# Identyfikacja białek z wykorzystaniem techniki *Peptide Mass Fingerprinting* (PMF)

## Część II – algorytmy scoringu

# The identification of proteins by Peptide Mass Fingerprinting (PMF)

## Part II – the scoring algorithms

*Hanna Kamińska[1, 2], Halina Podbielska[1]*

[1] Instytut Inżynierii Biomedycznej i Pomiarowej, Wydział Podstawowych Problemów Techniki, Politechnika Wrocławska, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, tel. +48 (71) 320 28 25, e-mail: hanna.kaminska@pwr.wroc.pl

[2] MedicWave AB, Stålverksgatan 1, SE-302 45 Halmstad, Sweden, tel. +46 35 133 600, e-mail: hanna.kaminska@medicwave.com

### Streszczenie

**Postęp w dziedzinie komputerów oraz rozwój Internetu zrewolucjonizował, proces identyfikacji białek oraz przyczynił się do szybkiego wzrostu proteomicznych baz danych. Krótko po wprowadzeniu pierwszej technologii identyfikacji białek z widm spektrometrów masowych PMF (*Peptide Mass Fingerprinting*) okazało się, że algorytmy wykorzystywane do wyszukiwania w bazie danych protein odpowiadających wynikom eksperymentu mają kluczowe znaczenie dla wysokiej poprawności identyfikacji. Rozwój metody PMF był zatem uwarunkowany nie tylko przez usprawnienia techniczne schematu, ale przede wszystkim przez zastosowanie rozmaitych metod matematycznych i statystycznych (tzw. algorytmów scoringu) przy wyszukiwaniu poprawnych rozwiązań. Kolejnym krokiem w informatycznym usprawnieniu identyfikacji było opracowanie metod walidacji jej rezultatów na podstawie istniejących baz danych lub też symulacji. Walidacja rezultatów pozwoliła na wyeliminowanie większości błędów pierwszego rodzaju w identyfikacji metodą PMF. Przez wzgląd na powszechność stosowania metody, a także jej ulepszenia autorzy postanowili podsumować obecny stan wiedzy w tym zakresie. Praca została podzielona na dwie części: w pierwszej przedstawiono opis historii powstania metody PMF wraz z charakterystyką jej części eksperymentalnej i opisem najpopularniejszych baz danych stosowanych przy identyfikacji, natomiast druga część jest poświęcona zagadnieniom algorytmicznym związanym z wyszukiwaniem w bazie danych protein najlepiej odzwierciedlających białko analizowane w próbce. Bioinformatyczne ujęcie identyfikacji białek w drugiej części nawiązuje do specyfikacji eksperymentu, omówionej w części pierwszej publikacji. Druga część pracy w szczegółowy sposób opisuje główne aspekty porównywania mas teoretycznych i eksperymentalnych, tj. trawienie *in silico*, rozpoznawanie modyfikacji białek, dopasowywanie mas oraz kalibrację poprawnych dopasowań. Opisane zostały także sposoby budowania funkcji scoringowych oraz algorytmy walidacji ich wartości. Dodatkowo, w pracy przedstawiono najbardziej znane funkcje scoringowe oraz pełny przegląd oprogramowania do identyfikacji białek metodą PMF.**

**Słowa kluczowe**: proteomika, identyfikacja protein, spektrometria masowa, *peptide mass fingerprinting*, schematy scoringu

### Abstract

The internet and computer science progress have revolutionized the process of protein identification and contributed to the growth of proteomics databases. Just after discovering the first technology for protein identification from the mass spectra PMF (peptide mass fingerprinting), it appeared that the algorithms searching databases for proteins corresponding to experiment results have crucial meaning for the sensitivity and specificity of the identification procedure. Therefore, the development of PMF method was conditioned by both the technological improvements in the PMF scheme and the application of various mathematical and statistical methods (so called: scoring algorithms) to the searching of correct identifications. The next step in the development of an identification procedure was to work out the methods for identification results validation, according to the proteomics databases content or simulations. The results validation allowed to eliminate the most of unwanted false positives in the PMF identification. Regarding the method common use, as well as its improvements which are still present, the authors decide to summarize the current level of knowledge related to this topic. The publication is divided into two parts. The first one is devoted to the origins of PMF scheme, the characteristics of its experimental part and a description of the most popular databases used in the identification procedure. The second part relates to the algorithmic issues of searching the database protein, which reflects the sample content best. From the bioinformatics point of view the protein identification in the second part of publication refers to the experiment specification described in the first part. The second part of the publication describes in details the aspects of theoretical and experimental masses comparison, i.e. *in silico* digestion, the discrimination of protein modifications, the pairing of masses and the calibration of matches. Moreover, the scoring functions building manners and the algorithms for scoring functions values validation were also taken into the consideration. Additionally, we present the most known scoring schemes with the comprehensive review of the PMF protein identification software.

**Keywords**: proteomics, identification of proteins, mass spectrometry, peptide mass fingerprinting, scoring schemes

### Wstęp

Identyfikacja białek jest złożonym procesem uwzględniającym zarówno etap analizy eksperymentalnej, jak i komputerowej. W poprzedniej części publikacji mogliśmy przyjrzeć się dokładnie jej części eksperymentalnej, w odniesieniu do techniki identyfikacji białek określanej jako PMF (*Peptide Mass Fingerprinting*). Nazwa metody wiąże się z obrazowaniem na spektrum masowym mas peptydów, będących wynikiem trawienia nieznanego białka [1]. Masy peptydów, w większości różne, tworzą charakterystyczną dla każdego białka kompozycję, która stanowi swego rodzaju identyfikator białka, jego masowy *odcisk palca* [2].

Po eksperymentalnym pozyskaniu identyfikatora masowego białka jest on porównywany z utworzonymi *sztucznie* zbiorami mas peptydów. Zbiory te są budowane na podstawie sekwencji aminokwasów, odpowiadających białkom opisanym w proteomicznych bazach danych [3]. Przez wzgląd na ogrom danych przechowywanych w bazach (przykładowo: 526 969 sekwencji aminokwasów opisanych w bazie danych Swiss-Prot [4] i 14 555 721 w bazie TrEMBL [5] na dzień 5 kwietnia 2011) proces porównywania identyfikatorów masowych białek nie może być przeprowadzany manualnie. Co więcej, czas, w jakim odbywa się przeszukiwanie bazy danych, powinien być jak najkrótszy [6]. Automatyzacja zadania wzajemnego dopasowywania mas peptydów eksperymentalnych oraz tak zwanych mas *teoretycznych* (powstałych w wyniku teoretycznego trawienia sekwencji przechowywanych w bazie) jest możliwa dzięki złożonym algorytmom komputerowym [7]. Każdy algorytm składa się z trzech podstawowych części [8, 9]:

1. Obliczania ilości mas peptydów odpowiadających sobie w białku eksperymentalnym i proteinie z bazy.
2. Przypisywania każdej proteinie z bazy danych pewnej sumy punktów (*score*), oceniającej jej podobieństwo do białka eksperymentalnego.
3. Weryfikacji istotności najlepszego wyniku dopasowania białek, tj. czy wynik faktycznie można uznać za poprawny.

Etapy te zostały określone wspólnym mianem *scoringu*. Nazwa ta wywodzi się od słowa *score* oznaczającego w tym kontekście liczbę punktów, jaką uzyskała proteina z bazy przy porównywaniu z białkiem eksperymentalnym [8]. Analogicznie, algorytmy poszukujące w bazie danych najbardziej prawdopodobnego rezultatu identyfikacji nazywamy *algorytmami scoringowymi.*

Głównym tematem drugiej części publikacji są właśnie algorytmy scoringowe oraz różnorodne przykłady ich realizacji, wraz z opisem najbardziej popularnego oprogramowania przeprowadzającego pełen proces identyfikacji w schemacie PMF. W sekcji 2 opisany zostanie proces trawienia teoretycznego bazy danych oraz sposób, w jaki algorytmy radzą sobie z obecnością różnorodnych modyfikacji w białkach. Podane zostaną także podstawowe zasady parowania mas teoretycznych i eksperymentalnych oraz sposoby kalibracji wyników takiego parowania. W sekcji 3 przedstawiona zostanie budowa funkcji scoringowej oraz sposoby walidacji wartości funkcji scoringowych. Sekcja: *Algorytmy scoringu i oprogramowanie* będzie natomiast dotyczyć istniejących probabilistycznych i nieprobabilistycznych modeli scoringowych oraz oprogramowania przeprowadzającego identyfikację białek w schemacie PMF.

