidu pro mobilní zařízení). Z unixových systémů dnes právě verze Linuxu hrají největší roli. Vyznačují se otevřeným zdrojovým kódem, modularitou, dynamickým vývojem úzce spojeným s komunitou uživatelů a především licenční politikou, která je velice přívětivá pro provádění různých úprav a modifikací.

Právě licenční podmínky a snadná upravitelnost dělají ze systémů na bázi Linuxu velice atraktivní nástroj pro akademické prostředí, protože systém lze snadno přizpůsobit pro různé specifické účely. Unixovým systémům se v akademickém prostředí obecně daří, byť to české je výjimkou. Možná i proto, že zde mají svůj původ – vznik větve BSD [BSD] je úzce spojen s univerzitou v Berkley a i Linux vznikl jako studentský projekt. Dokonce existují speciální verze právě pro vědecké účely, jak uvedu dále.

Z opravdu významných distribucí Linuxu je vhodné zmínit dva velké hráče v této oblasti.

Především ve Spojených Státech je rozšířena komerční distribuce **Red Hat Enterprise Linux** (často zkracováno na RHEL), kterou vyvíjí společnost RedHat [RH]. Vyznačuje se tím, že i když je samotný Linux zdarma, poskytuje RedHat placené služby – plně profesionální a kvalitní technickou podporu (aktualizace, opravy chyb, helpline). Placená podpora je vlastnost, která dovoluje seriozní nasazení do profesionálních prostředí – když je problém, existuje oficiální helpline, kam se lze obrátit. Právě RHEL a odvozené varianty (např. CentOS a Fedora) jsou často používány v akademickém prostředí a je z nich odvozena i distribuce přímo určená pro potřeby vědy, která je označovaná jako **ScientificLinux** [SciLi]. Ta je vyvíjená mj. i známým výzkumným střediskem CERN u Ženevy a podobně zaměřeným střediskem Fermilab u Chicaga. Ve skutečnosti si každá z organizací vytváří ještě svoji modifikaci – Fermi Linux [FeLi] a CERN Linux [CERNLi].

Druhým velkým hráčem je distribuce **Debian** [Debian], která je na rozdíl od RedHatu komunitní a plně svobodná. Její výhodou je bezplatnost, nicméně vyžaduje schopnějšího administrátora, protože v případě problémů si musí poradit bez komerční podpory, jen s

podporou komunity. Význam Debianu spočívá v tom, že poskytuje stabilní platformu pro řadu systémů (servery, superpočítače) a je z něj odvozena celá řada dalších distribucí, například celá rodina populárních systémů Ubuntu. Jednou z variant Debianu/Ubuntu je distribuce **NEBC Bio-Linux** [BioLi], která obsahuje softwarový balík **BioLinux**, který je přímo určený pro potřeby bioinformatiky.

Z uvedeného je tedy zřejmé, že mezi dovednostmi bioinformatika by měla patřit i schopnost používat a administrovat i jiné operační systémy než je Windows. Ať už z důvodu dostupnosti nativního unixového softwaru, tak pro schopnosti, které Unixové/Linuxové systémy nabízejí navíc proti běžnému desktopovému mainstreamu.

Současným trendem v oblasti výpočetní techniky a technologií ke zpracování dat obecně je směřování k postupům pro hromadné zpracování dat, k práci s velkými objemy dat a k paralelním výpočtům, které umožňují rozdělit úlohy do několika menších částí, které lze zpracovat odděleně. Z hlediska běžných uživatelů přitom dochází k trendu směrovanému na malá mobilní zařízení, což je v konfliktu s potřebou narůstajícího výzkum velké výpočetní kapacity. Je tedy otázkou co přinese dalších pár let v této oblasti.

3.10 Shrnutí

Dnes je bioinformatika pro řadu biologických odvětví již prakticky nepostradatelnou pomocnou vědou. Výzkum proteinů, genetiky, potravinářství, biochemie obecně, medicíny a dalších biologických oborů by bez využití výpočetní techniky nebyl možný. Bioinformatika tento výzkum zjednodušuje a urychluje, protože nové přístroje a metody zpracován dat umožňují získávat výsledky stále rychleji.

Co lze očekávat do budoucna?

Trend je celkem jasný a týká se většiny oblastí IT, nejen bioinformatiky. Bude pokračovat další růst objemu dat, porostou požadavky na jejich efektivní zpracování. S tím jak porostou potřeby zpracovávat velké množství dat, porostou i požadavky na nástroje, které (bio)informatika využívá. To paradoxně přinese další tlak na zlepšování postupů, jak data účinně třídit a zpracovávat na použitelné výsledky (např. samoučící se systémy, neuronové sítě, biologické algoritmy), tak aby se z informací stalo vědění.

Technologie samotné se budou ubírat mnohem více k paralelnímu zpracování, k většímu využití výpočetních clusterů i k v současnosti tolik módnímu cloudu[7]. Tyto změny v přístupu by měly částečně kompenzovat současný stav, kdy další růst výkonu výpočetní techniky začíná stále více narážet na technologicko-fyzikální limity miniaturizace, pracovních teplot a další omezení. Otázkou například je, kolik změn do oblasti relativně konzervativního vědeckého prostředí zanese např. rozvoj mobilních zařízení[8]. To vše bude v budoucnu vyžadovat změnu řady přístupů jak na poli hardwaru, tak softwaru.

V každém případě si bioinformatika svůj význam nejen udrží, ale zřejmě i posílí.

7 Jde o řešení kdy služby a data jsou na vzdáleném serveru a uživatel k nim přistupuje jen pomocí rozhraní. Příkladem je například naprostá většina emailových služeb (např. Gmail) – data jsou uložena na poštovním serveru, uživatel má jen nenáročné zařízení (webový prohlížeč v počítači, tablet, smartphone). Podobně pracuje řada současných služeb – YouTube, Flickr, Facebook, Dropbox. Ačkoliv je cloud snadné řešení pro pohodlné uživatele, jedná se o obrovský problém z hlediska soukromí, protože data jsou mimo kontrolu majitele.

8 Během stáže na TU Vídeň jsem zaznamenal, že je zde relativně rozšířené využívání např. tabletů.

4 ProteinCutter

*Webová aplikace **ProteinCutter** je určena pro predikci sekvencí peptidů, které mohou teoreticky vzniknout při štěpení proteinu. Uživateli umožňuje zadat vstupní data (protein), na kterých chce provést predikci štěpení a stanovit pravidla, podle kterých bude teoretické štěpení určováno.*

Výstupem jsou potom data v podobě jednotlivých peptidových řetězců, které dle zadaných pravidel mohou vniknout. Získaná i vkládaná data jsou v jednotlivých krocích doplňována dalšími informacemi o hodnotách chemickofyzikálních veličin získaných řetězců.

4.1 Sekvencování proteinů a bioinformatika

Sekvencování (též sekvenování) je v biochemii označení pro proces, při němž se určuje primární struktura (pořadí chemických jednotek) daného biopolymeru. Výsledkem je lineární kód označovaný jako sekvence, která shrnuje uspořádání sekvencované molekuly.

Významnou roli hraje sekvencování pomocí hmotnostní spektrometrie [Řehulka 2007]. Při analýze vzorku, u kterého chceme zjistit sekvenci, se postupuje tak, že původní řetězec štěpíme (např. pomocí enzymů) na kratší řetězce, které buď identifikujeme dle shody s již známými řetězci, nebo dále štěpíme. Identifikace probíhá pomocí hmotnostního spektrometru, kde ve vzorku s rozštěpenými fragmenty původního řetězce vyhledáváme píky, které odpovídají hmotnosti jednotlivých fragmentů. Tyto hmotnosti poté porovnáváme s databází už identifikovaných sekvencí aminokyselin.

Problémem je, že množství kombinací aminokyselinových sekvencí je natolik velké (n^{20}, kde n je délka řetězce), že jen malá část z nich je identifikovaná na základě výskytu v reálných proteinech. Proto je třeba některé sekvence štěpit opakovaně až na velice krátké segmenty, které buď již identifikované jsou, nebo je možné je teoreticky vypočítat.

Do výzkumu proteinů se proto zapojuje i bioinformatika. Jednou z cest je evidence známých struktur v databázích. Ať už jde o nukleotidové GenBank, ENA/EMBL-Bank, DDBJ, proteinové PIR, SWISS-PROT, TrEMBL, PSD-Kyoto, PRF [PRF], NRL-3D [NRL-3D] nebo kompozitní NCBInr, MSDB a OWL [OWL].

Druhou cestou k zapojení bioinformatiky je samotná predikce konkrétních sekvencí pomocí počítače. Rychlost počítačů umožňuje v přijatelném čase vypočítat teoreticky možné kombinace aminokyselinových sekvencí, např. peptidů vzniklých štěpením proteinů a určit jejich

hmotnosti. Takto získané teoretické výsledky (hmotnost-sekvence) při porovnání s reálnými výsledky ze spektrometru (hmotnost) umožňují přiřadit změřeným píkům odpovídající sekvence aminokyselin.

Takové aplikace jsou k dispozici – PeptideMass [PMass], PeptideCutter [PCutter], MS-Digest [MSDig] nebo Protein Calculator [PCalc].

4.2 Trocha algoritmizace

Výpočet sekvencí lze provádět dvěma směry. Buď vypočítáme teoretické kombinace metodou hrubé síly[9], nebo můžeme vycházet z původního štěpeného řetězce a vypočítat všechny možnosti, které z něj mohou vzniknout.

V prvním případě necháme počítač, aby nám vypočetl všechny možné sekvence aminokyselin, které mají odpovídající hmotnost. To má svoje úskalí v tom, že narážíme na složitost výpočtu. Počet kombinací je úměrný délce řetězců. Při dvaceti aminokyselinách bude mít sekvence délky *n* celkem 20^n možností. Sekvence o dvou znacích má 400 kombinací. Sekvence o třech znacích již 8 000. Pokud bychom měli počítač, který dokáže vyhodnotit 10 000 fragmentů za sekundu, jak dlouhý řetězec bychom mohli spočítat za určitý čas? To nám ukazuje Tabulka 3.

délka řetězce	**počet kombinací**	**délka výpočtu při rychlosti 10 000 řetězců za sekundu**	
1	20	0,002 s	
2	400	0,04 s	
3	8 000	0,8 s	
4	160 000	16 s	
5	3 200 000	320 s	5 minut a 20 s
10	$1{,}024 \cdot 10^{13}$	1 024 000 000 s	32 roků a 164 dnů
50	$1{,}125899907 \cdot 10^{65}$	$1{,}125899907 \cdot 10^{60}$ s	asi až do konce vesmíru ?

Tabulka 2: Závislost mezi délkou sekvence a rychlostí výpočtu všech kombinací

Z tabulky je tedy zřejmé, že metodou hrubé síly můžeme vypočítat jen velice krátké řetězce. Řetězec délky 5 by počítač počítal 5 minut, na délku 6 znaků by to už byly téměř dvě hodiny. Sedm znaků by byly téměř dva dny, osm znaků šest týdnů, devět znaků víc než rok a půl.

Uvedená hodnota 10 000 řetězců je modelový příklad. Reálně možná hodnota závisí na použitém hardwaru, programovacím jazyce a konstrukci kódu a může se lišit i v rozsahu řádů.

9 Metoda hrubé síly je řešení, kdy se mechanicky (většinou pomocí počítače) vyzkouší všechny možné kombinace. Jedná se o poměrně neefektivní metodu, která prostě jen zkouší všechny možnosti a předpokládá se, že se dříve či později narazí na správné řešení, nebo po projití všech možností potvrdí neřešitelnost. Metoda je použitelná jen do určitého rozsahu řešeného problému (počtu kombinací) a používá se jen v případě, že prohledávaná množina potenciální řešení je relativně malá a/nebo pokud neexistuje efektivnější řešení.

To ovšem neovlivní zásadní fakt, že prostým zvýšením výkonu nemůžeme kompenzovat exponenciální náročnost úlohy.

I kdybychom zvýšili rychlost počítače stokrát či tisíckrát, dostali bychom se o jeden, dva nebo tři znaky dále. Pokud bychom dokázali řešit milion sekvencí za sekundu, tak narazíme na problém ihned na dalších dvou znacích, protože přidání dvou znaků zvýší náročnost 400 krát. Přidat jeden znak navíc vyžaduje pro stejný čas výpočtu dvacetinásobné(!!) zvýšení výpočetního výkonu. V rozumném čase je tedy metoda hrubé síly zcela nepoužitelná pro jakékoliv delší řetězce.

Je to bohužel fakt, který vyplývá ze zákonů matematiky. V takových případech je v praxi složité vysvětlit nezasvěcenému člověku, že v těchto případech metoda „koupíme dvakrát rychlejší počítač“ prostě nefunguje. V rámci současné techniky je řešení hrubou silou nerealizovatelné. Tento problém snad vyřeší až konstrukce kvantového počítače, nicméně to je prozatím hudba budoucnosti.

Druhou možností je, že počítači zadáme sekvenci proteinu, který chceme štěpit a program nám vypočte všechny podřetězce aminokyselin, které mohou vzniknout. Je pochopitelně vyžadováno, abychom sekvenci znali. Pak jsou možné dva postupy.

První použijeme, pokud nevíme pravidla, podle kterých se bude štěpit. V tom případě musíme vypočítat štěpy všech délek. Výpočet probíhá tak, že začneme od začátku původního řetězce a otestujeme všechny podřetězce, které mohou vzniknout. Následně se přesuneme o další písmeno a pokračujeme v testování. Získané řetězce poté porovnáváme s výsledky ze spektrometru a hledáme shodu v hmotnosti.

Zde je celá situace příznivější co s týče časové náročnosti. Z předem známého řetězce délky n můžeme tímto postupem vytvořit teoreticky celkem $(n^2+n)/2$ řetězců[10]. Reálná hodnota by byla ovšem nižší, protože by se pravděpodobně vyskytovaly duplicitní řetězce. Nicméně ani toto řešení není příliš efektivní co se týče počtu operací. Počet vzniklých řetězců by byl stále vysoký, takže i další zpracování výstupu (vyhledání shod v hmotnostním spektru) by vyžadovalo slušný výpočetní výkon.

10 Z řetězce o délce n znaků lze vytvořit celkem n podřetězců (včetně plné délky) tak že postupně odebíráme poslední znak a zbytek řetězce prohlásíme za podřetězec. Tak dostaneme všechny podřetězce začínající prvním znakem. Následně odebereme první znak a dostaneme řetězec délky $n-1$. Z něj opět odebíráním posledních znaků vytvoříme řetězce, které tentokrát začínají druhým znakem z původního řetězce. Takových znaků dostaneme $n-1$. A opět pokračujeme odebráním prvního znaku, tím dostaneme řetězec délky $n-2$ a z něj vytvoříme $n-2$ podřetězců. Takto pokračujeme až na řetězec délky 1, který je sám o sobě jediným podřetězcem. Celkový počet řetězců je tedy $n + (n-1) + (n-2) + ... + 1$. To je ovšem součet posloupnosti čísel známý z historky o mladém Friedrichu Gaussovi, který bleskurychle sečetl všechna čísla od 1 do 100 tak, že sečetl první číslo s posledním a vynásobil výsledek počtem dvojic. Součet se tedy vypočte ze vztahu $(n+1)\cdot(n/2)$ což je $(n^2+n)/2$.

Nicméně je to řešení, které lze použít, pokud známe původní sekvenci, ale neznáme pravidla, dle kterých se štěpilo.

Ještě lepších výsledků můžeme dosáhnout, pokud známe pravidla, dle kterých se štěpí, resp. pokud známe enzym, který v souběžném experimentu štěpení katalyzuje. Každý z enzymů má svoje specifické účinky a ke štěpení dochází jen před, za, nebo mezi určitými konkrétními aminokyselinami.

Pokud dokáže software simulovat štěpení podle zadaných pravidel, počet štěpů opět výrazně poklesne. Nicméně požadavky na počátku výpočtu jsou ještě striktnější – musíme znát nejen počáteční řetězec, ale i pravidla, podle kterých se štěpení uskuteční.

Software ProteinCutter pracuje právě s posledně uvedenou možností.

4.3 Proč vznikl ProteinCutter

Hlavní motivací pro vytvoření ProteinCutteru byly chybějící funkce u stávajících programů pro simulaci štěpení. Existující software buď nepočítal výsledky dostatečně přesně a/nebo nedisponoval potřebnými funkcemi. Žádný software nedisponoval všemi potřebnými funkcemi dohromady. Navíc ani jeden nebyl schopen simulovat štěpení proteinů pomocí enzymů, které se používaly při výzkumu na Katedře biochemie PřF UP.

Proto bylo přistoupeno k vývoji vlastní aplikace, které jednak měla spojit dohromady schopnosti stávajících aplikací, přidat něco nového a především pokrýt potřeby Katedry biochemie PřF UP.

4.4 Uživatelské rozhraní

Z uživatelského hlediska jde o čistě webovou aplikaci přístupnou prostřednictvím webového prohlížeče, což zajišťuje platformovou nezávislost. K využití webového rozhraní jsem přistoupil z toho důvodu, že od uživatelů nevyžaduje absolutně žádné zvláštní nároky. Postačí webový prohlížeč a funkční připojení k internetu.

Práci s uživatelským rozhraním [Obr 11, Obr 12] lze rozdělit do čtyř kroků – vložení dat, nastavení možností štěpení, nastavení pravidel štěpení a nastavením filtrů.

4.4.1 Vložení dat

Vložení dat je možné provést buď manuálně zapsáním sekvence proteinu do formuláře, nebo importem ze vstupního souboru. Formát vstupu odpovídá běžným standardům, tj. jako řetězec jednopísmenných zkratek aminokyselin.

Při vkládání dat jsou vedle okna formuláře zobrazovány hodnoty chemickofyzikálních veličin vloženého řetězce, které jsou automaticky aktualizovány a v případě vložení chybného symbolu je uživatel na tuto chybu neprodleně upozorněn.

V případě, že má uživatel data připraveny v souboru, je možné provést import ze souboru. Aplikace zvládá soubory ve formátu FASTA a TXT (tzv. „plain-text“). V obou možných případech vstupu ještě aplikace provádí ošetření vstupu tak, aby byly odstraněny znaky, které do řetězce nepatří (např. bílé znaky).

Aplikace také dokáže přeložit sekvence DNA.

4.4.2 Nastavení možností (modifikací)

Nastavení možností (modifikací) umožňuje zvolit, jakým způsobem budou data vypočítávána. Poskytované možnosti jsou:

- ionizační stav vzniklého peptidu – $[M+H]^+$, M, $[M-H]^-$
- modifikace cysteinu – (bez modifikace, karbamidomethylace CysCAM, karboxymethylace CysCM)
- variabilní oxidace methioninu na methioninsulfoxid
- vynechaná dělení – vypočítá všechny možnosti, které mohou vynecháním vzniknout

První tři možnosti modifikují hmotnost fragmentu. Vynechaná dělení slouží k simulaci situace, kdy nedojde (např. z důvodu prostorové nepřístupnosti) ke štěpení ve všech možných místech štěpení. Taková „chyba“ se při štěpení stává kupodivu docela často, což potvrdily i zkušenosti s používáním ProteinCutteru. V grafickém rozhraní ProteinCutteru lze nastavit až čtyři vynechaná dělení, interní funkce umožňují neomezený počet (resp. počet možný pro daný řetězec).

4.4.3 Nastavení pravidel štěpení

Nastavit pravidla pro štěpení lze dvěma způsoby. Buď vybrat konkrétní látku, která bude štěpení iniciovat (pravidla se nastaví automaticky), nebo si uživatel nastaví vlastní pravidla.

Pro výběr je v aplikaci připraveno devatenáct nejčastěji používaných látek. V nabídce je Trypsin, Arg-C, Asp-N, Asp-N_ambic, Chymotrypsin, CNBr, CNBr+Trypsin, Formic_acid, Lys-C, Lys-C/P, PepsinA, Tryp-CNBr, TrypChymo, Trypsin/P, V8-DE, V8-E, semiTrypsin, LysC+AspN, Prolyl_endoprotease. V případě požadavku lze do aplikace další štěpící látky přidat snadnou úpravou rozhraní.

Pro manuální vložení je k dispozici jednoduché rozhraní, kde lze pravidla „naklikat“. Nebo může šikovnější uživatel pravidla přímo zapsat do připraveného okénka formuláře. Zápis pravidel je jednoduchý – pravidla jsou oddělena čárkami, pomlčky značí místa dělení. Např. „U-,X-Y,-Z“ znamená štípat za „U“, štípat mezi „X“ a „Y“ a štípat před „Z“.

4.4.4 Nastavení filtrů

V části určené pro nastavení filtrů je k dispozici rozhraní, které umožní nastavit, které řetězce aminokyselin budou zobrazeny. Pro každou z šesti veličin (monoizotopická a průměrná hmotnost, délka, hydropatie, NPS a izoelektrický bod) je možné nastavit intervaly hodnot **`větší než`**, **`menší než`** nebo rozsah **`od-do`**. Ve výpisu se potom objeví jen takové sekvence řetězců, které splní všechny podmínky definované ve filtrech.

4.4.5 Výstup výpočtu

Výstupem je tabulka s jednotlivými sekvencemi aminokyselin. U každého z řetězců jsou zobrazeny hodnoty veličin. V případě potřeby je možné u každého z řetězců pokračovat v dalším štěpení.

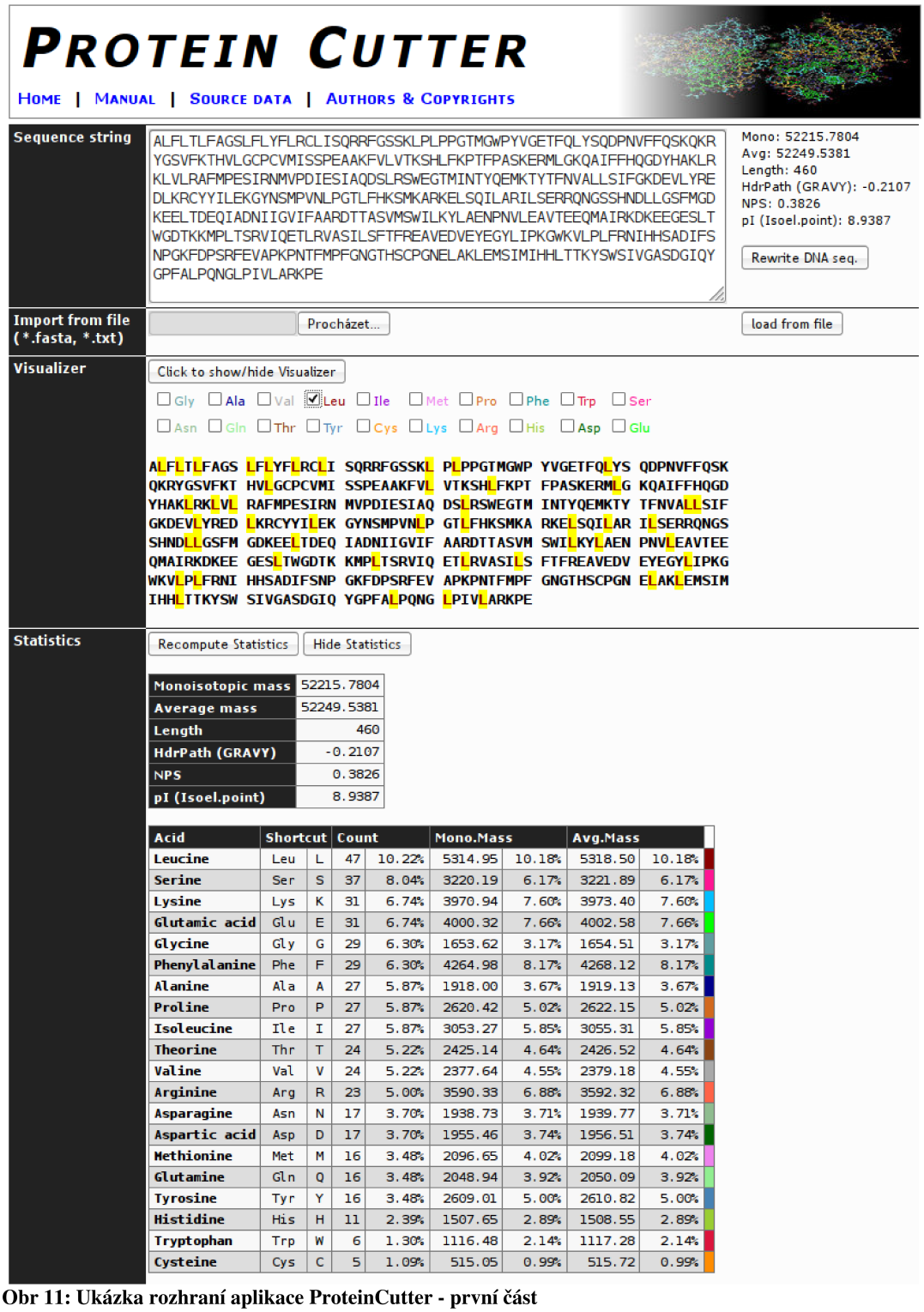
Pro přehlednost výstupu je možné nechat výstup seřadit vzestupně nebo sestupně podle hodnot jednotlivých veličin a zapnout filtry. K dispozici je i zobrazení dat vhodné pro tisk.

4.5 Vypočítávané fyzikálně-chemické parametry

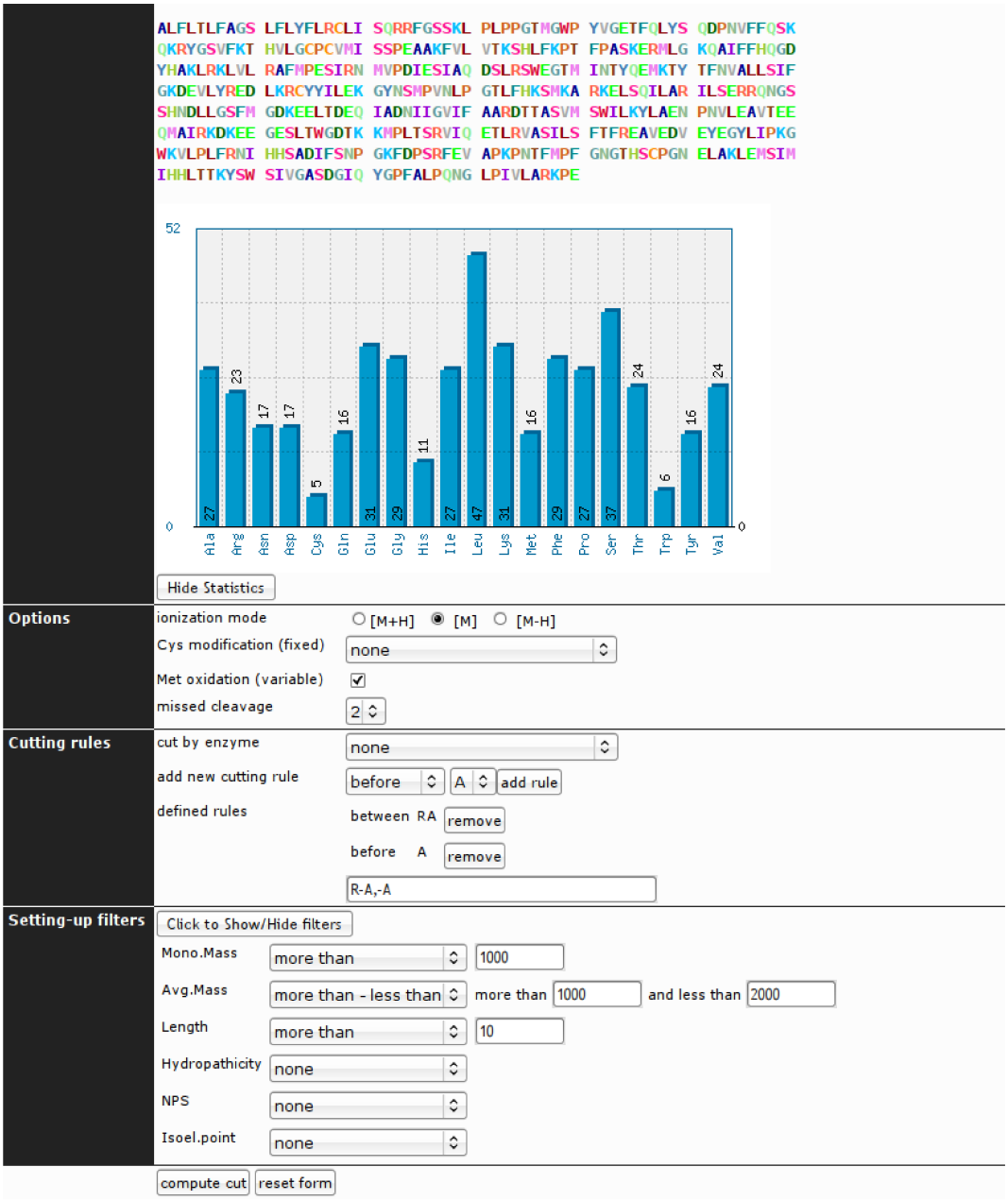
Aplikace je schopna pro každý řetězec vypočítávat následující hodnoty:

- monoizotopická hmotnost (monoisotopic mass)
- průměrná hmotnost (average mass)
- hydropatie (hydropathicity)
- NPS
- izoelektrický bod

U hmotností se na základě nastavení započítávají modifikace methioninu a cysteinu a oxidace methioninu.



Obr 11: Ukázka rozhraní aplikace ProteinCutter - první část



Obr 12: Ukázka rozhraní aplikace ProteinCutter - druhá část

4.6 Technická realizace

4.6.1 Postup při zpracování dat

ProteinCutter fakticky pracuje s textovými řetězci. Aby splnil požadavky, kvůli kterým byl vytvořen, musí být schopen vykonat dvě úlohy. Tou první je rozštěpení vstupního textového řetězce na základě stanovených pravidel. Tou druhou je schopnost vypočítat hodnoty veličin a to opět na základě zadaných požadavků (např. M+H, M, M−H, CysCAM, CysCM …).

Zpracování vstupu začíná sjednocením textového vstupu na jednotný formát. Jsou odstraněny prázdné bílé znaky, jsou odstraněny symboly, které nejsou symbolem pro aminokyselinu a všechny zbývající znaky jsou převedeny na velká písmena.

Poté dojde k samotnému „štěpení“, kde prvním krokem je označení míst, kde se bude štěpit (použije se vložení pomlčky). Označení míst je poněkud komplikovanější a vyžaduje zohlednění pravidel „štěpit před“, „štěpit mezi dvěma“ a „štěpit před“. Celé štěpení navíc komplikuje přítomnost Prolinu (v některých případech zabraňuje štěpení). Algoritmus pracuje tak, že musí procházet znak po znaku a kontrolovat předchozí i následující znak. Po označení štěpných míst následuje samotné rozštěpení. Výsledkem je seznam (datová struktura) obsahující po sobě jdoucí sekvence fragmentů.

Tyto sekvence se musí dále zpracovat, pokud se provádí i výpočet vynechaného štěpení. Vynechané štěpení se simuluje tak, že se postupně berou dvojice, trojice, čtveřice a delší n-tice (podle zvoleného počtu vynechaných štěpení) a slučují se do nových řetězců. Tím vzniknou neúplně či nesprávně rozštěpené fragmenty, který by vznikly při nedokonalém štěpení.

Po dokončení simulace štěpení je jsou ze získaného seznamu fragmentů odstraněny duplicitní výskyty. Poté je již možné vypočítat všechny fyzikálně-chemické veličiny, stanovit pozici fragmentu v původním řetězci a nakonec uživateli vypsat výsledky v podobě tabulky.

4.6.2 Technologie aplikace

ProteinCutter je webová aplikace, z čehož vyplývají její specifika i použité technologie. Aplikace pracuje na webovém serveru, kde probíhají veškeré výpočty. Uživatelé ji používají pomocí webového rozhraní (což je webová stránka). Pro její používání tedy uživateli stačí jakýkoliv moderní webový prohlížeč a přístup k internetu, není nutné nic do počítače instalovat. Aplikaci ovšem není možné provozovat v off-line režimu, tj. bez připojení k internetu

(resp. k webového serveru). Samotné stránky mají z technického hlediska dvě části – část na straně uživatele (to co je ve webovém prohlížeči) a část na straně serveru.

Část na straně uživatele je jen výstupem ze serverové části. Je vytvořena v jazycích XHTML, CSS a JavaScript (který musí být v prohlížeči povolen). Důležitou součást aplikace tvoří JavaScriptová knihovna **jQuery**[jQuery], která aplikaci poskytuje funkce AJAXu. AJAX (**a**synchronní **Ja**vaScript a **X**ML) tvoří velmi významný prvek aplikace, protože dodává aplikaci možnosti Web2 (viz. dále). Uživatelská část je validována dle webových standardů W3C, což zajišťuje korektní zobrazení ve všech moderních webových prohlížečích, které tyto standardy dodržují.

Serverová část aplikace je zcela nezávislá na použitém webovém prohlížeči uživatele. Je vytvořena ve skriptovacím jazyku PHP5. Pro samotný provoz je nutné použít webový server se zcela standardní podporou jazyka PHP5. Nejvhodnějším typem webového serveru prot tyto účely je Apache, který je celosvětově nejrozšířenější, nicméně je možné použít i jiný (např. IIS). Pro samotný běh není třeba na serveru ani v konfiguraci PHP5 provádět žádných nestandardních nastavení nebo úprav.

Při tvorbě aplikace bylo pro jádro použito tzv. objektového programování a došlo tak k oddělení vzhledu od logiky aplikace. To v budoucnu umožní použít naprogramované funkce v jiné aplikaci buď bez úprav, nebo jen s malými úpravami v kódu.

4.7 Čím je aplikace ProteinCutter zajímavá

4.7.1 Množství funkcí, přesnost, univerzálnost

V první řadě je třeba říci, že ProteinCutter v sobě spojuje možnosti, které dosud byly umístěny v několika nezávislých aplikacích. To umožňuje získat potřebné informace jediným výpočtem na jednom místě a používat jen jeden nástroj.

Důležitá je i přesnost vypočítávaných veličin. Protože stávající aplikace dávaly při stejných vstupech různé výsledky, bylo dbáno na to, aby použité algoritmy byly co nejpřesnější. Řada konstant a vzorců používaných při výpočtech byla ověřena z více zdrojů. Přesnost výsledků byla ověřena a prokázána i v praktickém použití.

Nezanedbatelným prvkem je i univerzálnost aplikace. Nabízí uživateli možnost nadefinovat si libovolná vlastní pravidla štěpení, což jiné aplikace neumožňují. Pro vědecké pracoviště Katedry biochemie UP je důležitým prvkem možnost provádět výpočet štěpení s několika zde používanými enzymy, které dosavadní aplikace nenabízely.

4.7.2 Uživatelské rozhraní a Web2

Z vědeckého hlediska nedůležitým, ale pro uživatele příjemným zpestřením je řešení uživatelského rozhraní. Protože mám zkušenosti s webdesignem a webovými aplikacemi, navrhl jsem rozhraní s ohledem na uživatelskou přívětivost a přehlednost. Uspořádání prvků je takové, že uživatele v podstatě vede od zadání dat na začátku přes veškeré nastavení až po samotný výpočet, který proběhne na konci.

Velice důležitou roli v návrhu hrálo využití prvků, které se označují jako Web2. V nich je opuštěna klasická koncepce webu, kdy každé kliknutí na aktivní prvek ve stránce vyžaduje její opakované načtení. Použití technologií AJAX umožňuje aplikaci komunikovat se serverem na pozadí a provádění výpočtů tak probíhá bez opakovaného načítání stránky. Podobným způsobem funguje např. našeptávač Google nebo internetové pohyblivé mapy.

V ProteinCutteru se tyto technologie používají např. v automatickém výpočtu hodnot veličin vedle okna pro vkládání vstupního řetězce, ale i jinde (zobrazování tabulek, definice štěpících pravidel atd...).

4.8 Shrnutí a publikace

Aplikace ProteinCutter se ukázala jako použitelný a užitečný nástroj pro predikci teoretického štěpení proteinů. Přesnost a správnost teoretických výpočtů byla ověřena při praktickém výzkumu samotnými uživateli, stejně jako použitelnost a funkčnost programu.

Význam ProteinCutteru nespočívá v klasickém vědeckém přínosu jak je běžně chápán na akademické půdě, ale jde o nástroj, jehož význam tkví v možnostech, které poskytuje pro další výzkum. Jde o komplexní nástroj pro predikci štěpení proteinů, zaplňuje prázdné místo ve stávající kolekci softwaru a svými možnostmi zjednodušuje práci uživatelů.

Aplikace byla uvedena ve dvou publikacích. První z článků se věnuje výsledkům, k jejichž získání byl PoteinCutter využit. Druhý článek je věnován samotné aplikaci. Oba články jsou přílohou této práce.

- Sebela M., Rehulka P., Kabrt J., Rehulkova H., Ozdian T., **Raus M.**, Franc V., Chmelik J. (2009) *Identification of N-glycosylation in prolyl endoprotease from Aspergillus niger and evaluation of the enzyme for its possible application in proteomics.* J. Mass Spectrom. 44 (11), 1587-1595
- **Raus M**, Kopečný D, Šebela M (2013) *Program application for the prediction of results of protein digestion by proteolytic enzymes [Programová aplikace pro predikci výsledků štěpení proteinu proteolytickými enzymy]*. Chem. Listy 107 (1), 44-53.

5 Izotop

*Aplikace **Izotop** vznikla pro výpočet rychlosti reakce a látkové výměny ve sledovaných rostlinách, konkrétně při biosyntéze cytokininů. Cílem bylo nahradit dosud používané řešení výpočtu pomocí tabulkového kalkulátoru, které bylo pomalé a pracné.*

5.1 Měření rychlosti biosyntézy v rostlinách

Měření rychlosti biosyntézy se provádělo metodou značení deuteriem in vivo (deuterium in vivo labeling). Rostliny jsou pěstovány ve sterilních podmínkách v tekutém kultivačním médiu, které obsahuje těžkou vodu (s deuteriem místo vodíku). Voda se snadno dostává do všech částí rostliny, je distribuována do všech kompartmentů buněk, kde se účastní biochemických pochodů. Pomocí hmotnostní spektrometrie je zjišťováno, jaká je míra inkorporace deuteria do nově syntetizovaných molekul.

Na začátku experimentu není v rostlině žádné deuterium. S probíhající látkovou výměnou se deuterium dostává do rostlinných pletiv. Měření rychlosti potom probíhá tak, že se sleduje narůstající poměr deuteria v molekulách, jež jsou předmětem našeho zájmu. Protože je mezi jednotlivými rostlinami rozdíl v rychlosti transportu vody, vyžaduje každý experiment velké množství biologických replikátů, což přináší poměrné velké soubory dat.

Dosavadní řešení využívalo připravenou tabulku se vzorci pro tabulkový kalkulátor (MS Excel), která byla připravená pro omezený počet vstupů. Hodnoty se ručeně vkládaly do tabulky a výpočet proběhl automaticky. Použití této tabulky sice dávalo správné výsledky, ale zpracování dat bylo zdlouhavé a pracné.

Proto vznikl požadavek na co nejvíce automatizovaný nástroj, který by dokázal data zpracovat hromadně. To byl impuls pro vznik aplikace Izotop.

5.2 Aplikace

Výsledná aplikace [Obr 14] je relativně jednoduchá, nicméně je obecnější a univerzálnější než původní řešení. Představuje výrazné zlepšení v několika směrech.

V první řadě se výrazně zlepšil způsob zadávání dat reprezentujících standardy. V původním řešení musel uživatel zadat právě tři vzorové standardy a z nich se vypočítával průměr pro samotné zpracování dat. Nově lze zadat neomezené množství standardů. Z nich se průměr

počítá automaticky. Ale v případě, že uživatel požadovaný vzorový průměr již zná, může ho nastavit manuálně bez potřeby vkládat více standardů. V případě potřeby je možné načíst standardy hromadně ve formátu CSV. To vše značně zjednodušuje práci.

Další výrazné vylepšení nastalo při práci se samotnými daty. Zatímco tabulka umožňovala zpracovat jen omezené množství vstupů, Izotop zvládá přepočet hromadně. Jednak umožňuje vkládat neomezené množství vstupů, ale také poskytuje možnost importu z CSV souborů. A navíc jako bonus dává možnost rozdělit vložené vstupy do několika skupin (samostatných tabulek), což zvyšuje přehlednost. Vypočítaná data jsou potom k dispozici ke stažení i v podobě CSV souboru.

Velkou výhodou aplikace je především její komplexnost. Klíčový význam má potom rychlost a pohodlnost při importu a exportu dat. Používaný formát CSV („comma separated values“ – hodnoty oddělené částkou) je velice vhodný, protože ho lze bez větších problémů používat i pro jiné aplikace, především pro tabulkové kalkulátory (MS Excel, OO/LO Calc), které s ním nemají problém a jsou často používány vědeckými pracovníky.

5.3 Algoritmus

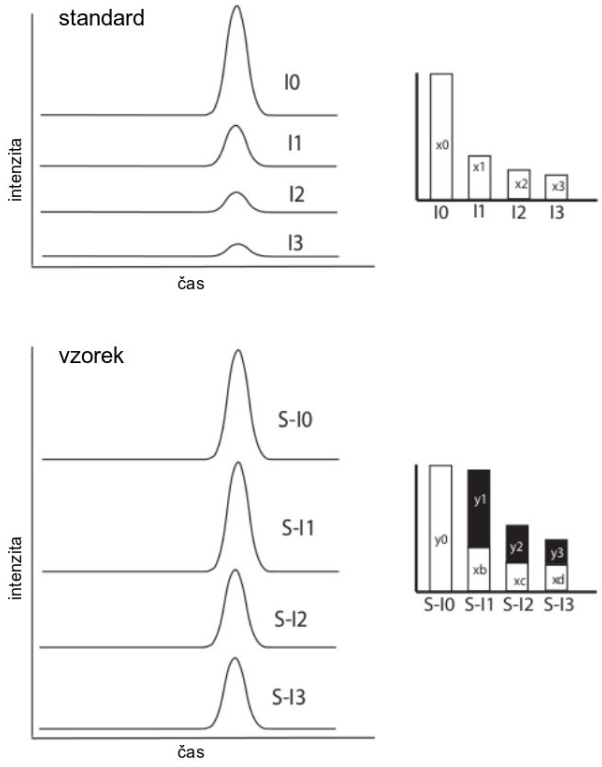
Aplikace přijímá na vstup vzorové standardy a analyzované vstupy. Vzorové **standardy** reprezentují přirozené zastoupení izotopů v přírodě. Analyzované **vstupy** obsahují laboratorně změřené vzorky, které již obsahují změny způsobené inkorporovaným deuteriem. Protože se při experimentech používá deuterium jen ve stopovém množství (z důvodu finanční náročnosti) a měření je tedy výrazně ovlivněno chybami, provádí se větší počet měření, která je nutné zprůměrovat pro eliminaci chyb. Aplikace tedy vyhodnocuje velký počet vstupů.

Výsledkem je výpočet relativních změn v poměru izotopů vůči přirozenému stavu. Toho dosáhneme tak, že vypočteme pro každý zkoumaný vzorek relativní odchylku vůči standardu a následně hodnoty zprůměrujeme pro celou skupinu vzorků. Tím dosáhneme eliminace chyb způsobených měřením na hranici detekce.

Při přepočtu vzorových standardů jsou vstupem uspořádané čtveřice. Každý standard je tvořen čtveřicí $[i_1, i_2, i_3, i_4]$. Tento standard je normalizován do tvaru $[1, i_2/i_1, i_3/i_1, i_4/i_1]$. Ze všech normalizovaných standardů se vypočte průměrný standard $[s_1, s_2, s_3, s_4]$, kde jednotlivé složky s1 až s4 odpovídají průměru jednotlivých složek normalizovaných vzorových standardů.

Přepočet analyzovaných vstupů, které jsou porovnávány s průměrným standardem $[s_1, s_2, s_3, s_4]$, pobíhá podobně. Ze zkoumaného vstupního vzorku $[v_1, v_2, v_3, v_4]$ se podle standardu vypočítají

teoretické hodnoty jednotlivých složek $[x_1, x_2, x_3, x_4]$ kde $x_2 = v_1 \cdot s_2$, $x_3 = v_1 \cdot s_3$, $x_4 = v1 \cdot s_4$ a nakonec $x_1 = s_1 + s_2 + s_3 + s_4 - v_2 - v_3 - v_4$. Tím dostaneme hodnoty, které by měl mít vzorek, kdyby byl standardní.

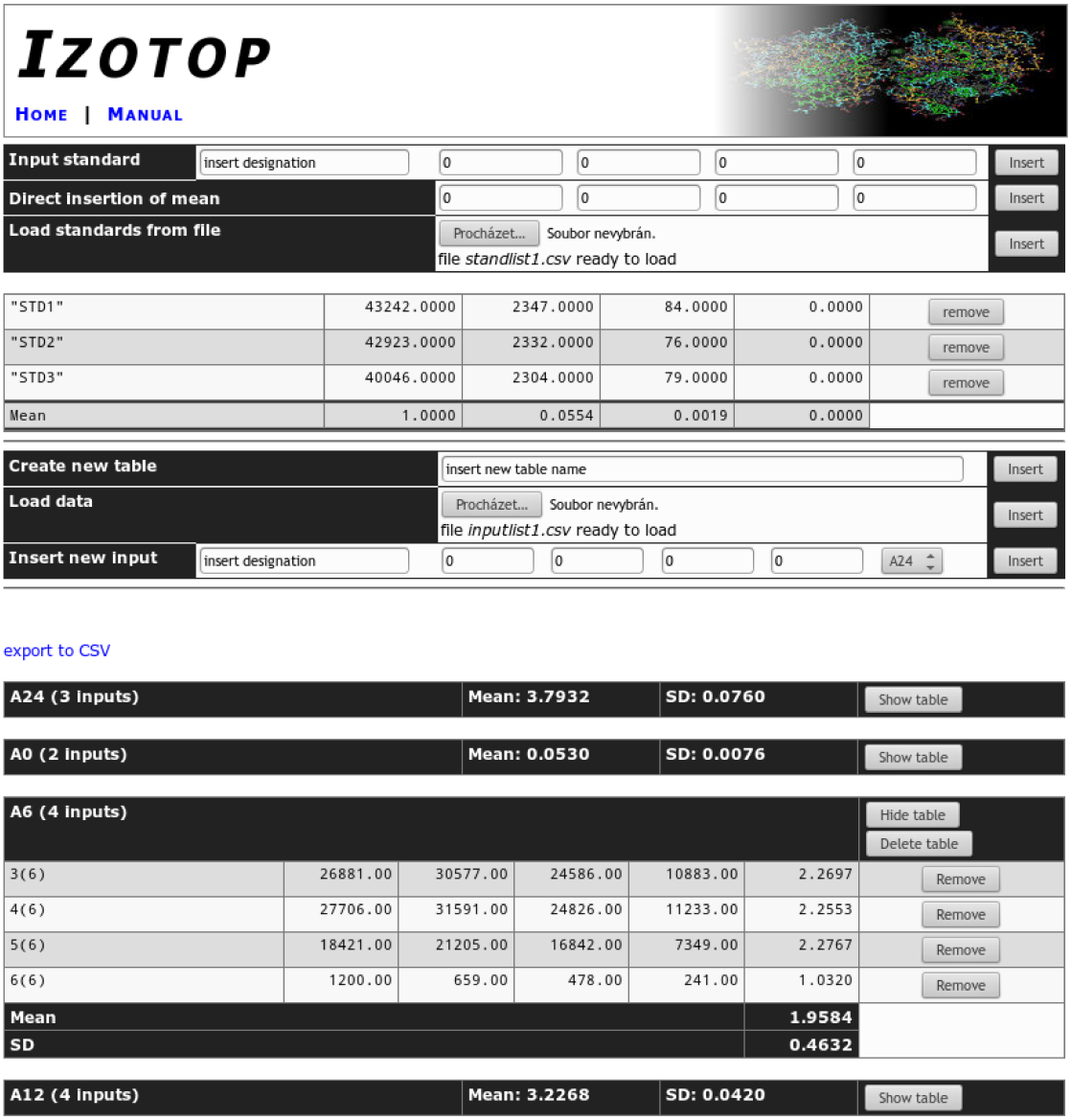


Obr 13: Srovnání standardu a deuterovaného vzorku

Standard [I0, I1, I2, I3] má přirozený pomět izotopů [x0, x1, x2, x3]. Vzorek [S-I0, S-I1, S-I2, S-I3] je ovlivněný inkorporací deuteria. Cílem je vypočítat relativní přírůstek izotopů (černé oblasti y1, y2 a y3) proti přirozenému stavu, jaký by byl v případě, kdyby vzorek žádné deuterium neobsahoval (oblasti y0, xb, xc, xd).
zdroj: Nordström A., Cytokinins in Arabidopsis, Tools, Pathways and Interaction with Auxin, str. 29 (2004) Swedish University of Agricultural Sciences, Umeå [online] http://pub.epsilon.slu.se/601/1/Silvestria_317.PDF [26. 6. 2013]

S pomocí tohoto mezivýsledku (standardnímu poměru pro daný vzorek) nyní musíme vypočítat poměr mezi reálně změřenými hodnotami a standardem. Tento poměr *p* pro celý vzorek se vypočte ze vztahu $p = (v_2 + v_3 + v_4 - x_2 - x_3 - x_4) / (v_1 + x_2 + x_3 + x_4)$, kde veličiny *v* jsou skutečné vstupní hodnoty vzorku a veličiny *x* jsou teoreticky vypočtené hodnoty, které by vzorek měl mít, kdyby byl standardní (bez deuteria).

Tyto operace se provedou hromadně nad celou množinou vstupů. Pro celou sadu vzorků se společný poměr vypočítá jako průměr hodnot jednotlivých vzorků. Aplikace vypočte a zobrazí i průměrnou odchylku pro sadu vzorků.



Obr 14: Rozhraní aplikace Izotop

V horní části rozhraní jsou textová okna pro manuální zadání standardu (pojmenování a čtveřice hodnot). V případě, že standardy známe, můžeme je zadat přímo (Direct insertion…) nebo je můžeme načíst v podobě CSV souboru. Pod vstupním formulářem pro standardy se již vložené standardy zobrazují společně s průměrným standardem.
Pod částí zobrazující standardy je formulář pro zadávání vstupů. Uživatel si může vytvořit samostatnou tabulku pro každou řadu vstupů. Vstupy lze zadávat manuálně jako čtveřici hodnot, nebo načíst hromadně z CSV souboru. Jednotlivé tabulky s přepočítanými vstupy se zobrazují pod formulářem. Do jednotlivých tabulek je možné přidávat další vstupy nebo tyto vstupy editovat (přepsat stejným označením řádku) či vstupy odebírat. Pro přehlednost je možné tabulky skrýt pomocí tlačítka a zůstane zobrazen jen průměr a odchylka. Výsledky je možné exportovat v podobě CSV souboru.

5.4 Technologie

S ohledem na flexibilitu použití je aplikace vytvořena jako webová. Díky tomu je při dostupném připojení k Internetu snadno použitelná odkudkoliv a na jakémkoliv OS.

Samotná aplikace je vytvořena v jazycích HTML, JavaScript, CSS a PHP [PHP]. Statickou část zajišťuje jazyk HTML (obsah) a CSS (vzhled). Dynamická část na serveru je vytvořena v PHP, dynamickou část na straně klienta (webový prohlížeč) obstarává JavaScript. Pro komunikaci mezi klientem a serverem je využíváno technologií AJAX. Pro jejich snadnější použití je využívána JavaScritptová knihovna jQuery [jQuery], která je k dispozici pod svobodnou licencí.

Pokud jde o požadavky na zařízení nutné k provozu, tak pro serverovou část je nutný webový server (nejčastěji Apache) s podporou PHP. Od samotného klienta (uživatel) se vyžaduje jen webový prohlížeč a přístup k Internetu.

5.5 Shrnutí a publikace

Aplikace Izotop je jednoúčelová aplikace, která se osvědčila jako nástroj, který dokáže výrazně zjednodušit zpracování velkého množství dat. Byla uvedena v jedné publikaci spolu s výsledky, které byly za její pomoci zpracovány.

- Tarkowski P., Flokova K., Václavikova K., Jaworek P., **Raus M.**, Nordström A., Novak O., Dolezal K., Sebela M., Frebortova J. (2010) An improved in vivo deuterium labeling method for measuring the biosynthetic rate of cytokinins. Molecules 15 (12), 9214-9229.

Publikace se věnuje výsledkům, které byly zpracovávány i za pomoci Izotopu. Text publikace je přílohou této práce.

6 Biospean

*Webová aplikace **Biospean** je zaměřena na analýzu a porovnání velkého množství hmotnostních spekter. Aplikace umožňuje hmotnostní spektra zkoumat, porovnávat s ostatními spektry, hledat znaky společné pro celé skupiny spekter. Aplikaci je možné používat i ke katalogizaci spekter a sdílet je v rámci pracovního týmu.*

6.1 Hmotnostní spektrometrie

Hmotnostní spektrometrie (v biochemii se používá například varianta MALDI, resp. MALDI-TOF [Havliš 1999]) je metoda založená na rozdělení částic podle hmotnosti, resp. *m/z*. Slouží ke kvantitativní analýze, tj. zjištění množství jednotlivých složek ve zkoumaném vzorku.

Jedná se o metodu, která je velmi citlivá a rychlá. Lze ji použít pro hmotnostní analýzu peptidů, bílkovin, nukleových kyselin i nízkomolekulárních látek (organických a anorganických) a její význam v oblasti biochemie neustále roste.

Výstupem ze spektrometru je hmotnostní spektrum. Lze je vizualizovat jako spojnicový graf, kde osa x (horizontální) zobrazuje hmotnost částic (*m/z*) a osa y (vertikální) je jejich množství (intenzita). Samotný spektrometr je zařízení napojené na počítač, který provede analýzu dat a nabídne je výzkumníkovi ve formátu vhodném pro další zpracování.

Většinou se data z hmotnostního spektrometru ukládají v podobě strukturovaného textového souboru (ASCII), ale lze použít i jiné formátu, např. CSV, XML, různé nativní formáty. Samotna data jsou fakticky jen diskrétních body, souřadnice ve tvaru [*m/z*, intenzita].

6.2 Proč vytvořit vlastní aplikaci

Vývoj aplikace Biospean iniciovala potřeba porovnávat podobnost různých hmotnostních spekter. Převedeno do praktického života – máme nově získané hmotnostní spektrum a chceme najít v databázi vzorky, které mu jsou nejpodobnější. Cílem může být identifikace konkrétního druhu vzorku nebo porovnání se známými vzorky (např. jak moc se od sebe liší jednotlivé vzorky stejného druhu).

Toto prakticky není možné provádět bez zapojení výpočetní techniky. Jmenovitě tedy bez softwaru, který umí porovnávat hmotnostní spektra. A v takové situaci lze buď sáhnout po řešení, které existuje, nebo vyvinout vlastní aplikaci.

S existujícím řešením se ale ukázal problém. Většina stávajícího „hmotnostního“ softwaru (např. mMass) totiž není schopna hromadného vyhledávaní mezi mnoha spektry. Skutečně hromadné porovnávání zvládá jen pár programů.

Praktické zkušenosti na katedře jsou pouze se softwarem Bruker Biotyper [Biotyper]. Ten ovšem ne zcela vyhovuje potřebám katedry (viz. níže). K podobným účelům slouží i software AnagnosTec Saramis [Sar]. S ním ovšem nejsou vůbec žádné zkušenosti, resp. víme o něm jen to, co je na oficiálních stránkách. Oba uvedené programy jsou komerční.

Ze situace tedy nakonec vyplynuly důvody, proč jít vlastní cestou.

V první řadě v případě existujícího softwaru jsme odkázání na komerční řešení, což s sebou nese finanční náročnost a licenční omezení. Především v případě Biotyperu představuje pořízení relativně vysokou položku v rozpočtu. Vybavit několik pracovních stanic tímto softwarem není z finančního hlediska optimální. Náklady na vývoj vlastní aplikace, která by byla jednodušší a přitom by poskytla podobnou, resp. pro požadované účely přesněji zaměřenou, funkčnost, by se ve skutečnosti vyplatily již při malém počtu uživatelů.

Komerční produkty jsou většinou silně ovlivněny i svojí striktní licenční politikou – placený upgrade i update, technická podpora jen na určitou dobu (omezena jen na několik posledních verzí programu) a pro předem omezený počet stanic, absolutní zákaz úprav a modifikací.

Dalším důvodem vlastního vývoje je jistá nevhodnost stávajícího softwaru pro specifické potřeby Katedry biochemie Přf UP a Centra Regionu Haná (dále jen KBC a CRH). Zde si musíme uvědomit, že např. součástí Biotyperu je vedle detekčního softwaru i velice důležitá databáze vzorků, se kterou se provádí srovnání. Tato databáze je součástí aplikace, ale je zaměřená specificky například na medicínskou biochemii a pro potřeby určitých výzkumů se nehodí. V případě Saramisu bude situace obdobná.

Pracovníci KBC a CRH potřebují porovnávat svoje vlastní skupiny zkoumaných spekter a nikoliv spektra v databázi Biotyperu/Saramisu. Takže důležitým požadavkem na vlastní aplikaci je možnost vytvořit si vlastní databázi spekter. A nejlépe takovým způsobem, aby bylo možné tuto databázi sdílet s kolegy v rámci týmu a centrálně spravovat.

Třetím významným důvodem je snadná použitelnost. Program, který má mnoho funkcí a další jsou s probíhajícím vývojem přidávány se časem dostane do stavu, kdy už je příliš komplexní. Vždy však zůstanou uživatelé, kterým pro jejich potřeby bude stačit jen základní funkčnost. Proto bylo cílem vytvořit aplikaci, která by byla snadno použitelná a intuitivní.

V konečném výsledku tedy vyplynul požadavek na aplikaci vytvořenou tzv. „na míru“, která by se držela pravidla „méně znamená více“, byla by jednodušší, přímočařejší a intuitivnější na ovládání a umožnila by rychlé vložení nového spektra, jeho úpravu a katalogizaci a následné porovnání.

Pokud tedy shrneme hlavní důvody k vývoji vlastní aplikace, jsou to nedostatek finančně dostupného softwaru, vysoké náklady komerčního řešení, nutnost porovnávat vzorky s vlastní databází spekter, využití pro týmovou práci a v neposlední řadě potřeba funkčně méně komplexního, ale uživatelsky přívětivějšího řešení.

6.3 Shrnutí požadavků na aplikaci

Z hlediska výzkumu je po aplikaci požadováno, aby byla schopna analyzovat nová hmotnostní spektra a hledat shody se spektry, která již byla do databáze vložena.

V praxi to tedy znamená, že aplikace musí být schopna nalézt charakteristické znaky spektra, kterými jsou v tomto případě pozice píků, lokálních maxim, které signalizují významné množství částic se shodnou molekulární hmotností. Při této analýze se fakticky dostáváme zcela mimo oblast biochemie, protože analýza spektra je ve skutečnosti zpracováním digitálně zaznamenaného elektromagnetického signálu. Samotná analýza tedy musí využít nástrojů, postupů a poznatků z oblasti matematiky, statistiky, fyziky a elektroinženýrství.

Zde je třeba říci, že jedno softwarově analyzované spektrum není samo o sobě příliš zajímavé. Pro analýzu jednoho spektra postačí často jen grafická vizualizace, zvýraznění charakteristických znaků (píků) a jeho prostá vizuální kontrola uživatelem. Toto dokáže řada softwarových nástrojů.

Aby mělo skutečně smysl zkoumat spektra komplexněji pomocí softwaru, je třeba, aby software dokázal zpracovávat spektra hromadně. Tj. porovnat získané spektrum i s dalšími získanými spektry a také zobrazit nakolik a v jakém množství charakteristických znaků se spektra shodují, resp. liší.

Jestliže má program mít schopnost analyzovat a porovnávat spektra, je vyžadováno, aby dokázal sloužit i jako úložiště, databáze a knihovna již analyzovaných spekter. A samozřejmě provádět další operace, např. porovnávat cíleně jen s vybranou skupinou spekter, např. vybraných referenčních vzorků. Tím vyvstává požadavek na použití databáze, což sice nesouvisí se spektry jako takovými, ale přináší určité technologické požadavky na samotné softwarové řešení.

Pro potřeby výzkumných týmů také vyvstává potřeba sdílet analyzovaná spektra mezi jednotlivými členy týmu. Výzkumník si do systému uloží svoje spektra a chce je využít pro práci s kolegy. To s sebou přináší požadavek na nástroje, který umožní vytvářet pracovní skupiny uživatelů (např. v rámci jednoho projektu) a sdílet mezi sebou spektra.

Požadavky na aplikaci tedy zahrnují schopnost analyzovat, srovnávat a katalogizovat spektra. V souvislosti s tím je vyžadována i schopnost data uchovávat pro další využití, např. jako referenční databázi. Užitečné mohou být i nástroje pro týmovou práci, umožňující sdílení vybraných spekter mezi uživateli.

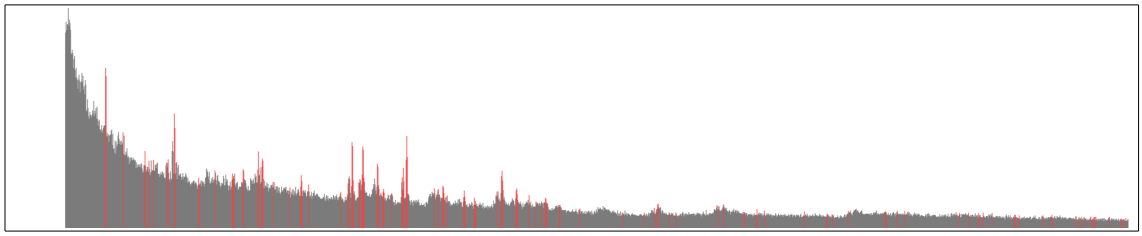
6.4 Detekce píků, princip a problémy s tím související

Pokud chceme nalézt ve spektru píky, musíme si nejprve uvědomit, jak takový pík ve skutečnosti vypadá, resp. jak se projeví v záznamu digitalizovaného spektra. pík je svým způsobem množina sousedících bodů, které mají více či méně extremální hodnoty vůči zbytku spektra, nebo alespoň vůči svému blízkému okolí. Protože spektrum může mít v každé své části jiné rozsahy hodnot jak u šumu, tak u píků, musí algoritmus pro detekci píků projít celé spektrum, analyzovat hodnoty v každém bodě a rozhodnout, zde jde o pík či nikoliv.

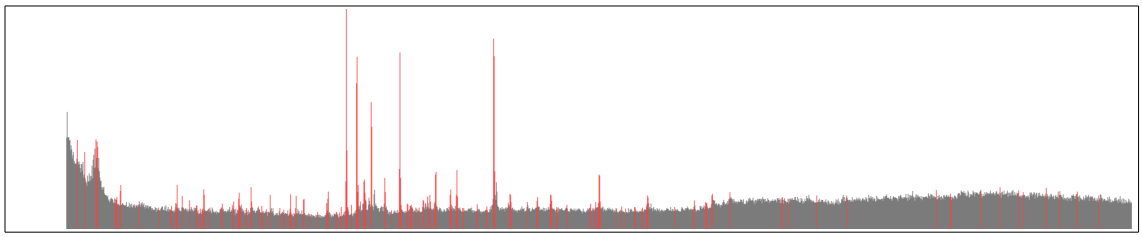
Při snaze analyzovat takové spektrum přitom narazíme na řadu problémů. Ty souvisejí už se samotným signálem, který často obsahuje velké množství mnohdy velmi nerovnoměrného šumu [Obr 15, Obr 16]. Ten spolu s málo intenzivním odezvou sledovaných složek ze vzorku značně komplikuje samotnou detekci píků. Pokud je šum nerovnoměrný, může při slabé odezvě vzorku v některých částech spektra výrazně převýšit samotné hodnoty píků, které se vyskytují v částech, které jsou šumem méně zasažené.

Problém způsobuje samotný způsob digitálního záznamu spojený s limity (či lépe řečeno způsobem fungování) současné výpočetní techniky. Zatímco signál ze senzoru ve spektrometru je analogový a spojitý, z výstupu dostaneme záznam už v digitalizované podobně. Digitalizací rozumíme převod analogového signálu na číselnou reprezentaci, kterou tvoří diskrétní body reprezentované dvojicí hodnot [*m/z*, intenzita].

Digitalizovaný záznam ovšem na rozdíl do analogového spojitý není. Jedná se řadu diskrétních bodů, které jsou uloženy ve výstupním souboru. A právě množství těchto dat je problém. Abychom dosáhli dostatečné přesnosti, musíme analogový signál digitalizovat s dostatečným počtem bodů. Datový objem takto zpracovávaných dat je, v závislosti na typu a nastavení spektrometru, relativně vysoký.



Obr 15: Spektrum s nepravidelným šumem 1
Nepravidelný šum který v určitých částech výrazně převyšuje intenzitu červeně zvýrazněných píků



Obr 16: Spektrum s nepravidelným šumem 2
Spektrum s nerovnoměrným šumem, který je v efektivní středové části nízký, ale na počátku a na konci roste.

Soubor s naměřenými daty má velikost v řádu jednotek megabytů a obsahuje někdy i více než 200 tisíc diskrétních hodnot. Rychlé zpracování takového objemu dat vyžaduje poměrně velkou výpočetní kapacitu. Ačkoliv samotné vyhodnocení a porovnávání jednoho bodu vyžaduje relativně velmi jednoduché operace, je nutné tyto operace provést na všechny body spektra. To dohromady dává velké množství výpočetních kroků.

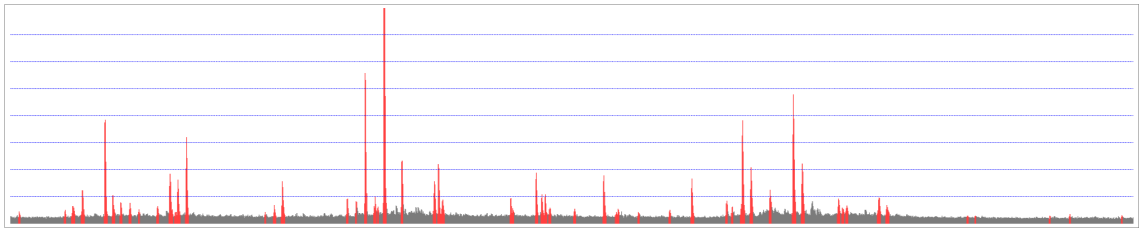
6.5 Detekce píků za přítomnosti šumu

Základním klíčem k funkčnosti softwaru Biospean je schopnost detekovat píky. Na vytvoření postupu pro jeho detekci jsem vyzkoušel několik metod než jsem nalezl funkční řešení. Samotné vytvoření postupu pro detekci píků tedy vznikalo po dílčích krocích, od první slepé uličky přes různě funkční mezikroky až po funkční finální řešení.

