

Uczenie maszynowe w bioinformatyce

Wykład 13: proteomika

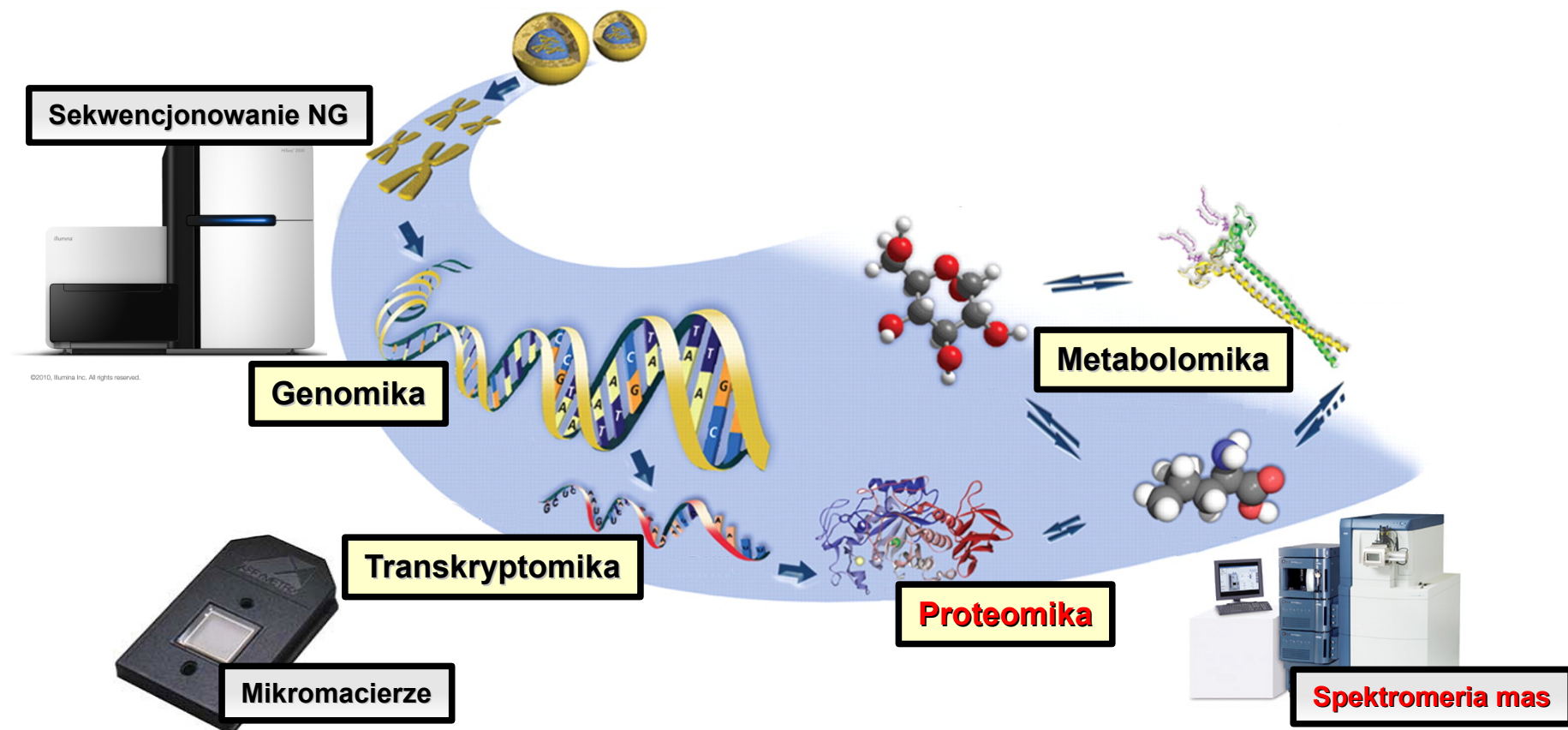
Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

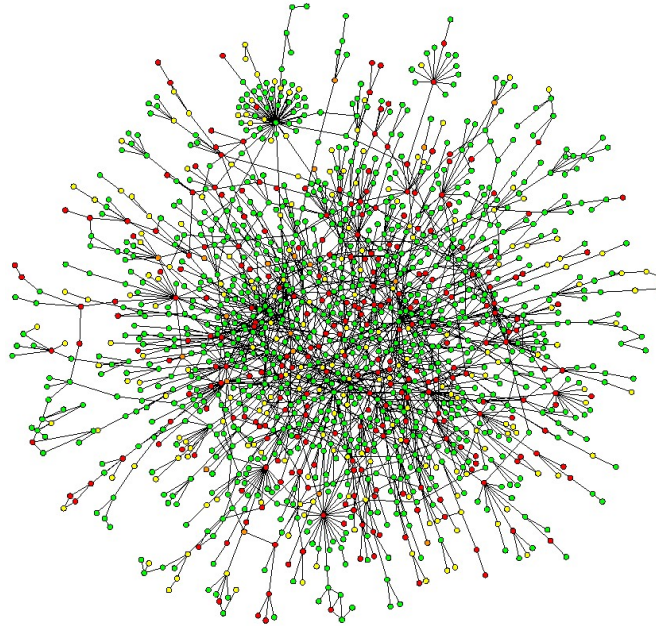
Proteomika

Proteom jest to całkowity zbiór białek kodowanych przez genom organizmu, wraz z ich możliwymi modyfikacjami. Termin ten może być stosowany również w węższym znaczeniu i rozumiany jako zbiór białek ulegających ekspresji w danym rodzaju komórek i w danym czasie. Różnica pomiędzy tymi definicjami wynika z faktu, że proteom jest tworem dynamicznym: jego chwilowy skład różni się pomiędzy komórkami organizmu i zmienia się w czasie, w zależności od fazy ich rozwoju, czynników środowiskowych oraz oddziaływań z innymi komórkami.

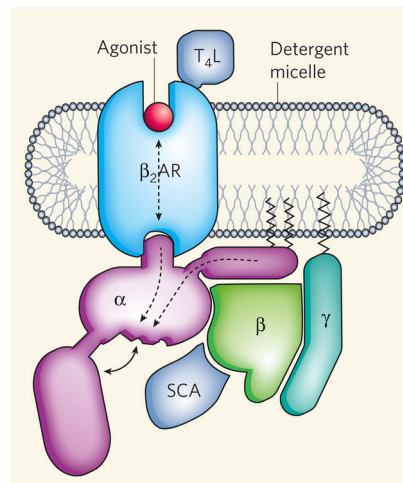
Proteomika jest dziedziną biologii molekularnej zajmującą się badaniem proteomu, zarówno w sposób jakościowy, jak i ilościowy.



Złożoność i różnorodność funkcji biologicznych białek jest przyczyną powstawania rozbudowanych sieci wzajemnych zależności pomiędzy nimi.