## Porównywanie mas

**Pojęcie trawienia teoretycznego *in silico***

Jak już zostało opisane w poprzedniej części publikacji, bazy danych proteomicznych, do których odwołujemy się, chcąc porównać efekt eksperymentu z ich zawartością, przechowują przede wszystkim sekwencje aminokwasów oznaczające poszczególne proteiny [3]. Końcowym efektem eksperymentu z wykorzystaniem spektrometru są natomiast masy peptydów zaobserwowanych w próbce. Kluczowym zadaniem jest więc odpowiednie przekształcenie sekwencji aminokwasów przechowywanej w bazie danych na peptydy, a następnie obliczenie ich mas, aby móc porównać je z efektem eksperymentu [10]. Proces ten jest określany mianem *trawienia teoretycznego, in silico* (czyli za pomocą komputera), ponieważ w sposób teoretyczny próbujemy w nim zasymilować trawienie łańcucha białkowego, znając właściwości enzymu trawiennego, który był wykorzystany w eksperymencie na danych rzeczywistych [11].

Przykładowo, jeżeli wiemy, że enzymem trawiennym była trypsyna, to sekwencję proteiny z bazy danych należy podzielić na fragmenty za każdą występującą w łańcuchu lizyną lub też argininą, o ile nie występuje za nimi prolina. W wyniku takiego podziału otrzymujemy pewien jednoznaczny, teoretyczny zbiór sekwencji peptydowych, na jakie z największym prawdopodobieństwem rozpadło się białko. Oczywiście taki zbiór sekwencji otrzymamy przy założeniu pełnej swoistości trypsyny, jednak z praktycznego punktu widzenia wiadomo, że w eksperymencie rzeczywistym mogą zajść warunki, które w niepożądany sposób wpłyną na zdolności trawienne enzymu. Może się zatem zdarzyć sytuacja, w której nie we wszystkich przewidywanych miejscach w łańcuchu aminokwasów nastąpi przerwanie wiązania peptydowego. Niewytrawione fragmenty łańcucha pojawiają się często w miejscach, w których aminokwasy bezpośrednio sąsiadujące z potencjalnym miejscem trawienia są dodatnio lub ujemnie naładowane [7]. Trawienie *in silico* pozwala nam uwzględnić takie niepożądane zachowanie enzymu. W takiej sytuacji popularnym podejściem jest wygenerowanie wszystkich możliwych sekwencji z dopuszczalną jedną lub więcej niewytrawioną pozycją w łańcuchu aminokwasów. Dopuszczenie w przeszukiwaniu bazy danych jednego niewytrawionego miejsca jest zalecane, ponieważ ma to znaczący wpływ na zwiększenie poprawności identyfikacji [12]. W praktyce, w większości przypadków nie obserwuje się w eksperymentach peptydów o więcej niż dwóch stronach z nierozerwanym wiązaniem peptydowym [12]. Analogicznie, możemy mieć także do czynienia ze zjawiskiem białka wytrawionego nadmiernie, w miejscach nieprzewidywanych. Zdarza się to jednak najczęściej w sytuacji zanieczyszczenia próbki eksperymentalnej innym enzymem [7].

Po przeprowadzaniu trawienia *in silico* dla każdej proteiny opisanej w bazie danych należy obliczyć masy wszystkich peptydów ją tworzących. Takie masy będziemy dalej określać mianem mas teoretycznych. Na początku zadanie to wydaje się trywialne, ponieważ wiąże się ono z dodaniem dla każdego peptydu znanych mas tworzących go aminokwasów. Dodatkowo, w sumie należy uwzględnić masy wodoru (H) oraz grupy wodorotlenowej (OH) znajdujących się odpowiednio przy N-końcu i C-końcu peptydu, powstałych w wyniku rozerwania wiązania peptydowego [13]. W rzeczywistości jednak o wiele więcej czynników ma wpływ na masy peptydów zaobserwowanych na widmie masowym i wszystkie te czynniki należy uwzględnić przy obliczaniu mas peptydów wchodzących w skład protein z bazy danych.

**Rozpoznawanie modyfikacji białek**

Podstawowym czynnikiem, który może wpłynąć na masy peptydów w eksperymencie rzeczywistym, są ich modyfikacje – zarówno te intencjonalne, jak i różnorodne modyfikacje posttranslacyjne. Niemniej jednak mimo wielu rodzajów modyfikacji, jakie mogą wystąpić w sekwencji aminokwasów, w większości przypadków zastosowanie spektrometrii mas do analizy białek pozwala w prosty sposób je wykryć [14]. Z każdą modyfikacją wiąże się bowiem zmiana masy aminokwasu, którego ta modyfikacja dotyczy. Dla wybranych rodzajów modyfikacji (np. przy fosforylacji) masa aminokwasów będzie się zawsze różnić o pewną stałą wartość (dodatnią lub ujemną) [15]. Ponieważ podstawowym zadaniem spektrometrii w schemacie PMF jest analiza masy, to takie modyfikacje na spektrum będą widoczne jako przesunięcia pików odpowiednio w prawo lub lewo o określony interwał masowy.

Modyfikacje celowo wprowadzane w białkach wiążą się najczęściej ze schematem przygotowywania próbki do eksperymentu spektrometrem masowym. Do najbardziej popularnych modyfikacji należą: karbamidometylizacja cystein, pojawiająca się w przypadku wykorzystania jodoacetamidu, lub ich karboksymetylizacja, jeżeli w eksperymencie użyto kwasu jodooctowego [16]. Te modyfikacje zmieniają średnią masę cysteiny odpowiednio o +57,072 Da i +58,037 Da. Jeżeli którakolwiek z tych dwóch modyfikacji ma miejsce, to dotyczy zawsze blisko 100% cystein obecnych w próbce [7]. Inną ustaloną modyfikacją jest tworzenie tzw. aminokwasów PEC [17] z cysteiny za pomocą 4-winylopirydyny i wiąże się ona ze zwiększeniem masy cysteiny o 105,139 Da. Cysteiny mogą być modyfikowane także przez obecny w próbce akrylamid, podstawowy składnik żelu wykorzystywanego w elektroforezie żelowej [18], który po dołączeniu do cysteiny zmienia jej masę o +71,079 Da. Do bardzo często

spotykanych modyfikacji, które nie były efektem laboratoryjnej interwencji do podstawowych, należą [15, 19]:

- acetylowanie N-końców lub lizyny (+42,037 Da),
- fosforylacja seryny, treoniny lub tyrozyny (+79,980 Da),
- metylowanie C-końców, kwasu asparaginowego, kwasu glutaminowego (+14,027 Da),
- tworzenie mostków dwusiarczkowych (–2 Da),
- dołączenie biotyny do N-końców lub lizyny (+226,293 Da),
- formylowanie N-końca (+28,010 Da),
- utlenianie histydyny, metioniny, tryptofanu (+15,999 Da).

Tego typu modyfikacje, jeżeli już się pojawiają, najczęściej nie są obecne na 100% aminokwasów, których dotyczą [7]. W przypadku zarówno ustalonych, jak i dodatkowych modyfikacji białek należy pamiętać, że uwzględnienie w oprogramowaniu obliczającym masy peptydów z bazy danych wszystkich możliwych modyfikacji znacząco zwiększy liczebność zbiorów mas teoretycznych. Masy teoretyczne trzeba natomiast porównać z masami eksperymentalnymi. W konsekwencji program realizujący algorytm dopasowania mas teoretycznych i eksperymentalnych może nie podołać w sensownym czasie rozwiązaniu zadania skojarzenia tych mas [20]. Dlatego wybierając przy przeszukiwaniu bazy danych możliwe modyfikacje, należy wybrać minimum tego, co uznamy za możliwe do pojawienia się w analizowanej próbce. Dodatkowo, należy pamiętać, iż w przypadku oceny rozwiązania dopasowania mas z bazy do białka z eksperymentu, jeżeli równie wysoko zostanie ocenione rozwiązanie z modyfikacjami i bez nich, za bardziej prawdopodobne będzie uznane to drugie. Optymalnym rozwiązaniem jest zatem najpierw wyszukanie odpowiadającego białka z minimalną liczbą możliwych modyfikacji, a następnie kolejne przeszukiwanie bazy danych z poszerzonym zbiorem modyfikacji w przypadku braku satysfakcjonującego rozwiązania.