6.5.1 Horizontální řez

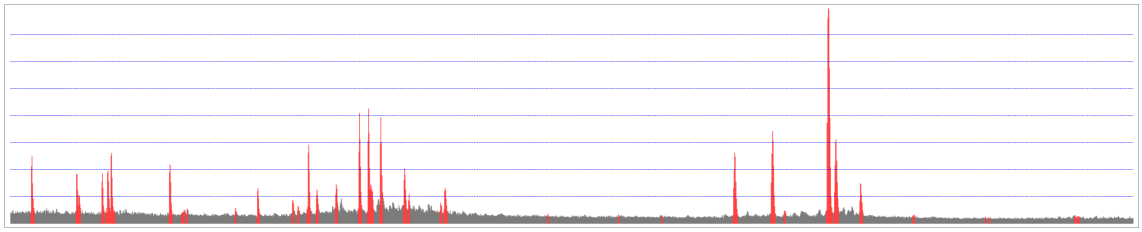
Při detekci píků byla nejprve vyzkoušena metoda, kterou lze označil za metodu „horizontálního řezu“. Metoda zaměřuje horizontální intervaly [Obr 17, Obr 18, Obr 19, Obr 20], které vzniknou protnutím spektra v určité výšce. Srovnáním polohy $[x_1,x_2]$ těchto intervalů a jejich intenzity se poté vyhledají shody proti ostatním spektrům.

Dvě spektra se porovnávala tak, že se při analýze v prvním spektru nalezl interval $[x_{11},x_{12}]$ a v druhém spektru interval $[x_{21},x_{22}]$. Tyto intervaly se následně navzájem porovnávaly, tj. shoda (průnik) intervalů nastala pokud $x_{12}>x_{21}$ a $x_{22}>x_{11}$. Potom délka průniku odpovídá $\max(x_{11},x_{21}) - \min(x_{21},x_{22})$.



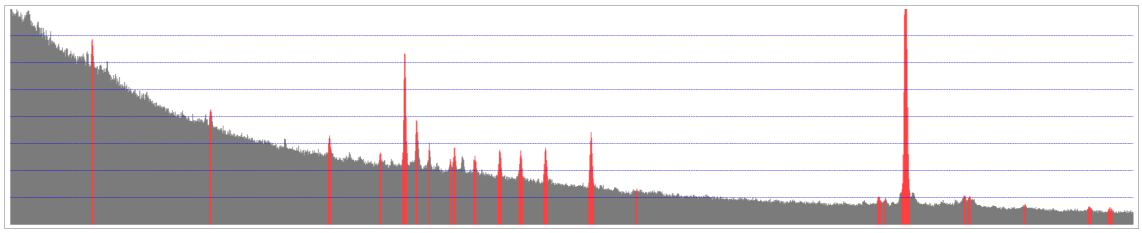
Obr 17: Použití horizontálních intervalů 1

Software zaměřuje pozice průsečíků mezi píky (červeně) a horizontálními liniemi „řezů“ (modře). Výška a pozice průsečíku mezi píkem a hladinou řezu zaznamenává pozici píku a jeho relativní intenzitu. Tato pozice se použije pro porovnání spekter.



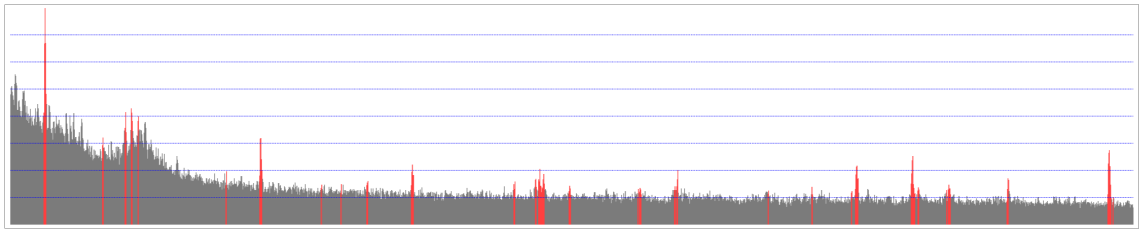
Obr 18: Použití horizontálních intervalů 2

Problém je, že intenzita píků je různá – intenzita nejsilnějšího z píků nastavuje ralativně hladiny řezů a ovlivní tak počet průsečíků. I píky, které jsou pro nás významné, mohou mít tak malou intenzitu, že je nemusí horizontální řez zachytit.



Obr 19: Použití horizontálních intervalů 3

V případě nerovnoměrného šumu pak nemají intervaly vůbec smysl, protože intenzivní šum na začátku spektra zasahuje do horizontálních „řezů“. Pokud by byly příbuzné vzorky změřeny jednou s takovýmto šumem a podruhé s menším(např. jako výše), nebylo by porovnání prakticky možné. Bez odstranění šumu tedy nemají horizontální řezy vůbec smysl.



Obr 20: Použití horizontálních intervalů 4

Opět příklad spektra, kde kvůli šumu i intenzitě jednoho z píků nelze intervaly efektivně použít. Na začátku spektra je kromě šumu i jeden vysoký pík, který nastavuje relativní výšku ostatních píků na tak malou hodnotu, že se dostanou pod úroveň naprosté většiny řezů.

Tato metoda, ačkoliv účinná na uměle vytvořených testovacích spektrech, se ovšem ukázala nepoužitelná pro praktické použití na spektrech skutečných. Důvodem byly jednak různé intenzity píků v porovnávaných spektrech, kde poloha píků byla shodná, ale intenzita signálu

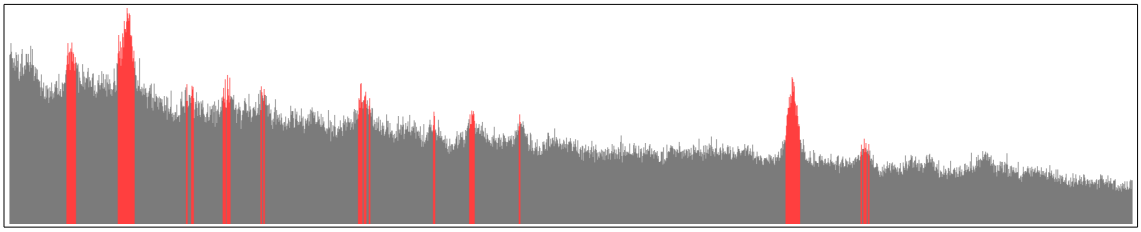
různá. Druhým, mnohem významnějším, důvodem byl velký vliv šumu, který v mnoha případech učinil tento postup zcela neúčinným. Bez možnosti efektivně odstranit šum, nebo lépe řečeno při absenci schopnosti pracovat za přítomnosti šumu, neměla tato metoda vůbec smysl.

6.5.2 Globální průměrování a eliminace šumu

Nakonec se ukázalo, že prakticky jediným řešením je ignorovat absolutní hodnotu intenzity píků a porovnávat jen jejich pozice na ose x. Tím vznikl druhý postup, který se ukázal jako v omezené míře použitelný a po úpravě vedl nakonec ke použitému řešení. Mimochodem podobně, tj. pouze s x-ovými souřadnicemi, pracuje i známý Biotyper.

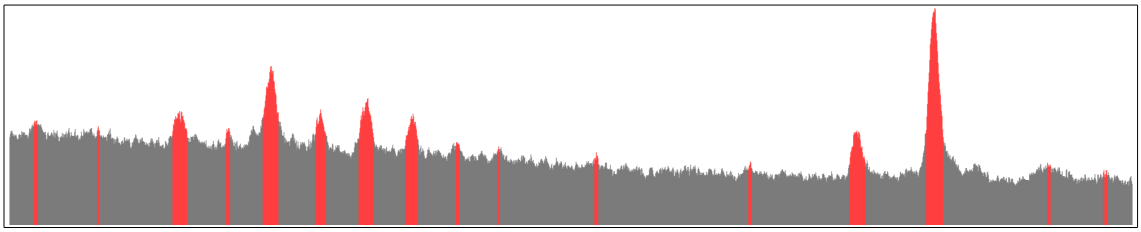
Výhodou tohoto řešení je, že je imunní vůči různé intenzitě píků, protože je nebere v úvahu. Nicméně samotné vyhledání x-ové polohy stále neodstraňuje problém se zkreslením způsobené šumem. Šum může snadno způsobit, že nedostatečně intenzivní pík nepůjde vůbec rozpoznat. Metoda se ukázala jako mnohem úspěšnější a dostatečně efektivní, nicméně je závislá na dalším důležitém kroku – na odstranění samotného šumu.

Šum může mít různý profil a charakter, takže šum v různých spektrech i při grafickém zobrazení může zblízka vypadat různě [Obr 21, Obr 22, Obr 23].



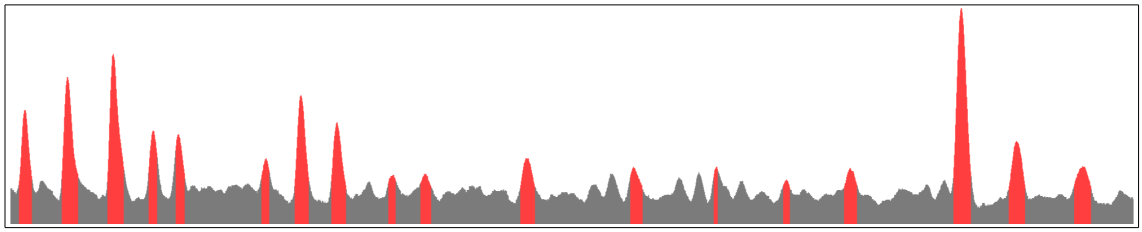
Obr 21: Spektrum s vysokým šumem

Vrcholky jsou špičky s rychlými ostrými přechody, je velký rozdíl mezi minimální a maximální hodnotou šumu. Místy lze pozorovat ne zcela jasnou hranici mezi píkem a šumem. Intenzita největších píků je dostatečná, ale menší píky se ztrácí v šumu. Je problém určit, co je pík a to i lidským pohledem. Množství chybných detekcí (pozitivních i negativních) bude vysoké.



Obr 22: Spektrum s šumem, který výrazně neovlivňuje píky

píky jsou celkem jasné, s odstupem od šumu. Skutečně sporných hraničních hodnot je jen několik. Záleží na koeficientu citlivosti, zda budou převažovat nesprávné pozitivní detekce (není pík, ale je označen), či nesprávné negativní detekce (je pík a není označen).



Obr 23: Spektrum s minimálním množstvím šumu (resp. s šumem vyhlazeným)
Kvalitní spektrum, nebo spektrum, které bylo dodatečně vyhlazeno. Změny signálu jsou plynulé, se správností pozitivní detekce není problém. Vše, co je označeno jako pík také píkem je. V některých místech se signál tvaruje do výrazné sinusoidy, což zřejmě vynechalo některé píky.

Odstraňování šumu ze signálu je přitom problém, který se řeší v mnoha oborech, nicméně v každém je přístup trochu jiný [Gunnlaugsson 2012], [DNR], [ANR], [GiNR]. Obecně můžeme šum definovat jako vysokofrekvenční vlnění s nízkou energií, které je přítomno ve sledovaném signálu, ale nemá spojitost s informací, daty, či energií, které jsou signálem přenášeny.

U analogových signálů lze uvést např. řešení pro zajištění konstantního napětí u napájecích zdrojů, kde je žádoucí (např. u citlivých laboratorních přístrojů), aby stejnosměrné napětí bylo co nejméně zvlněné. Zde se problém řeší především pomocí kondenzátorů a diod. Stejné řešení se používá i u některých přepěťových zásuvek a u všech záložních zdrojů UPS [UPS] pro stabilizaci napájení ze sítě.

Z digitálních technologií můžeme uvést např. odstraňování šumu ve snímacích čipech digitálních fotoaparátů a kamer, kde s rostoucí citlivostí (tzv. ISO) při nízké úrovni osvětlení výrazně roste šum, který se projevuje zrnitostí výsledné fotografie. V obraze se potom mohou objevit různé projevy šumu, například známý „sůl a pepř“ [Švanda 2011], [Švihlík 2008]. Zde se problém řeší většinou tzv. vyhlazováním, kdy se snímek „vyčistí“ pomocí průměrování přes tzv. masku [Hlaváč 2010]. Za cenu snížení detailů lze dosáhnout do jisté míry čistého obrazu. Tato metoda má samozřejmě svoje limity. Pokud šum přesáhne určitou hranici, tak ani sebelepší vyhlazování nepomůže. Metoda průměrování, resp. její modifikace, se dá na zpracování hmotnostních spekter použít. Výsledkem by ovšem nebylo odstranění šumu, ale pouze jeho vyhlazený průběh.

Další příkladem, kde se provádí odstranění šumu je audio průmysl, který řeší, jak se zbavit nežádoucího šumu v audiozáznamech a hudebních nahrávkách [Polzer 2008]. Vyčištění audionahrávky se provádí tak, že se nejprve získá vzorek šumu v části nahrávky, kde nehraje žádná hudba a kde se vyskytuje pouze šum. Tento vzorek se poté použije pro odstranění šumu z celé nahrávky.

V případě hmotnostních spekter by tato metoda znamenala, že by se muselo nejprve provést na spektrometru měření bez vzorku a to v celém měřeném rozsahu. Změřením spektrometru naprázdno bychom dostali profil šumu, který by se posléze použil k odstranění šumu ze vzorku. Nicméně takové řešení je nepohodlné a v praxi většinou nepoužitelné.

Jako příklad, který je naopak využitím umělého šumu, můžeme uvést například oblast vojenství, kde se šumu a inverzního signálu užívá při elektronickém boji k rušení např. radarů či komunikačních prostředků [Tůma 2011], [Havránek at al., 2007]. Pokud frekvenční pásmo, ve kterém pracuje sledovací radar či vysílač, zahltíme velkým množstvím energie, ztratí se ve výsledném šumu signál ze sledovaných objektů. Jde o analogickou situaci jaká nastává při přehlušení píků ve spektrometru – šum rušičky (šum ve spektrometru) zahltí sledované frekvence tak vysokým šumem, že nelze rozeznat odrazy cíle (slabé píky).

Naopak odstranit rušení je snadné, pokud známe jeho strukturu a dokážeme ho uměle odstranit jakýmsi „protisignálem“. Vysláním signálu s opačnou fází původní signál potlačíme (dojde k interferenci). Takto také fungují „inteligentní“ vojenské rušičky – vysíláním signálu je možné zarušit pouze radarový odraz objektu aniž by bylo poznat, že k rušení dochází. Podobného principu, ale pro změnu se zvukem, se používá např. u interiérů luxusních osobních vozů či kabin letadel, které mají aktivní rušení hluku. Rovněž podobný princip se používá u dražších modelů sluchátek, které mají aktivní potlačení hluku z okolí [Harris]. V tomto případě se využívá mikrofon sledující hluk (motoru, vrtulí, okolního prostředí) a reproduktor který vysílá rušící zvuk s opačnou fází. Za předpokladu, že rušený hluk je periodický a předvídatelný, bývají výsledky velice dobré.

Tím se dostáváme k technickému řešení.

U všech těchto možností je zcela samozřejmé, že využívají matematických metod. A to buď nějakou formu statistického průměrování nebo nástroje pro interpolaci [NAfE5] či aproximaci polynomu, který by modeloval křivku reprezentující samotný profil šumu. Takové metody jsou k dispozici, např. Newtonův či Lagrangeův polynom nebo tzv. metoda nejmenších čtverců [NAfE6]. Pro samotné rozpoznání píku (což je v podstatě nějaká extremální hodnota) lze potom využít jak prostých statistických metod (průměr, maximum, odchylka), tak složitějších nástrojů, např. diferenciálního a integrálního počtu, kde lze např. sledovat sklon a rychlosti změny v křivce reprezentující spektrum.

Z tohoto plyne, že pro samotné odstranění šumu je možné vybírat z řady postupů a technik.

V případě odstraňování šumu ze spektra nám jako použitelná možnost vychází z uvedených jen jediná a tou je průměrování. S ohledem na to, že program má zpracovávat už změřená spektra, není možné nikde získat vzorek šumu. Univerzální profil šumu neexistuje a nelze po uživatelích požadovat, aby ke každému spektru přiložili spektrum šumu, které získali ze spektrometru proměřením „naprázdno“.

Rovněž tak není možné použít pokročilé matematické metody pro odhad šumu profilu. Ty jsou totiž vhodné převážně pro malý počet bodů a jejich výpočetní složitost většinou poměrně rychle roste s počtem interpolačních/aproximačních bodů. S ohledem na to, že počet bodů ve spektru je opravdu značný (stovky tisíc bodů), je využití těchto metod velmi problematické jak z výpočetního hlediska, tak z hlediska získaného výsledku.

Zvláště nejpoužívanější interpolační metody mají už při relativně nízkém počtu bodů schopnost generovat polynomy vysokých stupňů a z nich získané křivky se (nevědecky řečeno) hodně „vlní“ a chovají se velice nepředvídatelně.

Tedy i čistě z hlediska matematiky se nejedná o problém, který by měl jednoznačné řešení pomocí nějakého magického „vzorce na šum“. Nicméně nejjednodušší interpolační metodou je průměr a nebyl důvod ho nezkusit, už jen ze zvědavosti.

Odstranění šumu, když se ukázalo jako nezbytné, bylo tedy nejprve vyzkoušeno s prostým globálním zprůměrováním (udělá se průměrný odhad šumu z celého rozsahu spektra). Tím byla zjištěna přibližná úroveň šumu v celém spektru a z něj stanoven základ pro hranici, od kterého byla naměřená hodnota považována za pík. Nicméně tento způsob odstranění šumu fungoval pouze na spektrech s nízkým šumem a i zde se často projevila necitlivost, kdy méně intenzivní píky nebyly nalezeny.

Globální průměrování zcela selhávalo u spekter, kde byl šum proměnlivý a intenzita píků nebyla dostatečná. V případě, že šum v určité části spektra převýšil intenzitu píků v jiné části, byla správná detekce nemožná.

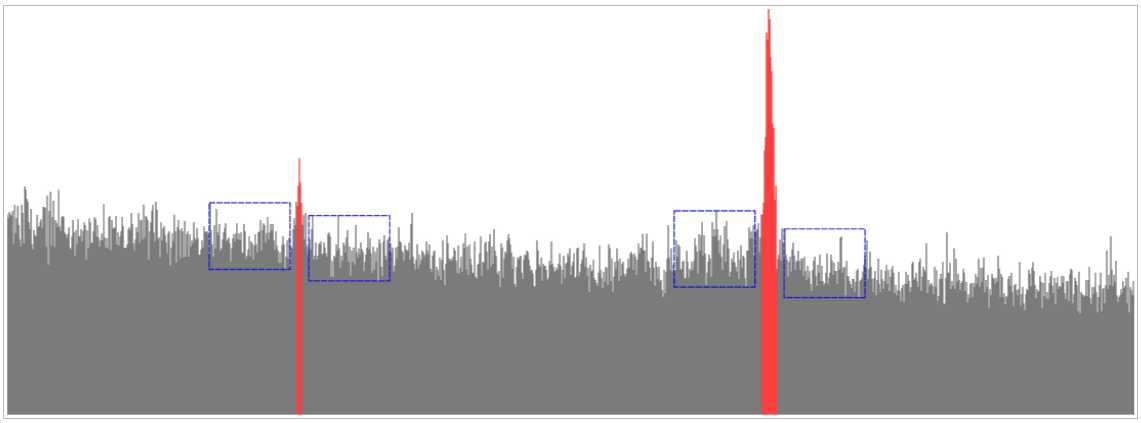
6.5.3 Lokální průměrování

S ohledem na to, že velká necitlivost na píky a chybovost při odstraněním šumu vznikala právě kvůli globálnímu průměrování celého spektra, byl učiněn pokus o obejití celého problému pomocí lokálního průměrování. A to se nakonec ukázalo jako funkční a velice úspěšné.

Výsledné řešení pro detekci píku tedy pracuje s metodou, kterou bych označil jako „kopečková metoda“. Dala by se přirovnat k situaci, kdy turista v horách chce zjistit, zda stojí na kopečku

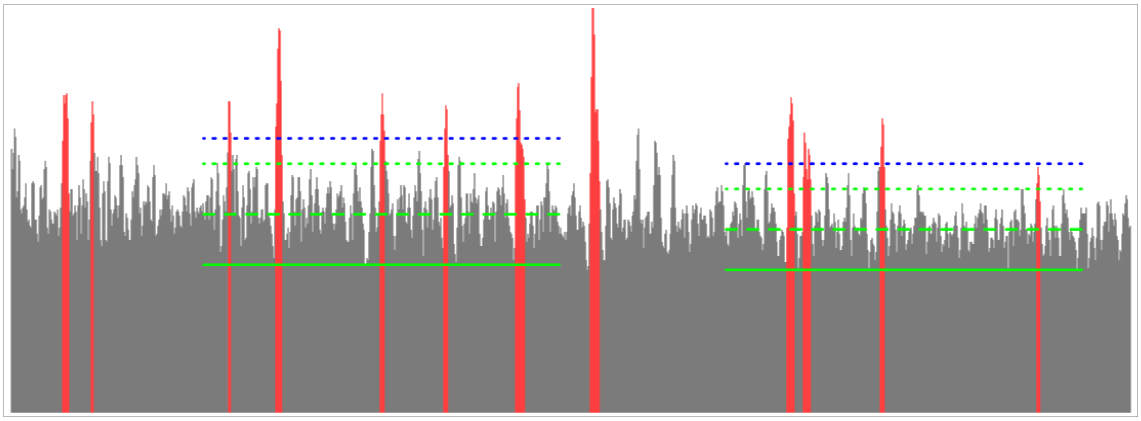
(píku). Turista se rozhlédne pouze po blízkém okolí (vzdálené kopce bude ignorovat) a pokud uvidí, že stojí výrazněji výše, než jaká je výška blízkého okolí, tak prohlásí, že stojí na kopečku (píku). A to i když v dálce bude vidět kopce mnohem vyšší.

V principu tedy algoritmus hledá lokální maxima. Hledání píku v bodě [x] probíhá tak, že se provede lokální kontrola levého a pravého okolí [Obr 24]. Při něm se stanoví nejvyšší, nejnižší a průměrná hodnota v okolí a ta se porovná s hodnotami ve zkoumaném bodě x [Obr 25]. Pokud hodnota v bodě dostatečně převyšuje průměrnou hladinu šumu (i s přihlédnutím k jeho rozptylu), je bod x prohlášen za pík.



Obr 24: Lokální průměrování - oblast hledání

Software při vyhledávání píku bere ohled pouze na blízké okolí píku. Aby nebyly průměrné hodnoty šumu ovlivňovány vysokými hodnotami píku, zahrnuje se do výpočtu minima a průměru jen okolí kolem píku (vyznačené obdélníky). Parametricky lze stanovit šířka obdélníků a odstup od testovaného píku.



Obr 25: Lokální průměrování - hladiny šumu důležité pro detekci píku

Lokálně se ze zkoumané oblasti vypočítá minimum šumu, což je relativní hodnota 0 (zelená plná čára), fakticky baseline. Dále se vypočte průměr šumu, což je relativní 1 (zelená čárkovaná). Z toho se odhadne teoretické maximum šumu (zelená tečkovaná), což odpovídá relativní hladině 2.
Mez citlivosti pro detekci píků (tj. hodnota od které je pozice považována za pík) je volitelně stanovena uživatelem/správcem. V základním nastavení je stanovena výška píku na relativní hodnotu 2,5 (modrá tečkovaná).
Důležité je si uvědomit, že detekce je založena na relativních poměrech mezi baseline a rozptylem šumu, tudíž absolutní hodnoty pro detekci píku jsou v každé oblasti spektra jiné. Hladinu detekce píků ovlivňuje např. i to, pokud je v okolí píků více. Může se stát, že seskupení více píků vedle sebe navýší průměr šumu a vyřadí tak z detekce méně intenzivní píky. Toto by bylo možné eliminovat dvoukrokovou detekcí.

Zde je na místě uvést dva důležité faktory, které rozeznání komplikují. Prvním faktorem je nepravidelná struktura šumu. Druhým potom skutečnost, že jednak právě testovaný pík nemusí být jediným píkem v intervalu, který je prohledáván a také, že v intervalu naopak žádný pík být nemusí. Výpočet musí obě tyto skutečnosti zohledňovat.

K prvnímu problému. Šum má ve sledované oblasti určitou, řekněme průměrnou, hladinu. Stejný vzorek lze změřit několikrát s různým šumem a získaná spektra sice budou zachycovat píky se stejnou intenzitou, ale s rozdílným šumem. Zatímco jedno spektrum může mít šum na tak nízké úrovni, že ho můžeme zanedbat, jiné spektrum může mít tak vysoký šum, že v určitých částech může zcela převýšit i maximální píky nacházející se v jiné části spektra.

Nelze tedy zobecnit, že skutečný pík je pouze např. dvojnásobek průměrného šumu. Takto nastavené kritérium prostě nebude fungovat. Detekovat pík můžeme jen pokud bude dostatečně vystupovat z okolního šumu. Tj. skutečný pík rozeznáme jen pokud bude výrazně silnější než šum. Pokud bude pík slabý, tak ho od šumu pravděpodobně nerozeznáme.

A kromě toho má šum ještě mnohem horší problém – určitý rozptyl kolem své průměrné hladiny. Tento rozptyl má periodický charakter, srovnatelný např. se sinusoidou. Na vizualizaci spektra se tento rozptyl projevuje „roztřesením“ nebo „zubatěním“ šumu.

Nepříjemné je, že rozptyl šumu je rozdílný nejen pro každé spektrum, ale často i pro různé úseky jednoho spektra. Některé spektrum má šum jako plynulou křivku s minimálním rozptylem a píky jsou jen malá výchylka nad hladinu šumu. V takovém případě musíme hranici nutnou pro rozpoznání píku nastavit jako velice malou. Jejím zvýšením by aplikace píky nerozpoznala. Jiné spektrum ovšem může mít rozptyl šumu velký a potom by nízká hranice způsobila, že i vrcholky rozptylu šumu by byly detekovány jako píky. Řešením je nastavit citlivost nižší, tak aby byly detekovány jen silné píky i za cenu toho, že slabé píky detekovány nebudou.

Takže hranice citlivosti proti rozptylu šumu je pro každé spektrum individuální a může být odlišná i pro určité části stejného spektra.

Způsob detekce tedy musí dynamicky reagovat jak na průměrnou hladinu šumu, tak na jeho rozptyl. Nic nelze nastavit staticky či absolutně. Hodnoty šumu a jeho rozptylu se vůči píkům musí počítat relativně a je poměrně obtížné je z hlediska statistiky přesně podchytit. Obecné řešení je poměrně jednoduché, ale jak efektivně bude řešení fungovat na skutečná spektra závisí především na nastavení konkrétních parametrů. A to je největší problém při odstranění šumu.

6.5.4 Je tedy daná pozice píkem?

Abychom rozhodli, zda pozici považovat za pík, musíme vypočíst ve zkoumaném bodě a jeho okolí nejen průměrnou hladinu šumu, ale i jeho odchylku. Tj. výsledkem je stanovení průměrné hodnoty šumu y_n (což dostaneme jako běžný průměr) a k tomu rozptyl šumu Δ (delta) nad i pod průměrnou hodnotu. Pozice, které budou výrazně nad tento rozptyl potom prohlásíme za píky.

Zůstává otázkou, jak stanovit hodnotu Δ. Rozhodně není dobré stanovit ji jako průměrnou odchylku od průměru, protože potřebujeme zachytit celý rozsah šumu. Průměr odchylky nepostihuje skutečný rozptyl šumu. Zde platí Gaussovská křivka rozložení hodnot a průměrem odchylek bychom nezískali filtr na veškerý šum. Potřebujeme tedy ve skutečnosti zjistit maximální rozptyl od průměrného šumu, ale musíme z něj vynechat píky, které jdou směrem nahoru nad maximální hladinu šumu. To znamená, že nemůžeme zvolit ani prosté maximum, které se v daném intervalu vyskytuje. píky, pokud jsou přítomny, totiž posouvají hodnotu maxima na celém intervalu směrem nahoru. Musíme najít hodnotu menší než je maximum, ale větší než je průměr odchylky.

Jednoduchou úvahou se ale dostaneme k nápadu, že když se nejde podívat nahoru na maxima, můžeme se podívat směrem dolů na minima. Protože šum je relativně periodická sinusoida, můžeme úvahou předpokládat, že maximální odchylka šumu směrem nahoru bude zhruba podobná maximální odchylce směrem dolů pod průměr šumu.

Nakonec jsem tedy po úvaze rozhodl stanovit hodnotu Δ jako rozdíl mezi y_n a y_{min}, kde y_{min} je nejnižší hodnota šumu na celém intervalu. I když jsem toto exaktně neověřil (resp. zkusil jsem si to ověřit jen prohlédnutím vzhledu spekter a výpisem číselných hodnot z databáze), jde o logicky podložený odhad, který, jak se ukázalo, v praxi funguje velice dobře.

Takto stanovený rozptyl nám tedy rozsah šumu ohraničí do intervalu $\langle y_n - \Delta, y_n + \Delta \rangle$ a cokoliv nad tento interval bude potenciálním kandidátem na pík.

Detekce píku je potom velice snadná. Kritériem pro zařazení mezi píky jsou hodnoty, které překročí prahovou hranici směrem nad hodnotu $y_n + \Delta$. V praxi používám postup, kdy se hodnota „ořízne“ o dolní hladinu šumu y_{min} (minimum šum) a tato hodnota se znormalizuje vůči průměrnému šumu y_n.

Matematicky relativní výšku píku p_{rel} vůči šumu y_n získáme vztahem $p_{rel} = (h - y_{min}) / (y_n - y_{min})$ kde *h* je absolutní výška píku. Minimální možná relativní hodnota p_{rel} je tedy 0 (což odpovídá baseline) a může teoreticky být do nekonečna, průměrný šum y_n má hodnotu 1, maximum šumu hodnotu 2.

Vše nad hodnotu 2 pak lze teoreticky považovat za pík. Nicméně pro zvýšení jistoty pozitivní detekce píku jsem se rozhodl nastavit určitý „odstup“ od šumu a používám prahovou hodnotu pro pozitivní detekci píku na hodnotě zhruba kolem 2,5. To z toho důvodu, že vypočtený průměr střední hladiny šumu je jen přesný průměr nepřesných čísel a nemusí být optimální, pokud spektrum bude velice nerovnoměrné. Hodnota 2, která by teoreticky měla být postačující, při pokusech dávala řadu falešných píků, které ve skutečnosti píkem být neměly.

Řešením bylo určité snížení citlivosti, tj. zvýšení prahové hodnoty. Pokud je totiž citlivost větší než 3, ukázaly pokusy, že je už téměř jisté, že jde o pík. Nicméně u některých spekter, zvláště u těch s vysokým rozptylem šumu, se už hodnota 3 ukázala jako příliš necitlivá a některé píky rozpoznány nebyly. Proto jsem se postupně uchýlil ke kompromisní hodnotě kolem 2,5, která dávala subjektivně nejlepší poměr správných a nesprávných detekcí.

Ještě se musíme pozastavit nad možností, že ve zkoumaném intervalu žádný pík není, nebo naopak je tam píků více. Zde je na místě uvést, že se musíme vyvarovat svádivé touze hledat píky jen podle maxima a minima. Použití metody „jen maximum“ by nevedlo k dobrým výsledkům. Testovaná souřadnice x totiž může být píkem, ale nemusí nutně být maximem na testovaném intervalu, pokud tam je píků více. A současně i když je souřadnice x maximem v daném intervalu, nemusí být píkem, pokud není dostatečně intenzivní s odstupem od šumu.

A tady je štěstím, že zvolená metoda na odstranění šumu a rozpoznání píku tento problém řeší zcela elegantně. Na intervalu může být současně více x-ových souřadnic, které mají relativní hodnotu vůči šumu větší než je zvolená prahová hodnota (2,5).

Pro sledovaný pík se vždy počítá s maximální hodnotou sledované x-ové souřadnice, nikoliv celého intervalu a ostatní maxima (píky) v intervalu se vůbec neberou v potaz. A odchylka šumu delta se počítá z minima intervalu. Současně, pokud se všechny hodnoty sledovaného intervalu dostanou do rozsahu $\langle 0; 2{,}5 \rangle$, nebude žádná hodnota detekována jako pík a interval tak bude zcela bez píků.

6.6 Způsob porovnávání dvou spekter

Než vysvětlím způsob jak posoudit podobnost dvou spekter, je třeba si objasnit, jak počítače vlastně srovnávají objekty [GRV], [Kanade 1977]. Počítačové metody [Štarnclová 2012] pro hledání podobnosti totiž nepracují tak, jak vnímá lidský mozek [Ledvoň 2009]. Lidský mozek hodnotí porovnávaný objekt jako celek a teprve poté přechází k detailům (pokud k tomu vůbec dojde). Navíc zapojuje i zkušenosti a odhad. Důkazem je např. periferní vidění, nebo různé populární zábavné obrázky, kdy obrazce vytváří optické klamy. Takové řešení ale není u počítače technicky možné.

Počítačové porovnávání pracuje tak, že se snaží nalézt shodu ve vybraných znacích, které charakterizují daný objekt. Přitom porovnání celku se ve skutečnosti realizuje pomocí porovnání více dílčích znaků. Jeden znak nám prohledávanou skupinu možných řešení rozdělí nejméně na dvě (většinou různě velké) části. To ale nemusí stačit, ve skutečnosti to vůbec nestačí. Pokud ale zkombinujeme podmínek více, tak se nám množina správných výsledků velice rychle zmenší. Podmínkou pro efektivní porovnání podobnosti je přitom volba správných znaků a také jejich reprezentace.

6.6.1 Příklad z praxe pro objasnění problematiky porovnání

Uveďme si jednoduchý příklad. Pokud bychom si chtěli simulovat rozpoznávání osoby, jak bychom postupovali? Můžeme rozeznávat osoby podle pohlaví, věku, výšky, váhy, barvy vlasů, očí... Jsou tyto znaky postačující?

Samy o sobě ne. Pokud budeme rozlišovat osoby dle pohlaví, zařadíme je do příslušné skupiny poměrně snadno, ale vždy dostaneme jako výsledek hledání polovinu populace. To pro jednoznačnou identifikaci není moc dobré.

Tak o zkusme znovu. Co použít identifikaci podle barvy vlasů? Tady se nám nabízí rozdělení na černá, hnědá, blond... a otázka je jak popsat různé odstíny? Jak rozlišíme hodně tmavou hnědou od černé? Kde končí hnědá a začíná blonďatá? A co třeba barva očí? Modrá, zelená, hnědá, černá... máme lidi s modrozelenýma očima a některé, kteří mají každé oko v jiné barvě.

Podobné dilema nastane v případě výšky. Pokud si uvědomíme, že průměrná výška dospělého člověka je někde kolem 175 cm, jak budeme hledat člověka, pokud nám svědek zločinu popíše, že pachatel byl asi 180 cm vysoký? Máme hledat mezi lidmi, kteří mají 175 - 180 cm? Jaký je ten „správný“ rozsah? Změřit výšku osoby přesně na centimetr není dost dobře možné (stačí

napnout svaly nebo mírně ohnout páteř) a přesně ji určit pouhým okem je prakticky nemožné. Totéž platí při určení barvy očí, vlasů atd...

Naproti tomu, třeba u věku, dokážeme skutečnou hodnotu u dané osoby určit zcela přesně dle data narození, ale často ji nejsme schopni přesně určit pohledem. Zatímco věk např. u dítěte určíte poměrně přesně, u dospělých je to už složitější. A věk je navíc zajímavý tím, že jde o kritérium, které nám nerozdělí prohledávanou skupinu rovnoměrně. Kolik osob bude vyhledáno ve věku 20, 50 a 100 let? Zatímco dvacetiletých i padesátiletých bude hodně, stoletých najdeme v populaci jen velmi málo.

Na těchto příkladech ukazuji, že některá porovnávací kritéria nemusí být zcela jednoznačná. Takže některé znaky dokonce nemůžeme ani přesně zařadit způsobem „ano-ne“ a při vyhledávání podobností je třeba je „zaškatulkovat“ a porovnávat v nějakém rozsahu tolerance „od-do“, či „spíše ano-spíše ne “. Tímto se zabývá tzv. fuzzy logika.

A k naší smůle dodejme, že ani věkem, ani uvedenou výškou a ani barvou očí stále nejsme schopni identifikovat konkrétní osobu zcela jednoznačně.

Abychom mohli pomocí počítače identifikovat konkrétní osobu na základě jejího popisu, musíme provést postupnou selekci podezřelých osob podle více kritérií. Tj. rozdělením dle pohlaví zmenšíme počet kandidátů na polovinu. Z této poloviny poté můžeme vybrat např. jen ty, kteří mají podprůměrnou výšku, zelené oči a černé vlasy. Každé z těchto kritérií nám dále zmenší skupinu těch, kteří odpovídají popisu.

Pokud bychom pokračovali dále, tak např. biometrické systémy porovnávají přibližné proporce lidské tváře – polohu očí, úst, délku a šířku nosu atd... Každé z těchto kritérií je založeno na nalezení určitých bodů a prvků v lidské tváři, které počítač dokáže identifikovat. Vždyť něco podobného umí i každý slušný fotoaparát s režimem „portrét“ – automaticky identifikuje polohu obličeje na fotografii a zaostřuje právě na obličej.

Přitom platí, že čím více znaků se porovnává, tím menší a přesnější skupinu výsledků dostaneme. Zde samozřejmě musíme zmínit ještě jednu klíčovou věc – existuje riziko, že můžeme správný výsledek vyřadit tím, že nesprávně identifikujeme jeden znak. Co by se stalo, kdyby si hledaná osoba obarvila vlasy? Měl by náš systém být schopen identifikovat, že tato osoba je potenciální kandidát i když neodpovídá v jednom znaku? Jak to udělat?

Na tohle existuje jen jediná odpověď – aby byl systém funkční, musí srovnávat více znaků, než je nutné minimum a pro spolehlivou pozitivní identifikaci musí stačit splnění dostatečného

počtu znaků. Přitom kombinace porovnávaných znaků musí být vícenásobně průkazná (redundantní), aby bylo možné osobu identifikovat i při chybné detekci některého znaku.

Kritéria porovnání a hodnoty, kterých mohou kritéria nabývat také musí být zvolena správně vzhledem k prohledávané množině. Každé kritérium musí co nejvíce snížit počet potenciálních kandidátů. Například by nemělo smysl hledat osobu pomocí popisu „je menší než 150 cm“ pokud bychom hledali v mateřské školce, protože tam mají všechny děti méně než 150 cm.

Například pokud máme 6 miliard osob, tak musíme mít tolik dostatečně průkazných kritérií a jejich kombinací, aby bylo možné identifikovat konkrétní osobu. Ve skutečnosti nám musí porovnávaná kritéria vytvořit mnohem více kombinací, než 6 miliard. Už jen proto, aby bylo možné eliminovat výše uvedené nepřesné detekce.

Zde se musíme opět podívat na statistiku. Pokud jedno kritérium vytvoří dvě možnosti, spojení dvou takových kritérií vytvoří čtyři kombinace. Pokud bychom měli kritérium se třemi možnostmi, kombinací dvou takových kritérií už dostaneme devět variant. Pokud takových správně zvolených kritérií spojíme např. třicet, dostaneme množství kombinací přesahující počet žijících osob na této planetě a pokud budou kritéria správně zvolena (budou efektivně dělit prohledávanou skupinu na podskupiny), nebude problém danou osobu identifikovat.

Příkladem takové identifikace je daktyloskopie [Fürbach 2008], která je pomocí zhruba deseti až patnácti bodů (kritérií) schopna identifikovat s prakticky absolutní přesností jakéhokoliv člověka. A to jen díky tomu, že množství kombinací papilárních linií na prstech je natolik vysoké a vytváří tolik kombinací, že nesprávná identifikace je téměř vyloučená.

6.6.2 Jak uvedené aplikovat na porovnání hmotnostních spekter

A teď k otázce, jak výše uvedené aplikovat na porovnání hmotnostních spekter. V první řadě si musíme uvědomit, co jsou znaky charakteristické pro dané spektrum. Každé spektrum obsahuje píky a šum. Šum nás nemusí trápit, pro porovnání nenese žádnou informaci a je eliminován při vyhledávání píků. Postup k nalezení píků jsme si už uvedli výše.

Zůstanou nám píky. Ty totiž skutečně popisují, co spektrum obsahuje. píky mají dvě vlastnosti – polohu odpovídající *m/z* a intenzitu. Dalo by se uvažovat i o šířce píku, pokud bychom pík chápali jako objekt, ale široký pík lze chápat také jako několik úzkých píků těsně vedle sebe. Biospean pracuje s širokými píky právě tímto způsobem, protože je takové řešení pro počítač jednodušší.

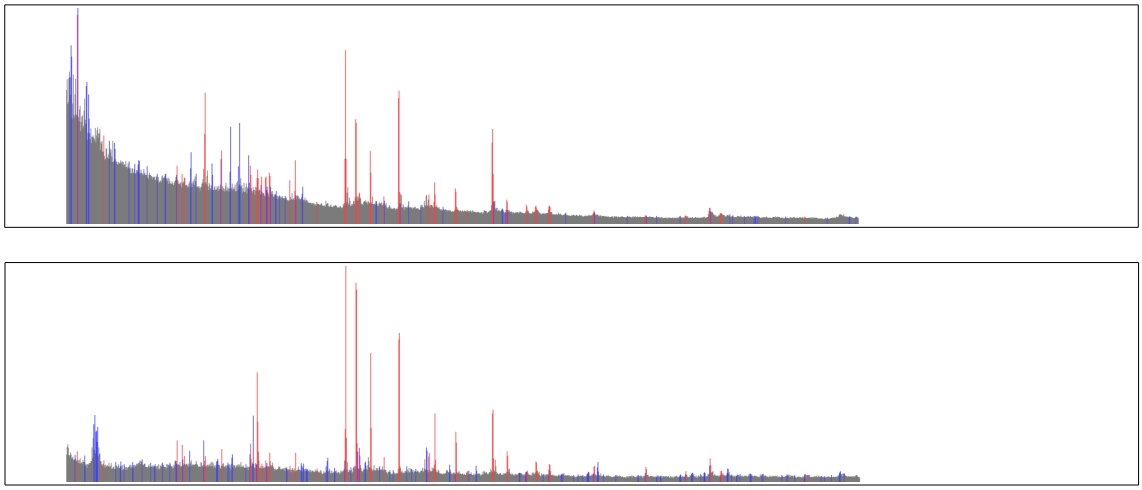
Samotná **poloha píků** charakterizuje celkem jednoznačně každý druh/skupinu vzorků, resp. jejich složení. Tj. každá *m/z* pozice je samostatným znakem, který posuzujeme pouze jako boolovskou hodnotu (ano/ne – buď to je pík nebo není). To je rozdíl proti příkladu, kde jsme u osob měli znak výšku, znak barvu vlasů atd. a hodnoty byla čísla nebo určené druhy (barva). V případě spektra máme přímo vlastnost „*je na pozici XY pík?*“, možnosti jsou ano či ne a těchto znaků je stejný počet jako je celý rozsah sledovaného spektra, tj. v případě Biospeanu až 22 tisíc znaků, které jsou na sobě fakticky nezávislé. Už jen tento počet nám dává dostatečný předpoklad, že by možný počet kombinací znaků měl být dostatečný aby zajistil identifikaci.

Pokud si uvědomíme, že většina spekter má efektivní (při výzkumu sledovaný) rozsah někde mezi 2 000 - 15 000 *m/z* a v tomto rozsahu identifikujeme řádově stovky píků (většinou více než 200 pozic), máme dostatečně velkou kombinaci *m/z* pozic, které nám dávají charakteristický „podpis“ (píklist) daného spektra. Teoreticky je na rozsahu 10 000 *m/z* možné vytvořit 2^{10000} kombinací, což je ovšem pro praxi nesmyslné a těžko představitelné číslo – $1{,}995 \cdot 10^{3010}$ kombinací (dvojka a za ní tři tisíce nul).

Nás v praxi mnohem více zajímá, kolik kombinací lze vytvořit např z 10, 50, 100 nebo 200 píky. Pokud budeme mít jen jeden pík, máme na rozsahu 10 000 *m/z* celkem 10 000 možností kam tento pík umístit. Při dvou pících to bude $(10\,000 \cdot 9\,999)$ kombinací, při třech už $(10\,000 \cdot 9\,999 \cdot 9\,998)$.

Ano, jde o skutečně vysoký počet kombinací pro onen hledaný píklist, který umožňuje spektrum jednoznačně identifikovat. Pokud připustíme určité zjednodušení výše uvedeného vztahu, bavíme se při deseti pících o téměř $10\,000^{10}$ kombinacích, což je $(10^4)^{10}$ neboli 10^{40} možností. A to pouze pro deset píků. U počtu 100 píků pak lze vytvořit řádově $10\,000^{100}$ kombinací. To je 10^{400}, což je jednička a za ní 400 nul, což je hodnota skutečně astronomického rozsahu.

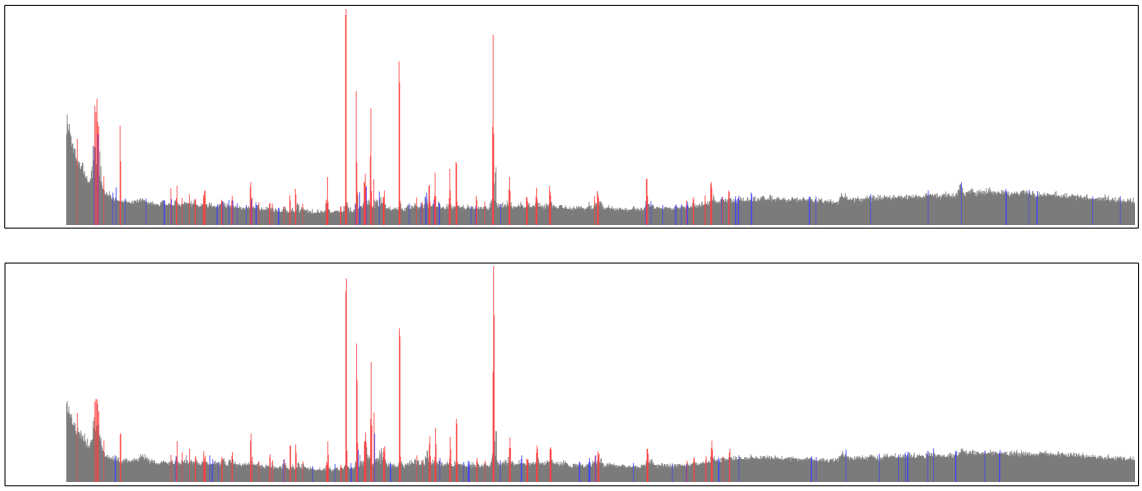
Takže pokud dvě různá spektra mají pozice píků víceméně shodná [Obr 26, Obr 27], můžeme směle prohlásit, že jsou si nějakým způsobem (více či méně) podobná. Hledání podobných spekter potom probíhá tak, že vezmeme charakteristický píklist z testovaného vzorku a v databázi se pokusíme najít spektra, která mají píklist co nejvíce podobný, tj. hledáme spektra s co největší shodou na *m/z* pozicích.



Obr 26: Spektra se shodnými pozicemi píků a velmi odlišným šumem

Zde je vidět, že obě spektra mají velké množství společných píků (červeně). Šum obsažený v prvním spektru sice komplikuje detekci, ale shoda je dostatečná pro identifikaci podobností. Některé píky jsou rozdílné (modře), ale jejich přítomnost nezabránila nalézt shodu spekter, byť s menší procentuální podobností.

Přitom si můžeme dovolit dostatečnou toleranci při identifikaci píků, protože počet kombinací je tak vysoký, že při stovkách píků (200 – 500) určitý počet nesprávně identifikovaných píků sice sníží procentuální podobnost, ale toto snížení, jak se při praktických pokusech ukázalo, není většinou natolik kritické, aby znemožnilo identifikaci.



Obr 27: Spektra se stejnou pozicí píků, které ovšem mají rozdílné intenzity

Na těchto velice podobných spektrech můžeme vidět, že zohledňovat intenzitu píků může být zrádné. Ačkoliv shoda na pozicích píků (červeně) je vysoká, neshod je jen málo (modré píky) a i samotný profil šumu je velice podobný, tak intenzita píků na stejných pozicích je často odlišná a nelze ji brát jako primární faktor pro porovnání.

Přitom i shoda na menším počtu pozic je postačující pro jednoznačnou identifikaci. To bylo také na Biospeanu ověřeno, že v praxi shoda na 100 % nastane jen při vložení stejného datového souboru dvakrát. Ani spektra pocházející z různých dvou měření stejného vzorku neukáží 100% shodu. To ale vůbec nevadí, protože praktické pokusy ukázaly, že většinou pro správnou identifikaci druhu spektra stačí shoda jen na 30-70 % pozic. Plnou 100% shodu, ale jen

jednostrannou, vykazují pouze virtuální spektra (viz. dále), což je logický důsledek toho, jak se virtuální spektrum tvoří.

Otázkou je, zda se může náhodně vyskytnout kombinace píků (a šumu nesprávně identifikovaného jako píky) i u nepříbuzných spekter tak, aby došlo k vysokému procentu shody. Odpověď je, že čistě statisticky je to dosti nepravděpodobné. Jednalo by se o situaci, kdy by se v jednom spektru nacházelo například 200 píků, v druhém třeba 250 a tyto píky by se nacházely zcela náhodně na rozsahu řekněme 15 000 jednotek *m/z*. Pro 20%, resp. 25% podobnost by musela nastat shoda alespoň na 50ti pících. Pro jednoznačnou identifikaci ovšem musíme počítat s větší podobností, alespoň 30-50%. Pravděpodobnost, že taková shoda nastane opravdu náhodně je poměrně nízká, počet všech možných kombinací je dosti vysoký.

Nicméně je zde ještě biochemické hledisko. A to je faktor, který nelze jednoznačně popsat. Určité látky a molekuly se vyskytují s různou pravděpodobností. Některé látky jsou prostě statisticky častější než jiné. Tudíž určité *m/z* pozice se mohou vyskytovat statisticky častěji a u více vzorků a zvyšovat tak shodu u nepříbuzných spekter. Technické řešení je možné – přidělit *m/z* pozicím tzv. váhy (koeficient důležitosti) a zvýhodňovat píky na méně častých pozicích. Nemám ovšem žádné statistické informace o četnosti konkrétních *m/z* pozic, na základě kterých bych takové hodnocení mohl provádět.

V praktických testech se ukázalo, že nesprávná identifikace píků (především z důvodu šumu) sice do jisté míry ovlivňuje výsledek porovnání, ale nejedná se natolik výrazné zhoršení, aby chyba zcela vyloučila správnou identifikaci (tj. aby došlo k nesprávné negativní identifikaci). Důležitější je, že se mi nepovedlo zaznamenat případ, že by bylo některé spektrum identifikováno nesprávně jako shodné (tj. nesprávná pozitivní identifikace). Zaznamenaná shoda u zcela nepříbuzných spekter byla nejvýše v řádu jednotek procent, což pro pozitivní identifikaci nestačí.

Poloha (*m/z*) píků je tedy na identifikaci spekter velice vhodné kritérium. Je dostatečně robustní na chyby a i při relativně vysokém počtu nesprávně identifikovaných píků je dostatečně spolehlivá, aby identifikovala příbuzná spektra.

Druhá vlastnost píků, intenzita, udává v jakém množství je daná složka zastoupena. Samotná intenzita přitom není pro srovnání příliš vhodná, protože intenzitu píků ovlivňuje například koncentrace látek ve vzorku, množství šumu a další faktory (např. v různých fázích reakce se jednotlivé složky mohou vyskytovat v různých koncentracích). Další složitou otázkou v případě

intenzity je, zda porovnávat absolutní intenzitu (závislou na skutečné koncentraci) nebo relativní intenzitu (závislou na okolním šumu).

Jinými slovy, zatímco pozice píků je (při stejném nastavení spektrometru a použité matrici) celkem jednoznačná a dostatečně vypovídající, porovnávat intenzitu už může být záludné. Proto jsem se po počátečních pokusech rozhodl intenzitu píků zcela ignorovat a porovnávat pouze na základě polohy píků.

6.6.3 Technické řešení výpočtu podobnosti

Pokud chceme porovnat, zda se sobě dvě spektra podobají, jsme odkázáni na srovnání pozic píků. Tj. srovnáváme, zda pozice píků z jednoho spektra se shodují s pozicí píků spektra druhého, resp. zda má první spektrum na pozicích svých píků odpovídající protějšek ve druhém spektru.

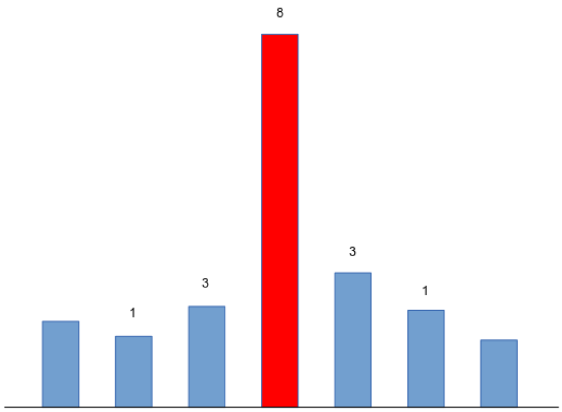
Srovnání je přitom jednosměrné, resp. srovnávat můžeme vždy jedno spektrum vůči druhému. Například pokud spektrum A má 152 píků a spektrum B má 178 píků a z toho 125 píků je společných, tak spektrum A má na 82% (125 ze 152) shodu se spektrem B a naopak spektrum B má v 70% (125 ze 178) shodu s píky ze spektra A.

Výsledkem porovnání spektra A a B jsou tedy dvě čísla, která udávají vzájemnou procentuální shodu obou spekter. Jedno z čísel udává shodu A ku B a druhé B ku A. V souvislosti s tím Biospean nabízí několik možností řazení výsledků, počínaje prostým průměrem přes použijí jen jedné z hodnot či výběr menší/větší z nich až po poměr mezi těmito hodnotami a experimentální koeficient, který současně zohledňuje průměr a poměr mezi hodnotami.

Porovnávání spekter by také mělo být schopné kompenzovat jistou odchylku v pozicích píků. Například program Biotyper je schopen rozpoznat posun o jednu až dvě pozice *m/z*. Minimálně jedna jednotka *m/z* je absolutní minimum, protože Biospean pracuje v případě pozice s celými zaokrouhlovanými čísly a při měření a následném přepočtu není problém, vinou zaokrouhlování a zobecňování, pík posunout o jednu pozici vedle. Pokud například bude mít pík maximum na pozici 4356.6, snadno se při vyhodnocování dostane na pozici 4356 a/nebo 4357. Případně dojde k detekci dvou píků na každé z pozic. Program proto musí mít nastavitelnou toleranci a vyhodnocovací algoritmus se musí s touto tolerancí dokázat vypořádat.

Řešení tohoto problému se ukázalo jako relativně snadné – přidělit píkům body. A to tím způsobem, že skutečný pík dostane určitý počet bodů a jeho sousední souřadnice, která už detekovaným píkem není, o něco menší počet bodů [Obr 28]. Pokud bude více detekovaných

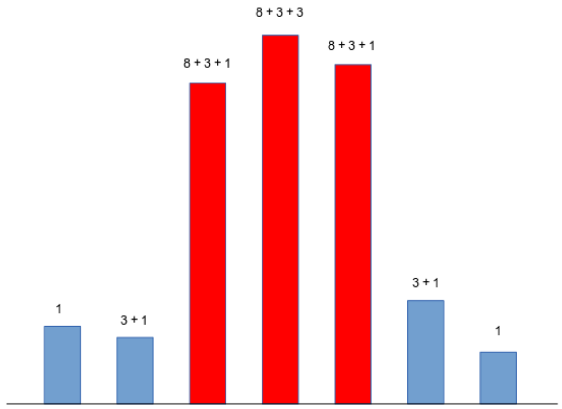
píků vedle sebe, může pík dostat body jak za ohodnocení sebe sama, tak za ohodnocení souseda [Obr 29]. Díky tomu více píků (ve skutečnosti jeden dostatečně široký pík) vedle sebe způsobí, že píky v takovémto seskupení budou mít více bodů, než běžné píky bez sousedů.



Obr 28: Bodové ohodnocení osamoceného píku

Pozice detekovaná jako pík je ohodnocena 8-mi body. Těsně sousedící pozice obdrží 3 body. O jednu pozici vzdálenější pozice obdrží 1 bod.

Ohodnocování sousedních píků menším bodovým ziskem způsobí, že pokud se při porovnání dvou spekter netrefíme přesně na plnohodnotný pík, setkáme se alespoň s píkem, který má nižší ohodnocení a budeme mít alespoň částečnou shodu.



Obr 29: Bodové ohodnocení sousedících píků

Pokud se vedle sebe nachází více píků, hodnocení sousedních píků se sčítají. Prostřední pík dostal svých 8 bodů a od každého souseda další 3 body. Krajní píky dostaly svých 8 bodů, k tomu 3 body od prostředního a 1 bod od protilehlého píku. Pozice ležící těsně vedle píků dostaly 1 bod od prostředního a 3 body od krajního píku. O jednu pozici vzdálenější pozice dostaly 1 bod od krajních píků.

Další výhodou bodového ohodnocování je, že můžeme, podle toho jak nastavíme bodování při vyhodnocení spektra, ovlivnit toleranci posunu. Dále se pomocí bodování dá nastavit i tolerance už zpracovaného spektra a to tak, že do srovnání zahrneme jen píky s určitým počtem bodů

(např. více než 4 body). Tím jsme dokonce schopni omezit vyhledávání např. jen na píky, které mají šířku větší než dvě/tři x-ové pozice.

Fakticky můžeme porovnávat spektra různými způsoby – buď poměrem shodných píků (jak to Biospean počítá), anebo by bylo také možné počítat výsledek jako poměr bodovaných píků na jednotlivých pozicích. Případně zvýhodňovat při výpočtu shodu píků se stejným hodnocením a píky s rozdílným bodovým ohodnocení započítávat méně významným způsobem (jedná se o použití vah). Je tedy zřejmé, že by bylo možné vytvořit více způsobů jak podobnost počítat.

Bodové hodnocení píků navíc také poskytuje příjemný bonus – umožňuje uživatelskou korekci. Pokud program automaticky nedokáže rozeznat všechny píky, nebo naopak některé píky označí nesprávně, může uživatel využít nástrojů pro korekci bodového hodnocení a přidělit konkrétním píkům body, kladné nebo záporné, kterými buď označí nerozpoznaný pík, nebo naopak odebere pík, který systém označil nesprávně. Tím na dané pozici dokáže vytvořit nový pík, nebo ho odstarnit.

Bodování navíc umožňuje udělat prozatím neralizovanou modifikovanou variantu porovnávání – srovnání nejen na základě pozic, ale i ohodnocení. Samotné porovnání bodovaných píků by potom probíhalo podobně jako při s rovnání x-ových pozic. Rozdíl by byl v tom, že ve skutečnosti by se neporovnávala jen pozice píků, ale započítával by se poměr v bodovém ohodnocení na každé souřadnici souřadnici. A opět obousměrně, první spektrum s druhým a naopak. Tato varianta srovnání prozatím není realizovaná, ale její přidání není problémem, protože klíčové prvky pro takového srovnání už program obsahuje.

Nakonec zůstává otázka, jak efektivně a účinně funguje porovnávání spekter. Praktické testy ukázaly, že srovnání na základě pozice píků poměrně přesně odpovídá shodě mezi dvěma spektry. Kvalita výsledku samozřejmě závisí na kvalitě rozeznání píků, což je logické. Proto je žádoucí, aby uživatelé věnovali pozornost kvalitě spekter a jejich případnému dodatečnému nastavení, pokud automatická identifikace není z nějakého důvodu ideální.

Pro spolehlivou identifikaci přitom není nutné dosahovat stoprocentní shody. Při vývoji a následném dokončování jsem měl k dispozici zhruba 1 400 různých spekter a aplikace byla vždy schopna i při plně automatické detekci píků najít skupinu nejpodobnějších spekter. Program se přitom dokázal vypořádat i s nesprávně detekovanými píky, resp. velký počet identifikovaných píků zajišťoval, že podobná spektra byla nalezena i v ne zcela přesných datech. Podobnost spekter stejného druhu se pohybovala od zhruba 20 % až vysoko nad 90 %.

Podobnost pod 25 % občas zahrnula i nepříbuzné spektrum. Hodnoty nad 40 % naopak vykazovaly velmi vysoký stupeň podobnosti.

Stoprocentní shoda ve skutečnosti není s reálnými spektry ani možná. Vždy je někde nějaká menší odchylka, špatně rozeznaný pík nebo jen jeho jiná šířka. Pokud nastala shoda 100 %, jednalo se buď o stejný dvakrát (duplicitně) vložený vstupní soubor, nebo šlo o jednostrannou shodu s virtuálním spektrem, což je teoretické uměle vytvořené spektrum určené pro detekci konkrétních znaků (viz. část o praktickém nasazení).

Zajímavé je, že se podařilo při testech ověřit, že program opravdu funguje jak má. To se stalo ve chvíli, když jsem dostal možnost porovnat vzorky monitorující v čase probíhající fáze reakce. Když jsem vytvořil tabulky (celkem jich bylo šest pro různé druhy fermentací) srovnávající podobnost mezi jednotlivými časovými stavy, tak se ukázalo, že vzorky jsou nejvíce podobné právě s nejbližšími stavy. Jak rostla doba mezi časy jednotlivých měření, podobnost postupně klesala. O tomto detailně pojednávám v části věnované praktickému použití Biospeanu.

6.7 Postup algoritmu pro detekci píků

Před tím, než může uživatel spektrum používat, je nutné předzpracování dat. To probíhá v několika krocích a trvá zhruba minutu. Kroky jsou následující:

1. **Vytvoření nového spektra a načtení datového souboru**

   Uživatel pomocí webového rozhraní vloží nové spektrum do systému a aplikace automaticky nahraje soubor se vstupními daty. Systém je schopen pracovat se soubory ve formátu TXT a CSV. Velikost souborů je omezena na 10MB, nicméně je možné ji zvýšit. V databázi je nové spektrum uloženo jako nový záznam a uživatel je o načtení informován.

   Další fáze – samotná analýza – už proběhnou nezávisle. V aktuální verzi se automaticky po uložení spektra do databáze spustí na pozadí vyhodnocovací skript, ale systém lze snadno upravit tak, aby ponechával zpracování spekter na pozdější dobu a zpracování se spustilo jiným podnětem (například pomocí cronu[11]).

11 Zvláštní proces v Unixových operačních systémech, který v počítači běží na pozadí (bez vědomí uživatele) a stará se o pravidelné spouštění určitých procesů a úloh v nastaveném čase, např. v určitou hodinu, v určitý den týdne nebo pravidelně např. po pěti minutách.

2. Načtení zdrojových dat do databáze

Z vloženého souboru se veškerá data načtou do databáze, což umožní jejich další zpracování. Při tomto kroku se také provede filtrace vstupních hodnot a přeskočí se vše, co se netýká samotných dat (např. komentáře).

3. Normalizace dat

Při samotné analýze se nejprve provede normalizace dat. Běžná data obsahují údaje s různou hustotou bodů. Body jsou ve tvaru [x,y], kde x je *m/z* a hodnota y je intenzita. Problém je v tom, že x není celé číslo a kvůli různé hustotě můžeme na intervalu <x1,x2> nalézt u dvou spekter různý počet hodnot.

Proto aplikace nejprve provede normalizaci dat dle osy x (*m/z*), kde vypočte pro každou celou hodnotu x maximum, minimum a průměr tak, že sloučí vstupní hodnoty kolem x. Tím dostaneme data ve tvary $[x_{norm}, y_{min}, y_{max}, y_{avg}]$, kde x_{norm} je celočíselná souřadnice *m/z* a hodnoty y_{min}, y_{max} a y_{avg} jsou minimum, maximum a průměr všech hodnot z původního vstupu, jejichž absolutní hodnota x je rovna normované x_{norm}. Jinými slovy do souhrnu pro hodnotu 10 budou zahrnuty všechny hodnoty v intervalu ⟨9,5; 10,5).

4. Výpočet intervalů

Ve chvíli, kdy máme sloučené x-ové souřadnice, je dalším krokem výpočet intervalů, které budou použity pro rozpoznávání píků. Postupně se vytváří pětice ve tvaru $[x_1, x_2, y_{min}, y_{max}, y_{avg}]$ kde x_1 a x_2 jsou meze intervalu a y_{min}, y_{max}, y_{avg} jsou analogicky minimum, maximum a průměr na celém intervalu. Startovacím stavem jsou intervaly délky jedna, tj. takové, kde $x_1 = x_2$.

Z důvodu složitosti (viz. dále) je používáno binární slučování, kdy se vytváří intervaly, jejichž délky jsou druhou mocninou čísla 2. Začíná se tedy vytvořením intervalů délky 1 (tj. 2^0), spojením sousedních se v druhém kroku vytvoří intervaly délky 2 (tj. 2^1), poté 4 (tj. 2^2), 8 (tj. 2^3) atd. To se provádí tak dlouho, dokud nejsou vypočteny intervaly potřebné délky.

5. Výpočet koeficientů

A nyní se už můžeme pustit do testování jednotlivých x-ových souřadnic (*m/z*), zda obsahují žádaný pík. Pokud chceme pro souřadnici x_{norm} a příslušný pík $[x_{norm}, y_{min}, y_{max}, y_{avg}]$ zjistit relativní výšku píku vůči šumu, vezmeme intervaly příslušné

délky, které sousedí se souřadnicí x_{norm} zleva a zprava a vypočteme jejich společné minimum a průměr. Ten porovnáme s maximem kontrolovaného píku (dle výše uvedeného vzorce v části věnované šumu) a zjistíme, zda se jedná o hledaný pík nebo jen o šum.

Samotný volba píků je v aplikaci parametricky nastavitelná, takže lze kromě délky intervalů například nastavit i to, aby nebyly brány intervaly přímo sousedící s píkem, ale intervaly, které jsou o několik pozic vzdálenější. Tím se dosáhne toho, že výsledek není zkreslen „náběhem“ intenzity u paty píku.

Různé volby délky intervalů a odstupu spolu s nastavením citlivosti na hranici detekování píku umožňuje připravit několik vyhledávacích kritérií, přičemž každé může být nastaveno na jiný typ spektra. Například jedno nastavení může být vhodné pro vysoký šum, jiné naopak pro široké nebo úzké píky. Díky parametrické nastavitelnosti je možné stejné spektrum vyhodnotit několika různými způsoby nastavení a z jejich spojených výsledků vytvořit společný výsledek.

6. **Bodové ohodnocení píků**

   Posledním krokem je rozpoznání a bodové ohodnocení píků. V této části dojde ke kontrole všech píků a ty, které jsou identifikovány jako píky, jsou ohodnoceny body. Stejně tak jsou body přiděleny i píkům sousedícím se skutečnými píky (uvedeno v části o toleranci pro posuv píku). U shluků píků dojde k nasčítání bodů od sousedních píků.

Takto připravená množina píků je již připravena na vyhledávání shod s ostatními píky tak jak bylo popsáno výše.

6.8 Náročnost výpočtu

Zásadním problémem vyhodnocování píků je bohužel náročnost výpočtů. Problém není ani tak v jejich složitosti, ale v objemu dat. Pojďme chvíli počítat.

