AKADEMIA MEDYCZNA W GDAŃSKU

Wydział Farmaceutyczny

Tomasz Bączek

USPRAWNIENIE IDENTYFIKACJI PEPTYDÓW W PROTEOMICE Z WYKORZYSTANIEM CHEMOMETRYCZNEJ ANALIZY DANYCH

Rozprawa habilitacyjna

GDAŃSK 2006

Wydano za zgodą
Senackiej Komisji Wydawnictw Akademii Medycznej w Gdańsku

© Copyright by Medical University of Gdańsk

Wydawca*: Akademia Medyczna w Gdańsku*
Druk*: Dział Wydawnictw AMG*
ul. Marii Skłodowskiej-Curie 3a,
Zlecenie KW/340/06

Natura jest świątynią, kędy słupy żywe
Niepojęte nam słowa wymawiają czasem.
Człowiek wśród nich przechodzi jak symbolów lasem,
One mu zaś spojrzenia rzucają życzliwe.
Charles Baudelaire *„Oddźwięki”*

Nie można przyrody zwyciężyć inaczej niż przez to, że się jej słucha.
Francis Bacon *„Novum organum” (1620), aforyzm 3*

Spis publikacji oryginalnych będących przedmiotem rozprawy habilitacyjnej:

[1] T. Bączek, Fractionation of peptides in proteomics with the use of *pI*-based approach and ZipTip pipette tips, *J. Pharm. Biomed. Anal.*, 34 (2004) 851-860. **(IF = 1,425)**

[2] T. Bączek, Fractionation of peptides and identification of proteins from *Saccharomyces cerevisiae* in proteomics with the use of reversed-phase capillary liquid chromatography and *pI*-based approach, *J. Pharm. Biomed. Anal.*, 35 (2004) 895-904. **(IF = 1,425)**

[3] T. Bączek, A. Buciński, A.R. Ivanov, R. Kaliszan, Artificial neural network analysis for evaluation of peptide MS/MS spectra in proteomics, *Anal. Chem.*, 76 (2004) 1726-1732. **(IF = 5,250)**

[4] R. Kaliszan, T. Bączek, A. Cimochowska, P. Juszczyk, K. Wiśniewska, Z. Grzonka, Prediction of high-performance liquid chromatography retention of peptides with the use of quantitative structure-retention relationships, *Proteomics*, 5 (2005) 409-415. **(IF = 5,483)**

[5] T. Bączek, P. Wiczling, M. Marszałł, Y. Vander Heyden, R. Kaliszan, Prediction of peptide retention at different HPLC conditions from multiple linear regression models, *J. Proteome Res.*, 4 (2005) 555-563. **(IF = 6,917)**

[6] T. Bączek, Chemometric evaluation of relationships between retention and physicochemical parameters in terms of multidimensional liquid chromatography of peptides, *J. Sep. Sci.*, 29 (2006) 547-554. **(IF = 1,829)**

[7] R. Put, M. Daszykowski, T. Bączek, Y. Vander Heyden, Retention prediction of peptides based on uninformative variable elimination by partial least squares, *J. Proteome Res.*, 5 (2006) 1618-1625. **(IF = 6,917)**

SPIS TREŚCI

1. Wstęp

Prawdopodobnie jedną z najbardziej znaczących zmian w życiu człowieka w ostatnim stuleciu stało się podwyższenie średniej długości życia z 45 do 75 lat. W istotnym stopniu przyczynił się do tego postęp w zakresie odkrywania nowych leków. Wzrost średniej długości życia stał się również nowym wyzwaniem rozwijającej się cywilizacji ludzkiej. Wraz ze starzejącym się społeczeństwem pojawiają się częstsze zachorowania na nowotwory, przypadki choroby Alzheimera, choroby Parkinsona. Pojawiają się także nowe choroby, takie jak AIDS, choroba Kreuzfelda-Jacoba, lekooporne infekcje bakteryjne, wirusowe i grzybicze. Pogarsza to znacznie jakość dłuższego życia. Wyzwaniem nauki XXI wieku stają się nieustanne poszukiwania nowych, skutecznych i bezpiecznych leków. Aby był to proces wystarczająco efektywny, niezbędne wydaje się poznanie i zrozumienie na poziomie molekularnym procesów fizjologicznych występujących w organizmach żywych [1].

Poszukiwanie nowych leków jest złożonym i wieloetapowym procesem. Rozpoczyna się on od identyfikacji odpowiedniego punktu uchwytu działania potencjalnego leku. Punktem uchwytu jest zwykle określone białko. Następnie przeprowadzana jest walidacja tego punktu uchwytu z wykorzystaniem modelu zwierzęcego lub kultury tkankowej, po czym prowadzone są badania przesiewowe związków małocząsteczkowych modulujących aktywność białka. Zoptymalizowane pod względem struktury chemicznej wybrane związki są potem testowane pod kątem ich skuteczności działania i toksyczności. Ostatecznie wyselekcjonowany związek poddawany jest badaniom klinicznym [2].

Ostatnie badania wskazują, że mniej niż 500 punktów uchwytu działania leków (receptory, enzymy, kanały jonowe) rozważanych jest obecnie w przemyśle farmaceutycznym. Z drugiej strony, zakrojone na skalę przemysłową badania zmierzające do odkrywania nowych leków dotyczą około 100 jednostek chorobowych. Liczba genów współdziałających podczas powstawania różnych fenotypów tej samej choroby waha się pomiędzy 5 a 10. Generuje to współdziałanie od 500 do 1000 potencjalnych białek będących produktami ekspresji genowej. W tym procesie każde z białek może wchodzić w oddziaływania z 3-10 innymi białkami. Daje to około 3000-10000 białek, które mogłyby być teoretycznie rozważane jako potencjalne punkty uchwytu działania dla związków biologicznie aktywnych [3,4]. Dlatego, przypuszcza się, że skuteczność analizy całościowego składu białkowego (proteomu) organizmu będzie

wkrótce również determinowała efektywność poszukiwania nowych związków o wartości terapeutycznej [5].

Poznanie możliwie największej liczby białek organizmów żywych powinno pomóc w opracowaniu oddziałujących na nie nowych związków chemicznych, w tym leków [6]. Przykładowo, związki oddziałujące z białkami kodowanymi przez geny czynne w procesach nowotworowych mogą stać się nowymi lekami powodującymi mniej działań ubocznych w porównaniu do obecnie stosowanych [7]. Z drugiej strony, białka te mogą stać się również biomarkerami diagnostycznymi wskazującymi na ryzyko powstawania nowotworów oraz monitorującymi proces chorobowy [8].

Do celów poszukiwania nowych leków na każdym etapie tego procesu pomocna może być proteomika [9]. Proteomika dąży do poznania współzależności możliwie wszystkich białek w danej komórce, tkance, organizmie. Obejmuje ona identyfikację i analizę ilościową całościowego składu białkowego, a także lokalizację poszczególnych białek, ich interakcje, aktywność i funkcje. Ma wypełnić lukę informacyjną pomiędzy wiedzą na temat kodu genetycznego a produktami ekspresji genów, czyli białkami [10]. Proteomika ma szanse stać się w przyszłości, obok genomiki, metabolomiki i innych nowoczesnych strategii biologiczno-chemicznych, jedną z wiodących dziedzin nauk o życiu. Sukces tej nowej dziedziny będzie jednak ściśle zależny od możliwości zaprojektowania i wykorzystania nowoczesnych i często nowatorskich, analitycznych i bioinformatycznych narzędzi badawczych umożliwiających studiowanie w szybki, efektywny i dokładny sposób ogromnej liczby biocząsteczek występujących w organizmach żywych [11].

Niniejsza rozprawa przedstawia wyniki studiów nad usprawnieniem identyfikacji peptydów i białek w proteomice, wykorzystując do tego celu chemometryczną analizę otrzymywanych danych. W badaniach wykorzystywano wysokosprawną chromatografię cieczową, technikę ogniskowania izoelektrycznego w roztworze, spektrometrię mas z jonizacją przez desorpcję laserową w stałej matrycy z analizatorem czasu przelotu oraz spektrometrię mas z jonizacją przez rozpylanie w polu elektrycznym. Podczas analizy otrzymywanych danych stosowano analizę korelacyjną i regresyjną, sztuczne sieci neuronowe oraz multiwariacyjne metody analizy danych. Krytycznej ocenie poddano zaproponowane, nowatorskie strategie analityczne i bioinformatyczne mające zastosowanie podczas analizy proteomicznej. Zamiarem autora było przedyskutowanie nowych rozwiązań służących do frakcjonowania peptydów i przetwarzania użytecznej analitycznie informacji w proteomice wykorzystując chemometryczną analizę danych.

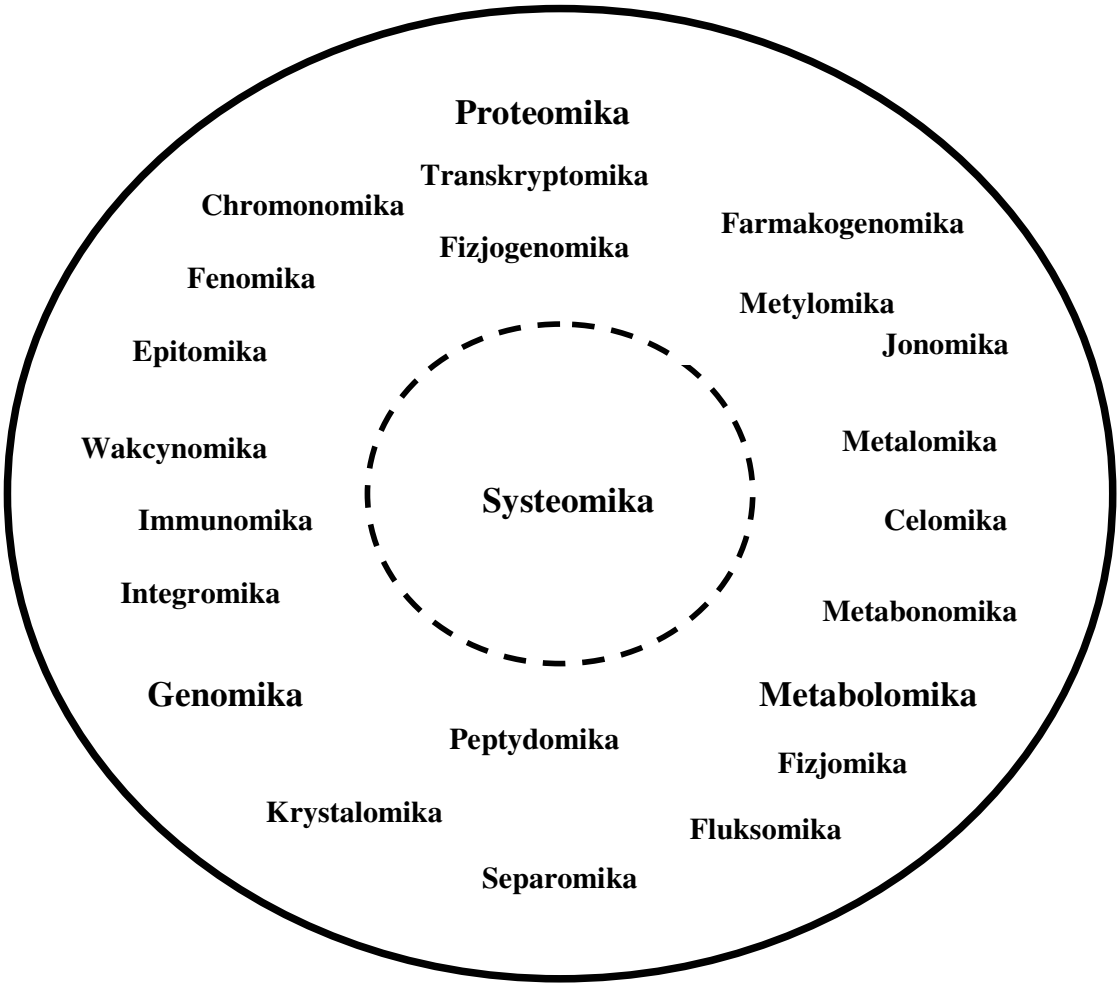
2. Podstawowe założenia proteomiki

Proteomika zajmuje się badaniem całościowego składu białkowego (proteomu) danej komórki lub danego organizmu, zapisanego w postaci informacji genetycznej zawartej w genomie, i obserwowanego w danym momencie czasu. Terminy „proteomika" i „proteom" zostały po raz pierwszy zaproponowane w latach 90-tych XX wieku [12]. Nawiązują one do uprzednio wprowadzonych terminów „genomika" i „genom", opisujących całkowity zestaw genów danej komórki lub danego organizmu. Techniki analityczne stosowane w proteomice umożliwiają pomiar ekspresji i aktywności białek oraz ocenę zjawisk fizykochemicznych i biologicznych na poziomie molekularnym, w które zaangażowane są białka. Ważną cechą białek, będących głównymi katalizatorami funkcji biologicznych, jest fakt, że odzwierciedlają one aktualny a nie potencjalny, jak w przypadku informacji zawartej w materiale genetycznym, stan komórki lub organizmu. W ten sposób proteomika może przyczynić się w bezpośredni sposób do zrozumienia stanów choroby i zdrowia na poziomie molekularnym i może być pomocna podczas odkrywania nowych leków [13,14]. Badania w zakresie genomiki, proteomiki, peptydomiki, transkryptomiki, metabolomiki, metabonomiki, fenomiki i innych nowoczesnych technologii biologiczno-chemicznych oraz ich integracja w formę interdyscyplinarnej systeomiki, wydają się obecnie niezbędne do zrozumienia procesów biologicznych zachodzących w organizmach żywych (rycina 1). Końcówka „-omika" (ang. *-omics*) symbolizować ma w wymienionych dziedzinach nową filozofię działania zmierzającego do całościowego zrozumienia funkcjonowania układów żywych.

Komórka może zawierać tysiące genów, które mogą przejawiać różnorodną ekspresję. Stąd też życie i śmierć komórki uzależnione są od ekspresji tych genów i aktywności ich produktów, czyli białek. Każde białko ma istotne znaczenie biologiczne, ale tylko w kontekście wszystkich innych funkcjonalnie aktywnych, współzależnych białek oraz innych składników komórki. Spojrzenie na układy żywe poprzez pryzmat „omiki", sprowadza się do zrozumienia danego układu biologicznego jako współdziałającej, zintegrowanej wewnętrznie całości, a nie do traktowania w sposób oddzielny poszczególnych jego składników [15].

Zarówno proteomika, jak i klasyczna biochemia zajmują się poznawaniem białek. Biochemia białek zajmuje się zasadniczo studiowaniem struktury białek, poznawaniem ich funkcji i jest silnie związana z biochemią fizyczną i enzymologią. Badania biochemiczne obejmują całościową analizę sekwencyjną i poznanie budowy przestrzennej białek. Celem tych badań

jest wyjaśnienie, jak struktura białka wpływa na jego funkcje biologiczne. Biochemicy poddają szczegółowym studiom zwykle pojedyncze białko lub kompleks białek ściśle ze sobą powiązanych. Do niedawna, biochemicy i biolodzy molekularni badali indywidualne geny i białka oraz poszczególne składniki różnych szlaków biochemicznych. Było to spowodowane faktem, że dostępne techniki analityczne umożliwiały jednoczesne badanie stosunkowo niewielkiej liczby genów lub białek.



Rycina 1. Proteomika i inne nowoczesne, badawcze strategie biologiczno-chemiczne (na podstawie [11]).

Proteomika jest nowym podejściem naukowym, którego celem jest badanie złożonych układów mieszanin białek w sposób kompleksowy. Bierze się pod uwagę wzajemne relacje jak największej liczby białek, traktowanych jako część całego układu biologicznego. Badania proteomiczne ukierunkowane są bezpośrednio na poznawanie złożonych mieszanin białek. Identyfikacja poszczególnych białek oparta jest na fragmentarycznej analizie sekwencyjnej, wystarczającej do oszacowania istnienia danego białka na podstawie wykorzystywanych

strategii analitycznych i bioinformatycznych. Celem proteomiki jest globalne scharakteryzowanie całego proteomu, a nie wybiórcze analizowanie poszczególnych jego elementów [13,14].