**Wzajemne dopasowywanie mas teoretycznych i eksperymentalnych**

Przy porównywaniu dwóch zbiorów: mas eksperymentalnych $e = (e^1, e^2, ..., e_n)$ i teoretycznych $t = (t^1, t^2, ..., t_n)$ stwierdzamy, że dwie wybrane masy ($e_i$ i $t_j$) pasują do siebie, wtedy i tylko wtedy gdy różnica pomiędzy nimi zawiera się w pewnym z góry założonym przedziale (1).

$$\exists_i \in [1 .. n] \;\; \exists_j \in [1 .. m] \qquad |ei - tj| < \delta \qquad (1)$$

Dla niewytrawionych fragmentów łańcucha aminokwasów powyższe równanie może przyjąć następującą postać (2):

$$\exists_i \in [1 .. n] \;\; \exists_j \in [1 .. m] \;\; \exists_k \in [1 .. m] \;\; j \neq k \qquad (2)$$
$$|e_i - (t_j + t_k)| < \delta$$

W przypadku modyfikacji należy natomiast wziąć pod uwagę zmienną $\lambda_{jk}$ reprezentującą sumę wybraną ze wszystkich możliwych sum mas aminokwasów zmodyfikowanych $l$ dla peptydu zgodnie z określonym wcześniej zbiorem modyfikacji (3):

$$\exists_i \in [1 .. n] \;\; \exists_j \in [1 .. m] \;\; \exists_k \in [1 .. l] \;\; j \neq k \qquad (3)$$
$$|e_i - t_j + \lambda_{jk}| < \delta$$

Nie są to jednak wszystkie problemy, którym powinien umieć sprostać algorytm dokonujący takiego dopasowania mas. W trzech powyższych równaniach (1–3) niezwykle istotną rolę pełni dobór parametru $\delta$. Parametr ten powinien być dobierany zależnie od dokładności spektrometru masowego, na którym był wykonywany eksperyment. Ponieważ błąd systematyczny pomiaru mas w spektrometrach masowych zwiększa się wraz ze wzrostem badanej masy cząsteczki [7], istotne jest, aby wartość $\delta$ zmieniała się liniowo wraz ze wzrostem dopasowywanych mas. Nieuwzględnienie tej właściwości spektrometru masowego może spowodować, że dwie stosunkowo duże cząsteczki (peptyd eksperymentalny i teoretyczny) nie zostaną ze sobą skojarzone ze względu na zbyt małą tolerancję dla różnicy ich mas [21, 22].

Kolejny problem, z jakim mamy do czynienia przy dopasowaniu mas teoretycznych i eksperymentalnych, to zjawisko dublowania się mas peptydów. Dublowanie mas peptydów może nastąpić w dwóch przypadkach. Po pierwsze, w zbiorze wszystkich aminokwasów istnieje izoleucyna (I) i leucyna (L), które mają identyczną masę molekularną (113,08 Da). Oznacza to, że dwa przykładowe peptydy *EGI* i *EGL* będą miały jednakową masę molekularną (299,86 Da), należąc do **dwóch różnych białek**. Również podobieństwo mas tryptofanu i peptydu *EG* oraz mas peptydów *HP* i *FS* może powodować niejednoznaczności w analizie sekwencji [8]. Po drugie, takie peptydy, należąc do **tego samego białka**, będą odpowiadały również pewniej (jednej) masie zaobserwowanej w próbce eksperymentalnej.

Niektóre proteiny z bazy przy trawieniu *in silico* uzyskują kilka mas o takiej samej wartości [14]. W konsekwencji mogą zatem zyskać więcej prawdopodobnych dopasowań z wybraną masą eksperymentalną, w stosunku do innych protein z bazy. W eksperymencie spektrometrem masowym nie mamy możliwości rozróżnienia, z jakich peptydów (sekwencji) pochodzi dana masa, i każdy element zbioru mas eksperymentalnych jest różny. Inaczej jest w zbiorze mas teoretycznych, który powinien być raczej określany mianem multizbioru, ponieważ masy w nim zawarte mogą należeć do innych sekwencji. Aby nie faworyzować sekwencji o dużej liczbie jednakowych mas, multizbiór mas teoretycznych traktuje się jako zwykły zbiór mas niepowtarzających się. Widzimy zatem, że tak jak w przypadku prawdziwych cech biometrycznych, odcisk palca (*fingerprint*) w szczególnych przypadkach nie daje jednoznacznej informacji o osobie (np. bliźniakach), analogicznie – masowy odcisk peptydu (*peptide mass fingerprint*) nie rozróżnia jednoznacznie sekwencji białkowych.

**Kalibracja poprawnych dopasowań**

Po dopasowaniu do siebie mas ze zbioru mas eksperymentalnych i zbiorów mas teoretycznych następuje kalibracja poprawnych dopasowań. Operacja ta jest wykonywana w celu pozbycia się fałszywych trafień z już skonstruowanych zbiorów dopasowanych mas oraz sprostania problemowi kilku mas teoretycznych pasujących na podstawie przyjętej tolerancji do jednej masy eksperymentalnej [8]. Dla otrzymanych zbiorów mas po kalibracji mogą iteracyjnie następować kolejne przeszukiwania bazy, mające na celu ulepszenie rezultatów dopasowania mas. Podstawową ideą wykorzystaną przy kalibracji wyników przeszukiwania bazy danych jest użycie korelacji między dopasowanymi masami eksperymentalnymi ($e$) i teoretycznymi ($t$). Na podstawie tej korelacji można wyznaczyć linię regresji, zgodnie z którą modyfikowane są pierwotne masy eksperymentalne. Kolejne przeszukiwanie bazy danych jest już przeprowadzane dla poprawionych mas eksperymentalnych ($e^*$).

Pierwsze podejście do kalibracji rezultatów przeszukiwania bazy danych zostało zaproponowane przez Egelhofera w programie MSA [23]. Algorytm kalibracji składa się z następujących kroków:

1. Z bazy danych wyszukiwane są wszystkie białka rozbijające się przy trawieniu na przynajmniej 5 peptydów i których masy są bliskie masom eksperymentalnym z uwzględnieniem sporej tolerancji $\delta$ (np. ± 500 ppm).
2. Dla $k$-tej sekwencji z bazy danych obliczany jest błąd względny dla każdej masy peptydu dopasowanego ($t_d$) ze zbioru mas teoretycznych $t^k_d$:

$$b_i = (m_i - t^k_{di}) / mi \qquad (4)$$

3. Następnie obliczane są średnia ($\mu$) i odchylenie standardowe ($\sigma$) dla każdego z błędów odpowiadających masom peptydów dla odrębnych sekwencji. Wszystkie dopasowania $t^k_{di}$ dla każdej proteiny, których błąd nie należy do przedziału $[\mu - 2\sigma, \mu + 2\sigma]$, są odrzucane jako niepoprawne.

4. Na podstawie regresji liniowej określana jest prosta $y = a + bm$ definiująca korelację błędów względem mas eksperymentalnych. Odległości między względnymi błędami dopasowań a linią regresji są obliczane i ponownie ze zbioru dopasowanych mas eliminowane są masy, których błędy odstają od przyjętego kryterium: np. znajdują się w większej odległości od prostej niż $2\sigma$.

Zupełnie inne podejście prezentuje algorytm wykorzystywany w popularnym otwartym programie do identyfikacji białek Aldente [24, 25]. Polega on na budowie wykresu zależności mas teoretycznych od eksperymentalnych w taki sposób, że każdej zaznaczonej na osi odciętych masie teoretycznej przyporządkowane są wszystkie punkty zaznaczone na osi rzędnych odpowiadające masom eksperymentalnym. Następnie z wykorzystaniem transformacji Hougha konstruowana jest prosta na wykresie zależności mas w taki sposób, aby jak największa ilość punktów należących do wykresu nie była oddalona od prostej dalej niż o błąd wewnętrzny spektrometru, na jakim przeprowadzany był eksperyment [26]. Warto wspomnieć, że prosta definiująca dopasowanie mas jest wyznaczana przy użyciu błędów kalibracji spektrometru masowego. Taki sposób dopasowania mas rozwiązuje problem wielokrotnego dopasowania mas teoretycznych do mas eksperymentalnych.

## Scoring

**Budowa funkcji scoringowej**

Kiedy dla wszystkich protein z bazy zostanie określony zbiór mas peptydów pasujących do mas uzyskanych w eksperymencie, to dla każdej takiej proteiny powinna zostać wyznaczona pewna wartość liczbowa określająca, z jak dużym prawdopodobieństwem wybrany zestaw dopasowań należy do białka eksperymentalnego. Ta wartość liczbowa jest określana jako wynik (*score*), natomiast procedura estymacji wyniku jest określana mianem scoringu. Głównym celem przy projektowaniu schematu scoringu, bądź też tzw. *funkcji scoringowej*, jest uzyskanie jak najwyższego wyniku dla sekwencji z bazy danych reprezentującej poprawne białko i zauważalnie gorszych, pod względem wartości, wyników dla pozostałych protein z bazy. Wynik uzyskany przez proteinę z bazy mierzy stopień jej podobieństwa w stosunku do białka eksperymentalnego.