Vstupní soubor s daty ve formátu TXT nebo CSV obsahuje řádově kolem 100 tisíc *m/z* souřadnic, velikost takového souboru je v řádech jednotek megabyte. Tyto data při vizualizaci představují polohu na ose x (*m/z*) a k nim odpovídající intenzitu. Biospean je nastaven tak, aby byl schopen pracovat v rozsahu od 0 do 22 000 *m/z*. Tento rozsah je dostačující pro všechna spektra, která byla při vývoji použita k testům. Většinou rozsah začíná někde u hodnoty 1000 *m/z* a končí mezi 15 - 18 tisíci *m/z*. Pro výzkumné účely jsou přitom nejčastěji využívány

rozsahy zhruba 3 000 - 10 000 *m/z*. Pro zjednodušení výpočtu tedy předpokládejme že máme u běžného spektra rozsah zhruba 15 000 pozic (což reprezentuje rozsah např. 1 000 - 16 000 *m/z*).

V prvním kroku výše uvedeného aplgoritmu se data načtou do serveru a založí se nový záznam do databáze. Jedná se o operace víceméně na úrovni operačního systému (přesun dat po síti, založení souboru a logovacího záznamu), takže samotný Biospean v tu chvíli moc víc dělat nemůže.

Při zpracování musí Biospean v druhém kroku načíst zdrojová data ze souboru do databáze. To je v podstatě opět mechanická činnost a závisí jen na rychlosti počítače. Zdrojový soubor se čte řádek po řádku a jednotlivé dvojice [*m/z*, intenzita] se ukládají se do databáze. Urychlit či nějak optimalizovat se tato úloha opět nedá.

Třetí krok je zpracování a normalizace na *m/z* pozice. Systém musí ze vstupních přibližně 100 tisíc bodů vypočítat maximum, minimum a průměr pro všech 15 tisíc normovaných *m/z* pozic. Zde opět mnoho možností pro optimalizaci není. Program prostě musí mechanicky přepočítat vložené hodnoty. Naštěstí není množství operací tak vysoké, aby způsobovalo výrazné zpomalení a při výpočtu výrazně pomáhá SQL databáze, která je na takové úkoly optimalizovaná.

Čtvrtým krokem je výpočet intervalů. A tady už problém je. Pokud chceme vyhodnotit všech 15 tisíc pozic a chceme počítat s intervaly o délce 100 pozic nalevo a napravo od sledované pozice, musíme pro každou z pozic vyhodnotit minimum, maximum a průměr pro 200 sousedních pozic. Složitost je v tomto případě 15 000 · 200, což dává obrovskou režii. A pokud má být vyhodnocení provedeno podle více kritérií (viz. výše uvedené nastavení intervalů), roste doba výpočtu s každým dalším kritériem.

Tato část se při testech ukázala jako úzké hrdlo, které způsobovalo neúměrnou délku výpočtu. Prvním verzím algoritmu trval výpočet pro jediné spektrum 15 až 25 minut. To by v praxi bylo nepoužitelné. Bylo tedy žádoucí přijít s nějakým lepším řešením řešením. Buď nějaké nekonvenční řešení, nebo zjednodušení za cenu přijatelného zhoršení kvality.

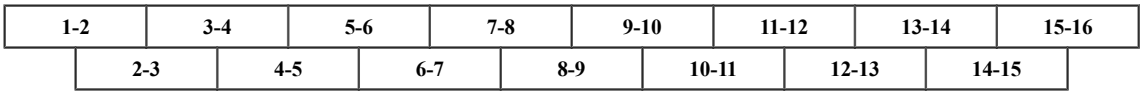
A nekonvenční řešeni se našlo. Celý vtip je v tom, že pokud už máme vypočtené dva sousední intervaly stejné délky, můžeme je triviálním výpočtem spojit do nového intervalu o dvojnásobné délce. Zákony matematiky a statistiky nám toto dovolují. Výpočet intervalů se potom změní na binární slučování kratších intervalů, což je fakticky průměr, minimum a maximum dvojic čísel.

Zjednodušeně řečeno, pokud máme dva sousedící intervaly reprezentované dvěma pěticemi $[x1_1, x1_2, y1_{min}, y1_{max}, y1_{avg}]$ a $[x2_1, x2_2, y2_{min}, y2_{max}, y2_{avg}]$ kde pro každou z pětic je $\langle x*_1; x*_2)$ interval na ose x (*m/z*), trojice $[y*_{min}, y*_{max}, y*_{avg}]$ reprezentuje minimum, maximum a průměr na daném intervalu a dále platí že $x1_2 = x2_1 + 1$ a platí, že $abs(x1_1 - x1_2) = abs(x2_1 - x2_2)$[12], potom sloučením těchto dvou sousedících intervalů vznikne nový interval s dvojnásobnou délkou ve tvaru $[x1_1; x2_2; min(y1_{min},y2_{min}); max(y1_{max},y2_{max}); avg(y1_{avg},y2_{avg})]$.

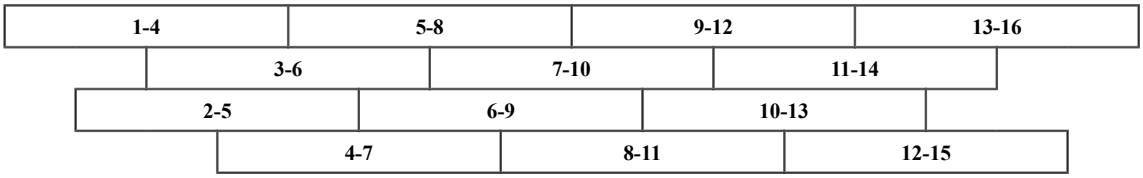
Lépe situaci demonstruje schéma:

Na počátku máme první úroveň, šestnáct intervalů délky 1.

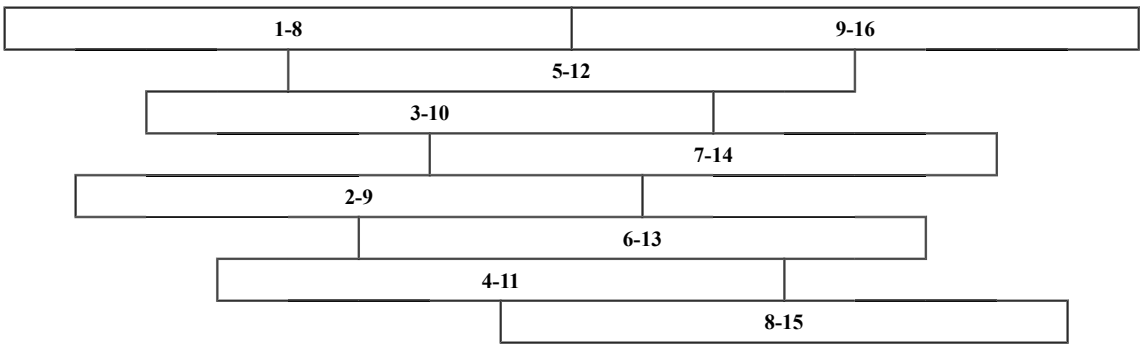
Když spojíme vždy dva sousedící intervaly, dostaneme patnáct intervalů délky 2.



Dalším krokem slučujeme intervaly délky 2 na intervaly délky 4. Pro vysvětlení – první dvě řady vznikly slučováním intervalů délky 2 z prvního řádku a dolní dvě řady vznikají z intervalů pocházejících z dolního řádku.



Z intervalů délky 4 potom vytvoříme intervaly délky 8.



Takto bychom mohli snadno pokračovat k dalším úrovním.

12 Matematicky nepříliš srozumitelně vyjádřeno, že intervaly musí být stejně dlouhé.

Nevýhodou je, že takto přímo můžeme vytvářet pouze intervaly, jejichž délka je mocninou dvojky (1, 2, 4, 8, 16, 32 …). Jiné délky by bylo třeba složit, například délka 20 by se složila z intervalů 16 a 4 se zohledněním poměru jejich délek (váhy). Nicméně v praxi se ukázalo, že vypočítávat takto nerovnoměrné intervaly není třeba a rozdíl v průměrech na intervalu např. délky 100 a délky 128 je v konečném výsledku malý.

Výhodou tohoto řešení je, že můžeme s minimální režií vypočítat intervaly pro různé délky. Interval v prvním kroku dostaneme z x-ových (*m/z*) pozic. A to je u běžného spektra přibližně 15 000 hodnot. Interval druhého kroku (délka 2) dostaneme z intervalů prvního stupně, tj. opět jen spojíme 15 tisíc hodnot ze sousedících intervalů. Další stupeň jsou intervaly o délce 4 a ty dostaneme z intervalů délky 2. Obecně tedy pro intervaly délky n potřebujeme řádově $d \cdot (\log_2 n)$ operací, kde d je délka prohledávaného rozsahu ve spektru (počet x-ových pozic).

Z toho odvodíme, že pokud bychom chtěli použít intervaly o délce zhruba 200 pozic (100 nalevo a napravo od píku), tak z původní složitosti 15 000 · (2 · režie pro výpočet průměru, minima a maxima ze 100 hodnot) se binárním slučováním dostaneme zhruba na složitost 15 000 · 7 · (režie na sloučení dvou intervalů). Hodnota 7 proto, že interval o délce 128 odpovídá 2^7, což je nejblíže ke 100. Režie na sloučení dvou intervalů je průměr z obou průměrů, maximum z obou maximálních hodnot a minimum z obou minimálních hodnot. Tj. výrazně méně operací než počítat dvakrát průměr, minimum a maximum ze 100 hodnot.

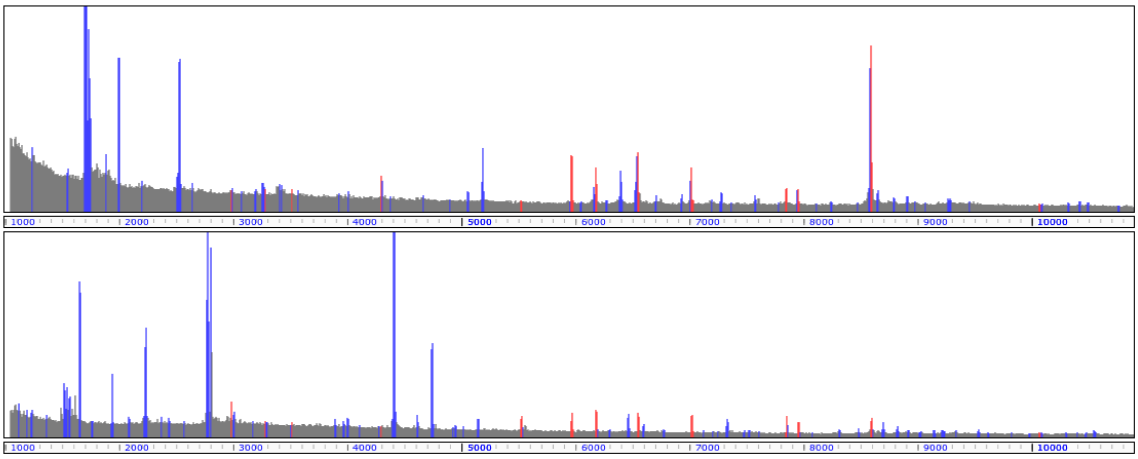
To je velice výrazné zlepšení proti původnímu stavu a výsledkem je, že časová náročnost kompletní analýzy nového spektra (při vložení) se z původních 15 až 25 minut dostala na čas zhruba kolem jedné minuty (samozřejmě v závislosti na hardwaru), což je pro praktické účely zcela postačující.

Samotná detekce píků a bodové ohodnocení jejich souřadnic je potom opět operace s lineální složitostí, kde se musí mechanicky projít všech 15 000 pozic a zkontrolovat, zda jde o píky a přidělit jim příslušné body. Respektive vyhledání píků je realizováno dotazem do databáze na seznam pozic, které mají dostatečně vysokou hodnotu. Získané množině se následně nastaví bodové ohodnocení. Ve skutečnosti je pak díky indexům v databázi vyhledání a ohodnocení prováděno s menší než lineární složitostí (provádí se pouze na podmnožině všech pozic).

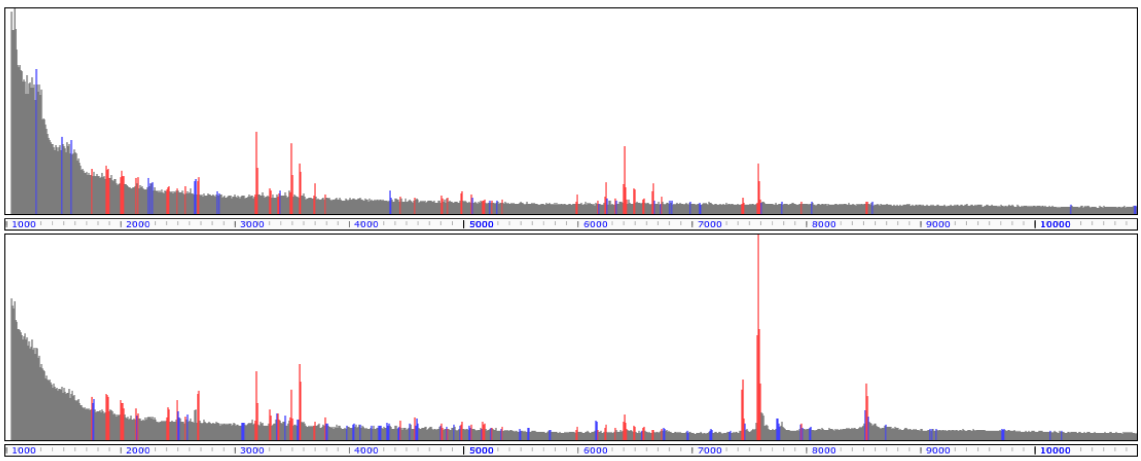
6.9 Efektivita vyhledávání píků a porovnání dvou spekter

Na níže uvedených obrázcích [Obr 30, Obr 31, Obr 32, Obr 33, Obr 34] je demonstrována schopnost detekce píků a jejich vzájemné srovnání s podobným spektrem. Obrázky jsou z grafického rozhraní Biospeanu. Červeně označené píky jsou společné pro obě spektra. Modře označené píky jsou jen v jednom z nich.

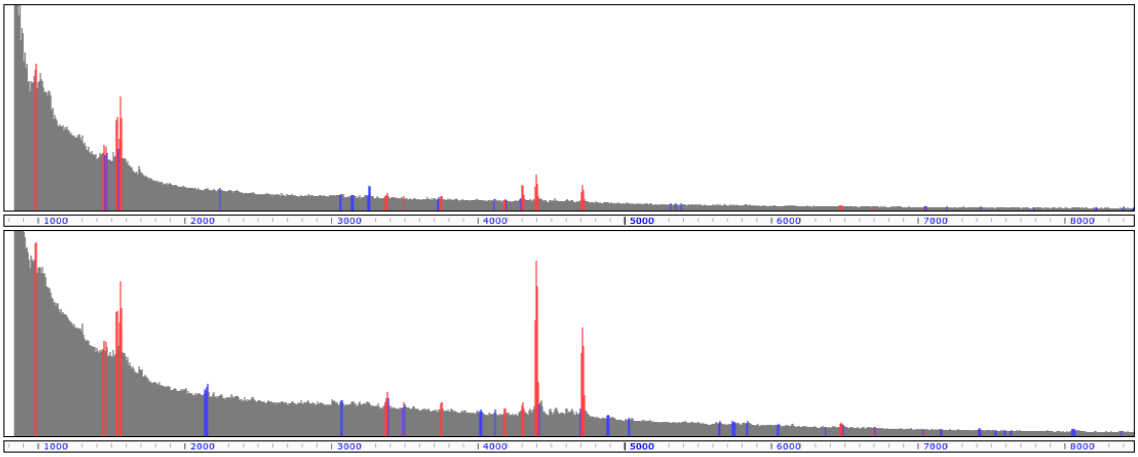
Použita byla spektra ze souboru spekter, která jsem obdržel pro testování. Konkrétní označení porovnávaných spekter je uvedeno pod obrázkem. Spektra byla detekována pouze automatickou detekcí bez dodatečné poloautomatické a manuální korekce (viz. dále).



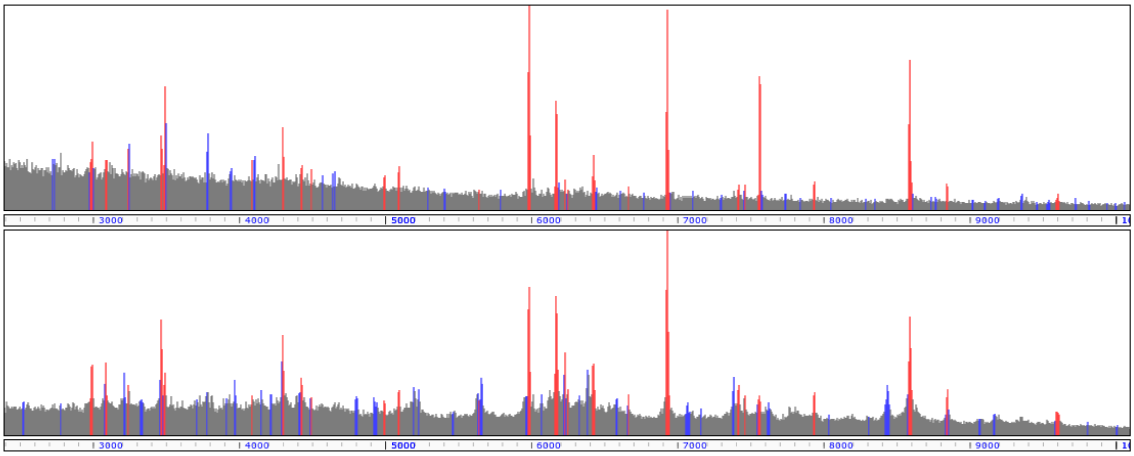
Obr 30: Detekce a porovnání píků 1
vzor_005_110118_P.halstedii_conc_FA_SA_1_1_matrix_ACN_2.5_perc_TFA_96_1_0_H6_1
vzor_016_110929_P.viticola_FA_SA_1_3_ACN_2_5_TFA_1_1_A5_1



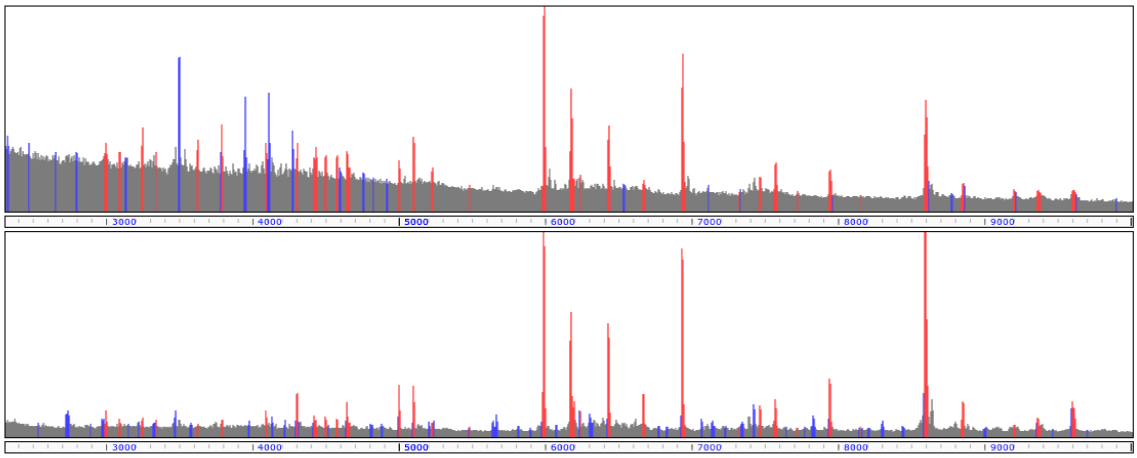
Obr 31: Detekce a porovnání píků 2
120320_Blumeria_graminis_hordei_2567_FA_SA_bremia_biot_met
120321_Blumeria_graminis_hordei_7557_FA_SA_bremia_biot_met



Obr 32: Detekce a porovnání píků 3
121108_Botrytis_cinerea_F-16_in_glycerol_FA_SA.txt
121102_Botrytis_cinerea_F-16_FA_SA.txt



Obr 33: Detekce a porovnání píků 4
100428_Bremia_lactuceae_LP_FA_matrix_10mg_mL_7_3_ACN_1_uL_0_C1_1
B.lactucae_CHCA_10_mg_in_ACN_2.5_TFA_Bruker



Obr 34: Detekce a porovnání píků 5
100909_B.lactucae_LP_FA_matrix_10mg_mL_7_3_ACN_2.5_TFA_1uL_1_0_A4_1
110127_B.lactucae_FA_SA_1_1_v_v_10_30_mg

Z uvedených obrázků je poměrně dobře vidět, že Biospean správně najde a identifikuje naprostou většinu píků. Z klíčových intenzivních píků jsou přitom identifikovány všechny. U méně výrazných píků lze nalézt takové, které jsou evidentně na hraně detekce. Jejich případné označení za píky je potom otázkou poloautomatické a manuální korekce (viz. dále).

6.10 Technologie

Z technologického hlediska se výběr nástrojů odvíjí od způsobu použití aplikace. Zde bylo možné zvolit dvě základní řešení. Prvním by byla klasická lokálně instalovaná aplikace, kterou by možné nainstalovat a použít na libovolném osobním počítači. Druhým řešením bylo vytvoření centralizované webové aplikace, která by trvale běžela na serveru a uživatelé by ji využívali prostřednictvím webového prohlížeče.

Výhody a nevýhody obou řešení lze shrnout do následující tabulky. Jednoznačně výhodnější řešení je zvýrazněno kurzívou.

Kritérium	Instalované řešení - PC	Webové řešení - Server
Požadavky na výbavu uživatele	Je nutné instalovat celou aplikaci do PC uživatele spolu s databází v řádu gigabytů.	*V počítači uživatele je pouze webový prohlížeč dodržující základní standardy, nicméně je nutno být připojen k síti. Kromě počítače lze použít i jiné zařízení, smartphone nebo tablet.*
Operační systém uživatele	Každý OS vyžaduje svoji verzi aplikace, uživatel si systém vybírá sám. Je extrémně náročné udržovat verze aplikace pro různé OS.	*Řešení je nezávislé na OS, který uživatel používá. Lze použít i mobil či tablet. Na centrálním serveru běží operační systém dle výběru tvůrců/správců.*
Aktualizace a údržba	Náročná distribuce změn	*Úprava dat probíhá přímo na serveru a změny jsou ihned k dispozici všem uživatelům. Všichni mají aktuální data.*
Centrální zranitelný bod	*Ne – každá instalace je samostatná*	Ano – centrální server
Výkon	Závisí na PC uživatelů, nemůžeme přímo ovlivnit.	*Závisí na výkonu serveru, můžeme přímo ovlivnit výběrem serveru*
Rychlost reakce a odezva	Prakticky okamžitá, pokud má uživatel dostatečně výkonný počítač.	Obecně pomalejší s ohledem na nutnost komunikace mezi serverem a webovým prohlížečem.

Sdílení dat mezi zařízeními, práce na různých počítačích, přenesení profilu a dat	Problematické, vyžadoval by export/import dat mezi databázovými servery	*Díky webovému řešení přístupnému odkudkoliv a centrálnímu uložení dat na serveru není nutné řešit.*
Sdílení dat v rámci týmu / týmová práce na více zařízeních	Problematické při použití nějakého centrálního úložiště, nepoužitelné při použití instalace na jedno zařízení.	*Bezproblémové. Databázový a webový server zajišťují uložení a dostupnost dat odkudkoliv a pro všechny uživatele současně.*
Práce offline	*ano*	ne
Bezpečnost a ochrana soukromí	*V případě zabezpečeného počítače velmi vysoká. Data jsou pouze v konkrétním počítači.*	Závisí na mnoha faktorech. Kritickým faktorem je přenos dat po síti. S ohledem na umístění dat na serveru obecně nesplňuje požadavky pro systém s vysokými požadavky na ochranu dat (např. data pacientů v nemocnici).

Z tabulky poměrně jednoznačně vyplývá, že většina důvodů hovoří spíše pro použití webového řešení, které také bylo zvoleno. Hlavním a nejpraktičtějším důvodem pro použití webového řešení je centrální správa, stále aktuální databáze spekter a možnost týmové spolupráce více uživatelů (z několika počítačů). Samotné webové řešení lze navíc používat i lokálně bez nutnosti síťového připojení, pokud se instalace provede na počítač, se kterým se bude přímo pracovat. V takovém případě uživatelé opět použijí webový prohlížeč, ale pracovat budou na své domovské adrese (localhost).

Odpůrce webového řešení může namítnout, že „konkurenční“ Biotyper používá lokální instalaci. Nicméně v jeho případě je významným důvodem instalovaného řešení fakt, že jde o komerční aplikaci, kdy se licence prodává na každou pracovní stanici a jeho využití v podobě webově přístupné aplikace je z komerčního hlediska spíše nežádoucí.

Dále s ohledem na způsob použití Biotyperu je v některých případech lokální instalace vyloženě nutná, protože např. nemocniční systémy pracují (z důvodu bezpečnosti a ochrany dat pacientů) zásadně bez přístupu k Internetu.

Navíc vývojáři v případě použití Biotyperu nemusí příliš řešit výkon počítače, alespoň relativně. Vzhledem k ceně Biotyperu několikrát přesahující cenu běžného PC by nemělo dojít na situaci, kdy by se tato aplikace instalovala na nevýkonný počítač. Většinou jsou k tomuto účelu vhodné (výkonné) počítače cíleně vybírány a zakoupeny. Při jeho návrhu se tedy nemusí řešit, jestli uživatel nebude chtít program používat na starém pomalém notebooku s minimem paměti.

V případě Biospeanu, který je v toto chvíli zdarma, ovšem nemůžeme požadovat, aby si uživatel pořizoval výkonný počítač s nejnovější generací procesoru.

Pro výběr webového řešení ovšem přispěly i technologie a postupy, které musela aplikace využívat. S ohledem na to, že tento typ aplikace se neobejde bez centrální databáze a je vhodné používat nástroje pro paralelizaci, automatické spouštění úloh atd., je řešení s webovým serverem rozumnější než instalovaná aplikace.

V podstatě se tak dospělo k řešení, které je osvědčené a dostatečně robustní – fyzický server používající operační systém Linux na kterém je spuštěn webový server Apache, který pro ukládání dat využívá osvědčený databázový server MySQL. Uživatelé potom mohou aplikaci využívat odkudkoliv jen pomocí svého oblíbeného webového prohlížeče, což je pro ně snadné a bezproblémové.

Výhoda systémů/serverů na bázi Linuxu spočívá v efektivní správě procesů, v bezpečnosti, snadné údržbě a stabilitě a také v řadě dalších funkcí, které systém podporuje (např. výše zmíněný cron). V neposlední řadě je výhodou také opensource licence, konzistentně probíhající vývoj a nulové licenční náklady, což je pro akademické prostředí ideální volba.

Klíčovou součástí celé aplikace je přitom databázový server s relační databází, který skladuje veškerá data a také vykonává velkou část výpočtů. Pro hromadné zpracování dat, získávání agregovaných výsledků (minima, maxima, průměry) a filtraci výsledků je databáze mocným a ideálním nástrojem.

Zde se od začátku nabízela otázka který databázový server použít. Volby v podstatě byly na začátku vývoje možné dvě, ve chvíli dokončování projektu tři. Buď použít výkonný freewarový server MySQL [MySQL], který je rychlý a donedávna široce podporovaný, nebo komunitní opensource server PostgreSQL [PgSQL], který je silný především v pokročilých funkcích a při práci s velkým objemem dat. V poslední době se jako další řešení nabídla i databáze MariaDB [MariaDB], což je nezávislý svobodný fork MySQL, který vznikl jako reakce komunity na celkový vývoj kolem MySQL a od počátku si získal velký ohlas.

Volba na začátku vývoje Biospeanu padla na databázi MySQL, ale v současnosti se počítá s přechodem na MariaDB pro její lepší perspektivu a licenční politiku. Nicméně ani použití PostgreSQL není žádným problémem. Její nasazení je v případě potřeby možné po snadné úpravě kódu. PostgreSQL je proti MySQL otevřenější a do budoucna by její použití mělo smysl, pokud by požadavky na funkce a především objem dat databázi narostly.

Je zajímavé, že i samotný Biotyper používá databázi PostgreSQL, která se instaluje současně a funguje nezávisle jako paralelní úložiště dat[13]. S ohledem na tento fakt můžeme usuzovat, že i vnitřní fungování Biotyperu bude podobné naší aplikaci a výpočty budou masivně využívat funkcí databáze.

Pro samotné grafické prostředí (webové stránky) Biospeanu byl využit jazyk HTML, přičemž byla snaha se držet standardu HTML5. Dynamická část stránek je vytvořena v jazyce PHP (část na straně serveru) a v JavaScriptu (část na straně webového prohlížeče). Jde o běžně používané a široce rozšířené technologie, nad kterými není třeba se rozsáhleji pozastavovat.

Dále byly využity svobodně použitelné knihovny jQuery [jQuery] a phpMyGraph [phpMyGraph], které zajišťují část služeb Ajaxu, resp. generování statistických grafů.

6.11 Funkce uživatelského rozhraní a jejich technické řešení

V této části si objasníme některé funkce aplikace Biospean, jejich smysl pro uživatele a jejich technickou realizaci.

6.11.1 Nastavení vlastností spektra

Každé spektrum si uživatel může pojmenovat, doplnit o poznámky a zařadit je do skupin (které lze sdílet s přáteli), což jsou možnosti, které mu zjednoduší práci. Co se týče technických vlastností, tak mezi nimi je možnost nastavit použitou matrici, efektivní rozsah spektra, koeficient pro detekci píků a také manuálně přidat či odebrat píky. Ke správnému nastavení mohou uživateli výrazně pomoci i statistické křivky a píklist.

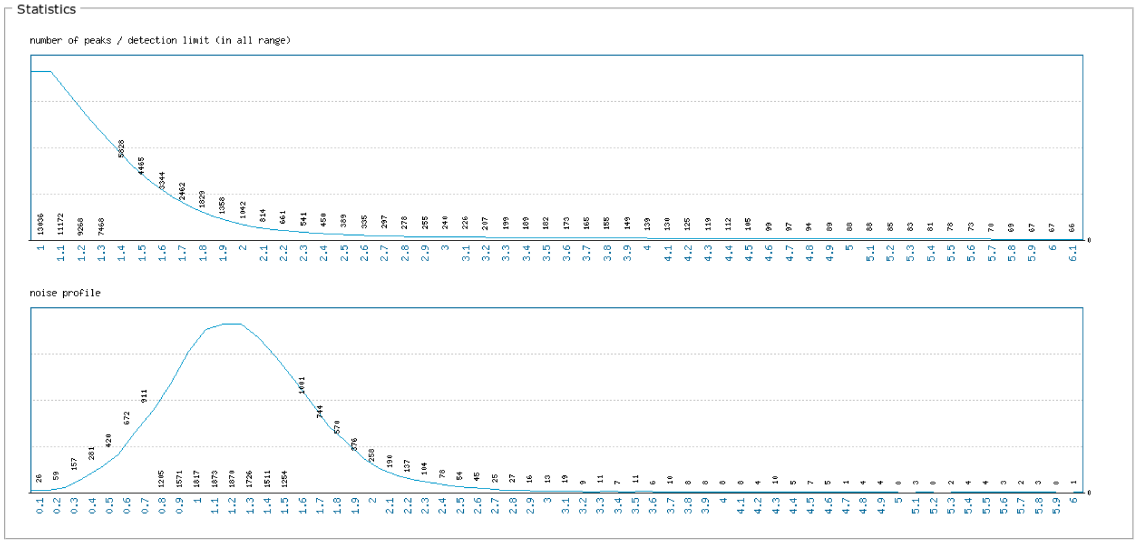
Nastavení použité **matrice** má smysl jak z organizačního hlediska (pro přehlednost), tak pro vyhledávání. Ačkoliv Biospean při porovnání spekter nijak nezohledňuje použitou matrici, tak je její nastavení žádoucí ze dvou důvodů. Jednak umožňuje hledat jen mezi spektry se zvolenou matricí, což urychluje vyhledávání (hledá se mezi méně spektry). Druhým důvodem je, že každý druh matrice svým charakteristickým způsobem ovlivňuje výsledné spektrum a porovnávat spektra měřená s různými druhy matrice může vést k nepěkným výsledkům.

Velice užitečnou vlastností je nastavení **efektivního rozsahu spektra**. Zde si uživatel zvolí jen interval „od – do“, který ho ve spektru opravdu zajímá. Nastavení efektivního rozsahu umožňuje vypočítávat podobnost mezi spektry jen na zvoleném intervalu a vyloučit tak ze srovnání píky, které jsou mimo efektivní rozsah. To velice výrazně zlepšuje výsledky při

13 Tuto skutečnost jsem odhalil náhodou náhledem na seznam instalovaného softwaru v počítači na KBC PřF UP, kde jediná nám dostupná instalace Biotyperu běží. Po chvilce pátrání jsem zjistil, že podobně jsou využívány např. i části komerčního výpočetního nástroje Maple.

porovnávání spekter, protože neporovnává píky, které jsou mimo oblast zájmu. To navíc i výrazně zrychluje celý výpočet podobnosti, protože srovnání probíhá jen na zvoleném intervalu, který je menší než celý rozsah.

Volba **koeficientu pro detekci píků** je v podstatě druhým krokem při detekci píků. Prvním krokem je automatická detekce, kterou provede software automaticky. Může se ovšem stát, že automatický výběr píků není příliš dobrý, nebo uživatel chce striktnější nebo naopak volnější podmínky pro detekci. Pokud je ve spektru hodně proměnlivého šumu, je vhodné se podívat jak vypadají křivky **statistických grafů** [Obr 35] a na základě nich stanovit vhodnější koeficient pro detekci. Tato uživatelem zadaná změna provede přepočet ohodnocení píků a při správné volbě může výrazně zlepšit kvalitu píklistu.

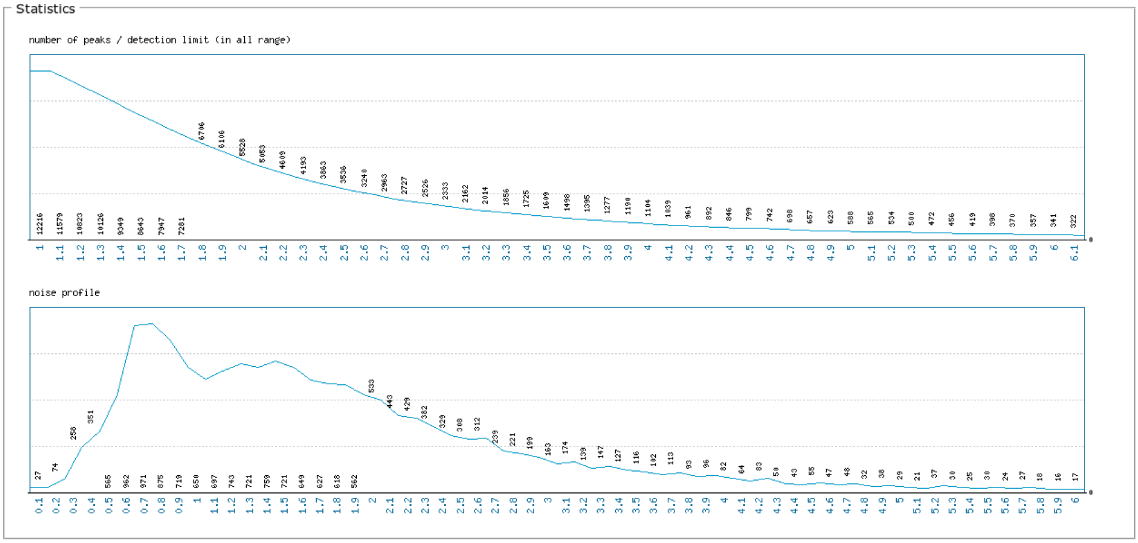


Obr 35: Statistické informace o šumu a pících ve spektru
Obrázek ukazuje statistické rozložení rozptylu šumu / píků ve spektru dle zvoleného koeficientu (osa x). Horní graf ukazuje kolik celkem píků bude označeno jako pík, pokud bude koeficient pro jejich detekci zvolen na uvedenou hodnotu. Graf začíná na maximální hodnotě a se zvyšujícím se koeficientem počet píků klesá.
Spodní graf ukazuje statistické rozložení šumu/píků v celém spektru. Udává kolik [m/z] pozic má relativní hladinu odpovídající danému koeficientu. Tento graf má tvar Gaussovy křivky – nejvíce šumu se vyskytuje zhruba kolem hodnoty 1 (průměrná hladina šumu) a směrem k 0 (baseline) klesá. Stejně tak nad hladinou 2 (teoreticky maximální úroveň šumu) je již šumu málo a převažují píky.
Tyto grafy mohou uživateli výrazně usnadnit manuální výběr koeficientu pro detekci píků. Snížením hodnoty koeficientu dosáhneme vyššího počtu píků za cenu, že bude mezi píky zahrnut i šum. Nastavením na vyšší hodnotu můžeme naopak vynechat kromě šumu i píky s malou intenzitou. Z uvedeného grafu je vidět, že pro uvedené spektrum by rozumná volba koeficientu byla někde mezi hodnotami 2,4 až 2,7.

Statistické grafy umožňují uživateli nejen snadno nastavit koeficient pro detekci píků, ale poskytují mu i určitou informaci o kvalitě samotného spektra. Kvalitní spektrum (správně zpracované a změřené) by mělo mít profil šumu odpovídající Gaussovské křivce s vrcholem někde mezi 1,0 až 1,5 jak můžeme vidět na obrázku výše [Obr 35]. Při změření spektrometru „naprázdno“ (bez vzorku) bychom měli mít jen šum a profil takovéhoto „ideálního šumu“ by

(teoreticky) měl mít vrchol při koeficientu 1,0 a rozsah od 0 do 2,0. Přítomnost píků a skutečnost, že koeficienty pro šum se při vyhodnocení počítají „zespodu“ (mezi baseline a průměrem, nikoliv na základě maxima) posouvají vrchol křivky mírně nad hodnotu 1. Proto také není moudré volit koeficient pro rozeznání píků přímo na hodnotu 2, kde ještě zůstávají maxima šumu, ale vybrat např. 2,5, což je hodnota, kterou jsem vybral jako výchozí.

Samotné přesné stanovení koeficientu je však v některých případech složité. Pokud je spektrum nekvalitní (z jakéhokoliv důvodu – příprava vzorku, čistota, kvalita a kalibrace přístroje atd…), nepomůže sebelepší nastavení, jak ukazuje níže uvedená statistika ze spektra, které je evidentně nekvalitně změřené [Obr 36].



Obr 36: Statistika u spektra, které není kvalitní
Na těchto grafech je na první pohled vidět, že hmotnostní spektrum není kvalitní – křivka šumu má ke klasickému tvaru Gaussovy křivky daleko, počet píků stejně jako šum klesají jen pozvolna a nerovnoměrně.

Z důvodu existence nepravidelných spekter jsem také při návrhu upustil od myšlenky plně automatické detekce koeficientu. Plně automatická detekce by na základě matematických metod vyhledala „bod zlomu“, kde se průběh Gaussovy křivky „narovnává“ po rychlém poklesu. Pro ideálně změřené spektrum by to bylo řešitelné, ale existence nemalého počtu nekvalitních spekter mne vedla k použití konzervativnějšího řešení založeného na automatické detekci s možnosti individuální korekce každého spektra dle přání uživatele.

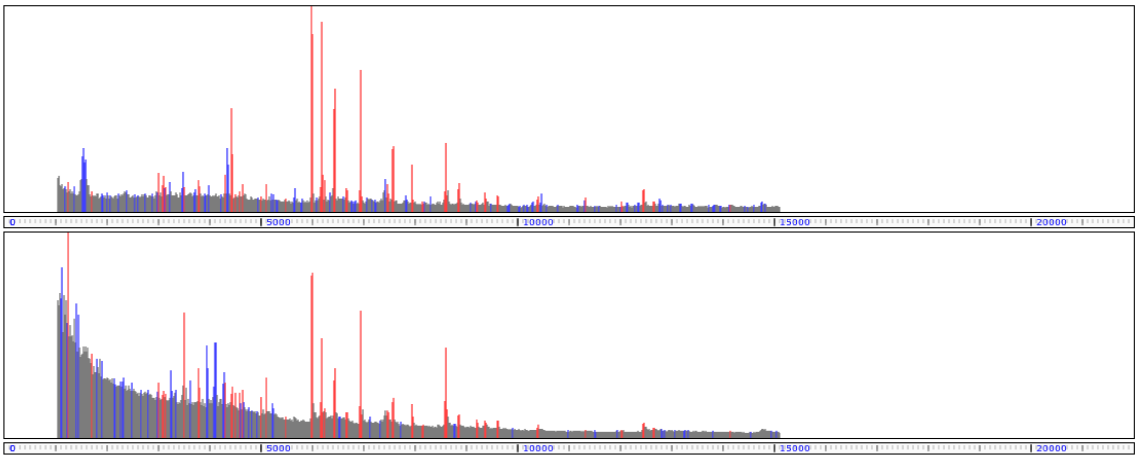
Třetím a posledním krokem pro nastavení píklistu je možnost **manuálně přidávat či odebírat píky**. Pokud se nepodaří správně nastavit vhodný koeficient pro detekci všech píků (koeficientem toto nemusí být ani možné), je nutné poslední odchylky v píklistu zeditovat ručně. Toto je v podstatě funkce poslední záchrany, která umožňuje manuálně korigovat veškeré nepřesně detekované píky a to s naprostou přesností.

Uživatel pomocí rozhraní postupně zadá pozice píků, nebo intervaly píků, které chce přidat nebo naopak odebrat. Tento nástroj je přitom zcela nezávislý na automatické detekci a umožňuje odstranit i takové píky, které byly z hlediska struktury signálu správně detekované, ale ve vzorku je nechceme (např. příměsi). Dokonce umožňuje označit jako píky i oblasti, které píkem nejsou.

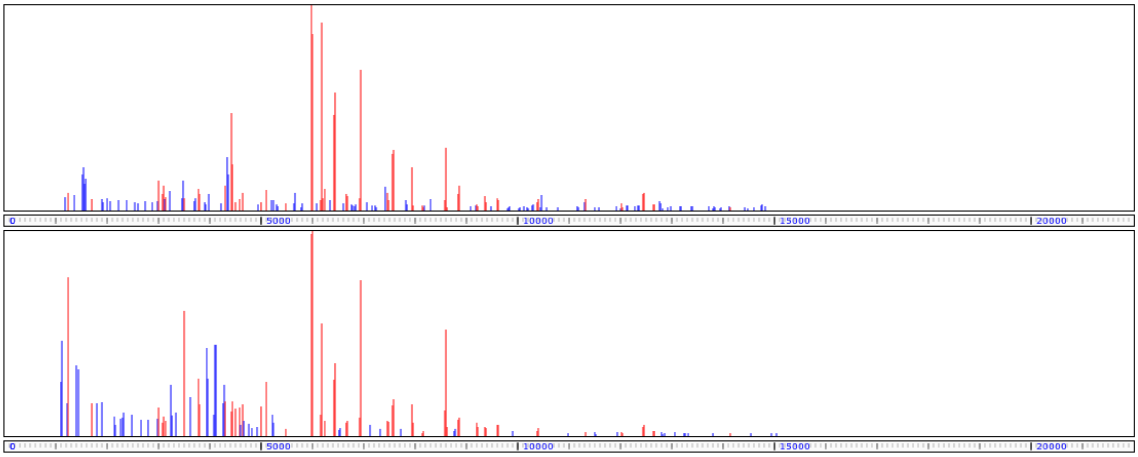
Systém tuto korekci provádí pomocí dodatečných opravných bodů, které přidá ke stávajícímu ohodnocení (pro přidání píků nastaví kladné body, pro odebrání záporné). Díky tomu, že opravné body se ukládají odděleně od automatického bodování, je možné je opětovně odebrat či znovu modifikovat. Stejně tak je možné např. i po korekci koeficientu dohledat, které píky v předchozím nastavení uživatel modifikoval.

6.11.2 Detailní porovnání dvou spekter

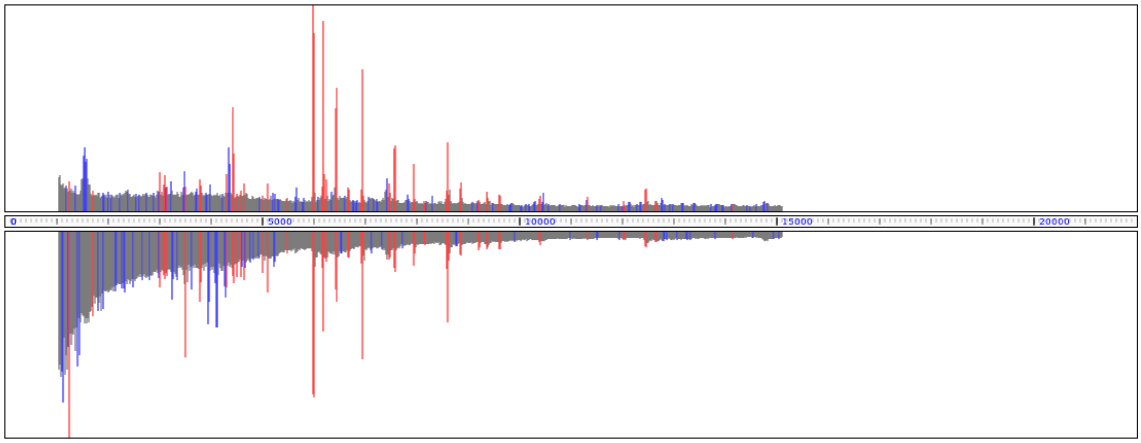
Biospean umožňuje uživatelům detailní porovnání dvou zvolených spekter. Volitelný způsob zobrazení přitom poskytuje dostatek možností, jak srovnání vizualizovat. Různé způsoby zobrazení lze vidět na obrázcích [Obr 37][Obr 38][Obr 39][Obr 40]. Spektra lze zobrazit s šumem nebo bez, je možné zapnout zvýraznění píků pomocí barev (červená – společné píky, modrá – píky jen v daném spektru), je možné zobrazit spektra nad sebou nebo zrcadlově převrácená. Spektra lze také zoomovat, což umožňuje detailně zkontrolovat konkrétní pozice.



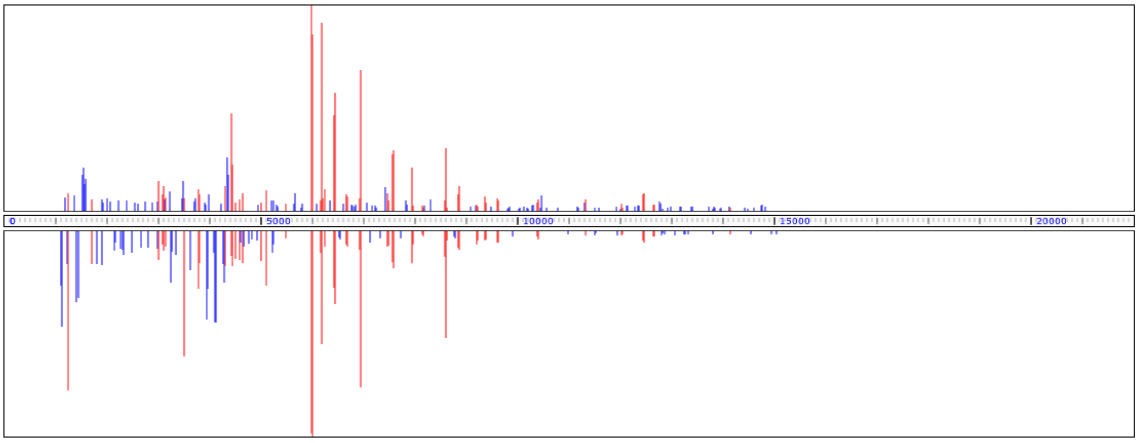
Obr 37: Konvenční zobrazení dvou spekter.
Na tomto obrázku jsou vidět červeně zvýrazněné píky, které jsou společné pro obě spektra. Modře zabarvené píky jsou pouze v jednom z spekter. Šedivé oblasti jsou šum.



Obr 38: Srovnání spekter bez šumu.
Toto zobrazení ukazuje spektra bez šumu. Zůstávají jen červené píky společné pro obě spektra a modré, na kterých shoda není. Intenzita píků je zobrazena po odečtení šumu, tj. píky na začátku spektra, kde je nejvyšší šum a které jsou při zobrazení s šumem nejvyšší, jsou po odečtení vlivu šumu nižší než intenzivní píky uprostřed spektra.



Obr 39: Zrcadlové porovnání dvou spekter.
Zrcadlové zobrazení je výhodné pro porovnání pozic píků.

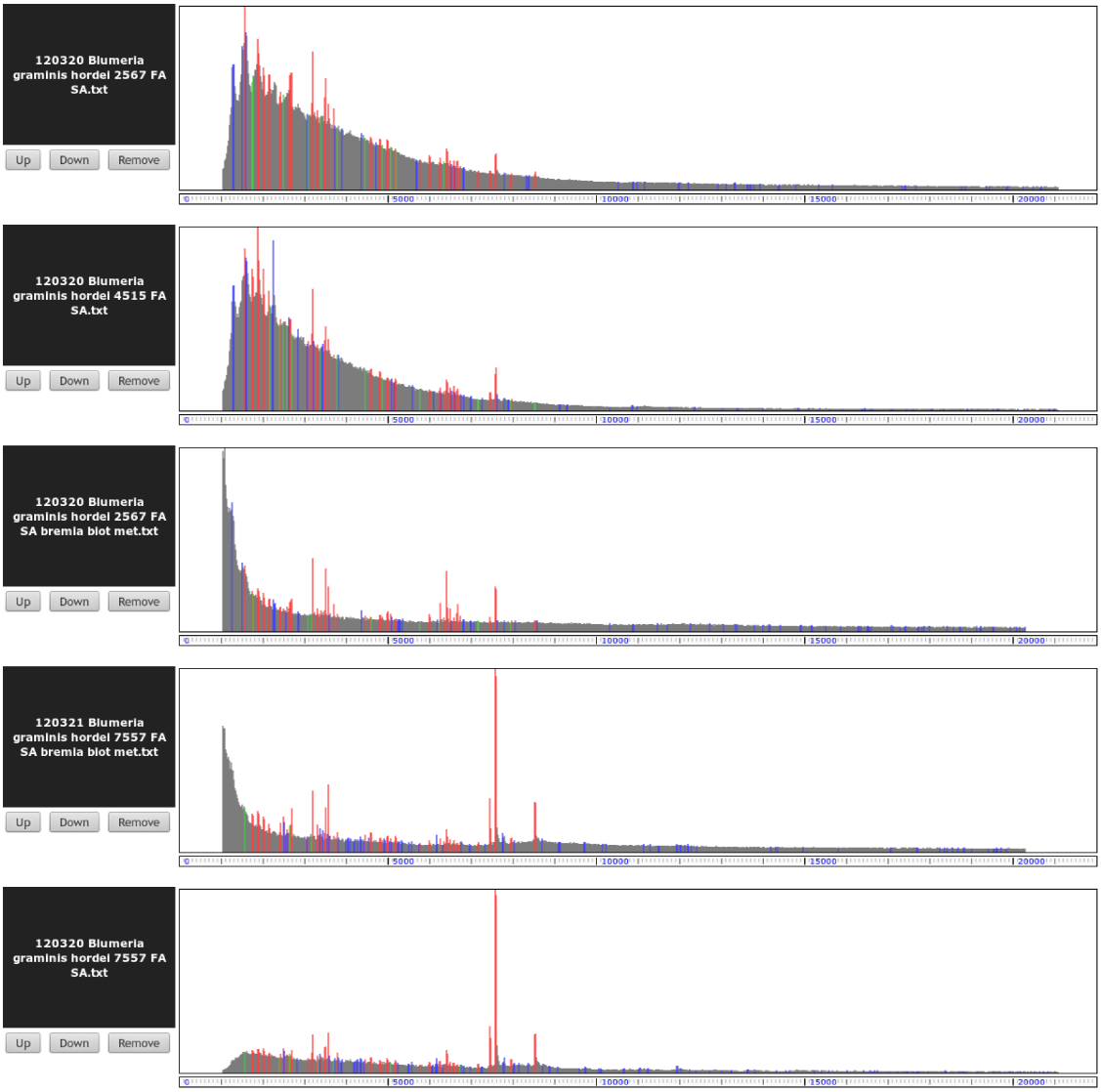


Obr 40: Zrcadlové zobrazení bez šumu.
A takto vypadá zrcadlové zobrazení bez šumu. Shodné pozice je snadné zkontrolovat pouhým pohledem.

6.11.3 Porovnání více spekter

Další užitečnou funkcí je nástroj na vizuální porovnání více spekter. Ten umožňuje přidávat do srovnání vybraná spektra jednotlivě nebo přidat najednou spektra z celé skupiny. U zobrazených spekter je možné měnit pořadí nebo je ze seznamu srovnávaných spekter vyjmout.

Podstatná je i vizuální stránka zobrazení. Grafy spekter zobrazují pomocí barev, na kterých pozicích je shoda mezi všemi zvolenými spektry. Červené jsou píky, které se nachází ve zvolené většině spekter. Modré jsou takové píky, které jsou jen v konkrétním spektru. A pokud si uživatel nastaví poměr pro detekci menší než 100 %, objeví se i zelené píky – to jsou píky, které jsou detekovány ve zvoleném procentu spekter, ale chybí v tom konkrétním [Obr 41].



Obr 41: Srovnání více spekter
Červené píky jsou společné pro většinu spekter (množství zvolené procentuálně). Zelené píky jsou takové, které jsou společné pro většinu spekter, ale nevyskytují se v tom konkrétním spektru. Modré jsou píky, které nejsou společné.

Poměr pro detekci umožňuje nastavit v kolika procentech spekter se musí pík vyskytovat, aby byl označen za společný pro všechna spektra a označen červeně. Pokud je nastaveno např. 80 %, tak musí být daná pozice alespoň v 80 % spekter píkem, aby byla označena jako společný pík. Ve zbylých 20 % spekter, kde píkem není, bude taková pozice označena zeleně.

Může se to zdát jako zbytečné, ale tato funkce ve skutečnosti dává uživateli do rukou velice mocný nástroj. Pokud má např. několik měření ze stejného vzorku, ale s méně kvalitním signálem, je schopen pomocí toho nástroje eliminovat část nekvalitních vzorků a získat kvalitní píklist, který obsahuje většinu píků, která je společná pro většinu vzorků.

Schopnost najít průnik píků společných pro více spekter otevřela prostor pro další nástroj – virtuální spektra.

6.11.4 Virtuální spektra

Schopnost detekovat průnik píků je dotažena do úplného konce v možnosti vygenerovat umělé **virtuální spektrum**, což je seznam společných píků vygenerovaný v podobě ASCII souboru, který lze následně znovu načíst do systému jako zcela nové spektrum. Takto získané nové virtuální spektrum vykazuje (jednostranně) plnou shodu se všemi spektry, které byly použity jako výchozí množina spekter, ze které se virtuální spektrum vytvářelo (pokud bylo kritérium pro průnik nastaveno na 100 %).

Jak tohoto nástroje využít? Představme si situaci, kdy se např. bakterie projevuje v organizmu některými specifickými látkami a my chceme najít něco, podle čeho ji jasně detekovat. Bakterie se v hmotnostním spektru projeví přítomností píků na specifických pozicích. Ve spektru ovšem budou i další píky, které budou pocházet z okolního prostředí (tkáně). A z různých tkání s danou bakterií dostaneme různá spektra s různými píky, ale píky typické pro hledanou bakterii by měly být ve všech spektrech stejné.

Potřebujeme tedy ze všech těchto spekter vyfiltrovat takové píky, které se vyskytují ve všech vzorcích. Nejprve tedy vezmeme různá spektra změřená s danou bakterií. Pomocí průniku píků si vytvoříme v Biospeanu umělé virtuální spektrum. To bude v ideálním případě obsahovat jen píky, které jsou typické pro danou bakterii. A budeme doufat, že pocházejí pouze z hledané bakterie, protože ze samotných dat to nelze nijak poznat.

Získané virtuální spektrum vložíme do Biospeanu jako nové spektrum. Toto spektrum tedy bude obsahovat jen píky, které vznikly průnikem píků vzorových spekter. Pokud toto spektrum porovnáme s jakýmkoliv spektrem, ze kterého bylo vytvořeno, bude detekována stoprocentní

(jednostranná) shoda píků virtuálního spektra s píky z libovolného vzorového spektra (za předpokladu, že jsme požadovali 100% zastoupení píku ve vzorových spektrech).

Pro detekci přítomnosti bakterie potom postačí vložit do Biospeanu data ze vzorku, který chceme na přítomnost bakterie prověřit a porovnat tento vzorek s virtuálním spektrem. Pokud bude mít virtuální spektrum úplnou nebo téměř úplnou shodu, je důvod se domnívat, že bakterie je přítomna.

6.11.5 Skupiny a přátelé – týmová spolupráce

Zvláštností Biospeanu na rozdíl od jiných softwarů je schopnost fungovat jako **prostředek týmové spolupráce**. Pro výzkumné týmy je určitě užitečné, pokud mohou členové týmu sdílet svoje získaná data. Proto byla do Biospeanu přidána možnost vytvořit si seznam svých přátel (dalších uživatelů) a těm poskytnout přístup k vybraným spektrům.

Této funkčnosti je dosaženo pomocí **friendlistu** (seznamu přátel) a **skupin spekter**.

Friendlist funguje tak, že uživatel si z dostupných uživatelů vybere ty, které chce mít v seznamu přátel. Nově zařazený uživatel ovšem musí k „přátelství“ udělit souhlas. Pokud souhlas udělí, je přidán do friendlistu. Přidání do friendlistu je přitom jednostranné – uživatel A si přidá do svého seznamu přátel nového uživatele B a on mu toto potvrdí. Nicméně nově přidaný uživatel B si souhlasem automaticky nepřidá uživatele A do svého friendlistu. Aby uživatel A mohl mít uživatele B ve friendlistu a uživatel B měl ve svém friendlistu uživatele A, tak jednak si A musí přidat B do svého friendlistu a uživatel B mu musí udělit souhlas a jednak si B musí přidat A do svého friendlistu a souhlas mu musí dát A. Uživatele si lze samozřejmě z friendlistu také odebrat.

Skupiny spekter slouží jednak k logickému členění a rozdělení spekter a jednak k udělení přístupu. Uživatel si může vytvořit skupiny, do kterých může svá spektra zařadit. Jedno spektrum může být zařazeno i ve více skupinách. Rozhraní nabízí možnost filtrovat výsledky hledání jen na vybrané skupiny spekter, což zlepšuje přehlednost a především urychluje vyhledávání (hledá se jen mezi spektry z vybraných skupin).

Další velkou výhodou skupin je možnost přidělit přátelům z friendlistu přístup do těchto skupin. Vlastník skupiny má možnost přidat právo přístupu (jen pro četní) pro vybrané přátele ze svého friendlistu. Přátelé, kteří budou do takové skupiny přiřazeni, mohou číst (prohlížet, porovnávat a využívat k hledání) všechna spektra, která majitel skupiny do skupiny zařadí. To umožňuje snadnou spolupráci vědeckého týmu, kdy každý uživatel má svoje spektra a může je nabídnout

k nahlednutí svým kolegům. Kolegové nemají oprávnění spektra nijak měnit, ale mohou je prohlížet a porovnávat se svými vlastními spektry.

6.12 Příklady užití Biospeanu v praxi

Níže uvedené příklady praktického využití softwaru Biospean jsem realizoval na povinné tříměsíční vědeckovýzkumné stáži na TU Vídeň (listopad - prosinec 2012, únor - duben 2013).

6.12.1 Detekce stavu fermentace

Prvním úkolem, který jsem na stáži řešil byl pokus vyzkoušet, zda Biospean dokáže rozeznat stav fermentační reakce podle hmotnostního spektra.

Testovací data mi dodala vídeňská strana (Helmel M., Marchetti-Deschmann M.) a bylo mi sděleno, je v nich zaznamenán experiment sledující produkci penicilinu plísní *Penicillium chrysogenum* [Posch at al., 2013] (fermentace). Byl jsem požádán o porovnání podobnosti naměřených spekter a vyzkoušení, zda je ze spekter možné rozlišit jednotlivé fáze fermentace (počátek, průběh, ustálení). Spektra byla získána ze vzorků obsahujících buňky zkoumaných plísní. Samotný experiment jsem po biochemické stránce blíže nezkoumal, můj úkol spočíval ve vyhodnocení dat.

Zkoumaná spektra byla rozdělena do šesti skupin, přičemž každá skupina odpovídala jednomu sledovanému vzorku. Každá skupina spekter byla získána tak, že probíhal sledovaný proces kvašení/fermentace a z něj byly průběžně odebírány vzorky a analyzovány hmotnostním spektrometrem. Každá testovaná skupina tedy byla složena z několika po sobě následujících spekter (jejich pořadí známe), která zaznamenávají postupné změny ve sledovaném vzorku.

Z teoretického hlediska lze porovnání založit pouze na předpokladu, že spektrum zachycující stav reakce bude podobné některému ze spekter, která budeme mít jako referenční vzorky z minulosti. Hmotnostní spektrum se během reakce průběžně mění. Intenzita píků se zvětšuje či zmenšuje, vznikají nové píky nebo zanikají staré. Pro nalezení stavu, ve kterém se fermentace nachází, je v tedy nutné k testovanému vzorku najít nejpodobnější referenční vzorek, resp. několik nejvíce podobných vzorků.

Pokud referenční spektra označme jako x_1 až x_n, teoreticky bychom měli při zkoumání nového spaktra dostat největší podobnost se vzory x_i a x_{i+1}, přičemž testovaný vzorek se nachází ve stavu „někde mezi x_i a x_{i+1}“. Další nejvíce podobné vzory by potom měly být vzory x_{i-1} a x_{i+2}. S rostoucí (časovou) vzdáleností od testovaného vzorku by potom měla postupně klesat i

podobnost testovaného vzorku s předloženými vzory. Takže podobnost testovaného vzorku s x_{i+1} by měla být větší než s x_{i+2} a podobnost s x_i by měla být větší než s x_{i-1}. Podle vzorů x_i a x_{i+1} a dalších blízkých vzorů bychom tedy měli být schopni snadno určit fázi fermentace, ve kterém se testovaný vzorek nachází.

Tento předpoklad se v principu potvrdil.

Zpracování pomocí Biospeanu proběhlo poměrně konvenčně. Data byla načtena jako běžná spektra. Samotné porovnání jednotlivých spekter s časově nejbližšími spektry ukázalo, že detekce stavu je víceméně možná. Při porovnání většiny vybraných vzorků se ukázala největší podobnost právě se sousedními spektry.

Pro větší přehlednost jsem vídeňským kolegům za pomoci upravených funkcí z Biospeanu vytvořil z analyzovaných dat tabulky (pro každou skupinu spekter jednu) se vzájemnou podobností jednotlivých spekter.

V příloze D jsou uvedeny získané tabulky, kde je přehledně vidět vzájemná podobnost spekter. V každé tabulce jsou spektra z dané skupiny seřazena tak, jak šla po sobě v čase. Na průsečíku řádku a sloupečku je uvedena procentuální podobnost spektra z řádku vůči spektru ve sloupečku.

Na hlavní diagonále tabulky jsou hodnoty 100 %, protože se porovnávají spektra sama vůči sobě. Tabulka přitom není symetrická. Na průsečíku spekter A ku B je pochopitelně podobnost jiná než mezi B ku A (viz. popis výpočtu podobnosti u Biospeanu).

Co můžeme z tabulek vyčíst a jaké bychom měli očekávat výsledky? Pro příklad si uveďme nejméně rozsáhlou skupinu spekter [Tabulka 3]:

%	04	05	06	07	08	09	10	11	12	13
BCB1_1-sample_04	100,00	20,95	8,75	8,54	8,55	7,95	8,29	8,71	5,98	2,73
BCB1_1-sample_05	61,68	100,00	31,00	29,18	29,81	26,66	28,00	30,21	19,33	10,10
BCB1_1-sample_06	54,21	65,24	100,00	81,71	84,39	80,59	76,48	77,20	55,89	24,55
BCB1_1-sample_07	45,79	53,17	70,74	100,00	78,40	77,19	67,43	70,81	52,64	22,83
BCB1_1-sample_08	45,33	53,65	72,17	77,44	100,00	86,45	79,44	82,58	58,96	22,32
BCB1_1-sample_09	42,52	48,41	69,53	76,92	87,21	100,00	78,26	82,95	62,65	25,45
BCB1_1-sample_10	45,79	52,54	68,17	69,43	82,80	80,86	100,00	83,04	56,94	19,70
BCB1_1-sample_11	43,93	51,75	62,82	66,55	78,57	78,23	75,80	100,00	61,25	26,16
BCB1_1-sample_12	31,78	34,92	47,96	52,18	59,17	62,33	54,82	64,60	100,00	37,17
BCB1_1-sample_13	12,62	15,87	18,33	19,69	19,49	22,03	16,50	24,00	32,34	100,00

Tabulka 3: Podobnost spekter ze vzorku BCB1_1

Spektra jsou seřazena tak jak byla měřena v průběhu času. S ohledem na to, že víme, že spektra v tabulkách byla měřena po sobě, známe tím jejich správné pořadí a předpokládáme, že největší podobnost by měla být vždy se sousedícími spektry a postupně se s rostoucí časovou vzdáleností zmenšovat. V tabulkách se můžeme přesvědčit, že výsledky odpovídají tomu, co bylo teoreticky očekáváno.

Pokud se podíváme na hodnoty 100 % na hlavní diagonále a podíváme se na hodnoty napravo, nalevo, nad i pod tyto hodnoty, jasně uvidíme, že podobnost mezi spektry se snižuje s rostoucí vzdáleností od hlavní diagonály. Nejmenší podobnost je potom mezi prvním a posledním spektrem, tj. mezi levým spodním a pravým horním rohem tabulky.

Zde zřejmě dojde oprávněně k námitce, že na několika místech tabulky tomu tak není a že se zde vyskytují mírně odlišné hodnoty. A to přesto, že je pokles podobnosti směrem od hlavní diagonály do okrajů zcela zřejmý a v obecně platí.

Je tomu skutečně tak a vysvětlení je zcela prosté – na vygenerování tabulek byla použita data zpracovaná pouze prvním (plně automatickým) stupněm detekce. Tj. nebyla použita ani poloautomatická ani manuální korekce výsledků (druhý a třetí stupeň analýzy spektra, viz. popis Biospeanu). V případě dodatečné poloautomatické a následné manuální korekce by byl výsledek mnohem bližší k ideálnímu stavu.

Důvodem proč jsem nepoužil druhý a třetí stupeň je ten, aby byla demonstrována schopnost vypořádat se automaticky i s poměrně nekvalitními spektry. Analyzovaná spektra přitom obsahovala poměrně velké množství šumu a protože nebyla provedena zmíněná dodatečná korekce, můžeme odchylky v rádu jednotek procent považovat za velmi uspokojivé.

Pro šťouravého čtenáře zdůrazňuji, že u každé tabulky se jedná o časosběrná spektra získaná z jediného vzorku, tj. jejich podobnost je opravdu vysoká. V praxi už shoda kolem 40 % prakticky na jistou říká, že spektra jsou příbuzná a odchylky v řádech jednotek procent může způsobit i pouhá odchylka při měření, kdy se menší píky ztratí v šumu spektrometru.