Fakt identyfikacji ludzkiego kodu genetycznego ogłoszono w lutym 2001 roku, a dokonali tego niezależnie badacze Projektu Badania Ludzkiego Genomu [16] oraz firmy biotechnologicznej Celera Genomics [17]. Dowiedziono wówczas, że genom człowieka zawiera około 30-40 tysięcy genów, mogących potencjalnie kodować białka. Obecnie, liczbę tę zredukowano do około 20-25 tysięcy [18].

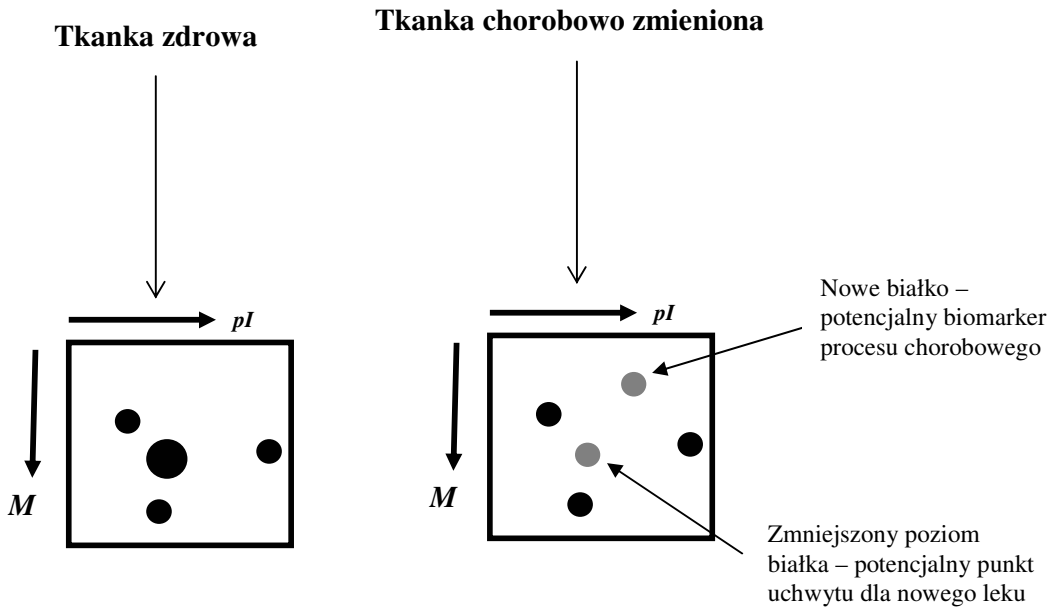
Analiza genomów z wykorzystaniem łańcuchowej reakcji polimerazy (ang. *polymerase chain reaction*, PCR), ukierunkowanej mutagenezy i sekwencjonowania DNA, należą obecnie do podstawowych metod inżynierii genetycznej. Występuje jednak słaba korelacja pomiędzy ekspresją genów a ostateczną ekspresją białek. Utrudnia to znacznie analizę proteomiczną. Komplikacje powodowane są ciągłymi zmianami stężenia białek w komórce oraz występowaniem tego samego białka w wielu różnych możliwych postaciach na skutek tzw. modyfikacji potranslacyjnych (np. fosforylacji, glikozylacji, hydroksylacji itp.). Szacując liczbę genów człowieka na ok. 20-25 tysięcy [18], przypuszcza się, że na ich podstawie może potencjalnie powstać co najmniej kilka-kilkanaście razy więcej białek. Zadanie zbadania tak ogromnej liczby białek komplikuje dodatkowo szeroki zakres stężeń, w którym mogą one występować w organizmie (rozpiętość 6-10 rzędów wielkości), zróżnicowanie ich właściwości fizykochemicznych i biologicznych oraz brak metod powielania białek lub peptydów w sposób analogiczny do replikacji DNA poprzez PCR [15,19].

Chociaż badania proteomu rozpoczęły się stosunkowo niedawno, to są one zaawansowane pod względem technicznym i informatycznym. Generalnie, badania w proteomice mogą dotyczyć identyfikacji składników złożonych mieszanin białek (ang. *mining*) oraz oceny ekspresji białek (ang. *protein-expression profiling*). Dotyczą również badań złożonych kompleksów białkowych (ang. *protein-network mapping*) oraz modyfikacji potranslacyjnych w obrębie poszczególnych białek (ang. *mapping of protein modifications*) [15,19].

Identyfikacja wszystkich (lub możliwe największej liczby) białek w danej próbce stanowi najprostszy typ badań proteomicznych i pozwala na poznanie danego proteomu. Przykładowo, charakteryzowano białka występujące w limfoblastach typu B [20], cieczy szklistej oka [21], wydzielinie z drzewa oskrzelowo-pęcherzykowego [22], płynie mózgowo-rdzeniowym

[23], tkankach zęba i przyzębia [24], surowicy i osoczu krwi [25,26], alergenach pochodzenia białkowego [27].

Ocena ekspresji białek stanowi bardziej zaawansowany etap badań w proteomice. Dokonywana jest ona w zależności od określonego stanu fizjologicznego lub patofizjologicznego organizmu bądź komórki, lub w funkcji ekspozycji danego układu biologicznego na zidentyfikowany czynnik zewnętrzny (na przykład lek lub inny ksenobiotyk). Najczęściej wykonywana jest analiza różnicowa polegająca na porównaniu dwóch badanych proteomów (rycina 2).



Rycina 2. Analiza różnicowa tkanki zdrowej i chorobowo zmienionej z wykorzystaniem dwuwymiarowej elektroforezy żelowej (na podstawie [8]).

Przykładowo, proteom z komórki zdrowej może być porównywany z proteomem z komórek uznanych za chorobowo zmienione celem oceny, jakie białka są charakterystyczne dla stanu zdrowia, a jakie dla choroby. Dlatego też, informacja uzyskana podczas badań proteomicznych może być pomocna w identyfikacji biomarkerów stanu chorobowego lub potencjalnych punktów uchwytu działania dla nowych leków.

3. Strategie analityczne i bioinformatyczne w proteomice

Pomimo trudności natury analitycznej oraz w zakresie przetwarzania ogromnej ilości informacji uzyskiwanej podczas badań proteomicznych, poznawanie proteomów jest obecnie realizowane. Stało się to możliwe dzięki integracji podstawowych narzędzi badawczych proteomiki: technik rozdzielczych, spektrometrii mas oraz algorytmów wykorzystywanych podczas przetwarzania danych ze spektrometrii mas i bioinformatycznych baz danych. Narzędzia te stanowią czuły i specyficzny instrument do identyfikacji i porównywania proteomów [28].

3.1. Techniki rozdzielcze w proteomice

Podstawowym narzędziem analitycznym wykorzystywanym w badaniach proteomicznych są techniki służące do frakcjonowania i rozdzielania białek i peptydów. Poprzez rozdzielenie złożonej mieszaniny uzyskuje się frakcje składające się z mniejszej liczby białek lub peptydów. Proces rozdzielenia pozwolić może na zaobserwowanie ewentualnego zróżnicowania pod względem składu białkowego dwóch porównywanych próbek. Możliwa jest także selektywna ekstrakcja danego białka z mieszaniny [15,29].

Jednokierunkowa elektroforeza żelowa w żelu poliakryloamidowym (ang. *one-dimensional polyacrylamide gel electrophoresis*, 1D-PAGE) oraz dwukierunkowa elektroforeza żelowa w żelu poliakryloamidowym (ang. *two-dimensional polyacrylamide gel electrophoresis*, 2D-PAGE) należą tradycyjnie do podstawowych technik rozdzielczych w proteomice. Pomimo wielu trudności napotykanych podczas analiz tymi technikami, pozostają one najpopularniejszymi technikami służącymi do rozdzielania złożonych mieszanin białek [20-23,29-34].

Dwuwymiarowa elektroforeza żelowa jest wciąż często stosowaną techniką rozdzielczą w proteomice [29,30]. Dotychczasowa popularność i częstość wykorzystywania dwukierunkowej elektroforezy żelowej jest związana z możliwością rozdzielania za jej pomocą znacznej liczby białek. Rozdzielenie w 2D-PAGE realizowane jest na podstawie różnic ładunku elektrycznego (różnic w zakresie punktu izoelektrycznego) rozdzielanych białek w pierwszym wymiarze oraz na podstawie wielkości cząsteczki (różnic w zakresie masy cząsteczkowej) w drugim wymiarze.

Wykorzystanie proteomiki w badaniach biomedycznych wymaga technik rozdzielczych pozwalających analizować znaczną ilość próbek w stosunkowo krótkim czasie. Elektroforeza

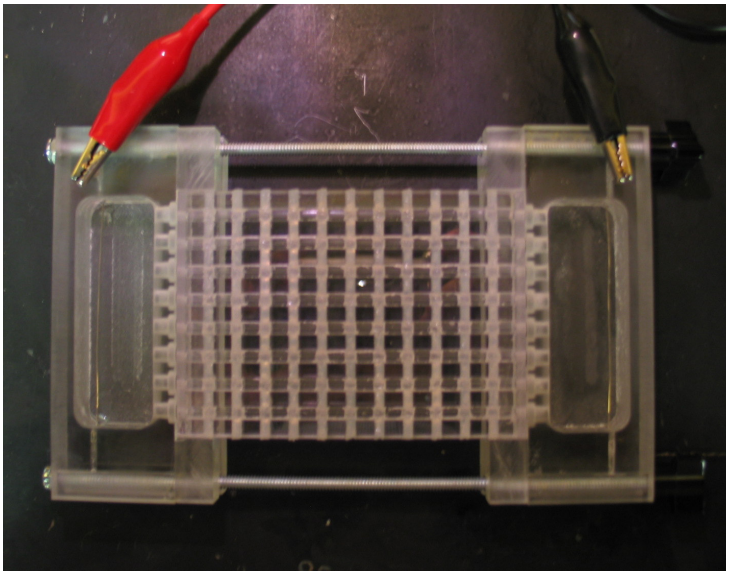
żelowa nie jest jednak narzędziem, które można wykorzystywać do takich celów w łatwy sposób. Nawet ostatnie udoskonalenia pozostawiają wciąż tę technikę czasochłonną i pracochłonną. Dwuwymiarowa elektroforeza żelowa może być obecnie częściowo zautomatyzowana. Konieczny jest jednak odpowiedni czas na przeprowadzenie ogniskowania izoelektrycznego w pierwszym wymiarze, następnie wykonanie doświadczenia w drugim wymiarze i wizualizację plam rozdzielonych białek. Głównym jednak minusem dwukierunkowej elektroforezy żelowej jest niemożność identyfikacji z jej udziałem całego proteomu. Duże i bardziej hydrofobowe białka słabo przemieszczają się wewnątrz żelu, a białka o charakterze kwasowym i zasadowym są gorzej rozdzielane. Białka występujące w małych stężeniach w materiale biologicznym zwykle są poniżej granicy wykrywalności. To ograniczenie wydaje się być najbardziej istotne, ponieważ wiele białek regulacyjnych, odgrywających istotną rolę w procesach chorobowych i mogących stać się miejscami uchwytu dla nowych leków, znajduje się w komórkach w bardzo małych stężeniach.

Oprócz elektroforezy żelowej, także inne techniki oparte na ogniskowaniu izoelektrycznym (ang. *isoelectric focusing*, IEF) [35-39] oraz wysokosprawna chromatografia cieczowa [40-47] i elektroforeza kapilarna [48-50] są stosowane w badaniach proteomicznych do frakcjonowania i rozdzielania białek i peptydów. Techniki te stają się obecnie istotną konkurencją dla elektroforezy żelowej.

W dwuwymiarowej elektroforezie żelowej rozdzielenie oparte jest na różnicach w zakresie punktów izoelektrycznych i masy rozdzielanych białek [51-54]. Ogniskowanie izoelektryczne jest techniką elektroforetyczną, w której białka lub peptydy rozdzielane są w zależności od różnic w zakresie ich punktów izoelektrycznych (*pI*). Białka i peptydy jako związki amfoteryczne i w zależności od pH środowiska, w którym się znajdują, posiadają dodatni, ujemy lub zerowy ładunek elektryczny. Średni, wypadkowy ładunek elektryczny białka lub peptydu jest sumą wszystkich dodatnich i ujemnych lokalnych ładunków w cząsteczce. Za punkt izoelektryczny uznaje się takie pH środowiska, w którym wypadkowy ładunek białka lub peptydu równa się zero. *pI* należy do jednych z najważniejszych deskryptorów charakteryzujących dane białko lub peptyd pod względem fizykochemicznym. Wielkość *pI* może być również pomocna podczas identyfikacji białek w proteomice [38,39].

Dwuwymiarowa elektroforeza żelowa ma znaczącą siłę rozdzielczą. Jest jednak także obarczona licznymi wadami wymienionymi uprzednio. Z drugiej strony, technika ogniskowania izoelektrycznego może być przeprowadzona nie tylko na pasku żelowym z immobilizowanym gradientem pH (ang. *immobilized pH gradient*, IPG) lub w rurce żelowej, ale też

bezpośrednio w roztworze. Gradient pH w tych metodach ogniskowania izoelektrycznego wytwarzany jest za pomocą roztworów amfolitów charakteryzujących się określonym zakresem pH lub roztworów tzw. immobilin (ang. *immobilines*), charakteryzujących się określonym pH i immobilizowanych w sieci żelu poliakryloamidowego [38]. Po przyłożeniu napięcia generowany jest wówczas stabilny gradient pH, umożliwiający rozdzielenie białek lub peptydów zgodnie z posiadanymi przez nie wartościami punktu izoelektrycznego.



Rycina 3. Urządzenie do ogniskowania izoelektrycznego w roztworze (sIEF), wykonane przez autora rozprawy na podstawie [37] i wykorzystane podczas badań przedyskutowanych w [38,39].

Rozdzielanie oparte na różnicach punktu izoelektrycznego analitów może być realizowane bezpośrednio w roztworze z wykorzystaniem kilku nowatorskich urządzeń. Do takich należy rotofor [55], urządzenie do mikroskalowego ogniskowania izoelektrycznego w roztworze (ang. *microscale solution isoelectrofocusing device*, μsol-IEF) [36], urządzenie do ogniskowania izoelektrycznego typu off-gel (ang. *off-gel isoelectric focusing*) [56], elektrolizer wielokompartmentowy (ang. *multicompartment electrolyzer*) [35,57]. Na uwagę zasługuje także ogniskowanie chromatograficzne (ang. *chromatofocusing*) [58], w którym wykorzystuje się technikę chromatograficzną rozdzielania białek opartego na różnicach w zakresie punktów izoelektycznych. Podczas procesu ogniskowania izoelektycznego białek może jednak zachodzić ich agregacja i wytrącanie [15]. Tego niebezpieczeństwa nie ma w przypadku peptydów. Właściwości fizykochemiczne peptydów, otrzymanych po trawieniu proteolitycznym białek, są mniej zróżnicowane niż oryginalnych białek. Ponadto, większość peptydów jest łatwo

rozpuszczalna w wodzie lub rozworach wodno-organicznych. Dlatego rozdzielanie peptydów w oparciu o różnice ich punktów izoelektrycznych możliwe jest obecnie także poprzez zastosowanie kapilarnego ogniskowania izoelektrycznego (ang. *capillary isoelectric focusing*, cIEF) [59,60] lub ogniskowania izoelektrycznego w roztworze (ang. *isoelectric focusing in solution*, sIEF) [37-39] (rycina 3).

Jedną z najważniejszych technik rozdzielczych, wykorzystywanych w celach naukowych i w rutynowych pomiarach laboratoryjnych, jest obecnie wysokosprawna chromatografia cieczowa. Technika ta jest powszechnie wykorzystywana w analizach farmaceutycznych i chemicznych. Jest ona wartościowym narzędziem analitycznym w laboratoriach klinicznych, a także podczas pomiarów właściwości fizykochemicznych związków chemicznych [62,63]. Znajduje ona również szerokie zastosowanie w badaniach proteomicznych [15,64].