Informacja o ilości dopasowań mas teoretycznych do eksperymentalnych jest kluczowa przy konstrukcji funkcji wyznaczającej taki wynik, jednakże nie można go wyznaczyć tylko na tej podstawie. Przeszkodą jest chociażby fakt, że wiele białek z bazy danych może mieć identyczną liczbę dopasowań mas peptydów do białka eksperymentalnego [27]. Pierwsze zaprezentowane funkcje scoringowe opierały się tylko na liczbie dopasowanych mas [28]. Bardziej skomplikowane metody, w których liczba poprawnych dopasowań była jednym z głównych czynników wpływających na wynik, zostały zaprezentowane później [13, 29]. Funkcje scoringowe są zatem złożone, a każdy ich komponent ma odpowiednio mniejszy lub większy wpływ na wartość wyniku dopasowania. Najczęściej każdemu komponentowi funkcji scoringowej jest przypisywana odpowiednia waga, opisująca jego wpływ na dopasowanie mas [30]. Wszystkie funkcje scoringowe uwzględniają pewien podstawowy zbiór parametrów mających wpływ na wynik dopasowania. Do najbardziej oczywistych parametrów wpływających na wartość wyniku należą [8]:

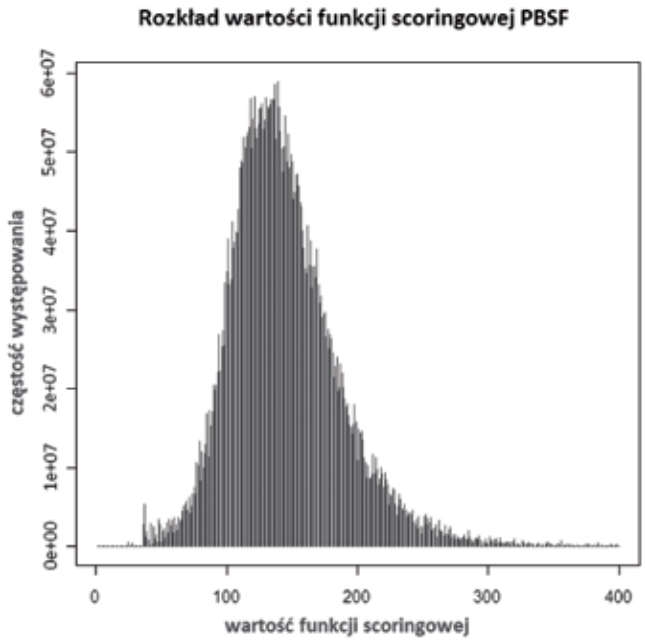
1. liczba mas, które zostały uznane za odpowiadające sobie w zbiorze mas eksperymentalnych i teoretycznych,
2. liczba mas eksperymentalnych,
3. liczba mas teoretycznych,
4. procent pokrycia przez dopasowane masy sekwencji proteiny z bazy danych,
5. oszacowana liczba różnego typu białek w próbce,
6. zróżnicowanie wartości błędów dopasowań mas peptydów eksperymentalnych i teoretycznych,
7. różnice w dopasowanych masach,
8. liczba dopuszczalnych niewytrawionych fragmentów sekwencji,
9. uwzględnienie modyfikacji w masach teoretycznych.

Należy pamiętać o tym, że wartości większości z tych parametrów zależą od wybranej bazy danych, zatem rezultat identyfikacji jednoznacznie wiąże się z doborem bazy danych szukanych sekwencji. Nawiązując do wyżej wymienionych parametrów funkcji scoringowych, im większa ilość poprawnie dopasowanych mas, tym większe prawdopodobieństwo, że identyfikacja jest poprawna. Przez wzgląd na objętość bazy danych protein i mnogość sekwencji, dla których należy obliczyć wartość funkcji scoringowej, białka z bazy danych niezawierające przynajmniej dwóch poprawnie dopasowanych mas nie są uwzględnianie w dalszej analizie. Liczby mas eksperymentalnych i teoretycznych mają kluczowe znaczenie dla wartości funkcji scoringowej, ponieważ określają istotę ilości mas poprawnie dopasowanych. Jeżeli liczba mas eksperymentalnych jest bardzo mała, jest to sytuacja niekorzystna, ponieważ mogą to być mało różnorodne masy, które będą pasowały do wielu peptydów teoretycznych pochodzących z różnych protein jednocześnie. W przypadku mas teoretycznych duże białka mają ich znacznie więcej, czym zwiększają swoje prawdopodobieństwo na dopasowanie do mas eksperymentalnych. Im większy procent pokrycia sekwencji z bazy przez masy eksperymentalne, tym więcej punktów taka proteina powinna zyskiwać względem pozostałych. Im więcej białek znajdowało się w próbce, która była analizowana spektrometrem masowym, tym większe prawdopodobieństwo, że część analizowanych mas eksperymentalnych powinniśmy traktować jako zanieczyszczenie i rozważyć kilka wyników identyfikacji jednocześnie [31]. Duże zróżnicowanie lub duże wartości błędów dopasowanych mas eksperymentalnych i teoretycznych mogą świadczyć o całkowitym niedopasowaniu białek, gdyż w przypadku poprawnej identyfikacji błędy te powinny być podobnej i stałej wartości dla wszystkich dopasowanych mas peptydów. Funkcja scoringowa powinna także premiować najprostszy scenariusz identyfikacji, czyli pasujące białka z bazy danych z jak najmniejszą liczbą niewytrawionych fragmentów oraz modyfikacji. Potencjalne niewytrawione miejsca sekwencji oraz modyfikacje zwiększają liczbę sekwencji dobrze pasujących do eksperymentalnego białka, zmniejszając tym samym wiarygodność identyfikacji.

**Walidacja wyniku**

Obliczenie wartości funkcji scoringowej oraz znalezienie białka w bazie danych posiadającego największy wynik nie kończy jednak procedury identyfikacji. Bo co dzieje się w przypadku, gdy taki sam co do wartości najlepszy wyniki posiada jednocześnie kilka białek? A co tak naprawdę znaczy fakt, że najwyższy wynik bardzo niewiele różni się od wszystkich pozostałych? Na takie pytania pozwala odpowiedzieć procedura walidacji wartości funkcji scoringowych przypisywanych białkom [32]. Jest ona kluczowym krokiem przy sprawdzaniu istotności wyników, ponieważ identyfikację możemy uznać za poprawną tylko w przypadku, gdy jej wynik w sposób istotny różni się od innych wartości uzyskanych przez pozostałe białka [33].

Istotność wyniku identyfikacji jest często mierzona przez estymację *p-wartości*, określanej dla danej proteiny z bazy jako prawdopodobieństwo zdobycia wyniku $s$ lub wyższego przez przypadek. Przyjmując za hipotezę zerową $H_0$, że wynik $s$ uzyskany przez proteinę z bazy jest losowy, natomiast za hipotezę alternatywną $H_1$, że $s$ ma wartość istotną, określamy poziom istotności $s$, czyli $p$-wartość w próbie. Jeżeli $p$-wartość jest mniejsza lub równa pewnemu ustalonemu poziomowi (zwykle przyjmujemy 0,05) odrzucamy hipotezę zerową, a wartość $s$ przyjmujemy za istotną. Ze względu na poziom skomplikowania funkcji scoringowych, często niemożliwe jest obliczenie $p$-wartości korzystając z postaci funkcji, stosuje się metodę estymacji $p$-wartości bazującą na symulacji.




***Rys. 1*** *Rozkład wartości funkcji scoringowej Probability Based Scoring Function (PBSF) w bazie danych Swiss-Prot (wersja: 15.12 z 15 grudnia 2009)*
***Źródło:*** *Opracowanie własne.*

Najbardziej powszechną metodą wyznaczania *p*-wartości wykorzystującą symulację jest skonstruowanie rozkładu najwyższych wartości funkcji scoringowej w wybranej bazie danych, dla symulowanych widm masowych [34]. Metoda polega na wygenerowaniu od kilku do kilkunastu tysięcy zestawów mas eksperymentalnych, a następnie ich identyfikacji w bazie danych [35]. Oczywiście spektra losowe muszą być wygenerowane w sposób wiarygodny, aby w jak najmniejszy sposób różnić się spektrów otrzymywanych w realnym eksperymencie. Przykładowo: nie mogą być na nich obecne masy peptydów niemożliwe do zaobserwowania w środowisku naturalnym lub zmierzenia spektrometrem masowym. Dobrym sposobem na wygenerowanie takich spektrów jest wylosowanie wartości mas tych peptydów z rozkładu mas wszystkich peptydów istniejących w rzeczywistej bazie danych protein. Następnie z największych wartości wyników identyfikacji, dla każdego spektrum z osobna, budujemy rozkład, który może być dopasowany do któregoś ze znanych rozkładów prawdopodobieństwa. Dla znanych rozkładów prawdopodobieństwa wyznaczenie *p*-wartości, nie będzie już zadaniem trudnym. Przykład takiego rozkładu, dla funkcji scoringowej *Probabilty Based Scoring Function* (PBSF), która bardziej szczegółowo zostanie omówiona w kolejnej sekcji – Algorytmy scoringu i oprogramowanie – można zobaczyć na rysunku 1. Podobne metody obliczania istotności wartości funkcji scoringowej, opierające się o symulacje rozkładów jej wartości, zostały zaprezentowane w literaturze przedmiotu [27, 33].