Pro zcela jednoznačnou odpověď na otázku přesnosti detekce píků by muselo být měření prováděno na skupině spekter, který by musela být ideálně kvalitní. Takovou skupinu spekter ovšem k dispozici nemám.

I přes problémy s kvalitou spekter a lokální odchylky můžeme na základě výsledků porovnání prohlásit, že data z uvedených tabulek podala poměrně silný důkaz o tom, že použité řešení je funkční a aplikace Biospean pracuje tak jak má.

Pro samotné vyřešení úkolu (detekce stádia fermentační reakce podle spektra) se tedy Biospean osvědčil. Ve většině případů největší ukazoval podobnost právě s referenčními spektry, která byla nejblíže k testovanému vzorku. Podle nejpodobnějších spekter bylo možné fázi určit celkem spolehlivě. Obtížnější byla identifikace fáze jen v případě, kdy se testovaný vzorek nacházel na hranici mezi dvěma stádii. Tam bylo určení konkrétního stádia pochopitelně obtížnější, což je logické, protože spektra v těchto místech obsahují znaky obou sousedících stádií.

Druhé, ještě efektivnější, řešení bylo založeno na využití a porovnání virtuálních spekter. Zde jsem vytvořil přímo virtuální spektra odpovídající jednotlivým fázím. Pro každou fázi každé skupiny spekter jsem vytvořil jedno charakteristické virtuální spektrum. Takto bylo získáno osmnáct virtuálních spekter. Ty jsem následně porovnával s testovanými vzorky.

Výsledek byl výborný. Sporné detekce u přechodů mezi fázemi byly výrazně potlačeny a přesnost správné detekce byla výrazně zvýšena. Pokud byl test prováděn současně oběma metodami, byla detekce fáze prakticky bezchybná.

Závěr celého pokusu tedy je, že Biospean lze účinně použít pro detekci konkrétního stádia známé reakce, pokud máme k dispozici referenční vzorky. K identifikaci lze navíc použít dvě na sobě relativně nezávislé metody (porovnávat přímo referenční spektra nebo použít virtuální spektra), které se mohou vhodně doplňovat, což výsledky ještě zpřesňuje.

6.12.2 Identifikace spór rodu Fusarium

Druhým řešeným problémem byla identifikace znaků shodných pro jednotlivé druhy spór rodu *Fusarium*. Tento problém je součástí rozsáhlejšího výzkumu, které pracoviště na TU Vídeň řešilo již dříve [Marchetti-Deschmann at al., 2012].

K dispozici byla změřená spektra různých druhů spór rodu *Fusarium*. Celkem 735 spekter získaných ze 245 vzorků z šesti různých druhů spór. Spektra byla získána z různých lokací a spóry rostly na různých druzích obilnin (ječmen, pšenice, oves), jeden ze vzorků i na vlašských ořeších. Měření měla různou kvalitu, vzorky bohužel obsahovaly poměrně dost šumu a samotná data ze spekter byla k dispozici ve třech variantách s různým způsobem ošetření šumu (bez úpravy, vyhlazený šum, ořez na baseline).

Úkolem tedy bylo najít způsob jak přiřadit testovaný vzorek ke konkrétnímu druhu spór. Jinými slovy najít způsob, jak identifikovat znaky společné pro jednotlivé druhy spór napříč různým prostředím, ve kterém byly získány a z nich vytvořit soubory znaků (virtuální spektra), kterými

by bylo možné spóry identifikovat. Zde se jako velice účinná ukázala funkce, která byla vytvořena pro hledání společných píků pro více spekter zároveň. Jejím rozšířením vznikl generátor virtuálních spekter.

Nicméně při prvních pokusech se příliš nedařilo najít vhodné průniky. Některé skupiny spekter vykazovaly několik podobných znaků pro daný druh spór, ale u jiných skupin byl problém najít vůbec něco. Většinou se dařilo najít podobnosti až poté, co jsem ze skupiny vyjmul nejméně kvalitní spektra. Hlavní problém spočíval v samotných spektrech, konkrétně ve velké úrovni šumu a nedostatečné intenzitě píků. Problém byl zcela zřejmý už ze samotných spekter – dvě spektra ze stejného vzorku změřená dvěma různými studenty vypadala velice odlišně. Proto bylo třeba vymyslet něco, co by dodalo detekci jistou toleranci a robustnost, schopnost vypořádat se i se špatnými daty.

Jako fungující řešení se ukázala úprava funkce pro hledání průniků, která byla poněkud „oslabena“ ve striktnosti vyhledávání. Byl přidán uživatelsky nastavitelný poměr pro detekci, zadávaný jako procentuální hodnota, který umožnil označit i píky, které se vyskytují ve většině spekter, ale nemusí nutně být ve všech. Např. nastavením hodnoty tolerance na 80 % je možné označit všechny píky, které jsou alespoň v 80 % spekter.

Pro otestování identifikace jednotlivých druhů spór jsem tedy vytvořil takováto virtuální spektra se shodami píků 100 %, 80 %, 60 % a 40 %. Pro každý druh spór tedy byly vytvořeny čtyři virtuální spektra, pomocí nichž jsem potom testoval, zda je možné testované spektrum dostatečně přesně rozpoznat.

Výsledky vyhledávání ukázaly, že toto řešení je možné, shoda vzorků proti virtuálním spektrům byla velice dobrá a dařilo se správně identifikovat naprostou většinu rozeznávaných spekter, přičemž virtuální spektra s citlivostí od 60 % výše se ukázala pro detekci konkrétního druhu jako zcela postačující. Problém byl jen se spektry spór, které rostly na vlašských ořeších. Zde se našla jen minimální shoda s ostatními vzorky (které pocházely z obilovin) a identifikace vzorku byla velice obtížná. Zde můžeme pouze konstatovat skutečnost, že kvalitní výsledky vyžadují kvalitní vstupní data a ani sebelepší software nedokáže spočítat správné výsledky ze špatných dat.

Tento experiment prokázal, že Biospean je schopen vyhledávat charakteristické znaky pro celé skupiny spekter, vypořádat se do jisté míry i s nižší kvalitou spekter, vytvářet z takových skupin virtuální spektra charakterizující např. konkrétní druhy organizmů (spór, bakterií) a identifikovat nové vzorky za pomoci těchto uměle vytvořených virtuálních spekter.

6.12.3 Softwarové vyhodnocení tandemových hmotnostních spekter (MS/MS) v rámci studia IgA nefropatie

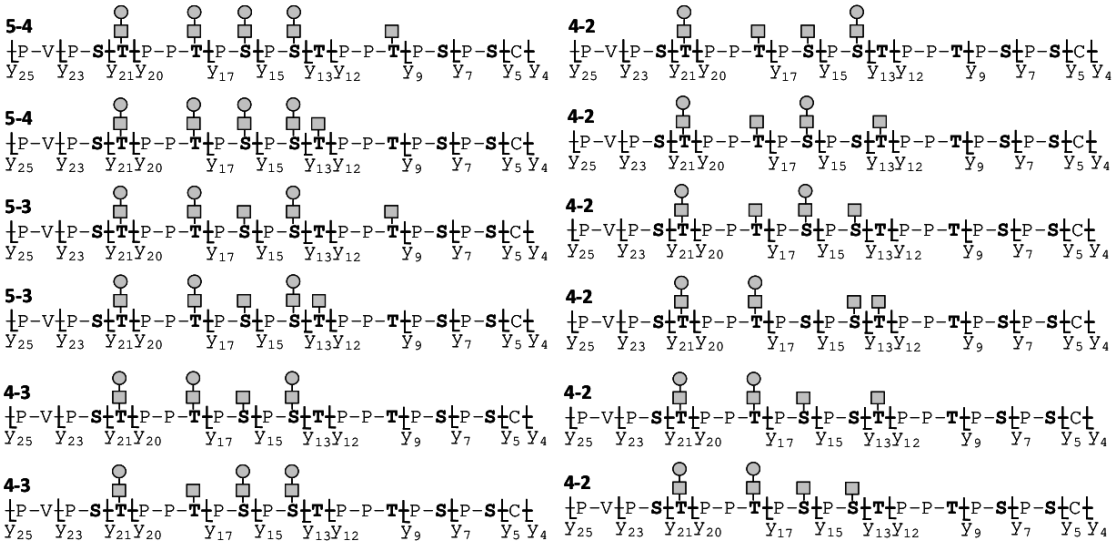
IgA nefropatie, známá také jako Bergerova choroba, je nejčastější primární chronickou glomerulonefritidou na světě a bývá častou příčinou transplantace ledvin. Ačkoliv byla tato nemoc popsána již v roce 1968, její detailní molekulární mechanismus je stále nejasný. Pacienti trpící IgA nefropatií produkují molekuly IgA1, které jsou charakteristické aberantní O-glykosylací v pantové oblasti proteinu. Podstatou této aberance je deficience galaktosy. Takto abnormálně glykosylované molekuly IgA1 jsou rozpoznávány imunitním systémem a tvoří imunokomplexy, které se ukládají v glomerulárním mesangiu ledvin a způsobují jejich poškození.

Výzkumem této choroby se na KBC PřF UP zabývá v rámci své diplomové a navazující disertační práce kolega Mgr. Vojtěch Franc [Franc 2011]. S ním jsem také spolupracoval na řešeném problému.

Mým úkolem bylo vyvinout softwarové řešení interpretace tandemových hmotnostních spekter získaných fragmentací rozdílně O-glykosylovaného tryptického peptidu z pantové oblasti těžkého řetězce IgA1. Zmíněný tryptický peptid obsahoval 12 potenciálních O-glykosylačních míst a vykazoval značnou míru mikroheterogenity. Počet O-glykanů vázaných na jednotlivých isomerech peptidu byl v rozmezí od jednoho až po šest glykanů, přičemž na každém potenciálním místě se mohl vyskytovat monosacharid N-acetylgalaktosamin (GalNAc), disacharid GalNAcGal nebo mohlo být místo neobsazeno. Interpretační problém MS/MS spekter spočíval v tom, že každý jednotlivý fragmentovaný peptid s určitým počtem O-glykanů se skládal z několika isobarických isomerů (peptid se stejnou hmotou, ale rozdílným umístěním O-glykanů), jejichž fragmenty byly obsaženy v jednom fragmentačním spektru.

To představovalo z pohledu manuální interpretace prakticky neřešitelný problém. Na základě manuální intepretace MS/MS spekter expertem byl počet potenciálních míst O-glykosylace zredukován na 6, nicméně počet fragmentů jednotlivých isoforem, které mohly teoreticky vznikat v MS/MS spektrech byl natolik vysoký [Obr 42], že tento analytický problém mohl vyřešit pouze počítač.

Prvním krokem počítačové analýzy byl výpočet všech možných kombinací. Struktura každého isomeru v MS/MS spektru mohla být teoreticky popsána sedmi klíčovými fragmenty. Pomocí skriptu byly vygenerovány všechny kombinace možných isomerů (celkem 262) a u každého z nich bylo vypočteno sedm fragmentů, které vznikly.



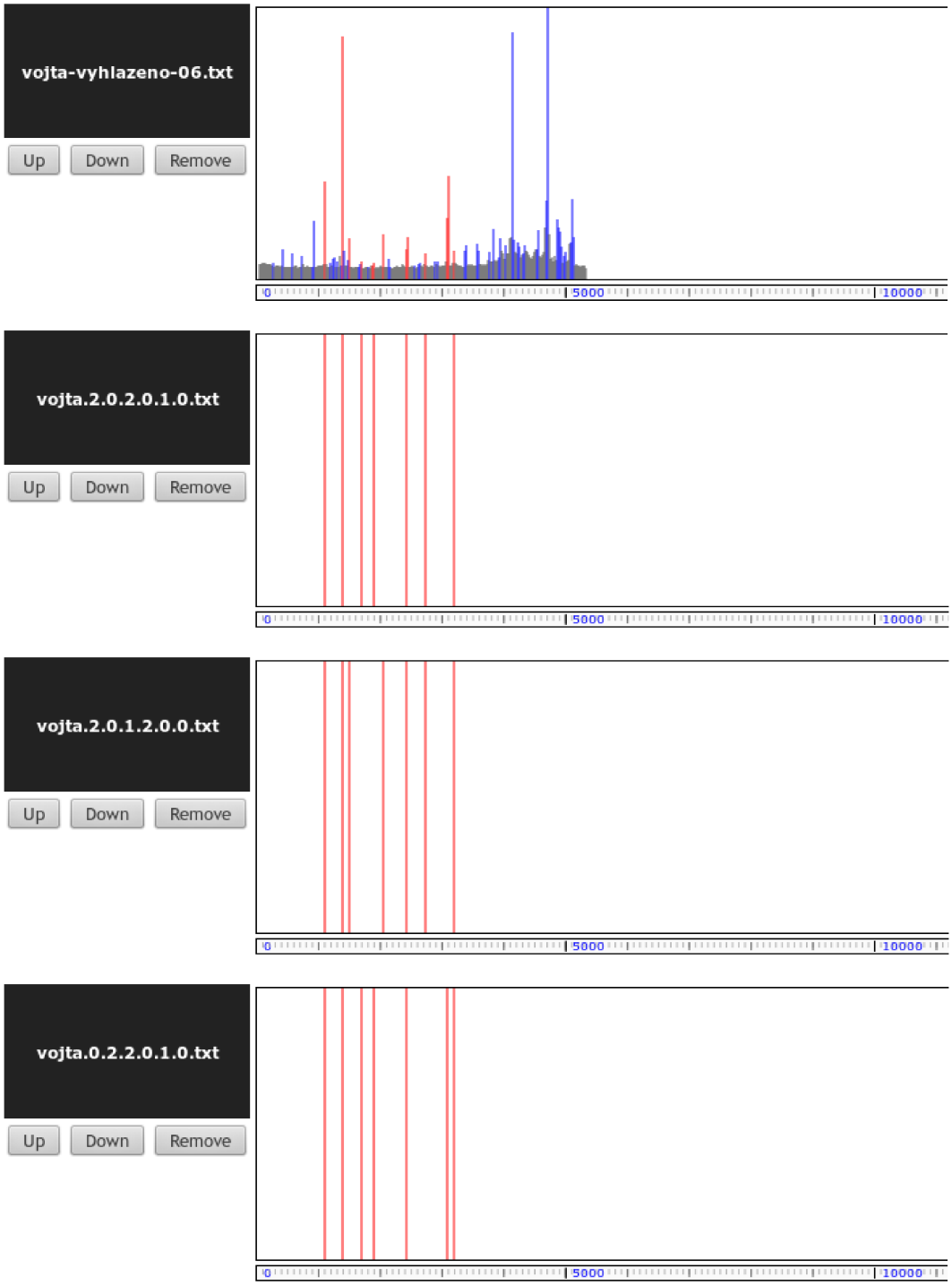
Obr 42: Některé z teoreticky možných isoforem

Na hlavní struktuře se může na celkem šesti různých místech navázat GalNAc (čtvereček) nebo GalNAcGal (čtvereček s kolečkem). Po interpretaci MS/MS spekter expertem, který provedl výběr potenciálně zajímavých kombinací (možné počty GalNAc a GalNAcGal), byl vytvořen počítačový skript, který byl použit pro vygenerování 262 teoreticky možných isoforem. Číselné kódy X-Y (např. 4-2 v druhém sloupečku) značí celkový počet všech vazeb (X) a počet vazeb s GalNAcGal (Y).

Z těchto sedmic byla skriptem vygenerována databáze 262 teoretických MS/MS spekter, přičemž každé obsahovalo sedm fragmentů, které vznikly z příslušného isomerů. Každý z teoretických isomerů byl tedy reprezentováno jedním virtuálním spektrem, které obsahovalo právě sedm virtuálních píků, které odpovídaly sedmi fragmentům vzniklým z rozštěpeného isomeru.

Tato knihovna 262 virtuálních MS/MS spekter byla vložena do programu Biospean a porovnána s 16 reálnými fragmentačními spektry laboratorně změřenými na hmotnostním spektrometru. Tím byl problém převeden na porovnání dvojic spekter, což Biospean dokáže snadno řešit.

Při porovnávání bylo využito pouze jednostranné shody, přičemž se testovalo kolik píků z virtuálního spektra se shoduje s píky ze spektra reálného. Shoda 100 % nebo této hodnotě velmi blízká (pokud přihlédneme k možné numerické odchylce píku o ± 0,5 *m/z*) znamenala, že všech sedm píků z virtuálního spektra má protějšek ve spektru reálném. Pokud software nalezl při porovnání reálného spektra a virtuálního spektra všech sedm klíčových fragmentů [Obr 43], které odpovídaly teoretické isoformě, byla existence této varianty přijata jako velmi pravděpodobná.



Obr 43: Porovnání reálného spektra s virtuálními spektry obsahujícími fragmenty

Obrázek demonstruje graficky shodu na pozicích píků mezi reálným spektrem (nahoře) a třemi virtuálními spektry, které reprezentují teoreticky vypočtené fragmenty. Každé z virtuálních spekter reprezentuje jeden z teoreticky možných isomerů a obsahuje sedm fragmentů, které by vznikly jeho rozštěpením. Na obrázku z Biospeanu je vidět, že v reálném spektru se současně vyskytují fragmenty ze všech tří porovnávaných isomerů, tj. zdrojem fragmentů není jen jediná isoforma forma.

Kódové označení virtuálních spekter obsahuje šestici čísel, přičemž každé z čísel reprezentuje jedno z možných vazebných míst. Hodnota čísel označuje glykan navázaný na daném místě (0 – nic, 1 – GalNAc, 2 – GalNAcGal).

Protože manuální porovnávání všech 16 reálných spekter s 262 virtuálními nebylo příliš přehledné, vytvořil jsem speciální skript, který využil interní funkce Biospeanu a pomocí něho jsem vygeneroval přehlednou tabulku o 16 sloupcích (reálná spektra) a 262 řádcích (virtuální spektra), kde bylo poměrně snadné nalézt podobnost jednotlivých reálných spekter se spektry virtuálními.

Tato přehledná tabulka už kolegovi umožnila snadno dokončit analýzu celého problém. Z výsledků vyplynulo, že v reálné situaci se může objevit jen část z teoreticky možných kombinací isoforem.

Výsledky tohoto výzkumu jsou publikovány v článku *Elucidating heterogeneity of IgA1 hinge-region O-glycosylation by use of MALDI-TOF/TOF mass spectrometry: role of cysteine alkylation during sample processing*, který je **Přílohou E** této práce.

6.13 Shrnutí

Software Biospean ukázal při testech, že je použitelný pro účely, ke kterým byl vytvořen. Testy potvrdily schopnost porovnávat spektra na základě pozic píků. Výsledky pokusů s detekcí stavu fermentace také ukázaly, že porovnávání je v pořádku jak po kvantitativní, tak po kvalitativní stránce – jednotlivá spektra vykazovala největší podobnost právě s časově sousedními spektry.

Zajímavé bylo u použití Biospeanu způsobem, který nebyl při jeho vytváření zamýšlen. Především se jedná o využití virtuálních spekter. A to jak při detekci fáze reakce pomocí MS spektra a při hledání společných znaků ve spektrech spór Fusarium, tak především při analýze spekter IgA nefropatie. Využití generátoru virtuálních spekter u prvních dvou a především externího skriptu v případě IgA pro vytvoření virtuálních spekter, která byla potom Biospeanem analyzována, demonstrovalo schopnost mnohem univerzálnějšího využití Biospeanu, než bylo z počátku zamýšleno.

6.14 Publikace

Poznatky získané využitím softwaru Biospean byly v době odevzdání této práce zaslány k recenzi v podobě článků:

- Franc V, Řehulka P, **Raus M**, Stulik J, Novak J, Renfrow MB, Šebela M. *Elucidating heterogeneity of IgA1 hinge-region O-glycosylation by use of MALDI-TOF/TOF mass spectrometry: role of cysteine alkylation during sample processing.* J. Proteomics

  revidovaná verze byla zaslána k recenzi 19. června 2013

- Chalupová J, **Raus M**, Šebela M. *Identification of fungal microorganisms by MALDI-TOF mass spectrometry.* Biotechnology Advances

  k recenzi zasláno 18. června 2013

7 Závěr

Předložená disertační práce je zaměřena na využití informatiky k řešení úloh z oblasti hmotnostní spektrometrie, resp. na vytvoření vhodných softwarových nástrojů. Řešeny byly úkoly ze tří specifických oblasti – teoretická predikce štěpení proteinů (ProteinCutter), výpočet rychlosti biosyntézy cytokininů (Izotop) a analýza a porovnání hmotnostních spekter (Biospean).

První z řešených úloh byla **teoretická predikce štěpení proteinů**. Požadavek zněl na schopnost teoreticky vypočítat na základě předloženého proteinového řetězce a znalosti pravidel štěpení (enzym) všechny možné produkty štěpení a vypočíst jejich chemicko-fyzikální vlastnosti. Tyto vypočítané výsledky měly být porovnávány s výsledky z laboratorních měření a být použity k identifikaci konkrétních štěpů na základě jejich hmotnosti.

Všechny tyto požadavky velmi úspěšně splnila aplikace ProteinCutter. Výsledná aplikace v jednotném a snadno použitelném rozhraní spojuje schopnosti, které byly dosud dostupné jen při použití několika různých aplikací. Navíc dodává i schopnosti, které jiné aplikace nemají (konkrétní štěpící enzymy, možnost zvolit si vlastní pravidla štěpení, simulace vynechaných štěpení) a přidává přesnost fyzikálně-chemických veličin. Přesnost a úplnost predikovaných štěpů byla ověřena na zkoumaných laboratorních výsledcích a praktické použití ukázalo, že jde o užitečný výzkumný nástroj odpovídající moderním požadavkům.

Aplikaci ProteinCutter tedy můžeme označit za úspěšnou a v podstatě i kompletně pokrývající potřeby, protože ani z hlediska uživatelů nejsou další požadavky na vylepšení, úpravy či rozšíření.

Druhým řešeným úkolem byl nástroj pro **výpočet rychlosti biosyntézy**, konkrétně cytokininů, z laboratorně získaných dat (metodou *deuterium in vivo labeling*). Požadavkem bylo vylepšení nevyhovujícího řešení, které využívalo tabulku v tabulkovém kalkulátoru a které sice bylo správné, ale pomalé na praktické použití. Primárním požadavkem pak byla možnost zpracovat data hromadně, což bylo u předchozího řešení neproveditelné.

K tomuto účelu byla vytvořena webová aplikace Izotop. Z hlediska funkčnosti nebo metody výpočtu nepřináší Izotop nic významného, nicméně jeho přínos spočívá v zobecnění a v univerzálnějším řešení, které umožňuje pracovat efektivněji. Hlavními přínosy je tedy možnost zpracovávat data snadno, hromadně a provádět přímý import vstupů a export výsledků pomocí souborů.

Ačkoliv samotná aplikace je relativně jednoduchá a vznikala k jednorázovému účelu, splnila účel, pro který byla vytvořena a je zde uvedena z důvodu publikovaného článku.

Třetím řešeným úkolem bylo vytvoření softwaru pro **analýzu a porovnání hmotnostních spekter**. Primární cíl byl poskytnout alternativu ke komerčnímu softwaru jako je např. Biotyper. Software měl být schopen analyzovat hmotnostní spektrum, označit píky a porovnat toto spektrum se vzorky dalších spekter v databázi.

Výsledkem je aplikace Biospean. Všechny požadované schopnosti se podařilo realizovat a navíc doplnit řadu dalších, které výrazně zvýšily užitnou hodnotu této aplikace. Z významných rozšíření je třeba zdůraznit schopnost práce s virtuálními spektry (jak jejich generování, tak zpracování), schopnost porovnávat velký počet spekter a hledat znaky společné pro všechna zvolená spektra. Významnou pomocí jsou i organizační nástroje. Každý uživatel může svá spektra členit do kategorií, provádět porovnání jen nad určitými skupinami, ale navíc i svoje spektra sdílet se svými kolegy. To umožňuje použití Biospeanu jako nástroje pro skupinovou práci výzkumného týmu.

Ačkoliv je Biospean v době odevzdání této práce stále ještě v testovacím provozu a probíhá shromažďování zkušeností z praktického používání, je možné ho označit za úspěšný minimálně po stránce splnění požadovaných cílů. Úkoly, které na něm byly během testů řešeny podaly poměrně přesvědčivé důkazy o tom, že aplikace funguje opravdu správně (detekce stádia reakce). Navíc se ukázalo (např. při studiu spekter z IgA nefropatie), že aplikace má potenciál i pro méně konvenční způsoby využití a rezervy pro další evoluci.

I v případě Biospeanu lze tedy prohlásit, že aplikace úspěšně splnila požadavky. Navíc umí i to, co v původním návrhu zamýšleno nebylo. Já sám bych tuto aplikaci nechtěl ani zdaleka prohlásit za uzavřenou kapitolu v okamžiku, kdy bude tato disertace odevzdána a obhájena. Není důvod nepokračovat v dalším vývoji. Aplikace má slibný potenciál pro další rozvoj do budoucna. Rozhodně by neměla skončit jako položka v univerzitní knihovně, ale měla by být dále využívána jako podpůrný nástroj pro další výzkum.

8 Seznam použitých zkratek

AJAX	*Asynchronous JavaScript and XML* - Skupina technologií, která umožňuje vytvářet webové stránky tak, aby se chovaly jako běžné aplikace, tj. bez nutnosti znovunačtení
Apache	Nejrozšířenější HTTP server, aplikace, která poskytuje webové stránky - návštěvník posílá ze svého webového prohlížeče příkaz k zobrazení webové stránky (umístěné na konkrétním serveru) a Apache (či jiný HTTP server) stránku sestaví a odešle do webového prohlížeče.
ASCII	*American Standard Code for Information Interchange* - Standardizovaná kódovací tabulka se znaky anglické abecedy.
BSD	*Berkeley Software Distribution* - Operační systém Unixového typu. Dnes je základem pro celou rodinu operačních systémů (FreeBSD, NetBSD, OpenBSD, Mac OS X, Solaris).
CSS	*Cascading Style Sheets* je jazyk pro zápis formátovacích stylů pro webové stránky.
CSV	*Comma-Separated Values* je jednuduchý textový formát pro zápis dat ve formě tabulky. Jednotlivé položky v řádku jsou odděleny čárkou, řádky buněk jsou umístěné na samostatný textový řádek.
HTML	*HyperText Markup Language* je značkovací jazyk pro hypertextové dokumenty, dnes využívaný ve formě webových stránek.
ISO	Veličina udávající citlivost snímače digitálních optických přístrojů (foťáky mobilů, fotoaparáty, digitální zrcadlovky, digitální videokamery).
LAMP	Zkratka pro nejrozšířenější platformu pro provoz webových stránek - Linux, Apache, MySQL, PHP (někdy uváděn i Perl nebo Python)
MALDI	*Matrix Assisted Laser Desorption Ionization* - ionizace laserem za přítomnosti matrice využívaná u hmotnostních spektrometrů typu MALDI-TOF
OO/LO Calc	*OpenOffice / LibreOffice Calc* - tabulkový kalkulátor který je přímým konkurentem Microsoft Office Excelu.

PHP	*PHP: Hypertext Preprocessor* - skriptovací interpretovaný jazyk vycházející z jazyka Perl, který je v současnosti nejpoužívanějším nástrojem pro tvorbu dynamických webových stránek.
SQL	*Structured Query Language* - standardizovaný dotazovací jazyk pro práci s databázovými systémy. Řada databázových systémů obsahuje jméno tohoto jazyka - MySQL, PostgreSQL, MSSQL...
TOF	*Time of Flight* - detekce doby letu používáná v hmotnosních spektrometrech MALDI-TOF ve spolupráci s ionizací MALDI
XHTML	*eXtensible HyperText Markup Language* - Značkovací jazyk pro webové stránky velmi podobný jazyku HTML, je ovšem striktnější v dodržování konvencí jazyka XML
XML	*Extensible Markup Language* - Velice obecný značkovací jazyk, který je ve standardizované podobě vhodný pro uchování a výměnu dat

9 Životopis

Mgr. Martin Raus

Trvalé bydliště	Pod Strážnicí 412, Velké Opatovice 679 63
E-mail	martin_raus@post.cz
Datum narození	28. října 1982
Rodinný stav	svobodný

Vzdělání

2008 – dnešek	doktorské studium, obor Biochemie, KBC PřF UP v Olomouci disertační práce: *Softwarová analýza a predikce v hmotnostní spektrometrii*
srpen 2008	dosažen **magisterský titul** v oboru Informatika diplomová práce: *Webová hra pro mnoho hráčů*
srpen 2006	dosažen **bakalářský titul** v oboru Informatika bakalářská práce: *Rozšiřující modul programu na vizualizaci gramatik a automatů*
2002 – 2008	magisterské studium, obor Informatika, KMI PřF UP v Olomouci
květen 2002	složena **maturitní zkouška**
1998 – 2002	čtyřleté gymnázium, obor s rozšířenou výukou informatiky, Gymnázium Jevíčko
1989 – 1998	Základní škola Velké Opatovice.

Pracovní zkušenosti

V rámci doktorského studia jsem absolvoval tříměsíční studijní stáž na Fakultě technické chemie, TU Vídeň, Rakousko v období listopad – prosinec 2012 a únor – duben 2013.

2008 – 2012	učitel informatiky a správce sítě, Gymnáziu Jevíčko
2007 – současnost	výuka předmětu ZPP (základy práce s počítačem) na Katedře biochemie PřF UP

Publikace

publikované články z oblasti biochemie/bioinformatiky

- Chalupová J, **Raus M**, Šebela M. Identification of fungal microorganisms by MALDI-TOF mass spectrometry. Biotechnology Advances
  v recenzi (zasláno 18.6. 2013)
- Franc V, Řehulka P, **Raus M**, Stulik J, Novak J, Renfrow MB, Šebela M. Elucidating heterogeneity of IgA1 hinge-region O-glycosylation by use of MALDI-TOF/TOF mass spectrometry: role of cysteine alkylation during sample processing. J. Proteomics, *v recenzi (revidovaná verze zaslána 19.6. 2013)*
- **Raus M.**, Kopečný D., Šebela M. (2013) Program application for the prediction of results of protein digestion by proteolytic enzymes [Programová aplikace pro predikci výsledků štěpení proteinu proteolytickými enzymy]. Chem. Listy 107 (1), 44-53.
- Tarkowski P., Flokova K., Václavikova K., Jaworek P., **Raus M.**, Nordström A., Novak O., Dolezal K., Sebela M., Frebortova J. (2010) An improved in vivo deuterium labeling method for measuring the biosynthetic rate of cytokinins. Molecules 15 (12), 9214-9229.
- Sebela M., Rehulka P., Kabrt J., Rehulkova H., Ozdian T., **Raus M.**, Franc V., Chmelik J. (2009) Identification of N-glycosylation in prolyl endoprotease from Aspergillus niger and evaluation of the enzyme for its possible application in proteomics. J. Mass Spectrom. 44 (11), 1587-1595

publikované články z oblasti informatiky

- **Raus M.**, Korbel M., Lór L. (2010) Gymnázium Jevíčko: S open source to jde lépe [online] http://www.linuxexpres.cz/business/gymnazium-jevicko-linux-ve-skole [26. 6. 2013]
- **Raus M.**, Korbel M., Lór L. (2010) Gymnázium Jevíčko používá OpenOffice.org [online] http://www.openoffice.cz/pouzivaji-openoffice-org/gymnazium-jevicko-pouziva-openoffice-org [26. 6. 2013]

Příspěvky na konferencích

- Franc V, Řehulka P, **Raus M**, Stulík J, Novak J, Renfrow MB, Šebela M. Analysis of heterogeneous hinge-region O-glycosylation of human IgA1 using MALDI-TOF/TOF mass spectrometry. 11th Discussions in Structural Biology, Academic and University Center, Nové Hrady, March 14-16, 2013. Lecture L4 by M. Šebela on 14th March. Abstract in: Materials Structure in Chemistry, Biology, Physics and Technology, Bulletin of the Czech and Slovak Crystallographic Association 2013, 20 (1):5
- Franc V, Řehulka P, **Raus M**, Stulík J, Šebela M, Novak J. MALDI-TOF/TOF mass spectrometry study of O-glycopeptides from the hinge region of human IgA1: role of precise cysteine alkylation during sample processing. MassSpec-Forum-Vienna-2013, 24th Mass

Spectrometric Discussion Meeting, February 19-20, 2013, University of Vienna, Austria. Lecture by V. Franc on 19th February 2013. Abstract in: Book of Abstracts, p. 30

- Šebela M, Řehulka P, Kábrt J, **Raus M**, Řehulková H, Oždian T, Franc, Chmelík J. Prolylendoproteasa: nový nástroj pro analýzu proteinů (Prolyl endoprotease: a new type of chainsaw for cutting proteins). 11. Škola hmotnostní spektrometrie (11th School of Mass Spectrometry), September 20-24, 2010, hotel Horizont, Pec pod Sněžkou, CZ. Lecture on 23rd September by M. Šebela; published in Sborník (Anthology), Holčapek M, Lísa M, Jirásko R, Eds.
- Sebela M, Rehulka P, Kabrt J, **Raus M**, Rehulkova H, Ozdian T, Franc V, Chmelik J. Identification of N-glycosylation in prolyl endoprotease from Aspergillus niger and evaluation of the enzyme for its possible application in proteomics. 18th International Mass Spectrometry Conference, August 30 – September 4, 2009, Bremen, Germany. Poster PMM-241 by M. Šebela. Program, p. 71
- Šebela M, Řehulka P, Kábrt J, Řehulková H, Oždian T, **Raus M**, Franc V, Chmelík J. Identification of N-glycosylation in prolyl endoprotease from Aspergillus niger and evaluation of the enzyme for its possible application in proteomics. Cukrblik 2009: Current Chemistry and Biochemistry of Carbohydrates, April 16, 2009. Masaryk University, Brno, Czech Republic; lecture by M. Šebela. Abstract in: Abstract Book, Masaryk University, Brno, p. 18, ISBN 978-80-210-4831-7

Další znalosti, schopnosti a dovednosti

jazykové schopnosti	čeština (rodilý mluvčí) angličtina (aktivně i pasivně, pokročilý)
počítačové znalosti	vystudován obor informatika specializace na webové technologie, tvorbu webů, webdesign... programovací jazyky PHP, Java, C/C++, Scheme, Python, SQL, HTML a další operační systémy GNU/Linux a Windows administrace serverů (webové, databázové...) a sítí a mnoho dalšího...
řidičské oprávnění	A, B
záliby	letecká a vojenská technika, počítače, fotografování, historie 20. století, exaktní vědy, literatura faktu, válečná a sci-fi literatura

10 Literatura a zdroje

[42FLSS]
42 of the Best Free Linux Scientific Software (2013) [online]
http://www.linuxlinks.com/article/20080803104017665/Scientific.html [staženo 26.6.2013]

[Altschul 2011]
Altschul S.F. (2011) Global and Local Sequence Alignment [online]
http://www.cs.umd.edu/class/fall2011/cmsc858s/Alignment.pdf [staženo 26.6.2013]

[ANR]
Texas Instruments (2013) LM1894(NRND) Dynamic Noise Reduction System DNR [online]
http://www.ti.com/product/LM1894 [staženo 26.6.2013]

[ASCII]
ASCII Table and Description [online] http://www.asciitable.com/ [staženo 26.6.2013]

[Baum 2008]
Baum D. (2008) Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups [online]
http://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956
[staženo 26.6.2013]

[Bioinformatics]
Bioinformatics Organization – Bioinformatics.Org [online] http://www.bioinformatics.org/
[staženo 26.6.2013]

[BioinformaticsWiki]
Bioinformatics Organization - Bioinformatics.Org Wiki [online]
http://www.bioinformatics.org/wiki/Bioinformatics [staženo 26.6.2013]

[BioLi]
Bio-Linux [online] http://envgen.nox.ac.uk/tools/bio-linux [staženo 26.6.2013]

[BioPerl]
BioPerl [online] http://www.bioperl.org/wiki/Main_Page [staženo 26.6.2013]

[BioPython]
Biopython [online] http://biopython.org/wiki/Main_Page [staženo 26.6.2013]

[BioRuby]
BioRuby [online] http://www.bioruby.org/ [staženo]

[Biotyper]
Bruker MALDI Biotyper [online]
http://www.bruker.com/en/products/mass-spectrometry-and-separations/maldi-Biotyper
[staženo 26.6.2013]

[BLAST homapage]
BLAST: Basic Local Alignment Search Tool [online] http://blast.ncbi.nlm.nih.gov/Blast.cgi
[staženo 26.6.2013]

[BSD]
BSD.org [online] http://www.bsd.org/ [staženo 26.6.2013]

[BW-ov]
Molecular biology database/biological database/bioinformatics database - an overview [online]
http://bioinformaticsweb.net/data.html [staženo 26.6.2013]

[BWord 2006]
Houser P., Cvrčková F. (2006) Bioinformatika - na půl cestě mezi algoritmy a životem [online] http://businessworld.cz/rozhovory/bioinformatika-na-pul-ceste-mezi-algoritmy-a-zivotem-3820 [staženo 26.6.2013]

[C&CPP]
C Programming and C++ Programming [online] http://www.cprogramming.com/ [staženo 26.6.2013]

[CERNLi]
CERN Linux [online] http://linux.web.cern.ch/linux/ [staženo 26.6.2013]

[CIF]
The Crystallographic Information File (CIF): a New Standard Archive File for Crystallography [online] http://www.iucr.org/__data/iucr/cif/standard/cifstd1.html [staženo 26.6.2013]

[Clustal homepage]
Clustal homepage [online] http://www.clustal.org/ [staženo 26.6.2013]

[COOT]
Crystallographic Object-Oriented Toolkit [online] http://www.biop.ox.ac.uk/coot/ [staženo 26.6.2013]

[CSD]
Cambridge Structural Database [online] http://www.ccdc.cam.ac.uk/pages/Home.aspx [staženo 26.6.2013]

[CTF] Symyx Solutions, Inc. (2005) CT File Formats, Symyx Solutions, Inc. [online] http://c4.cabrillo.edu/404/ctfile.pdf [staženo 26.6.2013]

[Cvrčková 2006a]
Cvrčková F. (2006) Úvod do praktické bioinformatiky, str. 38-46, Academia, Praha

[Cvrčková 2006b]
Cvrčková F. (2006) Úvod do praktické bioinformatiky, str. 42-46, Academia, Praha

[Cvrčková 2006c]
Cvrčková F. (2006) Úvod do praktické bioinformatiky, str. 25-34, Academia, Praha

[Cvrčková 2006d]
Cvrčková F. (2006) Úvod do praktické bioinformatiky, str. 25, Academia, Praha

[Cvrčková 2006e]
Cvrčková F. (2006) Úvod do praktické bioinformatiky, str. 26, Academia, Praha

[Cvrčková 2006f]
Cvrčková F. (2006) Úvod do praktické bioinformatiky, str. 108-125, Academia, Praha

[CzWiki Bi]
Bioinformatika [online] http://cs.wikipedia.org/wiki/Bioinformatika [staženo 26.6.2013]

[DDBJ]
DDBJ - DNA Data Bank of Japan [online] http://www.ddbj.nig.ac.jp/ [staženo 26.6.2013]

[Debian]
Debian - Univerzální operační systém [online] http://www.debian.org/ [staženo 26.6.2013]

[DEN1]
What does the dendrogram show, or what is correlation analysis? [online]
http://www.nonlinear.com/support/progenesis/samespots/faq/dendrogram.aspx
[staženo 26.6.2013]

[DEN2]
Dendrograms and Clustering [online]
http://stn.spotfire.com/spotfire_client_help/heat/heat_dendrograms_and_clustering.htm
[staženo 26.6.2013]

[DEN3]
How Dendrogram works [online]
http://edndoc.esri.com/arcobjects/9.2/net/shared/geoprocessing/spatial_analyst_tools/how_dendrogram_works.htm [staženo 26.6.2013]

[DNR]
Dynamic Noise Reduction [online] http://www.compolinc.com/dynamic.htm
[staženo 26.6.2013]

[EMBL]
EMBL-EBI [online] http://www.ebi.ac.uk/ [staženo 26.6.2013]

[EMBL-formát]
European Bioinformatics Institute: EMBL (2013) User Manual [online]
ftp://ftp.ebi.ac.uk/pub/databases/embl/doc/usrman.txt [staženo 26.6.2013]

[ExPASy]
ExPASy: SIB Bioinformatics Resource Portal [online] http://www.expasy.org/
[staženo 26.6.2013]

[FASTA 1]
What is FASTA format? [online] http://zhanglab.ccmb.med.umich.edu/FASTA/
[staženo 26.6.2013]

[FASTA 2]
Query Input and database selection [online]
http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml [staženo 26.6.2013]

[Fasta homepage]
FASTA Sequence Comparison at the U. of Virginia [online]
http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml [staženo 26.6.2013]

[FeLi]
Fermi Linux [online] http://fermilinux.fnal.gov/ [staženo 26.6.2013]

[Fklec]
Faradayova klec [online] http://kdf.mff.cuni.cz/vyuka/elektrodynamika/doku.php?id=experimenty:faradayova_klec [staženo 26.6.2013]

[Franc 2011]
Franc V. (2011) Diplomová práce – Studium proteinových biomarkerů a posttranslačních modifikací pomocí hmotnostní spektrometrie [online]
http://theses.cz/id/lo1nla/Franc_diplomova_prace_zkracena_2011.pdf [staženo 26.6.2013]

[Fürbach 2008]
Fürbach M. (2008) Uříznout si bříška prstů nestačí. Vzor pro otisky se vám vrátí [online]
http://goo.gl/QMbJH [staženo 26.6.2013]

[GenBank]
GenBank [online] http://www.ncbi.nlm.nih.gov/genbank/ [staženo 26.6.2013]

[GiNR]
Fix hot pixel noise in a photo or image with Gimp [online]
http://gimps.de/en/tutorials/gimp/picture-photo-image/pixel-noise/ [staženo 26.6.2013]

[GRV]
Jak Google využívá rozpoznávání vzorů [online]
http://www.google.com/intl/cs/policies/technologies/pattern-recognition/ [staženo 26.6.2013]

[Gunnlaugsson 2012]
Gunnlaugsson Ó. (2012) Noise Reductions Systems [online] http://audiotools.com/noise.html
[staženo 26.6.2013]

[Harris]
Harris W.: How Noise-canceling Headphones Work [online]
http://electronics.howstuffworks.com/gadgets/audio-music/noise-canceling-headphone3.htm
[staženo 26.6.2013]

[Havliš 1999]
Havliš J. (1999) Hmotnostní spektrometrie MALDI TOF [online]
http://www.vesmir.cz/clanek/hmotnostni-spektrometrie-maldi-tof [staženo 26.6.2013]

[Havránek at al., 2007]
Havránek P., Žejdlík T., Novák P. (2007) Systematické rušení české meteorologické radarové sítě CZRAD zařízeními RLAN [online]
http://pocasi.chmi.cz/meteo/olm/Let_met/_tmp/ruseni_radaru.htm [staženo 26.6.2013]

[Hlaváč 2010]
Hlaváč V. (2010) Předzpracování obrazu v lokálním okolí [online]
http://cmp.felk.cvut.cz/~hlavac/TeachPresCz/11DigZprObr/21ImagPreprocCz.pdf
[staženo 26.6.2013]

[Holčapek 2013]
Holčapek M. (2013) Hmotnostní spektrometrie [online]
http://holcapek.upce.cz/teaching/Mol_spek_2013/Mol_spek_prednaska6_MS.pdf
[staženo 26.6.2013]

[IBMdb]
IBM DB2 [online] http://www-03.ibm.com/software/products/us/en/category/SWB00
[staženo 26.6.2013]

[INSDC]
International Nucleotide Sequence Database Collaboration [online] http://www.insdc.org/
[staženo 26.6.2013]

[intJ]
Interpretovaný jazyk [online] http://python.wraith.cz/uvod-interpret.php [staženo 26.6.2013]

[ItBD]
Introduction toBiological Databases [online]
http://www.iasri.res.in/ebook/win_school_aa/notes/Biological_Databases.pdf
[staženo 26.6.2013]

[Java]
What is Java and why do I need it? [online]
http://www.java.com/en/download/faq/whatis_java.xml [staženo 26.6.2013]

[Jmol]
Jmol: an open-source Java viewer for chemical structures in 3D [online]
http://jmol.sourceforge.net/ [staženo 26.6.2013]

[jQuery]
jQuery [online] http://jquery.com/ [staženo 26.6.2013]

[Kanade 1977]
Kanade T. (1977) Computer recognition of human faces [online]
http://www.ri.cmu.edu/pub_files/pub3/kanade_takeo_1977_1/kanade_takeo_1977_1.pdf
[staženo 26.6.2013]

[König 2010]
König M. (2010) Biological Databases [online]
http://www.charite.de/sysbio/people/koenig/presentation/100505_koenig_biological_databases.
pdf [staženo 26.6.2013]

[Ledvoň 2009]
Ledvoň J. (2009) MOZEK - nejsložitější a nejorganizovanější hmota ve známém vesmíru…
[online] http://files.mzsb.cz/200006679-4ccdf4dc7f/Mozek%20a%20mysl1.pptx
[staženo 26.6.2013]

[Lindeberg 2001]
Lindeberg T. (2001) Edge detection, Encyclopedia of Mathematics [online]
http://www.encyclopediaofmath.org/index.php?title=Edge_detection [staženo 26.6.2013]

[Liu 2011]
Liu W. (2011) CUDA-BLASTP: accelerating BLASTP on CUDA-enabled graphics hardware.
[online] http://novatte.com/gtc_2011/gtcpresentations/weiguo_liu.pdf [staženo 26.6.2013]

[LoBD]
List of biological databases [online] http://en.wikipedia.org/wiki/List_of_biological_databases
[staženo 26.6.2013]

[LoSAS]
List of sequence alignment software [online]
http://en.wikipedia.org/wiki/List_of_sequence_alignment_software [staženo 26.6.2013]

[Maple]
Maple 17 - Technical Computing Software for Engineers, Mathematicians, Scientists,
Instructors and Students [online] http://www.maplesoft.com/products/Maple/
[staženo 26.6.2013]

[Marchetti-Deschmann at al., 2012]
Marchetti-Deschmann M., Winkler W., Dong H., Lohninger H., Kubicek Ch. P., Allmaier G.
(2012) Using Spores for Fusarium spp. Classification by MALDI-Based Intact Cell/Spore Mass
Spectrometry [online] http://hrcak.srce.hr/file/129390 [staženo 26.6.2013]

[MariaDB]
MariaDB - An enhanced, drop-in replacement for MySQL [online] https://mariadb.org/
[staženo 26.6.2013]

[MASCOT]
Mascot search engine | Protein identification software for mass spec data [online]
http://www.matrixscience.com/ [staženo 26.6.2013]

[MATLAB]
MATLAB - The Language of Technical Computing [online]
http://www.mathworks.com/products/matlab/ [staženo 26.6.2013]

[MEGA homepage]
MEGA :: Molecular Evolutionary Genetics Analysis [online] http://www.megasoftware.net/ [staženo 26.6.2013]

[mMass]
mMass - Open Source MS Tool [online] http://www.mmass.org/ [staženo 26.6.2013]

[Molnár et al., 1987]
Molnár Ľ., Češka M., Melichar B. (1987) Gramatiky a jazyky, Alfa, Bratislava

[MSDig]
MS-Digest [online] http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msdigest [staženo 26.6.2013]

[MSS]
Mass spectrometry software [online] http://en.wikipedia.org/wiki/Mass_spectrometry_software [staženo 26.6.2013]

[MTOF]
MALDI-TOF [online]
http://chemwiki.ucdavis.edu/Analytical_Chemistry/Instrumental_Analysis/Mass_Spectrometry/MALDI-TOF [staženo 26.6.2013]

[MySQL]
MySQL :: The world's most popular open source database [online] http://www.mysql.com/ [staženo 26.6.2013]

[NAfE5]
Harder D.W.: Numerical Analysis for Engineering: 5 Interpolation [online]
https://ece.uwaterloo.ca/~dwharder/NumericalAnalysis/05Interpolation/ [staženo 26.6.2013]

[NAfE6]
Harder D.W.: Numerical Analysis for Engineering: 6 Least Squares [online]
https://ece.uwaterloo.ca/~dwharder/NumericalAnalysis/06LeastSquares/ [staženo 26.6.2013]

[NDB]
Nucleic Acid Database [online] http://ndbserver.rutgers.edu/ [staženo 26.6.2013]

[NRL-3D]
PIR-NRL3D Sequence-Structure Database [online]
http://pbil.univ-lyon1.fr/pf_bioinfo/article203.html [staženo 26.6.2013]

[NVIDIA]
Bioinformatics and Life Sciences [online]
http://www.nvidia.com/object/bio_info_life_sciences.html [staženo 26.6.2013]

[Octave]
GNU Octave [online] http://www.gnu.org/software/octave/ [staženo 26.6.2013]

[Odb]
Oracle Database [online] http://www.oracle.com/cz/products/database/overview/index.html [staženo 26.6.2013]

[Opperdoes 1997]
Opperdoes F. (1997) Construction of a distance tree using clustering with the Unweighted Pair Group Method with Arithmatic Mean (UPGMA) [online]
http://www.icp.ucl.ac.be/~opperd/private/upgma.html [staženo 26.6.2013]

[OS4SA]
Online Software for Sequence Analysis [online]
http://www.science.co.il/biomedical/sequence-tools.asp [staženo 26.6.2013]

[OWL]
OWL [online] http://www.bioinf.man.ac.uk/dbbrowser/OWL/index.php [staženo 26.6.2013]

[PCalc]
Protein Calculator [online] http://www.scripps.edu/~cdputnam/protcalc.html
[staženo 26.6.2013]

[PCutter]
PeptideCutter [online] http://web.expasy.org/peptide_cutter/ [staženo 26.6.2013]

[PDB]
Protein Data Bank [online] http://www.wwpdb.org/ [staženo 26.6.2013]

[PDBdoc]
Protein Data Bank – Documentation [online] http://www.wwpdb.org/docs.html
[staženo 26.6.2013]

[Pearson & Lipman 1987]
Pearson W.R., Lipman D.J. (1987) Improved tools for biological sequence comparison [online]
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC280013/pdf/pnas00260-0036.pdf
[staženo 26.6.2013]

[Perl]
The Perl Programming Language - www.perl.org [online] http://www.perl.org/
[staženo 26.6.2013]

[PgSQL]
PostgreSQL: The world's most advanced open source database [online]
http://www.postgresql.org/ [staženo 26.6.2013]

[Phenix]
Phenix [online] http://www.phenix-online.org/ [staženo 26.6.2013]

[PHP]
PHP: Hypertext Preprocessor [online] http://php.net/ [staženo 26.6.2013]

[phpMyGraph]
Martijn Beulens (2011) phpMyGraph [online] http://phpmygraph.abisvmm.nl/
[staženo 26.6.2013]

[Phylo1]
What is Phylogeny? [online] http://tolweb.org/tree/learn/concepts/whatisphylogeny.html
[staženo 26.6.2013]

[Phylo2]
How to read a phylogenetic tree [online]
http://archosaurmusings.wordpress.com/2008/12/18/how-to-read-a-phylogenetic-tree/
[staženo 26.6.2013]

[Phylo3]
Learn about the Tree of Life! [online]
http://archive.peabody.yale.edu/exhibits/treeoflife/learn.html [staženo 26.6.2013]

[Pietrokovski 1999]
Pietrokovski S. (1999) Introduction to making and using protein multiple alignments [online]
http://bioinfo.weizmann.ac.il/~pietro/Making_and_using_protein_MA/ [staženo 26.6.2013]

[PIR]
Protein Information Resource [online] http://pir.georgetown.edu/ [staženo 26.6.2013]

[PMass]
PeptideMass [online] http://web.expasy.org/peptide_mass/ [staženo 26.6.2013]

[Polášek 2005]
Polášek M. (2005) Hmotnostní spektrometrie [online]
http://www.njh.cz/seminare/457864/1270588 [staženo 26.6.2013]

[Polzer 2008]
Polzer J. (2008) Jak vyčistit nahrávku od šumu a praskání v Audacity [online]
http://extrawindows.cnews.cz/jak-vycistit-nahravku-od-sumu-praskani-v-audacity
[staženo 26.6.2013]

[Posch at al., 2013]
Posch AE, Koch C, Helmel M, Marchetti-Deschmann M, Macfelda K, Lendl B et al. (2013) Combining light microscopy, dielectric spectroscopy, MALDI intact cell mass spectrometry, FTIR spectromicroscopy and multivariate data mining for morphological and physiological bioprocess characterization of filamentous organism., str. 1-11, Fungal Genet Biol 51

[Poustka 2007]
Poustka J. (2007) Hmotnostní spektrometrie - Mass Spectrometry (MS) [online]
http://web.vscht.cz/poustkaj/ISM%20MS%20PRINCIP%20A%20IONIZACE
%20%20102007.pdf [staženo 26.6.2013]

[PRF]
Protein Research Foundation [online] http://www.prf.or.jp/index-e.html [staženo 26.6.2013]

[PyMOL]
PyMOL [online] http://www.pymol.org/ [staženo 26.6.2013]

[Python]
Python Programming Language – Official Website [online] http://www.python.org/
[staženo 26.6.2013]

[RAPIs]
Gene Expression Atlas APIs [online] http://www.ebi.ac.uk/gxa/help/AtlasApis
[staženo 26.6.2013]

[Reusch 2013]
Reusch W. (2013) Mass Spectrometry [online]
http://www2.chemistry.msu.edu/faculty/reusch/VirtTxtJml/Spectrpy/MassSpec/masspec1.htm
[staženo 26.6.2013]

[RH]
Red Hat - The World's Open Source Leader [online] http://www.redhat.com/
[staženo 26.6.2013]

[Ruby]
Ruby Programming Language [online] http://www.ruby-lang.org/en/ [staženo 26.6.2013]

[Řehulka 2007]
Řehulka P., Řehulková H., Chmelík J. (2007) Měření proteomických dat hmotnostní spektrometrií a jejich bioinformatická interpretace, str. 280, Chemické Listy 101, 279−286,

[Sar]
AnagnosTec SARAMIS™ - Spectral ARchive And Microbial Identifications System [online]
http://www.anagnostec.eu/products-services/saramis.html [staženo 26.6.2013]

[SAS]
Sequence Alignment Software [online]
http://www.sequence-alignment.com/sequence-alignment-software.html [staženo 26.6.2013]

[SciLi]
Scientific Linux [online] https://www.scientificlinux.org/ [staženo 26.6.2013]

[SepM]
Separační metody [online] http://www.vscht.cz/ktk/www_324/lab/texty/ana/A02.pdf
[staženo 26.6.2013]

[shell]
Programování v (bash) shellu (2000) [online]
http://www.root.cz/clanky/programovani-v-bash-shellu/ [staženo 26.6.2013]

[SQLite]
SQLite Home Page [online] http://www.sqlite.org/ [staženo 26.6.2013]

[Šeda 2002]
Šeda M. (2002) Databázové systémy [online]
http://www.uai.fme.vutbr.cz/~mseda/DBS02_BS.pdf [staženo 26.6.2013]

[Štarnclová 2012]
Štarnclová J. (2012) AIL072 - Rozpoznávání vzorů (jednotlivé přednášky v PDF) [online]
http://www1.cuni.cz/~stancloj/ail072/ail072_index.html [staženo 26.6.2013]

[Švanda 2011]
Švanda M. (2011) Šum a jeho potlačení [online]
http://sirrah.troja.mff.cuni.cz/~svanda/AST001/F_sum_mrizky_prezentace.pdf
[staženo 26.6.2013]

[Švihlík 2008]
Švihlík J. (2008) Metody rekonstrukce obrazu a odstranění šumu z obrazu [online]
http://goo.gl/hET81 [staženo 26.6.2013]

[Thompson et al.,1997]
Thompson J.D., Gibson T.J. Plewniak F. Jeanmougin F., Higgins D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools [online] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC147148/pdf/254876.pdf
[staženo 26.6.2013]

[Tools 4 MP]
Tools for Multiple Alignments [online] http://pbil.univ-lyon1.fr/alignment.html
[staženo 26.6.2013]

[TOP500]
TOP500 Supercomputer Sites [online] http://www.top500.org/statistics/list/ [staženo 26.6.2013]

[Tresh]
Tresholding (Jain et al., Sections 3.2.1, 3.2.2, Petrou et al., Chapt 7) [online]
http://www.cse.unr.edu/~bebis/CS791E/Notes/Thresholding.pdf [staženo 26.6.2013]

[Tůma 2011]
Tůma M. (2011) Letadlo vs. radar – kdo vyhraje? [online]
http://vtm.e15.cz/letadlo-vs-radar-mateme-rusime-nicime [staženo 26.6.2013]

[UbSci]
UbuntuScience [online] https://help.ubuntu.com/community/UbuntuScience
[staženo 26.6.2013]

[UniProt]
UniProt [online] http://www.uniprot.org/ [staženo 26.6.2013]

[UPS]
Záložní zdroje UPS (2011) [online]
http://www.elektrotrh.cz/cs/elektricke-a-zalozni-zdroje-energie/zalozni-zdroje-ups
[staženo 26.6.2013]

[WApp1]
Webové aplikace [online] http://www.shopcentrik.cz/slovnik/webove-aplikace.aspx
[staženo 26.6.2013]

[WApp2]
O webových aplikacích [online] https://support.google.com/chrome/answer/1050586?hl=cs
[staženo 26.6.2013]

[Xiong 2006a]
Xiong J. (2006) Essential Bioinformatics, str. 31-94, Cambridge University Press, Cambridge

[Xiong 2006b]
Xiong J. (2006) Essential Bioinformatics, str. 57, Cambridge University Press, Cambridge

[Xiong 2006c]
Xiong J. (2006) Essential Bioinformatics, str. 52, Cambridge University Press, Cambridge

[Xu & Ma 2006]
Xu Ch., Ma B. (2006) Software for Computational Peptide Identification from MS-MS data
[online] http://www.csd.uwo.ca/~bma/pub/msreview.pdf [staženo 26.6.2013]

[yEd homepage]
yEd Graph Editor [online] http://www.yworks.com/en/products_yed_about.html
[staženo 26.6.2013]

[Zmasek]
Zmasek Ch.: Archaeopteryx: Visualization, Analysis, and Editing of Phylogenetic Trees
[online] https://sites.google.com/site/cmzmasek/home/software/archaeopteryx
[staženo 26.6.2013]

[Zuker 2004]
Zuker M. (2004) Alignment: Definitions and algorithms [online]
http://homepages.rpi.edu/~zukerm/MATH-4961/align/ [staženo 26.6.2013]

Seznam obrázků

Příloha A – publikované výsledky (ProteinCutter)

Sebela M., Rehulka P., Kabrt J., Rehulkova H., Ozdian T., **Raus M**., Franc V., Chmelik J. (2009) *Identification of N-glycosylation in prolyl endoprotease from Aspergillus niger and evaluation of the enzyme for its possible application in proteomics.* J. Mass Spectrom. 44 (11), 1587-1595

Research Article

Journal of MASS SPECTROMETRY

Received: 8 April 2009 Accepted: 21 August 2009 Published online in Wiley Interscience: 15 September 2009

(www.interscience.com) DOI 10.1002/jms.1667

Identification of N-glycosylation in prolyl endoprotease from *Aspergillus niger* and evaluation of the enzyme for its possible application in proteomics†

Marek Šebela,[a]* Pavel Řehulka,[a,b] Jaromír Kábrt,[a] Helena Řehulková,[b] Tomáš Oždian,[a] Martin Raus,[a] Vojtěch Franc[a] and Josef Chmelík[c]

An acidic prolyl endoprotease from *Aspergillus niger* was isolated from the commercial product Brewers Clarex™ to evaluate its possible application in proteomics. The chromatographic purification yielded a single protein band in sodium dodecyl sulfate polyacrylamide gel electrophoresis providing an apparent molecular mass of 63 kDa and a broad peak (*m/z* 58 061) in linear matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) mass spectrometry (MS) indicating the glycoprotein nature of the enzyme. Indeed, a colorimetric assessment with phenol and sulfuric acid showed the presence of neutral sugars (9% of weight). The subsequent treatment with *N*-glycosidase F released a variety of high-mannose type *N*-glycans, which were successfully detected using MALDI-TOF MS. MALDI-TOF/TOF tandem MS analysis of glycopeptides from a tryptic digest of prolyl endoprotease unraveled the identity of the N-glycosylation site in the primary structure. The data obtained also show that the enzyme is present in its processed form, i.e. without putative signal and propeptide parts. Spectrophotometric measurements demonstrated optimal activity at pH 4.0–4.5 and also high thermostability for the cleavage at the C-terminal part of proline residues. In-solution digestion of standard proteins (12–200 kDa) allowed to evaluate the cleavage specificity. The enzyme acts upon proline and alanine residues, but there is an additional minor cleavage at some other residues like Gly, Leu, Arg, Ser and Tyr. The digestion of a honeybee peptide comprising six proline residues (apidaecin 1A) led to the detection of specific peptides terminated by proline as it was confirmed by MALDI postsource decay analysis. **Copyright © 2009 John Wiley & Sons, Ltd.**

Keywords: *Aspergillus niger*; Brewers Clarex; digestion; glycan; MALDI-TOF; proline; prolyl endoprotease

Introduction

The filamentous fungus *Aspergillus niger*, a member of the black aspergilli,[1] is known as an infectious agent causing invasive pulmonary aspergillosis, particularly in immunosuppressed patients such as transplant recipients. This fungal disease that usually develops after inhaling airborne conidia is very devastating, resulting in high mortality.[2–4]

In contrast, *A. niger* has a great significance for biotechnology. In Asia, it is widely used in the manufacture of local fermented foods and beverages.[1,5] However, the most beneficial use of the fungi resides in the production of enzymes and organic acids by fermentation, which has a long tradition. Many of these products have obtained "generally regarded as safe" status.[6] *A. niger* exhibits a remarkably versatile metabolism. It can grow on solid or in liquid media under various environmental conditions. In their natural habitat, *A. niger* strains secrete large amounts of enzymes to release nutrients from biopolymers, and this property has been exploited by industry.[7] *A. niger* is also used for the production of recombinant proteins[8] and it has a big potential for bioremediation due to its ability of degrading a range of xenobiotics through various oxidative, hydroxylation and demethylation reactions.[9] The genomic DNA sequences of *A. niger* strains ATCC 1015, ATCC 9029[10] and CBS 513.88,[11] which is an early ancestor of ATCC 1015 used in industrial enzyme production, have recently been published. This achievement will definitely facilitate the development of new products, improved strains and more efficient industrial fermentation processes.[11]

A few years ago, *A. niger* strain CBS 513.88 was used as a host for the overproduction of a proline-specific endoprotease (PSE), which had previously been identified in an *A. niger* strain using *N*-benzyloxycarbonyl-glycyl-prolyl-*p*-nitroanilide (Z-Gly-Pro-*p*NA) as a chromogenic peptide substrate for activity screening in molds. PSE has an acidic pH optimum and preferentially cleaves its peptide and protein substrates at the C-terminus of proline residues.[12,13] Peptide bonds C-terminal of alanine can also be cleaved, but with a lower efficiency.[12] As large amounts of pure PSE can be produced

* *Correspondence to: Marek Šebela, Department of Biochemistry, Faculty of Science, Palacký University, Šlechtitelů 11, CZ-783 71 Olomouc, Czech Republic. E-mail: marek.sebela@upol.cz*

† *This article was published online on 15 September 2009. An error was subsequently identified in Figure 7. This notice is included in the print and online versions to indicate that both have been corrected; 14 October 2009*

a *Department of Biochemistry, Faculty of Science, Palacký University, Šlechtitelů 11, CZ-783 71 Olomouc, Czech Republic*

b *Institute of Molecular Pathology, Faculty of Military Health Sciences, University of Defence, Třebešská 1575, CZ-50001 Hradec Králové, Czech Republic*

c *Institute of Analytical Chemistry, Czech Academy of Sciences, Veveří 97, CZ-602 00 Brno, Czech Republic*

Copyright © 2009 John Wiley & Sons, Ltd.

relatively cheaply, the enzyme has found its current application in debittering of protein hydrolyzates[12] and prevention of chill-haze in beer by degradation of proline-rich proteins.[14]

Proline-specific peptidases have been studied in many organisms and they include endopeptidases (postproline cleaving enzyme and proline-specific endopeptidase), N-terminal exopeptidases (postproline dipeptidyl aminopeptidase, proline iminopeptidase and aminopeptidase P), C-terminal exopeptidases (prolylcarboxypeptidase and carboxypeptidase P) and dipeptidases (prolyl dipeptidase and proline dipeptidase).[15–17] To our best knowledge, their use in mass spectrometry (MS) of proteins and peptides is not frequent.[18]

In this work, the commercial product Brewers Clarex™[19] was used to obtain PSE. The enzyme was purified by liquid chromatography, and its molecular and kinetic properties were characterized by biochemical methods. Stability against autolysis of the enzyme was tested by incubating in solutions of different pH values, followed by mass spectrometric analyses of the autolyzates. For the analysis of N-glycosylation, PSE was treated with *N*-glycosidase F and the released *N*-glycans were subjected to matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS and tandem MS/MS on a TOF/TOF instrument. To demonstrate PSE applicability in protein analysis, standard proteins of molecular mass from the interval 12–200 kDa were subjected to in-solution digestion with PSE in 50 mM ammonium acetate. The resulting peptide mixtures were analyzed by MALDI-TOF MS and MALDI-TOF/TOF MS/MS. All these analyses confirmed the previously reported cleavage specificity, which is reflected in the predominant production of peptides C-terminated by proline and alanine. In addition, a minor action on other amino acid residues was found.

Materials and Methods

Chemicals

Unless otherwise specified, the chemicals used were purchased from Sigma-Aldrich Chemie (Steinheim, Germany). Brewers Clarex was from DSM Food Specialties (Seclin, France). Z-Gly-Pro-*p*NA was purchased from Bachem (Weil am Rhein, Switzerland). Tris base was obtained from MP Biomedicals (Aurora, OH, USA). A recombinant *N*-glycosidase F from *Elizabethkingia meningoseptica* (Calbiochem brand, Cat. No. 362 185), with a specific activity of 25 000 U mg^{-1}, was purchased from Merck (Darmstadt, Germany). Protein standards used in MALDI-TOF peptide mass fingerprinting (PMF) [horse heart cytochrome *c*, hen egg lysozyme, horse heart myoglobin, bovine carboanhydrase, bovine serum albumin (BSA), baker's yeast alcohol dehydrogenase and β-amylase from sweet potato] were from Sigma-Aldrich Chemie. MALDI matrices were from Bruker Daltonik (Bremen, Germany). A synthetic honeybee peptide apidaecin 1A was provided by Dr Marek Petřivalský, Faculty of Science, Palacký University.