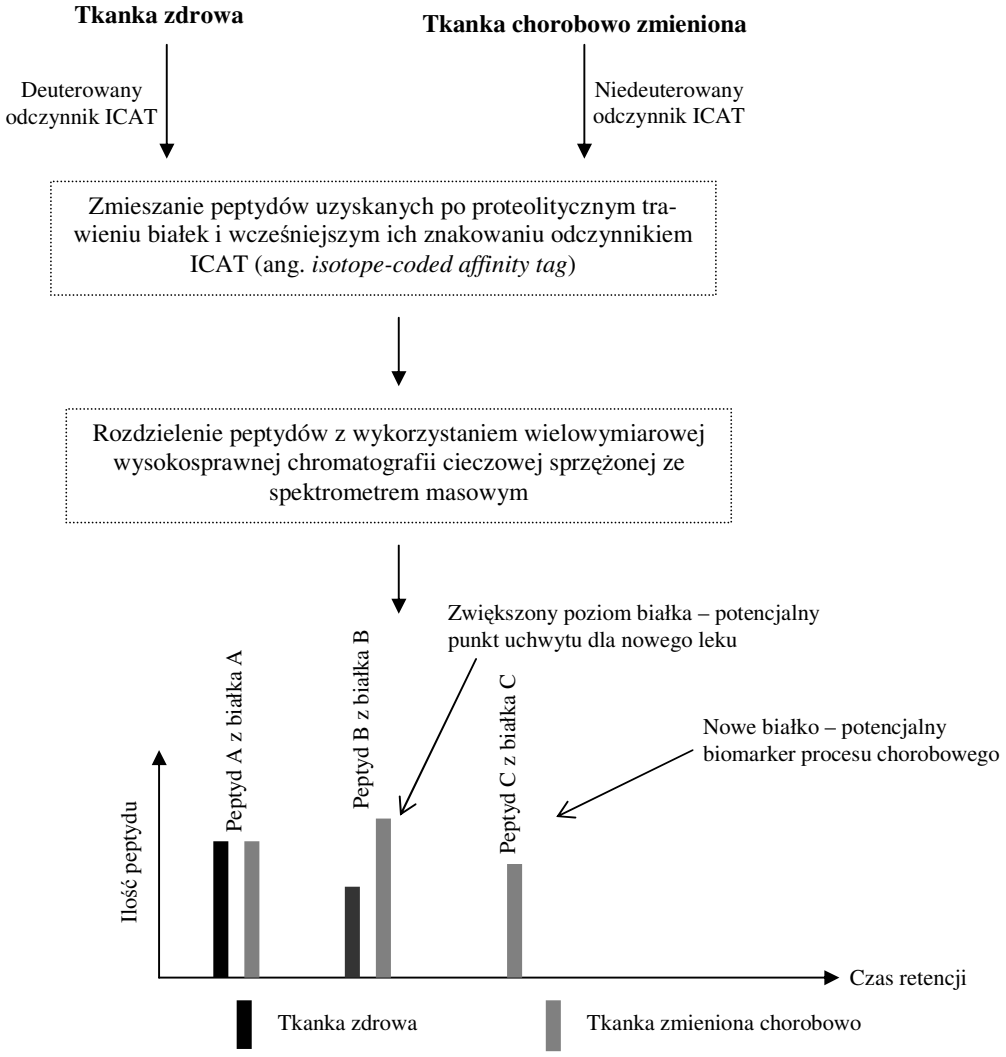
Techniki chromatograficzne mają wiele zalet w porównaniu do technik opartych na elektroforezie żelowej. Mogą być wykorzystywane zarówno do wstępnego frakcjonowania złożonej mieszaniny, jak i późniejszego rozdzielania białek i peptydów. Duża pojemność upakowań kolumn stosowanych w wysokosprawnej chromatografii cieczowej (ang. *high-performance liquid chromatography*, HPLC) jest cenna z punktu widzenia wymogów rozdzieleń preparatywnych. Przy użyciu kolumn chromatograficznych możliwe jest także zatężanie analitów z dużych objętości próbek bez utraty ich rozdzielczości. Technika HPLC pozwala na zautomatyzowanie procesu wprowadzania próbki, rozdzielania i detekcji oraz zbierania frakcji [61]. Szeroki wybór rodzajów technik chromatograficznych umożliwić może uzyskanie dobrego rozdzielenia i detekcję różnych białek i peptydów, które trudno jest analizować techniką elektroforezy żelowej. Chromatografia w odwróconym układzie faz oraz chromatografia oddziaływań hydrofobowych należą do podstawowych technik rozdzielania białek i peptydów opartego na ich właściwościach hydrofobowych i polarności. Chromatografia kationowymienna może być wykorzystywana do rozdzielania białek o charakterze zasadowym, a anionowymienna – białek o charakterze kwasowym. Obie techniki służą także do rozdzieleń peptydów w oparciu o różnice w zakresie ładunków elektrycznych. Chromatografia wykluczania objętościowego (chromatografia sitowa) wykorzystywana jest do wstępnego frakcjonowania białek, a chromatografia powinowactwa jest pomocna podczas izolacji specyficznych białek lub peptydów [15,61].

Wstępna charakterystyka genomu człowieka wykazała, że zawiera on porównywalną liczbę genów kodujących białka w stosunku do genomu myszy domowej (*Mus musculus*) czy rzodkiewnika pospolitego (*Arabidopsis thaliana*) [16-18,65]. Na proteom organizmu

człowieka w przeciągu jego całego życia składać się może natomiast od około kilkudziesięciu tysięcy do nawet kilkuset tysięcy białek. Charakterystyka całego zestawu białek w określonych warunkach fizjologicznych lub patofizjologicznych jest podstawowym celem eksperymentu proteomicznego. W przypadku badań przeprowadzanych z wykorzystaniem technik chromatograficznych jako metody rozdzieleń, początkowo przeznaczone do rozdzieleń białka traktowane są odpowiednim enzymem proteolitycznym (np. trypsyną), generując powstania licznych peptydów. Peptydy mogą być następnie rozdzielane przed ostateczną ich identyfikacją na podstawie widm masowych uzyskanych za pomocą spektrometrii mas. Trawienie proteolityczne z wykorzystaniem trypsyny generuje około 20-50 peptydów z jednego białka. Oznacza to, że ostateczna próbka, np. proteom drożdży piekarniczych (*Saccharomyces cerevisiae*), które mają około 6 tysięcy genów kodujących białka, zawierać może przynajmniej 120 tysięcy peptydów. Nawet najlepszy pojedynczy chromatograficzny układ rozdzielczy nie jest w stanie rozdzielić w sposób satysfakcjonujący tak złożonej próbki. Dlatego też bardzo często spotyka się w proteomice złożone zestawy układów chromatograficznych, w których proces rozdzielenia opiera się na różnych właściwościach fizykochemicznych peptydów. Umożliwić to ma ostateczne uzyskanie odpowiedniego do celów identyfikacyjnych rozdzielenia złożonych mieszanin peptydów [15,29].

W przypadku dwukierunkowej elektroforezy żelowej występują dwa niezależne od siebie układy rozdzielcze. Działają one w oparciu o różnice ładunku elektrycznego (punktu izoelektrycznego), i jest to rozdzielenie białek w pierwszym wymiarze, oraz na podstawie wielkości cząsteczki (masy cząsteczkowej), i jest to rozdzielenie w drugim wymiarze. W przypadku rozdzielenia realizowanego z wykorzystaniem chromatografii, aby uzyskać porównywalne do dwuwymiarowej elektroforezy żelowej możliwości rozdzielcze, należy skorzystać z dwóch technik chromatograficznych (układ dwuwymiarowy) różniących się znacznie mechanizmem rozdzielenia pod względem fizykochemicznym. Stosuje się więc chromatografię jonowymienną odzwierciedlającą pierwszy wymiar w dwukierunkowej elektroforezie żelowej. Rozdzielenie oparte jest wówczas na różnicach w zakresie ładunków elektrycznych peptydów, oraz chromatografię w odwróconym układzie faz, w której rozdzielenie oparte jest na różnicach w zakresie hydrofobowości peptydów. W praktyce, podczas korzystania z dwuwymiarowego układu chromatograficznego, próbka wędruje najpierw na kolumnę jonowymienną. Z tej kolumny poszczególne frakcje peptydów są wymywane eluentem o wzrastającym etapowo stężeniu roztworu soli (np. 350 mM KCl [66-69]), według uprzednio zaprogramowanego gradientu. Frakcje te zatrzymywane są na prekolumnie pracującej w odwróconym układzie

faz. Po zmianie kierunku przepływu fazy ruchomej dokonanej poprzez układ zaworów, peptydy wymywane z prekolumny rozdzielane są na kolumnie pracującej w odwróconym układzie faz i następnie analizowane w spektrometrze mas.



Rycina 4. Analiza porównawcza tkanek zdrowych i zmienionych chorobowo w proteomice z wykorzystaniem dwuwymiarowej chromatografii cieczowej (na podstawie [8,70]).

Wielowymiarowa chromatografia cieczowa (ang. *multidimensional liquid chromatography*), w połączeniu z kompatybilnymi z nią technikami spektrometrii mas (tandemową spektrometrią mas z jonizacją przez rozpylanie w polu elektrycznym i spektrometrią mas z jonizacją przez desorpcję laserową w stałej matrycy), stają się obecnie coraz częściej wykorzystywanymi narzędziami w badaniach proteomicznych [66-76]. Przykładowo, analizę różnicową próbek pochodzących z tkanek zdrowych i chorych można dokonać nie tylko z wykorzystaniem elektroforezy żelowej, lecz i dwuwymiarowej chromatografii cieczowej (rycina 4).

3.2. Spektrometria mas w proteomice

Istotnym narzędziem analitycznym w proteomice jest spektrometria mas (ang. *mass spectrometry*, MS). Spektrometria mas to technika analityczna, w której wykorzystywana jest jonizacja analitów w fazie gazowej oraz rozdzielanie otrzymanych jonów w polu elektrycznym i magnetycznym. Rozdzielanie jonów zależy od stosunku wartości ich masy do liczby ładunków (*m/z*), a identyfikacja analitów dokonywana jest na podstawie widm masowych. Spektrometria mas, osiągając wysoki poziom techniczny, stała się obecnie czułym i wiarygodnym narzędziem podczas analizy biocząsteczek. Technika ta jest bardzo użyteczna w proteomice. Przede wszystkim, dzięki spektrometrii mas możliwe stało się otrzymywanie bardzo dokładnych wartości mas cząsteczkowych peptydów i białek. Jest to aktualnie wiodąca metoda pomiaru mas cząsteczkowych białek i peptydów, wypierająca pomiary oparte na różnicach w migracji biocząsteczek w żelu poliakryloamidowym. Nie mniej jednak, nawet najdokładniejszy pomiar masy cząsteczkowej (szczególnie, gdy rozważane są złożone mieszaniny białek, będące obiektem badań proteomicznych) jest często niewystarczający do jednoznacznego zidentyfikowania danego białka lub peptydu. Jednakże spektrometria mas może być także stosowana w analizie sekwencyjnej peptydów poprzez wykorzystanie widm masowych typu MS/MS. Analiza widm MS/MS jest uważana obecnie za metodę umożliwiającą jednoznaczną identyfikację peptydów. W związku z tym, możliwa jest również identyfikacja białka, którego fragment struktury pierwszorzędowej stanowią te peptydy [28,77].

Analityczna identyfikacja białek w proteomice oparta jest na wstępnym ich proteolitycznym trawieniu do peptydów, określeniu sekwencji otrzymanych peptydów i wykorzystaniu tej sekwencji do identyfikacji białek poprzez przeszukiwanie odpowiednich, bioinformatycznych baz danych sekwencji peptydowych. Badania proteomiczne rozpoczynają się od mieszaniny białek charakteryzujących się zróżnicowanymi masami cząsteczkowymi, rozpuszczalnością i modyfikacjami potranslacyjnymi. Aby otrzymać mieszaninę peptydów z pierwotnej mieszaniny białek, należy poddać badane białka trawieniu proteolitycznemu. Jest to konieczne, gdyż aktualnie dostępna spektrometria mas dokonuje najdokładniejszych pomiarów mas cząsteczkowych właśnie w przypadku peptydów, w przeciwieństwie do samych cząsteczek białek. Również widma MS/MS, niezbędne do wiarygodnych identyfikacji peptydów, są otrzymywane dla peptydów.

Nowoczesne spektrometry mas są w stanie mierzyć masy cząsteczkowe dla względnie złożonych mieszanin peptydów. Uproszczenie złożoności tych mieszanin jest jednak konieczne w przypadku bardzo skomplikowanych próbek. Aby więc efektywnie analizować taką

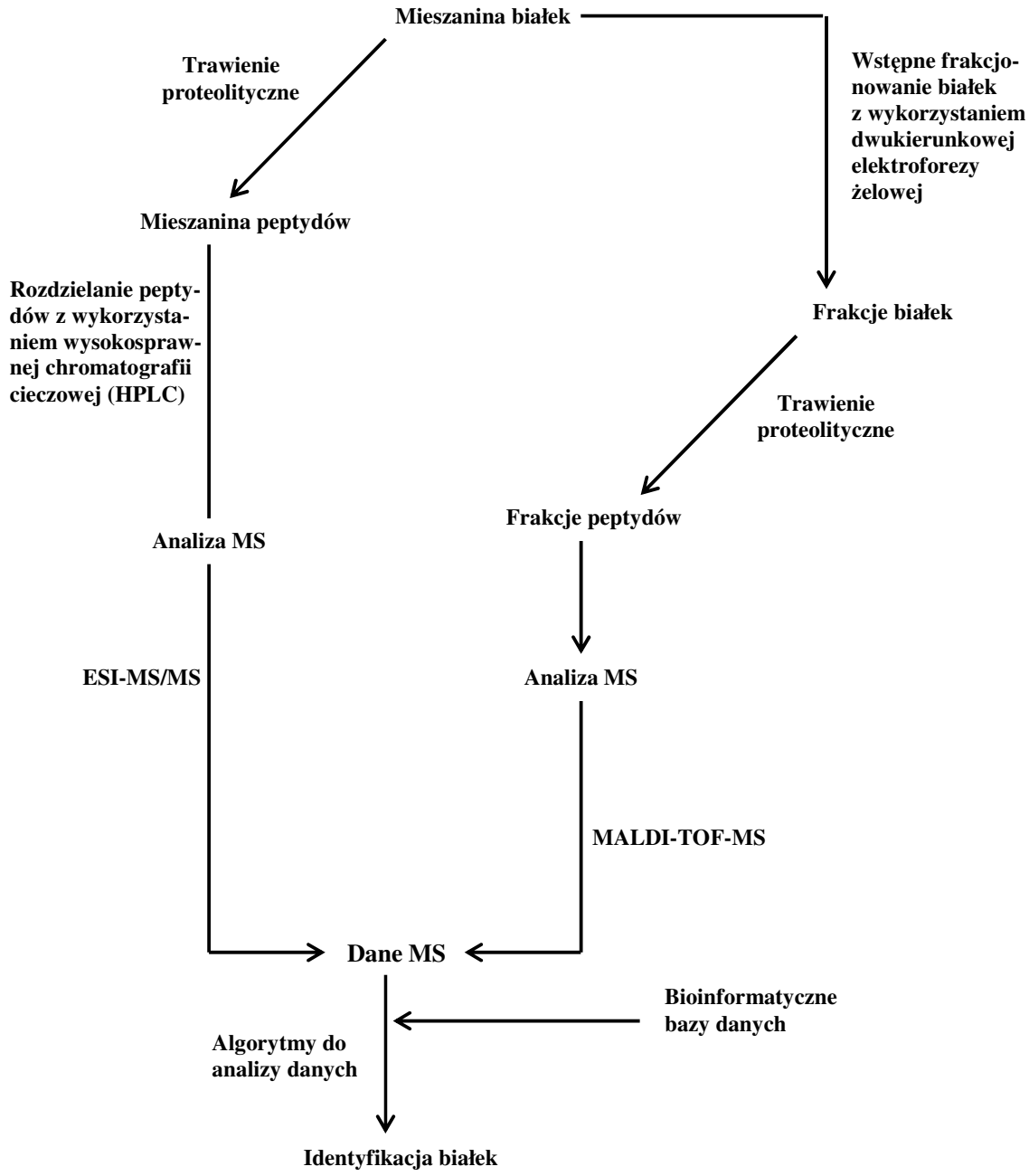
mieszaninę peptydów z wykorzystaniem spektrometrii mas, mieszanina ta musi być wstępnie rozdzielona na frakcje zawierające mniejszą ilość składników. Przyjęta strategia wstępnego frakcjonowanie może przy tym przebiegać na dwa sposoby. Najpierw rozdzielana może być mieszanina białek, a następnie rozdzielane białka trawione są do peptydów i analizowane za pomocą spektrometrii mas. Często spotykanym postępowaniem jest także wstępne trawienie białek do peptydów w początkowej mieszaninie. Rozdzielanie dokonywane jest następnie dla mieszaniny peptydów [15,69,78] (rycina 5).