## Algorytmy scoringu i oprogramowanie

Na świecie istnieje obecnie wiele programów, zarówno komercyjnych, jak i darmowych do identyfikacji białek metodą PMF, wykorzystujących funkcje scoringowe bazujące na zróżnicowanych parametrach. Wszystkie te funkcje możemy rozdzielić na dwie kategorie: probabilistyczne i nieprobabilistyczne. Pierwsze są zawsze związane z budową pewnego rozkładu mas (protein bądź peptydów) w bazie danych. Ich wartość reprezentuje ocenione na podstawie różnych parametrów prawdopodobieństwo tego, że masy zaobserwowane na spektrum masowym należą do wybranej proteiny z bazy danych. Funkcje nieprobabilistyczne nie analizują prawdopodobieństwa, lecz opierają się przede wszystkim na chemicznych własnościach mas i komponentów próbki. W dalszej części artykułu przedstawione zostaną oba rodzaje wyżej wymienionych funkcji oraz zaprezentowane zostanie związane z nimi oprogramowanie do identyfikacji białek. Ze względu na liczbę rozmaitych funkcji scoringowych, szczegółowo zaprezentowane zostaną te najbardziej znane i charakterystyczne. Inne opisane będą mniej szczegółowo, wraz z listą oprogramowania przeprowadzającego identyfikację PMF, w sekcji: Oprogramowanie.

### Probabilistyczne funkcje scoringowe

#### MOWSE i Mascot

Prekursorem definicji funkcji scoringowej (jak również wprowadzenia określenia *peptide mass fingerprint*) był Darryl Pappin, który w roku 1993 zaproponował pierwszy schemat scoringu pod nazwą MOWSE [2]. MOWSE nie był przykładem typowo probabilistycznej funkcji, jednak dał podwaliny pod obecnie najbardziej popularny, probabilistyczny program do identyfikacji protein w schemacie PMF – Mascot [36]. Aby wyznaczyć wartość funkcji MOWSE, należy najpierw obliczyć częstotliwość występowania w wybranej bazie danych mas poszczególnych protein oraz mas peptydów je tworzących, według przyjętego wcześniej enzymu trawiennego. By to zrobić, ustalamy pewne przedziały masowe, w których zliczamy masy peptydów i protein. Według tych przedziałów występowania poszczególnych mas tworzy się tak zwaną *tabelę MOWSE*, składającą się z kolumn reprezentujących masy protein w przedziałach co 10 kDa oraz wiersze reprezentujące przedziały oddalone od siebie co 100 Da, reprezentujące masy molekularne zliczanych peptydów.

***Tabela 1*** *Tabela częstości występowania mas MOWSE*

| Masy peptydów | 10 kDa | 20 kDa | ... | J*10 kDa | ... | |
|---|---|---|---|---|---|---|
| | | | | | | 100 Da |
| | | | | | | 200 Da |
| | | | | | | ... |
| | | | | $n_{ij}$ | | I*100 Da |
| | | | | | | ... |
| | Masy protein | | | | | |

***Źródło****: Opracowanie własne: [4].*

W tabeli 1 wartość $n_{ij}$ oznacza liczbę peptydów, których masy należą do przedziału $[(i-l) \times 100Da . i \times 100Da]$. Oczywiście peptydy $n_{ij}$, należące do *j*-tej kolumny tabeli, są wynikiem trawienia protein, których masy należą do przedziału $[(i-l) \times 10kDa . i \times 10kDa]$. Wartości *i* i *j* są ograniczone przez największe masy protein i peptydów występujące w bazie danych. Następnie na postawie tabeli MOWSE tworzy się tabelę częstości występowania poszczególnych mas, przez normalizację wartości każdej komórki (5):

$$f_{ij} = \frac{n_{ij}}{\max_i n_{ij}} \qquad (5)$$

Ostatecznie, wartość funkcji scoringowej dla każdego białka z bazy danych jest obliczana według poniższej formuły (6):

$$Score_{MOWSE} = \frac{50\,000}{m_t \prod_{k=1}^{r} f_{ck}} \qquad (6)$$

gdzie $m_t$ jest masą molekularną proteiny pochodzącej z bazy danych, *r* reprezentuje ilość dopasowań mas teoretycznych i eksperymentalnych, a $f_{ck}$ jest wartością $f_{ij}$ z tabeli częstotliwości mas, odpowiadającą *k*-temu peptydowi teoretycznemu dopasowanemu poprawnie do mas eksperymentalnych. Z postaci formuły funkcji scoringowej MOWSE możemy wyciągnąć kilka wniosków:

- Dopasowania takich mas peptydów, które występują relatywnie często przy trawieniu białek (co możemy zauważyć przez wyższą częstotliwość w znormalizowanej tabeli MOWSE), dają niższy wynik proteinie kandydującej.

- Wartość funkcji scoringowej, jaką może uzyskać białko kandydujące, jest odwrotnie proporcjonalna do jego masy, ponieważ białka o dużej masie mają większe prawdopodobieństwo uzyskania znacznej liczby dopasowań mas ich peptydów do mas eksperymentalnych.
- Przyjmuje się, że wartość funkcji scoringowej rośnie wraz ze wzrostem liczby poprawnych dopasowań mas eksperymentalnych i teoretycznych białka kandydującego.

Podstawowa formuła MOWSE (równanie 6) nie jest zbyt skomplikowana, jednak uwzględnienie w niej kilku kluczowych zależności między wartością funkcji a masami peptydów i protein z bazy pozwoliło uzyskać zadowalające wyniki identyfikacji.

MOWSE, chociaż wartości jego funkcji opierały się o pewien rozkład mas w bazie danych, nie był typowo probabilistyczny. Dopiero w roku 1999 Darryl Pappin z zespołem wykorzystał schemat MOWSE do budowy scoringowej funkcji probabilistycznej, włączonej w pełne oprogramowanie do identyfikacji białek metodą PMF – **Mascot** [37]. Mascot jest obecnie najbardziej znanym komercyjnym narzędziem do identyfikacji białek. Zarówno kod aplikacji, jak i podstawy matematyczne funkcji scoringowej Mascota są niejawne, wiadomo natomiast, że wynik dopasowania mas teoretycznych i eksperymentalnych w programie Mascot bazuje na funkcji MOWSE, a jego wartość odzwierciedla prawdopodobieństwo sytuacji, że dopasowanie białka eksperymentalnego i białka z bazy jest przypadkowe.

Oznaczanie wartości funkcji scoringowej poprzez prawdopodobieństwo może być dla odbiorcy mało intuicyjne, ponieważ bardzo małe prawdopodobieństwo oznacza w rzeczywistości wysoki wynik. Dodatkowo, przedział możliwych wartości prawdopodobieństwa jest na tyle mały, że ich szybkie rozróżnianie jest utrudnione. W konsekwencji nie tylko wartość funkcji scoringowej Mascota, ale też wartości wielu innych funkcji probabilistycznych są wyrażone jako ujemny logarytm z obliczonego prawdopodobieństwa (7):

$$Score_{MOWSE} = -10 \times log_{10}(P) \qquad (7)$$

***Probability Based Scoring Function* (PBSF)**

Mascot nie jest jednak jedyną funkcją scoringową powstałą w oparciu o model MOWSE. Inną funkcją scoringową poszerzającą możliwości funkcji MOWSE jest tak zwana ***Probability Based Scoring Function*** (PBSF). W 2007 roku Zhao Song wraz z zespołem zaproponowali kilka różnych, nowych funkcji scoringowych opartych na tabeli rozkładu mas peptydów względem mas protein w bazie danych (MOWSE), w tym: PBSF [13]. W odróżnieniu od Mascota, PBSF ma jawny model matematyczny, umożliwiający niezależną implementację funkcji bez korzystania z rozwiązań komercyjnych. Tak jak Mascot, PBSF szacuje prawdopodobieństwo wystąpienia zaistniałego dopasowania mas teoretycznych i eksperymentalnych między dwoma białkami. Oznaczmy $H_k$ jako zbiór mas dopasowanych pomiędzy spektrum eksperymentalnym i pewną $k$-tą proteiną z bazy. Załóżmy także, że $n_{ij}^k$ oznacza liczbę peptydów komórki $(i, j)$ tabeli MOWSE dla $k$-tej proteiny [1]. Jeżeli $N_j$ reprezentuje sumę wszystkich peptydów $j$-tej kolumny tabeli, to $m_{ij}$ określa średnią liczbę wystąpień peptydów w komórce $(i, j)$ tabeli MOWSE dla pojedynczej proteiny z bazy danych (8):

$$m_{ij} = n_{ij} \setminus N_j \qquad (8)$$

gdzie $n_{ij}$ oznacza liczbę zliczonych peptydów w komórce $(i, j)$ tabeli MOWSE. Dalej oznaczmy jako $M_j$ liczbę wszystkich średnich wystąpień peptydów w $j$-tej kolumnie w następujący sposób (9):

$$M_j = \sum_{i-1}^{r} m_{ij} \qquad (9)$$

gdzie $r$ oznacza liczbę wierszy tabeli MOWSE. Dla tak zdefiniowanych wartości ostateczny wzór funkcji scoringowej PBSF przedstawia się następująco (10):

$$Score_{MOWSE} = \prod_{i=R(l),\, l \in H_k} \left[ 1 - \left( 1 - \frac{m_{ij}}{M_i} \right)^{n_{ij}^k} \right] \qquad (10)$$

gdzie $R(l)$ definiuje wiersz odpowiadający masie $l$-tego peptydu eksperymentalnego. Warto zauważyć, iż w powyższym wzorze formuła $\frac{m_{ij}}{M_i}$ opisuje częstotliwość występowania mas peptydów, która jednak w znaczny sposób różni się od częstotliwości występowania mas zdefiniowanej w podstawowej wersji funkcji MOWSE. Testy pokazały przewagę metody PBSF w stosunku do podstawowej metody MOWSE, wiążącą się przede wszystkim z bardziej złożonym i dokładnym podejściem do rozkładu wszystkich mas peptydów oraz rozkładu mas dopasowanych [13].