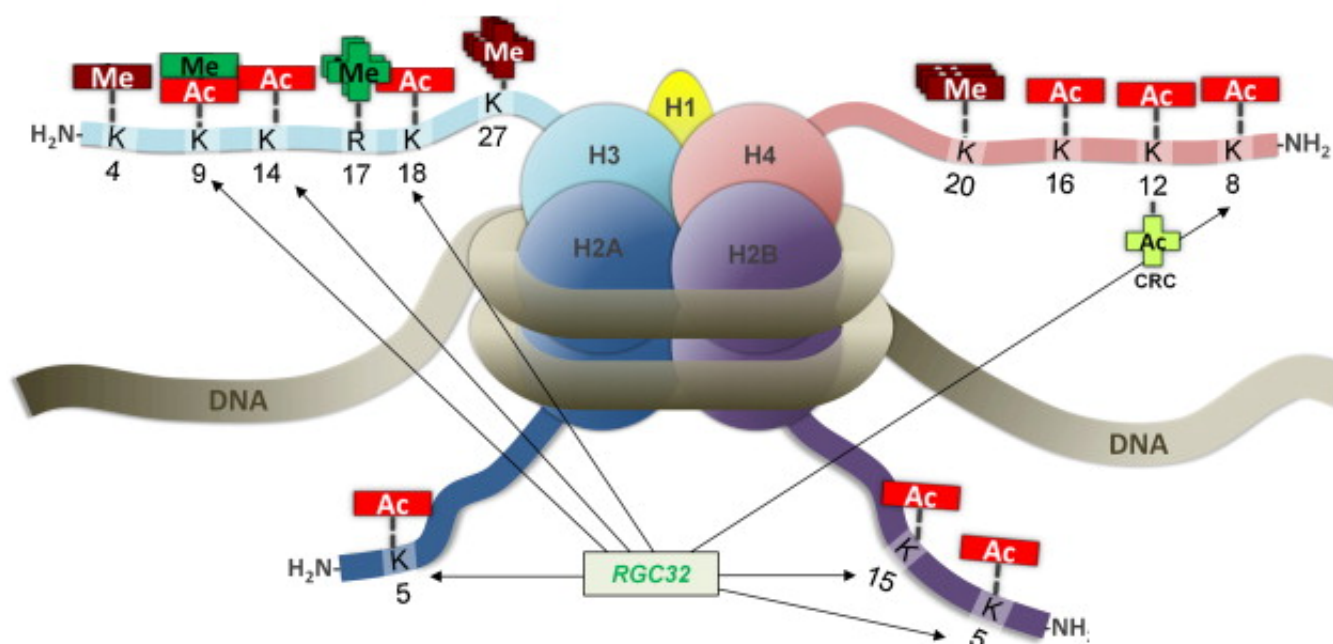
Często również ściśle współdziałają one ze sobą, tworząc kompleksy białkowe.



Badanie tego typu interakcji jest jednym z głównych celów proteomiki.

Na właściwości białek i ich aktywność biologiczną wpływają również **modyfikacje potranslacyjne (PTM – Posttranslational Modifications)**, polegające na przyłączeniu do aminokwasów cząstek niewielkich związków chemicznych.

Przykładami roli modyfikacji potranslacyjnych mogą być acetylacja (dodanie grupy acetylowej COCH_3 do lizyn) i metylacja (dodanie jednej, dwóch lub trzech grup metylowych CH_3 do lizyn lub arginin) białek histonowych. Modyfikacje te zmieniają lokalne właściwości chromatyny, wpływając tym samym na ekspresję genów.



Inną bardzo ważną modyfikacją jest fosforylacja (dodanie reszty fosforanowej PO_4 do seryn, treonin i tyrozyn), mająca kluczowe znaczenie dla sygnalizacji komórkowej.

Wielkoskalowe metody biologii molekularnej

Wykład 13: spektrometria mas (MS)

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

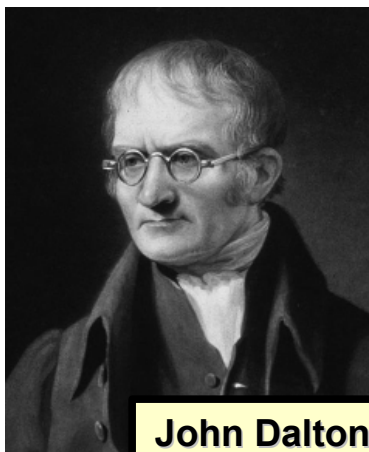
Spektrometria mas

Podstawowym narzędziem proteomiki jest obecnie **spektrometria mas (MS – Mass Spectrometry)** – technika analityczna umożliwiającą pomiar mas cząstek po ich uprzedniej jonizacji i przetransferowaniu do fazy gazowej.

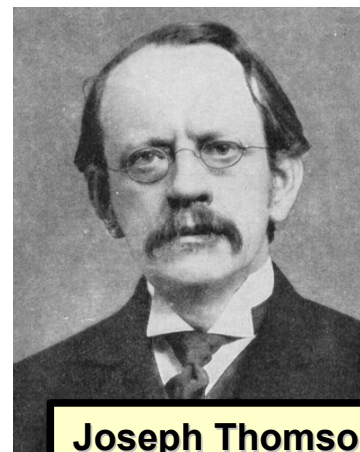
Masa wyrażana jest w **ujednoliconych jednostkach masy atomowej (u)** lub **daltonach (Da)**. Obie jednostki są równoważne i wynoszą $1,6605402 \cdot 10^{-27}$ kg, co odpowiada 1/12 masy pojedynczego atomu izotopu węgla ^{12}C .

W spektrometrii mas badane są jony, tak więc pomiar nie dotyczy bezpośrednio masy: mierzona jest wartość m/z , czyli wartość **stosunku masy jonu do jego ładunku**.

Jednostką stosunku masy do ładunku jest m/z **thompson (Th)**, zdefiniowany jako $1 \text{ Da}/e$, gdzie e jest elementarnym ładunkiem elektrycznym. Jednak bardziej powszechne jest stosowanie bezwymiarowej wartości m/z . W takim przypadku m oznacza względną masę jonu (odniesioną do 1 Da), a z jest stopniem naładowania, czyli liczbą niesionych przez jon ładunków.



John Dalton

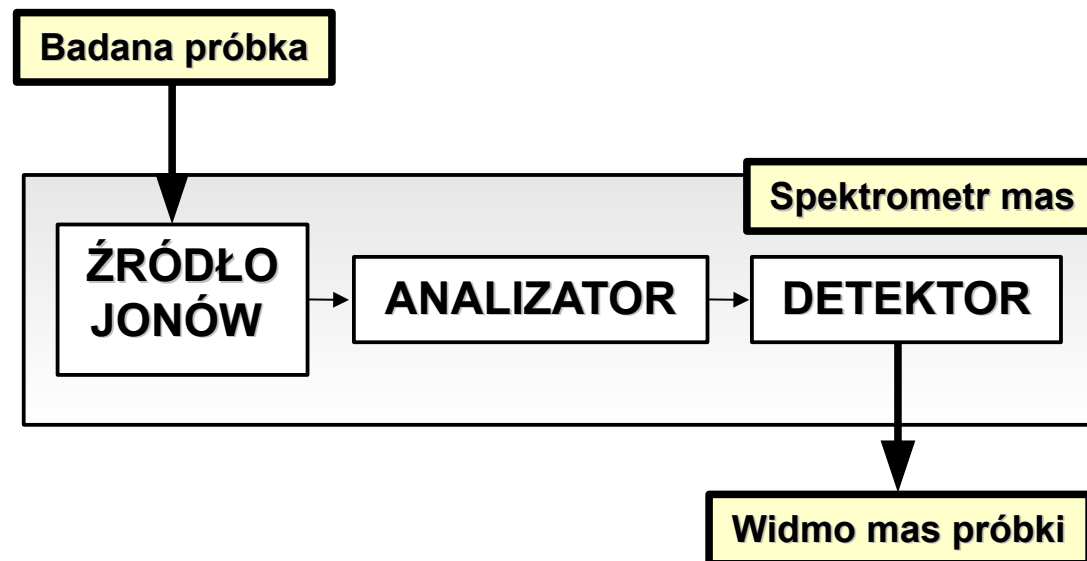


Joseph Thomson

Spektrometria mas: budowa spektrometru mas

W największym uproszczeniu można powiedzieć, że spektrometr mas składa się z trzech podstawowych części:

- **źródła jonów**, w którym obojętne cząstki zamieniane są na jony;
- **analizatora**, odpowiedzialnego za rozdzielanie jonów o różnych stosunkach masy do ładunku;
- **detektora**, zliczającego liczby jonów o poszczególnych wartościach m/z , które są zapamiętywane w postaci **widma mas**.