Identyfikacja peptydów w proteomice dokonywana jest najczęściej z wykorzystaniem dwóch typów spektrometrów mas: spektrometru mas z jonizacją przez desorpcję laserową w stałej matrycy z analizatorem czasu przelotu (ang. *matrix-assisted laser desorption ionisation time of flight mass spectrometry*, MALDI-TOF-MS) oraz tandemowego spektrometru mas z jonizacją przez rozpylanie w polu elektrycznym (ang. *electrospray ionisation tandem mass spectrometry*, ESI-MS/MS) [79,80].

3.3. Bioinformatyka i chemometria w proteomice

Niezbędnymi narzędziami podczas badań proteomicznych są algorytmy umożliwiające analizę danych ze spektrometrii mas w powiązaniu z informacją z bioinformatycznych baz danych. Analiza ta służy do identyfikacji białek na podstawie mas cząsteczkowych i sekwencji aminokwasowych analizowanych peptydów. Do najbardziej znanych i wykorzystywanych bioinformatycznych baz danych należą kolekcje sekwencji białkowych Universal Protein Resource (UniProt Knowledgebase) [81] oraz kolekcje sekwencji białkowych National Center for Biotechnology Information (NCBI) [82]. Cenne pod względem praktycznym są ponadto proteomiczne bazy danych European Bioinformatics Institute (EBI) [83].

Do efektywnego analizowania informacji ze spektrometii mas i bioinformatycznych baz danych niezbędne są odpowiednie, specjalistyczne algorytmy. Do najbardziej znanych należą: Mowse w programie komputerowym Mascot [84], Sequest w programie BioWorks [85] oraz programy MS-FIT i MS-TAG [86].



Rycina 5. Podstawowe strategie analityczne w proteomice (na podstawie [78]).

Właściwa interpretacja sekwencji peptydów na podstawie tych informacji nie jest jednak zadaniem łatwym. Szczególnie, jeżeli weźmie się pod uwagę, że w przypadku złożonych mieszanin białek odpowiedniej interpretacji podlegać muszą tysiące widm masowych dla setek peptydów. Generalnie, widma masowe są analizowane poprzez dopasowanie eksperymentalnie uzyskanych sekwencji aminokwasowych analizowanych peptydów do sekwencji

teoretycznie możliwych. Specjalistyczne algorytmy pozwalają na automatyczne analizowanie dużej liczby widm masowych. Informacja pochodząca z tych widm wykorzystywana jest później do identyfikacji białek.

Niektóre algorytmy oparte są na informacji uzyskanej eksperymentalnie odnoszącej się do wartości mas cząsteczkowych peptydów („podejście peptydowego odcisku palca", ang. *peptide mass fingerprinting approach*). Inne wykorzystują informacje z tandemowych widm masowych (widmo MS/MS), odzwierciedlających sekwencje aminokwasowe analizowanych peptydów i potwierdzających identyfikację danego białka (podejście przeszukiwania jonów fragmentarycznych MS/MS, ang. *MS/MS ions search approach*) [87]. Dane eksperymentalne są porównywane z wyliczonymi masami cząsteczkowymi peptydów lub wartościami mas jonów fragmentarycznych, otrzymanymi poprzez zastosowanie teoretycznego trawienia białek według ściśle określonych reguł. Wartości mas, a także widma MS/MS, są następnie interpretowane w kontekście identyfikacji peptydów i białek.

Należy również wspomnieć, że w bioinformatyce, oprócz wykorzystywania rozwiązań typowo informatycznych związanych z komputerowo wspomaganym przeszukiwaniem baz danych, stosowane są także rozwiązania oparte na metodach statystycznych i chemometrycznych. Chemometria jest dziedziną zajmującą się wydobywaniem użytecznej informacji z wielowymiarowych danych pomiarowych, wykorzystującą metody statystyki i matematyki [88]. Do jednych z podstawowych zadań chemometrii należy przewidywanie wartości zmiennej zależnej (odpowiedzi) badanego zjawiska na podstawie wartości zmiennych objaśniających. Proces modelowania danego zjawiska polega na tym, że aby przewidzieć wartość odpowiedzi należy stworzyć wiarygodny model matematyczny. Podstawowym, najczęściej stosowanym sposobem wyznaczania parametrów modelu jest metoda regresyjna. Analiza regresji wykorzystywana jest w badaniach zależności pomiędzy zmiennymi i pozwala na przewidywanie wartości jednej zmiennej, nazywanej zmienną zależną, na podstawie jednej lub większej liczby innych zmiennych, nazywanych zmiennymi niezależnymi. Należy w tym miejscu zaznaczyć, że w przeciwieństwie do metody regresyjnej, metody korelacyjne służą jedynie do wykrycia związku pomiędzy dwiema zmiennymi lub większą ich liczbą oraz oszacowania siły i istotności statystycznej tego związku.

Ilościowe zależności struktura-retencja (ang. *quantitative structure-retention relationships*, QSRR) są na ogół wyprowadzane z użyciem analizy liniowej regresji wieloparametrowej (ang. *multiple regression analysis*, MRA). Są one statystycznie wyprowadzonymi zależnościami

pomiędzy parametrami chromatograficznymi i deskryptorami charakteryzującymi strukturę molekularną analitów.

Obecnie coraz częściej pojawiają się także inne metody budowania i identyfikacji modeli. Sztuczne sieci neuronowe (ang. *artificial neural networks*, ANN), to nowoczesna metoda obliczeniowo-predykcyjna, która może znaleźć zastosowanie w takich przypadkach. Większość tradycyjnych analiz statystycznych zorientowanych jest na stworzenie użytecznego modelu na bazie pewnych założeń i rozważań teoretycznych (np. o liniowości zależności pomiędzy zmiennymi). Podejście do problemu z pomocą sieci neuronowych wolne jest od większości standardowych założeń i nadaje się bardzo dobrze do modelowania złożonych, nieliniowych zależności.

W praktyce chemometrycznej występuje często także potrzeba analizy struktury wewnętrznej wielowymiarowego zbioru danych. Często gromadzone dane dotyczą zmiennych, które nie tylko są ze sobą skorelowane, ale także zawierają bardzo dużą liczbę przypadków. Powodować to może problemy związane z interpretacją danych oraz może utrudniać wykrycie ich struktury. Metoda analizy głównych składowych (ang. *principal component analysis*, PCA) ułatwia wykonanie tych zadań poprzez przekształcenie oryginalnych zmiennych do mniejszej liczby nowych, nieskorelowanych zmiennych (tzw. głównych składowych). Z drugiej strony, metoda cząstkowych najmniejszych kwadratów (ang. *partial least squares*, PLS) jest metodą regresji liniowej, która w charakterze zmiennych niezależnych używa także nowych składowych (zmiennych ukrytych, ang. *latent factors*). Regresja metodą cząstkowych najmniejszych kwadratów stanowi jednak rozszerzenie modelu liniowej regresji wieloparametrowej i umożliwia analizę wpływów dużej liczby zmiennych niezależnych dowolnego typu na dużą liczbę zmiennych zależnych. Dalszym rozszerzeniem możliwości samej metody PLS jest metoda cząstkowych najmniejszych kwadratów z eliminacją zmiennych niewnoszących istotnej informacji (ang. *uninformative variable elimination by partial least squares*, UVE-PLS).

4. Usprawnienie identyfikacji peptydów w proteomice z wykorzystaniem chemometrycznej analizy danych (prace własne)

Poznanie genomów umożliwić może lepsze zrozumienie funkcji biologicznych organizmów żywych. Jednak badania samego genomu dostarczają wciąż jeszcze ograniczonego wglądu w poszczególne, szczegółowo rozpatrywane, procesy komórkowe. Dlatego też dalsze badania ukierunkowane na wszystkie biocząsteczki występujące w komórce, w tym białka, analizowane z wykorzystaniem nowoczesnych strategii bioanalitycznych, są obecnie w fazie gruntownych rozważań i wielopłaszczyznowej oceny. Jednym z kluczowych aspektów badań proteomicznych jest poszukiwanie strategii analitycznych i bioinformatycznych pozwalających na całkowicie jednoznaczną identyfikację białek. Badania będące przedmiotem niniejszej rozprawy wpisują się w ten obszar zainteresowań proteomiki. Mianowicie, zaprezentowano przykłady wykorzystania chemometrycznej analizy danych podczas przetwarzania danych analitycznych otrzymywanych z użyciem technik rozdzielczych i spektrometrii mas [38-44].

Celem pracy było usprawnienie identyfikacji peptydów w proteomice. Aby zrealizować ten cel, zaproponowano nowatorskie podejścia badawcze umożliwiające, w sposób efektywny i wiarygodny, analizę danych otrzymywanych podczas badań proteomicznych. Wykorzystano przy tym nowe rozwiązania w zakresie rozdzielania peptydów oraz chemometryczną analizę danych.

W pierwszym etapie badań rozważano możliwość wykorzystania informacji analitycznej otrzymanej podczas rozdzieleń peptydów techniką opartą na różnicach pomiędzy ich punktami izoelektrycznymi. Celem pracy było przebadanie możliwości analitycznych i bioinformatycznych frakcjonowania peptydów opartego na ogniskowaniu izoelektrycznym w roztworze. Technika ta rozważana była jako alternatywna metoda rozdzielania mieszanin peptydów w proteomice. W pierwszej kolejności, analiza identyfikacyjna białek przeprowadzana była z wykorzystaniem tandemowej spektrometrii mas z jonizacją przez desorpcję laserową w stałej matrycy z analizatorem czasu przelotu. Do oceny efektywności i dokładności rozdzielenia, opartego na ogniskowaniu izoelektrycznym, połączonego z identyfikacją peptydów za pomocą tandemowych widm masowych, użyto mieszaninę peptydów otrzymaną w wyniku proteolitycznego trawienia osoczowej albuminy wołowej oraz mieszaninę peptydów otrzymaną w wyniku proteolitycznego trawienia pięciu białek (osoczowej albuminy wołowej, albuminy

kurzej, β-laktoglobuliny, mioglobiny i β-kazeiny) [38]. Dodatkowo, celem pracy było przebadanie możliwości identyfikacji białek drożdży piekarniczych (*Saccharomyces cerevisiae*) wykorzystując do rozdzieleń peptydów ogniskowanie izoelektryczne w roztworze sprzężone z chromatografią cieczową w odwróconym układzie faz. W tym przypadku, identyfikację peptydów przeprowadzano korzystając z tandemowego spektrometru mas z jonizacją przez rozpylanie w polu elektrycznym. Zaproponowane podejście oparte na ogniskowaniu izoelektrycznym w roztworze przebadano pod kątem efektywności i dokładności alternatywnej metody ogniskowania izoelektrycznego w roztworze rozważanej jako pierwszy wymiar podczas proteomicznych rozdzieleń dwuwymiarowych do frakcjonowania złożonych mieszanin peptydów. Ponadto, informacja analityczna uzyskiwana podczas analizy tą techniką była przedyskutowany jako dodatkowe narzędzie bioinformatyczne usprawniające identyfikację peptydów w proteomice [39].

W kolejnym etapie badań przedyskutowano zastosowanie sztucznych sieci neuronowych do wiarygodnej oceny widm masowych MS/MS peptydów, rozdzielonych uprzednio z wykorzystaniem wysokosprawnej chromatografii cieczowej w odwróconym układzie faz sprzężonej z tandemowym spektrometrem mas z jonizacją przez rozpylanie w polu elektrycznym. Przeprowadzone zostały szczegółowe rozważania dotyczące zastosowania sztucznych sieci neuronowych do automatycznej klasyfikacji tandemowych widm masowych peptydów. Identyfikowano białka z komórek drożdży piekarniczych (*Saccharomyces cerevisiae*) opierając się na tandemowych widmach masowych uprzednio rozdzielonych peptydów, otrzymanych w wyniku proteolitycznego trawienia białek w początkowym etapie tego eksperymentu. Przeanalizowano możliwości zaprojektowanej sztucznej sieci neuronowej w kontekście automatycznej klasyfikacji, jako „poprawne" lub „niepoprawne", poszczególnych tandemowych widm masowych peptydów w porównaniu do uprzednio przeprowadzonej indywidualnej oceny ich jakości [40].

Ostatni etap badań dotyczył przedyskutowania użyteczności przewidywań czasów retencji peptydów rozdzielanych z wykorzystaniem wysokosprawnej chromatografii cieczowej w odwróconym układzie faz w badaniach proteomicznych. W celu przewidywania retencji chromatograficznej peptydów wyprowadzono odpowiednie ilościowe zależności struktura-retencja. Do ilościowego scharakteryzowania struktury cząsteczkowej peptydów, i następnie, przewidywania gradientowych czasów retencji w danych warunkach chromatograficznych, wykorzystano następujące parametry strukturalne: logarytm sumy czasów retencji aminokwasów budujących dany peptyd, *log* Sum_{AA}, logarytm objętości van der Waalsa danego peptydu, *log* VDW_{Vol}, i logarytm obliczonego współczynnika podziału *n*-oktanol/woda danego peptydu, *clog P*. Pierwszy

z deskryptorów oparty był na danych chromatograficznych otrzymanych dla naturalnie występujących aminokwasów. Dwa pozostałe deskryptory obliczono na podstawie struktury cząsteczkowej peptydów z wykorzystaniem metod modelowania molekularnego [41]. Zaproponowaną strategię przeanalizowano również w aspekcie przewidywania czasów retencji peptydów w odpowiednio scharakteryzowanych, zróżnicowanych układach chromatograficznych [42]. Ponadto strategia ta została wykorzystana podczas porównywania właściwości fizykochemicznych kolumn chromatograficznych, dokonywanego z pomocą chemometrycznej analizy danych retencyjnych rozdzielanych peptydów [43]. Uzupełnieniem i dalszym rozszerzeniem analizy danych, podczas przewidywania retencji HPLC peptydów w proteomice, z wykorzystaniem strategii chemometrycznych, były studia nad wykorzystaniem metody cząstkowych najmniejszych kwadratów z eliminacją zmiennych niewnoszących istotnej informacji [44].

4.1. Frakcjonowanie peptydów w proteomice z wykorzystaniem różnic ich punktów izoelektrycznych

Poszukując alternatywnej, w stosunku do jedno- i dwuwymiarowej elektroforezy żelowej, techniki rozdzieleń, wykorzystywanej w pierwszym etapie analizy proteomicznej, zaproponowano i przetestowano nowatorską technikę ogniskowania izoelektrycznego, przeprowadzanego bezpośrednio w roztworze (ang. *in-solution isoelectric focusing*, sIEF), służącą do frakcjonowania peptydów w oparciu o różnice wartości ich punktów izoelektrycznych (*pI*) [38,39]. Kompleksowej ocenie podlegało zarówno analityczne zastosowanie pod względem frakcjonowania złożonych mieszanin peptydów, jak również możliwość wykorzystania użytecznej bioinformatycznie informacji dotyczącej eksperymentalnie uzyskiwanych wartości punktów izoelektrycznych rozdzielanych peptydów w aspekcie usprawnienia poprawności identyfikacji peptydów i białek. Metoda ogniskowania izoelektrycznego w roztworze została przedstawiona jako alternatywna metoda rozdzieleń w proteomice złożonych mieszanin peptydów otrzymanych z białek trawionych trypsyną. Do frakcjonowania peptydów wykorzystano zminiaturyzowane urządzenie dwunastokomorowe (o objętości każdej z komór wynoszącej 75 μL), wykonane przez autora rozprawy, przeznaczone do ogniskowania izoelektrycznego w roztworze z membranami z żelu poliakryloamidowego utrzymującymi ściśle określony zakres pH w obrębie poszczególnych komór (rycina 3).