**OLAV-PMF**

Inną znaną scoringową funkcją probabilistyczną, opracowaną w roku 2004, jest **OLAV-PMF** [38]. Funkcja scoringowa OLAV-PMF prezentuje dość skomplikowany model matematyczny oparty na testowaniu hipotez statystycznych. W funkcjach scoringowych stosowanych w bioinformatyce popularnym podejściem jest wykorzystywanie funkcji logitowych [39]. Są one również obecne w OLAV-PMF, ponieważ wynik dopasowania białka eksperymentalnego do kandydującego białka z bazy jest obliczany na podstawie poniższej formuły (11):

$$log\left(\frac{P(E \mid H_1)}{P(E \mid H_0)}\right) \qquad (11)$$

W powyższym wyrażeniu (11) $E$ jest zdarzeniem opisującym wartość wzajemnego dopasowania protein lub peptydów. $H_1$ reprezentuje hipotezę, że dopasowanie to jest poprawne, natomiast hipoteza zerowa $H_0$ mówi o tym, że dopasowanie nastąpiło przypadkowo. Analogicznie, $P(E \mid H_1)$ opisuje zatem prawdopodobieństwo zdarzenia, że dopasowanie, które jest poprawne, skutkujące pewną konkretną wartością ($E$), a $P(E \mid H_0)$ – że pewna wartość dopasowania wystąpiła w sposób przypadkowy. Funkcja OLAV-PMF składa się z trzech głównych niezależnych komponentów:

- pokrycia sekwencji kandydującego białka ($C_1$),
- ułożenia aminokwasów w peptydach ($C_2$),
- występowania modyfikacji posttranslacyjnych ($C_3$).

Z każdym z wymienionych komponentów jest związane pewne prawdopodobieństwo, które obliczane jest dla obu hipotez ($H_0$ i $H_1$). Aby obliczyć te prawdopodobieństwa, dla obu hipotez konstruowane są zbiory uczące, odpowiednio $T_0$ dla $H_0$ i $H_1$ dla $H_1$. $T_1$ zawiera w sobie tylko poprawnie dopasowane białka, natomiast zbiór $T_0$ losowo wygenerowane sekwencje, które są następnie dopasowywane do białek eksperymentalnych wykorzystanych przy generowaniu $T_1$. Na podstawie tych dwóch zbiorów ($T_0$ i $T_1$) generowane są rozkłady prawdopodobieństwa, dla których wyznacza się prawdopodobieństwa poszczególnych komponentów funkcji scoringowej dla różnych hipotez. Ostatecznie wartość funkcji scoringowej OLAV-PMF, dla $k$-ego białka z bazy danych, opisuje następująca formuła:

$$Score_{OLAV\text{-}PMF}(k) = log(C_1 C_2 C_3) \qquad (12)$$

**Nieprobabilistyczne funkcje scoringowe**

**ChemScore**

Przykładem znanej nieprobabilistycznej metody scoringowej jest **ChemScore**. ChemScore opisany w 2002 roku przez Kennetha C. Parkera jako część programu ChemApplex [40] prezentuje podejście, w którym wartość funkcji scoringowej zależy od różnorodnych chemicznych własności eksperymentu spektrometrii masowej, jego warunków, procesu jonizacji cząstek i trawienia białek. Funkcja ChemScore uwzględnia w swoim

sformułowaniu intensywność sygnału peptydów na spektrum masowym, błąd dopasowania mas teoretycznych i eksperymentalnych oraz tzw. główny komponent. Wartość głównego komponentu ChemScore jest wyznaczana na podstawie następujących założeń:

- Technika jonizacji MALDI wykrywa pewne rodzaje peptydów bardziej wydajnie, z drugiej strony jednak jest wrażliwa na występowanie w próbce pewnych rodzajów modyfikacji.
- Różna ilość trypsyny dodanej do próbki skutkuje różnym stopniem wytrawienia sekwencji w próbce.
- Wskutek warunków eksperymentalnych pewne peptydy ulegają modyfikacjom.
- Spektrometry prezentują różny zakres mas obserwowanych na spektrum, a masy peptydów przekraczające maksimum w tym zakresie otrzymują 0 punktów w ChemScore.

Warto wspomnieć, że ChemScore jest metodą przeznaczoną do identyfikacji białek za pomocą spektrometrów MALDI i próbek trawionych trypsyną, ponieważ skupia się ona na specyficznych własnościach chemicznych tych metod. ChemScore bierze pod uwagę charakterystyczną tylko dla MALDI siłę sygnału pewnych peptydów, jest zatem również funkcją, której nie da się uogólnić do wykorzystania w innych typach analiz. Ostateczna postać funkcji scoringowej zaimplementowana w oprogramowaniu Chemperek ma następującą postać (13):

$$Score_{CombChemScore} = \frac{Intensity \times ChemScore}{Abs(ppm\ error)} \qquad (13)$$

**Oprogramowanie**

W Internecie dostępnych jest bardzo wiele programów przeprowadzających pełną identyfikację PMF danych pozyskanych z wykorzystaniem spektrometru masowego [41, 42]. Programy uwzględniają przeszukiwanie baz danych, scoring oraz walidację wyników, jednak na wejściu muszą otrzymywać listy pików (dane z widm masowych), już po wstępnej obróbce.

Kategorie oprogramowania możemy podzielić według różnych czynników. Po pierwsze, przez wzgląd na sposób użytkowania, czyli dostępne zdalnie, poprzez interfejs webowy oraz umożliwiające lokalną instalację na komputerze. Aplikacje webowe mają tę przewagę nad rozwiązaniami stacjonarnymi, że identyfikację można przeprowadzić z dowolnego komputera podłączonego do Internetu. W takim oprogramowaniu mamy także za każdym razem gwarancję podłączenia do wyszukiwania w najnowszej wersji bazy danych. W przypadku oprogramowania stacjonarnego zaletą może być szybkość otrzymania wyniku wyszukiwania. Musimy jednak pamiętać o konieczności posiadania na dysku komputera wybranej proteomicznej bazy danych, jeżeli nie mamy z nią połączenia.

Po drugie, możemy podzielić oprogramowanie ze względu na licencję, z jaką jest wydawane. Część oprogramowania jest darmowa, niektóre z tych programów posiadają otwarte źródła, a metody matematyczne, na których bazują, są jawne i opublikowane. Korzystanie z takich programów umożliwia naukowcom nie tylko weryfikowanie poprawności algorytmów, ale także – w przypadku oprogramowania o otwartych źródłach – samodzielne ich ulepszanie. Dodatkowo istnieje również wiele komercyjnych rozwiązań o zamkniętych źródłach i niejawnych algorytmach. Takie rozwiązania często mają przewagę nad darmowymi, ponieważ ich producenci dają możliwość skorzystania ze zdalnych serwerów o dużej mocy obliczeniowej, dzięki którym efekt przeszukiwania bazy danych, nawet z uwzględnieniem kilku modyfikacji, jest bardzo szybki [43].