Spektrometria mas: budowa spektrometru mas (źródło jonów)

Wybór sposobu jonizacji jest podyktowany właściwościami badanych substancji. Peptydy i białka, jako związki zawierające w swej strukturze grupy funkcyjne łatwo ulegające protonacji, **jonizowane są przez przyłączenie jednego lub większej liczby protonów.**

Przy tego typu jonizacji, jeżeli względna masa cząstki wynosi M_0 , a stopień naładowania równy jest z , to wartość m/z jonu można określić z zależności:

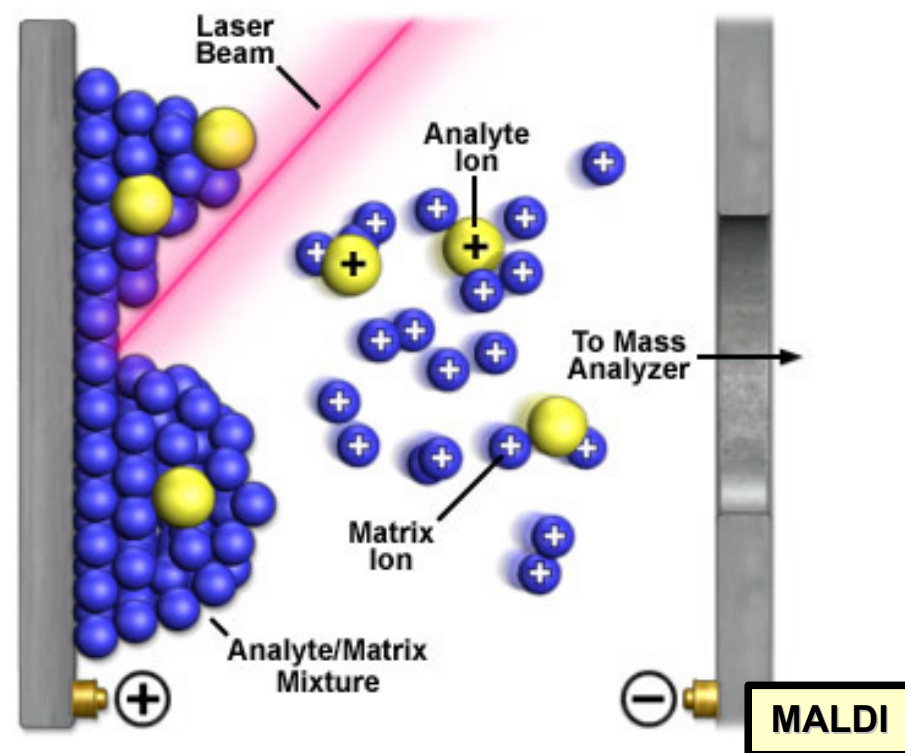
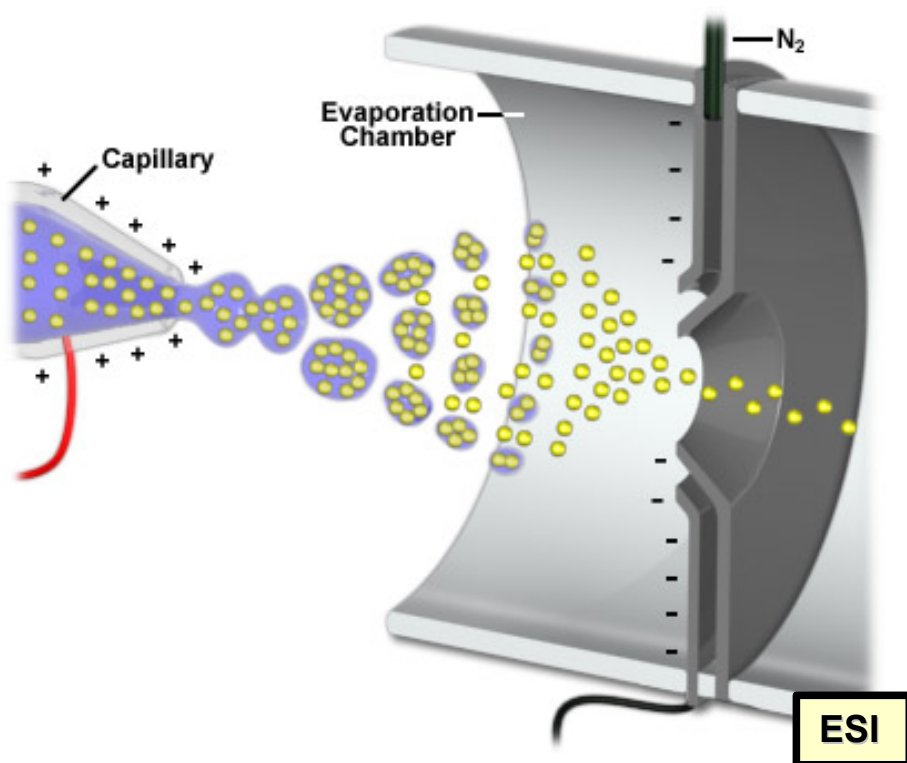
$$m_0 = \frac{M_0 + z M_p}{z}$$

gdzie M_p jest masą protonu.

Jonizacja i przetransferowanie do fazy gazowej dużych biomolekuł, takich jak białka i peptydy nie jest zadaniem prostym, gdyż są to cząstki nielotne i łatwo ulegające rozpadowi pod wpływem temperatury. Aby nie doprowadzić do ich niekontrolowanej fragmentacji cząstek konieczne jest stosowanie metod „łagodnej jonizacji”.

Spektrometria mas: budowa spektrometru mas (źródło jonów)

Dwie najczęściej stosowane w badaniach proteomicznych techniki jonizacji są **elektrozpylanie (ESI – *ElectroSpray Ionisation*)** oraz **jonizacja przez desorpcję laserową w matrycy (MALDI – *Matrix Assisted Laser Desorption/ Ionization*)**.

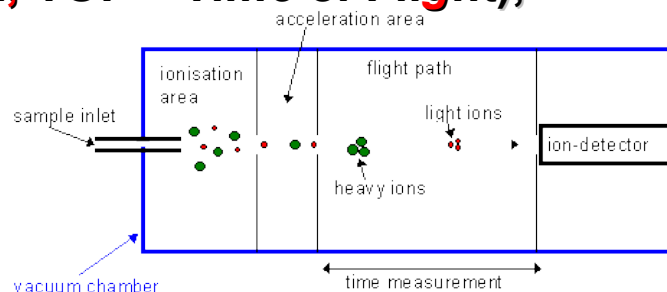


[nationalmaglab.org]

Spektrometria mas: budowa spektrometru mas (analyzer)