Do wstępnej oceny efektywności i dokładności frakcjonowania z wykorzystaniem sIEF połączonego z równoczesną identyfikacją peptydów na podstawie widm MS/MS wykorzystano dwie próbki zawierające mieszaniny peptydów otrzymane z trawionych trypsyną białek

[38]. W pierwszej z próbek były peptydy pochodzące z albuminy wołowej. Druga próbka zawierała mieszaninę peptydów z pięciu białek: albuminy wołowej, albuminy jaja kurzego, β-laktoglobuliny, mioglobiny i β-kazeiny. W pierwszej fazie badań analiza identyfikacyjna peptydów została przeprowadzona z wykorzystaniem tandemowej spektrometrii mas z jonizacją przez desorpcję laserową w stałej matrycy z analizatorem czasu przelotu (MALDI-TOF/TOF-MS). Otrzymywane widma masowe były analizowane z wykorzystaniem programu komputerowego Mascot.

Początkowo dokonywano frakcjonowania mieszaniny peptydów z albuminy wołowej. Badania przeprowadzone dla tej stosunkowo prostej mieszaniny peptydów (zidentyfikowano łącznie 24 peptydy w oparciu o widma MS i MS/MS) dowiodły możliwości dokładnego ogniskowania peptydów głównie w jednej z poszczególnych komór zastosowanej aparatury sIEF w zależności od ich wartości *pI*. Dodatkowo przeprowadzona analiza korelacyjna pomiędzy obliczonymi wartościami punktów izoelektrycznych (pI_{calc}) i oszacowanymi eksperymentalnie (pI_{exp}) wykazała istnienie wysokiej korelacji między tymi wartościami o współczynniku korelacji 0,9697. Dowiedziono, że informacja ta może mieć wartość analityczną, stanowiąc dodatkowy, identyfikacyjny filtr ograniczający (eliminujący) podczas procesu identyfikacji peptydów na podstawie ich widm masowych. W tym przypadku, porównując wyniki identyfikacji otrzymane na podstawie podejścia „peptydowego odcisku palca" (widma MS) w porównaniu do podejście opartego na przeszukiwaniu jonów fragmentarycznych (widma MS/MS), poprawność identyfikacji była potwierdzana znajomością wartości punktów izoelektrycznych dla danych peptydów znajdowanych w poszczególnych komorach urządzenia sIEF. W ten sposób, za pomocą dodatkowego identyfikacyjnego filtru ograniczającego (eliminującego) opartego na punktach izoelektrycznych peptydów zwiększona może być jednoznaczność identyfikacji peptydów.

Następnie, frakcjonowanie za pomocą sIEF sprzężono z frakcjonowaniem z zastosowaniem mikrokolumn ZipTip. Mikrokolumny ZipTip zawierają złoże z fazą stacjonarną pracującą w odwróconym układzie faz i są wykorzystywane zasadniczo do odsalania i zagęszczania próbek przed analizą MALDI-MS. W tym przypadku użyto je również do frakcjonowania peptydów na podstawie różnic w zakresie ich hydrofobowości wykorzystując eluent z rosnącą zawartością rozpuszczalnika organicznego (acetonitrylu). Analizowano próbkę zawierającą mieszaninę peptydów otrzymaną po trawieniu trypsyną pięciu białek (albuminy wołowej, albuminy jaja kurzego, β-laktoglobuliny, mioglobiny i β-kazeiny). Z tym, że obie metody frakcjonowania (zastosowane osobno lub łącznie) wykorzystano do rozdzieleń mieszaniny

peptydów otrzymanych z białek, które występowały w jednej z próbek w tym samym stężeniu, a w drugiej próbce różniły się między sobą w zakresie stężeń obejmujących cztery rzędy wielkości. Identyfikacja z pomocą widm MS/MS, otrzymanych w wyniku analizy MALDI-TOF/TOF-MS, przeprowadzona bez zastosowania wcześniejszego frakcjonowania mieszaniny peptydów z pięciu białek, umożliwiła identyfikację ograniczonej liczby białek na podstawie nielicznych peptydów. Szczególnie w przypadku próbki, w której stężenia białek różniły się między sobą w zakresie czterech rzędów wielkości możliwe było zidentyfikowanie tylko albuminy wołowej, czyli białka występującego w najwyższym stężeniu. Poprawę identyfikacji peptydów i odpowiednich białek, osiągnięto po wcześniejszym zastosowaniu frakcjonowania metodą sIEF. Natomiast kombinacja frakcjonowania za pomocą sIEF z frakcjonowaniem za pomocą mikrokolumn ZipTip przed analizą za pomocą spektrometrii mas umożliwiły identyfikację największej liczby peptydów, wszystkich pięciu białek w próbce pierwszej i czterech białek w próbce drugiej. Kombinacja frakcjonowania opartego na sIEF i mikrokolumnach ZipTip może stanowić w związku z tym prostą, alternatywną metodę analizy próbek zawierających niezbyt skomplikowaną mieszaninę białek, bez potrzeby stosowania chromatograficznych metod rozdzieleń. Strategia ta może być szczególnie użyteczna podczas analiz, w których zastosowana jest spektrometria mas MALDI-MS, i związanych z potwierdzaniem skuteczności całkowitego rozdzielenia białek (np. po uprzednim zastosowaniu elektroforezy żelowej do rozdzielenia ich mieszaniny).

W wyniku przeprowadzonych badań udowodniono, że zaproponowane podejście oparte na ogniskowaniu izoelektrycznym w roztworze może stanowić efektywną, alternatywną metodą frakcjonowania peptydów otrzymywanych z białek trawionych trypsyną. Ponadto, informacja analityczna uzyskiwana podczas frakcjonowania opartego na różnicach w zakresie punktów izoelektrycznych peptydów może być traktowana jako cenny, dodatkowy filtr ograniczający (eliminujący) podczas identyfikacji białek w procesie analizy bioinformatycznej danych proteomicznych. Zaproponowana strategia, oparta na różnicach w zakresie punktów izoelektrycznych peptydów, może stanowić wartościowe analitycznie i bioinformatycznie narzędzie do frakcjonowaniem peptydów w proteomice.

Celem kolejnego etapu pracy było przestudiowanie skuteczności i wiarygodności identyfikacji białek pochodzących z proteomu drożdży piekarniczych (*Saccharomyces cerevisiae*) wykorzystując zaproponowaną metodę frakcjonowania peptydów opartą na ogniskowaniu izoelektrycznym w roztworze (sIEF) sprzężoną z kapilarną wysokosprawną chromatografią cieczową w odwróconym układzie faz (RP-HPLC) i spektrometrią mas [39]. W tym przypadku,

analiza identyfikacyjna peptydów przeprowadzona została z wykorzystaniem tandemowej spektrometrii mas z jonizacją za pomocą rozpylania w polu elektrycznym i analizatorem typu pułapki jonowej (ESI-IT-MS/MS). Otrzymywane widma masowe analizowane były z wykorzystaniem algorytmu Sequest. W przeprowadzonych badaniach dla widm peptydów obdarzonych pojedynczym ładunkiem dodatnim, za poprawne uważano widma charakteryzujące się wartościami korelacji krzyżowej pomiędzy obserwowanym fragmentem widma masowego a widmem teoretycznie przewidzianym (X_{corr}) większymi niż 2,0. Z kolei dla widm peptydów obdarzonych podwójnym ładunkiem dodatnim, za poprawne uważano widma charakteryzujące się wartościami X_{corr} większymi niż 1,5, a dla widm peptydów obdarzonych potrójnym ładunkiem za poprawne uważano widma charakteryzujące się wartościami X_{corr} większymi niż 3,3 [66]. Akceptowane były tylko widma charakteryzujące się wartością różnicy pomiędzy znormalizowanymi wartościami korelacji krzyżowej pomiędzy pierwszym i drugim zidentyfikowanym peptydem (ΔC_n) przekraczającą 0,08. Na podstawie widm masowych typu MS/MS zidentyfikowano wstępnie 851 białek wchodzących w skład analizowanej próbki proteomu drożdży piekarniczych. Następnie, zgodnie z zaleceniami odnośnie właściwej interpretacji widm MS/MS, poddano je indywidualnej ocenie jakości widm uznanych za poprawnie zidentyfikowane, weryfikując ich prawidłowość według zasad zaproponowanych przez Linka i współpracowników [72]. Po tej weryfikacji pozostało 542 zidentyfikowane białka. Niestety tylko 17,2% spośród tych białek było zidentyfikowanych na podstawie większej liczby peptydów niż jeden, co sugerować może dużą liczbę białek zidentyfikowanych fałszywie pozytywnie. Celem zwiększenia wiarygodności uzyskanych identyfikacji postanowiono wprowadzić, proponowany wcześniej, dodatkowy, identyfikacyjny filtr ograniczający (eliminujący) oparty na różnicach w zakresie punktów izoelektrycznych badanych peptydów. Zastosowanie tego filtra spowodowało istotną redukcję w zakresie całkowitej liczby zidentyfikowanych białek wynoszącej w tym momencie 187. Pozwoliło to również na zwiększenie procentu białek zidentyfikowanych na podstawie większej liczby peptydów niż jeden osiągając 26,7%. Ostatecznie, zastosowano jeszcze jeden dodatkowy filtr ograniczający oparty na przeszukiwaniu bazy danych dla białek proteomu drożdży piekarniczych podczas jednoczesnego porównania otrzymanych identyfikacji po wprowadzeniu funkcji określającej zastosowany enzym trawiący i bez określania tego enzymu. Osiągnięto liczbę zidentyfikowanych białek 126, przy czym 39,7% z nich było zidentyfikowanych na podstawie większej liczby peptydów niż jeden.

Podsumowując, można stwierdzić, że zastosowanie metody ogniskowania izoelektrycznego w roztworze (sIEF) jako jednego z etapów rozdzielenia złożonej mieszaniny peptydów z proteomu drożdży piekarniczych, wskazuje na zasadność i istotne korzyści wynikające ze stosowania metody rozdzieleń peptydów w proteomice opartej na różnicach w zakresie ich punktów izoelektrycznych.

4.2. Poprawa indywidualnej oceny jakości widm masowych peptydów w proteomice z wykorzystaniem sztucznych sieci neuronowych

W dotychczasowych badaniach związanych z zastosowaniem metody ogniskowania izoelektrycznego w roztworze (sIEF) przy identyfikacji białek wykorzystywano algorytm Mowse zawarty w programie Mascot (pierwsza część tych badań dotycząca albuminy wołowej i mieszaniny pięciu białek) lub Sequest zawarty w programie BioWorks (druga część badań obejmująca analizę proteomu drożdży piekarniczych). Podczas analizy widm MS/MS przy identyfikacji proteomu drożdży piekarniczych, zgodnie z zaleceniami odnośnie właściwej interpretacji widm MS/MS, poddano je wstępnie standardowej procedurze indywidualnej oceny jakości widm MS/MS uznanych za poprawnie zidentyfikowane [72]. Procedurę tę przeprowadzono zgodnie z zaleceniami odnośnie właściwej interpretacji widm masowych zaproponowanymi przez Linka i współpracowników [69]. Obejmuje ona w szczególności weryfikację poprawności każdego z widm MS/MS, określanych poprzez algorytm Sequest jako identyfikacja odpowiednich peptydów.

W 1994 roku opracowany został algorytm korelacji krzyżowej (ang. *cross-corellation algorithm*) do identyfikacji białek na podstawie odpowiednich bioinformatycznych baz danych z wykorzystaniem informacji dla peptydów otrzymanej za pomocą tandemowej spektrometrii masowej [89-91]. Otrzymany na jego podstawie algorytm Sequest zawarty w programie komputerowym BioWorks stanowi obecnie integralną część kompleksowych, komercyjnie dostępnych platform rozdzielczo-identyfikacyjnych typu LC/MS/MS służących w badaniach proteomicznych. Program ten jest ciągle udoskonalany i, obok programu Mascot, jest jednym z najczęściej wykorzystywanych programów do identyfikacji białek w praktyce proteomicznej. Podstawową cechą algorytmu Sequest jest założenie, że sekwencja aminokwasowa peptydów może być zdefiniowane za pomocą tandemowego widma masowego (MS/MS). Zastosowany algorytm automatyzuje proces oceny widm dopasowując widma dostępne w bazie danych do widm eksperymentalnie otrzymanych. Na początku, sekwencje aminokwasowe są

szybko oceniane za pomocą wstępnego algorytmu numerycznego, co ułatwia eliminację sekwencji niepoprawnych. Następnie, stosowany jest bardziej zaawansowany algorytm korelacji krzyżowej, który ocenia bardziej szczegółowo widma otrzymane eksperymentalnie z widmami teoretycznymi [92]. Wykorzystywanie algorytmu Sequest jest ściśle związane z odpowiednią interpretacją danych parametrycznych z analizy bioinformatycznej. Parametry te warunkują identyfikację peptydów, dla których dopasowane zostały widma eksperymentalne [91]. Zestaw informacji statystycznej ułatwiać ma przy tym klasyfikację zidentyfikowanych peptydów. Wstępnie wykorzystywana jest różnica pomiędzy znormalizowanymi wartościami korelacji krzyżowej (ΔC_n) pomiędzy pierwszym i drugim zidentyfikowanym peptydem. Służy ona do wskazania poprawności wyboru sekwencji aminokwasowej. Następnie, za pomocą dodatkowych parametrów, jak wartości korelacji krzyżowej pomiędzy obserwowanym fragmentem widma masowego a widmem teoretycznie przewidzianym (X_{corr}), wartości wstępnego rankingu liczby jonów w teoretycznym widmie MS/MS, które znajdują się w widmie eksperymentalnym (S_p), rankingu dopasowania widm (RS_p), wartości liczby jonów (I) w teoretycznym widmie MS/MS, które znajdują się jednocześnie w widmie eksperymentalnym, dokonywana jest dalsza korekta poprawności identyfikacji. Jako podstawowy filtr ograniczający (eliminujący) służą w szczególności wartości korelacji krzyżowej pomiędzy obserwowanym fragmentem widma masowego a widmem teoretycznie przewidzianym (X_{corr}), wyznaczane oddzielnie dla peptydów o określonej wartości ładunku dodatniego na cząsteczce. W przeprowadzonych badaniach [40] dla widm peptydów obdarzonych pojedynczym ładunkiem dodatnim, za poprawne uważano widma charakteryzujące się wartościami X_{corr} większymi niż 2,0. Z kolei dla widm peptydów obdarzonych podwójnym ładunkiem dodatnim, za poprawne uważano widma charakteryzujące się wartościami X_{corr} większymi niż 1,5, a dla widm peptydów obdarzonych potrójnym ładunkiem za poprawne uważano widma charakteryzujące się wartościami X_{corr} większymi niż 3,3 [72]. Akceptowane były tylko widma charakteryzujące się wartością ΔC_n przekraczającą 0,08.

Rekomendowana jest także dodatkowo indywidualna ocena jakości widm MS/MS peptydów uznanych za poprawnie zidentyfikowane. Tu właśnie włącza się procedurę zaproponowaną przez Linka i współpracowników [72], dotyczącą właściwej oceny widm masowych i ostatecznie weryfikacji ich „dobroci". Według tych kryteriów rozpatruje się, czy jakość widma jest odpowiednia dla jednoznacznej wizualizacji jonów fragmentarycznych powyżej linii bazowej widma. Ponadto, potwierdza się spójność kontynuacji występujących serii *b* i serii *y* jonów oraz ocenia się, czy jony odpowiadające resztom proliny są wystarczająco intensywne.