Enzyme purification

PSE was purified from Brewers Clarex (20 ml) by two steps of low-pressure liquid chromatography performed at 0–5 °C. First, the liquid commercial product was dialyzed against 2 × 3 l of 20 mM Tris buffer, pH 8.0 (buffer A) at 4 °C overnight. After centrifugation at 10 000*g* for 20 min, the dialyzate was loaded onto a Macro-Prep High Q (Bio-Rad, Hercules, CA, USA) column, 2.5 cm internal diameter (i.d.) × 30 cm, previously equilibrated with buffer A and operating at a flow rate of 2 ml min^{-1}. The ionex column was connected to a low-pressure chromatographic system consisting of a peristaltic pump P-1, a gradient mixer GP-250, a single path monitor UV-1 operating at 280 nm, a fraction collector Frac-920 and a two-channel recorder REC-112 (GE Healthcare, Uppsala, Sweden). After a washing step to remove unbound impurities (buffer A), the retained proteins were eluted by 50 ml of linear gradient of 0-1 M NaCl in buffer A. Fractions containing PSE activity were pooled and dialyzed against the buffer A at 4 °C. The dialyzate was concentrated to a volume of 10 ml by ultrafiltration using a 10-kDa cutoff filter (Millipore, Bedford, MA, USA). The subsequent gel chromatography was performed on a Sephacryl S-200 HR (GE Healthcare) column, 2.5 cm internal diameter (i.d.) × 50 cm, equilibrated with the buffer A and operated at a flow rate of 1 ml min^{-1}. Active fractions were pooled and dialyzed against 0.1% (v/v) formic acid. After concentration as above aliquots of the enzyme solution were frozen and lyophilized. The final enzyme preparation was stored at −80 °C.

Determination of enzyme activity and protein assay

Spectrophotometric determination of PSE activity was performed with Z-Gly-Pro-*p*NA as a substrate using a modification of the previously published protocols.[20,21] The standard reaction mixture, thermostated at 37 °C, contained 0.1 M citric acid/0.2 M disodium phosphate buffer, pH 5.0, 20% (v/v) acetonitrile and 50 µl of 30 mM Z-Gly-Pro-*p*NA solution in dimethylsulfoxide, all in a total volume of 1.48 ml. The cleavage reaction was initiated by the addition of 20 µl of the enzyme solution. The release of *p*-nitroaniline by PSE reaction was monitored by increased absorbance at 410 nm taking a molar extinction coefficient of 8 800 M^{-1} for calculations.[20,21] One activity unit (1 nkat) is defined here as the enzyme activity that liberates 1 nmol of *p*-nitroaniline in 1 s under described reaction conditions. McIlvaine's buffers,[22] pH 3.0–7.0, made by mixing 0.1 M citric acid and 0.2 M disodium phosphate in a proper ratio, were used for activity assay in determining the pH optimum for PSE. Thermostability of the enzyme was evaluated by monitoring the changes in activity upon incubating enzyme aliquots (in the assay buffer, pH 5.0) at temperatures from 30 °C up to 50 °C for 30 min, followed by rapid cooling in water–ice bath. Protein content was determined using bicinchoninic acid[23] with BSA as a standard.

Molecular mass determination

One-dimensional sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS–PAGE) was run using 10% T running and 4% T stacking gels (2.6% C),[24] where T and C represent the total acrylamide concentration and crosslinking ratio, respectively. MALDI-TOF MS for intact molecular mass determination was performed on a Microflex LRF20 MALDI-TOF mass spectrometer (Bruker Daltonik) equipped with a MicroScout ion source and a 337-nm nitrogen laser. Prior to analysis, enzyme samples were dialyzed against 25 mM ammonium bicarbonate and lyophilized. After dissolving in deionized water to obtain a protein solution of 10 mg ml^{-1}, 1 µl of sample was mixed with 3 µl of matrix [10 mg ml^{-1} sinapinic acid in acetonitrile/0.1% trifluoroacetic acid (TFA), 1 : 1, v/v]. Aliquots (0.6 µl) were placed onto an MSP AnchorChip™ 600/96 target plate (Bruker Daltonik) previously modified by a thin layer of sinapinic acid[25] and allowed to dry.

 Copyright © 2009 John Wiley & Sons, Ltd.

Spectra were acquired in the linear mode for positive ions and calibrated externally using the following molecular ions of BSA (average masses): $[M+H]^+$ *m/z* 66 431; $[M+2H]^{2+}$ *m/z* 33 216; dimer $[M+H]^+$ *m/z* 132 859. Parameters of the instrument were as follows: an acceleration voltage of 20.0 kV, an extraction voltage of 18.0 kV, a lens voltage of 8.7 kV and a delayed extraction of 1000 ns.

MALDI-TOF MS

Aliquots (1 nmol) of protein standards in 930 µl of 50 mM ammonium bicarbonate were mixed with 20 µl of 100 mM dithiothreitol (DTT) and incubated at 70 °C for 15 min. After cooling to laboratory temperature, 20 µl of 200 mM iodoacetamide was added and the mixture was kept in the dark at laboratory temperature for 30 min. The alkylation reagent was quenched by adding 30 µl of 100 mM DTT and subsequent incubation at laboratory temperature for 10 min. The alkylated protein solutions were evaporated on a vacuum centrifuge Concentrator plus (Eppendorf, Hamburg, Germany). The solid residues were dissolved in 1 ml of 50 mM ammonium acetate and digested with 100 pmol PSE at 37 °C overnight. In-solution tryptic digestion of 100 µg PSE after reduction/alkylation as above was performed in 50 mM ammonium bicarbonate using a trypsin-to-PSE ratio of 1 : 20. Aliquots of the digests (10 µl) were mixed with 10 µl of 0.1% (v/v) TFA. Then a desalting operation was performed using ZipTip C_{18} pipette tips (Millipore) and indirect elution according to manufacturer's protocol. Alternatively, the digestion of standard proteins by PSE was performed in 50 mM ammonium acetate without prior reduction/alkylation. Positive-ion mass spectra of peptides were measured in the reflectron mode on the Microflex mass spectrometer using an acceleration voltage of 19 kV, an extraction voltage of 16.1 kV, a lens voltage of 9.1 kV, a reflectron voltage of 20 kV and a delayed extraction of 250 ns. MALDI probes were prepared using the dried droplet method with an α-cyano-4-hydroxycinnamic acid matrix (CHCA).[26] The instrument was calibrated externally using a mixture of peptide standards (Bruker Daltonik). The acquired spectra were processed by flexAnalysis 2.4 and Biotools 3.0 software (Bruker Daltonik). Protein identification was achieved using the database search engine Mascot Server 2.2 (Matrix Science, London, UK) and nonredundant protein databases SwissProt and NCBInr (versions available in March 2009). For the database search in PMF, PSE cleavage specificity was set at the C-termini of proline and alanine in Mascot configuration. Other search parameters were as follows: no taxonomic restriction; one missed cleavage; variable modification – methionine oxidation; fixed modification – carbamidomethyl cysteine; peptide tolerance – 150 ppm; $[M+H]^+$; monoisotopic masses. Postsource decay (PSD) MALDI-TOF experiments were performed as described previously.[27] PSD spectra were processed by the software flexAnalysis 2.4, Biotools 3.0 and RapiDeNovo (Bruker Daltonik).

Tandem MS

MALDI-TOF/TOF measurements in the positive reflectron mode were performed with an ABI 4800 Proteomics Analyzer (Applied Biosystems, Framingham, MA, USA). The instrument was equipped with a neodymium-doped : yttrium aluminium garnet laser (355 nm) of 3 – 7 ns pulse and 200-Hz firing rate. Both MS and MS/MS spectra were acquired using dual-stage reflectron mirror. Accelerating voltages applied for MS and MS/MS measurements were 20 kV and 8 kV, respectively. In MS/MS mode, a collision energy of 1 kV was applied and air was used as a collision gas. For measurements of protein digests, the matrix solution was CHCA (5 mg ml^{-1}) dissolved in acetonitrile/0.1% TFA (3 : 2, v/v); a mixture of six standard peptides was used for external calibration. Aliquots of in-solution digests desalted using ZipTip C_{18} (Millipore) were spotted onto the previously deposited matrix layer and left to dry. MS and MS/MS data were further processed using 4000 Series Explorer (Applied Biosystems). Peak lists were created using Peaks to Mascot tool included in this software and submitted to a local database search engine Mascot. The following search parameters were used: database – SwissProt (version 56.5 of 25.11.2008); taxonomy – all entries; enzyme – none; allowed missed cleavages – 0; fixed modifications – carbamidomethyl cysteine; variable modifications – none; peptide tolerance – 30 ppm; MS/MS tolerance – 300 mmu; peptide charge – (+1); monoisotopic masses and instrument – MALDI-TOF/TOF.

Carbohydrate analyses

Neutral sugar content was estimated by an established colorimetric method with phenol and sulfuric acid.[28] Enzyme samples containing 500 – 1000 µg of protein in 0.1 ml water were mixed with 1 ml of 5% (w/v) phenol and 5 ml of sulfuric acid. After a 30-min incubation at laboratory temperature, the absorbance at 490 nm was read against blank. Calibration was made using 10 – 100 µg of mannose. Enzymatic deglycosylation of PSE was performed at pH 7.5 using *N*-glycosidase F under reducing conditions and in the presence of detergents according to manufacturer's instructions. Control mixture did not contain the glycosidase. After incubation at 37 °C, the released *N*-glycans were separated from the reaction mixture by ultrafiltration using 10-kDa cutoff Microcon centrifugation cartridges (Millipore). The filtrate was desalted on porous graphitized carbon material obtained from Hypercarb SPE cartridges (Thermo Fisher Scientific, Walthan, MA, USA) and packed into GELoader pipette tips (Eppendorf) using C8 disc as a frit to retain the carbon sorbent. The desalting procedure was carried out essentially as previously described.[29] MALDI-TOF mass spectra of the purified *N*-glycans were acquired on the Microflex instrument operating in the reflectron mode for positive ions with the same voltage parameters as given above for measuring peptides. MALDI-TOF/TOF MS and MS/MS were performed on the ABI 4800 Proteomics Analyzer operating in the reflectron mode. The 2,5-dihydroxybenzoic acid (DHB, 20 mg ml^{-1} in acetonitrile/0.1% TFA, 1 : 2, v/v) or 6-aza-2-thiothymine (ATT, 1 mg ml^{-1} in ethanol/20 mM diammonium hydrogencitrate, 1 : 1, v/v) were used as matrices. Sample preparation was carried out by the dried droplet method using direct mixing on the target (0.5 µl sample plus 0.5 µl matrix solution).

Results and Discussion

Chromatographic purification of PSE

Brewers Clarex was a pale brown solution of high density (it probably contains a stabilizer-thickener like e.g. glycerol) and pH 4.2. The protein content was 100 mg ml^{-1}. Enzyme activity of PSE in the commercial product was 11.7 nkat ml^{-1} (measured using Z-Gly-Pro-*p*NA as a substrate) giving a specific activity value of 0.12 nkat mg^{-1}. Prior to chromatography, the crude enzyme was dialyzed against solutions of different pH values and the dialyzates

 Copyright © 2009 John Wiley & Sons, Ltd.

Table 1. An overview of PSE purification from 20 ml of Brewers Clarex

Purification step	Volume (ml)	Total activity (nkat)	Total protein (mg)	Specific activity (nkat/mg)	Enrichment factor (fold)	Yield (%)
Brewers Clarex	20	234	2000	0.117	1	100
Ionex chromatography	100	230	1800	0.128	1.09	98
Gel chromatography	60	179	1040	0.172	1.47	76

Enzyme activity was assayed spectrophotometrically with Z-Gly-Pro-*p*NA as a substrate.[20,21] Protein content was estimated using bicinchoninic acid with BSA as a standard.[23]

kept at 4 °C for 7 days before repeated activity assay to evaluate stability. PSE was found stable in 0.1% (v/v) formic acid, 20 mM Bistris-HCl, pH 6.0 and in buffers of basic pH like 20 mM Tris–HCl, pH 8.0 or 50 mM ammonium bicarbonate, pH 8.0. Based on theoretical *p*I values of around 4 deduced from the translated amino acid sequence (GenBank/EMBL accession no. AX458699–calculated without the signal peptide at the positions 1–22 and propeptide at 23–41) and described in the literature,[12] the low-pressure chromatography on Macro-Prep High Q anion exchanger at pH 8.0 was the purification step of first choice. PSE was bound to the ion exchanger and eluted with a linear salt gradient. The second purification step involved gel chromatography on Sephacryl S-200 HR. The final PSE preparation had a specific activity of 0.17 nkat mg^{-1}, an enrichment factor of 1.5 and a yield of 76% (Table 1). This clearly shows that Brewers Clarex represents a solution of highly pure enzyme. The low enrichment factor value achieved would indicate that the minor nonactive impurities probably originated from PSE autolysis (results discussed subsequently). The final enzyme preparation was dialyzed against 0.1% (v/v) formic acid, concentrated using ultrafiltration (a 10 kDa cutoff filter) and stored as a white lyophilizate at −80 °C.

Characterization of PSE properties

Homogeneity of the final enzyme preparation was checked by SDS–PAGE, which showed a single fuzzy band corresponding to an apparent molecular mass of 63 kDa (Fig. 1, inset). A very similar band of 60 kDa was independently obtained by size-exclusion chromatography on a Superdex 200 HR column (GE Healthcare) performed following a protocol already described.[30] The original Brewers Clarex contained impurities as demonstrated by additional bands on SDS–PAGE gels (Fig. 1, inset). However, MALDI-TOF PMF of these impurities after in-gel digestion by modified trypsin confirmed the presence of PSE fragments (results not shown).

Protein intact mass measurement using MALDI-TOF MS revealed two major peaks at *m/z* 29 027 and 58 061, characterized by a low resolution, which could be assigned to doubly charged and singly charged molecular ions of PSE, respectively (Fig. 1). The width of the peaks suggested a certain degree of glycosylation, which probably influenced the migration of PSE samples in SDS–PAGE. MALDI-TOF PMF on Microflex after in-gel tryptic digestion confirmed the identity of PSE (Fig. 2). The protein was assigned to the accession number XP_001392567 in the NCBInr protein database (this amino acid sequence differs only in a few amino acids from that translated from the nucleotide sequence deposited under the accession number AX458699); 20 matched peptides covered 38% of the amino acid sequence and the probability-based score was 118. MS/MS spectra of precursor peptides with *m/z* 2129.0, 2200.0 and 2271.1

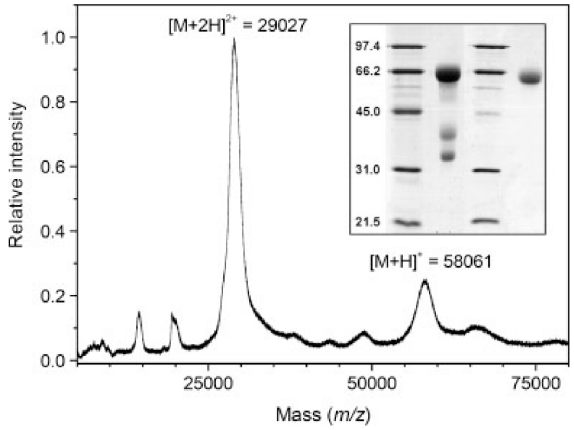


Figure 1. Molecular mass determination of PSE. MALDI-TOF MS of intact PSE (main panel) was performed on the Microflex LRF-20 instrument operating in the linear mode for positive ions. Sinapinic acid was used as a matrix. The inset shows a scan of two SDS–PAGE gels run with PSE samples; lanes from the left: protein standards with the indicated molecular mass (kDa), Brewers Clarex (8 µg protein), the same protein standards and purified PSE (5 µg). The gels were stained by Bio-Safe Coomassie (Bio-Rad).

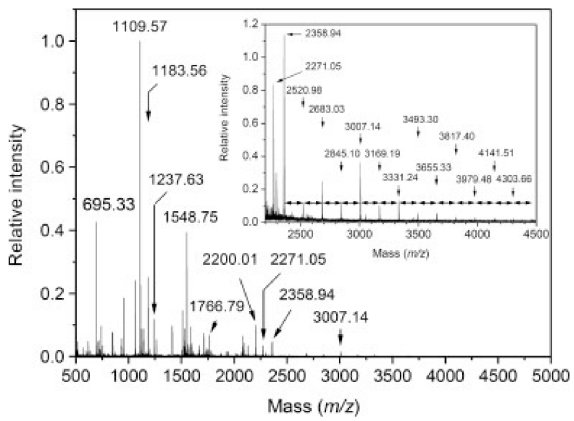


Figure 2. MALDI-TOF PMF of purified PSE. After previous reduction and alkylation, a sample of PSE in 50 mM ammonium bicarbonate was subjected to in-solution digestion by modified trypsin. The main panel shows a MALDI-TOF mass spectrum of the digest that was acquired on the ABI 4800 Proteomics Analyzer operating in the reflectron mode for positive ions (CHCA matrix). The inset shows a magnification of the spectrum region with well pronounced signals of glycopeptides where the mass differences of 162 Da (representing one hexose unit) are indicated by double-headed horizontal arrows.

Copyright © 2009 John Wiley & Sons, Ltd.

yielded the sequences TTGEAYFEQLLDHHNPEK, ATTGEAYFEQLLDHHNPEK and AATTGEAYFEQLLDHHNPEK (not shown). At the same time, no peptide was registered originating from the putative signal and propeptide sequences.[12] This would indicate that the enzyme is present in Brewers Clarex in its processed form – a mature peptidase unit (MEROPS database, http://merops.sanger.ac.uk/, accession number MER093133), which is released from the zymogen by an activation cleavage. The presence of variable N-termini might suggest an autocatalytic role of PSE in the activation processing. The enzyme has been reported to cleave peptide bonds C-terminal of alanine, but with a lower efficiency when compared to its action on proline.[12]

With respect to the possible glycosylation, neutral sugar content of PSE was estimated using a spectrophotometric method with phenol and sulfuric acid, which provided an average value of 9% (w/w). Taking the molecular mass of 58 kDa from MALDI-MS into consideration, a difference of about 5 kDa can be calculated, which is in rough accordance with the predicted molecular mass of 54 kDa of the processed enzyme. The amino acid sequence of PSE comprises seven potential N-glycosylation sites suggesting a possible N-glycosylation. Indeed, a ladder of glycopeptides was observed in tryptic digest of the enzyme with *m/z* values gradually increasing between 1872 and 4952 by a mass difference of 162 Da (illustrated in Fig. 2, inset). The observed difference clearly indicated the presence of variable hexose units referring to high-mannose type glycans. MS/MS analysis of glycopeptides with *m/z* 2358.9, 2521.0 and 2683.0 yielded the same fragment ion corresponding to the peptide sequence (K)-LQFDNSTR-(S), which allowed localization of the respective N-glycosylation site (underlined). This peptide containing one *N*-acetylglucosamine residue was also registered as a peak with *m/z* 1183.6 (confirmed by MS/MS). So far no information on PSE glycosylation has appeared in the literature.

Enzymatic deglycosylation of PSE by a recombinant *N*-glycosidase F provided a molecular mass of 54 kDa in SDS–PAGE (Fig. 3, inset). Autolytic fragments of PSE, which appeared in the reaction mixture, were also deglycosylated as demonstrated by the observed mass difference. The released *N*-glycans were recovered from the reaction mixture using porous graphitic carbon, eluted by 25% acetonitrile and analyzed by MALDI-MS. Figure 3 shows a mass spectrum of PSE-derived *N*-glycans acquired in the reflectron mode for positive ions using DHB as a matrix[31] (similar spectra were recorded using ATT[32]). All sugars were registered in the form of sodium adduct peaks $[M+Na]^+$ accompanied by minor peaks of potassium adducts $[M+K]^+$. The corresponding mass list was analyzed by the GlycoMod software[33] (http://www.expasy.ch/tools/glycomod) resulting in high-mannose type structures[34,35] containing 0–21 hexose units in addition to the pentasaccharide core structure $(GlcNAc)_2(Man)_3$. The glycan type was additionally confirmed by MALDI-TOF/TOF MS/MS analysis of the peak with *m/z* 1419.49 (not shown). When measuring the same sample in the negative reflectron mode with ATT as a matrix, the following peaks were observed (*m/z* values): 989.30, 1151.37, 1313.44, 1475.50, 1637.55, 1799.62, 1961.67, 2123.73 and 2285.78. Based on GlycoMod interpretation, such peaks belong to deprotonated high-mannose type sugars of a general formula $(GlcNAc)_2(Man)_3(Hex)_n$, where $n = 0–8$, plus one phosphate group. MS/MS spectra of the negatively charged sugars confirmed the presence of phosphorylation by recording *m/z* 79/97 phosphate-specific ions (not shown). In summary, PSE contains high-mannose type *N*-glycans that are partially phosphorylated.

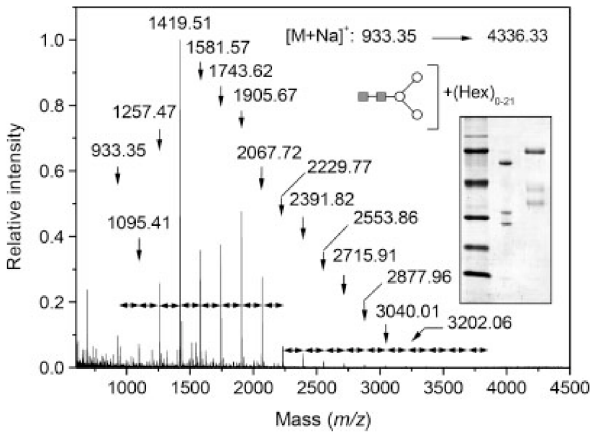


Figure 3. Analysis of neutral *N*-glycans released from PSE by *N*-glycosidase F. The main panel shows a MALDI-TOF mass spectrum of PSE-derived glycans that was acquired using the ABI 4800 Proteomics Analyzer operating in the reflectron mode for positive ions (DHB matrix). The mass differences of 162 Da (representing one hexose unit) are indicated by double-headed horizontal arrows. The inset demonstrates evaluation of the performance of *N*-glycosidase F by SDS–PAGE; lane 1: protein molecular weight marker (from the top −97.4, 66.2, 39.2, 26.6, 21.5 and 14.4 kDa), lane 2: deglycosylated PSE (1 µg) and lane 3: untreated PSE (1 µg). The gels were stained by Bio-Safe Coomassie (Bio-Rad).

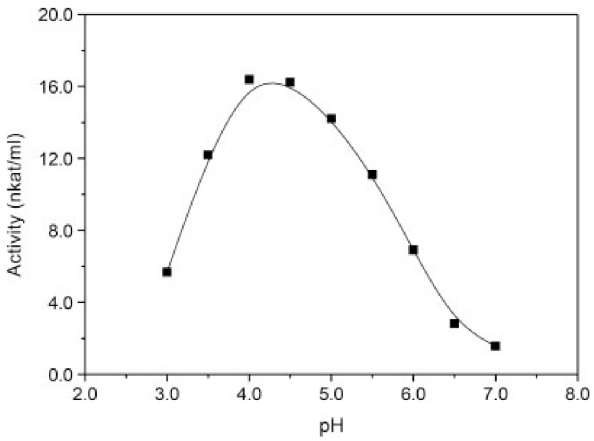


Figure 4. PSE activity under various pH conditions. Based on the measured profile, PSE shows optimal activity at pH 4.0–4.5. Spectrophotometric activity assays were performed using Z-Gly-Pro-*p*NA as a substrate.

For kinetic determination of pH optimum, McIlvaine's buffers of pH 3.0–7.0 were used. The enzyme showed highest activity at pH 4.0–4.5 (Fig. 4), which is in agreement with previously published data.[12] Interestingly, PSE exhibited high thermostability. When dissolved in McIlvaine's buffer, pH 5.0, the enzyme could be incubated at 65 °C for 30 min without any loss of activity.

PSE autolyzate and its analysis

PSE autolysis was evaluated by incubation of the enzyme in solutions of various pH values with the subsequent measurements of MALDI-TOF mass spectra of the released peptides. PSE solutions (1.5 µM) were made in 0.1% (v/v) formic acid (pH 2.4), 50 mM ammonium acetate (pH 4.5) and 50 mM ammonium

Copyright © 2009 John Wiley & Sons, Ltd.

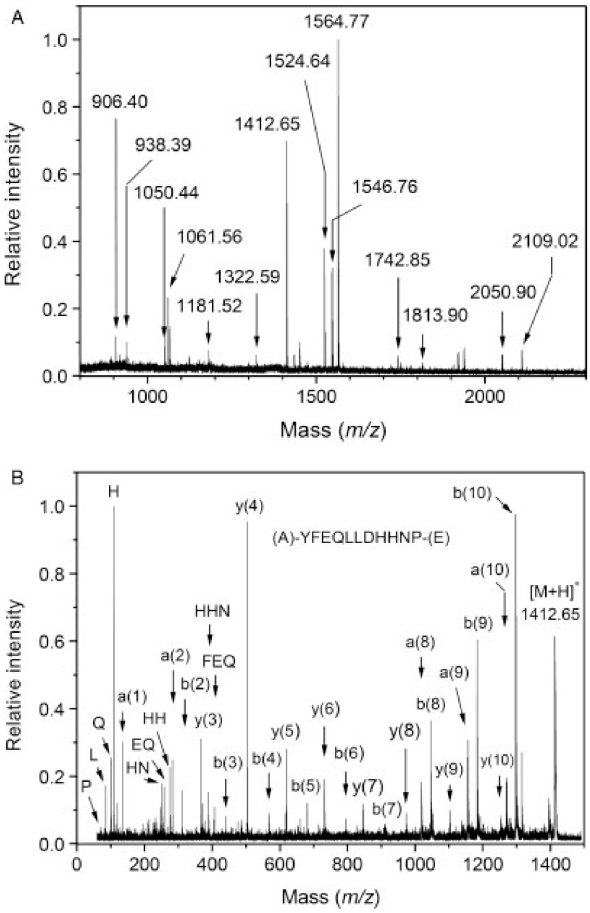


Figure 5. Mass spectrometric analysis of PSE autolysis. (A) MALDI-TOF PMF of an autolyzate obtained by PSE incubation in 50 mM ammonium acetate. The fingerprint spectrum was acquired on the ABI 4800 Proteomics Analyzer operating in the reflectron mode for positive ions (CHCA matrix). (B) MALDI-TOF/TOF tandem MS of the peptide (A)-YFEQLLDHHNP-(E) from the PSE autolyzate. The peptide was observed at an experimental *m/z* value of 1412.65 in the peptide mass fingerprint of the autolyzate (compare with A). The spectrum was measured on the ABI 4800 Proteomics Analyzer using CHCA as a matrix. Fragment ions are appropriately labeled; single capital letters denote immonium or internal fragment ions.

bicarbonate (pH 8.0) and incubated at 37 °C overnight. For prediction of results, a web application Protein Cutter (available at: http://biochemie.upol.cz/software/proteincutter) was developed, which cleaves *in silico* a user-entered protein sequence with PSE (taking the published cleavage specificity P-X and A-X into account) and computes the masses of the generated peptides. The software certainly allows to choose another common proteolytic enzyme or user-entered cutting rules. No autolysis was observed in the diluted formic acid (pH 2.4) and ammonium bicarbonate (pH 8.0). Conversely, PSE was subjected to a significant autolysis when incubated at the pH optimum. Figure 5(A) shows a peptide mass fingerprint of PSE autolyzate obtained in 50 mM ammonium acetate (pH 4.5). However, except for the peptides with *m/z* 1050 and 1412, the measured masses did not agree with the calculated prediction indicating the presence of nonspecific peptides in the autolyzate.

To obtain further information with respect to the cleavage specificity of PSE, a series of MS/MS experiments were performed. Collision-induced fragmentation of the autolytic peptides performed on a MALDI-TOF/TOF instrument provided peptide sequence data, which are summarized in Table 2 and illustrated by an example of MS/MS spectrum in Fig. 5(B). The results clearly indicate that PSE shows a broad cleavage site preference involving not only the sites P-X and A-X as has been previously published,[12] but also the C-termini of other amino acids: Gly, Ile, Asn, Gln, Arg, Ser and Tyr.

MALDI-TOF PMF and MS/MS analysis of proteins and peptides digested with PSE

The applicability of PSE for specific cleavage of protein samples was tested using a set of protein standards with molecular masses spanning the interval of 12–200 kDa. In the first trial, the proteins were subjected to in-solution digestion (the molecular mass of 60 kDa of PSE precluded in-gel digestion) that was performed without prior reduction/alkylation in 50 mM ammonium acetate at 37 °C overnight. Then aliquots of the digests were transferred on the target plate (a MicroScout 600 µm AnchorChip™ from Bruker Daltonik), mixed with CHCA matrix and dried out for MALDI-TOF PMF. Figure 6 shows a typical fingerprint mass spectrum of myoglobin digested with PSE. From the other spectra, especially those digests obtained from carbonic anhydrase, BSA and alcohol dehydrogenase were rich in peptides (not shown). To perform database search using the program Mascot Server 2.2, a cleavage specificity of PSE directed toward the C-termini of proline and alanine was set up in Mascot Configuration: Enzymes. The acquired mass spectrum of myoglobin was evaluated by searching the respective peak list against the database SwissProt. Myoglobin was assigned unambiguously (accession no. P68082 i.e. MYG_HORSE), when 13 peptides were identified giving a sequence coverage value of 65% and a score of 156. Bovine carbonic anhydrase was correctly assigned to SwissProt (accession no. P00921 i.e. CAH2_BOVIN) only when the taxonomy was restricted to mammals. Then ten peptides were identified, the sequence coverage was 49% and score 73. Fingerprints of the other standard proteins comprised numerous unspecific peptides and for that reason, their direct identification by database search was unsuccessful.

In the second experiment with protein digestion, samples were reduced and alkylated at a weakly basic pH of 8.0, then transferred to pH 4.5 and digested with PSE in a molar ratio of protein-to-PSE of 10 : 1 at 37 °C overnight. After a reversed-phase purification using ZipTip C_{18}, the digests were treated with CHCA matrix using the dried droplet method and finally MALDI-TOF mass spectra and MS/MS spectra were measured. In general, mass spectra containing many peaks were obtained and several of the most intense peaks were selected for MALDI-TOF/TOF tandem MS. A summary of MS/MS identified peptides is provided in Table 3. The results show that the cleavage after proline and alanine residues sometimes occurs together with a minor PSE action on several other amino acids. In addition to those cleavage sites uncovered using PSE autolyzate, several new targets were found (C-termini of Asp, Lys, Leu and Met). That is why the MS/MS analysis of PSE-derived peptides is necessary step for correct interpretation of obtained data. In contrast, the action on nonspecific position of analyzed proteins can aid in the proteomic analysis of protein sequences (or their parts) that are not accessible using standard proteomic approach involving, e.g. trypsin as digestion enzyme.

The digestion performance of PSE was also evaluated using a proline-rich peptide apidaecin 1A, naturally occurring in

Copyright © 2009 John Wiley & Sons, Ltd.

Table 2. MS/MS sequencing of PSE autolytic peptides

No.	Sequence	Observed mass (Da)[a,b]	Calculated mass (Da)	Difference (Da)
1	(Q)-YDPWRDS-(G)	938.387	937.393	−0.014
2	(G)-AVEHFDDFA-(A)	1050.438[c]	1049.445	−0.015
3	(N)-TTRLIWTNG-(Q)	1061.564	1060.567	−0.009
4	(I)-WTNGQYDPWR-(D)	1322.587	1321.584	−0.005
5	(A)-YFEQLLDHHNP-(E)	1412.651	1411.652	−0.008
6	(I)-WTNGQYDPWRDS-(G)	1524.644	1523.643	−0.007
7	(N)-TTRLIWTNGQYDP-(W)	1564.769	1563.768	−0.007
8	(N)-EVKQIKEWVEEYY-(A)	1742.854	1741.856	−0.010
9	(N)-EVKQIKEWVEEYYA-(—)	1813.897	1812.893	−0.004
10	(G)-TFSQRYWWSTEYWGGP-(G)	2050.900	2049.901	−0.008
11	(N)-TTRLIWTNGQYDPWRDS-(G)	2109.018	2108.008	0.003

The peptides were individually selected from peptide mass fingerprint of PSE autolyzate (Fig. 5(A)) and then subjected to the collision-induced dissociation by air on 4800 Proteomics Analyzer MALDI-TOF/TOF mass spectrometer. The MS/MS data were processed using 4000 Series Explorer software.
[a] Monoisotopic mass, $[M+H]^+$.
[b] Further peptides matching the sequence: *m/z* 906.395, 1181.523, 1434.629, 1450.611, 1546.757, 1721.775 and 1937.829.
[c] Isobaric with a predicted peptide (A)-VEHFDDFAA-(V).

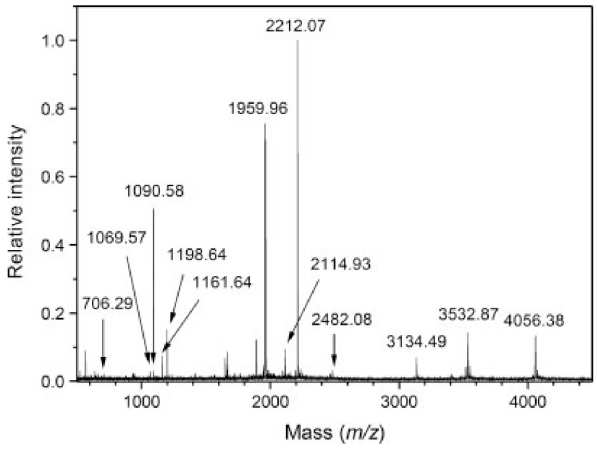


Figure 6. MALDI-TOF PMF using PSE. Horse heart myoglobin was in-solution digested with PSE without prior reduction/alkylation. The presented mass spectrum of peptides was acquired on the Microflex LRF-20 instrument operating in the reflectron mode for positive ions. CHCA was used as a matrix.

bee hemolymph, where it has an important role in humoral defense against microbial invasion.[36] The apidaecin 1A sequence consists of 18 amino acids including six proline residues (GNNRPVYIPQPRPPHPRI). In MALDI-TOF MS, the peptide provides a positive quasimolecular ion $[M+H]^+$ with *m/z* 2108.2. An aliquot of apidaecin in 50 mM ammonium acetate was in-solution digested in a molar ratio of peptide-to-PSE of 5 : 1 at 37 °C overnight. The digest was concentrated in vacuum centrifuge and purified by ZipTip C_{18}. MALDI-TOF mass spectra showed specific peptides with *m/z* 522.3, 557.3, 594.3, 716.4, 1029.5 and 1254.7 (Fig. 7) in accordance with a prediction by the web application Protein Cutter. The cleavage peptides were analyzed by MALDI PSD yielding the sequences HPRI, GNNRP, QPRPP (inset in Fig. 7; *m/z* 594.3), VYIPQP, GNNRPVYIP and GNNRPVYIPQP, respectively. This clearly demonstrated that the enzymatic cleavage took place at the C-terminus of almost every proline residue. Predicted small dipeptide and tripeptide fragments could not be assigned reliably because of high signal background in the area of matrix adducts formation.

Conclusions

A. niger PSE with an acidic pH optimum was isolated from Brewers Clarex. Our purification procedure represents a protocol for removing additives and autolytic products from the commercial material. During characterization of the final enzyme preparation, a significant degree of glycosylation was discovered. Mass spectrometric analysis of PSE-derived glycopeptides and *N*-glycans released by the use of *N*-glycosidase F allowed determination of sugar structures. PSE contains high-mannose type *N*-glycans that are partially phosphorylated. The enzyme was found useful for proteomic research involving protein sample digestion. Due to its relatively high molecular mass and significant glycosylation, the enzyme is applicable for in-solution digestion only. The cleavage specifity of PSE covers a major action on the C-termini of proline and alanine residues, which is accompanied by a minor action on other cleavage sites. For this reason, the use of PSE for MALDI-TOF PMF seems to be problematic but there is a potential of its applicability in tandem MS of peptides including *de novo* sequencing, for example in shotgun proteomics[37] of samples containing proline-rich proteins. Nonspecific peptides released by PSE digestion may provide benefits in studying posttranslational modifications especially in such cases where the modification site is not present on a properly sized tryptic peptide.

Acknowledgements

This work was supported by the grant MSM6198959216 from the Ministry of Education, Youth and Sports, Czech Republic and institutional research plan MO0FVZ0000501 of Faculty of Military Health Sciences, University of Defence, Czech Republic. The authors thank Dr Marek Petřivalský from the Department of Biochemistry, Faculty of Science, Palacký University in Olomouc for providing us with apidaecin sample. Our results are dedicated to the initiator of the study, Dr Josef 'Pepa' Chmelík, who passed away in 2007.

Copyright © 2009 John Wiley & Sons, Ltd.

Table 3. MS/MS sequencing of peptides derived from PSE digestion of protein standards

Identified protein/accession no.	Sequence	Observed mass (Da)[a]	Calculated mass (Da)	Difference (Da)
Cytochrome c/P00004	(P)-NLHGLFGRKTG-(Q)	1199.674	1198.657	0.010
	(L)-MEYLENPKKYIP-(G)	1524.786	1523.769	0.009
	(L)-IAYLKKATNE-(−)	1150.650	1149.639	0.004
Lysozyme C/P00698	(G)-KVFGRCELA-(A)	1079.577	1078.559	0.010
	(P)-GSRNLCNIP-(C)	1030.521	1029.503	0.011
Myoglobin/P68082	(P)-ETLEKFDKFK-(H)	1284.689	1283.676	0.005
	(P)-ETLEKFDKFKHLKTEA-(E)	1964.056	1963.041	0.007
	(A)-SEDLKKHGTVVLTALG-(G)	1667.938	1666.925	0.006
	(G)-GILKKKGHHEA-(E)	1217.717	1216.704	0.005
	(S)-HATKHKIP-(I)	931.552	930.540	0.005
	(P)-IKYLEFISDA-(I)	1198.639	1197.628	0.004
	(A)-IIHVLHSKHP-(G)	1180.703	1179.688	0.008
Carbonic anhydrase 2/P00921	(P)-EHWHKDFP-(I)	1095.501	1094.493	0.000
	(P)-IANGERQSPVDIDTK-(A)	1642.841	1641.832	0.001
	(P)-IANGERQSPVDIDTKAVVQDP-(A)	2252.160	2251.144	0.008
	(P)-LKNRQVRG-(F)	970.592	969.583	0.001
	(P)-LKNRQVRGFP-(K)	1214.714	1213.704	0.002
Serum albumin/P02769	(A)-SIQKFGERA-(L)	1035.563	1034.551	0.005
	(S)-VARLSQKFP-(K)	1045.621	1044.608	0.006
	(A)-RLSQKFP-(K)	875.508	874.502	−0.001
Alcohol dehydrogenase 1/P00330	(P)-LVGGHEGAGVVVG-(M)	1150.651	1149.614	0.029
	(P)-LVGGHEGAGVVVGM-(G)	1281.693	1280.655	0.031
	(K)-IGDYAGIK-(W)	836.475	835.444	0.023
	(S)-IGGEVFIDFTK-(E)	1225.676	1224.639	0.029
	(M)-EKGQIVGRY-(V)	1049.602	1048.567	0.028
Beta-amylase/P10537	(S)-GLTYLRLSDDLLQTD-(N)	1722.897	1721.884	0.006
	(D)-ELMEATKGSRPFPWYDVTDMPVD-GSNPFD-(−)	3301.543	3300.474	0.061
	(P)-FPWYDVTDMPVDGSNPFD-(−)	2101.879	2100.877	−0.006

Peptides were selected from MALDI-TOF MS spectra of digests purified using ZipTip C_{18} pipette tips and then subjected to the collision-induced dissociation by air on 4800 Proteomics Analyzer MALDI-TOF/TOF mass spectrometer. Only signals confirmed by successful MS/MS analysis are reported in the table. Accession numbers refer to SwissProt database.
[a] Monoisotopic mass, $[M+H]^+$.

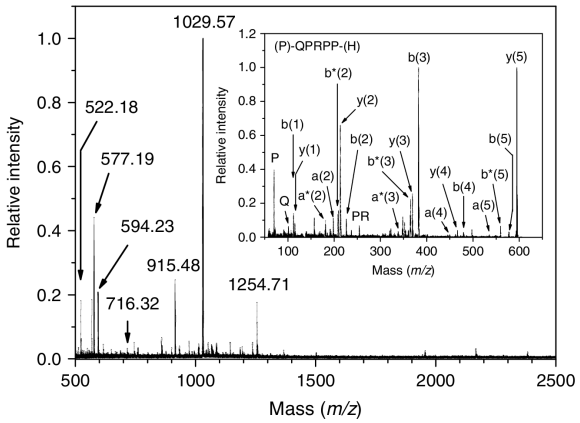


Figure 7. In-solution digestion of apidaecin 1A by PSE. The main panel shows a MALDI-TOF mass spectrum of a PSE digest of apidaecin that was acquired using the Microflex LRF-20 instrument operating in the reflectron mode for positive ions (CHCA matrix). The inset displays a PSD spectrum of the peptide fragment with *m/z* 594.23, which yields the sequence (P)-QPRPP-(H) [Correction made here after initial online publication]. The complete apidaecin 1A is characterized by the amino acid sequence GNNRPVYIPQPRPPHPRI.

References

[1] M. L. Abarca, F. Accensi, J. Cano, F. J. Cabañes. Taxonomy and significance of black aspergilli. *Antonie Van Leeuwenhoek* **2004**, *86*, 33. DOI: 10.1023/B:ANTO.0000024907.85688.05.

[2] M. A. Nalesnik, R. L. Myerowitz, R. Jenkins, J. Lenkey, D. Herbert. Significance of *Aspergillus* species isolated from respiratory secretions in the diagnosis of invasive pulmonary aspergillosis. *Journal of Clinical Microbiology* **1980**, *4*, 370.

[3] D. A. Stevens, V. L. Kan, M. A. Judson, V. A. Morrison, S. Dummer, D. W. Denning, J. E. Bennett, T. J. Walsh, T. F. Patterson, G. A. Pankey. Practice guidelines for diseases caused by *Aspergillus*. *Clinical Infectious Diseases* **2000**, *30*, 696. DOI: 10.1086/313756.

[4] J. Winkler, U. Müller, P. Nenoff, H.-J. Seyfarth, M. Vogtmann, G. Borte, W. Pönisch, T. Kahn, H. Wirtz, J. Schauer, G. Hoheisel. Treatment of invasive pulmonary Aspergillosis in neutropenic patients by additional bronchoscopic amphotericin B instillation. *Respiration* **2007**, *74*, 663. DOI: 10.1159/000105385.

[5] T. Tamaki, Y. Takamiya, T. Miyagi, T. Nishiya. Studies on the maturation of Awamori. VII. Changes in ester compounds and higher alcohols of Awamori during aging. *Journal of Fermentation Technology* **1986**, *64*, 17. DOI: 10.1016/0385-6380(86)90052-X.

[6] E. Schuster, N. Dunn-Coleman, J. C. Frisvad, P. W. M. van Dijck. On the safety of *Aspergillus niger* – a review. *Applied Microbiology and Biotechnology* **2002**, *59*, 426. DOI: 10.1007/s00253-002-1032-6.

[7] S. J. Romero-Gómez, C. Augur, G. Viniegra-González. Invertase production by *Aspergillus niger* in submerged and solid-

Copyright © 2009 John Wiley & Sons, Ltd.

state fermentation. *Biotechnology Letters* **2000**, *22*, 1255. DOI: 10.1023/A:1005659217932.

[8] P. J. Punt, N. van Biezen, A. Conesa, A. Albers, J. Mangnus, C. van den Hondel. Filamentous fungi as cell factories for heterologous protein production. *Trends in Biotechnology* **2002**, *20*, 200. DOI: 10.1016/S0167-7799(02)01933-9.

[9] T. S. Bhalerao, P. R. Puranik. Biodegradation of organochlorine pesticide, endosulfan, by a fungal soil isolate, *Aspergillus niger*. *International Biodeterioration and Biodegradation* **2007**, *59*, 315. DOI: 10.1016/j.ibiod.2006.09.002.

[10] S. E. Baker. *Aspergillus niger* genomics: past, present and into the future. *Medical Mycology* **2006**, *44*, S17. DOI: 10.1080/13693780600921037.

[11] H. J. Pel, J. H. de Winde, D. B. Archer, P. S. Dyer, G. Hofmann, P. J. Schaap, G. Turner, R. P. de Vries, R. Albang, K. Albermann, M. R. Andersen, J. D. Bendtsen, J. A. E. Benen, M. van den Berg, S. Breestraat, M. X. Caddick, R. Contreras, M. Cornell, P. M. Coutinho, E. G. J. Danchin, A. J. M. Debets, P. Dekker, P. W. M. van Dijck,, A. van Dijk, L. Dijkhuizen, A. J. M. Driessen, C. d'Enfert, S. Geysens, C. Goosen, G. S. P. Groot, P. W. J. de Groot, T. Guillemette, B. Henrissat, M. Herweijer, J. P. T. W. van den Hombergh, C. A. M. J. J. van den Hondel, R. T. J. M. van der Heijden, R. M. van der Kaaij, F. M. Klis, H. J. Kools, C. P. Kubicek, P. A. van Kuyk, J. Lauber, X. Lu, M. J. E. C. van der Maarel, R. Meulenberg, H. Menke, M. A. Mortimer, J. Nielsen, S. G. Oliver, M. Olsthoorn, K. Pal, N. N. M. E. van Peij, A. F. J. Ram, U. Rinas, J. A. Roubos, C. M. J. Sagt, M. Schmoll, J. Sun, D. Ussery, J. Varga, W. Vervecken, P. J. J. van de Vondervoort, H. Wedler, H. A. B. Wösten, A. P. Zeng, A. J. J. van Ooyen, J. Visser, H. Stam. Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. *Nature Biotechnology* **2007**, *25*, 221. DOI: 10.1038/nbt1282.

[12] L. Edens, P. Dekker, R. van der Hoeven, F. Deen, A. de Roos, R. Floris. Extracellular prolyl endoprotease from *Aspergillus niger* and its use in the debittering of protein hydrolysates. *Journal of Agricultural and Food Chemistry* **2005**, *53*, 7950. DOI: 10.1021/jf050652c.

[13] D. Stepniak, L. Spaenij-Dekking, C. Mitea, M. Moester, A. de Ru, R. Baak-Pablo, P. van Veelen, L. Edens, F. Koning. Highly efficient gluten degradation with a newly identified prolyl endoprotease: implications for celiac disease. *American Journal of Physiology Gastrointestinology and Liver Physiology* **2006**, *291*, G621. DOI: 10.1152/ajpgi.00034.2006.

[14] M. Lopez, L. Edens. Effective prevention of chill-haze in beer using an acid proline-specific endoprotease from *Aspergillus niger*. *Journal of Agricultural and Food Chemistry* **2005**, *53*, 7944. DOI: 10.1021/jf0506535.

[15] R. Walter, W. H. Simmons, T. Yoshimoto. Proline specific endo- and exopeptidase. *Molecular and Cellular Biochemistry* **1980**, *30*, 111. DOI: 10.1007/BF00227927.

[16] J. Gass, C. Khosla. Prolyl endopeptidases. *Cellular and Molecular Life Sciences* **2007**, *64*, 345. DOI: 10.1007/s00018-006-6317-y.

[17] J. S. Rosenblum, J. W. Kozarich. Prolyl peptidases: a serine protease subfamily with high potential for drug discovery. *Current Opinion in Chemical Biology* **2003**, *7*, 496. DOI: 10.1016/S1367-5931(03)00084-X.

[18] J. Leprince, D. Cosquer, G. Bellemère, D. Chatenet, H. Tollemer, S. Jégou, M. C. Tonon, H. Vaudry. Catabolism of the octadecaneuropeptide ODN by prolyl endopeptidase: identification of an unusual cleavage site. *Peptides* **2006**, *27*, 1561. DOI: 10.1016/j.peptides.2005.11.024.

[19] M. G. Jones. The first filamentous fungal genome sequences: Aspergillus leads the way for essential everyday resources or dusty museum specimens? *Microbiology* **2007**, *153*, 1. DOI: 10.1099/mic.0.2006/001479-0.

[20] S. Blumberg, V. I. Teichberg, J. L. Charli, L. B. Hersh, J. F. McKelvy. Cleavage of substance P to an N-terminal tetrapeptide and a C-terminal heptapeptide by a post-proline cleaving enzyme from bovine brain. *Brain Research* **1980**, *192*, 477. DOI: 10.1016/0006-8993(80)90898-7.

[21] K. Hauzer, T. Barth, L. Servitova, K. Jost. Post-proline endopeptidase – partial purification and characterization of the enzyme from pig kidneys. *Collection of Czechoslovak Chemical Communications* **1984**, *49*, 1846.

[22] T. C. McIlvaine. A buffer solution for colorimetric comparison. *Journal of Biological Chemistry* **1921**, *49*, 183.

[23] P. K. Smith, R. I. Krohn, G. T. Hermanson, A. K. Mallia, F. H. Gartner, M. D. Provenzano, E. K. Fujimoto, N. M. Goeke, B. J. Olson, D. C. Klenk. Measurement of protein using bicinchoninic acid. *Analytical Biochemistry* **1985**, *150*, 76. DOI: 10.1016/0003-2697(85)90442-7.

[24] U. K. Laemmli. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **1970**, *227*, 680. DOI: 10.1038/227680a0.

[25] O. Vorm, P. Roepstorff, M. Mann. Improved resolution and very high sensitivity in MALDI TOF of matrix surfaces made by fast evaporation. *Analytical Chemistry* **1994**, *66*, 3281. DOI: 10.1021/ac00091a044.

[26] H. Thomas, J. Havliš, J. Peychl, A. Shevchenko. Dried-droplet probe preparation on AnchorChip™ targets for navigating the acquisition of matrix-assisted laser desorption/ionization time-of-flight spectra by fluorescence of matrix/analyte crystals. *Rapid Communications in Mass Spectrometry* **2004**, *18*, 923. DOI: 10.1002/rcm.1427.

[27] P. Halada, C. Leitner, P. Sedmera, D. Haltrich, J. Volc. Identification of the covalent flavin adenine dinucleotide-binding region in pyranose 2-oxidase from *Trametes multicolor*. *Analytical Biochemistry* **2003**, *314*, 235. DOI:10.1016/S0003-2697(02)00661-9.

[28] M. DuBois, K. A. Gilles, J. K. Hamilton, P. A. Rebers, F. Smith. Colorimetric method for determination of sugars and related substances. *Analytical Chemistry* **1956**, *28*, 350. DOI: 10.1021/ac60111a017.

[29] N. H. Packer, M. A. Lawson, D. R. Jardine, J. W. Redmond. A general approach to desalting oligosaccharides released from glycoproteins. *Glycoconjugate Journal* **1998**, *15*, 737. DOI: 10.1023/A:1006983125913.

[30] D. Kopečný, C. Pethe, M. Šebela, N. Houba-Hérin, C. Madzak, A. Majira, M. Laloue. High-level expression and characterization of *Zea mays* cytokinin oxidase/dehydrogenase in *Yarrowia lipolytica*. *Biochimie* **2005**, *87*, 1011. DOI: 10.1016/j.biochi.2005.04.006.

[31] D. I. Papac, A. Wong, A. J. S. Jones. Analysis of acidic oligosaccharides and glycopeptides by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Analytical Chemistry* **1996**, *68*, 3215. DOI: 10.1021/ac960324z.

[32] Y. Huang, Y. Mechref, M. V. Novotny. N-Linked oligosaccharide structures in the diamine oxidase from porcine kidney. *Carbohydrate Research* **2000**, *323*, 111. DOI: 10.1016/S0008-6215(99)00254-2.

[33] C. A. Cooper, E. Gasteiger, N. H. Packer. GlycoMod – a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* **2001**, *1*, 340. DOI: 10.1002/1615-9861(200102)1:2<340::AID-PROT340>3.0.CO;2-B.

[34] A. Dell, H. R. Morris. Glycoprotein structure determination by mass spectrometry. *Science* **2001**, *291*, 2351. DOI: 10.1126/science.1058890.

[35] S. M. Haslam, H. Morris, A. Dell. Mass spectrometric strategies: providing structural clues for helminth glycoproteins. *Trends in Parasitology* **2001**, *17*, 231. DOI: 10.1016/S1471-4922(00)01860-2.

[36] K. Casteels-Josson, T. Capaci, P. Casteels, P. Tempst. Apidaecin multipeptide precursor structure: a putative mechanism for amplification of the insect antibacterial response. *EMBO Journal* **1993**, *12*, 1569.

[37] G. T. Cantin, J. R. Yates III. Strategies for shotgun identification of post-translational modifications by mass spectrometry. *Journal of Chromatography A* **2004**, *1053*, 7. DOI: 10.1016/j.chroma.2004.06.046.

 Copyright © 2009 John Wiley & Sons, Ltd.

Příloha B – článek o aplikaci ProteinCutter

Raus M, Kopečný D, Šebela M (2013) *Program application for the prediction of results of protein digestion by proteolytic enzymes [Programová aplikace pro predikci výsledků štěpení proteinu proteolytickými enzymy]*. Chem. Listy 107 (1), 44-53.

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

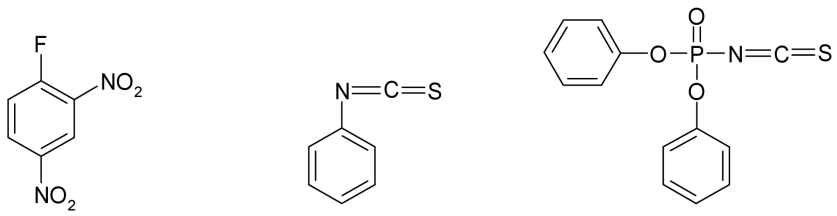
Obsah

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



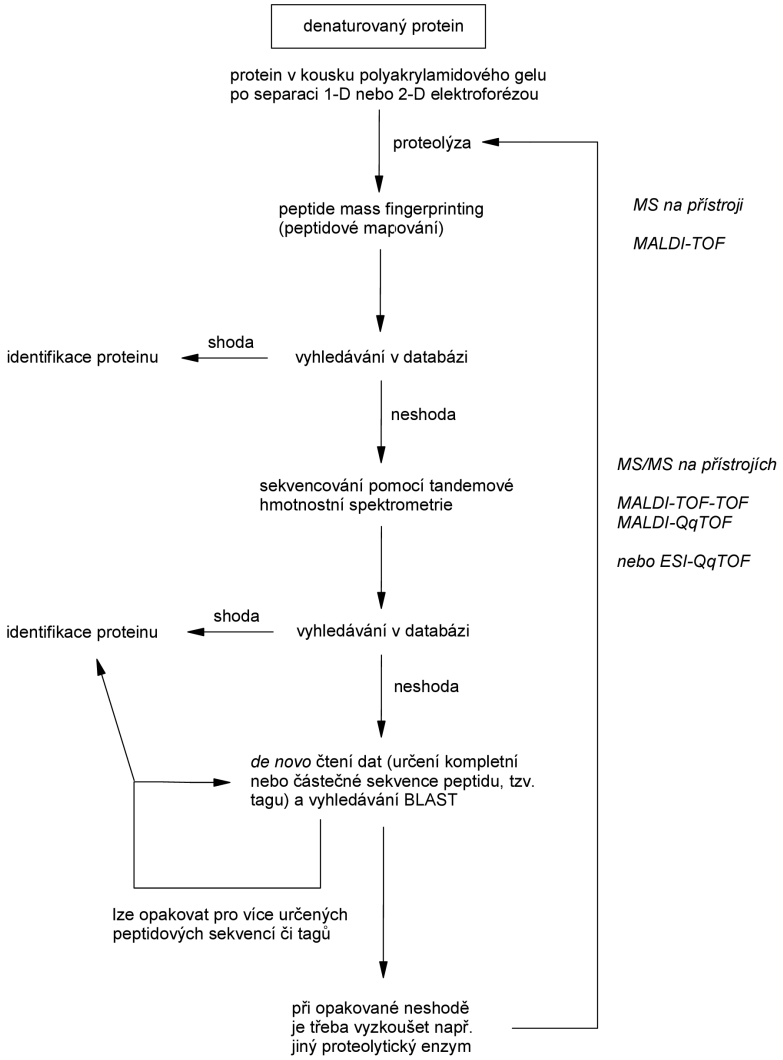
Obr. 1. **Chemická činidla pro sekvenční analýzu proteinů: 2,4-dinitro-1-fluorbenzen (DNFB), fenylisothiokyanát (PITC) a difenylfosforylisothiokyanát (DPPITC)**

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].

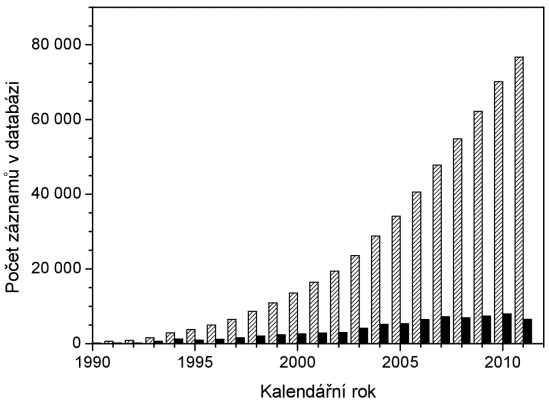
S rostoucím množstvím dat v centralizovaných databázích se zrodila i nová vědecká disciplína. Bioinformatika byla původně úzce spojena s genetikou a genomikou, a to díky genomovým projektům a výsledkům automatizovaného čtení a skládání částečných sekvencí komplementárních DNA (angl. „expressed sequence 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

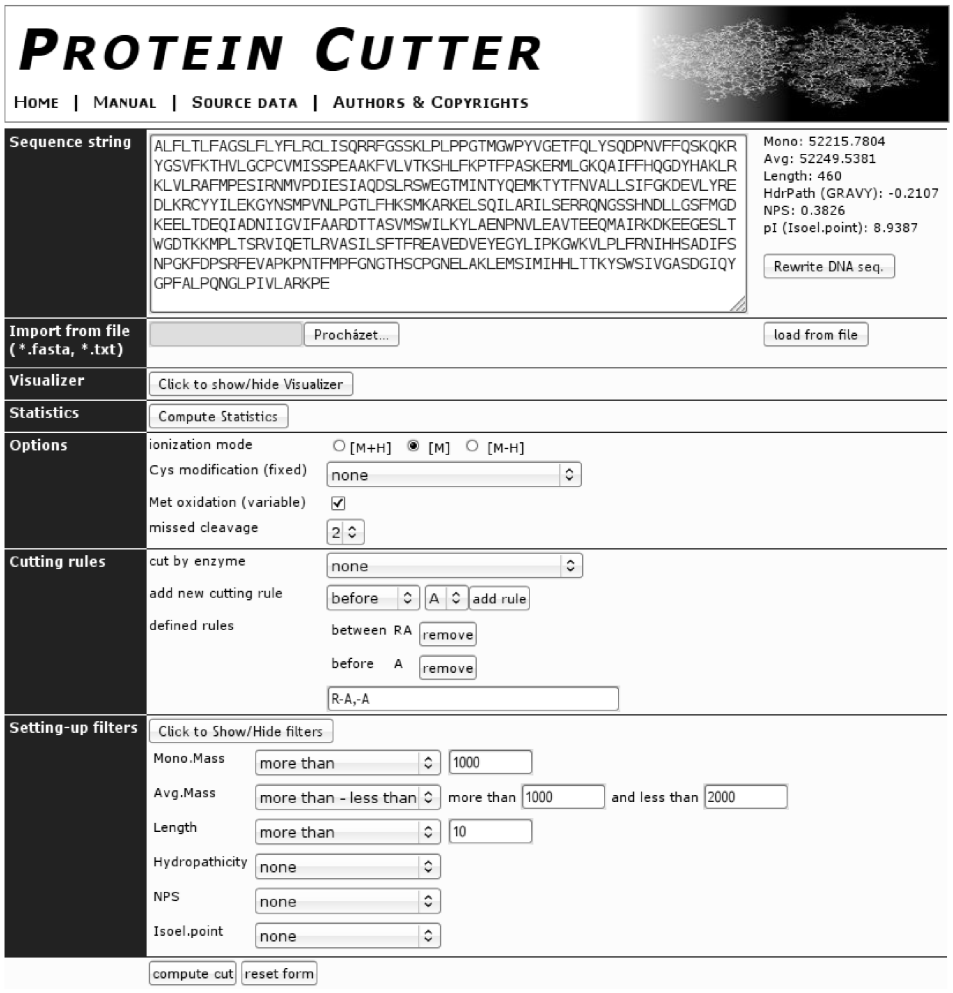
6.1. Představení aplikace

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

Příloha C – publikované výsledky (Izotop)

Tarkowski P., Flokova K., Václavikova K., Jaworek P., **Raus M**., Nordström A., Novak O., Dolezal K., Sebela M., Frebortova J. (2010) *An improved in vivo deuterium labeling method for measuring the biosynthetic rate of cytokinins*. Molecules 15 (12), 9214-9229.

Molecules **2010**, *15*, 9214-9229; doi:10.3390/molecules15129214

OPEN ACCESS

molecules

ISSN 1420-3049

www.mdpi.com/journal/molecules

Article

An Improved *in Vivo* Deuterium Labeling Method for Measuring the Biosynthetic Rate of Cytokinins

Petr Tarkowski [1,2,*], Kristýna Floková [2], Kateřina Václavíková [2], Pavel Jaworek [2], Martin Raus [2], Anders Nordström [3], Ondřej Novák [4], Karel Doležal [1,4], Marek Šebela [1,2] and Jitka Frébortová [1,4]

[1] Centre of the Region Haná for Biotechnological and Agricultural Research, Faculty of Science, Palacký University, Šlechtitelů 11, 783 71 Olomouc, Czech Republic; E-Mails: karel.dolezal@upol.cz (K.D.); marek.sebela@upol.cz (M.Š.); jitka.frebortova@upol.cz (J.F.)

[2] Department of Biochemistry, Faculty of Science, Palacký University, Šlechtitelů 11, 783 71 Olomouc, Czech Republic; E-Mails: DrobRangers@seznam.cz (K.F.); katka.vaclavik@seznam.cz (K.V.); p.jaworek@seznam.cz (P.J.); martin_raus@post.cz (M.R.)

[3] Department for Oncology–Pathology, Karolinska Biomics Center, Karolinska Institutet, Z5:02, 17176 Stockholm, Sweden; E-Mail: anders.nordstrom@ki.se (A.N.)

[4] Laboratory of Growth Regulators, Palacký University and Institute of Experimental Botany ASCR, Šlechtitelů 11, 783 71 Olomouc, Czech Republic; E-Mail: ondrej.novak@upol.cz (O.N.)

* Author to whom correspondence should be addressed; E-Mail: petr.tarkowski@upol.cz; Tel.: +420-585-634-928; Fax: +420-585-634-933.

Received: 10 October 2010 / Accepted: 14 December 2010 / Published: 15 December 2010

Abstract: An improved method for determining the relative biosynthetic rate of isoprenoid cytokinins has been developed. A set of 11 relevant isoprenoid cytokinins, including zeatin isomers, was separated by ultra performance liquid chromatography in less than 6 min. The iP-type cytokinins were observed to give rise to a previously-unknown fragment at *m/z* 69; we suggest that the diagnostic (204-69) transition can be used to monitor the biosynthetic rate of isopentenyladenine. Furthermore, we found that by treating the cytokinin nucleotides with alkaline phosphatase prior to analysis, the sensitivity of the detection process could be increased. In addition, derivatization (propionylation) improved the ESI-MS response by increasing the analytes' hydrophobicity. Indeed, the ESI-MS response of propionylated isopentenyladenosine was about 34% higher than that of its underivatized counterpart. Moreover, the response of the derivatized zeatin ribosides was about 75% higher than that of underivatized zeatin ribosides. Finally, we created a web-based

calculator (IZOTOP) that facilitates MS/MS data processing and offer it freely to the research community.

Keywords: cytokinin; deuterium labelling; biosynthetic rate; UPLC; MS

1. Introduction

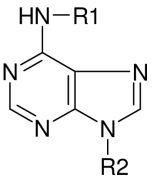
The cytokinins are a group of plant hormones that play a central role in the regulation of cell division and differentiation. They control processes as diverse as apical dominance, root formation, leaf senescence, stomatal behavior, and chloroplast development [1]. Although their biological significance has been known for decades, the basic molecular mechanisms of cytokinin action have only recently been elucidated [2]. Structurally, cytokinins are adenine derivatives containing either an isoprenoid or an aromatic chain at the N6 position of the adenine ring. They mainly occur as free bases, nucleosides, nucleotides, and as a number of sugar conjugates substituted at the N7 and N9 positions of the purine ring (*N*-glucosides) or at the hydroxylated side chain (*O*-glucosides). However, only the free bases are thought to have significant biological activity. Accordingly, the interaction between the cytokinin ligand and its receptor is strongly influenced by the character of the side-chain. Nevertheless, cytokinin-receptor assays have shown that at least some receptors also respond to nucleosides and nucleotides [3].

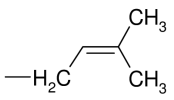
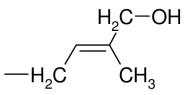
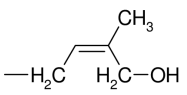
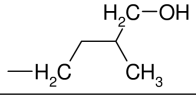
The first dedicated step in cytokinin biosynthesis - *N*-prenylation of adenosine 5'-phosphates (AMP, ADP or ATP) with dimethylallyl diphosphate or hydroxymethylbutenyl diphosphate as side-chain donor - is catalyzed by the appropriate isopentenyltransferase (IPT; EC 2.5.1.27). The substrate specificity of IPTs varies depending on their origin and the species from which they are derived. In *Arabidopsis thaliana*, IPTs predominantly use ADP or ATP rather than AMP as prenyl acceptors, forming isopentenyladenosine 5'-diphosphate and 5'-triphosphate [4,5]. The isopentenyl side-chain can subsequently be *trans*-hydroxylated by cytochrome P450 monooxygenases to form *trans*-zeatin-type nucleotides [6]. In addition, tRNA IPTs, another group of cytokinin-generating enzymes, catalyze the prenylation of specific tRNAs and are responsible for the production of cis-zeatin-type cytokinins [7].

Plant tissue extracts are complex multi-component mixtures that contain cytokinins in minute quantities (fmol to pmol per gram of fresh weight) along with other compounds with similar structures and/or physico-chemical properties. Therefore, mapping the cytokinin composition of plants requires the use of sensitive and highly selective analytical methods. Hyphenated techniques such as ultra performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS) and capillary electrophoresis-tandem mass spectrometry (CE-MS/MS) have revolutionized trace analysis of cytokinins [8]. These methods combine a high-resolution separation system with a powerful detection and characterization technique. A high resolving power is necessary to separate cytokinin isomers, which differ significantly in their biological activity; mass spectrometric detection is essential because the sample matrix contains potentially interfering compounds at much higher concentrations than the target analytes.

The most common method to assess biosynthetic activity in biological systems involves the use of labeled precursors. Previously, both radioactive and stable-isotope precursors have been applied to elucidate cytokinin biosynthesis in various plant species [9,10]. Because the downstream analyses, scintillation and mass spectrometry, are very sensitive, labeled precursors can be added to the biological material at very low concentrations. This represents a significant advantage over classic feeding experiments, which use labeled precursors in concentrations high enough to trigger the activation of catabolic or interconversion enzymes and thus alter the metabolism of the studied compounds. Åstot and co-workers used an alternative method –*in vivo* deuterium labeling [11]. By growing plants on liquid cultivation media containing 30% D_2O, deuterium was incorporated into general metabolic pathways and newly synthesized molecules thus labeled without altering their steady state levels [12]. The cytokinins were then isolated and the extent of their labeling was analyzed by liquid chromatography/frit-fast atom bombardment mass spectrometry. To increase the hydrophobicity of the target molecules, cytokinins were derivatized by propionylation prior to the final analysis [13]. A modified version of this approach, in which fast-atom bombardment was replaced by electrospray ionization, was subsequently used to study the crosstalk between auxins and cytokinins [14]. In the work described in this paper, we have focused on isoprenoid cytokinin bases, nucleosides and nucleotides (Table 1). We improved the method described above by using ultra performance liquid chromatography to separate 11 relevant cytokinins, including zeatin isomers.