Do najbardziej popularnych programów do identyfikacji PMF należą [41, 44, 45]:

- **ProFound** znany darmowy program do identyfikacji białek, dostarczany przez nowojorskie uniwersytety, w tym The Rockefeller University. Program bazuje na bayesowskim algorytmie scoringowym, uwzględniając także indywidualne cechy białek [29]. Formularz dostępny na stronie www umożliwia przeszukiwanie wybranych proteomicznych baz danych według standardowych parametrów: enzymu trawiennego, maksymalnej liczby niewytrawionych fragmentów, przynależności gatunkowej białka oraz jego przybliżonej masy i punktu izoelektrycznego. W przeszukiwaniu bazy danych algorytm może także uwzględnić modyfikacje białka, z podziałem na takie, które dotyczą całej jego sekwencji lub jej części. Rezultat przeszukiwania bazy danych jest wyświetlany w postaci listy protein uporządkowanej według wyników uzyskanych na podstawie funkcji scoringowej. Dodatkowo program oblicza wartość oczekiwaną dla każdej z wartości funkcji scoringowej oraz podaje procent pokrycia sekwencji przez masy eksperymentalne dla każdej z protein na liście.
- **Aldente** lub **PeptIdent** jest programem do identyfikacji białek PMF dostępnym jako formularz na stronie www proteomicznego serwera ExPASy [46], zarządzanego przez Szwajcarski Instytut Bioinformatyczny (*Swiss Institute of Bioinformatics*, SIB). Jak zostało już wcześniej opisane w sekcji: *Kalibracja poprawnych dopasowań*, główną zaletą wykorzystanego w Aldente algorytmu jest kalibracja dopasowań mas teoretycznych i eksperymentalnych [47] bazująca na transformacji Hougha [26]. W ramach licencji Aldente jest także dostępny w wersji stacjonarnej, skomercjalizowanej przez firmę GeneBio [48]. W porównaniu z innymi programami Aldente oferuje więcej opcji konfiguracyjnych mających wpływ na przeszukiwanie bazy danych. Oprócz tych standardowych, związanych z eksperymentalnymi własnościami próbki, program uwzględnia także charakterystykę spektrometru, jaki posłużył do wykonania eksperymentu. Wynik przeszukiwania bazy danych również składa się z listy protein uporządkowanej według wartości funkcji scoringowej.
- **MS-Fit** jest częścią pakietu komputerowych narzędzi do analiz proteomicznych, figurujących pod nazwą ProteinProspector [49]. ProteinProspector jest dystrybuowany przez grupę badawczą Uniwersytetu Kalifornijskiego. Wszystkie narzędzia należące do pakietu są dostępne za darmo w Internecie i można z nich korzystać poprzez formularze na stronach www. Oprócz podstawowych omówionych do tej pory opcji, formularz MS-Fit umożliwia także wybór kategorii spektrometru, na jakim został przeprowadzony eksperyment, oraz odfiltrowania z listy mas tych, które oznaczone są jako zanieczyszczenie. Rezultaty identyfikacji są sortowane według wybranego przez użytkownika porządku, a funkcja scoringowa wykorzystywana w algorytmie jest zmodyfikowaną, ulepszoną postacią funkcji MOWSE (sekcja: *MOWSE i Mascot*).
- **pepMapper** jest narzędziem do identyfikacji białek w schemacie PMF wydawanym przez Uniwersytet w Manchesterze. Narzędzie jest darmowe, dostępne na stronie www w postaci formularza. Program jest podzielony na kilka formularzy w zależności od parametrów, które chce określić użytkownik podczas wyszukiwania. Umożliwiają one na przykład przeprowadzenie porównawczej identyfikacji białek dla dwóch lub trzech eksperymentów. Oprócz standardowych parametrów przeszukiwania bazy danych, program umożliwia określenie N-końca kandydującego białka, jeżeli został on sprawdzony eksperymentalnie. Rezultat identyfikacji jest wyświetlany w postaci listy protein uporządkowanej według wyników funkcji scoringowej, jednak w przypadku pepMappera jej wartość jest odwrotnie proporcjonalna do istotności kandydującej proteiny z bazy danych.
- **Mascot** wykorzystywany przez większość naukowców, najbardziej znany komercyjny program do identyfikacji białek metodą PMF [37]. Dla ograniczonej ilości przesyłanych danych, Mascot może być testowany przez interfejs

webowy dostępny na stronie www. Formularz na stronie umożliwia użytkownikowi wybór bazy danych, w której mają być wyszukiwane kandydujące sekwencje białek. Można równie wybrać maksymalną możliwą liczbę niewytrawionych fragmentów sekwencji, enzym trawienny wykorzystany w eksperymencie oraz pochodzenie gatunkowe identyfikowanych białek i ich przybliżoną masę. Dodatkowo można wybierać pomiędzy wieloma ustalonymi i przypadkowymi rodzajami modyfikacji protein, które zostaną wzięte pod uwagę przy przeszukiwaniu bazy danych. Rezultat identyfikacji jest wyświetlany w postaci listy protein uporządkowanej według wyników uzyskanych na podstawie funkcji scoringowej. Dodatkowo Mascot oblicza poziom istotności dla najwyższego wyniku (*p*-wartość) i informuje nas na podstawie histogramu, jak wiele białek otrzymało niskie i wysokie wyniki w fazie scoringu.

## Podsumowanie

Mimo że w dzisiejszych czasach coraz bardziej popularne stają się metody identyfikacji białek z wykorzystaniem spektrometrów tandemowych i algorytmów sekwencjonowania, to metoda *peptide mass fingerprinting* jest ciągle bardzo powszechna. Świadczy o tym fakt nieprzerwanego udoskonalania, zarówno procedury eksperymentalnej, jak i algorytmicznej techniki. Tylko w ciągu dwóch ostatnich lat (2009–2011) naukowcy zaprezentowali udoskonalenia identyfikacji PMF takie, jak: poprawa efektywności z wykorzystaniem ujemnej jonizacji cząstek [50], dostosowanie algorytmu PMF do identyfikacji złożonej mikstury białek [51], nowe funkcje scoringu (funkcja opierająca się na teście Kolmogorova-Smirnova [52], metoda iMOWSE [53]) oraz inne propozycje walidacji wyniku identyfikacji [54].

Postęp w dziedzinie informatyki pozwala coraz bardziej skomplikowanym algorytmom identyfikacji PMF działać w znacznie bardziej wydajny sposób, niż miało to miejsce na przykład 10 lat temu. Również tempo rozrostu proteomicznych baz danych przyczyniło się do poprawy dokładności identyfikacji, ponieważ naukowcy mogą porównywać zawartość swoich próbek eksperymentalnych z nieporównywalnie większą liczbą białek, niż miało to miejsce około 15 lat temu, kiedy metoda zaczęła być stosowana. Nie bez wpływu na upowszechnianie się metody PMF pozostał rozwój Internetu, który pozwolił na korzystanie z wielu, dostępnych również za darmo, aplikacji webowych do identyfikacji białek schematem PMF (sekcja: *Oprogramowanie*).

Ulepszenia techniczne w zakresie analizy spektrometrem masowym umożliwią metodzie PMF dalszy rozwój. Nawet jeżeli w przyszłości wykorzystanie tandemowych spektrometrów masowych stanie się głównym sposobem identyfikacji białek, to metoda PMF przez wzgląd na mniejszy nakład kosztów, czasu i pracy, a także dzięki znacznie szybszemu działaniu algorytmów będzie wciąż stanowić wsparcie dla bardziej złożonych systemów definiowania zawartości białkowej próbek [55]. ■