Ostatecznie, nieidentyfikowalne, intensywne jony fragmentaryczne powinny odpowiadać jonom fragmentarycznym obdarzonym ładunkiem dodatnim, bądź utracie jednego lub dwóch aminokwasów z końców danego peptydu.

Ostatnio został opisany i zastosowany algorytm SVM (ang. *support vector machine*), którego celem jest szybsza i dokładniejsza ocena wyników uzyskanych z wykorzystaniem algorytmu Sequest [91]. Polega na rozróżnianiu poprawnie i niepoprawnie zidentyfikowanych peptydów w oparciu o subtelne różnice w obrębie złożonego zestawu danych wejściowych. Nie jest on jednak wciąż w stanie wyeliminować etapu czasochłonnej i pracochłonnej, indywidualnej oceny jakości każdego z widm MS/MS.

Do rozwiązania problemu indywidualnej oceny jakości widm MS/MS peptydów zaproponowano nowatorskie podejście, którego podstawę stanowią sztuczne sieci neuronowe (ang. *artificial neural networks*, ANN). Sztuczne sieci neuronowe są metodą analizy danych, która odzwierciedlać ma sposób pracy mózgu. Różnią się one od klasycznych programów komputerowych tym, że mają zdolność „uczenia się", a informacja jest zakodowana w sile połączeń „synaptycznych" sieci [93-95]. W zaproponowanym podejściu [40] wykorzystane zostały dane uzyskane z tandemowej spektrometrii mas ESI-MS/MS dla peptydów otrzymanych podczas trawienia trypsyną zespołu białek pochodzących z komórek drożdży piekarniczych (*Saccharomyces cerevisiae*), rozdzielonych, w poprzednim etapie badań, za pomocą ogniskowania izoelektrycznego w roztworze (sIEF) i kapilarnej wysokosprawnej chromatografii cieczowej w odwróconym układzie faz (RP-HPLC). Do sprawdzenia przydatności sztucznych sieci neuronowych wszystkie widma MS/MS musiały być początkowo poddane indywidualnej ocenie ich jakości.

Celem pracy było udowodnienie, że na podstawie widm, dla których przeprowadzono indywidualnę ocenę jakości widm MS/MS peptydów uznanych za poprawnie lub niepoprawnie zidentyfikowane, zebranych w zbiorze uczącym i walidacyjnym, możliwe jest zaprojektowanie odpowiedniej sztucznej sieci neuronowej pozwalającej na klasyfikację widm zebranych w zbiorze testowym na „poprawne" i „niepoprawne" dokonywaną w sposób zautomatyzowany, analogicznie do klasycznie przeprowadzanej indywidualnej oceny ich jakości. Poprawność lub brak poprawności identyfikacji peptydów była oceniana na podstawie kryteriów opublikowanych przez Linka i współpracowników [72]. Każdy peptyd charakteryzowany był za pomocą zestawu zmiennych, niezbędnych do ich rozpoznawania przez sztuczną sieć neuronową jako „poprawny" lub „niepoprawny". Poprawność została potwierdzona poprzez indywidualną ocenę jakości widm MS/MS zgodnie w powyższymi założeniami. Wśród zmiennych

opisujących peptydy znalazły się wartości parametrów X_{corr}, ΔC_n, S_p i *RSp* z algorytmu Sequest, oraz liczba charakteryzująca ładunek, jakim obdarzony jest dany peptydu (*CH*), masa cząsteczkowa peptydu (*MW*), i parametry obliczone dla poszczególnych peptydów, takie jak hydrofobowość (*H*), wartość punktu izoelektrycznego (*pI*) i wartość nachylenia funkcji opisującej relacje między *H* i *z*, *dz*/*dpH*, gdzie *z* to ładunek. W wyniku przeprowadzonej analizy zaprojektowano i zastosowano sztuczną sieć neuronową o architekturze perceptronu wielowarstwowego, składającego się z 10 neuronów w warstwie wejściowej, 23, 10 i 7 neuronów w odpowiednio trzech kolejnych warstwach ukrytych oraz jednego neuronu w warstwie wyjściowej. Wykorzystano nadzorowaną metodę uczenia stosując algorytm wstecznej propagacji błędu i algorytm gradientów sprzężonych. Rejestrowano zmiany błędu średniokwadratowego (ang. *root mean square*, RMS) dla zbioru uczącego i walidacyjnego podczas procesu uczenia. Ostateczne wykorzystano sieć charakteryzującą się najmniejszą wartością RMS.

Dane wejściowe zostały podzielone w sposób losowy w programie komputerowym Statistica Neural Networks (StatSoft, Tuls, OK, USA) na zbiór uczący, walidacyjny i testujący. Analizie poddano łącznie 2094 widm MS/MS peptydów. W zbiorze uczącym było 1048 widm peptydów, a w zbiorze walidacyjnym i zbiorze testującym po 523 widm peptydów. Spośród 704 widm dla peptydów przypisanych jako „poprawne", w zbiorze uczącym wybrana sztuczna sieć neuronowa 588 widm sklasyfikowała właściwie. Natomiast właściwie przypisanych widm w zbiorze uczącym dla danych oznaczonych wstępnie jako „niepoprawne" było 281 na 344 rozważanych widm. Bardzo podobne wyniki otrzymano dla zbioru walidacyjnego. Spośród 368 widm „poprawnych" i 155 widm "niepoprawnych", odpowiednio 282 widma i 113 widm było zaklasyfikowanych właściwie. W przypadku zbioru testującego, na 362 widma właściwie sklasyfikowane były 272 widma, oznaczone jako „poprawne". Natomiast 130 widm „niepoprawnych" zostało sklasyfikowanych właściwie spośród 161 widm rozważanych w tym zbiorze. Stanowi to odpowiednio 75,14% i 80,75% poprawności klasyfikacji uzyskanych zaprojektowaną i zastosowaną siecią neuronową. Z praktycznego punktu widzenia, wyniki te należy wciąż traktować pilotowo, nie mniej jednak skuteczność klasyfikacji można uznać za zadowalającą i obiecującą w perspektywie dalszego zastosowania zaproponowanej strategii. W przypadku zastosowania innego zbioru danych uczących i walidacyjnych oraz zaprojektowania architektury odpowiedniej sieci neuronowej, zaproponowana strategia przyczynić się może do zoptymalizowania zautomatyzowanego procesu oceny jakości widm MS/MS w analizie proteomicznej.

Równolegle z podstawową statystyką poprawności przewidywań zaprojektowanej sieci neuronowej przeprowadzono analizę wrażliwości (ang. *sensitivity analysis*) dla danych wejściowych. Pozwala ona na rozróżnienie ważnych zmiennych od takich, które niewiele wnoszą do wyniku działania sieci. Analiza wrażliwości daje wgląd w użyteczność poszczególnych zmiennych wejściowych. Wskazuje zmienne, które, bez straty jakości przewidywanej sieci mogą być pominięte i zmienne kluczowe, których nie należy pomijać. Analiza ta pozwoliła na klasyfikację zmiennych pod względem ich użyteczności w kontekście automatycznej klasyfikacji widm MS/MS na „poprawne" i „niepoprawne". Zidentyfikowane parametry najbardziej istotne to S_p, oraz w nieco mniejszym stopniu X_{corr}, *MW*, *I* oraz ΔC_n. Wyraźnie mniej istotne dla właściwej klasyfikacji widm okazały się RS_p, *CH*, *H*, *pI* i *dz/dpH*. W celach porównawczych, oprócz analizy wrażliwości przeprowadzonej w ramach analizy sztucznych sieci neuronowych, dane ze zbioru uczącego poddano także analizie za pomocą testu Fishera [91]. Wykazano, że najbardziej użytecznym predykcyjnie parametrem jest X_{corr}, a kolejnymi istotnymi parametrami są S_p i *I*. Dopiero na kolejnych miejscach znalazły się ΔC_n i RS_p. Podobnie jak poprzednio *CH*, *H*, *pI*, i *dz/dpH* okazały się najmniej istotne. Pomimo pewnych różnic w ocenie uzyskanej z wykorzystaniem obu metod, jednoznaczna jest przewaga parametrów pochodzących z algorytmu Sequest. Nie mniej jednak, obie analizy potwierdzają również fakt dużego znaczenia parametru X_{corr} podczas oceny jakości widm MS/MS, co znalazło swoje odzwierciedlenie w literaturze poświęconej kryteriom klasyfikacyjnym widm w kontekście identyfikacji peptydów i białek [40].

Podsumowując, można stwierdzić, że odpowiednio wyuczona i zwalidowana sztuczna sieć neuronowa wykazuje czułość i specyficzność odnośnie dokładnego i wydajnego analizowania widm masowych MS/MS peptydów pod względem ich „dobroci". Zaproponowana strategia oparta na wykorzystaniu sztucznych sieci neuronowych dostarcza przewidywań mogących w sposób wiarygodny wskazywać, czy dane widmo MS/MS może być traktowane jako „poprawne" lub „niepoprawne". Poprzez to, zredukować można potrzebę indywidualnej oceny jakości ogromnej liczby widm MS/MS zwykle rozważanych w praktyce proteomicznej.

4.3. Przewidywanie retencji chromatograficznej peptydów z wykorzystaniem ilościowych zależności struktura-retencja (QSRR) do celów analizy proteomicznej

Nierozwiązanym dotychczas w sposób satysfakcjonujący problemem analizy proteomicznej jest pełne wykorzystanie w celach bioinformatycznych informacji analitycznej uzyskiwanej

podczas rozdzielania peptydów z pomocą technik chromatograficznych, w szczególności wysokosprawnej chromatografii cieczowej w odwróconym układzie faz (RP-HPLC). Dla danego eksperymentu chromatograficznego (określona faza ruchoma i faza stacjonarna, temperatura, pH itd.), czas retencji jest parametrem charakterystycznym dla danego analitu. W połączeniu z informacją uzyskiwaną z widm MS/MS, przewidywanie czasu retencji dla danej struktury peptydu może być użyte w celu usprawnienia identyfikacji peptydów i zwiększenia liczby poprawnie zidentyfikowanych peptydów. Co więcej, przy odpowiednio wysokiej dokładności pomiarów wartości mas cząsteczkowych, przewidywania retencji peptydów mogą również ograniczyć potrzebę uwzględniania danych z widm MS/MS [95].

Zasadniczo wszystkie prowadzone dotychczas próby oszacowania retencji peptydów, w tym także dla celów analizy proteomicznej [95-105], oparte były na prostych zależnościach wynikających ze znajomości składu aminokwasowego danego peptydu. Zaproponowana nowatorska procedura przewidywania retencji chromatograficznej peptydów wykorzystuje ilościowe zależności struktura-retencja (QSRR) [41,106]. Równania QSRR wyprowadzone były z użyciem wieloparametrowej analizy regresji. Do ilościowej charakterystyki struktury molekularnej peptydów wykorzystano następujące deskryptory: logarytm z sumy czasów retencji aminokwasów budujących dany peptyd, $log\ Sum_{AA}$, logarytm z objętości van der Waalsa, $log\ VDW_{Vol}$ oraz logarytm ze współczynnika podziału *n*-oktanol/woda, *clog P* [41]:

$$t_R = k_1 + k_2\ log\ Sum_{AA} + k_3\ log\ VDW_{Vol} + k_4\ clog\ P \qquad \text{Równanie 1}$$

gdzie: t_R to czas retencji peptydu rozdzielanego w elucji gradientowej RP-HPLC, a k_1-k_4 to współczynniki równania regresji.

Pierwszy z parametrów strukturalnych otrzymywany był eksperymentalnie, natomiast dwa pozostałe były obliczane na podstawie modelowania molekularnego. Podczas gdy parametr $log\ Sum_{AA}$ odnosi się do udziału retencji poszczególnych aminokwasów w ogólnej retencji peptydu, parametry $log\ VDW_{Vol}$ i *clog P* należy traktować jako parametry korekcyjne podwyższające wiarygodność przewidywania retencji peptydu. Dodatkowo, deskryptory $log\ VDV_{Vol}$ i *clog P* odzwierciedlają różnice w retencji dla tych samych sekwencji aminokwasowych, spowodowane wypadkową struktury peptydu oraz modyfikacjami potranslacyjnymi. Przewidywanie retencji RP-HPLC peptydów za pomocą QSRR zostało przetestowane na zróżnicowanej strukturalnie grupie peptydów oraz w zróżnicowanych eksperymentalnych warunkach procesu HPLC.

W pierwszym etapie badań przeprowadzono eksperyment dla wybranych 35 peptydów, które potraktowano jako zbiór modelowy do wyprowadzenia równania QSRR, charakteryzującego dany układ chromatograficzny. Po zastosowaniu elucji gradientowej z czasem gradientu równym 20 min uzyskane równanie QSRR miało następującą formę [41]:

$$t_R = 7{,}52\ (\pm\ 3{,}12) + 15{,}24\ (\pm\ 1{,}54)\ log\ Sum_{AA} - 5{,}83\ (\pm\ 1{,}84)\ log\ VDW_{Vol} +$$
$$p = 0{,}022 \qquad p = 4\text{x}10^{-11} \qquad p = 0{,}003$$

$$+ 0{,}26\ (\pm\ 0{,}08)\ clog\ P \qquad \text{Równanie 2}$$
$$p = 0{,}004$$
$$n = 35;\ R = 0{,}966;\ F = 144;\ s = 1{,}06;\ p < 3\text{x}10^{-18}$$

W przypadku zastosowanego układu chromatograficznego charakterystyka t_R poprzez zaproponowane parametry strukturalne jest satysfakcjonująca. Wszystkie współczynniki przy trzech deskryptorach strukturalnych są istotne statystycznie ($p \leq 0{,}003$). Istotne statystycznie jest również całe równanie ($p = 3\text{x}10^{-18}$). Współczynnik korelacji, R, oraz wartość testu F są wysokie, a standardowy błąd estymacji, s, odpowiednio niewielki.

Wykorzystując równanie 2 można było w kolejnym etapie przewidywać czasy retencji dla pozostałych 66 peptydów niewykorzystywanych do wyprowadzenia modelu QSRR. Otrzymano wysoką korelację pomiędzy czasami retencji uzyskanymi eksperymentalnie i obliczonymi z wykorzystaniem odpowiedniego równania QSRR, wynoszącą 0,963. Średnia różnica pomiędzy czasami retencji uzyskanymi eksperymentalnie i obliczonymi z wykorzystaniem odpowiedniego równania QSRR wynosiła 0,76 min. Przy tym, czasy retencji analizowanych peptydów znajdowały się w szerokim zakresie wartości: od 2,32 min to 17,72 min.

Wyniki uzyskane podczas tych badań były obiecujące i w następnym etapie przeprowadzono rozszerzone studia na temat przewidywań retencji peptydów [42,43]. W pierwszym rzędzie przeprowadzono badania w różnych warunkach chromatograficznych. Następnie przetestowano ogólny wpływ innych parametrów strukturalnych na retencję peptydów w odniesieniu do zastosowanych w wyprowadzonym równaniu QSRR oraz w aspekcie klasyfikacji kolumn chromatograficznych stosowanych do rozdzieleń peptydów.

Na początku, korzystając z 98 peptydów, wyprowadzono odpowiednie równania QSRR dla układów chromatograficznych różniących się zastosowaną kolumną chromatograficzną, czasem gradientu i temperaturą kolumny [42]. Otrzymano 18 istotnych statystycznie równań

QSRR, charakteryzujących poszczególne układy chromatograficzne. Umożliwiło to przeprowadzenie analizy porównawczej przydatności zastosowanej strategii w testowanych układach chromatograficznych podczas przewidywań retencji RP-HPLC peptydów. Generalnie, najlepsze korelacje pomiędzy eksperymentalnymi i przewidywanymi czasami retencji otrzymano w przypadku mniej polarnych kolumn chromatograficznych (XTerra, LiChrospher RP-18, PLRP-S i Chromolith). Wyniki mniej satysfakcjonujące odnotowano dla bardziej polarnych kolumn (Discovery Amide C16, LiChrospher CN i Discovery HS FS-3). W kontekście badania oddziaływań międzycząsteczkowych, popartego rozważaniami nad parametrami strukturalnymi analitów, jest to wynik potwierdzający ogólnie większą trudność ilościowego wyjaśnienia specyficznych oddziaływać polarnych w porównaniu do oddziaływań niespecyficznych mającymi przewagę w przypadku niepolarnych faz stacjonarnych [63,106].