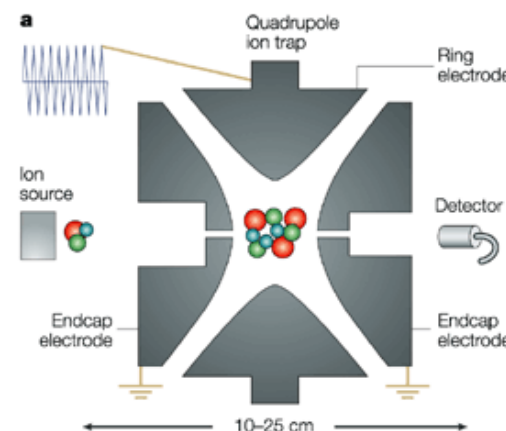
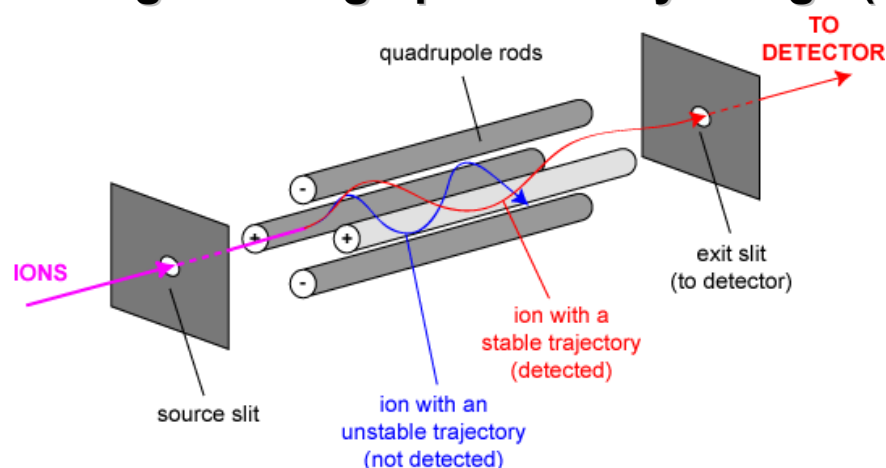
Zadaniem **analyzera** jest rozdzielenie jonów pod względem wartości m/z i skierowanie ich skupionej wiązki do detektora. W zależności od typu analyzera, podstawą rozdzielania jonów charakteryzujących się różnymi wartościami może być:

- pomiar prędkości z jaką poruszają się jony po przyspieszeniu w polu elektrycznym (**analyzer czasu przelotu, TOF – Time of Flight**),



[alevelnotes.com]

- selektywna destabilizacja torów lotu lub orbit jonów pod wpływem superpozycji zmiennego i stałego pola elektrycznego (**kwadrupol i pułapka jonowa**),



- pomiaru częstotliwości cyklotronowej ruchu obrotowego jonów w polu magnetycznym (analyzerze **cyklotronowego rezonansu jonów z transformacją Fouriera, FT-ICR – Fourier Transform Ion Cyclotron Resonance**)

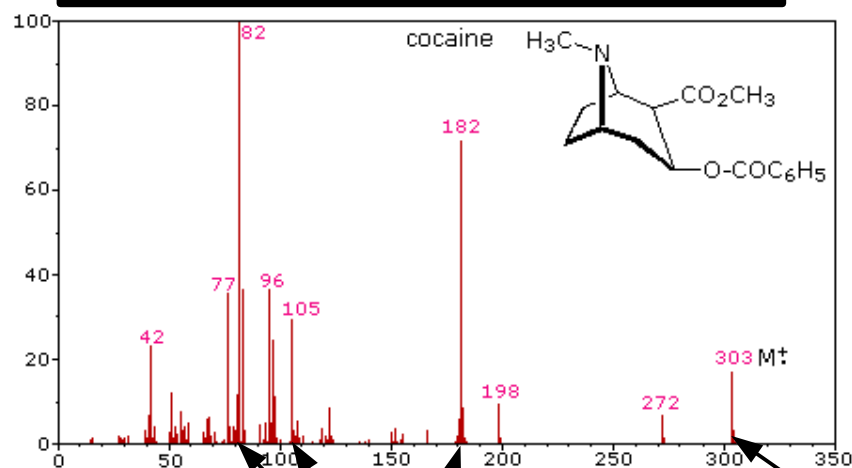
Spektrometria mas: widmo mas

Rozdzielone w analizatorze jony o różnych wartościach m/z są zliczane w detektorze, a ich liczby zapamiętywane są w postaci **widma mas**.

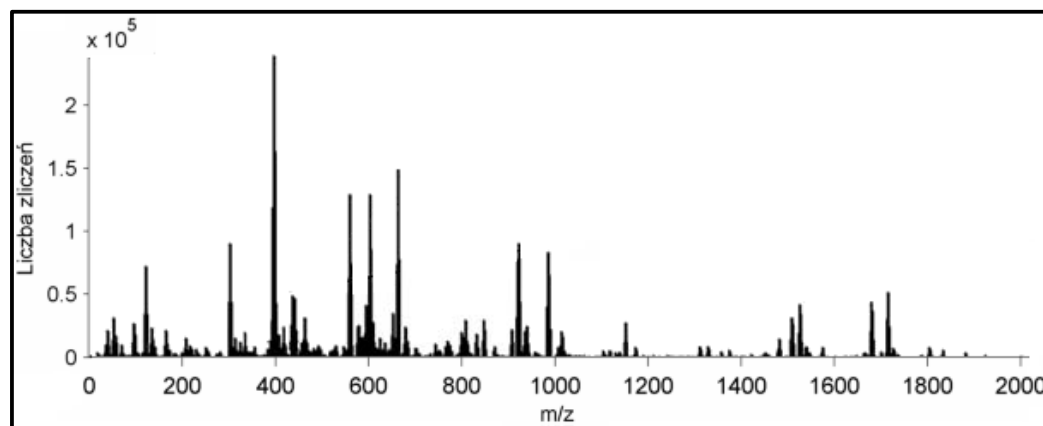
Widmo mas jest zwykle prezentowane w postaci wykresu liczby zarejestrowanych jonów o poszczególnych wartościach m/z . Często wykres jest znormalizowany wobec najwyższego piku.

Występujące w widmie piki mogą odpowiadać całym jonom molekularnym związków występujących w badanej próbce lub jonom ich mniejszych fragmentów.

Znormalizowane widmo mas związku chemicznego ($C_{17}H_{21}NO_4$, $M = 303.147$)



Widmo mas próbki biologicznej złożonej z dużej liczby białek



Pik jonu molekularnego

Piki fragmentów jonu molekularnego

Spektrometria mas: tandemowa spektrometria mas (MS/MS)

W przypadku cząstek o dużych masach pomiar m/z jonów molekularnych może nie być wystarczający do jednoznacznej identyfikacji. Przykładowo, nawet bardzo dokładany pomiar masy peptydu pozwala jedynie określić skład aminokwasowy i nie niesie informacji o kolejności występowania reszt aminokwasowych w sekwencji:

Zmierzona masa:	Skład chemiczny:	Możliwe sekwencje:
1861.032250	$\text{H}_{136} \text{C}_{84} \text{N}_{26} \text{O}_{22}$	APIRPDIVNFVHTNLR PIRPDIVNFVHTNLR IRPDIVNFVHTNLRAP ... RLNTHVFNVIDPRIPA