Table 1. Structures, common names and abbreviations of the compounds used in this study.



R_1	R_2	Common name	Abbreviation
	H	isopentenyladenine	iP
	R	isopentenyladenosine	iPR
	RMP	isopentenyladenosine-5′-monophosphate	iPMP
	H	*trans*-zeatin	*t*Z
	R	*trans*-zeatin riboside	*t*ZR
	RMP	*trans*-zeatin riboside-5′-monophosphate	*t*ZMP
	H	*cis*-zeatin	*c*Z
	R	*cis*-zeatin riboside	*c*ZR
	RMP	*cis*-zeatin riboside-5′-monophosphate	*c*ZMP
	H	dihydrozeatin	DHZ
	R	dihydrozeatin riboside	DHZR

H: hydrogen; R: β-D-ribose; RMP: β-D-ribose-5′-monophosphate.

In addition, by treating the cytokinin nucleotides with alkaline phosphatase followed by immunoaffinity purification of the dephosphorylated ribosides, we increased the sensitivity of the mass spectrometric detection. This method was used to measure the rate of synthesis of selected isoprenoid cytokinins isolated from the cytokinin-overproducing Arabidopsis line *PGA22*, and the results so obtained were compared to those obtained using an alternative method based on the analysis of underivatized cytokinins. Finally, we created a web-based calculator to process mass spectrometric data and convert them to tracer:tracee ratios (t/t), as described by Åstot *et al.* [11].

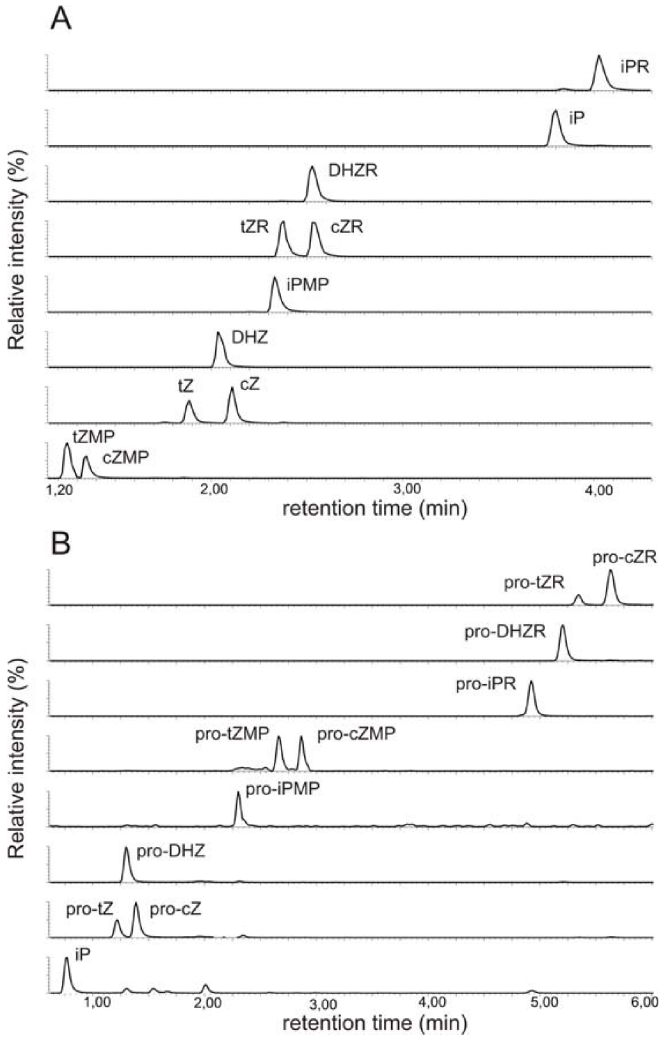
2. Results and Discussion

2.1. Implementation of ultra performance liquid chromatography expedites cytokinin analysis

As analytical methods evolve, it becomes necessary to perform increasing numbers of separations. To make it easier to cope with this growing need, we sought to improve on the current HPLC method for the separation of propionylated cytokinins [11]. Generally, there are three main approaches to reducing the time required for a given LC separation: (i) the use of monolith columns, (ii) liquid chromatography at high temperatures and (iii) liquid chromatography at ultra-high pressures using columns packed with sub-2-micron particles (UPLC) [15]. We have previously used UPLC with a 50 mm column (BEH C18, 50 × 2.1 mm; 1.7 μm particle size; Waters) and 15% ammonium formate/methanol as the mobile phase buffers to separate 21 underivatized cytokinins [17]. However, isomer co-elution made it impossible to separate propionylated cytokinins under these chromatographic conditions. The various isomers of the zeatin-type cytokinins have identical molecular masses and give rise to identical basic fragments under MS/MS conditions, highlighting the need for efficient methods for their chromatographic separation.

On the basis of this experience, we sought to develop a UPLC method using the same 50 mm column for the separation of propionylated cytokinins. A gradient elution program was designed using the Acquity UPLC columns calculator (Masslynx v.4.1, Waters) and slightly adjusted to achieve a baseline separation of 11 isoprenoid cytokinins, including free bases, nucleosides and nucleotides. The HPLC method previously used for such separations employed a 150 mm column and took 56 min; the new UPLC method uses a 50 mm column and has a chromatographic separation time of less than 6 min. The gradient elution buffers used in this method (water and acetonitrile) typically contain 3% of added formic acid [16], which causes ion-source pollution. This decreases the sensitivity of the mass spectrometer and necessitates daily cleaning of the ion-source. By reducing the formic acid content of the mobile phase to 0.7%, we were able to significantly reduce the contamination of the mass spectrometer ion-source without affecting the quality of the chromatographic separation. Decreasing the formic acid content of the mobile phases to less than 0.7% or raising their pH resulted in the loss of chromatographic resolution. Although the selectivity of this chromatographic system differs from that of the system developed for the separation of underivatized cytokinins, both approaches can be used to separate 11 cytokinins in less than 6 min (Figure 1).

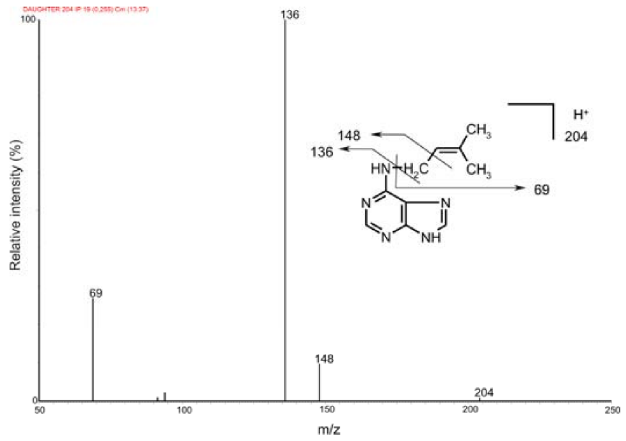
Figure 1. Separation of a mixture of cytokinin standards by ultra-performance liquid chromatography (UPLC). The figure shows reconstituted mass chromatograms of 11 underivatized (A) and 11 propionylated (B) cytokinins; each trace represents one MRM transition. The mixture contained 1 pmol of each metabolite.



2.2. Mass spectrometric detection

We began by recording positive electrospray-ionization mass spectra for all 11 propionylated cytokinins. The background-subtracted ESI+ spectra of all of the studied compounds contained base peaks corresponding to the quasi-molecular ions [M+H]+ and no adduct formation was observed under the acidic conditions (data not shown). Moreover, the fragmentation patterns were consistent with previously acquired FAB MS spectra [13]. However, we observed a previously neglected sidechain fragment at *m/z* 69 in the spectra of various iP-type cytokinins, including the free bases, ribotides, ribosides and glucosides (Figure 2).

Figure 2. Collision mass spectrum of isopentenyladenine and its fragmentation pattern.



To confirm the identity of this fragment we performed an exact mass determination of the fragments of isopentenyladenine (iP) and its deuterium-labeled counterpart (D_6-iP) by quadrupole-time-of-flight mass spectrometry. We reasoned that if unlabelled iP gives a side-chain fragmnent at *m/z* 69, D_6-iP should give rise to an analogous fragment at *m/z* 75 due to the incorporation of six deuterons into the terminal methyl groups of the isopentenyl moiety. The results obtained are summarized in Table 2.

Table 2. The identification of an isopentenyladenine side-chain fragment by exact mass determination using a Q-TOF mass spectrometer. Experimental values are compared to those calculated from the fragments' chemical formulae. The difference between the two values (in ppm) was calculated as a measure of fidelity.

Features	**iP**	**D_6-iP**
Selected fragment formula	$C_5H_9^+$	$C_5H_3D_6^+$
Experimental *m/z*	69.0703	75.1081
Calculated *m/z*	69.0704	75.1081
Difference (ppm)	-1.4	0.0

The diagnostic transition 204-69 could potentially be used to monitor the labeling of isopentenyladenine with deuterium. It is better suited for this purpose than is the 204-136 transition, whose magnitude reflects the extent of deuterium incorporation into the adenine moiety rather than the labeling of the sidechain; while labeling of the adenine moiety may be of interest in other contexts, it is not relevant in studies of cytokinin biosynthesis. The diagnostic quasi-molecular ion – fragment ion transition at *m/z* 69 was observed in the mass spectra of other iP type cytokinins and so may be of use as a general indicator of deuterium labeling in these species. Because the focus of this study was on the primary products of cytokinin biosynthesis, *i.e.* cytokinin nucleotides and their ribosides, we did not investigate the extent of deuterium incorporation into the free bases.

The *in vivo* deuterium labeling method exploits the ability of intact plants to take up the tracer, which easily penetrates all cell compartments and enters the general metabolic pathways, giving rise to labeled precursors. Åstot and co-workers showed that the extent to which deuterium is incorporated into the main structural components of the cytokinins depends on the relative biosynthetic rates of the different pathways that produce them (note that "biosynthetic rate" here and throughout this paper

refers to the measured degree of labeling of the compound rather than the exact rate of its biosynthesis) [11]. Labeling was observed in the propionyl ribose, the cytokinin base, and the isoprenoid side-chain. Since the first step in cytokinin biosynthesis is the prenylation of ADP and/or ATP leading to the formation of isopentenyladenosine-5'-diphosphate and/or isopentenyladenosine-5'-triphosphate [4,5], we selected the quasi-molecular ion of pro-iPR at *m/z* 504 as the precursor ion and the free base fragment at *m/z* 204 as the product ion for multiple reaction monitoring of isotopomer I0 (the isotopomer with lowest *m/z*), to exclude the signal of the ribose moiety. The isopentenyl moiety can be trans-hydroxylated by cytochrome P450 monooxygenases to form trans-zeatin-type nucleotides [6]. The rate of biosynthesis of the zeatin-type cytokinins was monitored using the analogous *m/z* 576-276 diagnostic transition for the I0 isotopomer. This transition corresponds to the dissociation of pro-zeatin riboside to its aglycone. An overview of the diagnostic transitions examined in this work is given in Table 3.

Table 3. Diagnostic transitions used to measure the rate of synthesis of selected cytokinins.

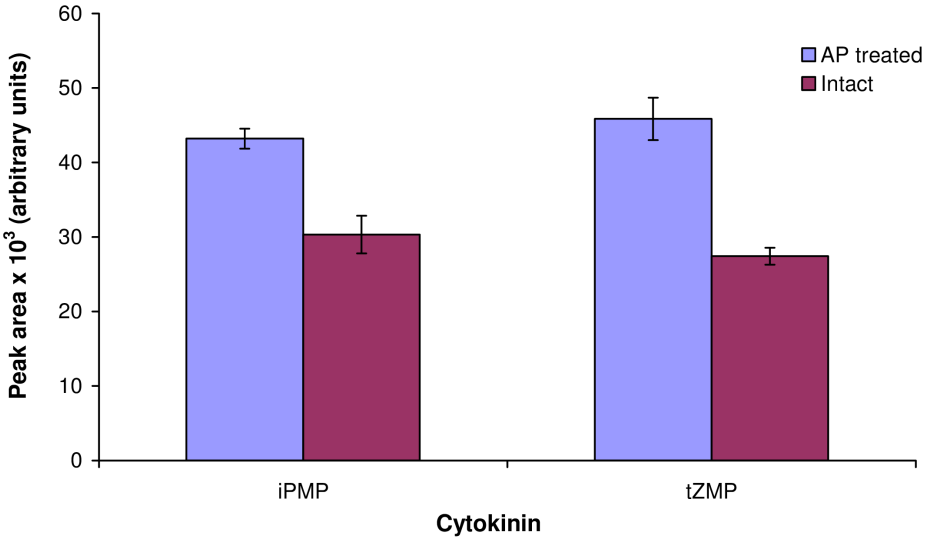
Derivatized cytokinins			**Underivatized cytokinins**		
Cytokinin	**Isotopomer**	**Diagnostic transition**	**Cytokinin**	**Isotopomer**	**Diagnostic transition**
pro-iPR	I_0	504-204	iPR	I_0	336-204
	I_1	505-205		I_1	337-205
	I_2	506-206		I_2	338-206
	I_3	507-207		I_3	339-207
pro-ZR	I_0	576-276	ZR	I_0	352-220
	I_1	577-277		I_1	353-221
	I_2	578-278		I_2	354-222
	I_3	579-279		I_3	355-223

2.3. Enzymatic treatment of cytokinin nucleotides improves the sensitivity of mass spectrometric detection

Generally, cytokinin nucleotides can be analyzed by LC-MS either as intact monophosphates [16,17] or as the corresponding ribosides, prepared by treatment with alkaline phosphatase (AP) [18]. The ribosides can be further purified by immunoaffinity chromatography using a broad-spectrum monoclonal anti-cytokinin antibody that recognizes free bases, 9-ribosides and 9-glucosides but not the nucleotides (L. Spichal, unpublished data). To determine which approach gives the best sensitivity, we prepared two sets of cytokinin standards (5 nmol of tZMP and iPMP). The first set of samples (five replicates) was treated with alkaline phosphatase, the reaction products were purified by immunoaffinity chromatography (IAC) and the eluate was evaporated to the dryness. The second set of samples was directly evaporated to dryness and both sets were derivatized by propionylation [13]. All samples were re-dissolved in the initial mobile phase and 0.5 pmol of each analyte was injected onto the chromatographic column. The samples were analyzed by the UPLC-MS/MS method described above and the signals corresponding to the first isotopomer (I0) in each case were compared (Figure 3). The signals of iPMP and tZMP treated with alkaline phosphatase were about 42% and 67% higher than those of the non-treated nucleotides, respectively, which is in good agreement with our previously published data on the quantitative analysis of cytokinins [16]. Generally, the ESI response is higher for more hydrophobic molecules because they have a greater affinity for the droplet surface [19]. Because

cytokinin ribosides are more hydrophobic than cytokinin monophosphates, their detection limit is almost 5-fold lower [16]. Moreover, propionylation of cytokinin ribosides gave slightly higher yields (data not shown). On the other hand, the recovery of IAC ranges between 30 and 70% [17]. Altogether, the data indicate that the use of enzymatic treatment and purification by IAC significantly improves the ESI-MS signal. Additionally, when AP treatment is omitted, only the monophosphates are detected and analyzed. However, experiments using recombinant IPTs suggest that the primary products of CK biosynthesis are isopentenyladenosine 5'-diphosphate and 5'-triphosphate [4,5]. AP will cleave all of the various forms of the nucleotides (*i.e.* mono-, di-, and tri-phosphates), making it possible to analyze all of them at once; while this approach is obviously less informative than an analysis of the individual nucleotides, it is very convenient and efficient. We recently published a HPLC-MS method for the determination of underivatized intact cytokinin nucleotides in human leukemia cells [20]. However, the sensitivity of this method is restricted by the poor ionization efficiency of cytokinin di- and tri-phosphates. Moreover, an efficient procedure for the purification of cytokinin nucleotides isolated from plant tissue remains to be developed; current purification methods offer recoveries ranging from 34% to 65% [16,17]. It is evident that despite ongoing progress in method development, an optimal technique for the analysis of individual cytokinin nucleotides isolated from plant tissue is not yet available.

Figure 3. Cytokinin nucleotides analyzed by UPLC-MS/MS as intact monophosphates and after AP treatment. 0.5 pmol of each cytokinin was injected onto the chromatographic column and detected by MS using multiple reaction monitoring mode. The intensities of the mass spectrometric signals corresponding to the first isotopomer (I0) are shown.

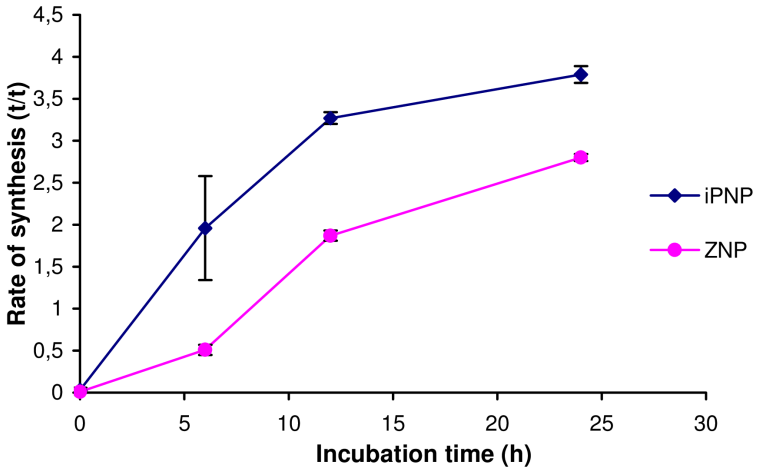


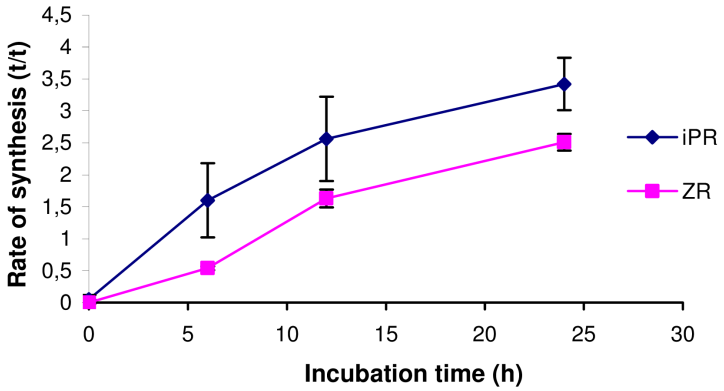
2.4. Application of the new method

The newly-developed UPLC-MS/MS method was employed for the analysis of the relative rate of biosynthesis of cytokinins in the cytokinin-overproducing Arabidopsis line *pga22*, which carries the

AtIPT8 gene under the control of the 17-β-estradiol-inducible promoter/enhancer [20]. It has previously been reported that over the course of a 24-hour induction period, iPMP and iPR levels in this mutant increased more than 19- and 38-fold, respectively, but only minor increases in the levels of zeatin-type monophosphates and ribosides were observed [20]. Three-week-old Arabidopsis seedlings were incubated in liquid growth medium enriched with 30% deuterium oxide with 0 (control) or 5 μM 17-β-estradiol for 24 hours, after which the nucleotides were extracted, subjected to AP treatment and purification by IAC, and analyzed. Cytokinins including iPNP, tZNP, iPR and tZR, showed significant deuterium enrichments, *i.e.* increased ratios of labeled to unlabeled cytokinin after correcting for the natural isotope distribution (tracer/tracee ratio; Figure 4).

Figure 4. The biosynthetic rate (tracer:tracee ratio) of iP-type nucleotides (IPNP), tZ-type nucleotides (tZNP), iPR and tZR isolated from Arabidopsis line *pga22* plotted against the incubation time. Data were acquired in MRM mode and include corrections to compensate for the natural isotope distribution.





As expected, induction of AtIPT8 expression triggered an increase in the biosynthetic rate. The iP-type nucleotides (iPNP) were more extensively labeled than were the tZ-type nucleotides (tZNP) throughout the experiment. This is consistent with iPNPs being the primary products of IPT action and tZNP being synthesized from them by trans-hydroxylation. In addition, both ribosides were less extensively labeled than their phosphorylated counterparts, confirming that *in vivo* dephosphorylation occurs at a relatively late stage in cytokinin biosynthesis. Interestingly, the absolute t/t-ratios were similar to those obtained in a bacterial IPT overexpression system [12]. The tracer:tracee ratio of the *cis*-Z-type cytokinins remained at the basal level throughout the experiment (data not shown).

2.5. Comparison to existing methods

Finally, we compared the newly-developed UPLC-MS/MS method to that developed by Dobrev and co-workers, [21] which was designed for the analysis of underivatized cytokinins. The underivatized cytokinins were resolved on a BEH C18 chromatographic column (50 × 2.1 mm, particle size 1.7 μm; Waters) using 15 mM ammonium formate (pH 4) and methanol as mobile phase buffers [17]. The diagnostic transitions utilized for mass spectrometric detection of derivatized and underivatized cytokinins are summarized in Table 3. A quasi-molecular ion of the protonated riboside was selected as the precursor ion, and the most intense fragment (a free base) was selected as the product ion for multiple reaction monitoring of isotopomer I0. This diagnostic transition allowed us to exclude signals arising from the ribose moiety. The response of underivatized iPR was about 34% lower than that of its propionylated counterpart. The response of the underivatized zeatins was about 75% lower than those of propionylated zeatins (Figure 5). The difference between the values obtained for both cytokinin types reflects the fact that the different analyte molecules incorporate different numbers of propionyl groups during derivatization: whereas iPR carries propionyl groups on the ribose moiety only, zeatins carried an additional one on the hydroxylated side-chain. Derivatization increased the ESI response of the analytes by augmenting their hydrophobicity and also increased their molecular mass. This increase in mass shifted the signals arising from the derivatized molecules into an *m/z* region with relatively little chemical noise, resulting in a slightly higher MS-response. The sensitivity of MS detection is very important when designing specific biological experiments. Previously, at least 1 g of fresh plant tissue was required for measurements of the cytokinin biosynthetic rate [22], which is problematic when valuable information on tissue and organ specific cytokinin biosynthesis is desired. For the sake of comparison, the concentrations of the auxins (another class of phytohormones) in plant tissues are more than ten times higher than those of cytokinins and consequently, only a few miligrams of fresh plant material is necessary for accurate measurements of the biosynthetic rate of auxins, making experimental resolution of auxin concentrations at the cellular level possible [23]. Although, the proposed UPLC-MS/MS method for cytokinins cannot be used for analysis at the cellular level, the organ level would be attainable if one were to use other, larger plants than Arabidopsis. Derivatization thus increases the sensitivity of mass spectrometric detection to the point that much less plant tissue is required for analysis.

Both approaches, with or without derivatization, can be used to measure the relative biosynthetic rate of cytokinins. In both cases, analysis of peaks corresponding to the loss of the ribose moiety should yield similar results. To confirm this assumption we measured the relative biosynthetic rates of

iPNP and iPR in the *pga22* line by both methods. Cytokinins were isolated by harvesting Arabidopsis seedlings 6 hours after chemical induction of AtIPT8, fractionating them to separate the ribosides from the nucleotides, treating the latter with AP, and finally purifying them by IAC. The results are summarized in Table 4.

Figure 5. Comparison of ESI-MS response. 10 pmol of derivatized and underivatized cytokinin standards were analyzed using UPLC-MS/MS. Peak areas (arbitrary units) of the first isotopomer are compared. Error bars indicate the standard deviation (n = 4).

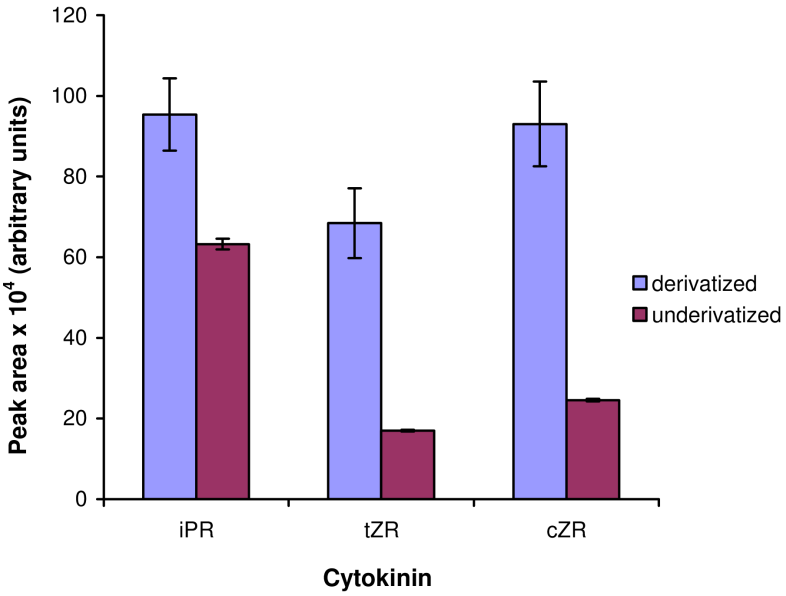


Table 4. The relative biosynthetic rate (t/t) of iP-type nucleotides (iPNP) and isopentenyladenosine (iPR) isolated from Arabisopsis seedlings 6h after chemical induction of *AtIPT8*.

Cytokinin	Derivatized		Underivatized	
	Rate of synthesis (t/t)	SD	Rate of synthesis (t/t)	SD
iPNP	1.6843	0.2099	1.7961	0.1750
iPR	1.2965	0.1334	1.4538	0.1525

As expected, no significant differences were observed in the relative biosynthetic rates of iP-type cytokinins. Both UPLC-MS/MS methods are rapid, achieving complete separation in less than 6 min. Sample preparation remains the most time consuming part of the entire method; the derivatization in particular requires at least 10 hours (encompassing derivatization, incubation, and concentration *in vacuo*). Nonetheless, this step is worthwhile because it greatly increases the sensitivity of the mass spectrometric detection. Since all of the relevant cytokinin metabolites are well resolved by UPLC, it is possible to analyze the whole set at once if necessary. The method desribed in this paper could potentially be used to analyze deuterium enrichment in all of the various kinds of cytokinins, including

nucleotides, ribosides, free bases and glucosides, provided that one could obtain enough material. In this work, we decided to study the rather small plant *Arabidopsis thaliana,* which is a very popular model organism in plant sciences. However, due to the low mass of the plant seedlings and the relatively low levels of physiologically active CK bases, we chose to focus exclusively on nucleotides and ribosides in this instance.

We have also created a web-based calculator (IZOTOP), which helps the user to convert signals obtained by MS/MS measurements into relative biosynthetic rate (t/t ratio) values. The procedure used in performing these calculations was adapted from the work of Åstot *et al.* [11]. This calculator is freely available (www.biochemie.upol.cz/software/izotop).

3. Experimental

3.1. Chemicals

Authentic cytokinin standards were purchased from Olchemim (Olomouc, Czech Republic), formic acid from Merck (Darmstadt, Germany), and DEAE-Sephadex A-25 from GE Healthcare (Uppsala, Sweden). Deionised (Milli-Q) water was obtained from a Simplicity 185 water system (Millipore, Bedford, MA, USA). All other chemicals were of analytical grade or higher purity and were purchased from Sigma-Aldrich Chemie (Steinheim, Germany).

3.2. Biological material

Arabidopsis thaliana line *pga22* carrying the *AtIPT8* gene under the control of the 17-β-estradiol-inducible promoter/enhancer [18] was grown in 250-mL Erlenmeyer flasks containing 50 mL of full Murashige and Skoog (MS) basal growth medium, 3% sucrose, pH 5.6 (25 seeds per flask). The flasks were agitated and maintained at 23 °C under long day conditions with 18 h light and 6 h darkness. After three weeks, the plants were transferred to half-strength MS medium containing 1.5% sucrose, 30% deuterium oxide and 0 or 5 μM 17- β -estradiol to induce cytokinin biosynthesis.

3.3. Sample preparation

Samples were extracted and purified essentially as outlined by Novák *et al.* [18]. Briefly, frozen plant material (500 mg fresh weight) was homogenized using an MM 301 vibration mill (Retsch, Haan, Germany) at a frequency of 30 Hz for 2 min and extracted overnight in methanol–chloroform–formic acid–water (12:5:1:2, v/v/v/v). The extract was first passed through a cation (SCX-cartridge) and then an anion [DEAE-Sephadex combined with an SPE(C18)-cartridge] exchanger to yield fraction 1, which contained the cytokinin free bases, ribosides, and glucosides, and fraction 2, which contained the riboside-5'-phosphates. Fraction 2 was first treated with alkaline phophatase (4U/sample, 37 °C, 60 min) and both fractions were further purified by immunoaffinity chromatography based on generic monoclonal anticytokinin-antibodies. For propionylation, the samples were dissolved in 10 μL of dimethylformamide. 6 μL of N-methylimidazole and 2 μL of propionic anhydride were added and the reaction mixtures were heated at 37 °C for 60 min, after which they were evaporated under reduced pressure.

3.4. HPLC-Q-TOF MS identification

A hybrid Q-TOF micro™ mass spectrometer (Waters MS Technologies) was used for the high-resolution identification and confirmation of the side-chain fragments at *m/z* 69 and *m/z* 75. Electrospray ionization in the positive ion mode was used with the following parameters: source block/desolvation temperature, 100 °C/350 °C; capillary/cone voltage, 2500/25 V; and spray/cone gas flow (N_2), 500/50 L/h. In the product ion scan of the selected precursors (*m/z* 204 and *m/z* 210), data were acquired in the mass range *m/z* 50–250, with a cycle time of 33 ms, a scan time of 2.0 s, and collision energy of 40 eV, 50 eV and 60 eV. For the exact mass determination experiments, a lock spray was used for external calibration with a mixture of 0.1 M NaOH/10% formic acid (v/v) and acetonitrile (1:1:8, v/v/v) as a reference. Accurate masses were calculated and used for the determination of the elementary composition of the analytes with fidelity of 5 ppm.

3.5. UPLC-MS/MS

An Acquity UPLC™ System (Waters, Milford, MA, USA) coupled to a Xevo triple-stage quadrupole mass spectrometer (Waters MS Technologies, Manchester, UK) was utilized for the cytokinin measurements. The system was controlled by Masslynx software (version 4.0, Waters, Manchester, UK).

Derivatized cytokinins were separated on a chromatographic column (BEH C18, 50 × 2.1 mm, particle size 1.7 μm; Waters) using 0.7% HCOOH in water and 0.7% HCOOH in acetonitrile as solvents A and B, respectively. At a flow-rate of 0.5 mL min^{-1}, the following elution profile was used: an isocratic step at 15% B for 0.5 min preceded a linear gradient from 15 to 45% B for 5.5 min. At the end of the gradient there was a washing step at 100% B (for 2 min) and then the column was equilibrated to the initial conditions for 1.9 min. The column was thermostated at 40 °C. Tandem mass spectra of all cytokinins examined in this study were acquired by continuous infusion of 10^{-3} mol L^{-1} solution in solvent A/solvent B (1:1, v/v) at a flow-rate of 15 μL min^{-1}.The capillary voltage, cone voltage, collision cell energy, and ion source temperature were optimized for each individual compound. The mass spectrometer settings were as follows: a capillary voltage of 3 kV, a cone voltage of 33 V, a source temperature of 120 °C, a desolvation temperature of 550 °C, a cone gas flow of 70 L h^{-1}, a desolvation gas flow of 600 L h^{-1}, collision cell energy of 19 eV, and a collision gas flow of 0.2 mL min^{-1}.

Underivatized cytokinins were separated on chromatographic column (BEH C18, 50 × 2.1 mm, particle size 1.7 μm; Waters) using 15 mM HCOOH (pH 4.0, adjusted by NH_4OH) and methanol as solvent A and B, respectively. At a flow-rate of 0.5 mL min^{-1}, the following elution profile was used: an isocratic step at 10% B for 0.5 min preceded a linear gradient from 12 to 45% B for 3.5 min. At the end of the gradient there was a washing step at 100% B (for 0.3 min) and then the column was equilibrated to the initial conditions for 1.2 min. The column was thermostated at 40 °C. The capillary voltage, cone voltage, collision cell energy, and ion source temperature were optimized for each individual compound. The mass spectrometer settings were as follows: a capillary voltage of 3 kV, a cone voltage of 30 V, a source temperature of 120 °C, a desolvation temperature of 575 °C, a cone gas

flow of 70 L h^{-1}, a desolvation gas flow of 600 L h^{-1}, collision cell energy of 19 eV, and a collision gas flow of 0.2 mL min^{-1}.

4. Conclusions

We have developed a UPLC-MS/MS method for determining the relative rates of biosynthesis of cytokinin-type phytohormones. The method is based on a rapid and efficient separation of derivatized analytes by ultra-performance liquid chromatography combined with sensitive and selective mass spectrometric detection. We show that the biosynthetic rate of isopentenyladenine can be measured using a previously neglected diagnostic 204-69 transition. Cytokinin nucleotides are analyzed after treatment with alkaline phosphatase, which enhances the sensitivity of the detection process. The new method proved to be more sensitive than one based on the analysis of underivatized cytokinins, although the relative synthetic rate values measured by the two both methods were similar. Finally, we created a web-based calculator (IZOTOP) that facilitates MS/MS data processing and have made it freely available to the research community.

Acknowledgements

We thank Nam-Hai Chua for providing the PGA22/AtIPT8 Arabidopsis seeds. We thank Danny Vereecke for critical reading and linguistic correction of the text. The work presented was supported by grants MSM 6198959216 from the Ministry of Education, Youth and Sports of the Czech Republic, 522/08/0920 from Czech Science Foundation and ED0007/01/01 Centre of the Region Haná for Biotechnological and Agricultural Research.

References and Notes

1. Mok, M.C. Cytokinins and plant development-an overview. In *Cytokinins: Chemistry and Function*; Mok, D.W.S., Mok, M.C., Eds.; CRC Press: Boca Raton, FL, USA, 1994.
2. Sakakibara, H. Cytokinins: activity, biosynthesis, and translocation. *Annu. Rev. Plant Biol.* **2006**, *57*, 431-449.
3. Spíchal, L.; Rakova, N.Y.; Reifler, M.; Mizuno, T.; Romanov, G.A.; Strnad, M.; Schmulling, T. Two cytokinin receptors of *Arabidopsis thaliana*, CRE1/AHK4 and AHK3, differ in their ligand specificity in a bacterial assay. *Plant Cell Physiol.* **2004**, *45*, 1299-1305.
4. Kakimoto, T. Identification of plant cytokinin biosynthetic enzymes as dimethylallyl diphosphate: ATP/ADP isopentenyltransferases. *Plant Cell Physiol.* **2001**, *42*, 677-685.
5. Takei, K.; Sakakibara, H.; Sugiyama, T. Identification of genes encoding adenylate isopentenyltransferase, a cytokinin biosynthesis enzyme, in *Arabidopsis thaliana*. *J. Biol. Chem.* **2001**, *276*, 26405-26410.
6. Takei, K.; Yamaya, T.; Sakakibara, H. *Arabidopsis CYP735A1* and *CYP735A2* encode cytokinin hydroxylases that catalyze the biosynthesis of *trans*-zeatin. *J. Biol. Chem.* **2004**, *279*, 41866-41872.
7. Miyawaki, K.; Tarkowski, P.; Matsumoto-Kitano, M.; Kato, T.; Tarkowska, D.; Tabata, S.; Sandberg, G.; Kakimoto, T. *In planta* roles of Arabidopsis ATP/ADP isopentenyltransferases and

tRNA isopentenyltransferases in cytokinin biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 16598-16603.

8. Tarkowski, P.; Ge, L.Y.; Young, J.W.H.; Tan, S.N. Analytical methods for cytokinins. *TrAC. Trends Anal. Chem.* **2009**, *28*, 323-335.
9. Hocart, C.H.; Letham D.S. Biosynthesis of cytokinin in germinating seeds of *Zea mays*. *J. Exp. Bot.* **1990**, *41*, 1525-1528.
10. Sakakibara, H.; Kasahara, H.; Ueda, N.; Kojima, M.; Takei, K.; Hishiyama, S.; Asami, T.; Okada, K.; Kamiya, Y.; Yamaya, T.; Yamaguchi, S. *Agrobacterium tumefaciens* increases cytokinin production in plastids by modifying the biosynthetic pathway in the host plant. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 9972-9977.
11. Åstot, C.; Dolezal, K; Moritz, T.; Sandberg, G. Deuterium *in vivo* labelling of cytokinins in *Arabisopsis thaliana* analysed by capillary liquid chromatography/frit-fast atom bombardment mass spectrometry. *J. Mass Spectrometry* **2000**, *35*, 13-22.
12. Åstot, C.; Dolezal, K; Nordström, A.; Wang, Q.; Kunkel, T.; Moritz, T.; Chua, N.H.; Sandberg, G. An alternative cytokinin biosynthesis pathway. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 14788-14783.
13. Åstot, C.; Dolezal, K; Moritz, T.; Sandberg, G. Precolumn derivatization and capillary liquid chromatographic/frit-fast atom bombardment mass spectrometric analysis of cytokinins in *Arabidopsis thaliana*. *J. Mass Spectrometry* **1998**, *33*, 892-902.
14. Nordström, A.; Tarkowski, P.; Tarkowska, D.; Norbaek, R.; Åstot, C.; Dolezal, K.; Sandberg, G. Auxin regulation of cytokinin biosynthesis in *Arabidopsis thaliana*: factor of potential importance for auxin-cytokinin-regulated development. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 8039-8044.
15. Nováková, L.; Vlčková, H. A review of current trends and advantages in modern bio-analytical methods: Chromatography and sample preparation. *Anal. Chim. Acta* **2009**, *656*, 8-35.
16. Nordström, A.; Tarkowski, P.; Tarkowská, D.; Dolezal, K.; Åstot, C.; Sandberg, G.; Moritz, T. Derivatization for LC-electrospray ionization-MS: a tool for improving reversed-phase separation and ESI response of bases, ribosides and intact nucleotides. *Anal. Chem.* **2004**, *76*, 2869-2877.
17. Novák, O.; Hauserová, E.; Amakorová, P.; Doležal, K.; Strnad, M. Cytokinin profiling in plant tissues using ultra-performance liquid chromatography-electrospray tandem mass spectrometry. *Phytochemistry* **2008**, *69*, 2214-2224.
18. Novák, O.; Tarkowski, P.; Tarkowská, D.; Doležal, K.; Lenobel, R.; Strnad, M. Quantitative analysis of cytokinins in plants by liquid chromatography-single-quadrupole mass spectrometry. *Anal. Chim. Acta* **2003**, *480*, 207-218.
19. Cech, N.B.; Enke, C.G. Effect of affinity for droplet surface on the fraction of analyte molecules charged during electrospray droplet fission. *Anal. Chem.* **2001**, *73*, 4632-4639.
20. Béres, T.; Zatloukal, M.; Voller, J.; Niemann, P.; Gahsche, M.C.; Tarkowski, P.; Novák, O.; Hanuš, J.; Strnad, M.; Doležal, K. Synthesis and LC-MS identification and quantification of cytokinin nucleotides in K-562 human leukemia cells. *Anal. Bioanal. Chem.* **2010**, *398*, 2071-2080.
21. Sun, J.Q.; Niu, Q.W; Tarkowski, P. Zheng, B.L.; Tarkowska, D.; Sandberg, G.; Chua, N.H.; Zuo, J. The Arabidopsis *AtIPT8/PGA22* gene encodes an isopentenyl transferase that is involved in de novo cytokinin biosynthesis. *Plant Physiol.* **2003**, *131*, 167-176.

22. Dobrev, P.I.; Novák, O., Doležal, K.; Trčková, M.; Kamínek, M. Determination of phytohormone biosynthesis in wheat grains by LC/MS. In *Proceedings of 34th Symposium on High-Performance Liquid Phase Separations and Related Techniques*, Dresden, Germany, June 28–July 2, 2009; Dizdarevic, A., Huber, C., Eds.; GDC: Dresden, Germany, 2009; p. 723.
23. Stirk, W.A.; Novák, O.; Václavíková, K.; Tarkowski, P.; Strnad, M.; van Staden, J. Spatial and temporal changes in endogenous cytokinin in developing pea roots. *Planta* **2008**, *227*, 1279-1289.
24. Petersson, S.V.; Johansson, A.I.; Kowalczyk, M.; Makoveychuk, A.; Wang, J.Y.; Moritz, T.; Grebe, M.; Benfey P.N.; Sandberg, G.; Ljung, K. An auxin gradient and maximum in the Arabidopsis root apex shown by high-resolution cell-specific analysis of IAA distribution and synthesis. *Plant Cell* **2009**, *21*, 1659-1668.

Sample Availability: Samples of the compounds of interest are available from the authors.

© 2010 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).

Příloha D – srovnání spekter fermentace (Biospean)

Příloha obsahuje výsledky porovnání hmotnostních spekter jednotlivých stavů fermentace (produkce penicilinu plísní *Penicillium chrysogenum*, viz. 6.12.1 Detekce stavu fermentace, str. 95) vypočtené pomocí aplikace Biospean.

Každá z níže uvedených tabulek představuje jeden sledovaný fermentační experiment. Z těchto experimentů byly v průběhu času odebírány menší vzorky buněk plísně, které byly analyzovány hmotnostním spektrometrem. Tím vznikla řada hmotnostních spekter, které zachycují jednotlivé stavy procesu.

Z těchto posloupností spekter byly vytvořeny jednotlivé níže uvedené tabulky. Každá z tabulek tedy obsahuje spektra jedné fermentace, ale v různých časových okamžicích.

Jednotlivé řádky a sloupečky představují jednotlivá spektra v pořadí, v jakém byla v průběhu času postupně změřena. Průsečík řádku a sloupečku obsahuje podobnost spektra v řádku ku spektru ve sloupečku. Podobnost mezi spektry A a B je procentuální hodnota, která udává, kolik píků ve spektru A má odpovídající pík ve spektru B. Podobnost A ku B je tedy jiná než B ku A.

Na hlavní diagonále tabulek, kde se porovnávají hmotnostní spektra sama se sebou, jsou hodnoty 100 % (shoda je úplná). S ohledem na to, že porovnávaná spektra reprezentují stavy probíhající fermentace, je logické předpokládat, že čím bude větší časový rozdíl mezi odběry jednotlivých spekter, tím se budou výsledky podobnosti lišit.

Tento předpoklad se skutečně potvrzuje – podobnost spekter mezi sebou se obecně směrem od hlavní diagonály snižuje. Pokud si vybereme hodnotu na hlavní diagonále a podíváme se na změny hodnot v dalších buňkách do všech stran, zjistíme, že hodnoty mají obecně trend se postupně snižovat. Nejmenší hodnoty podobnosti nalezneme v levém dolním a pravém horním rohu tabulky – tam se porovnávají první spektra ze začátku reakce s posledními spektry z konce fermentace, proto jsou podobnosti nejnižší.

V některých buňkách tabulek se vyskytují odchylky, která klesající řadu hodnot porušují. To je způsobeno jednak nepříliš vysokou kvalitou spekter (vysoký šum, kvalita měření) a jednak tím, že pro rozpoznání píků byla použita pouze automatická detekce. V případě dodatečné poloautomatické a případně i manuální korekce (viz. 6.11.1 Nastavení vlastností spektra, str. 87) by výsledky byly lepší.

Hodnoty porovnání jsem ponechal úmyslně pouze v automatickém režimu, protože dávají určitou představu o efektivitě i rezervách automatické detekce – algoritmus počítá podobnosti správně, nicméně kvalita spekter způsobuje odchylky.

%	04	05	06	07	08	09	10	11	12	13
BCB1_1-sample_04	100,00	20,95	8,75	8,54	8,55	7,95	8,29	8,71	5,98	2,73
BCB1_1-sample_05	61,68	100,00	31,00	29,18	29,81	26,66	28,00	30,21	19,33	10,10
BCB1_1-sample_06	54,21	65,24	100,00	81,71	84,39	80,59	76,48	77,20	55,89	24,55
BCB1_1-sample_07	45,79	53,17	70,74	100,00	78,40	77,19	67,43	70,81	52,64	22,83
BCB1_1-sample_08	45,33	53,65	72,17	77,44	100,00	86,45	79,44	82,58	58,96	22,32
BCB1_1-sample_09	42,52	48,41	69,53	76,92	87,21	100,00	78,26	82,95	62,65	25,45
BCB1_1-sample_10	45,79	52,54	68,17	69,43	82,80	80,86	100,00	83,04	56,94	19,70
BCB1_1-sample_11	43,93	51,75	62,82	66,55	78,57	78,23	75,80	100,00	61,25	26,16
BCB1_1-sample_12	31,78	34,92	47,96	52,18	59,17	62,33	54,82	64,60	100,00	37,17
BCB1_1-sample_13	12,62	15,87	18,33	19,69	19,49	22,03	16,50	24,00	32,34	100,00

spektra skupiny BCB1_1

%	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17
BCB1_6-sample_01	100,00	46,15	22,16	21,92	14,70	13,69	6,17	3,01	2,11	7,21	2,29	1,21	0,88	1,18	0,99	0,41	3,57
BCB1_6-sample_02	63,83	100,00	56,29	37,44	26,52	26,62	15,74	9,03	7,08	8,69	6,56	5,17	4,23	2,13	1,98	1,64	1,43
BCB1_6-sample_03	39,36	72,31	100,00	52,51	39,07	40,30	29,01	17,17	14,46	12,38	13,13	11,21	9,17	9,00	6,91	4,93	2,14
BCB1_6-sample_04	51,06	63,08	68,86	100,00	70,25	71,86	53,09	26,90	25,30	23,84	22,75	20,17	15,52	10,66	8,40	5,54	5,00
BCB1_6-sample_05	43,62	56,92	65,27	89,50	100,00	93,92	71,60	38,76	35,24	30,50	31,30	27,24	21,69	17,06	11,36	7,60	6,43
BCB1_6-sample_06	38,30	53,85	63,47	86,30	88,53	100,00	69,44	37,88	34,79	30,68	31,45	27,59	20,99	17,06	10,62	7,19	7,14
BCB1_6-sample_07	21,28	39,23	56,29	78,54	83,15	85,55	100,00	44,60	44,13	39,74	38,93	35,52	28,57	25,36	14,07	13,35	6,43
BCB1_6-sample_08	18,09	39,23	58,08	69,41	78,49	81,37	77,78	100,00	62,35	49,91	64,27	61,21	51,50	44,08	24,44	25,67	9,29
BCB1_6-sample_09	14,89	36,15	57,49	76,71	83,87	87,83	90,43	73,27	100,00	57,67	68,70	56,72	47,27	43,36	24,69	25,05	13,57
BCB1_6-sample_10	41,49	36,15	40,12	58,90	59,14	63,12	66,36	47,79	46,99	100,00	48,24	52,07	39,68	35,55	20,49	19,92	8,57
BCB1_6-sample_11	15,96	33,08	51,50	68,04	73,48	78,33	78,70	74,51	67,77	58,41	100,00	81,38	69,14	60,43	32,84	33,47	15,71
BCB1_6-sample_12	7,45	23,08	38,92	53,42	56,63	60,84	63,58	62,83	49,55	55,82	72,06	100,00	79,54	63,98	38,77	40,86	10,71
BCB1_6-sample_13	5,32	18,46	31,14	40,18	44,09	45,25	50,00	51,68	40,36	41,59	59,85	77,76	100,00	79,86	51,11	55,03	15,71
BCB1_6-sample_14	5,32	6,92	22,75	20,55	25,81	27,38	33,02	32,92	27,56	27,73	38,93	46,55	59,44	100,00	56,05	50,10	22,86
BCB1_6-sample_15	4,26	6,15	16,77	15,53	16,49	16,35	17,59	17,52	15,06	15,34	20,31	27,07	36,51	53,79	100,00	61,19	20,71
BCB1_6-sample_16	2,13	6,15	14,37	12,33	13,26	13,31	20,06	22,12	18,37	17,93	24,89	34,31	47,27	57,82	73,58	100,00	14,29
BCB1_6-sample_17	5,32	1,54	1,80	3,20	3,23	3,80	2,78	2,30	2,86	2,22	3,36	2,59	3,88	7,58	7,16	4,11	100,00

spektra skupiny BCB1_6

%	05	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	29
FbL-L_05	100,00	34,48	14,33	8,08	10,40	9,89	6,63	2,75	7,54	9,81	8,88	8,31	14,01	10,25	12,03	4,35	6,49	6,06	6,00	10,34	12,87	10,27	7,58
FbL-L_07	34,29	100,00	34,21	32,69	25,99	29,28	18,37	10,07	12,38	8,86	9,17	8,55	12,56	8,07	9,96	4,35	5,77	5,36	4,55	8,81	10,89	9,38	7,20
FbL-L_08	28,00	67,24	100,00	67,69	47,40	48,29	31,12	16,70	18,38	14,24	15,19	12,24	17,87	9,94	12,45	5,03	6,25	6,76	5,59	10,73	13,37	11,61	7,20
FbL-L_09	12,00	48,85	51,46	100,00	57,19	55,51	34,95	18,08	16,25	10,13	12,89	10,39	14,49	7,45	9,13	5,03	4,81	5,13	4,35	7,66	7,92	7,59	6,06
FbL-L_10	19,43	48,85	45,32	71,92	100,00	76,05	52,81	32,95	29,59	16,46	16,05	13,86	19,81	13,04	14,11	7,32	8,41	8,39	7,87	12,26	12,87	13,84	11,36
FbL-L_11	14,86	44,25	37,13	56,15	61,16	100,00	53,83	32,27	26,50	16,77	16,05	13,86	18,36	14,29	14,52	8,70	9,13	9,79	8,07	12,64	13,37	13,39	12,12
FbL-L_12	14,86	41,38	35,67	52,69	63,30	80,23	100,00	49,43	40,81	23,73	20,63	19,40	22,22	20,50	20,33	11,44	13,70	12,35	11,59	18,01	18,32	19,64	15,53
FbL-L_13	6,86	25,29	21,35	30,38	44,04	53,61	55,10	100,00	45,84	18,35	16,33	16,40	19,32	19,25	20,33	10,76	12,02	12,82	12,42	18,39	17,82	17,86	14,77
FbL-L_14	22,29	36,78	27,78	32,31	46,79	52,09	53,83	54,23	100,00	48,73	40,11	42,96	39,13	41,93	36,10	25,40	23,32	28,21	25,05	35,63	35,64	36,16	24,24
FbL-L_15	17,71	16,09	13,16	12,31	15,90	20,15	19,13	13,27	29,79	100,00	63,32	59,12	74,40	66,77	65,56	43,71	42,79	45,69	40,17	60,15	67,33	65,18	44,32
FbL-L_16	17,71	18,39	15,50	17,31	17,13	21,29	18,37	13,04	27,08	69,94	100,00	56,58	68,60	62,11	68,05	46,91	43,51	51,52	42,65	63,60	66,83	68,30	51,14
FbL-L_17	20,57	21,26	15,50	17,31	18,35	22,81	21,43	16,25	35,98	81,01	70,20	100,00	85,51	84,47	85,06	58,81	58,89	62,94	58,18	75,48	84,16	81,70	60,98
FbL-L_18	16,57	14,94	10,82	11,54	12,54	14,45	11,73	9,15	15,67	48,73	40,69	40,88	100,00	51,86	61,83	33,41	37,74	38,23	33,54	60,15	68,32	63,84	47,73
FbL-L_19	18,86	14,94	9,36	9,23	12,84	17,49	16,84	14,19	26,11	68,04	57,31	62,82	80,68	100,00	84,65	54,23	55,05	60,37	53,42	82,38	85,64	79,91	55,30
FbL-L_20	16,57	13,79	8,77	8,46	10,40	13,31	12,50	11,21	16,83	50,00	46,99	47,34	71,98	63,35	100,00	42,79	48,56	49,18	43,69	76,63	85,64	82,59	54,92
FbL-L_21	10,86	10,92	6,43	8,46	9,79	14,45	12,76	10,76	21,47	60,44	58,74	59,35	70,53	73,60	77,59	100,00	79,57	75,99	73,50	85,82	80,20	83,48	76,89
FbL-L_22	15,43	13,79	7,60	7,69	10,70	14,45	14,54	11,44	18,76	56,33	51,86	56,58	75,85	71,12	83,82	75,74	100,00	72,26	71,01	85,44	83,66	85,71	82,58
FbL-L_23	14,86	13,22	8,48	8,46	11,01	15,97	13,52	12,59	23,40	62,03	63,32	62,36	79,23	80,43	87,55	74,60	74,52	100,00	71,01	90,04	88,61	88,84	77,65
FbL-L_24	16,57	12,64	7,89	8,08	11,62	14,83	14,29	13,73	23,40	61,39	59,03	64,90	78,26	80,12	87,55	81,24	82,45	79,95	100,00	90,04	91,09	89,29	87,50
FbL-L_25	15,43	13,22	8,19	7,69	9,79	12,55	11,99	10,98	17,99	49,68	47,56	45,50	75,85	66,77	82,99	51,26	53,61	54,78	48,65	100,00	88,12	86,16	60,61
FbL-L_26	14,86	12,64	7,89	6,15	7,95	10,27	9,44	8,24	13,93	43,04	38,68	39,26	66,67	53,73	71,78	37,07	40,63	41,72	38,10	68,20	100,00	74,55	53,03
FbL-L_27	13,14	12,07	7,60	6,54	9,48	11,41	11,22	9,15	15,67	46,20	43,84	42,26	69,08	55,59	76,76	42,79	46,15	46,39	41,41	73,95	82,67	100,00	59,85
FbL-L_29	11,43	10,92	5,56	6,15	9,17	12,17	10,46	8,92	12,38	37,03	38,68	37,18	60,87	45,34	60,17	46,45	52,40	47,79	47,83	61,30	69,31	70,54	100,00

spektra skupiny FbL-L

%	04	05	06	07	08	09	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Fb_M-04	100,00	56,23	43,11	11,58	6,65	4,54	3,71	5,17	9,24	9,16	10,12	11,71	9,15	9,79	7,54	9,63	8,56	7,63	9,14	7,49	9,41	7,75	8,10	8,89
Fb_M-05	49,30	100,00	49,56	8,38	5,19	4,23	4,50	6,97	8,43	7,69	8,50	10,92	9,71	9,65	8,45	9,09	9,09	7,77	8,34	6,78	7,89	6,55	8,10	9,71
Fb_M-06	41,18	53,99	100,00	10,65	7,19	4,84	4,39	6,37	9,91	7,81	9,18	10,60	10,12	11,06	8,19	9,22	10,29	8,17	11,13	7,13	9,03	6,82	7,96	12,31
Fb_M-07	31,37	25,88	30,21	100,00	50,73	31,96	22,27	17,61	23,56	17,19	18,35	17,56	13,45	14,04	12,74	13,24	14,57	13,22	16,42	13,56	14,63	13,10	13,02	16,28
Fb_M-08	20,45	18,21	23,17	57,60	100,00	62,80	52,19	27,36	36,01	28,62	31,98	27,53	24,69	28,51	22,76	29,14	24,73	22,07	24,77	21,76	24,17	20,72	22,72	23,12
Fb_M-09	12,61	13,42	14,08	32,78	56,74	100,00	77,39	30,15	38,69	32,81	35,09	29,11	28,99	29,22	25,62	30,88	27,41	24,66	24,50	23,19	23,66	22,33	23,30	21,20
Fb_M-10	9,24	12,78	11,44	20,48	42,26	69,35	100,00	36,32	40,29	36,31	39,41	31,17	33,70	32,20	29,91	34,63	28,61	26,98	26,75	25,80	26,08	25,00	26,77	23,12
Fb_M-12	14,57	22,36	18,77	18,30	25,05	30,54	41,06	100,00	45,38	36,99	40,35	30,38	33,29	31,21	29,00	33,69	26,07	24,52	27,42	25,21	26,46	21,93	26,77	27,09
Fb_M-13	19,33	20,13	21,70	18,20	24,50	29,13	33,86	33,73	100,00	62,22	69,23	69,46	57,28	56,17	48,37	50,00	46,26	47,00	46,09	40,07	42,37	41,04	40,96	43,09
Fb_M-14	22,69	21,73	20,23	15,72	23,04	29,23	36,11	32,54	73,63	100,00	81,65	77,22	75,03	67,80	63,20	61,63	56,28	58,45	55,36	53,98	53,56	51,87	53,69	51,85
Fb_M-15	21,01	20,13	19,94	14,06	21,58	26,21	32,85	29,75	68,67	68,44	100,00	81,49	75,03	72,91	61,64	66,18	61,90	58,58	58,15	50,89	52,54	50,94	52,53	50,75
Fb_M-16	20,73	22,04	19,65	11,48	15,85	18,55	22,16	19,10	58,77	55,20	69,50	100,00	71,01	70,64	60,86	61,90	59,36	60,22	58,54	49,70	52,04	52,81	50,65	53,08
Fb_M-17	18,49	22,36	21,41	10,03	16,21	21,07	27,33	23,88	55,29	61,20	73,01	81,01	100,00	82,55	76,98	75,94	67,78	72,62	63,44	60,05	61,58	60,70	63,97	55,40
Fb_M-18	19,33	21,73	22,87	10,24	18,31	20,77	25,53	21,89	53,01	54,07	69,37	78,80	80,72	100,00	72,04	79,14	64,71	68,80	69,54	60,64	63,36	61,90	63,97	61,01
Fb_M-19	16,25	20,77	18,48	10,13	15,94	19,86	25,87	22,19	49,80	54,98	63,97	74,05	82,11	78,58	100,00	79,81	69,65	78,20	69,40	69,56	69,97	70,59	75,98	60,88
Fb_M-20	20,17	21,73	20,23	10,24	19,85	23,29	29,13	25,07	50,07	52,15	66,80	73,26	78,78	83,97	77,63	100,00	72,46	76,84	75,23	68,25	70,48	68,58	71,78	63,89
Fb_M-21	17,93	21,73	22,58	11,27	16,85	20,67	24,07	19,40	46,32	47,62	62,48	70,25	70,32	68,65	67,75	72,46	100,00	75,07	66,75	57,31	62,60	58,96	59,48	54,17
Fb_M-22	15,69	18,21	17,60	10,03	14,75	18,25	22,27	17,91	46,18	48,53	58,03	69,94	73,93	71,63	74,64	75,40	73,66	100,00	74,30	73,25	73,54	76,07	74,10	62,24
Fb_M-23	19,33	20,13	24,63	12,82	17,03	18,65	22,72	20,60	46,59	47,29	59,24	69,94	66,44	74,47	68,14	75,94	67,38	76,43	100,00	68,49	75,19	74,33	72,36	78,52
Fb_M-24	17,65	18,21	17,60	11,79	16,67	19,66	24,41	21,09	45,11	51,36	57,76	66,14	70,04	72,34	76,07	76,74	64,44	83,92	76,29	100,00	81,93	88,90	90,16	69,22
Fb_M-25	20,73	19,81	20,82	11,89	17,30	18,75	23,06	20,70	44,58	47,62	55,74	64,72	67,13	70,64	71,52	74,06	65,78	78,75	78,28	76,58	100,00	78,74	83,65	70,31
Fb_M-26	16,25	15,65	14,96	10,13	14,12	16,83	21,03	16,32	41,10	43,89	51,42	62,50	62,97	65,67	68,66	68,58	58,96	77,52	73,64	79,07	74,94	100,00	83,94	66,89
Fb_M-27	15,69	17,89	16,13	9,31	14,30	16,23	20,81	18,41	37,88	41,97	48,99	55,38	61,30	62,70	68,27	66,31	54,95	69,75	66,23	74,08	73,54	77,54	100,00	61,42
Fb_M-28	18,21	22,68	26,39	12,31	15,39	15,63	19,01	19,70	42,17	42,87	50,07	61,39	56,17	63,26	57,87	62,43	52,94	61,99	76,03	60,17	65,39	65,37	64,98	100,00

spektra skupiny Fb_M

%	06	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	30	31
Fb_N-06	100,00	10,19	7,22	16,15	2,76	6,12	15,41	20,64	18,07	8,22	9,77	9,06	10,85	10,24	9,83	11,62	10,35	9,17	11,32	11,20	10,42	11,52	11,95
Fb_N-08	12,79	100,00	30,25	18,97	15,19	18,82	19,03	26,16	20,05	11,19	13,16	11,07	15,57	9,87	11,72	15,35	13,16	10,05	13,54	14,93	11,60	10,99	11,95
Fb_N-09	12,40	41,36	100,00	25,38	32,54	30,12	26,28	31,69	27,97	22,55	22,93	19,46	26,18	20,48	21,36	22,41	21,40	21,52	17,81	19,84	18,15	17,10	15,66
Fb_N-10	24,42	22,84	22,35	100,00	15,98	24,00	38,67	29,36	18,56	8,57	6,95	6,88	5,42	7,86	5,86	6,85	8,07	7,23	7,61	6,68	6,55	7,85	7,58
Fb_N-11	5,43	23,77	37,25	20,77	100,00	63,06	31,42	29,65	26,73	19,23	22,56	21,31	25,24	19,56	21,17	22,61	18,07	15,87	16,70	19,65	18,82	16,58	14,48
Fb_N-12	10,08	24,69	28,89	26,15	52,86	100,00	47,73	37,50	31,68	18,18	17,86	19,97	21,70	17,73	17,77	20,75	15,09	14,29	16,14	18,47	16,81	16,06	15,49
Fb_N-13	19,77	19,44	19,64	32,82	20,51	37,18	100,00	40,70	30,69	14,86	13,16	16,11	17,69	13,35	12,67	15,98	13,16	12,87	12,62	13,36	12,61	12,74	11,78
Fb_N-14	27,52	27,78	24,60	25,90	20,12	30,35	42,30	100,00	65,84	32,52	35,71	35,74	45,75	36,20	35,73	40,66	34,74	33,69	36,92	38,11	31,09	33,33	30,81
Fb_N-15	28,29	25,00	25,51	19,23	21,30	30,12	37,46	77,33	100,00	46,33	53,38	50,17	61,08	52,29	50,28	54,56	47,54	45,33	48,79	49,71	46,22	48,52	42,59
Fb_N-16	18,22	19,75	29,12	12,56	21,70	24,47	25,68	54,07	65,59	100,00	71,62	58,72	66,27	64,17	57,66	55,60	60,88	62,61	52,13	51,28	49,41	50,61	45,96
Fb_N-17	20,16	21,60	27,54	9,49	23,67	22,35	21,15	55,23	70,30	66,61	100,00	69,63	75,24	70,75	66,73	64,94	65,09	62,26	55,66	56,19	54,79	53,75	48,65
Fb_N-18	20,93	20,37	26,19	10,51	25,05	28,00	29,00	61,92	74,01	61,19	78,01	100,00	83,96	73,67	78,83	77,39	69,47	65,08	60,30	66,80	63,19	59,86	54,71
Fb_N-19	17,83	20,37	25,06	5,90	21,10	21,65	22,66	56,40	64,11	49,13	59,96	59,73	100,00	60,33	63,89	68,05	56,67	55,38	51,76	58,15	53,61	48,69	46,30
Fb_N-20	21,71	16,67	25,28	11,03	21,10	22,82	22,05	57,56	70,79	61,36	72,74	67,62	77,83	100,00	76,37	71,16	69,12	73,19	63,64	64,83	64,54	64,92	56,73
Fb_N-21	20,16	19,14	25,51	7,95	22,09	22,12	20,24	54,94	65,84	53,32	66,35	69,97	79,72	73,86	100,00	80,08	70,35	69,66	68,09	74,85	70,42	65,10	59,76
Fb_N-22	21,71	22,84	24,38	8,46	21,50	23,53	23,26	56,98	65,10	46,85	58,83	62,58	77,36	62,71	72,97	100,00	62,81	60,67	68,46	73,87	65,04	62,65	62,63
Fb_N-23	22,87	23,15	27,54	11,79	20,32	20,24	22,66	57,56	67,08	60,66	69,74	66,44	76,18	72,03	75,80	74,27	100,00	79,01	68,46	72,10	65,38	66,49	59,93
Fb_N-24	20,16	17,59	27,54	10,51	17,75	19,06	22,05	55,52	63,61	62,06	66,35	61,91	74,06	75,87	74,67	71,37	78,60	100,00	70,50	68,17	66,72	65,62	57,07
Fb_N-25	23,64	22,53	21,67	10,51	17,75	20,47	20,54	57,85	65,10	49,13	56,39	54,53	65,80	62,71	69,38	76,56	64,74	67,02	100,00	80,16	72,94	76,96	67,85
Fb_N-26	22,09	23,46	22,80	8,72	19,72	22,12	20,54	56,40	62,62	45,63	53,76	57,05	69,81	60,33	72,02	78,01	64,39	61,20	75,70	100,00	72,10	70,33	67,85
Fb_N-27	24,03	21,30	24,38	10,00	22,09	23,53	22,66	53,78	68,07	51,40	61,28	63,09	75,24	70,20	79,21	80,29	68,25	70,02	80,52	84,28	100,00	79,76	72,39
Fb_N-30	25,58	19,44	22,12	11,54	18,74	21,65	22,05	55,52	68,81	50,70	57,89	57,55	65,80	68,01	70,51	74,48	66,84	66,31	81,82	79,17	76,81	100,00	74,75
Fb_N-31	27,52	21,91	20,99	11,54	16,96	21,65	21,15	53,20	62,62	47,73	54,32	54,53	64,86	61,61	67,11	77,18	62,46	59,79	74,77	79,17	72,27	77,49	100,00

spektra skupiny Fb_N

%	04	05	06	07	08	09	10	11	12	13	14	15	16	17	19	20
Run_D-sample_04	100,00	36,00	23,08	13,71	10,36	4,59	4,42	3,45	3,11	3,54	3,43	3,64	3,75	3,44	3,27	0,98
Run_D-sample_05	11,69	100,00	13,68	8,12	5,18	2,44	2,28	2,07	1,24	1,42	1,33	2,01	1,56	1,52	1,57	0,00
Run_D-sample_06	35,06	64,00	100,00	41,62	21,36	9,61	6,98	4,14	3,55	4,45	3,91	6,02	5,32	4,05	4,60	1,57
Run_D-sample_07	35,06	64,00	70,09	100,00	43,04	16,07	11,82	6,91	6,75	7,48	7,44	10,16	9,07	7,29	8,11	4,13
Run_D-sample_08	41,56	64,00	56,41	67,51	100,00	27,69	23,65	18,92	13,68	13,35	13,54	20,33	12,30	14,49	16,34	6,69
Run_D-sample_09	41,56	68,00	57,26	56,85	62,46	100,00	59,83	43,78	35,97	45,80	39,28	50,19	35,56	38,20	47,34	13,58
Run_D-sample_10	40,26	64,00	41,88	42,13	53,72	60,26	100,00	63,67	47,34	52,98	46,62	63,74	44,11	49,54	58,47	10,43
Run_D-sample_11	32,47	60,00	25,64	25,38	44,34	45,48	65,67	100,00	38,99	40,55	33,84	55,08	32,74	44,68	57,99	7,87
Run_D-sample_12	45,45	56,00	34,19	38,58	49,84	58,11	75,93	60,64	100,00	81,60	73,78	73,40	68,30	74,67	72,64	29,92
Run_D-sample_13	45,45	56,00	37,61	37,56	42,72	64,99	74,64	55,39	71,67	100,00	76,84	70,14	68,82	65,75	68,40	27,56
Run_D-sample_14	46,75	56,00	35,04	39,59	45,95	59,11	69,66	49,03	68,74	81,50	100,00	71,77	80,92	71,43	63,32	26,57
Run_D-sample_15	37,66	64,00	41,03	41,12	52,43	57,39	72,36	60,64	51,95	56,52	54,53	100,00	55,89	60,59	62,83	13,19
Run_D-sample_16	46,75	60,00	43,59	44,16	38,19	48,92	60,26	43,37	58,17	66,73	73,98	67,25	100,00	65,65	54,84	28,15
Run_D-sample_17	44,16	60,00	34,19	36,55	46,28	54,09	69,66	60,91	65,45	65,62	67,21	75,03	67,57	100,00	80,63	23,82
Run_D-sample_19	35,06	52,00	32,48	34,01	43,69	56,10	68,80	66,16	53,29	57,13	49,86	65,12	47,24	67,48	100,00	23,62
Run_D-sample_20	6,49	0,00	6,84	10,66	11,00	9,90	7,55	5,52	13,50	14,16	12,87	8,41	14,91	12,26	14,53	100,00

spektra skupinyRun_D

Příloha E – publikované výsledky (Biospean)

Franc V, Řehulka P, **Raus M**, Stulik J, Novak J, Renfrow MB, Šebela M. *Elucidating heterogeneity of IgA1 hinge-region O-glycosylation by use of MALDI-TOF/TOF mass spectrometry: role of cysteine alkylation during sample processing.* J. Proteomics

revidovaná verze zaslaná k recenzi 19. června 2013

Elsevier Editorial System(tm) for Journal of Proteomics
Manuscript Draft

Manuscript Number: JPROT-D-13-00230R1

Title: Elucidating heterogeneity of IgA1 hinge-region O-glycosylation by use of MALDI-TOF/TOF mass spectrometry: role of cysteine alkylation during sample processing

Article Type: SI: Protein Modifications

Section/Category: Original Article

Keywords: human immunoglobulin A1 (IgA1); IgA nephropathy; O-glycosylation; glycopeptide; mass spectrometry; microgradient separation

Corresponding Author: Prof. Marek Sebela,

Corresponding Author's Institution: Palacky University

First Author: Vojtech Franc, M.Sc.