Jakość przewidywania retencji była również rozważana w zależności od zastosowanego czasu gradientu podczas analizy. Porównawcza analiza przewidywań przy czasie gradientu równym 20, 60 i 120 min została przeprowadzona z wykorzystaniem wyników otrzymanych na kolumnie LiChrospher RP-18. Zaobserwowano wzrost błędu średniokwadratowego obliczonego z zastosowaniem walidacji krzyżowej (ang. *cross-validated root mean square error*, RMSECV) z wartościami 0,98 min, 2,91 min i 7,06 min przy wzroście długości czasu gradientu, t_G, od odpowiednio 20 min, poprzez 60 min i kończąc na 120 min. Równocześnie nie odnotowano znaczącego spadku korelacji pomiędzy czasami retencji uzyskanymi eksperymentalnie i obliczonymi z wykorzystaniem danego równania QSRR: $R = 0{,}964$ ($t_G = 20$ min), $R = 0{,}951$ ($t_G = 60$ min) i $R = 0{,}913$ ($t_G = 120$ min).

Nie zauważono wyraźnych różnic w zakresie czasów retencji podczas chromatografowania peptydów w układach chromatograficznych różniących się temperaturą kolumny (40, 60 lub 80 °C) podczas przewidywania retencji za pomocą QSRR. Jedynie niewielkie zmniejszenie korelacji pomiędzy czasami retencji uzyskanymi eksperymentalnie i obliczonymi z wykorzystaniem odpowiedniego równania QSRR wraz ze wzrostem temperatury zaobserwowano w przypadku kolumny PLRP-S.

W kolejnym etapie badań, posługując się danymi retencyjnymi otrzymanymi dla serii zróżnicowanych fizykochemicznie peptydów w różnych układach chromatograficznych, przeprowadzono dodatkową ocenę chemometryczną zależności pomiędzy retencją chromatograficzną a deskryptorami strukturalnymi peptydów [43]. Ocenę tę wykonano z użyciem multiwariacyjnej metody analizy danych, włączającą analizę głównych składowych (ang. *principal component analysis*, PCA). Spośród 44 parametrów strukturalnych charakteryzujących

w sposób ilościowy 98 testowanych peptydów, 17 stanowiły logarytmy z sumy czasów retencji aminokwasów budujących dany peptyd, *log* Sum_{AA}, otrzymane dla wszystkich analizowanych układów chromatograficznych. Ten parametr, w wyniku przeprowadzonej analizy PCA, okazał się mieć największy wpływ na retencję peptydów. Rozważając udział innych deskryptorów w retencji peptydów, w kontekście przewidywania retencji z wykorzystaniem ilościowych zależności struktura-retencja, dowiedziono, że oprócz parametru *log* Sum_{AA}, istotne znaczenie mają także parametry strukturalne opisujące właściwości powierzchni i objętości cząsteczek peptydów (jak np. *log* VDW_{Vol}) oraz ich hydrofobowości (jak np. *clog P*). Na podstawie przeprowadzonej analizy potwierdzono dodatkowo słuszność wyboru zastosowanych uprzednio deskryptorów podczas przewidywań retencji za pomocą QSRR. Dodatkowo, wykorzystując analizę PCA, przeprowadzono klasyfikację testowanych kolumn chromatograficznych. Umożliwiło to ocenę podobieństw i różnic tych kolumn w zakresie rozdzieleń peptydów. Zaproponowane podejście zostało przedyskutowane jako prosta metoda porównywania kolumn chromatograficznych w celu usprawnienia identyfikacji ortogonalnych warunków HPLC podczas wielowymiarowych rozdzieleń złożonych mieszanin peptydów w analizie proteomicznej.

Ostatni etap badań obejmował analizę ilościowych zależności struktura-retencja przeprowadzonych z wykorzystaniem danych retencyjnych dla 90 peptydów otrzymanych z wykorzystaniem elucji gradientowej HPLC i wartości 1726 deskryptorów strukturalnych opisujących każdy z peptydów [44]. Rozważano duży zbiór deskryptorów strukturalnych reprezentujących szeroki wachlarz właściwości fizykochemicznych analitów, które mogą być potencjalnie użyteczne podczas budowania modelu QSRR. Etapem o decydującym znaczeniu podczas tego podejścia był obiektywny wybór istotnych statystycznie deskryptorów strukturalnych spośród danego zbioru deskryptorów. Celem niniejszych badań było przeanalizowanie możliwości efektywnej selekcji zmiennych przydatnych do analizy QSRR wykorzystując metodę cząstkowych najmniejszych kwadratów z eliminacją zmiennych niewnoszących istotnej informacji (ang. *uninformative variable elimination by partial least squares*, UVE-PLS). Do wyprowadzenia odpowiedniego modelu QSRR wykorzystano 63 peptydy. 27 peptydów stanowiło niezależną grupę testową służącą do oceny siły predykcyjnej zaproponowanego modelu. Nowy model QSRR, oparty na analizie UVE-PLS, wykazywał dobre właściwości przewidywania retencji peptydów niewykorzystanych uprzednio podczas konstruowania tego modelu. Wartość błędu średniokwadratowego przewidywań (ang. *predictive root mean square error*,

RMSEP) wynosiła 0,454 dla modelu z pięcioma zmiennymi ukrytymi obejmującymi 128 deskryptorów wyselekcjonowanych wstępnie z wykorzystaniem metody UVE-PLS.

Stwierdzono, że wyniki osiągnięte z użyciem metody UVE-PLS były lepsze w porównaniu do klasycznej metody PLS. Oprócz większej siły predykcyjnej, model UVE-PLS był również mniej złożony (zawierając pięć zmiennych ukrytych, w porównaniu do siedmiu w metodzie PLS). Zastosowanie metody cząstkowych najmniejszych kwadratów z eliminacją zmiennych niewnoszących istotnej informacji umożliwiło efektywną selekcję najbardziej istotnych statystycznie zmiennych w liczbie 128 spośród wstępnie rozważanych 1726 deskryptorów. Metoda UVE-PLS może być również rozważana jako metoda alternatywna w stosunku do metody krokowej regresji wieloparametrowej (ang. *stepwise regression*). Jest to spowodowane faktem wykazania możliwości efektywnej selekcji istotnych statystycznie deskryptorów strukturalnych w połączeniu z dużą siłą predykcyjną wyprowadzonego modelu z błędem przewidywań wynoszącym mniej niż 30 s.

Otrzymane wyniki [41-44] wskazują, że zaproponowana strategia identyfikacyjna peptydów oparta na analizie QSRR służącej do przewidywania retencji chromatograficznej, może być rozważana jako nowatorskie narzędzie umożliwiając pełniejsze wykorzystanie informacji analitycznej gromadzonej podczas rozdzieleń HPLC mieszanin peptydów przeprowadzanych w trakcie analizy proteomicznej. Dowiedziono, że przewidywanie retencji peptydów oparte na ilościowej charakterystyce peptydów z wykorzystaniem deskryptorów z modelowania molekularnego może stanowić nowy filtr ograniczający (eliminujący) w proteomice.

5. Podsumowanie

Podstawowym celem przeprowadzonych badań było przedyskutowanie możliwości usprawnienia identyfikacji peptydów w proteomice z wykorzystaniem chemometrycznej analizy danych. Kluczem do powstania nowych narzędzi umożliwiających bardziej efektywną analizę proteomiczną było rozwinięcie i zastosowanie zoptymalizowanych pod kątem analizy proteomicznej odpowiednich strategii rozdzielczych i chemometrycznych. Podczas tych badań dowiedziono przydatności opracowanej nowatorskiej metody rozdzieleń złożonych mieszanin peptydów opartej na izoelektrycznym ogniskowaniu w roztworze [38] oraz w sprzężeniu z mikrokapilarną wysokosprawną chromatografią cieczową podczas analizy próbek proteomicznych, w tym także do analizy proteomu drożdży piekarniczych (*Saccharomyces cerevisiae*) [39]. Do rozwiązania problemu interpretacji widm typu MS/MS peptydów zaproponowano nowatorskie podejście, którego podstawę stanowi wykorzystanie sztucznych sieci neuronowych. Wykorzystane zostały dane z tandemowej spektrometrii masowej dla peptydów i dowiedzione zostało, że odpowiednio zaprojektowana sztuczna sieć neuronowa jest w stanie usprawnić i ograniczyć proces indywidualnej oceny jakości widm masowych typu MS/MS dla peptydów w proteomice [40]. Zaprezentowany w kolejnej pracy algorytm oparty na ilościowych zależnościach struktura-retencja (QSRR) wykorzystano do przewidywania retencji chromatograficznej HPLC peptydów w odwróconym układzie faz metodą elucji gradientowej w badaniach proteomicznych [41]. W oparciu o parametry strukturalne przeprowadzone były także przewidywania retencji dla serii strukturalnie zróżnicowanych peptydów w warunkach gradientowych w scharakteryzowanych za pomocą modeli QSRR, różnych pod względem fizykochemicznym układach HPLC [42]. Na podstawie przeprowadzonej chemometrycznej analizy danych z wykorzystaniem analizy głównych składowych potwierdzono słuszność wyboru zastosowanych deskryptorów do przewidywania retencji za pomocą QSRR. Dodatkowo, przeprowadzono klasyfikację testowanych kolumn chromatograficznych w aspekcie poszukiwania podobieństw i różnic w zakresie rozdzieleń peptydów [43]. Ostatecznie, retencję peptydów skutecznie modelowano z wykorzystaniem metody cząstkowych najmniejszych kwadratów z eliminacją zmiennych niewnoszących istotnej informacji (UVE-PLS) [44].

Proteomika, jako jedna z dziedzin współczesnej biologii molekularnej, jest związana z osiągnięciami ostatnich lat związanymi z ustalaniem sekwencji genomów organizmów żywych. Dynamicznie rozwijająca się proteomika integruje także szereg innych dyscyplin

naukowych, wliczając szybkie i czułe techniki rozdzieleń białek, bioinformatykę, inżynierię materiałową, metody krystalograficzne i spektroskopowe [107]. Nie należy również zapominać o istnieniu i możliwościach poznawczych genomiki, transkryptomiki, metabolomiki i innych nowoczesnych technologii biomedycznych. Cała wiedza o funkcjonowaniu organizmów żywych staje się więc interdyscyplinarną, szeroko pojętą systeomiką [108]. Analizowanie całkowitych zespołów białkowych oraz identyfikacja i monitorowanie składników wielu kompleksów białkowych zaangażowanych w istotne procesy komórkowe, będących domeną proteomiki, będzie możliwe dzięki złożonym, zintegrowanym platformom analityczno-bioinformatycznym [8,11,78,108,109]. Wykorzystanie w pełni możliwości proteomiki zależeć będzie nie tylko od umiejętności opracowania i zastosowania nowych narzędzi analitycznych i bioinformatycznych pozwalających w sposób szybki i dokładny analizować tysiące białek występujących w komórkach organizmów żywych, lecz również od umiejętności zastosowania chemometrycznej analizy danych pozwalającej na bardziej efektywną i wiarygodną ich identyfikację.

6. Bibliografia

[1] D. Figeys, Proteomics approaches in drug discovery, *Anal. Chem.*, 74 (2004) 413A-419A.

[2] G.C. Terstappen, A. Reggiani, *In silico* research in drug discovery, *Trends Pharm. Sci.*, 22 (2001) 23-26.

[3] J. Drews, Genomic sciences and the medicine of tomorrow, *Nat. Biotechnol.*, 14 (1996) 1516-1518.

[4] J. Drews, S. Ryser, The role of innovation in drug development, *Nat. Biotechnol.*, 15 (1997) 1318-1319.

[5] S. Grabley, R. Thiericke (red.), *Drug Discovery from Nature*, Springer-Verlag, Berlin, 1999.

[6] A. Abbott, A post-genomic challenge: learning to read patterns of protein synthesis, *Nature*, 402 (1999) 715-720.

[7] K.A. Cole, D.B. Krizman, M.R. Emmert-Buck, The genetics of cancer – a 3D model, *Nat. Genet. Suppl.*, 21 (1999) 38-41.

[8] T. Bączek, Proteomika a nowotwory, *Farmacja Polska*, 62 (2006) 12-19.

[9] J.L. Walgren, D.C. Thompson, Application of proteomic technologies in the drug development process, *Toxicol. Lett.*, 149 (2004) 377-385.

[10] G. Stix, Parsing cells, *Sci. Am.*, 281 (1999) 35-36.

[11] T. Bączek, R. Kaliszan, *Nowoczesne techniki analityczne w proteomice*, w: *Miniaturyzacja w analityce* (red. Z. Brzózka), Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, 2005.

[12] M.R. Wilkins, J.C. Sanchez, A.A. Gooley, R.D. Appel, I. Humphery-Smith, D.F. Hochstrasser, K.L. Williams, Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it, *Biotechnol. Genet. Eng. Rev.*, 13 (1996) 41-50.

[13] V.C. Wasinger, G.L. Corthals, Proteomic tools for biomedicine, *J. Chromatogr. B*, 771 (2002) 33-49.

[14] S. Fields, Proteomics in genomeland, *Science*, 291 (2001) 1221-1224.

[15] D.C. Liebler, *Introduction to Proteomics*, Humana Press, Totowa, NJ, 2002.

[16] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, *Nature*, 409 (2001) 860-921.

[17] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural *et al.*, The sequence of the human genome, *Science*, 291 (2001) 1304-1351.

[18] F.S. Collins, E.S. Lander, J. Rogers, R.H. Waterson, Finishing the euchromatic sequence of the human genome, *Nature*, 431 (2004) 931-945.

[19] A. Pandey, M. Mann, Proteomics to study genes and genomes, *Nature*, 405 (2000) 837-846.

[20] M. Caron, N. Imam-Sghiouar, F. Poirier, J-.P. Le Caër, V. Labas, Raymonde Joubert-Caron, Proteomic map and database of lymphoblastoid proteins, *J. Chromatogr. B*, 771 (2002) 197-209.

[21] T. Nakanishi, R. Koyama, T. Ikeda, A. Shimizu, Catalogue of soluble proteins in the human vitreous humor: comparison between diabetic retinopathy and macular hole, *J. Chromatogr. B*, 776 (2002) 89-100.

[22] I. Noël-Georis, A. Bernard, P. Falmagne, R. Wattiez, Database of bronchoalveolar lavage fluid proteins, *J. Chromatogr. B*, 771 (2002) 221-236.

[23] A. Sickmann, W. Dormeyer, S. Wortelkamp, D. Woitalla, W. Kuhn, H.E. Meyer, Towards a high resolution separation of human cerebrospinal fluid, *J. Chromatogr. B*, 771 (2002) 167-196.

[24] M.J. Hubbard, J.C. Kon, Proteomic analysis of dental tissues, *J. Chromatogr. B*, 771 (2002) 211-220.

[25] N.L. Anderson, N.G. Anderson, The human plasma proteome, *Mol. Cell Proteomics*, 1 (2002) 845-867.

[26] J.N. Adkins, S.M. Varnum, K.J. Auberry, R.J. Moore, N.H. Angell, R.D. Smith, D.I. Springer, J.G. Pounds, Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry, *Mol. Cell Proteomics*, 1 (2002) 947-955.

[27] M. Tichá, V. Pacáková, K. Stulik, Proteomics of allergens, *J. Chromatogr. B*, 771 (2002) 343-353.

[28] R. Kellner, F. Lottspeich, H.E. Meyer, *Microcharacterization of Proteins*, Wiley-VCH, Weinheim, 1999.

[29] T. Wehr, Separation technology in proteomics, *LCGC North America*, 19 (2001) 702-711.

[30] T. Rabilloud, Two-dimensional gel elctrophoresis in proteomics: old, old fashioned, but still climbs up the mountains, *Proteomics*, 2 (2002) 3-10.