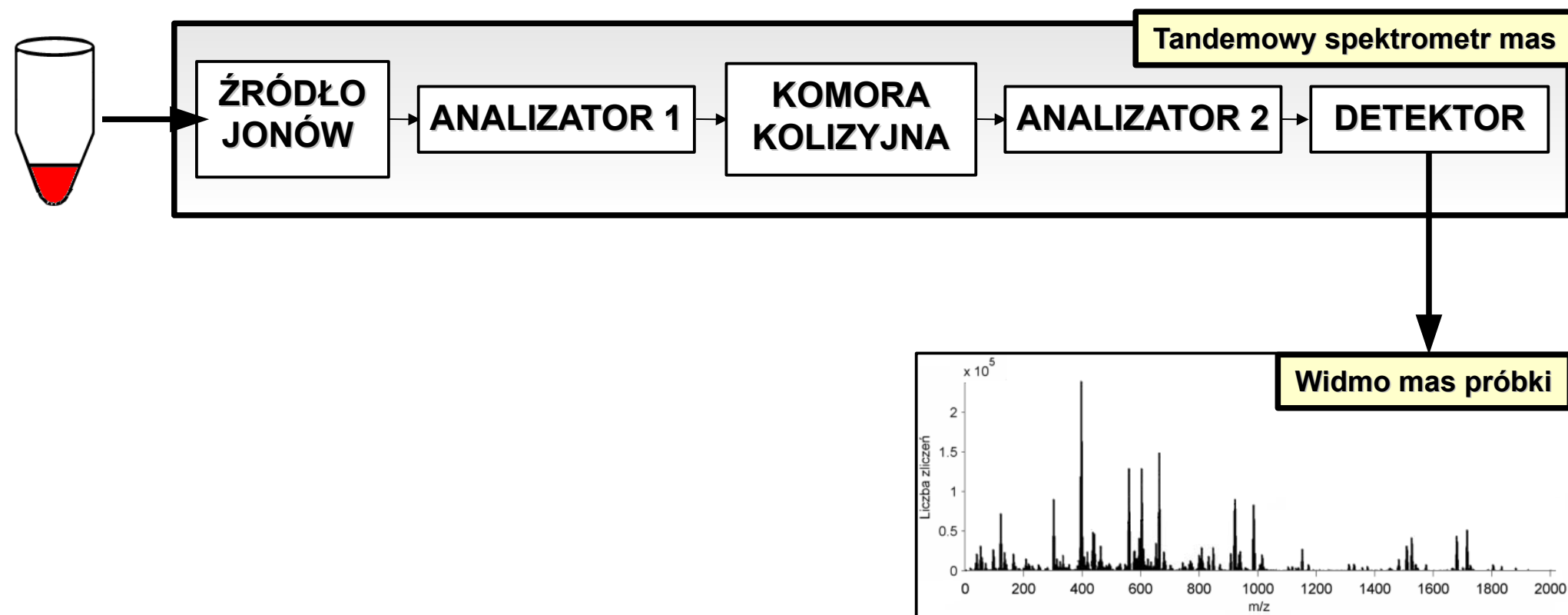
Wnioskowanie o kolejności reszt aminokwasowych w sekwencji możliwe jest dopiero po doprowadzeniu do dysocjacji wiązań w cząstce i zmierzeniu mas powstałych w ten sposób fragmentów. Wymaga to wykonania dwóch cykli analizy spektrometrycznej, pomiędzy którymi następuje **fragmentacja**, najczęściej poprzez **dysocjację wywołaną kolizjami (CID – Collision Induced Dissociation)** z cząstkami gazu obojętnego.

Spektrometry umożliwiające przeprowadzenie tego typu pomiaru nazywa się **spektrometrami tandemowymi**.

Spektrometria mas: tandemowa spektrometria mas (MS/MS)

Najprostszy tandemowy spektrometr mas składa się z dwóch połączonych szeregowo analizatorów (najczęściej kwadrupoli), przedzielonych komorą kolizyjną. Przyrząd taki może pracować w dwóch trybach: zbierania pełnego widma (**tryb MS**) oraz tandemowym (**tryb MS/MS**, w proteomice nazywany **trybem sekwencjonowania**).

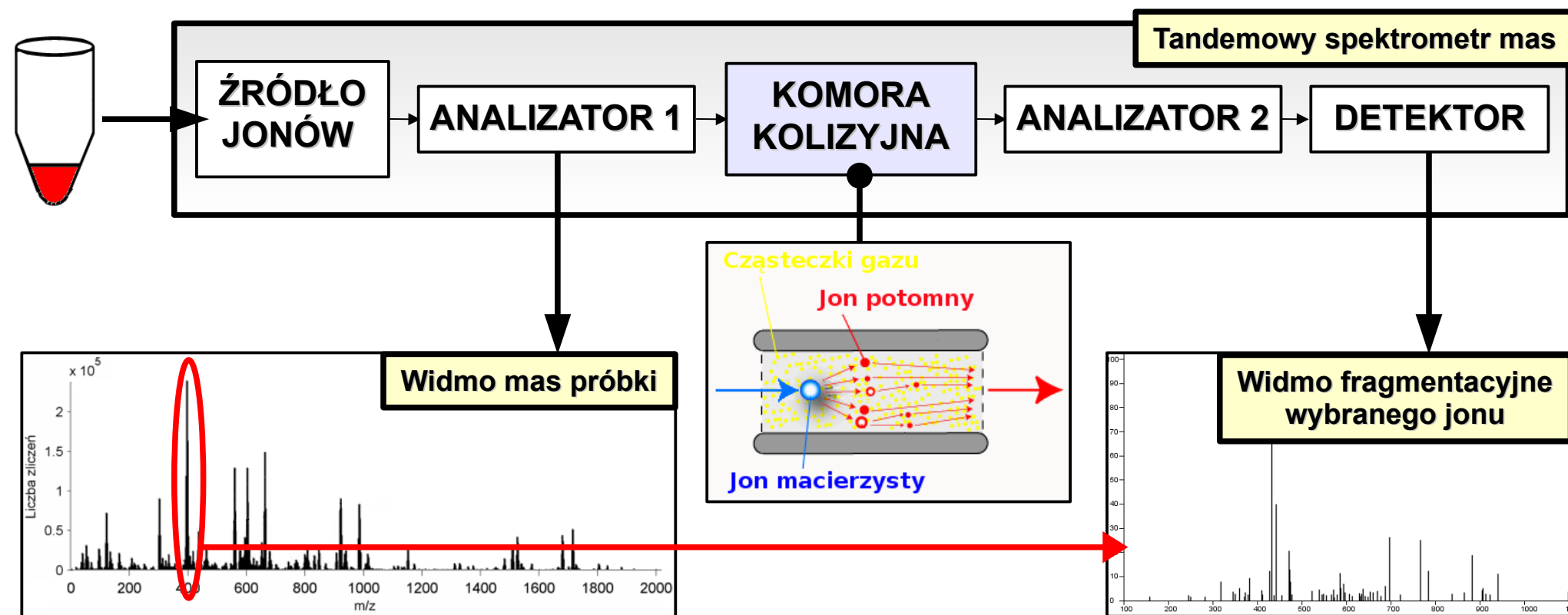
Tryb MS: pierwszy analizator pełni funkcję kolimatora wiązki, komora kolizyjna jest pusta, a separacja jonów następuje w drugim analizatorze (czyli przyrząd zachowuje się tak jak klasyczny spektrometr).



Spektrometria mas: tandemowa spektrometria mas (MS/MS)

Najprostszy tandemowy spektrometr mas składa się z dwóch połączonych szeregowo analizatorów (najczęściej kwadrupoli), przedzielonych komorą kolizyjną. Przyrząd taki może pracować w dwóch trybach: zbierania pełnego widma (**tryb MS**) oraz tandemowym (**tryb MS/MS**, w proteomice nazywany **trybem sekwencjonowania**).

Tryb MS/MS: pierwszy analizator przepuszcza tylko jony o określonym stosunku masy do ładunku, zwane **jonami macierzystymi**. Jony te ulegają fragmentacji w komorze kolizyjnej, która wypełniana jest na ten czas gazem szlachetnym (zwykle argonem). **Widmo fragmentacyjne (widmo MS/MS)**, czyli wartości m/z i liczby **jonów potomnych** będących efektem fragmentacji powstaje w drugim analizatorze.