## Literatura


1. W.J. Henzel, C. Watanabe, J.T. Stults: ***Protein identification: the origins of peptide mass fingerprinting***, Journal of the American Society for Mass Spectrometry, vol. 14, 2003, s. 931-942.
2. D.J. Pappin, P. Hojrup, A. Bleasby: ***Rapid identification of proteins by peptide-mass fingerprinting***, Current biology, vol. 3(6), 1993, s. 327-332.
3. R. Apweiler, A. Bairoch, C. Wu: ***Protein sequence databases***, Current opinion in chemical biology, vol. 8, 2004, s. 76-80.
4. Swiss-Prot, statystyka: http://www.expasy.org/sprot/relnotes/relstat.html
5. TrEMBL, statystyka: http://www.ebi.ac.uk/uniprot/TrEMBLstats/
6. I. Bogdan, R. Beynon, D. Coca: ***Reconfigurable computing solution for Peptide Mass Fingerprinting***, 2008 11th International Conference on Optimization of Electrical and Electronic Equipment, 2008, s. 57-62.
7. R. Matthiesen: ***Methods, algorithms and tools in computational proteomics: a practical point of view***, Proteomics, vol. 7, 2007, s. 2815-2832.
8. I. Eidhammer, K. Flikka, L. Martens, S.-O. Mikalsen: ***Computational methods for mass spectrometry proteomics***, Wiley-Interscience, Chichester 2008, s. 6-118.
9. H. Kamińska, H. Podbielska: ***Identyfikacja białek z wykorzystaniem techniki peptide mass fingerprinting (PMF). Część I – charakterystyka eksperymentu identyfikacji***, Inżynieria Biomedyczna – Acta Bio-Optica et Informatica Medica, vol. 17, 2011, s. 153-160
10. H. Kaltenbach, S. Böcker, S. Rahmann: ***Markov additive chains and applications to fragment statistics for peptide mass fingerprinting***, Joint RECOMB 2006 Satellite Workshops on Systems Biology and on Computational Proteomics, 2007, s. 29-41.
11. R. Aebersold, D.R. Goodlett: ***Mass spectrometry in proteomics, Chemical reviews***, vol. 101, 2001, s. 269-295.
12. B. Thiede, W. Höhenwarter, A. Krah, J. Mattow, M. Schmid, F. Schmidt, P.R. Jungblut: ***Peptide mass fingerprinting***, Methods, vol. 35, 2005, s. 237-247.
13. Z. Song, L. Chen, A. Ganapathy, X.-F. Wan, L. Brechenmacher, N. Tao, D. Emerich, G. Stacey, D. Xu: ***Development and assessment of scoring functions for protein identification using PMF data***, Electrophoresis, vol. 28, 2007, s. 864-870.
14. A.E. Ashcroft: ***Protein and peptide identification: the role of mass spectrometry in proteomics***, Natural Product Reports, vol. 20, 2003, s. 202-215.
15. M. Mann, O. Jensen: ***Proteomic analysis of post-translational modifications***, Nature biotechnology, vol. 21, 2003, s. 255-261.
16. A. Aitken, M. Learmonth: ***Carboxymethylation of cysteine using iodoacetamide/iodoacetic acid***, Humana Press, NY 2002, s. 455-456.
17. J. Cavins, M. Friedman: ***An internal standard for amino acid analyses: S-[beta]-(4-pyridylethyl)-L–cysteine***, Analytical Biochemistry, vol. 35(2), 1970, s. 489-493.
18. A. Chrambach, D. Rodbard: ***Polyacrylamide gel electrophoresis***, Science, vol. 172(3982), 1971, s. 440.
19. J. Garavelli: ***The RESID Database of Protein Modifications as a resource and annotation tool***, Proteomics, vol. 4, 2004, s. 1527-1533.
20. R. Craig, R. Beavis: ***A method for reducing the time required to match protein sequences with tandem mass spectra***, Rapid communications in mass spectrometry, vol. 17, 2003, s. 2310-2316.
21. A. Chernobrovkin, O. Trifonova, N. Petushkova, E. Ponomarenko, A. Lisitsa: ***Selection of the peptide mass tolerance value for protein identification with peptide mass fingerprinting***, Russian Journal of Bioorganic Chemistry, vol. 37, 2011, s. 119-122.
22. S.J. Hubbard: ***Systematic characterization of high mass accuracy influence on false discovery and probability scoring in peptide mass fingerprinting***, Analytical biochemistry, vol. 372, 2008, s. 156-166.
23. V. Egelhofer, K. Büsso, C. Luebbert, H. Lehrach, E. Nordhoff: ***Improvements in protein identification by MALDI-TOF-MS peptide mapping***, Analytical chemistry, vol. 72, 2000, s. 2741-2750.
24. http://www.expasy.org/tools/aldente/
25. M. Tuloup, C. Hernandez, I. Coro, C. Hoogland, P.-A. Binz, R.D. Appel: ***Aldente and BioGraph: An improved peptide mass fingerprinting protein identification environment***, Proceedings of the Swiss Proteomics Society 2003 Congress: ***Understanding Biological Systems through Proteomics***, 2003, s. 174-176.
26. R. Duda, P. Hart: ***Use of the Hough transformation to detect lines and curves in pictures***, Communications of the ACM, vol. 15(1), 1972, s. 11–15.
27. J. Eriksson, B.T. Chait, D. Fenyö: ***A statistical basis for testing the significance of mass spectrometric protein identification results***, Anal. Chem, vol. 72(5), 2000, s. 999-1005.
28. M. Mann, P. Hojrup, P. Roepstorff: ***Use of mass spectrometric molecular weight information to identify proteins in sequence databases***, Biological mass spectrometry, vol. 22(6), 1993, s. 338-345.
29. W. Zhang, B. Chait: ***ProFound: an expert system for protein identification using mass spectrometric peptide mapping information***, Analytical chemistry, vol. 72, 2000, s. 2482-2489.

30. R. Gras, M. Müller, E. Gasteiger, S. Gay, P.-A. Binz, W. Bienvenut, C. Hoogland, J.-C. Sanchez, A. Bairoch, D. F. Hochstrasser i in.: ***Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection***, Electrophoresis, vol. 20, 1999, s. 3535-3550.
31. Q. Ding, L. Xiao, S. Xiong, Y. Jia, H. Que, Y. Guo, S. Liu: ***Unmatched masses in peptide mass fingerprints caused by cross-contamination: an updated statistical result***, Proteomics, vol. 3, 2003, s. 1313-1317.
32. S. Damodaran, T.D. Wood, P. Nagarajan, R.A. Rabin: ***Evaluating peptide mass fingerprinting-based protein identification***, Genomics, proteomics & bioinformatics/Beijing Genomics Institute, vol. 5, 2007, s. 152-157.
33. J. Eriksson, D. Fenyö: ***A model of random mass-matching and its use for automated significance testing in mass spectrometric proteome analysis, Proteomics***, vol. 2, 2002, s. 262-270.
34. J. Eriksson, D. Fenyö i in.: ***Probity: a protein identification algorithm with accurate assignment of the statistical significance of the results***, Journal of Proteome Research, vol. 3(1), 2004, s. 32-36.
35. A. Ganapathy, X.-F. Wan, J. Wan, J. Thelen, D.W. Emerich, G. Stacey, D. Xu: ***Statistical assessment for mass-spec protein identification using peptide fingerprinting approach***, Conference proceedings: ***Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society***, Conference, vol. 4, 2004, s. 3051-3054.
36. http://www.matrixscience.com
37. D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell: ***Probability-based protein identification by searching sequence databases using mass spectrometry data***, Electrophoresis, vol. 20(18), 1999, s. 3551-3567.
38. J. Magnin, A. Masselot, C. Menzel, J. Colinge: ***OLAV-PMF: a novel scoring scheme for high-throughput peptide mass fingerprinting***, Journal of Proteome Research, vol. 3(1), 2004, s. 55-60.
39. J. Cramer: ***The origins and development of the logit model***, Logit models from economics and other fields, 2003, s. 1-19.
40. K.C. Parker: ***Scoring methods in MALDI peptide mass fingerprinting: ChemScore, and the ChemApplex program***, Journal of the American Society for Mass Spectrometry, vol. 13(1), 2002, s. 22-39.
41. D. Fenyö: ***Identifying the proteome: software tools***, Current opinion in biotechnology, vol. 11(4), 2000, s. 391-395.
42. J. Handley: ***Software for MS protein identification***, Analytical chemistry, vol. 74, 2002, s. 159A-162A.
43. I.A. Bogdán, J. Rivers, R.J. Beynon, D. Coca: ***High-performance hardware implementation of a parallel database search engine for real-time peptide mass fingerprinting***, Bioinformatics, vol. 24, 2008, s. 1498-1502.
44. R.C. Beavis, D. Fenyö: ***Database searching with mass-spectrometric information***, Trends in Biotechnology, vol. 18, 2000, s. 22-27.
45. W.-A. Joo, J.-B. Lee, M. Park, J.-W. Lee, H.-J. Kim, C.-W. Kim: ***Comparison of search engine contributions in protein mass fingerprinting for protein identification***, Biotechnology and Bioprocess Engineering, vol. 12(2), 2007, s. 125-130.
46. http://expasy.org/
47. E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, A. Bairoch: ***Protein identification and analysis tools on the ExPASy server***, The proteomics protocols handbook, vol. 112, 2005, s. 571-607.
48. http://www.genebio.com/
49. C. Jiménez, L. Huang, Y. Qiu, A. Burlingame: ***Searching Sequence Databases Over the Internet: Protein Identification Using MS-Fit***, Wiley Online Library, 2001.
50. T. Sanaki, M. Suzuki, S. H. Lee, T. Goto, T. Oe: ***A simple and efficient approach to improve protein identification by the peptide mass fingerprinting method: concomitant use of negative ionization***, Analytical Methods, vol. 2(8), 2010, s. 1144.
51. Z. He, C. Yang, W. Yu: ***A partial set covering model for protein mixture identification using mass spectrometry data***, IEEE/ACM transactions on computational biology and bioinformatics / IEEE, vol. 8(2), 2011, s. 368-380.
52. R. Jain, M. Wagner: ***Kolmogorov-Smirnov scores and intrinsic mass tolerances for peptide mass fingerprinting***, Journal of proteome research, vol. 9, 2010, s. 737-742.
53. S. K.-W. Tsui, K.-K. Leung: ***iMOWSE, a scoring scheme bridging in silico and in vitro digestion in peptide mass fingerprints***, 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop, 2009, s. 344.
54. Z. Song, L. Chen, D. Xu: ***Confidence assessment for protein identification by using peptide-mass fingerprinting data***, Proteomics, vol. 9, 2009, s. 3090-3099.
55. I. Shadforth, D. Crowther, C. Bessant: ***Search-space reduction of a non-redundant peptide database***, Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004, 2004, s. 450-451.


***otrzymano / received: 20.02.2011***
***zaakceptowano / accepted: 23.05.2011***