[31] P.G. Righetti, A. Castagna, B. Herbert, Prefractionation techniques in proteome analysis, *Anal. Chem.*, 73 (2001) 320A-326A.

[32] P.G. Righetti, A.V. Stoyanov, M.Y. Zhukov, *The Proteome Revisited. Theory and Practice of All Relevant Electrophoretic Steps*, Elsevier, Amsterdam, 2001.

[33] D. Jäger, P.R. Jungblut, U. Müller-Werdan, Separation and identification of human heart proteins, *J. Chromatogr. B*, 771 (2002) 131-153.

[34] R.C.M.Y. Liang, J.C.H. Neo, S.L. Lo, G.S. Tan, T.K. Seow, M.C.M. Chung, Proteome database of hepatocellular carcinoma, *J. Chromatogr. B*, 771 (2002) 303-328.

[35] B. Herbert, P.G. Righetti, A turning point in proteome analysis: sample prefractionation via multicompartment electrolyzer with isoelectric membranes, *Electrophoresis*, 21 (2000) 3639-3648.

[36] X. Zuo, P. Hembach, L. Echan, D.W. Speicher, Enhanced analysis of human breast cancer proteomes using micro-scale solution isoelectrofocusing combined with high resolution 1-D and 2-D gels, *J. Chromatogr.*, 782 (2002) 253-265.

[37] A. Tan, A. Pashkova, L. Zang, F. Foret, B.L. Karger, A miniaturized multichamber solution isoelectric focusing device for separation of protein digests, *Electrophoresis*, 23 (2002) 3599-3607.

[38] T. Bączek, Fractionation of peptides in proteomics with the use of *pI*-based approach and ZipTip pipette tips, *J. Pharm. Biomed. Anal.*, 34 (2004) 851-860.

[39] T. Bączek, Fractionation of peptides and identification of proteins from *Saccharomyces cerevisiae* in proteomics with the use of reversed-phase capillary liquid chromatography and *pI*-based approach, *J. Pharm. Biomed. Anal.*, 35 (2004) 895-904.

[40] T. Bączek, A. Buciński, A.R. Ivanov, R. Kaliszan, Artificial neural network analysis for evaluation of peptide MS/MS spectra in proteomics, *Anal. Chem.*, 76 (2004) 1726-1732.

[41] R. Kaliszan, T. Bączek, A. Cimochowska, P. Juszczyk, K. Wiśniewska, Z. Grzonka, Prediction of high-performance liquid chromatography retention of peptides with the use of quantitative structure-retention relationships, *Proteomics*, 5 (2005) 409-415.

[42] T. Bączek, P. Wiczling, M. Marszałł, Y. Vander Heyden, R. Kaliszan, Prediction of peptide retention at different HPLC conditions from multiple linear regression models, *J. Proteome Res.*, 4 (2005) 555-563.

[43] T. Bączek, Chemometric evaluation of relationships between retention and physicochemical parameters in terms of multidimensional liquid chromatography of peptides, *J. Sep. Sci.*, 29 (2006) 547-554.

[44] R. Put, M. Daszykowski, T. Bączek, Y. Vander Heyden, Retention prediction of peptides based on uninformative variable elimination by partial least squares, *J. Proteome Res.*, 5 (2006) 1618-1625.

[45] S.-L. Wu, H. Amato, R. Biringer, G. Choudhary, P. Shieh, W.S. Hancock, Targeted proteomics of low-level proteins in human plasma by LC/MS^n: using human growth hormone as a model system, *J. Proteome Research*, 1 (2002) 459-465.

[46] A. Premstaller, H. Oberacher, W. Walcher, A.M. Timperio, L. Zolla, J.-P. Chervet, N. Cavusoglu, A. van Dorsselaer, C.G. Huber, High-performance liquid chromatography-electophoresis ionization mass spectrometry using monolithic capillary columns for proteomics studies, *Anal. Chem.*, 73 (2001) 2390-2396.

[47] L.J. Licklider, C.C. Thoreen, J. Peng, S.P. Gygi, Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column, *Anal. Chem.*, 74 (2002) 3076-3083.

[48] J. Preisler, P. Hu, T. Rejtar, B.L. Karger, Capillary electrophoresis – matrix-assisted laser desorption/ionization time-of-flight mass spectrometry using a vacuum deposition interface, *Anal. Chem.*, 72 (2000) 4785-4795.

[49] B. Zhang, F. Foret, B.L. Karger, High-throughput microfabricated CE/ESI-MS: automated sampling from a microwell plate, *Anal. Chem.*, 73 (2001) 2675-2681.

[50] M. Minarik, K. Klepàrnik, M. Gilàr, F. Foret, A.W. Miller, Z. Sosic, B.L. Karger, Design of a fraction collector for capillary array electrophoresis, *Electrophoresis*, 23 (2002) 35-42.

[51] R. Joubert, J.-M. Strub, S. Zugmeyer, D. Kobi, N. Carte, A. van Dorsselaer, H. Boucherie, L. Jaquet-Gutfreund, Identification by mass spectrometry of two-dimensional gel electrophoresis-separated proteins extracted from lager brewing yeast, *Electrophoresis*, 22 (2001) 2969-2982.

[52] M. Perrot, F. Sagliocco, T. Mini, C. Monribot, U. Schneider, A. Shevchenko, M. Mann, P. Jeno, H. Boucherie, Two-dimensional gel protein database of *Saccharomyces cerevisiae* (update 1999), *Electrophoresis*, 20 (1999) 2280-2298.

[53] M. Poutanen, L. Salusjarvi, L. Ruohonen, M. Penttila, N. Kalkkinen, Use of matrix-assisted laser desorption/ionization time-of-flight mass mapping and nanospray liquid

chromatography/electrospray ionization tandem mass spectrometry sequence tag analysis for sensitive, *Rapid Commun. Mass Spectrom.*, 15 (2001) 1685-1692.

[54] L. Salusjarvi, M. Poutanen, J.-P. Pitkanen, H. Koivistoinen, A. Aristidou, N. Kalkkinen, L. Ruohonen, M. Penttila, Proteome analysis of recombinant xylose-fermenting *Saccharomyces cerevisiae*, *Yeast*, 20 (2003) 295-314.

[55] D.M. Lubman, M.T. Kachman, H. Wang, S. Gong, F. Yan, R.L. Hamler, K.A. O'Neil, K. Zhu, N.S. Buchanan, T.J. Barder, Two-dimensional liquid separations–mass mapping of proteins from human cancer cell lysates, *J. Chromatogr. B*, 782 (2002) 183-196.

[56] A. Ros, M. Faupel, H. Mees, J. Van Oostrum, R. Ferrigno, F. Reymond, P. Michel, J.S. Rossier, H.H. Girault, Protein purification by off-gel electrophoresis, *Proteomics*, 2 (2002) 151-156.

[57] M. Cretich, G. Pirri, G. Carrea, M. Chiari, Separation of proteins in a multicompartment electrolyzer with chambers defined by a bed of gel beads, *Electrophoresis*, 24 (2003) 577-581.

[58] X. Kang, D.D. Frey, Chromatofocusing using micropellicular column packings with computer-aided design of the elution buffer composition, *Anal. Chem.*, 74 (2002) 1038-1045.

[59] Y. Shen, F. Xiang, T.D. Veenstra, E.N. Fung, R.D. Smith, High-resolution capillary isoelectric focusing of complex protein mixtures from lysates of microorganisms, *Anal. Chem.*, 71 (1999) 5348-5353.

[60] Y. Shen, S.J. Berger, G.A. Anderson, R.D. Smith, High-efficiency capillary isoelectric focusing of peptides, *Anal. Chem.*, 72 (2000) 2154-2159.

[61] L.R. Snyder, J.J. Kirkland, J.L. Glajch, *Practical HPLC Method Development*, John Wiley & Sons, New York, 1997.

[62] J. Swadesh, *HPLC Practical and Industrial Applications*, CRC Press, Boca Raton, 1997.

[63] R. Kaliszan, *Structure and Retention in Chromatography. A Chemometric Approach*, Harwood Academic Publishers, Amsterdam, 1997.

[64] S.R. Pennington, M.J. Dunn (red.), *Proteomics. From Protein Sequence to Function.* BIOS Scientific Publishers, Springer-Verlag, New York, 2001.

[65] http://eugenes.org:7072/

[66] J. Peng, J.E. Elias, C.C. Thoreen, L.J. Licklider, S.P. Gygi, Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome, *J. Proteome Res.*, 2 (2003) 43- 50.

[67] H.J. Cortes, *Multidimensional Chromatography. Techniques and Applications*, Marcel Dekker, New York, 1990.

[68] F. Regnier, A. Amini, A. Chakraborty, M. Geng, J. Ji, L. Riggs, C. Sioma, S. Wang, X. Zhang, Multidimensional chromatography and the signature peptide approach to proteomics, *LCGC North America*, 19 (2001) 200-213.

[69] T. Wehr, Multidimensional liquid chromatography in proteomic studies, *LCGC North America*, 20 (2002) 954-962.

[70] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, R. Aebersold, Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, *Nat. Biotechnol.*, 17 (1999) 994-999.

[71] S.P. Gygi, B. Rist, T.J. Griffin, J. Eng, R. Aebersold, Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags, *J. Proteome Res*., 1 (2002) 47-54.

[72] A.J. Link, J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Morris, B.M. Garvik, J.R. Yates III, Direct analysis of protein complexes using mass spectrometry, *Nat. Biotechnol.*, 17 (1999) 676-682.

[73] M.P. Washburn, D. Wolters, J.R. Yates III, Large-scale analysis of the yeast proteome by multidimensional protein identification technology, *Nat. Biotechnol.*, 19 (2001) 242-247.

[74] M.P. Washburn, R. Ulaszek, C. Deciu, D.M. Schieltz, J.R. Yates III, Analysis of quantitative proteomic data generated via multidimensional protein identification technology, *Anal. Chem.*, 74 (2002) 1650-1657.

[75] G.J. Opiteck, K.C. Lewis, J.W. Jorgenson, R.J. Anderegg, Comprehensive on-line LC/LC/MS of proteins, *Anal. Chem.*, 69 (1997) 1518-1524.

[76] K. Wagner, T. Miliotis, G. Marko-Varga, R. Bischoff, K.K. Unger, An automated on-line multidimensional HPLC system for protein and peptide mapping with integrated sample preparation, *Anal. Chem.*, 74 (2002) 809-820.

[77] J.T. Watson, *Introduction to Mass Spectrometry*, Lippincott-Raven, Philadelphia, PA, 1997.

[78] T. Bączek, Techniki rozdzielcze i spektrometria masowa w badaniach ludzkiego proteomu. Proteomika – nowe narzędzie w naukach medyczno-farmaceutycznych, *Farmacja Polska*, 59 (2003) 204-213.

[79] J.R. Yates III, Mass spectrometry and the age of the proteome, *J. Mass Spectrom.*, 33 (1998) 1-19.

[80] R.M. Caprioli, A. Malorni, G. Sindona, *Selected Topics in Mass Spectrometry in the Biomolecular Sciences*, Kluwer Academic Publishers, Dordrecht, 1997.

[81] http://www.ebi.uniprot.org/index.shtml

[82] http://www.ncbi.nlm.nih.gov/

[83] http://www.ebi.ac.uk//

[84] http://www.matrixscience.com/help/scoring_help.html

[85] http://www.thermo.com/com/cda/product/detail/0,,16483,00.html?CA=bioworks

[86] http://prospector.ucsf.edu/

[87] D.N. Perkins, D.J.C. Pappin, D.M. Creasy, J.S. Cottrell, Probability-based protein identification by searching databases using mass spectrometry data, *Electrophoresis*, 20 (1999) 3551-3567.

[88] J. Mazerski, *Podstawy chemometrii*, Wydawnictwo Politechniki Gdańskiej, Gdańsk, 2000.

[89] J.K. Eng, A.L. McCormack, J.R. Yates III, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, *J. Am. Soc. Mass Spectrom.*, 5 (1994) 976-989.

[90] J.R. Yates, III, J.K. Eng, A.L. McCormack, D. Schieltz, Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database, *Anal. Chem.*, 67 (1995) 1426-1436.

[91] D.C. Anderson, W. Li, D.G. Payan, W.S. Noble, A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores, *J. Proteome Res.*, 2 (2003) 137-146.

[92] D.L. Tabb, W.H. McDonald, J.R. Yates, III, DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics, *J. Proteome Res.*, 1 (2002) 2-26.

[93] J. Zupan, J. Gasteiger, Neural networks: A new method for solving chemical problems or just a passing phase?, *Anal. Chim. Acta*, 248 (1991) 1-30.

[94] J. Zupan, J. Gasteiger, *Neural Networks for Chemists. An Introduction*, VCH, Weinheim, 1993.

[95] K. Petritis, L.J. Kangas, P.L. Ferguson, G.A. Anderson, L. Pasa-Tolic, M.S. Lipton, K.J. Auberry, E.F. Strittmatter, Y. Shen, R. Zhao, R.D. Smith, Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses, *Anal. Chem.*, 75 (2003) 1039-1048.

[96] J.L. Meek, Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition, *Proc. Natl. Acad. Sci, USA*, 77 (1980) 1632-1636.

[97] C.A. Browne, H.P.J. Bennett, S. Solomon, The isolation of peptides by high-performance liquid chromatography using predicted elution positions, *Anal. Biochem.*, 124 (1982) 201-208.

[98] V. Casal, P.J. Martin-Alvarez, T. Herraiz, Comparative prediction of the retention behaviour of small peptides in several reversed-phase high-performance liquid chromatography columns by using partial least squares and multiple linear regression, *Anal. Chim. Acta*, 326 (1996) 77-84.

[99] D. Guo, C.T. Mant, A.K. Taneja, J.M.R. Parker, R.S. Hodges, Prediction of peptide retention times in reversed-phase high-performance liquid chromatography. I. Determination of retention coefficients of amino acid residues of model synthetic peptides, *J. Chromatogr.*, 359 (1986) 499-518.

[100] D. Guo, C.T. Mant, A.K. Taneja, R.S. Hodges, Prediction of peptide retention times in reversed-phase high-performance liquid chromatography. II. Correlation of observed and predicted peptide retention times and factors influencing the retention times of peptides, *J. Chromatogr.*, 359 (1986) 519-532.

[101] C.T. Mant, N.E. Zhou, R.S. Hodges, Correlation of protein retention times in reversed-phase chromatography with polypeptide chain length and hydrophobicity, *J. Chromatogr.*, 476 (1989) 363-375.

[102] R.A. Houghten, S.T. DeGraw, Effect of positional environmental domains on the variation of high-performance liquid chromatographic peptide retention coefficients, *J. Chromatogr.*, 386 (1987) 223-228.

[103] N.E. Zhou, C.T. Mant, R.S. Hodges, Effect of preferred binding domains on peptide retention behavior in reversed-phase chromatography: amphipathic alpha-helices, *Pept. Res.*, 3 (1990) 8-20.

[104] M. Palmblad, M. Ramström, K.E. Markides, P. Håkansson, J. Bergquist, Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry, *Anal. Chem.*, 74 (2002) 5826-5830.

[105] M. Palmblad, M. Ramström, C.G. Bailey, S.L. McCutchen-Maloney, J. Bergquist, L.C. Zeller, Protein identification by liquid chromatography mass spectrometry using retention time prediction, *J. Chromatogr. B*, 803 (2004) 131-135.

[106] R. Kaliszan, *Quantitative Structure-Chromatographic Retention Relationships*, Wiley, New York, 1987.

[107] A. Kraj, J. Silberring (red.), *Proteomika*, Wydział Chemii Uniwersytetu Jagiellońskiego, Kraków, 2004.

[108] T. Bączek, R. Kaliszan, Proteomika a techniki separacyjne, *Farmacja Polska*, 57 (2001) 923-925.

[109] T. Bączek, Improvement of peptides identification in proteomics with the use of new analytical and bioinformatic strategies, *Curr. Pharm. Anal.*, 1 (2005) 31-40.

7. Dodatek 1: opublikowane prace oryginalne wchodzące w skład rozprawy habilitacyjnej