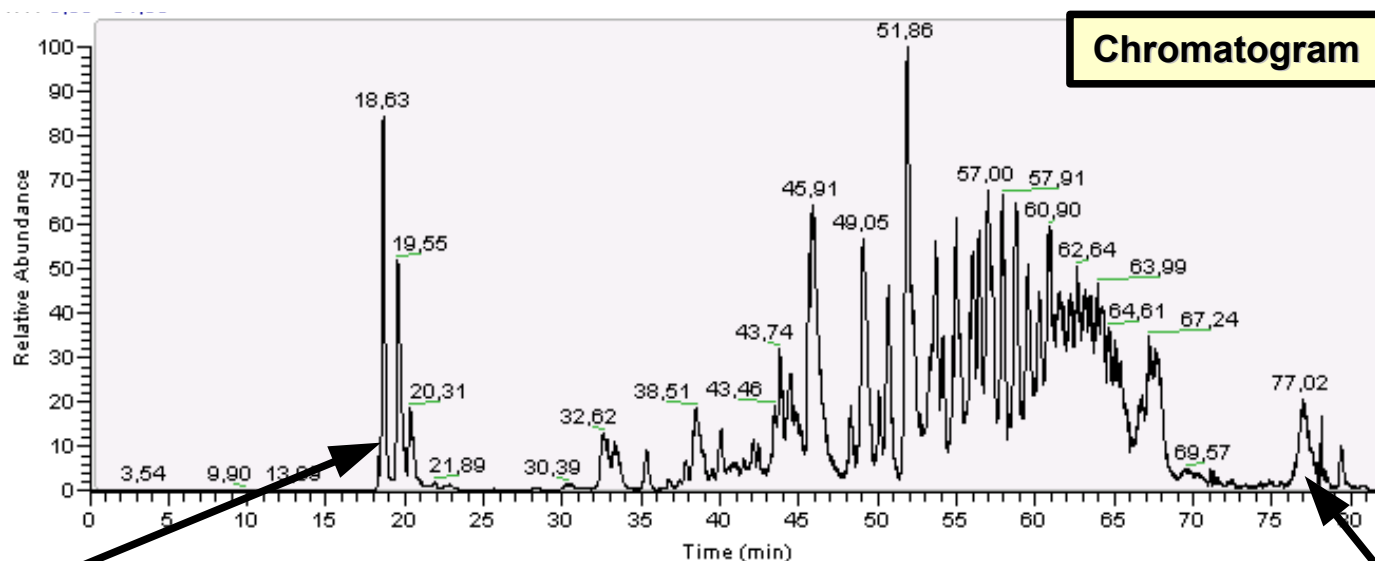


Spektrometria mas: sprzężanie MS z HPLC (LC-MS)

W przypadku badania złożonych mieszanin znaczną poprawę czułości analizy można uzyskać poprzez połączenie spektrometru z systemem chromatograficznym.

Chromatografia jest metodą rozdzielania składników mieszanin w wyniku ich różnego podziału pomiędzy fazę ruchomą i stacjonarną. Przemieszczanie się wzdłuż układu chromatograficznego jest możliwe tylko w fazie ruchomej, dlatego też dłużej będą w nim przebywać składniki mieszaniny silniej oddziałujące z fazą stacjonarną.

Rozdzielonym składnikom, opuszczającym układ chromatograficzny w różnym czasie, odpowiadać będą osobne piki w **chromatogramie** rejestrowanym przez detektor. Czas od wprowadzenia badanej mieszaniny do detekcji maksimum pików nazywany jest **czasem retencji**.

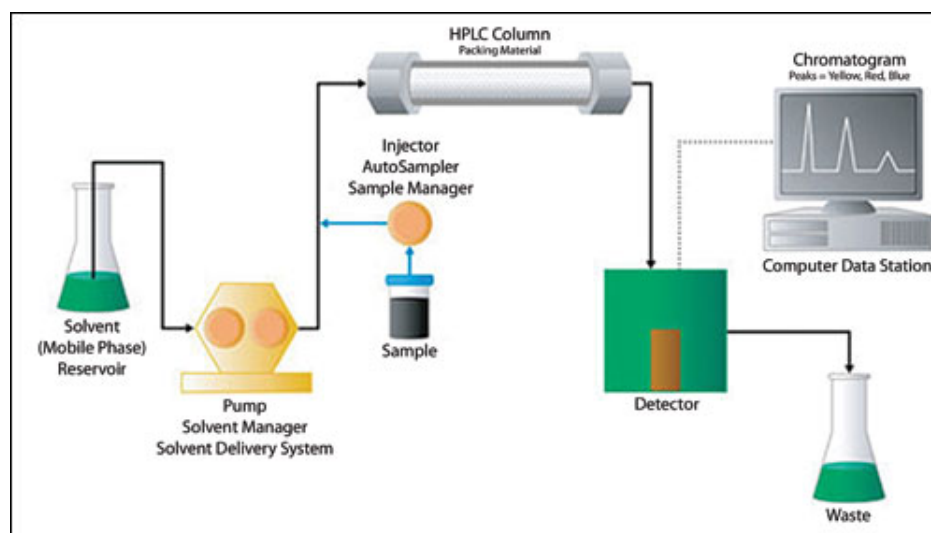


Piki odpowiadające cząstkom słabo oddziałującym z fazą stacjonarną

Piki odpowiadające cząstkom silnie oddziałującym z fazą stacjonarną

Spektrometria mas: sprzężanie MS z HPLC (LC-MS)

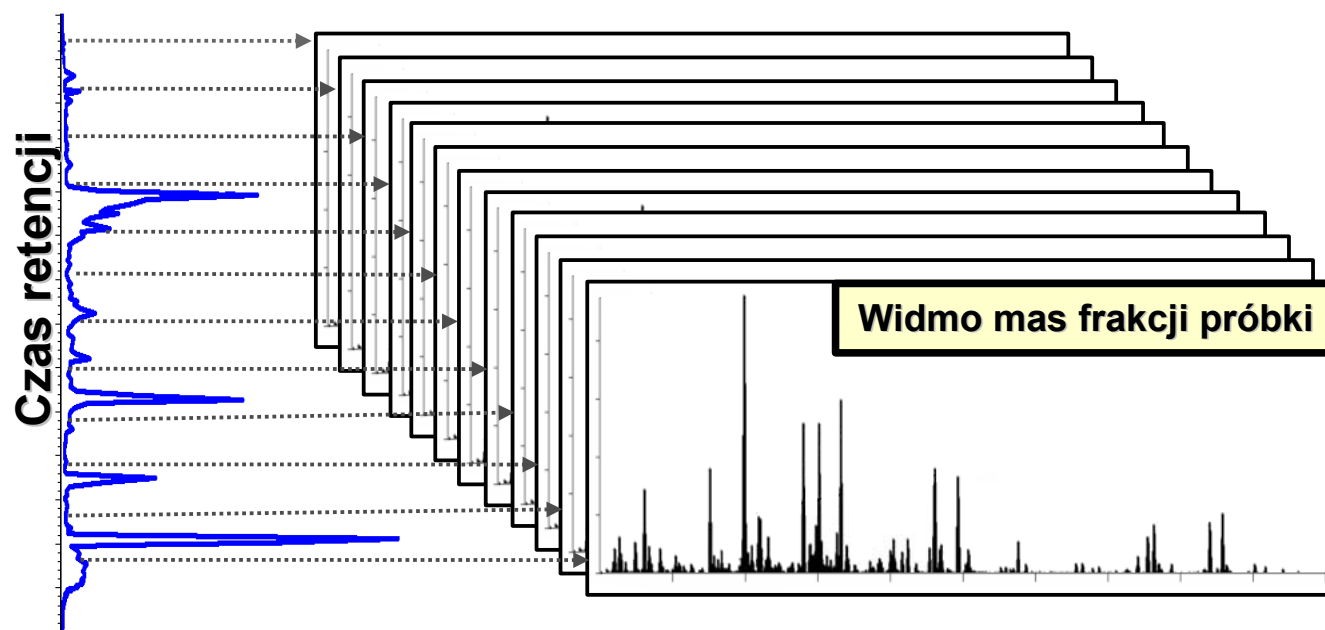
W zależności od stosowanej techniki chromatograficznej, fazą ruchomą może być gaz, ciecz lub fluid, a fazą stacjonarną ciało stałe lub ciecz. W przypadku peptydów i białek stosowana jest technika **LC-MS (Liquid Chromatography – Mass Spectrometry)**, w której wstępne rozdzielanie składników próbki następuje w sprzężonym ze spektrometrem systemie **wysokosprawnej chromatografii ciekowej (HPLC – High Performance Liquid Chromatography)**.