Order of Authors: Vojtech Franc, M.Sc.; Pavel Rehulka, Ph.D.; Martin Raus, M.Sc.; Jiri Stulik, Prof.; Jan Novak, Ph.D.; Matthew B Renfrow, Ph.D.; Marek Sebela

Abstract: Determining disease-associated changes in protein glycosylation provides a better understanding of pathogenesis. This work focuses on human immunoglobulin A1 (IgA1), where aberrant O-glycosylation plays a key role in the pathogenesis of IgA nephropathy (IgAN). Normal IgA1 hinge region carries 3 to 6 O-glycans consisting of N-acetylgalactosamine (GalNAc) and galactose (Gal); both sugars may be sialylated. In IgAN patients, some O-glycans on a fraction of IgA1 molecules are Gal-deficient. Here we describe a sample preparation protocol with optimized cysteine alkylation of a Gal-deficient polymeric IgA1 myeloma protein prior to in-gel digestion and analysis of the digest by MALDI-TOF/TOF mass spectrometry (MS). Following a novel strategy, IgA1 hinge-region O-glycopeptides were fractionated by reversed-phase liquid chromatography using a microgradient device and identified by MALDI-TOF/TOF tandem MS (MS/MS). The acquired MS/MS spectra were interpreted manually and by means of our own software. This allowed assigning up to six O-glycosylation sites and demonstration, for the first time, of the distribution of isomeric O-glycoforms having the same molecular mass, but a different glycosylation pattern. The most abundant Gal-deficient O-glycoforms were GalNAc4Gal3 and GalNAc5Gal4 with one Gal-deficient site and GalNAc5Gal3 and GalNAc4Gal2 with two Gal-deficient sites. The most frequent Gal-deficient sites were at Ser230 and/or Thr236.

Cover Letter

Centre of the Region Haná for Biotechnological and Agricultural Research - Department of Protein Biochemistry and Proteomics
Faculty of Science, Palacký University
Šlechtitelů 11
CZ-783 71 Olomouc
CZECH REPUBLIC

Journal of Proteomics

to Editor-in-Chief, Juan J. Calvette

June 18, 2013

Dear Mr. Editor-in-Chief,

I would like to submit a revised version of the manuscript entitled "**Elucidating heterogeneity of IgA1 hinge-region O-glycosylation by use of MALDI-TOF/TOF mass spectrometry: role of cysteine alkylation during sample processing"** by Vojtěch Franc, Pavel Řehulka, Martin Raus, Jiří Stulík, Jan Novak, Matthew B. Renfrow and Marek Šebela, which has recently been evaluated by two reviewers for publishing in the Journal of Proteomics. The manuscript is intended for a special issue (Protein Modification; Guest Editors: Tilman Grüne and Marek Šebela). Both reviewers suggested a minor revision and raised several questions and comments. We went through the corresponding reports and tried to improve the manuscript accordingly.

On separate sheets please find our detailed responses to all questions or comments.

Thank you in advance for considering our revised manuscript.

Sincerely yours

Marek Sebela
Department of Biochemistry
Faculty of Science
Palacký University
Slechtitelu 11
CZ-783 71 Olomouc
Czech Republic

Tel.: +420 585634927
Fax: +420 585634933
E-mail: marek.sebela@upol.cz
Web: http://biochemie.upol.cz

***Detailed Response to Reviewers**

RESPONSES TO REVIEWERS' COMMENTS AND QUESTIONS

REVIEWER 1

GRAPHICAL ABSTRACT
I think that the graphical abstract is maybe too technical for the non-expert reader. It would be helpful if the authors could somehow simplify it or embed a short explanation of the symbols shown.

The graphical abstract has been improved by adding both an explanatory sentence and legend to symbols for the attached glycans.

RESULTS:
1. Line 309: However, we observed a glycopeptide population where Thr225 is not occupied by a carbohydrate. Moreover, Ser230 and Thr236 are the 311 dominant sites of Gal-deficiency on this IgA1 myeloma protein. Based on the estimation of abundance using areas under the peaks, the two glycoforms (5-3, 4-3) with unambiguous monosaccharide assigned at Ser230 represent a substantial part of the studied sample (see Table 1) and match the previous analysis of this IgA1 myeloma protein by use of ECD MS/MS [6].
- The authors should try to discuss the biological properties (if known) of these alternatively glycosylated forms.

Inflammatory responses triggered by IgA1 or IgA2 are primarily mediated by the IgA-specific Fc *receptor, FcαRI, which is a transmembrane glycoprotein expressed on cells of the myeloid lineage* including eosinophils, neutrophils, monocytes*, and macrophages. In a complex with FcαRI, IgA1 binds in* such a way that the C_H2 N-*glycans of IgA1 may directly contact FcαRI and play a role in the FcαRI*-IgA1 interaction (Herr, A. B. et al., Nature 2003; 423: 614–620). The hinge-region O-glycans clearly do not participate in such an interaction as they are too distant from the site of contact between the two proteins. Aberrant O-glycosylation of IgA1 is involved in the pathogenesis of IgA nephropathy (IgAN) and the closely related Henoch-Schönlein purpura nephritis. Interestingly, the aberrantly glycosylated molecules are recognized by the immune system as evidenced by formation of specific antibodies (Suzuki H. et al., J. Clin. Invest.2009; 119:1668–1677) and, consequently, circulating immune complexes are formed that can deposit in the glomerular mesangium. Alternatively glycosylated forms of IgA1 thus generate a different immune response. Whether different glycoforms are better suited to particular immune functions at particular immune sites is not clear (Barratt J. et al., Nephrology 2007; 12:275-284).

This was added into the text.

2. Line 341: Fig. 6 summarizes the isomeric species which scored high in our method for the most abundant Gal-deficient O-glycoforms (see Table 1).

- Is it expected that these O-glycoforms will have distinct functions?

Because of an isomeric arrangement, physicochemical properties of the O-glycoforms will be different. At the moment, we could only speculate about specific functions of different IgA1 hinge-region O-glycoforms. Their existence seems to be a consequence of biosynthetic processes that are still unknown. Positions of O-glycans in IgA1 are dependent on the variety of GalNAc-transferases expressed in IgA-positive cells and their substrate specificities. Differential regulation of these enzymes may be implicated in the production of isomeric O-glycoforms of IgA1.

DISCUSSION
1. Line 358-367:
- Please discuss on the anticipated effects of the macro- and micro heterogeneity of O-glycosylation

and of Gal-deficiency on the function of the myeloma proteins.

IgA has a significant role in mucosal immunity, particularly in the intestinal tract. Its production in the mucosal compartments represents a process, which is important for homeostasis between the gut commensal microflora and the local immunological environment. More specifically, IgA can neutralize toxins and pathogenic microbes as well as restrict the gut commensal flora to the intestinal lumen (Macpherson AJ et al., Mucosal Immunol. 2008, 1: 11–22). As differences in O-glycosylation of the hinge region of IgA1 should not influence the binding sites of antigens in the variable regions, we might expect that all alternatively glycosylated IgA1 molecules will have the same primary function (antigen recognition) regardless of their macro- or microheterogeneity. It seems more important to consider the consequence of the observed heterogeneity (at both levels) for pathogenesis of specific diseases. We addressed this point in the answer to question 1 referring to Results.

2. Line 403: Previously published data obtained for the same protein (although different sample preparation) by use of RPLC-FT-ICR.
- Indicate in more details the differences of this study vs. studies reported in [6]

In 2010, Takahashi et al., Mol Cel Proteomics 9:2545-57, reported the existence of a mixture of isomeric and differently glycosylated O-glycopeptides based on results from ECD/ETD analyses of the most abundant pIgA1 (Ale) O-glycoforms. However, the authors have not analyzed the variable attachment sites and did not perform a quantification study, which would provide relative abundance data for the isomers. Later on, Takahashi et al., J Prot Res 2012; 11:692-702, described a detailed investigation of isomeric O-glycoforms in pIgA (Mce1) as well as in IgA1 from normal human serum. The quantification of relative abundance was based on chromatographic data (peak areas) and the glycan attachment sites were determined by interpretation of MS/MS spectra. Nevertheless, the range of the detected isomeric O-glycoforms is significantly higher in the present study, where also the O-glycoforms 4-2, 4-3, 4-4, 5-3 and 5-4 were found to exist in many isomeric variants.

This paragraph was added into the text.

REVIEWER 2

(1) Throughout the paper, please, stay consistent in the name (and symbol) of the sample to which the newly assessed data refer: In the lines 47 and 107 it is defined as a "Gal-deficient polymeric IgA1 myeloma protein" (designated by pIgA1 or Ale) being a model for Gal-deficient IgA1 in IgAN. In other places it is addressed simply as IgA1. It is not clear for the reader how far these proteins are identical. In this sense "IgA1" should be replaced by "pIgA1" in line 305.

Based on the suggested nomenclature for myeloma proteins, we introduced *"IgA1 (Ale) myeloma protein" we used and as it may exist in a polymeric or monomeric form, we underscored the fact that it* was the polymeric form. In general, the term pIgA1 should thus refer to the sample we used at the beginning of the experimental work. After the pIgA sample had been resolved by SDS-PAGE, the protein was no more polymeric and sensu stricto all digestions followed by MS analyses were started at the level of IgA1 heavy chain. A correction has been made as suggested and the whole text was checked accordingly.

In the discussion part the data assessed for pIgA1 are compared to the data of ref. 6: Didn't the authors of ref.6 use a slightly different sample so that one would expect certain differences in the glycopattern? Please, make a brief comment on this point. It might be essential for evaluating the the significance of these type of data (and interesting for the reader) to clearly work out what is the main variability in the O-glycopattern when comparing different IgA1 samples, e.g. pIgA1 from myeloma with IgA1 in nephropathy or other sources.

Thank you for this excellent comment. We specify in Materials & Methods (line 134) that the procedure used for the preparation of pIgA1 (Ale) in this work was the same as described in ref. [6]. Thus, we could expect that both preparations were very similar if not identical. There was a difference observed in comparison with ref. [6] as regards to relative abundance of the O-glycoforms with 5 and 4 occupied glycosylation sites, which we attribute to the use of different ionization techniques in our work and ref. [6]. Interestingly, when evaluating summarized percentage numbers for the most abundant O-glycoforms from this work and the study by RPLC-FT-ICR MS [6], they are comparable. If a difference in O-glycosylation between two different preparations of pIgA1 (Ale) would occur, the described method has a potential to recognize it.

(2) Did the authors find any sialylation of the glycans? (cf. line 45 and 135). The presence of SAs is usually of high significance for biological processes. If SAs were not found, can the authors exclude that the SAs got lost during sample preparation (cleavage by treatment under acidic conditions when using 0.1 % TFA; or loss by GELoader-tip washing and prefractionation?).

On page 2555 in Takahashi et al., Mol. Cell Proteomics 9:2545-*57, it is stated that "IgA1 (Ale) and other* IgA1 myeloma proteins mimic the Gal-deficient IgA1 in the serum of patients with IgAN as we have demonstrated in previous studies (citations). This IgA1 (Ale) myeloma protein is polymeric and Gal-deficient, reacts with glycan-specific antibodies from sera of IgAN patients, and forms immune complexes that can activate cultured mesangial cells (citations). **However, unlike the Gal-deficient IgA1 in patients with IgAN, IgA1 (Ale) has its O-glycans minimally sialylated.***" In the light of this knowledge, we did not* consider the existence of a sialylation in our sample.

(3) In the eyes of this reviewer, the final part of the discussion (beginning from line 433) becomes to some extent speculative in the sense that all these conclusions can hardly be derived from the data gained here. There would be a need for additional in vitro data measuring the effect of GalNAc-transferase activities when acting over various time intervals. This could be done by measuring time series using the method proposed. (But this would probably be a topic for an extra paper.)

Yes, this is true, we fully realize that there are rather speculative tones in the mentioned part of the Discussion section. But we discuss our results in context of a previous communication as it was necessary to discuss suggestions and ideas by Takahashi et al., 2010 (ref. 6). To indicate clearly that such thinking cannot stand firmly on results obtained in th*is work, we replaced at line 456 the words "we propose ..." by "When thinking about biological relevance of our results, we hypothesize that ....".*

Some formal points:

(-) Keywords: cysteine alkylation: is this a keyword one would look for ?

Because of the presence of Cys residues in the studied tryptic HR O-glycopeptides, the role of a quantitative cysteine alkylation was found very important to prevent from a splitting of the corresponding signals in MALDI-TOF/TOF MS analyses. For that reason we thought it would be a good idea to include a phrase *"cysteine alkylation" into the group of keywords. As suggested, new keywords "microgradient separation" were* now added as a substitute.

(-) Line 44: "change 3-6 O-glycans" to "3 to 6 O-glycans"

This has been corrected as suggested.

(-) Line 414: "within the off-line ____ resulted" obviously, there are some words missing.

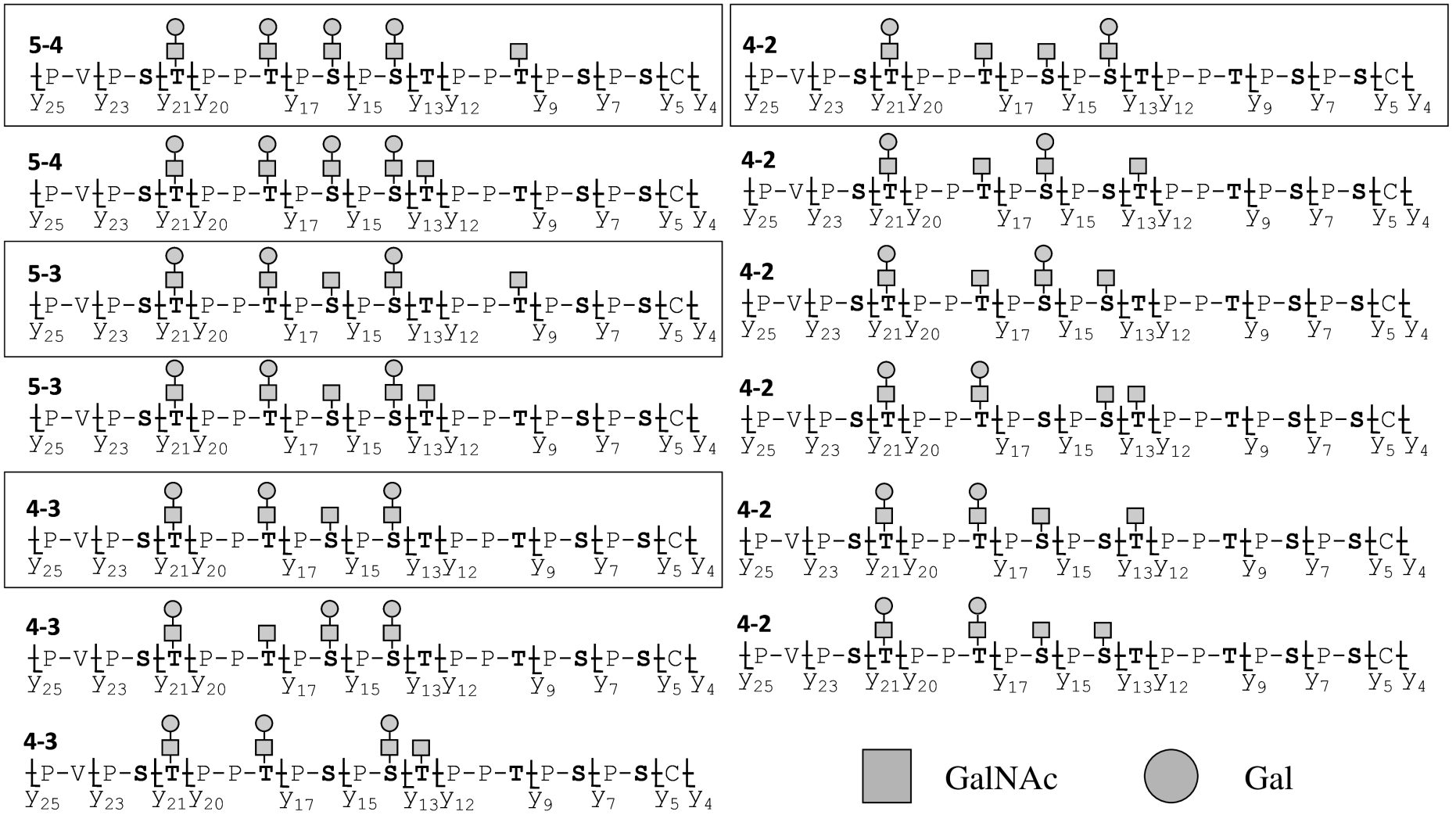
This has been corrected and the whole text checked for missing words.

(-) Line 469: We have introduced "a new approach". This is a too strong claim.

This has been changed to *read "a modified approach"*

***Graphical Abstract (for review)**

Amino-acid sequence segments of selected heavy-chain O-glycopeptide isoforms originating from the hinge region of human myeloma pIgA1 (Ale) protein, which were detected in the present study by a software processing of MALDI-TOF/TOF MS/MS data. Frames indicate abundant isoforms; characteristic y-type fragment ions observed in the corresponding MS/MS spectra are labeled. For each isoform, a code with the number of attached GalNAc and Gal residues is provided.



*Highlights (for review)

Franc V, et al.,

Elucidating heterogeneity of IgA1 hinge-region O-glycosylation by use of MALDI-TOF/TOF mass spectrometry: role of cysteine alkylation during sample processing

Highlights

- This work focuses on the O-glycosylation of immunoglobulin A1 (IgA1)
- IgA1 hinge-region glycopeptides were chromatographed using a microgradient device
- MALDI-TOF/TOF tandem mass spectrometry allowed assigning O-glycosylation sites
- The most frequent galactose-deficient sites were at Ser230 and/or Thr236
- The distribution of isomeric O-glycoforms is demonstrated using a new software

Franc V, et al.,

Elucidating heterogeneity of IgA1 hinge-region O-glycosylation by use of MALDI-TOF/TOF mass spectrometry: role of cysteine alkylation during sample processing

Significance

In this work, we studied the O-glycosylation in the hinge region of human immunoglobulin A1 (IgA1). Aberrant glycosylation of the protein plays a key role in the pathogenesis of IgA nephropathy. Thus identification of the O-glycan composition of IgA1 is important for a deeper understanding of the disease mechanism, biomarker discovery and validation, and implementation and monitoring of disease-specific therapies. We developed a new procedure for elucidating heterogeneity of IgA1 O-glycosylation. After running a polyacrylamide gel electrophoresis under denaturing conditions, the heavy chain of IgA1 was subjected to in-gel digestion by trypsin. O-glycopeptides were separated from the digest on capillary columns using a microgradient chromatographic device (replacing commonly used liquid chromatographs) and subjected to MALDI-TOF/TOF mass spectrometry (MS) and tandem mass spectrometry (MS/MS) involving post-source decay fragmentation. We show that the complete modification of cysteines by iodoacetamide prior to electrophoresis is critical for successful MS/MS analyses on the way to deciphering the microheterogeneity of O-glycosylation in IgA1. Similarly, removal of the excess of the reagent is equally important. The acquired MS/MS allowed assigning up to six O-glycosylation sites and identification of isomeric O-glycoforms. We show that our simplified approach is efficient and has a high potential to provide a method for the rapid assessment of IgA1 heterogeneity that is a less expensive and yet corroborating alternative to LC-(high-resolution)-MS protocols. The novelty and biological significance resides in demonstration, for the first time, of the distribution of the most abundant isoforms of HR O-glycopeptides of IgA1. As another new feature, we introduce a software solution for interpretation of MS/MS data of O-glycopeptide isoforms, which provides the possibility of fast and easier data processing.

*Manuscript
Click here to download Manuscript: JProt_Franc_et_al_IgA_revision_main_text.doc Click here to view linked References

Elucidating heterogeneity of IgA1 hinge-region O-glycosylation by use of MALDI-TOF/TOF mass spectrometry: role of cysteine alkylation during sample processing

Vojtěch Franc[1], Pavel Řehulka[2*], Martin Raus[3], Jiří Stulík[2], Jan Novak[4], Matthew B. Renfrow[5], and Marek Šebela[1,3***]**

1) Department of Protein Biochemistry and Proteomics, Centre of the Region Haná for *Biotechnological and Agricultural Research, Faculty of Science, Palacký University, Šlechtitelů 11,* CZ-783 71 Olomouc, Czech Republic; 2) Institute of Molecular Pathology, Faculty of Military Health *Sciences, University of Defence, Třebešská 1575, CZ*-500 01 Hradec Králové, Czech Republic; 3) *Department of Biochemistry, Faculty of Science, Palacký University, Šlechtitelů 11, CZ*-783 71 Olomouc, Czech Republic; 4) Department of Microbiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA; 5) Department of Biochemistry and Molecular Genetics, University of Alabama at Birmingham, Birmingham, AL 35294, USA

*Correspondence to: P. Řehulka, Institute of Molecular Pathology, Faculty of Military Health Sciences, University of Defence, Třebešská 1575, CZ- 500 01 Hradec Králové, Czech Republic. Tel.: +420 973255199.

**Correspondence to: J. Novak, Department of Microbiology, University of Alabama at Birmingham, 845 19th Street South, BBRB 761A, Birmingham, AL 35294, USA. Tel.: +1 (205) 934-4480.

***Correspondence to: M. Šebela, Department of Biochemistry, Faculty of Science, Palacký University, Šlechtitelů 11, CZ-783 71 Olomouc, Czech Republic. Tel.: +420 585634927; fax: +420 585634933.

E-mail addresses: marek.sebela@upol.cz (M. Šebela), jannovak@uab.edu (J. Novak), rehulka@pmfhk.cz (P. Řehulka).

Keywords

human immunoglobulin A1 (IgA1); IgA nephropathy; O-glycosylation; glycopeptide; mass spectrometry; microgradient separation

Abbreviations

ACN, acetonitrile; CHCA, α-cyano-4-hydroxycinnamic acid; DTT, dithiothreitol; ECD, electron capture dissociation; ETD, electron transfer dissociation; FEP, fluorinated ethylene propylene; FT-ICR, Fourier transform ion cyclotron resonance; IAM, iodoacetamide; IgA, immunoglobulin A; IgAN; IgA nephropathy; HR, hinge region; LC, liquid chromatography; MALDI-TOF, matrix-assisted laser desorption/ionization time-of-flight; MS, mass spectrometry; MS/MS, tandem mass spectrometry; PSD, post-source decay; RPLC, reversed-phase liquid chromatography; TFA, trifluoroacetic acid

Abstract

Determining disease-associated changes in protein glycosylation provides a better understanding of pathogenesis. This work focuses on human immunoglobulin A1 (IgA1), where aberrant O-glycosylation plays a key role in the pathogenesis of IgA nephropathy (IgAN). Normal IgA1 hinge region carries 3 to 6 O-glycans consisting of N-acetylgalactosamine (GalNAc) and galactose (Gal); both sugars may be sialylated. In IgAN patients, some O-glycans on a fraction of IgA1 molecules are Gal-deficient. Here we describe a sample preparation protocol with optimized cysteine alkylation of a Gal-deficient polymeric IgA1 myeloma protein prior to in-gel digestion and analysis of the digest by MALDI-TOF/TOF mass spectrometry (MS). Following a novel strategy, IgA1 hinge-region O-glycopeptides were fractionated by reversed-phase liquid chromatography using a microgradient device and identified by MALDI-TOF/TOF tandem MS (MS/MS). The acquired MS/MS spectra were interpreted manually and by means of our own software. This allowed assigning up to six O-glycosylation sites and demonstration, for the first time, of the distribution of isomeric O-glycoforms having the same molecular mass, but a different glycosylation pattern. The most abundant Gal-deficient O-glycoforms were $GalNAc_4Gal_3$ and $GalNAc_5Gal_4$ with one Gal-deficient site and $GalNAc_5Gal_3$ and $GalNAc_4Gal_2$ with two Gal-deficient sites. The most frequent Gal-deficient sites were at Ser230 and/or Thr236.

1. Introduction

Immunoglobulin A (IgA) is the most abundantly produced antibody with an important role in mucosal immunity. It occurs in two structurally and functionally distinctive subclasses IgA1 and IgA2 [1]. In contrast to IgA2, the hinge region (HR) of IgA1 (Fig. 1) contains two octapeptide repeats with multiple Pro, Ser, and Thr residues and 3 to 6 O-glycan chains [2-4]. IgA1 O-glycans consist of N-acetylgalactosamine (GalNAc) that may carry galactose (Gal) and/or sialic acid (Neu5Ac) giving rise to several different IgA1 O-glycoforms (Fig. 1) [5-9]. An abnormal glycosylation of serum immunoglobulins and other glycoproteins has been observed in several human diseases [10-15]. In 1968, IgA nephropathy (IgAN) was described as a clinical entity [16] and has been since recognized as the most common primary glomerulonephritis and an important cause of end-stage renal disease [17-19]. Although the precise mechanism of IgAN pathogenesis is still being elucidated, it clearly involves formation of immune complexes that contain IgA1 with Gal-deficient O-glycans [10,20]. IgAN is diagnosed based on evaluation of renal biopsy; no alternative noninvasive diagnostic method is currently available [21-23]. Patients with IgAN have elevated levels of circulatory IgA1 with Gal-deficient O-glycans; these molecules are bound in pathogenic immune complexes (for review see: [22,23]). Therefore, identification of the O-glycan composition of IgA1 is important for a deeper understanding of the disease mechanism, biomarker discovery and validation, and implementation and monitoring of disease-specific therapies [24].

O-glycosylated isoforms from a single biological source show a distinct distribution of heterogeneity with respect to the number and structure of glycan chains [25]. Many methodological approaches have been employed for the analysis of aberrant O-glycosylation in IgA1 [26]. The Human Disease Glycomics/Proteome Initiative associated with the Human Proteome Organisation recently coordinated a multi-institutional study that evaluated methodologies with a wide use for defining N-glycan content in glycoproteins [27]. Such activities have also been extended to O-glycans [28]. These studies have shown that mass spectrometry (MS) is the most powerful tool for both identification and quantification of N- and O-glycans. The possibility of a precise assessment of mucin-type O-glycans has been successfully demonstrated using tandem mass spectrometry (MS/MS) with electron capture dissociation (ECD) or electron transfer dissociation (ETD) [6,25,29-31].

So far, two main strategies have been adopted to assess the heterogeneity of O-glycans in HR of IgA1: lectin binding assays ideally combined with monosaccharide compositional analysis [32-34] and MS analysis. Notably, only a few studies have shown a direct assignment of multiple sites of O-glycan attachment [3,6,25,29]. There have been several reports based on matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) [7,35,36], but there is always a limitation in resolving power and mass accuracy. Every single O-glycoform of IgA1 represents a unique glycoprotein species with different abundance and possibly different biological role. Microscale solid-phase extraction methods have been used in proteomics for sample desalting, enrichment and fractionation. For that purpose, pipette tips with immobilized sorbents are used [37-39]. Another easy and economical approach for peptide fractionation involves a simple microgradient device for reversed-phase liquid chromatography (RPLC) coupled offline to MALDI-TOF MS [40].

In patients with IgAN, the O-glycosylation pattern includes Gal-deficiency but it is not fully understood whether and how it occurs at specific sites [24]. MS analysis of IgA1 O-glycosylation is complicated especially for two reasons. First, the O-glycosylation pattern of HR is rather complex. Secondly, the tryptic peptide containing HR is relatively large (38 amino acids) and thus the molecular mass of the corresponding O-glycopeptides usually exceeds 5 kDa. Recently, both glycan distribution and specific native sites of Gal deficiency for the major O-glycoforms within a single sample were elucidated by RPLC coupled to Fourier transform ion cyclotron resonance (FT-ICR) MS with ECD and ETD fragmentation. The sample was digested by a combination of trypsin and IgA1-specific proteases to obtain shorter glycopeptides [6].

In this study, a polymeric IgA1 (pIgA1) myeloma protein [20] was used that mimics Gal-deficient IgA1 in IgAN. We developed a novel experimental approach with optimized sample preparation, SDS-PAGE, in-gel tryptic digestion of the heavy chain of IgA1, RPLC separation of the obtained peptide mixture (using a simple microgradient device), and MALDI-TOF/TOF MS and MS/MS analyses of O-glycopeptides. The most important steps were both reduction/alkylation of the protein and RPLC separation of peptides. MALDI-TOF/TOF MS detected HR peptides with up to six O-glycans and the subsequent MS/MS analyses determined the sites of occupancy for most of them. Our data are consistent with recent results obtained using RPLC coupled online to FT MS [3]. The main

advantage of the present approach is in its simplicity, robustness, low cost, speed and reliability. The biological significance of this work resides in the demonstration of the distribution of HR glycopeptide isoforms having the same molecular mass, but a different O-glycosylation pattern (site-specific microheterogeneity).

2. Materials and methods

2.1. Chemicals and materials

High-quality solvents for liquid chromatography/MS and MALDI matrices were from Sigma-Aldrich Chemie (Steinheim, Germany). POROS Oligo R3 50 μm particles were obtained from PerSeptive Biosystems (Framingham, MA, USA) and packed into GELoader pipette tips (Eppendorf, Hamburg, Germany). Polyimide-coated fused silica capillaries (360 μm o. d., 200 μm i. d.; and 360 μm o. d., 50 μm i. d.) were purchased from Agilent Technologies (Santa Clara, USA), AQ-C8 5 μm particles and AQ-C18 5 μm particles for laboratory filled columns were purchased from Dr. Maisch GmbH (Ammerbuch, Germany). A microtight union (5/16-24, 360 μm o.d., cat. no. P-772) was obtained from UpChurch Scientific (Oak Harbor, USA), fluorinated ethylene propylene (FEP) tubing (1/16" x 0.25mm i.d.) was from Vici AG International (Schenkon, Switzerland).

2.2. Gal-deficient IgA1, cysteine alkylation, SDS-PAGE, and in-gel digestion

pIgA1 myeloma protein (Ale) was purified as described previously [6]. This myleoma protein has almost no sialic acid on the HR O-glycans. An aliquot (10 μl) of pIgA1 (Ale), 0.9 mg ml^{-1}, was mixed with 4 μl of 10 mM dithiothreitol (DTT) in 100 mM NH_4HCO_3. The mixture was incubated at 56 °C for 30 min. The reduced pIgA1 was then alkylated in the dark for 30 min by the addition of 2.5 μl of 50 mM iodoacetamide (IAM) in 100 mM NH_4HCO_3. After the carbamidomethylation, the sample solution was mixed in a volume ratio of 1:1 with 2x concentrated Laemmli sample buffer containing 200 mM DTT. Then the resulting excess of the reducing agent was quenched by adding 9 μl of 1 M IAM. Aliquots of this mixture (10.5 μl) were loaded on the top of a manually prepared discontinuous polyacrylamide gel slab (4% T/3.3%C stacking gel; 10% T/3.3% C resolving gel); 100 mm x 100 mm

x 1 mm in size. SDS-PAGE was performed by a standard method [41]. The separated antibody chains were visualized by Bio-Safe™ Coomassie Stain (Bio-Rad, Hercules, CA, USA). In-gel digestion of the heavy chain by trypsin was conducted using a previous protocol [42].

2.3. On-tip desalting and fractionation of digests

For desalting of tryptic digests, GELoader tips filled with POROS Oligo R3 50 µm particles were used [37]. The outlet of a GELoader tip was carefully squeezed with a Combi Tip syringe (Eppendorf, Hamburg, Germany) on flat surface and then 1 µl of POROS Oligo R3 suspension in acetonitrile (ACN) was loaded into the GELoader tip. The tip content was wetted with 10 µl 80% ACN/0.1% TFA (v/v) and equilibrated with 10 µl 2% ACN/0.1% TFA (v/v). A digest sample (40 µl) acidified with 5% (v/v) TFA (using a volume ratio of acid to sample of 1:10) was passed through the equilibrated tip and washed with 10 µl 2% ACN/0.1% TFA (v/v). Elution was achieved with 10 µl of 60% ACN/0.1% TFA (v/v), the eluate evaporated in a vacuum centrifuge and dissolved in 50 µl of 0.1% (v/v) TFA. Fractionation was done by a sequential elution of the retained sample using 12%, 24% and 35% ACN/0.1% TFA solutions (v/v; 10 µl each). The fraction containing HR O-glycopeptides (24% ACN/0.1% TFA) was subjected to MALDI-TOF/TOF MS for both O-glycopeptide profiling and estimating the relative abundance of O-glycoforms based on the isotope cluster area.

2.4. Preparation of capillary columns and RPLC separation using simple microgradient device

Capillary columns were prepared as previously described [43], except that two stationary phases were selected: either AQ-C8, 5 µm, for filling up precolumns or AQ-C18, 5 µm, for analytical columns. A simple microgradient device was then used for RPLC separation of O-glycopeptide fractions from tryptic digests (Supplementary Fig. 1). It consisted of a syringe infusion pump (Harvard Apparatus, Holliston, MA, USA) with a 100-µl microsyringe (SGE, Melbourne, Australia) connected to a 250 mm long capillary (360 µm o. d., 200 µm i. d.) using a piece of FEP tubing [43]. The capillary was connected to a home-made C8 precolumn (50 mm) through a 360-µm microtight union. The precolumn was finally connected using a second 360 µm microtight union to a home-made C18 analytical column (75 mm). The whole system was first wetted with 40 µl of 80% ACN/0.1% TFA

(v/v) and then equilibrated with 40 µl of 2% ACN/0.1% TFA (v/v) after aspiration of the appropriate solvent into the microsyringe through capillary disconnected from the precolumn. Desalted sample (dried and dissolved in 50 µl 0.1% TFA, v/v) was aspirated into the syringe and loaded on the precolumn disconnected from the analytical column just for the sample loading process. The sample-loaded precolumn was then washed (40 µl 2% ACN/0.1% TFA, v/v) and connected back to the analytical column. Prior to sample separation, the microsyringe with connecting capillary was filled consecutively by four ACN–0.1% TFA (v/v) mobile phases with gradually decreasing ACN content (30% ACN/0.1% TFA – 40 µl, 20% ACN/0.1% TFA – 10 µl, 10% ACN/0.1% TFA – 10 µl, 2% ACN/0.1% TFA – 18 µl). An S-shaped gradient of mobile phase was formed inside the microsyringe as a result of the turbulent mixing of solvents during aspiration [44]. Finally the microsyringe (with the attached capillary) placed in the syringe infusion pump was connected to the precolumn (that was further connected to analytical column) and separation was started. The flow rate was 2 µl min^{-1} during sample loading and 0.8 µl min^{-1} in all other cases. Eluate drops from the analytical column were manually deposited in 20-s intervals (~0.3 µl) on MALDI targets 22 min after separation had started. Immediately after eluate drop deposition, 0.5 µl of α-cyano-4-hydroxycinnamic acid (CHCA) as a MALDI matrix, 5 mg ml^{-1} in 60% ACN/1 mM diammonium hydrogen citrate, was added for crystallization.

2.5. MS and MS/MS analyses

MALDI-TOF/TOF MS and MS/MS analyses were performed in the positive mode on an ABI 4800 Proteomics Analyzer (Applied Biosystems, Framingham, MA, USA) equipped with a neodymium-doped: yttrium aluminum garnet laser (355 nm) of 3-7 ns pulse and 200-Hz firing rate. Both MS and MS/MS spectra were acquired using a dual-stage reflectron mirror. Accelerating voltages applied for MS and MS/MS measurements were 20 kV and 8 kV, respectively. In MS/MS mode, a potential of 1 kV was applied and fragmentation was achieved by post-source decay [45]. The matrix composition is provided above. A mixture of six standard peptides was used for external calibration. MS and MS/MS data were processed using 4000 Series Explorer (Applied Biosystems) and further manually interpreted using Data Explorer 4.8 (Applied Biosystems) software. The following parameters were set

up for estimating the relative abundance of O-glycoforms: Integration Baseline Settings: Valley to Baseline; Peak Detection Settings: %Centroid – 50, S/N Threshold – 6, Noise Window Width (m/z) – 250, Recalculate S/N from Cluster Area – Yes, Threshold after S/N Recalculation – 12, Spectrum Resolution – 10000 for m/z 1000, 16000 for m/z 2000, 13000 for m/z 3000 and 13000 for m/z 6500.

2.6. Software inspection of MS/MS spectra of O-glycopeptides

A home-made software BIOSPEAN (generated by M. Raus at Palacký University in Olomouc for intact cell/spore MS purposes) was used for a rapid inspection of the acquired MS/MS spectra. The software is a web program application and will be published separately. BIOSPEAN utilizes Linux as an operating system on the physical server, Apache as a web server software, MySQL as a database engine and PHP as a scripting language. BIOSPEAN allows detection of peaks in a mass spectrum and is able to evaluate its similarity with other spectra when searching over a spectral database. The principle of peak detection resides in a local scanning of intensity values around a particular m/z position taking the level of noise into consideration by adjusting signal-to-noise (S/N) ratio. The score value (expressed as a percentage) for a comparison of an inspected spectrum with another spectrum is derived from the number of identical peak positions found (assigned with an adjustable mass tolerance with the corresponding penalization) divided by the total number of detected peaks in the inspected spectrum. The two spectra may optionally be compared also in the opposite way and the score values obtained are then averaged.

3. Results

3.1. Subject of the study and experimental strategy

The following amino-acid residues in the HR of human IgA1 have been identified as the predominant sites of O-glycan attachment: Thr225, Thr228, Ser230, Ser232, Thr233 and Thr236 (Fig. 1) [3,6]. The range of O-glycosylation (3 to 6 O-glycans attached to 9 possible sites) and the composition of O-glycans at each binding site result in a complex heterogeneity of human IgA1. To elucidate this heterogeneity, it is necessary to include analyses of structural isomers of HR that occur based on the alternative location of O-glycans and their chemical composition. These requirements present a great challenge for current separation and detection methods. In this work, analysis of IgA O-glycan heterogeneity involved five steps: (1) optimal reduction/alkylation of a pIgA1 sample; (2) SDS-PAGE for the separation of the heavy chain of pIgA1; (3) in-gel tryptic digestion of the heavy chain; (4) fractionation and separation of the digest containing IgA HR O-glycopeptides using a simple microgradient device; (5) analysis of the separated O-glycopeptides by MALDI-TOF/TOF MS and PSD-MALDI-TOF/TOF MS/MS.

3.2. Preparation of IgA1 HR O-glycopeptides

Prior to the separation of heavy chain of pIgA1 by SDS-PAGE, the protein was reduced by DTT and the free cysteine thiol groups were then modified by IAM. After SDS-PAGE, the heavy chain was in-gel digested with trypsin (no further reduction/alkylation is needed during the procedure), generating HR glycopeptides of the following amino-acid sequence: HYTNPSQDVTVPCPVPSTPPTPSPSTPPTPSPSCCHPR. This fragment contains 13 Pro, 6 Thr, 6 Ser, and 3 Cys residues; Thr and Ser may be O-glycosylated. Quantitative in-solution cysteine reduction and alkylation by IAM was a crucial step for the optimal detection of O-glycopeptides by MALDI-TOF/TOF MS. Fig. 2 illustrates the effect of cysteine alkylation by IAM on the intensity of IgA1 HR O-glycopeptides in a set of typical mass spectra. Our data show that alkylating the sample just prior to the electrophoresis results in abundant HR O-glycopeptide ions in MALDI-TOF/TOF MS. The next step involved desalting or pre-fractionation of the tryptic digest. The resultant sample was

249 either used for a profiling of IgA1 HR O-glycopeptides and estimation of relative abundance of
250 individual isoforms (after pre-fractionation) or it was subjected to RPLC fractionation by a
251 microgradient device (after desalting). Fractions containing IgA1 HR O-glycopeptides were then
252 analyzed by MALDI-TOF/TOF MS and MS/MS.
253
254 **3.3. Profiling and relative quantification of IgA1 HR O-glycopeptides**
255 O-glycopeptide profiling was achieved after pre-fractionation of pIgA1 heavy-chain tryptic digests
256 on GELoader tips filled with POROS Oligo R3 50-µm particles. A fraction containing HR O-
257 glycopeptides (eluted in 24% ACN/0.1% TFA) was analyzed by MALDI-TOF/TOF instrument and
258 MS profiles were acquired for estimating the relative abundance of O-glycoforms (Fig. 3). The
259 estimation was based on processing isotope cluster areas for O-glycopeptides from 15 MS profiles
260 obtained from 5 independent tryptic digests (each sample was analyzed in 3 replicates). All spectra
261 were processed by the software Data Explorer (Applied Biosystems). The most abundant O-glycoform
262 was HR containing 4 GalNAc and 4 Gal residues (designated as 4-4), whereas the most abundant Gal-
263 deficient O-glycoforms were two variants, 4-3 and 5-4, with one Gal-deficiency and 5-3 and 4-2 with
264 two Gal-deficient sites (Table 1). Notably, HR O-glycoforms with a single Gal-deficiency were more
265 abundant than those with two Gal residues missing.
266
267 **3.4. RPLC separation of IgA1 HR O-glycopeptides and their MS/MS analysis**
268 To detect most HR O-glycopeptides, an aliquot of an in-gel tryptic digest of pIgA1 heavy chain
269 (from about 15 µg total IgA) was desalted on GELoader tips with POROS Oligo R3 50 µm particles.
270 The purified sample was reconstituted in 0.1%TFA and separated by RPLC in the C18 capillary
271 column attached to the microgradient device. Eluted fractions were collected in 20-s intervals (fraction
272 volume of ~ 0.3 µl), directly spotted on a MALDI target, and immediately overlaid with 0.5 µl of
273 CHCA matrix solution. MALDI-TOF/TOF MS analysis of all spotted fractions revealed the presence
274 of 19 unambiguous HR O-glycopeptide ion peaks in 15 fractions (Supplementary Fig. 2), and 15 of
275 them were selected as precursors for the subsequent MS/MS.

MS/MS data were essential for assigning the sites of O-glycan attachment in pIgA1 (Ale). Fig. 4a displays an MS/MS spectrum acquired for a tryptic O-glycopeptide with precursor m/z of 5638.7 [HR+ 5 GalNAc + 3 Gal + H]$^+$. During interpretation of MS/MS spectra, characteristic y-series fragment ions (marked by arrows in Fig. 4) allowed confirmation or exclusion of O-glycan chains at particular sites. In Fig. 4a, from the C-terminus, the y_9 ion at a theoretical molecular mass calculated from the sequence excludes the presence of O-glycans at Ser238 and Ser240. The y_{12} ion, which is shifted by 203 Da when compared with its expected mass, indicates the presence of GalNAc at Thr236. However, interpretation of the spectra is complicated by the presence of another y_{12} ion, without GalNAc at Thr 236. This would suggest the existence of two isomeric 5-3 O-glycoforms, where the second isoform has a GalNAc residue attached at Thr233. Alternatively, this could be a y_{12} ion that has lost its GalNAc due to the presence of a [M-GalNAc + H]$^+$ in the spectra. However based on previous reports of the utility of PSD-MALDI-TOF MS/MS in localizing sites of O-glycosylation [46] and previous reports of isomers existing between Thr233 and Thr236 [3], the fragmentation peaks in this MS/MS spectra would appear to represent true isomers. In similar fashion, all O-glycosylation sites were identified. The y_{15} ion confirms a disaccharide binding at Ser232, the y_{17} ion indicates a monosaccharide at Ser230, while the y_{20} and y_{21} ions prove a disaccharide at Thr228 and Thr225, respectively. Larger y ions, e.g., y_{27}, support an interpretation that no carbohydrates are present at the other potential sites of O-glycan attachment (Ser224, Thr217, Ser213 and Thr210).

Figs. 4b and 4c show PSD-MALDI TOF/TOF MS/MS spectra for precursor ions with m/z of 5435.6 [HR + 4 GalNAc + 3 Gal + H]$^+$ and m/z of 5273.5 [HR + 4 GalNAc + 2 Gal + H]$^+$, respectively. Again, there were multiple y-series fragments that allowed identification of occupied O-glycosylation sites. However, the analysis of isomeric O-glycoforms for these and other HR glycoforms was much more complicated with multiple fragment ions that could be assigned to several different possible combinations of GalNAc-Gal and GalNAc attachments. To assess these various PSD-MALDI fragment possibilities arising from the same precursor mass, a computational approach was developed, as described in the section 3.5 below. All MS/MS spectra referring to the O-glycoforms listed in Table 1 are provided in Supplementary Fig. 3.

3.5. Defining overall microheterogeneity of IgA1 HR O-glycans

The MS/MS data revealed a very complex heterogeneity for most pIgA1 (Ale) O-glycosylation sites (Thr225, Thr228, Ser230, Ser232, Thr233 and Thr236), in agreement with previous publication that used LC-FT MS/MS approach to study O-glycosylation in pIgA (Mce1) myeloma protein and serum IgA1 [3]. For the most common HR O-glycoforms of pIgA1 (Ale), the overall microheterogeneity is shown in Fig 5. The sites at Thr225, Thr228 and Ser232 are each modified predominantly with a GalNAcGal disaccharide. However, we observed a glycopeptide population, where Thr225 is not occupied by a carbohydrate. Moreover, Ser230 and Thr236 are the dominant sites of Gal-deficiency on this IgA1 myeloma protein. Based on the estimation of abundance using areas under the peaks, the two glycoforms (5-3, 4-3) with unambiguous monosaccharide assigned at Ser230 represent a substantial part of the studied sample (see Table 1) and match the previous analysis of this IgA1 myeloma protein by use of ECD MS/MS [6].

The complete set of MS/MS spectra (see in Supplementary Fig. 3) showed a notable feature: in addition to the intense y-series fragments that allowed direct assigning of the binding sites of O-glycans with a high probability, there were additional fragment ions in the MS/MS spectra, which complicated unambiguous elucidation. Based on the previously documented [3] and now confirmed existence of six O-glycosylation sites in the HR of IgA1 (see above) and the assumption that there are three to six GalNAc residues with or without Gal, one could consider multiple potential combinations. Moreover, a possible presence of numerous isomeric O-glycoforms and the potential for some loss of sugar residues off of individual fragments complicate the interpretation and straightforward conclusions.

To tackle this problem, we used a computational approach with the idea that there is a group of typical y-series fragment ions, which allow assigning of individual O-glycosylation sites (y_{12}, y_{13}, y_{15}, y_{17}, y_{20} and y_{21}). A computer script was written to generate all possible O-glycoforms covering the level of O-glycosylation from 3-1 up to 6-5 (Fig. 5). The number of combinations was 262 at the given setup, providing the same number of virtual MS/MS spectra comprising all typical fragment peaks, which may occur when considering that each of the O-glycosylation sites in an isomeric form may carry no carbohydrate, GalNAc, or GalNAc-Gal.

Twelve selected experimental MS/MS spectra were individually compared with the set of virtual spectra using the software BIOSPEAN. The agreement level in every single comparison was assessed by a score value derived from the number of overlapping peaks (considering a preset mass tolerance of 0.5 Da) with a S/N ratio of 3:1 for peak picking; see Materials and methods. Using another script, a large matrix of score values (262 rows x 12 columns) was generated, as shown on example in Table 2. For isomeric O-glycopeptide forms, the combinations showing a score value above 80 (green and yellow color in Table 2) would be considered as highly probable (under a condition that there is an agreement in precursor ion mass between the experimental and virtual spectrum). However, we cannot exclude the presence of other isoforms, with lower but still high enough score values (e.g., around 60-70). By decreasing S/N ratio, the number of matching pairs increases. Supplementary Table shows a complete data set obtained with a S/N of 2.25:1. Fig. 6 summarizes the isomeric species which scored high in our method for the most abundant Gal-deficient O-glycoforms (see Table 1). There were 2 isomeric species identified for each of the O-glycoforms 5-4 and 5-3 (for the latter case, compare with Fig. 4a). Three isomeric molecules were implicated for the O-glycoform 4-3 and evidence of six possible isomers were found in the PSD-MALDI MS/MS for the O-glycoform 4-2. Still, in the absence of separation, these identifications are based on interpretations of PSD-MALDI-TOF/TOF MS/MS spectra and would need further confirmation by higher resolution methods for LC separation and MS resolution or by use of standards of HR isomeric glycoforms. However, based on previous reports of isomeric IgA1 HR O-glycoforms, our custom matching of theoretical and observed MS/MS spectra provides targeted series of possible isomers that could be present in a ranked order.

4. Discussion

Approximately half of all proteins have glycosylation as a post-translational modification and thus studies of glycoprotein structures are important [47]. The aim of this work was to perform by the use of PSD-MALDI FOF/TOF MS identification and direct localization of O-glycosylation in the HR of polymeric IgA1 myeloma protein (Ale) that has served as a model of Gal-deficient IgA1 in IgAN [20,24,48]. For a deeper understanding and determination of O-glycoforms that are involved in the pathogenesis of IgAN, it is necessary to identify all sites with Gal-deficiency. Defining both macro-

and micro heterogeneity of O-glycosylation in the HR is one of the biggest analytical challenges in the analysis of IgA1. A recent study of pIgA1 myeloma protein (Ale) reported that ECD/ETD fragmentation combined with an IgA-specific protease digestion can define the distribution and specific sites of Gal-deficiency for a majority of the O-glycoforms within a single sample of IgA1. In that study, five O-glycosylation sites were identified at Thr225, Thr228, Ser230, Ser232 and Thr236. The most frequent Gal-deficient positions were localized at Ser230 and Thr236 [6]. In a more recent study, another IgA1 myeloma protein (Mce) and IgA1 from normal human sera were analyzed. The authors have reported the discovery of isoforms of HR glycopeptides and a sixth native site of O-glycosylation at Thr233 [3].

Inflammatory responses triggered by IgA1 or IgA2 are primarily mediated by the IgA-specific Fc receptor, FcαRI, which is a transmembrane glycoprotein expressed on cells of the myeloid lineage including eosinophils, neutrophils, monocytes, and macrophages. In a complex with FcαRI, IgA1 binds in such a way that the C_H2 N-glycans of IgA1 may directly contact FcαRI and play a role in the FcαRI-IgA1 interaction [49]. The hinge-region O-glycans clearly do not participate in such an interaction as they are too distant from the site of contact between the two proteins. Aberrant O-glycosylation of IgA1 is involved in the pathogenesis of IgAN and the closely related Henoch-Schönlein purpura nephritis. Interestingly, the aberrantly glycosylated molecules are recognized by the immune system as evidenced by formation of specific antibodies [48] and, consequently, circulating immune complexes are formed that can deposit in the glomerular mesangium. Alternatively glycosylated forms of IgA1 thus generate a different immune response. Whether different glycoforms are better suited to particular immune functions at particular immune sites is not clear [50].

In this work, we confirmed all of the five sites in pIgA1 (Ale) and additionally localized the sixth O-glycosylation site at Thr233 for the first time in this model protein. Our strategy relied on MALDI-TOF/TOF instrument and a key point in the procedure of sample preparation. SDS-PAGE was used for the separation of the heavy chain of pIgA1. This step allowed purification of the sample from impurities and reducing complexity of peptides after the subsequent digestion with trypsin as seen with previous reports [25]. However, a problem arose during sample modification by IAM. When we performed reduction/alkylation of the sample according to a protocol by Shevchenko et al. [42],

388 unspecific modifications of tryptic peptides by IAM were observed suggesting an "overalkylation"
389 because of side reactions. This was probably caused by an insufficient washing of gel pieces after
390 aspirating IAM solution. IAM is a very powerful modification agent. Its reaction with Cys under
391 suitable conditions is very specific. If used in excess, IAM might become nonspecific. Then amino
392 acids undergo N- as well as S-carbamidomethylation and even O-alkylation was detected [51]. For that
393 reason we have included a repeated washing of the gel pieces with 0.1 M NH_4HCO_3 after the
394 alkylation step. As a result, O-glycopeptides became readily detectable, although each of them was
395 still split into a group of peaks, which reduced intensity of possible precursor ions for MS/MS analysis
396 (Fig. 2b). This effect was due to the differential modification by free acrylamide in the polyacrylamide
397 gels of reduced Cys residues in the HR region of IgA1 during the electrophoresis. The subsequent
398 targeted modification by IAM generated differentially modified variants of each O-glycopeptide
399 (theoretically there are 4 variants for three modified cysteines: 3xCAM, 2xCAM+1xPAM,
400 1xCAM+2PAM and 3xPAM; Fig. 2b). There were two possible ways of correcting this problem:
401 either by performing reduction/alkylation prior to SDS-PAGE or using acrylamide as a modification
402 agent after SDS-PAGE. As alkylation with acrylamide is pH and time dependent and usually proceeds
403 for several hours [52], we selected the first approach. MALDI-TOF/TOF mass spectra of HR O-
404 glycopeptides after optimization of Cys modification exhibited well-defined ion peaks (Fig. 2c). The
405 complete modification of cysteines by IAM was found critical for successful MS/MS analyses on the
406 way to deciphering the microheterogeneity of O-glycosylation in pIgA1 (Ale). Similarly, removal of
407 the excess of the reagent was equally important.

408 After optimizing the sample preparation for MALDI-TOF/TOF MS, profiling of the O-
409 glycopeptide pattern of pIgA1 (Ale) was reproducible (Fig. 3). After a pre-fractionation of tryptic
410 digest on POROS Oligo R3 sorbent and subsequent MALDI-TOF/TOF MS analysis, reading the m/z
411 values of O-glycopeptide peaks allow determination of the overall glycan composition (Table 1). As
412 some of the detected peaks displayed remarkably high intensities when compared with the others in
413 the O-glycopeptide profile, we attempted to estimate their relative abundance. The most abundant O-
414 glycoforms were with 4-4, 4-3, 5-4, 5-3, 4-2, and 3-3 GalNAc-Gal residues (given in the descending

415 order of abundance). The glycoforms with 4 attached O-glycans represented about 70% of the detected
416 O-glycopeptides followed by those with 5 occupied glycosylation sites – 24%).

417 Previously published data obtained for the same protein (although different sample preparation) by
418 use of RPLC-FT-ICR MS, showed the following descending abundance series: 4-4, 5-4, 4-3, 5-3, 5-5,
419 4-2, and HR glycoforms with 5 attached glycans represented about 40% of total HR content, whereas
420 the O-glycoforms with 4 occupied sites represented ~55% [6]. While the rank orders are similar the
421 difference in relative abundance is likely due to the different ionization techniques (MALDI vs.
422 electrospray ionization). Moreover, MALDI is not a quantitative technique, although it can be used to
423 some extent for semi-quantitative comparisons based on the relative abundance of molecules within a
424 spectrum [53]. Overall, when evaluating summarized percentage numbers from this work and the
425 study by RPLC-FT-ICR MS, they are comparable.

426 The application of a microgradient separation of the IgA1 digests using RPLC on capillary
427 columns within the off-line resulted in suitable precursor ions for PSD-MALDI-TOF/TOF MS/MS,
428 which enabled us to decipher the microheterogeneity of O-glycosylation in pIgA1 (Ale). Fig. 5 and
429 Supplementary Fig. 2 show a comprehensive picture of the results. Fifteen unique ion species
430 corresponding to HR O-glycopeptides were subjected to a fragmentation analysis to localize O-
431 glycosylation sites. Fig. 5 indicates that the major sites of microheterogeneity appear between Thr225
432 up to Thr236, which is in accordance with previous findings [6]. Ser232 is almost exclusively
433 occupied by a disaccharide as seen with a different IgA1 myeloma protein [3]. Only three O-
434 glycoforms of low abundance (1-0, 1-1, 2-2) have no sugar attached at Ser232. The second site which
435 is predominantly occupied with a disaccharide is Thr225. Nevertheless, we observed a population of
436 low abundant position isomers of O-glycoforms (1-0, 1-1, 2-2, 3-1, 3-2 and 3-3) where Thr225 is not
437 occupied. Manual inspection of the MS/MS spectra resulted in completing a list of the most obvious
438 O-glycoforms (Fig. 5). However, the complexity of results is largely increased by the existence of
439 isomeric O-glycoforms having the same level of O-glycosylation but a different occupancy at
440 particular amino-acid sites. Because of the presence of more y-series fragment ions related to a single
441 fragmentation site in the given O-glycopeptide (Fig. 4), the manual inspection becomes almost
442 impossible and we suggested using a computational approach, which proved to be helpful for getting

443 the highest possible information content from our data (Table 2 and Fig. 6). Based on this analysis, the
444 most frequent Gal-deficient sites seem those at Ser230 and Thr236 in agreement with previous reports
445 [6]. In 2010, Takahashi et al. [6] reported the existence of a mixture of isomeric and differently
446 glycosylated O-glycopeptides based on results from ECD/ETD analyses of the most abundant pIgA1
447 (Ale) O-glycoforms. However, the authors have not analyzed the variable attachment sites and did not
448 perform a quantification study, which would provide relative abundance data for the isomers. Later on,
449 Takahashi et al. [3] described a detailed investigation of isomeric O-glycoforms in pIgA (Mce1)
450 protein as well as in IgA1 from a normal human serum. The quantification of relative abundance was
451 based on chromatographic data (peak areas) and the glycan attachment sites were determined by
452 interpretation of MS/MS spectra. Nevertheless, the range of the detected isomeric O-glycoforms is
453 significantly higher in the present study, where also the O-glycoforms 4-2, 4-3, 4-4, 5-3 and 5-4 were
454 found to exist in many isomeric variants.
455 Takahashi et al. (2010) proposed that the distribution of IgA1 (Ale) myeloma protein O-
456 glycosylation is dominated by a group of IgA1 O-glycoforms that may originate from a semi-ordered
457 carbohydrate additions that proceed from Thr225 to Thr236. However, they also proposed the
458 existence of a smaller subpopulation of IgA1 O-glycoforms that does not follow this semi-ordered
459 process, giving rise to differentially O-glycosylated isomeric species [6]. These findings are in a slight
460 contrast with our data. Here the MS/MS experiments revealed that the process of O-glycosylation in
461 IgA1 HR is probably initiated not only at Thr225 but also at Thr228 in a population of positional
462 isomers. Lectin domains of different GalNAc-transferases can recognize the same substrate (GalNAc-
463 glycopeptides) and this could result in different sites or order of GalNAc attachment, catalyzed by
464 several GalNAc transferases [54]. In more recent work, the same group that suggested Thr225 has
465 performed in vitro GalNAc-transferase reactions that suggests that Thr228 is the more likely
466 originating site of glycosylation (unpublished data). This matches an alternative explanation by
467 Takahashi et al. (2010) for the existence of O-glycoforms that does not follow the semi-ordered
468 mechanism of carbohydrate addition due to a low-level natural occurrence of Thr225 being “skipped”
469 by the GalNAc-transferase [6]. This hypothesis seems to be more appropriate in the light of our
470 results. There is a population of positional isomers, which probably follows the semi-ordered

mechanism proceeding from Thr228. Our current results would suggest that Thr225 is not initially glycosylated by GalNAc-transferase but is as low as fourth in rank order out of the six confirmed glycosylation sites (Thr225 is not occupied in some positional isomers of the O-glycoforms 1-0, 1-1, 2-2, 3-1, 3-2 and 3-3; Fig. 5, Supplementary Fig. 3).

Similar to previous results [6,7,35], the distribution of HR glycoforms for pIgA1 (Ale) focuses around O-glycopeptides comprising $GalNAc_4Gal_3$ (4-3), $GalNAc_4Gal_4$ (4-4) and $GalNAc_5Gal_4$ (5-4). We hypothesize that the species 4-4 represents a product of the addition of a Gal residue to the GalNAc monosaccharide at Ser230 of the O-glycoform 4-3. Alternatively, but to a lesser extent, GalNAc-transferase can add a single GalNAc monosaccharide to Ser236 giving rise to the less abundant O-glycoform 5-3. The O-glycoform 4-3 may arise from the addition of a Gal residue to at least two different positional isomers of the O-glycoform 4-2 (Figs. 5 and 6). The structure of the O-glycoform 3-3 might indicate that it is produced by a minor mechanism of galactosylation. The other mechanism would prefer adding Gal to GalNAc at Thr228 and Ser230 after adding GalNAc at Thr225. The GalNAc at Thr225 is then galactosylated to provide the O-glycoform 4-2, which could be the major substrate for GalNAc-tranferase in the production of the most abundant O-glycoform with two Gal-deficient sites (5-3). These similar findings in this study using a different IgA1 myeloma protein preparation and different MS platform provide strong corroborating evidence for the proposed semi-ordered addition of Gal for these dominant O-glycoforms of IgA1 [6].

5. Conclusions

In this work, we have introduced a modified approach for analyzing the microheterogeneity of O-glycosylation in human IgA1 at the level of glycosylation sites and O-glycan composition. The sample preparation procedure used involved in-gel digestion of a polymeric IgA1 model and relied on a key step of optimized cysteine alkylation prior to SDS-PAGE, which allowed to avoid future signal splitting of O-glycopeptides in MALDI-TOF/TOF MS. Mass spectrometric profiling of O-glycopeptides (after separation of the digests by RPLC on capillary columns using a microgradient device) was completed by MS/MS. Based on the acquired MS/MS spectra and their interpretation, which utilized a manual reading as well as computational processing to decipher the fragmentation patterns, we may conclude that the most abundant O-glycoforms in the model protein are those containing GalNAc-Gal levels of 4-4, 4-3, 5-4, 5-3, 4-2 and 3-3, which exist in isomeric variants with differentially occupied O-glycosylation sites. The most frequent Gal-deficient sites were localized at Ser230 and Thr236.

We show that our simplified approach is efficient and has a high potential to provide a method for the rapid assessment of IgA1 heterogeneity that is a less expensive and yet corroborating alternative to LC-(high-resolution)-MS protocols. The novelty and biological significance resides in demonstration, for the first time, of the distribution of the most abundant isoforms of HR O-glycopeptides of IgA1. Our data also allow a direct comparison of MALDI MS and FT MS results [3,6] and evaluation of advantages of individual approaches. Notably, the present procedure is based on a tryptic peptide, which may be difficult to fragment by ECD/ETD. As another new feature, we introduce a software solution for interpretation of MS/MS data of O-glycopeptide isoforms, which provides the possibility of fast and easier data processing.

Acknowledgments

This work was supported by OP RD&I grant no. ED0007/01/01 (Centre of the Region Haná for Biotechnological and Agricultural Research) and OP EC grant CZ.1.07/2.4.00/31.0130 (ChemPharmNet) from the Ministry of Education Youth and Sports, Czech Republic, plus a long-term organization development plan no. 1011 from the Faculty of Military Health Sciences, University of

Defence, CZ. JN and MBR were supported in part by grants from the National Institutes of Health DK082753, DK078244, DK083663, and GM098539. The authors appreciate technical assistance of Mrs. Stacy Hall and Rhubell Brown with the isolation of pIgA1 (Ale).

REFERENCES

[1] Mestecky J. Immunobiology of IgA. Am J Kidney Dis 1988; 12:378-83.

[2] Frangione B, Wolfenstein-Todel C. Partial duplication in the "hinge" region of IgA_1 myeloma proteins. Proc Natl Acad Sci USA 1972; 69:3673-6.

[3] Takahashi K, Smith AD, Poulsen K, Kilian M, Julian BA, Mestecky J, et al. Naturally occurring structural isomers in serum IgA1 O-glycosylation. J Prot Res 2012; 11:692-702.

[4] Tarelli E, Smith AC, Hendry BM, Challacombe SJ, Pouria S. Human serum IgA1 is substituted with up to six O-glycans as shown by matrix assisted laser desorption ionisation time-of-flight mass spectrometry. Carbohydrate Res 2004; 339:2329-35.

[5] Baenziger B, Kornfeld S. Structure of the carbohydrate units of IgA1 immunoglobulin. II. Structure of the O-glycosidically linked oligosaccharide units. J Biol Chem 1974; 249:7270-81.

[6] Takahashi K, Wall SB, Suzuki H, Smith AD, Hall S, Poulsen K, et al. Clustered O-glycans of IgA1: defining macro- and microheterogeneity by use of electron capture/transfer dissociation. Mol Cel Proteomics 2010; 9:2545-57.

[7] Novak J, Tomana M, Kilian M, Coward L, Kulhavy R, Barnes S, et al. Heterogeneity of O-glycosylation in the hinge region of human IgA1. Mol Immunol 2000; 37:1047-56.

[8] Field MC, Dwek RA, Edge CJ, Rademacher TW. O-linked oligosaccharides from human serum immunoglobulin A1. Biochem Soc Trans 1989; 17:1034-5.

[9] Mattu TS, Pleass RJ, Willis AC, Kilian M, Wormald MR, Lellouch AC, et al. The glycosylation and structure of human serum IgA1, Fab, and Fc regions and the role of N-glycosylation on Fcα receptor interactions. J Biol Chem 1998; 273:2260-72.

[10] Mestecky J, Tomana M, Crowley-Nowick PA, Moldoveanu Z, Julian BA, Jackson S. Defective galactosylation and clearance of IgA1 molecules as a possible etiopathogenic factor in IgA nephropathy. Contrib Nephrol 1993; 104:172-82.