Podstawą rozdzielania w HPLC są wzajemne oddziaływania pomiędzy analitem a niepolarną fazą stacjonarną i polarnym rozpuszczalnikiem. Częstki hydrofobowe (niepolarne), dążąc do zmniejszenia powierzchni wystawionej na działanie polarnego rozpuszczalnika, silniej oddziałują z fazą stacjonarną, natomiast cząstki bardziej hydrofilowe (polarne) spędzać będą więcej czasu w fazie ruchomej. Oznacza to, że rozdzielanie mieszaniny następuje zgodnie z hydrofobowością jej składników: w pierwszej kolejności kolumnę opuszczają cząstki hydrofilowe, podczas gdy cząstki hydrofobowe będą charakteryzować się większymi czasami retencji.

Spektrometria mas: sprzęganie MS z HPLC (LC-MS)

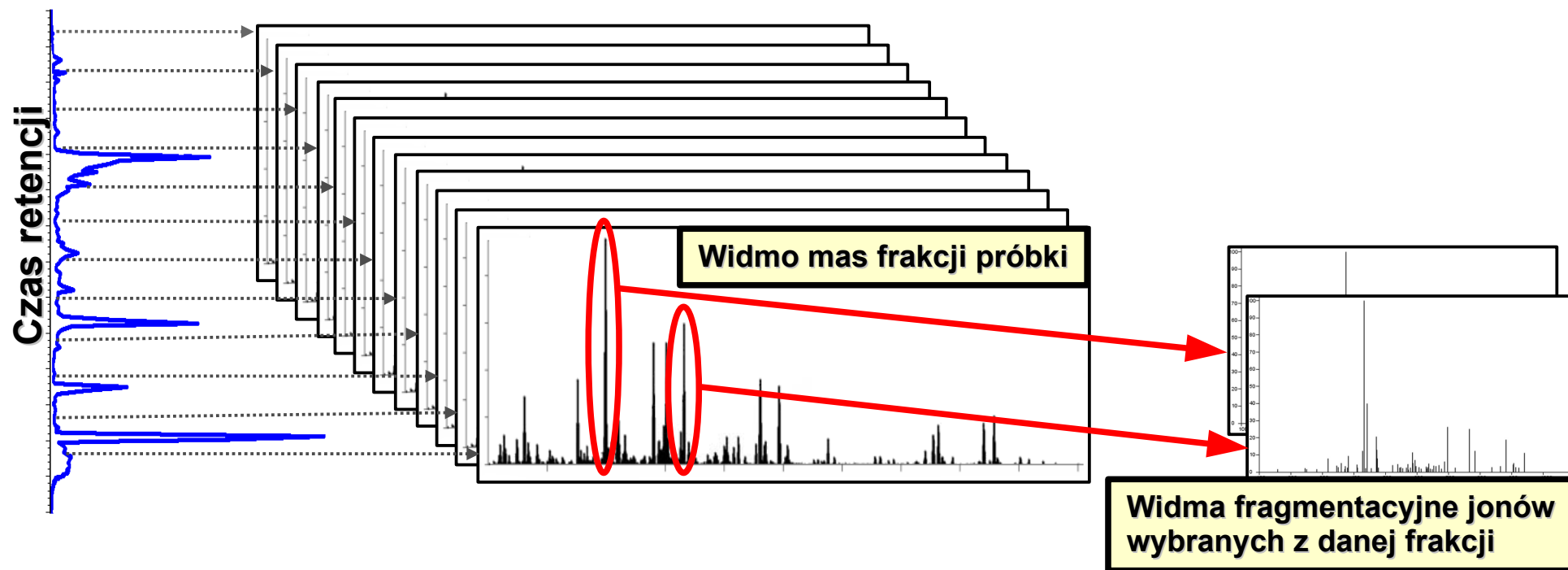
W czasie pomiaru techniką LC-MS badana próbka jest wstępnie rozdzielana pod względem hydrofobowości w systemie HPLC, a kolejne jej frakcje schodzące z kolumny chromatograficznej są poddawane jonizacji i wprowadzane do spektrometru. Pomiar widm mas następuje zwykle w regularnych odstępach czasu, zależnych od szybkości skanowania spektrometru.



Postać danych LC-MS: zbiór jednowymiarowych widm mas kolejnych frakcji próbki. Widmo w jednym skanie może zawierać nawet kilkaset tysięcy punktów, zaś liczba skanów może sięgać kilku tysięcy. W efekcie rozmiary plików z danymi LC-MS mierzone są w gigabajtach.

Spektrometria mas: sprzęganie MS/MS z HPLC (LC-MS/MS)

Możliwe jest również połączenie chromatografii cieczowej ze spektrometrią tandemową (**LC-MS/MS**), dzięki czemu rozdzielone chromatograficznie substancje mogą być identyfikowane na podstawie widm fragmentacyjnych. W tym przypadku z każdego zarejestrowanego widma mas wybierana jest pewna liczba jonów macierzystych (zwykle jest to kilka jonów o najwyższych pikach), które poddawane są fragmentacji, a ich widma MS/MS są mierzone w kolejnych skanach.



Postać danych LC-MS/MS: od kilku do kilkuset tysięcy widm MS/MS. Każdemu z widm fragmentacyjnych towarzyszy informacja o wartości m/z jonu macierzystego oraz numerze skanu (czyli pośrednio o czasie retencji), w którym nastąpiła fragmentacja.