[11] Moore JS, Wu X, Kulhavy R, Tomana M, Novak J, Moldoveanu Z, et al. Increased levels of galactose-deficient IgG in sera of HIV-1-infected individuals. AIDS 2005; 19:381-9.

[12] Rademacher TW, Williams P, Dwek RA. Agalactosyl glycoforms of IgG autoantibodies are pathogenic. Proc Natl Acad Sci USA 1994; 91:6123-7.

[13] Springer GF. Immunoreactive T and Tn epitopes in cancer diagnosis, prognosis, and immunotherapy. J Mol Med 1997; 75:594-602.

[14] Troelsen LN, Garred P, Madsen HO, Jacobsen S. Genetically determined high serum levels of mannose-binding lectin and agalactosyl IgG are associated with ischemic heart disease in rheumatoid arthritis. Arthritis Rheum 2007; 56:21-9.

[15] Stuchlová Horynová M, Raška M, Clausen H, Novak J. Aberrant O-glycosylation and anti-glycan antibodies in an autoimmune disease IgA nephropathy and breast adenocarcinoma. Cell Mol Life Sci 2013; 70:829-39.

574 [16] Berger J, Hinglais N. Intercapillary deposits of IgA-IgG. J Urol Nephrol (Paris). 1968; 74:694-5.
575
576 [17] Gharavi AG, Yan Y, Scolari F, Schena FP, Frasca GM, Ghiggeri GM, et al. IgA nephropathy, the
577 most common cause of glomerulonephritis, is linked to 6q22-23. Nat Genetics 2000; 26:354-7.
578
579 [18] Julian BA, Waldo FB, Rifai A, Mestecky J. IgA nephropathy, the most common
580 glomerulonephritis worldwide: a neglected disease in the United States?. Am J Med 1988; 84:129-32.
581
582 [19] D'Amico G. The commonest glomerulonephritis in the world: IgA nephropathy. Quart J Med
583 1987; 64:709-27.
584
585 [20] Novak J, Moldoveanu Z, Renfrow MB, Yanagihara T, Suzuki H, Raska M, et al. IgA
586 nephropathy and Henoch-Schoenlein purpura nephritis: aberrant glycosylation of IgA1, formation of
587 IgA1-containing immune complexes, and activation of mesangial cells. Contrib Nephrol 2007;
588 157:134-8.
589
590 [21] Tomino Y, Suzuki S, Imai H, Saito T, Kawamura T, Yorioka N, et al. Measurement of serum IgA
591 and C3 may predict the diagnosis of patients with IgA nephropathy prior to renal biopsy. J Clin Lab
592 Anal 2000; 14:220-3.
593
594 [22] Novak J, Julian BA, Mestecky J, Renfrow MB. Glycosylation of IgA1 and pathogenesis of IgA
595 nephropathy. Semin Imunopathol 2012; 34:365-82.
596
597 [23] Suzuki H, Kiryluk K, Novak J, Moldoveanu Z, Herr AB, Renfrow MB, et al. The
598 pathophysiology of IgA nephropathy. J Am Soc Nephrol 2011; 22:1795-803.
599
600 [24] Novak J, Julian BA, Tomana M, Mestecky J. IgA glycosylation and IgA immune complexes in
601 the pathogenesis of IgA nephropathy. Semin Nephrol 2008; 28:78-87.

602

603 [25] Renfrow MB, Mackay CL, Chalmers MJ, Julian BA, Mestecky J, Kilian M, et al. Analysis of O-
604 glycan heterogeneity in IgA1 myeloma proteins by Fourier transform ion cyclotron resonance mass
605 spectrometry: implications for IgA nephropathy. Anal Bioanal Chem 2007; 389:1397-407.

606

607 [26] Allen AC Methodological approaches to the analysis of IgA1 O-glycosylation in IgA
608 nephropathy. J Nephrol 1999; 12:76-84.

609

610 [27] Wada Y, Azadi P, Costello CE, Dell A, Dwek RA, Geyer H, et al. Comparison of the methods for
611 profiling glycoprotein glycans - HUPO Human Disease Glycomics/Proteome Initiative multi-
612 institutional study. Glycobiology 2007; 17:411-22.

613

614 [28] Wada Y, Dell A, Haslam SM, Tissot B, Canis K, Azadi P, et al. Comparison of methods for
615 profiling O-glycosylation: Human Proteome Organisation Human Disease Glycomics/Proteome
616 Initiative multi-institutional study of IgA1. Mol Cell Proteomics 2010; 9:719-27.

617

618 [29] Renfrow MB, Cooper HJ, Tomana M, Kulhavy R, Hiki Y, Toma K, et al. Determination of
619 aberrant O-glycosylation in the IgA1 hinge region by electron capture dissociation Fourier transform-
620 ion cyclotron resonance mass spectrometry. J Biol Chem 2005; 280:19136-45.

621

622 [30] Mirgorodskaya E, P. Roepstorff, R. Zubarev RA. Localization of O-glycosylation sites in
623 peptides by electron capture dissociation in a Fourier transform mass spectrometer. Anal Chem 1999;
624 71:4431-6.

625

626 [31] Mormann M, Paulsen H, Peter-Katalinic J. Electron capture dissociation of O-glycosylated
627 peptides: radical site-induced fragmentation of glycosidic bonds. Eur J Mass Spectrom (Chichester,
628 Eng) 2005; 11:497–511.

629

[32] Tomana M, Novak J, Julian BA, Matousovic K, Konecny K, Mestecky J. Circulating immune complexes in IgA nephropathy consist of IgA1 with galactose-deficient hinge region and antiglycan antibodies. J Clin Inves 1999; 104:73-81.

[33] Novak J, Julian BA, Tomana M, Mestecky J. Progress in molecular and genetic studies of IgA nephropathy. J Clin Immunol 2001; 21:310-27.

[34] Novak J, Vu HL, Novak L, Julian BA, Mestecky J, Tomana M. Interactions of human mesangial cells with IgA and IgA-containing immune complexes. Kidney Int 2002; 62:465-75.

[35] Hiki Y, Odani H, Takahashi M, Yasuda Y, Nishimoto A, Iwase H, et al. Mass spectrometry proves under-O-glycosylation of glomerular IgA1 in IgA nephropathy. Kidney Int 2001; 59:1077–85.

[36] Hiki Y, Tanaka A, Kokubo T, Iwase H, Nishikido J, Hotta K, et al. Analyses of IgA1 hinge glycopeptides in IgA nephropathy by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. J Am Soc Nephrol 1998; 9:577–82.

[37] Kussmann M, Nordhoff E, Rahbek-Nielsen H, Haebel S, Rossel-Larsen M, Jakobsen L, et al. Matrix-assisted laser desorption/ionization mass spectrometry sample preparation techniques designed for various peptide and protein analytes. J Mass Spectrom 1997; 32:593–601.

[38] Ishihama Y, Rappsilber J, Mann M. Modular stop and go extraction tips with stacked disks for parallel and multidimensional peptide fractionation in proteomics. J Proteome Res 2006; 5:988-94.

[39] Rappsilber J, Mann M, Ishihama Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc 2007; 2:1896–1906.

[40] Moravcová D, Kahle V, Řehulková H, Chmelík J, Řehulka P. Short monolithic columns for purification and fractionation of peptide samples for matrix-assisted laser desorption/ionization time-of-flight/time-of-flight mass spectrometry analysis in proteomics. J Chromatogr A 2009; 1216:3629–36.

[41] Laemmli UK. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 1970; 227:680-5.

[42] Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protoc 2007; 1:2856-60.

[43] Franc V, Šebela M, Řehulka P, Končitíková R, Lenobel R, Madzak C, et al. Analysis of N-glycosylation in maize cytokinin oxidase/dehydrogenase 1 using a manual microgradient chromatographic separation coupled offline to MALDI-TOF/TOF mass spectrometry. J Proteomics 2012; 75:4027-37.

[44] Kahle V, Vázlerová M, Welsch T. Automated microgradient system for capillary electrochromatography. J Chromatogr A 2003; 990:3–9.

[45] Spengler B. Post-source decay analysis in matrix-assisted laser desorption/ionization mass spectrometry of biomolecules. J Mass Spectrom 1997; 32:1019-36.

[46] Müller S, Goletz S, Packer N, Gooley A, Lawson A, Hanisch FG. Localization of O-glycosylation sites on glycopeptide fragments from lactation-associated MUC1: all putative sites within the tandem repeat are glycosylation targets in vivo. J Biol Chem 1997; 272:24780-24793.

[47] Apweiler R, Hermjakob H, Sharon N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochim Biophys Acta 2007; 1473:4–8.

[48] Suzuki H, Fan R, Zhang Z, Brown R, Hall S, Julian BA, et al. Aberrantly glycosylated IgA1 in IgA nephropathy patients is recognized by IgG antibodies with restricted heterogeneity. J Clin Inves 2009; 119:1668-77.

[49] Herr AB, Ballister ER, Bjorkman PJ. Insights into IgA-mediated immune responses from the crystal structures of human FcαRI and its complex with IgA1-Fc. Nature 2003; 423:614-20.

[50] Barratt J, Smith AC, Feehally J. The pathogenic role of IgA1 O-linked glycosylation in the pathogenesis of IgA1 nephropathy. Nephrology (Carlton) 2007, 12:275-84.

[51] Boja ES, Fales HM. Overalkylation of a protein digest with iodoacetamide. Anal Chem 2001; 73:3576-82.

[52] Luche S, Diemer H, Tastet C, Chevallet M, Van Dorsselaer A, Leize-Wagner E, et al. About thiol derivatization and resolution of basic proteins in two-dimensional electrophoresis. Proteomics 2004; 4:551-61.

[53] Alexandrov T. MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. BMC Bioinformatics 2012; 13(Suppl 16):S11

[54] Wandall HH, Irazoqui F, Tarp MA, Bennett EP, Mandel U, Takeuchi H, et al. The lectin domains of polypeptide GalNAc-transferases exhibit carbohydrate-binding specificity for GalNAc: lectin binding to GalNAc-glycopeptide substrates is required for high density GalNAc-O-glycosylation. Glycobiology 2007; 17:374-87.

FIGURE LEGENDS

Figure 1

Structure of human IgA1

(a) Localization of O-glycosylation in the hinge region and N-glycosylation in the C-terminal constant region of the heavy chain. (b) Variants of O-glycans occurring in the hinge region of human circulatory IgA1. The circle highlights galactose-deficient O-glycans. The drawing was adopted from previous publications [3,6].

Figure 2

The influence of sample alkylation on the intensity of O-glycopeptide peaks in mass spectra

a) a MALDI-TOF mass spectrum (m/z region of 5000 – 6200) of a tryptic digest of pIgA1 (Ale) heavy chain (the reduction/alkylation step was performed after SDS-PAGE), which was acquired without any prior desalting of the sample; b) a MALDI-TOF mass spectrum (m/z region of 5000 – 6200) of the same tryptic digest (the reduction/alkylation step was performed after SDS-PAGE) acquired after involving the desalting procedure on POROS Oligo R3; c) a MALDI-TOF mass spectrum (m/z region of 5000 – 6200) of the same tryptic digest (the reduction/alkylation step was performed both prior to and after SDS-PAGE) acquired after involving the desalting procedure on POROS Oligo R3. Besides the necessity of desalting, a complete alkylation of cysteine residues is crucial for obtaining the highest possible intensity of O-glycopeptide peaks. All MALDI-TOF/TOF mass spectra were acquired on 4800 Proteomics Analyzer operating in the reflectron mode for positive ions. CHCA was used as a matrix. The abbreviations CAM and PAM stand for carbamidomethylation and propionamidation of cysteine, respectively.

Figure 3

MALDI-TOF/TOF MS profile of O-glycopeptides from tryptic digest of pIgA1 (Ale)

A mass spectrum of O-glycopeptides is shown, which was acquired after pre-fractionation of pIgA1 heavy-chain tryptic digest on a GELoader tip filled with POROS Oligo R3 50-µm particles. The inset demonstrates the resolution level, which was achieved during MS measurements (for m/z 5435, a resolution value of 9100 was observed). The symbol “P” represents the HR tryptic peptide.

Figure 4

Positive ion MALDI-TOF/TOF MS/MS spectra of pIgA1 (Ale) HR O-glycopeptides

Three selected MS/MS spectra are shown, which were acquired for tryptic O-glycopeptides with precursor m/z of 5638.7 [HR+ 5 GalNAc + 3 Gal + H]$^+$ (a), 5435.6 [HR + 4 GalNAc + 3 Gal + H]$^+$ (b)

and 5273.5 [HR + 4 GalNAc + 2 Gal + H]$^+$ (c). In several instances, ambiguous fragments suggested the presences of isomeric IgA1 O-glycoforms. The measurements were carried out on the ABI 4800 Proteomics Analyzer operating in the reflectron mode for positive ions; CHCA was used as a matrix. Note that the signal intensity of the precursor ion [M+H]$^+$ was reduced in the spectra as the function "Metastable Suppresor" was active on the instrument. This means that after PSD fragmentation of a precursor ion in the collision cell and subsequent reacceleration of all ions (mostly non-fragmented), the remaining precursor ion becomes deflected from the second TOF analyzer to prevent both detector saturation and metastable fragmentation of the precursor ion prior to the reflectron.

Figure 5

O-glycoforms of pIgA1 (Ale) that were deduced from manual inspection of MS/MS spectra of tryptic O-glycopeptides

The HR tryptic peptide 208-HYTNPSQDVTVPCPVPSTPPTPSPSTPPTPSPSCCHPR-245 carrying O-glycosylation is shown only in part (underlined) for clarity. Three of the 15 IgA1 O-glycoforms that underwent PSD-MALDI TOF/TOF MS/MS fragmentation are shown in Fig. 4. The MS/MS data for remaining 12 glycoforms are provided in Supplementary Fig. 3.

Figure 6

Some of the isomeric O-glycoforms of pIgA1 (Ale) identified by computational analysis of MS/MS spectra of tryptic O-glycopeptides using BIOSPEAN software.

Experimental MS/MS spectra were individually compared with the set of virtual spectra using the software BIOSPEAN (see Materials and methods). The results are based on the number of overlapping peaks and allowed identification of many potential isomeric O-glycoforms based on peaks identified in the PSD-MALDI TOF/TOF MS/MS spectra for each predominant precursor ion species (see Table 2). Isomers of highest confidence are highlighted by frames.

Tables 1 and 2
Click here to download Table: JProt_Franc_et_al_IgA_revision_tables.doc

Table 1

O-glycopeptides deduced from MALDI-TOF/TOF mass spectra of pIgA1 digest

All O-glycopeptides were deduced based on mass calculations considering the presence of the HR tryptic peptide 208-HYTNPSQDVTVPCPVPSTPPTPSPSTPPTPSPSCCHPR-245.

Glycan structure[a]	Experimental mass[b] (Da)	Calculated mass[b] (Da)	Mass difference (Da)	Relative abundance[c] (%) ± S.D.
$GalNAc_6Gal_5$	6165.94	6165.63	-0.31	-
$GalNAc_6Gal_4$	6003.84	6003.58	-0.26	-
$GalNAc_5Gal_5$	5962.84	5962.55	-0.29	-
$GalNAc_5Gal_4$	5800.85	5800.50	-0.35	13.92 ± 2.87
$GalNAc_5Gal_3$	5638.68	5638.45	-0.23	10.16 ± 1.57
$GalNAc_4Gal_4$	5597.70	5597.42	-0.28	35.50 ± 3.83
$GalNAc_4Gal_3$	5435.60	5435.37	-0.23	32.21 ± 3.30
$GalNAc_4Gal_2$	5273.51	5273.31	-0.20	4.90 ± 2.94
$GalNAc_4Gal_1$	5111.45	5111.26	-0.19	-
$(GalNAc_4)$[b]	4950.54	4949.21	-1.33	-
$GalNAc_3Gal_3$	5232.50	5232.29	-0.21	3.27 ± 2.86
$GalNAc_3Gal_2$	5070.40	5070.23	-0.17	-
$GalNAc_3Gal_1$	4908.32	4908.18	-0.14	-
$(GalNAc_3)$	4746.37	4746.13	-0.24	-
$GalNAc_2Gal_2$	4867.30	4867.15	-0.15	-
$(GalNAc_2Gal_1)$	4705.35	4705.10	-0.25	-
$(GalNAc_2)$	4543.26	4543.05	-0.21	-
GalNAcGal	4502.29	4502.02	-0.27	-
$(GalNAc_1)$	4340.18	4339.97	-0.21	-
Naked peptide	4137.06	4136.89	-0.17	-

[a]Parentheses indicate those O-glycopeptides, where high-quality MS/MS spectra could not be acquired

[b]Monoisotopic mass, $[M+H]^+$

[c]Estimated from the isotope cluster area. The total area of the most abundant O-glycopeptide peaks was set at 100%.

Table 2

An excerpt from software calculation results for assigning MS/MS spectra to O-glycoforms

Peak lists from individual experimental spectra were searched against a database of virtual spectra generated by the program BIOSPEAN using theoretical masses of y-series fragments reflecting particular O-glycoforms. Numbers in the first column refer to the virtual spectra and indicate the presence of GalNAc-Gal (2), GalNAc (1) and no sugar (0) at the following glycosylation sites: Thr225, Thr228, Ser230, Ser232, Thr233 and Thr236 (in the order given); the total numbers of GalNAc and Gal residues for each virtual spectrum are provided in the second and third column, respectively. Experimental spectra are labeled with GalNAc-Gal codes (i.e., 3-1, 3-2 etc.) reflecting precursor-ion masses. Score values reflect the number of identical peak positions a color coding differentiates similarity levels. Green color represents the highest agreement, followed by yellow, orange, blue and white (in a descending way).

Virtual specra			**Score values for individual experimental spectra**											
Isoform code	**GalNAc**	**Gal**	**3-1**	**3-2**	**3-3**	**4-1**	**4-2**	**4-3**	**4-4**	**5-3**	**5-4**	**5-5**	**6-4**	**6-5**
2.1.1.2.0.0	**4**	**2**	42.86	71.43	57.14	51.43	100	82.86	65.71	42.86	28.57	28.57	28.57	0
2.1.1.2.0.2	**5**	**3**	14.29	14.29	14.29	14.29	14.29	14.29	14.29	40	65.71	57.14	28.57	14.29
2.1.1.2.2.0	**5**	**3**	28.57	28.57	28.57	28.57	28.57	28.57	28.57	54.29	65.71	57.14	28.57	0
2.1.1.2.2.2	**6**	**4**	14.29	22.86	14.29	20	14.29	14.29	14.29	14.29	28.57	28.57	42.86	14.29
2.1.2.0.0.0	**3**	**2**	28.57	62.86	42.86	42.86	42.86	54.29	42.86	28.57	42.86	28.57	28.57	0
2.1.2.0.1.0	**4**	**2**	42.86	42.86	42.86	37.14	100	68.57	51.43	42.86	42.86	28.57	28.57	0
2.1.2.0.1.2	**5**	**3**	14.29	14.29	14.29	14.29	14.29	28.57	14.29	54.29	68.57	28.57	28.57	14.29
2.1.2.0.2.0	**4**	**3**	28.57	57.14	57.14	28.57	42.86	82.86	68.57	42.86	42.86	42.86	28.57	0
2.1.2.0.2.1	**5**	**3**	14.29	14.29	14.29	14.29	14.29	28.57	14.29	68.57	68.57	28.57	28.57	14.29
2.1.2.0.2.2	**5**	**4**	14.29	14.29	17.14	14.29	14.29	14.29	14.29	14.29	68.57	57.14	28.57	14.29
2.1.2.1.0.0	**4**	**2**	42.86	57.14	42.86	51.43	100	68.57	51.43	28.57	28.57	28.57	28.57	0
2.1.2.1.0.2	**5**	**3**	14.29	14.29	14.29	14.29	14.29	28.57	14.29	54.29	68.57	42.86	28.57	14.29
2.1.2.1.2.0	**5**	**3**	28.57	28.57	28.57	28.57	28.57	42.86	28.57	68.57	68.57	42.86	28.57	0
2.1.2.1.2.2	**6**	**4**	14.29	22.86	14.29	14.29	14.29	14.29	14.29	14.29	28.57	28.57	42.86	14.29
2.1.2.2.0.0	**4**	**3**	28.57	71.43	71.43	42.86	57.14	97.14	82.86	42.86	42.86	28.57	28.57	0
2.2.1.2.0.1	**5**	**3**	14.29	14.29	28.57	14.29	28.57	42.86	28.57	97.14	68.57	28.57	28.57	14.29
2.2.1.2.0.2	**5**	**4**	14.29	14.29	17.14	14.29	14.29	14.29	14.29	14.29	82.86	57.14	28.57	14.29
2.2.1.2.1.0	**5**	**3**	28.57	28.57	42.86	28.57	42.86	57.14	42.86	97.14	68.57	28.57	28.57	0

Figure 1
Click here to download Figure: JProt_Franc_et_al_IgA_Fig_1.ppt

Franc, V. et al., Fig. 1

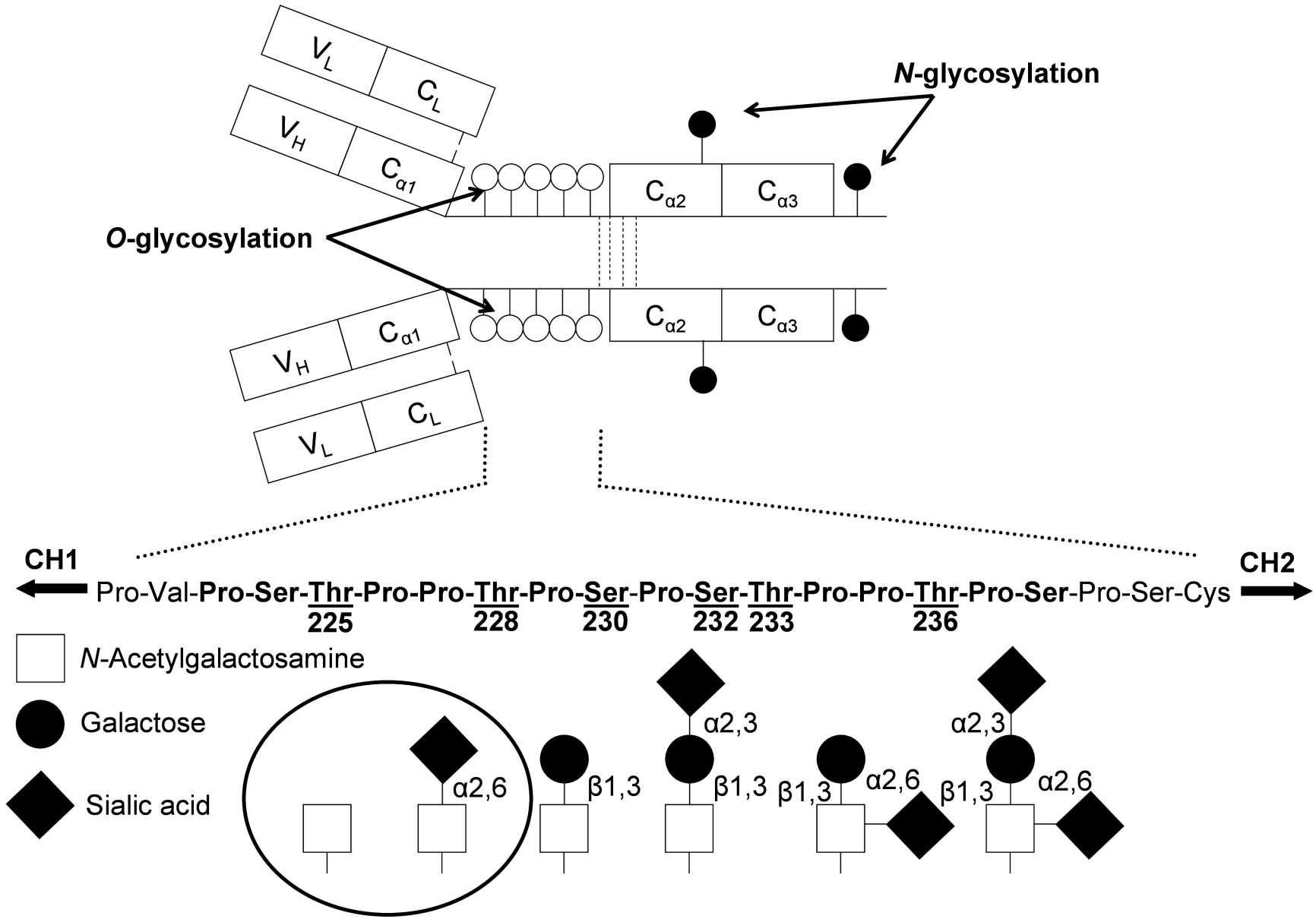


Figure 2
Click here to download Figure: JProt_Franc_et_al_IgA_Fig_2.ppt

Franc, V. et al., Fig. 2

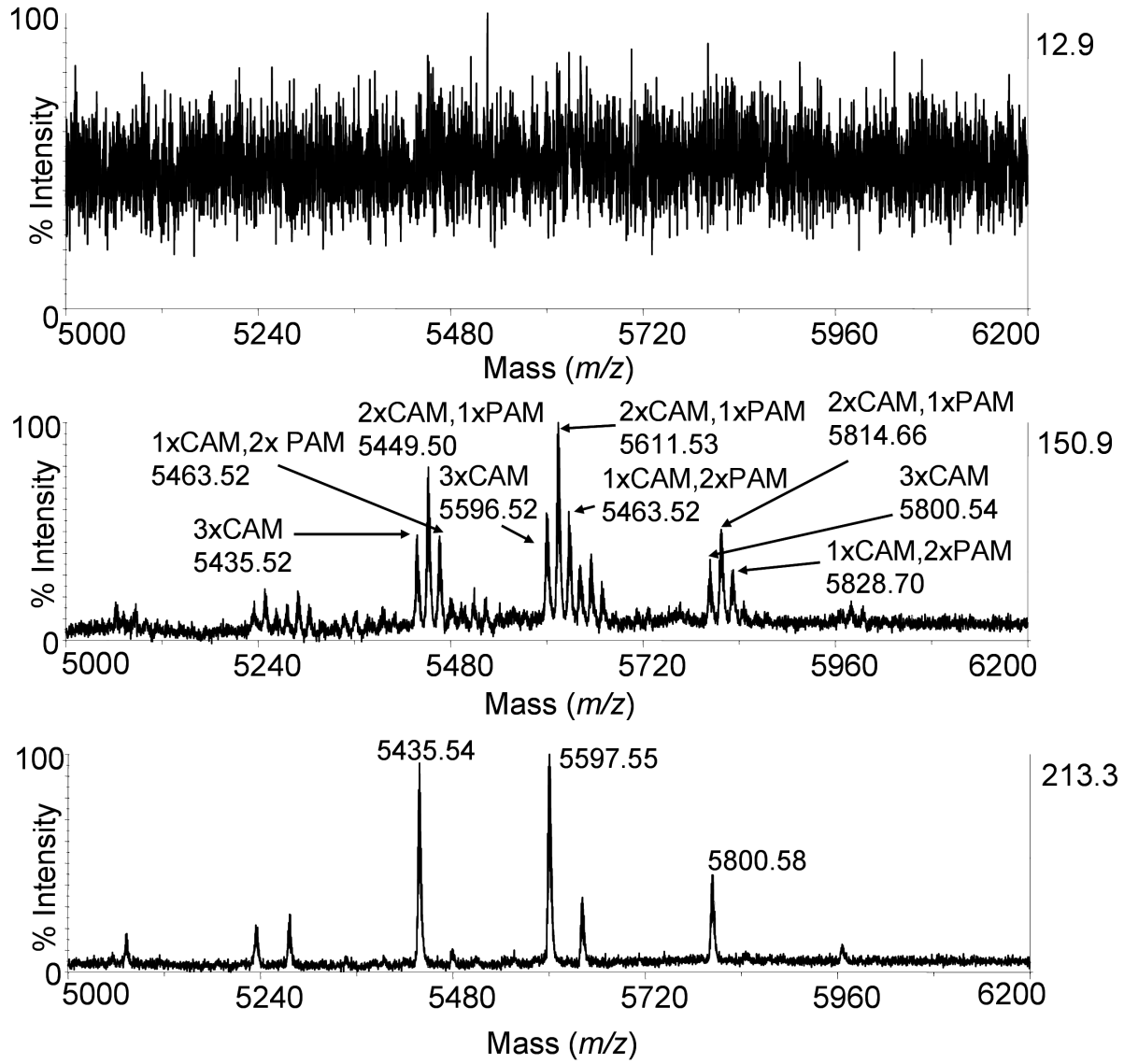


Figure 3
Click here to download Figure: JProt_Franc_et_al_IgA_Fig_3.ppt

Franc, V. et al., Fig. 3

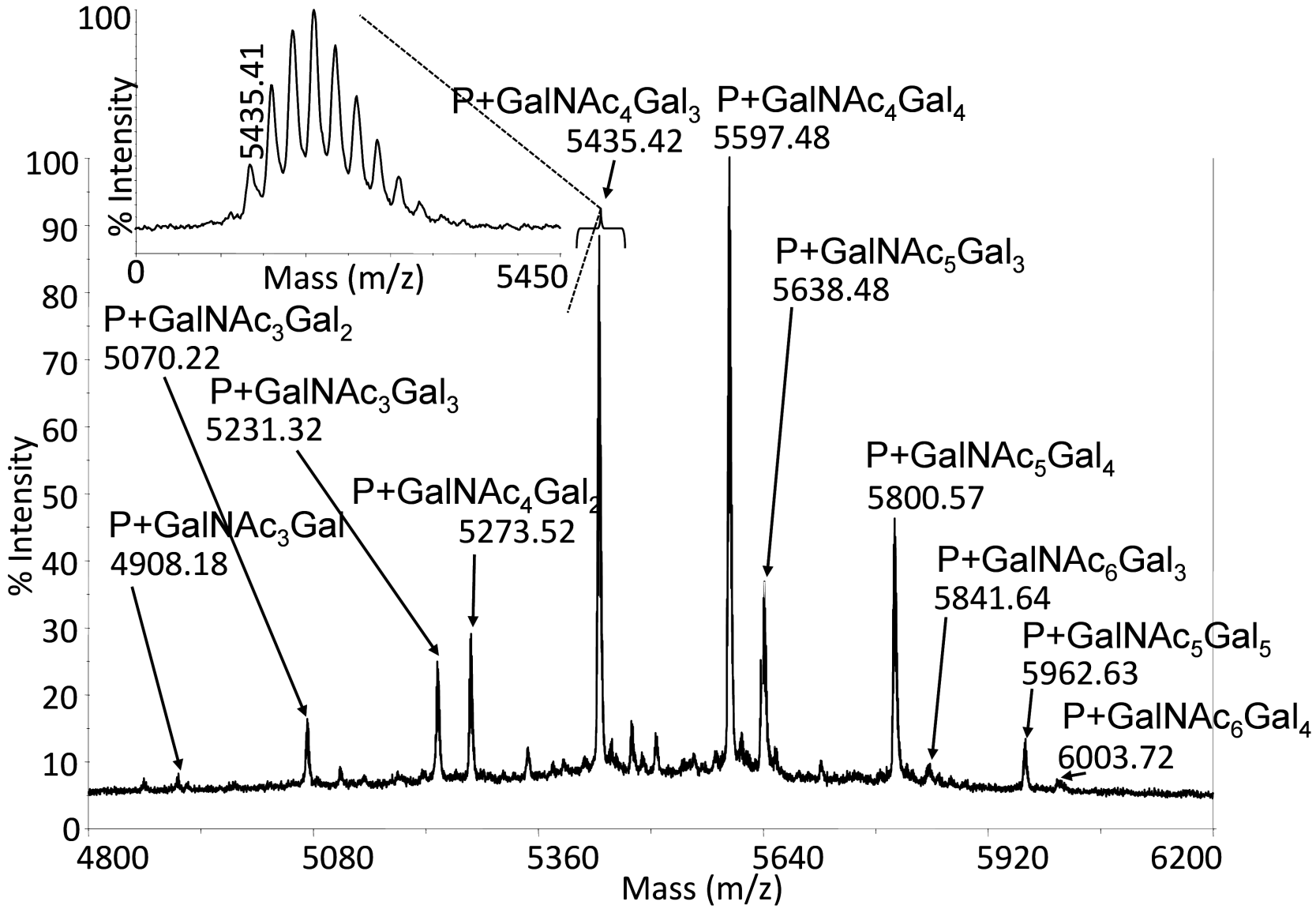


Figure 4

Click here to download Figure: JProt_Franc_et_al_IgA_Fig_4.ppt

Franc, V. et al., Fig. 4

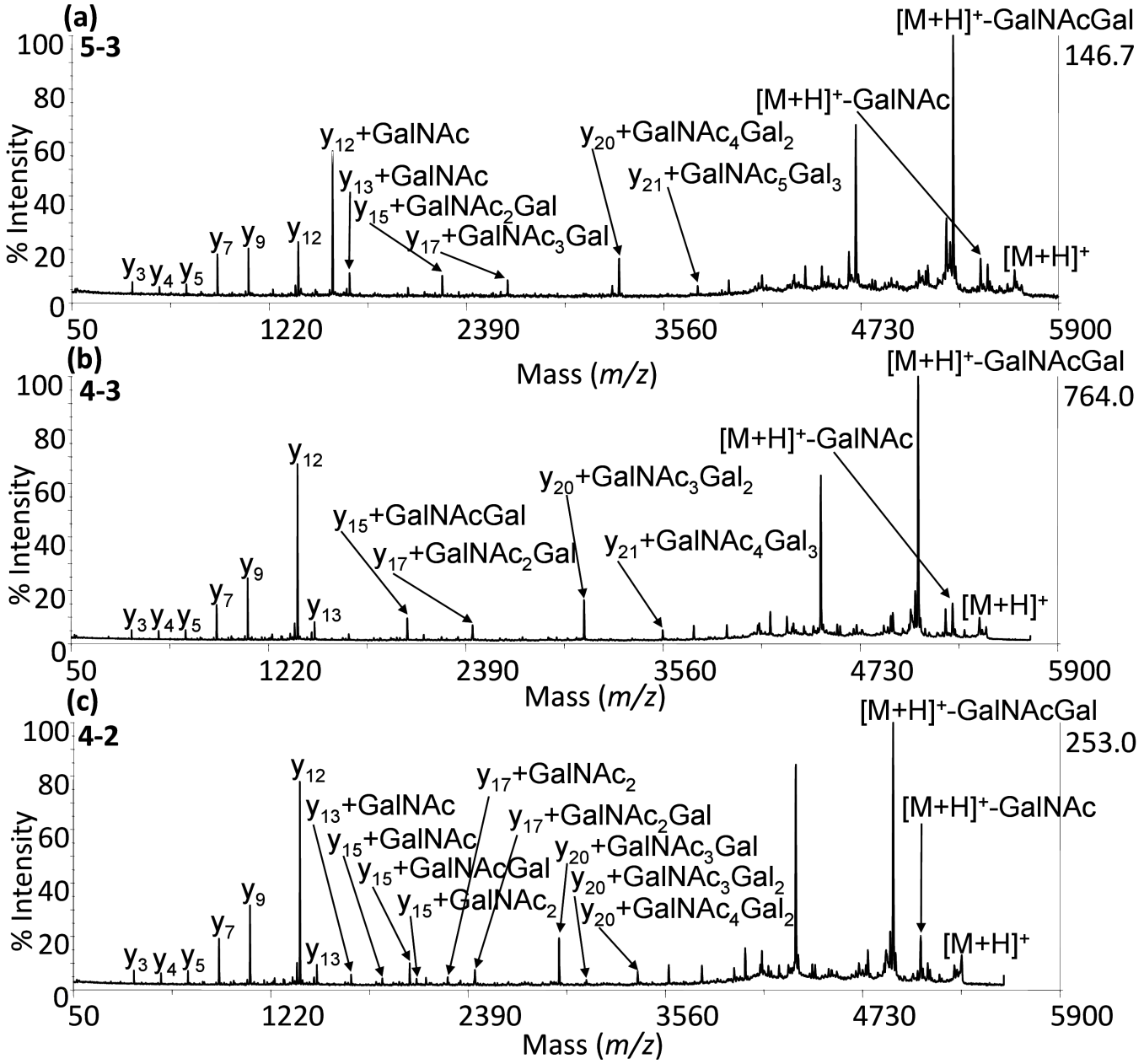


Figure 5

Click here to download Figure: JProt_Franc_et_al_IgA_Fig_5.ppt

Franc, V. et al., Fig. 5

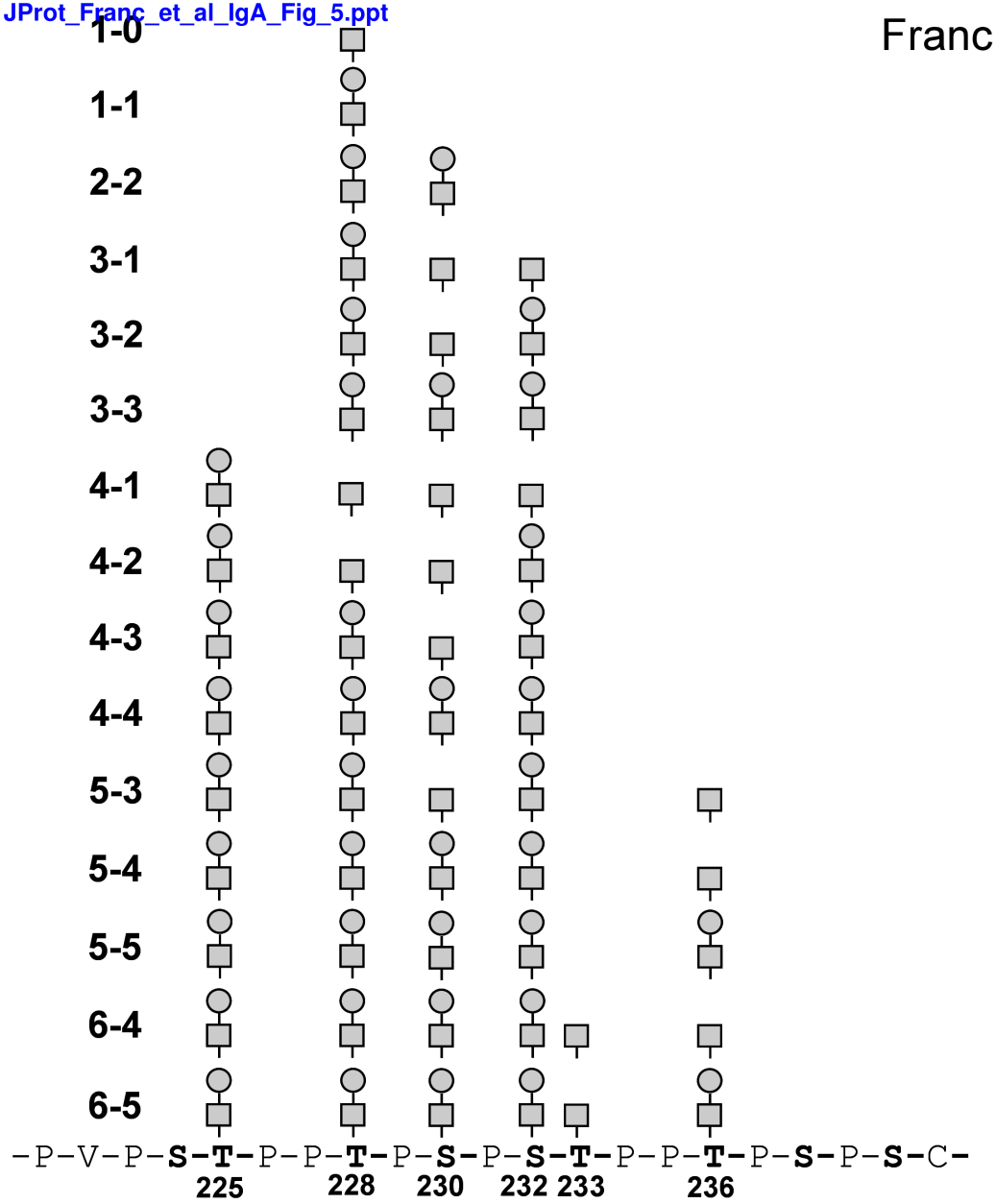
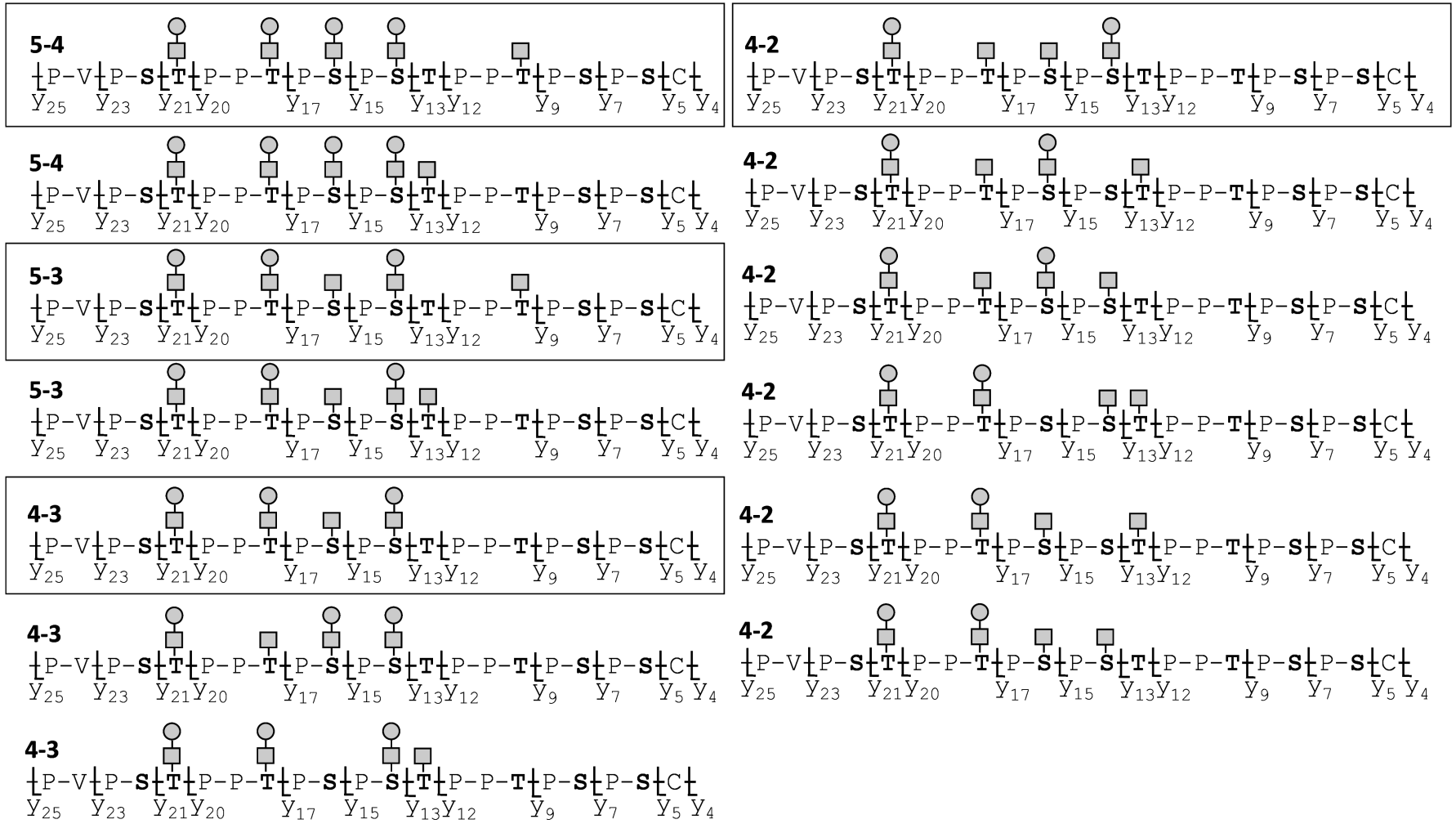


Figure 6
Click here to download Figure: JProt_Franc_et_al_IgA_Fig_6.ppt

Franc, V. et al., Fig. 6



Příloha F – referát ze stáže

S ohledem na povinnost absolvovat v rámci doktorského studia zahraniční vědeckovýzkumnou stáž jsem v období 19. listopadu 2012 – 19. prosince 2012 absolvoval stáž na Fakultě technické chemie Technologické univerzity Vídeň v Rakousku (Fakultät für Technische Chemie, Technische Universität Wien).

Protože tato stáž byla financována z projektu BIOTREND, je povinností přiložit referát ze stáže k této disertační práci.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Inovace ve vzdělávání v chemii a biologii s ohledem
na aktuální trendy v biomedicinálním výzkumu
reg. číslo: CZ.1.07/2.2.00/28.0184

REFERÁT ZE STÁŽE na Technologické univerzitě Vídeň, Rakousko

V rámci projektu BIOTREND (CZ.1.07/2.2.00/28.0184) jsem vycestoval na zahraniční stáž na pracoviště Fakultät für Technische Chemie, Technische Universität Wien (Fakulta technické chemie, Technologická univerzita Vídeň, Rakousko). Stáž proběhla ve dnech 19. 11. 2012 až 19. 12. 2012.

V původním plánu bylo jednak vylepšování, ladění a testování softwaru *Biospean* a jednak splnění úkolů o které měla zájem TU Vídeň. Nicméně vývoj situace na místě tyto úkoly postupně rozšířil a paralelní stáž dalších dvou kolegů z KBC UPOL spolu se spoluprací na dálku s pracovištěm v Olomouci přidaly další vědecké úkoly, které byly během stáže řešeny.

Vývoj softwaru *Biospean*

Tento úkol byl plněn jako součást práce na doktorandském studiu.

Biospean je software pro katalogizaci a analýzu hmotnostních spekter. Zaměřením a schopnostmi srovnání spekter patří do stejné kategorie jako např. *MALDI Biotyper*. Na rozdíl od něj se ale jedná o aplikaci webovou, multiplatformní a víceuživatelskou, určenou pro použití v rámci většího vědeckého týmu a umožňující skupinovou práci. Je provozována na platformě LAMP (Linux, Apache, MySQL, PHP).

V rámci pobytu na TU Vídeň byly v aplikaci provedeny následující změny:

1) S ohledem na specifika hmotnostních spekter používaných na TU Vídeň byla výrazně zvýšena citlivost aplikace při detekci méně výrazných peaků a schopnost se vypořádat se šumem.
2) Zvýšila se schopnost zpracovat data ze vstupních souborů obsahujících předem normalizované výsledky, které se používají na TU Vídeň. Dále byla přidána podpora pro data ve formátu CSV a zvýšila se „odolnost“ na vstup dat, která jsou v chybném formátu či nesouvisí se samotnými daty (např. komentáře vložené do zdrojového souboru).
3) Vedlejším efektem velkého počtu spekter, která TU Vídeň chtěla analyzovat, bylo, že aplikaci i databázi bylo možné otestovat na hranici výkonu. To mi výrazně pomohlo nalézt „úzká hrdla“, která aplikaci zpomalovala. Zátěžové testy databáze za přítomnosti velkého množství dat mi pomohly nalézt lepší nastavení webového i databázového serveru a proběhla optimalizace struktury databáze.

Tím se povedlo dosáhnout znatelného zrychlení při zpracování dat, ale především významného zrychlení při vyhledávání a vizualizaci spekter.

4) Přibyla schopnost vyhledat společné peaky po větší skupiny spekter. Tato funkce byla původně určena jen pro dvě spektra, nicméně v rámci praktických testů vyvstala potřeba hledat průniky pro větší skupiny spekter. Proto došlo k přepracování a rozšíření na schopnost zpracovat více spekter současně.

   Tato funkce byla posléze dále rozšířena o velice důležitou schopnost vyhledat a označit i takové peaky, které se vyskytují ve většině zkoumaných spekter, ale nelze je nalézt ve všech, např. z důvodu odchylek a chyb v měření, šumu atd... Touto „měkčí“ schopností detekce se získala aplikace schopnost zpracovat velkou skupinu spekter a přitom se vypořádat i s menšinou spekter, která jsou z různých důvodů nižší kvality.

5) Další novou funkcí se stala schopnost vytvářet a pracovat s virtuálními spektry, která lze s úspěchem použít pro hledání/ověřování charakteristických znaků v reálných spektrech či přímo k identifikaci reálného spektra (např. k jeho zařazení ke konkrétnímu druhu).

   Virtuální spektra vzniknou tak, že vytvoříme umělé spektrum, které obsahuje jen peaky, které jsou pro nás zajímavé. Při porovnání takovéhoto virtuálního spektra s reálnými spektry vypíše program reálná spektra, která obsahují charakteristické peaky typické právě pro hledané virtuální spektrum. To umožňuje rychlou a snadnou identifikaci reálného spektra např. pokud virtuální spektrum obsahuje znaky typické např. pro nějakou bakterii nebo chorobu.

6) Aplikace získala další možnost jak zobrazit data. Doposud si mohl uživatel zobrazit pouze stupeň shody jednoho spektra vůči ostatním spektrům (v podobě jednorozměrné tabulky). Nově byla přidána schopnost vygenerovat dvourozměrnou tabulku (v podobě tzv. kartézského součinu – porovnání každý prvek s každým) kde může uživatel na jednom místě vidět stupeň shody mezi všemi dvojicemi z vybrané množiny.

7) Během praktické práce s grafickým prostředím došlo k přidání a úpravě řady drobných prvků, k vylepšení ovládání a dalších maličkostí, které zpříjemňují uživateli práci a zlepšují ergonomii grafického prostředí.

8) Praktické využívání aplikace mi dodalo několik dalších podnětů a nápadů na další vylepšení, která mám v plánu zrealizovat.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Inovace ve vzdělávání v chemii a biologii s ohledem
na aktuální trendy v biomedicinálním výzkumu
reg. číslo: CZ.1.07/2.2.00/28.0184

Určení stadia fermentace vzorku

Tento úkol byl vyžádán ze strany TU Vídeň.

Situace: Máme změřených několik spekter z průběhu fermentace. Spektra reprezentují stav v jednotlivých časových okamžicích. Tato spektra rozdělíme do tří fází. První fází je začátek procesu fermentace, druhým stádiem je její průběh a třetím stadiem je ukončení a stabilizace vzorku.

Úkol: Zjistit, zda je *Biospean* schopen identifikovat stádium předloženého vzorku, pokud mu předložíme pro srovnání vzorky z předchozích reakcí. Pro testování mi byla poskytnuto šest skupin hmotnostních spekter, reprezentujících různé fermentace. U každé skupiny bylo známé pořadí, ve kterém byla měření prováděna a do které fáze fermentace vzorek patří.

Řešení: Identifikace stádia u vzorků se ukázala jako možná. Už verze *Biospeanu*, kterou jsem dovezl do Vídně, byla (po určité úpravě pro jejich specifická spektra) schopna celkem přesvědčivě identifikovat stádium fermentace. Následně bylo nalezeno a implementováno další řešení, které ukázalo ještě přesvědčivější výsledky.

První řešení bylo založeno na prostém faktu, že hmotnostní spektrum se během reakce proměňuje v závislosti na biochemických změnách.

Stádia fermentace nám reprezentují předložené vzory (označme si je jako x_1 až x_n). Pokud tedy chceme identifikovat stádium testovaného vzorku, musíme nalézt takové vzory, které vykazují největší shodu/podobnost s testovaným vzorkem. Z této podobnosti potom určíme stádium, ve kterém se testovaný vzorek nachází.

Teoreticky bychom měli dostat největší podobnost se vzory x_i a x_{i+1}, přičemž testovaný vzorek se nachází ve stavu mezi nimi. Další nejvíce podobné vzory by potom měly být vzory x_{i-1} a x_{i+2}. S rostoucí časovou vzdáleností od testovaného vzorku by postupně měla klesat podobnost testovaného vzorku s předloženými vzory. Podle vzorů x_i a x_{i+1} a dalších blízkých vzorů bychom tedy měli být schopni snadno určit stádium, ve kterém se testovaný vzorek nachází.

V praxi se tento předpoklad ukázal jako správný a funkční. *Biospean* ukazoval ve většině případů největší podobnost právě se vzorovými spektry, která byla nejblíže k testovanému vzorku.

V praktických výsledcích se občas vyskytovaly nepřesnosti, když bylo např. prohozeno pořadí nejbližších sousedů (např. x_{i+2} byl před x_{i+1}). Tyto odchylky však byly jednoznačně způsobené ne zcela optimálním nastavením *Biospeanu* (které jsem musel několikrát změnit kvůli specifikům sledovaných spekter) a především šumem v nepříliš kvalitně změřených

spektrech. Tuto nepřesnost nicméně bylo možné zkorigovat vizuální kontrolou výsledků a pro identifikaci nepředstavovaly zásadnější problém.

Obtížnější byla identifikace jen v případě, kdy se testovaný vzorek nacházel na hranici mezi dvěma stádii. Tam bylo určení konkrétního stádia pochopitelně obtížnější, což je logické, protože spektra v těchto místech obsahují znaky obou sousedících stádií .

Tento úkol vedl k prvním změnám v softwaru programu. Nejprve bylo třeba program upravit pro potřeby spekter z Vídně – byla nutná větší citlivost a rozdílné rozsahy hodnot. V dalším kroku pak byla pro potřeby lepšího zobrazení podobnosti vytvořena výše zmíněná možnost zobrazit výsledky ve dvourozměrné tabulce, kde bylo možné snadno vyčíst, jak si jsou navzájem jednotlivá spektra podobná.

Tyto pokusy rovněž poskytly velice důležitý důkaz o tom, že vyhledávací algoritmus *Biospeanu* pracuje skutečně správně. Pokud byly *Biospeanu* předloženy sekvence spekter, u kterých jsme znali skutečné pořadí (a tudíž i správné výsledky při testu podobnosti), porovnáním vypočtených výsledků mezi sebou ukázalo, že největší podobnost je téměř vždy právě se sousedními vzorky a s rostoucí časovou vzdáleností se podobnost postupně snižuje. To dokázalo, že srovnání spekter tak jak je navrženo, pracuje správně a občasné odchylky jsou spíše otázkou nastavení citlivosti a kvality spekter.

Druhé řešení pro nalezení fáze bylo vytvořeno až dodatečně, přičemž jsem využil funkce vytvořené pro vyřešení řešení jiného úkolu. Těmito funkcemi byly nově vytvořené nástroje pro generování virtuálních spekter.

Princip spočíval v tom, že jsem vytvořil ze všech vzorů daného stádia charakteristické virtuální spektrum. Z každé sekvence byly tedy vytvořeny tři virtuální spektra, přičemž každé reprezentovalo vždy jednu ze tří fází fermentace. Takto získaná spektra pro dané fáze jsem následně porovnával s testovanými vzorky.

Výsledek byl výborný. Chybné detekce prakticky zmizely a obtížnější detekce byla jen na rozhraní mezi fázemi, kde je to ale logické. Při společném použití s první metodou byla detekce fáze prakticky stoprocentní.

Závěr: *Biospean* lze s přijatelnou přesností použít pro detekci konkrétního stádia reakce. K identifikaci lze použít dvě na sobě nezávislé metody, které se mohou vhodně doplňovat.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Inovace ve vzdělávání v chemii a biologii s ohledem
na aktuální trendy v biomedicinálním výzkumu
reg. číslo: CZ.1.07/2.2.00/28.0184

Identifikace peaků u spekter plísní *blumeria*, *botrytis* a *bremia*

Tento úkol byl neplánovaný a vznikl při snaze pomoci kolegyni z KBC UPOL s identifikací peaků v jejich hmotnostních spektrech.

Úkol: Vyhledávání peaků v jednotlivých spektrech plísní. Upřesnění pozice peaků vzájemným porovnáním výsledků z programů *Biospean* a *mMass*. Nalezení peaků společných pro konkrétní druh plísně.

Provedení: Zatímco *mMass* vyznačí všechny nadprůměrné hodnoty podezřelé jako peak a ponechává na uživateli, jak je interpretuje, *Biospean* tyto peaky vybírá a označuje pro další zpracování. Chtěli jsme proto zjistit, zda je *Biospean* schopen sám najít charakteristické peaky v podobné kvalitě.

Výsledky z programu *Biospean* proto byly vizuálně porovnávány s výsledky z programu *mMass*. Z výsledků bylo vidět, že *Biospean* nemá žádné problémy s identifikací dostatečně intenzivních peaků. Nicméně v případě peaků se slabou intenzitou nacházejících se v šumu byla přesnost menší. Běžný byl výskyt tzv. „negative-false“ detekce, tedy peaku, který nebyl správně rozpoznán jako peak. Naopak se vůbec nevyskytly případy „pozitive-false“ detekce, tj. označení peaku, kterým peakem ve skutečnosti nebyl. Toto bylo zcela evidentně způsobeno tím, že citlivost programu byla nastavena na větší odstup od šumu. V případě snahy o detekci méně intenzivních peaků by se samozřejmě mezi peaky mohl dostat i intenzivní šum.

V průběhu testů vyvstala od kolegyně otázka, zda by bylo možné najít peaky společné pro všechna spektra daného druhu. Proto byla provedena úprava, která umožnila programu nalézt průniky mezi skupinami peaků. Z této funkce následně při řešení dalšího úkolu vznikla funkce pro práci s virtuálními spektry.

Identifikace druhu spór rodu *fusarium*

Tento úkol byl vyžádán ze strany TU Vídeň.

Situace: Máme změřená spektra různých druhů spór rodu fusarium. Celkem 735 spekter celkem šesti druhů spór. Spektra byla získána z různých lokací a spóry rostly na různých druzích obilnin (ječmen, pšenice, oves), jeden ze vzorků i na vlašských ořeších. Měření prováděli různí studenti na různém vybavení a spektra byla ve třech variantách s různým způsobem odstranění šumu.

Úkol: Najít způsob jak přiřadit testovaný vzorek ke konkrétnímu druh spór. Identifikovat znaky společné pro jednotlivé druhy spór.

Řešení: Zde se jako velice účinná ukázala funkce, která byla vytvořena pro hledání společných peaků pro více spekter. Bylo to logické řešení – najít ze všech spekter stejného druhu průnik společných peaků a z nich vytvořit vzor pro vyhledávání charakteristických peaků. Ze společných peaků se potom pomocí další funkce vytvořilo virtuální spektrum, které obsahovalo jen peaky charakteristické pro daný druh spór. Takové spektrum by při výpočtu shody s reálným spektrem ukázalo dostatečnou shodu, která by umožnila identifikaci příslušného vzorku.

Nicméně při prvních pokusech se vůbec nedařilo najít vhodné průniky. Některé skupiny spekter vykazovaly několik podobných znaků, ale u jiných skupin byl problém najít vůbec něco. Někdy se podařilo najít podobnosti až poté, co jsem ze skupiny vyjmul nejméně kvalitní spektra. Problém spočíval v samotných spektrech – velká úroveň šumu a hlavně nízká kvalita měření. Např. když dvě spektra ze stejného vzorku změřená dvěma různými studenty vypadala velice odlišně, bylo jasné, že kvalita měření spekter je velice zásadní problém.

Jako fungující řešení se ukázala úprava funkce pro hledání průniků, která byla poněkud „oslabena“ ve striktnosti vyhledávání. Byl přidán uživatelsky nastavitelný parametr (procentuální hodnota), který umožnil označit i peaky, které se vyskytují ve většině spekter, ale nemusí nutně být ve všech. Např. nastavením na 80 % je možné označit všechny peaky, které jsou alespoň v 80 % spekter.

Pro otestování identifikace jednotlivých druhů spór jsem tedy vytvořil virtuální spektra se shodami peaků 100 %, 80 %, 60 % a 40 %. Pomocí nich jsem potom testoval, zda je možné testované spektrum dostatečně přesně rozpoznat.

Výsledky byly povzbudivé – shoda proti virtuálním spektrům byla velice dobrá a dařilo se správně identifikovat naprostou většinu rozeznávaných spekter, přičemž virtuální spektra s citlivostí od 60 % výše se ukázala pro detekci druhu jako zcela postačující. Problém byl jen se spektry spór, které rostly na vlašských ořeších. Zde se našla jen minimální shoda s ostatními vzorky (které pocházely z obilovin) a identifikace vzorku byla velice obtížná.

Velkým povzbuzením byla skutečnost, že se díky schopnostem *Biospeanu* podařilo velice rychle najít způsob, jak v teoretické i praktické rovině vyřešit problém, který se vídeňské pracoviště marně snažilo najít delší dobu.

Závěr: *Biospean* je schopen vyhledávat charakteristické znaky pro celé skupiny spekter, vypořádat se s nepřesností části spekter, vytvářet z takových skupin virtuální spektra charakterizující např. konkrétní druhy organizmů (spór, bakterií) a identifikovat jiné vzorky za pomoci těchto uměle vytvořených virtuálních spekter.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Inovace ve vzdělávání v chemii a biologii s ohledem
na aktuální trendy v biomedicinálním výzkumu
reg. číslo: CZ.1.07/2.2.00/28.0184

Simulace peptidů souvisejících s Bergerovou chorobou

Tento úkol vznikl neplánovaně ve snaze pomoci kolegům z KBC UPOL, kteří pracují na výzkumu Bergerovy choroby, známé také jako IgA nefropatie. Konkrétně se jednalo o objasnění výskytu GalNAc a GalGalNAc v postranních řetězcích sledovaného proteinu.

Problém: Sledovaný řetězec umožňuje navázat na specifická místa deriváty galaktosy, konkrétně GalNAc a GalGalNAc. Cílem bylo zjistit, na kterých místech se navazují a jaké teoretické sekvence podřetězců mohou ve vzorcích vzniknout. Tyto výsledky následně porovnat s reálnými hmotnostními spektry a zjistit, které kombinace opravdu vznikají.

Velkou nepříjemností bylo, že počet kombinací a možných štěpů byl natolik vysoký, že v případě manuálního zpracování lidskými silami by šlo o velice zdlouhavý a náročný proces.

Nápad použít počítač vznikl spíše náhodou ve chvíli, kdy jsem viděl jak kolega zkouší vytvořit kombinace ručně a navrhl mu možnost použít automatický skript k vygenerování. Pak jsme jen přidávali další kroky až jsme se za pomoci počítačových skriptů a programu Biospean propracovali k výslednému řešení.

Řešení: Pro získání výsledků bylo třeba provést následující kroky.

Nejprve bylo třeba získat kombinace řetězců, které mohou z proteinu vzniknout. Proto jsem pomocí skriptu vytvořil reprezentaci pro všechny požadované kombinace proteinů. Vzniklo 262 kombinací. Z každé kombinace jsem následně vypočítal všechny možnosti rozštěpení, což vytvořilo 7 štěpů pro každou kombinaci, přičemž jeden štěp byl společný pro všechny kombinace.

U každé kombinace štěpů jsem vypočetl hmotnosti jednotlivých štěpů. Hmotnosti odpovídají x-ovým souřadnicím charakteristických peaků ve hmotnostním spektru. Z těchto sedmi souřadnic pro každou kombinaci jsem tedy skriptem vytvořil 262 virtuálních spekter, kdy každé virtuální spektrum obsahovalo právě těch sedm peaků charakteristických pro danou kombinaci.

Získaná virtuální spektra jsem načetl do *Biospeanu* stejně jako normální spektra a nechal jsem systém, aby vypočítal pozice peaků. Tím jsem získal charakteristická spektra, která bylo už možno porovnat se šestnácti reálnými spektry. Z výsledků vyplynulo, že v reálné situaci se může objevit jen část z teoreticky možných kombinací.

Teď už zbývalo jen vytvořit výstup v podobě, se kterou se dá pracovat. Využil jsem funkce vytvořené pro *Biospean* a vytvořil skript, který pro všech 16 reálných spekter udělal test podobnosti proti všem 262 virtuálním spektrům. Výsledky skript se vypsal v podobě

přehledné tabulky (resp. v podobě několika tabulek s různě nastavenou citlivostí), která umožnila další využití.

Získané výsledky výrazně pomohly – článek *Deciphering heterogeneity of Oglycans from the hinge region of human IgA1 using MALDI-TOF/TOF mass spectrometry: role of precise cysteine alkylation during sample processing (V. Franc, P. Řehulka, M. Raus, J. Stulík, M.Šebela, J. Novák)* o řešeném problému byl odeslán na recenzi v polovině ledna 2013.

Shrnutí

Ačkoliv jsem stáž bral ze začátku především jako povinnou součást doktorandského studia, kterou jako „nutné zlo“ prostě musím absolvovat, musím konstatovat, že v konečném byla po všech stránkách přínosem.

Především z toho důvodu, že došlo k praktickému využití vyvíjeného softwaru *Biospean* a v důsledku toho byla provedena řada úprav a vylepšení, která by byla realizována až později, po uvedení do zkušebního provozu. V samotném programu došlo k vylepšení výkonu a zvýšení robustnosti (odolnosti proti méně kvalitním datům). Bylo přidáno několik užitečných funkcí, které zvýšily schopnosti i užitnou hodnotu a během řešených úkolů byly schopnosti programu využity i způsobem, se kterým se zpočátku nepočítalo (např. virtuální spektra a simulace teoretických dat).

Z hlediska ověření správné funkce softwaru *Biospean* byly nejdůležitější výsledky získané při detekci stadia fermentace – první testovaná metoda dokázala správnou funkci detekce a porovnání spekter. Ačkoliv šlo o relativně snadný úkol, bylo důležité, že se poprvé pracovalo s daty, u kterých jsme přesně věděli, jak má jejich vzájemná podobnost vypadat. Tudíž bylo možné jasně říci, zda software vrátil správný výsledek. A získané výsledky skutečně vycházely téměř přesně tak, jak vycházet měly, což byl výborný důkaz, že software opravdu pracuje správně.

Po stránce vědeckého přínosu největší význam má vyřešení otázky peptidových štěpů u IgA nefropatie. Tyto výsledky se staly důležitou součástí probíhajícího výzkumu a fakticky završením předchozí práce kolegů. Počítačové vyhodnocení dat získaných v laboratoři výrazně zrychlilo a zjednodušilo zpracování dat a vedlo i k poněkud překvapivému závěru, kdy se z výsledků ukázalo, že řešený problém je zřejmě mnohem složitější než se původně očekávalo – slovy kolegy „máme víc otázek než odpovědí“. Získané výsledky zakončily několikaměsíční práci kolegů a staly se součástí článku, který šel prakticky ihned na recenzi.

Jako další vědecký přínos lze uvést i nalezení společných znaků ve spektrech spór rodu *fusarium*, což byl úkol, o který se vídeňské pracoviště marně snažilo delší dobu. Zde pro mne byla výhodou moje specializace (informatik-analytik-programátor), která mi dala možnost vyřešit problém z jiného úhlu pohledu, než jaký mají kolegové specializovaní na klasickou biochemii. Hlavní přínos přitom nespočívá v samotném rozeznávaní konkrétních spór, ale v přípravě mnohem obecnějšího řešení pro vyhledání shodných znaků ve spektrech, které je aplikovatelné i na jiné případy.

Celou stáž považuji za úspěšnou a přínosnou jak pro mne z hlediska profesního a studijního, tak pro kolegy z obou vědeckých pracovišť, kteří se účastnili řešených úkolů.

Mgr. Martin Raus

v Olomouci dne 7.1.2013