Uczenie maszynowe w bioinformatyce

Wykład 13: identyfikacja białek przy użyciu MS

Tymon Rubel

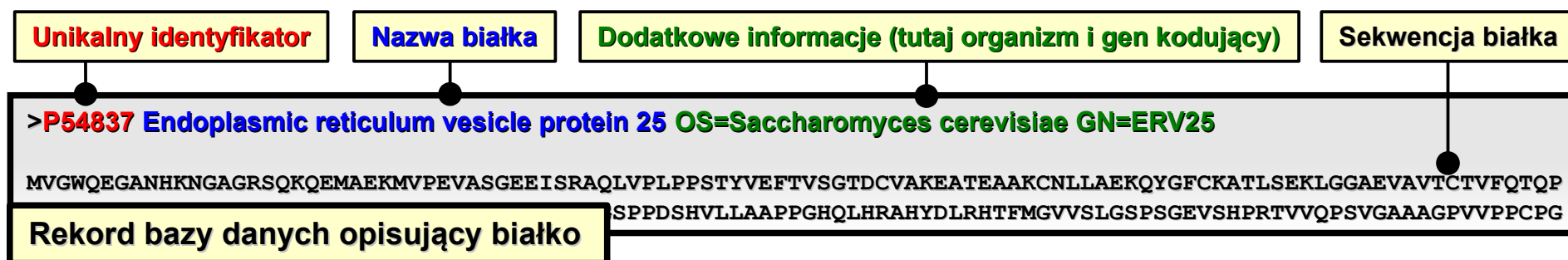
Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

MS w proteomice: identyfikacja białek

Podstawowym zadaniem spektrometrii mas w proteomice jest **identyfikacja białek** występujących w próbkach biologicznych. Należy jednak przy tym pamiętać, że:

- **Identyfikacja najczęściej rozumiana jest jako rozpoznawanie w próbce białek, o których mamy pewną wiedzę wstępną** – w szczególności znamy, przynajmniej częściowo, ich sekwencje.

Wyznaczanie sekwencji nieznanego białka jest bardzo rzadko wykonywane jedynie w oparciu o dane dostarczane przez spektrometr mas. Zdecydowanie częściej **identyfikacja wspomagana jest informacjami zgromadzonymi w bazach danych sekwencji białkowych**. W tym przypadku białko rozpatrywane jest jako liniowa sekwencja aminokwasów, opatrzona jednoznacznym identyfikatorem i zapisana jako rekord bazy danych.

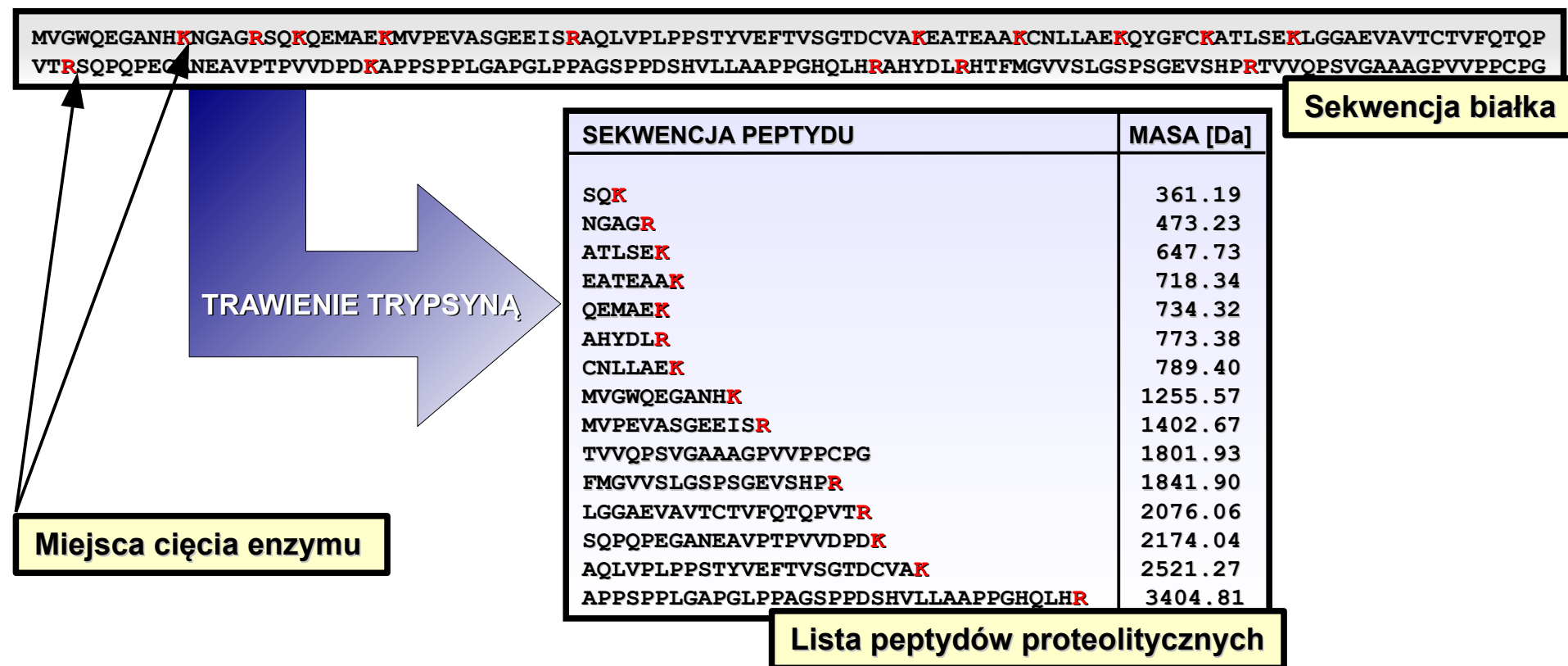


Przykładami często stosowanych baz danych są: **UniProt** (www.uniprot.org) oraz **NCBI** (www.ncbi.nlm.nih.gov/protein).

MS w proteomice: identyfikacja białek

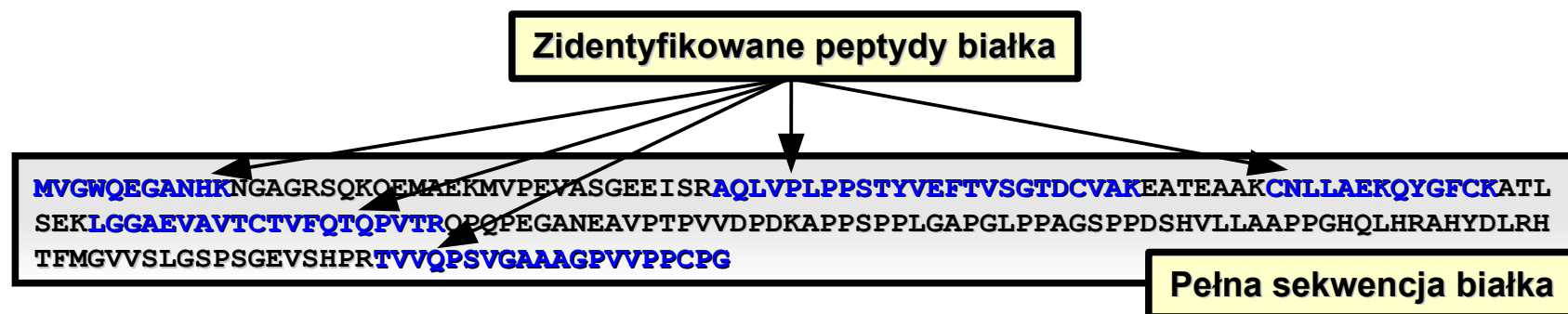
- Możliwości jednoznacznej identyfikacji tak dużych cząstek jak białka jedynie na podstawie pomiaru ich mas cząsteczkowych są mocno ograniczone. Dlatego też pomiarowi w spektrometrze poddawane są zwykle nie całe białka, a jedynie ich krótsze fragmenty (peptydy) powstałe w wyniku trawienia proteolitycznego.

Przed wprowadzeniem próbki do spektrometru znajdujące się w niej białka poddawane są działaniu **enzymu proteolitycznego**, który dzieli ich sekwencje w ściśle określonych miejscach. Przykładem takiego enzymu jest **trypsyna**, dokonująca podziału sekwencji po C-końcowych stronach arginin (R) i lizyn (K), z wyjątkiem miejsc gdzie następnym aminokwasem jest prolina (P).



MS w proteomice: identyfikacja białek

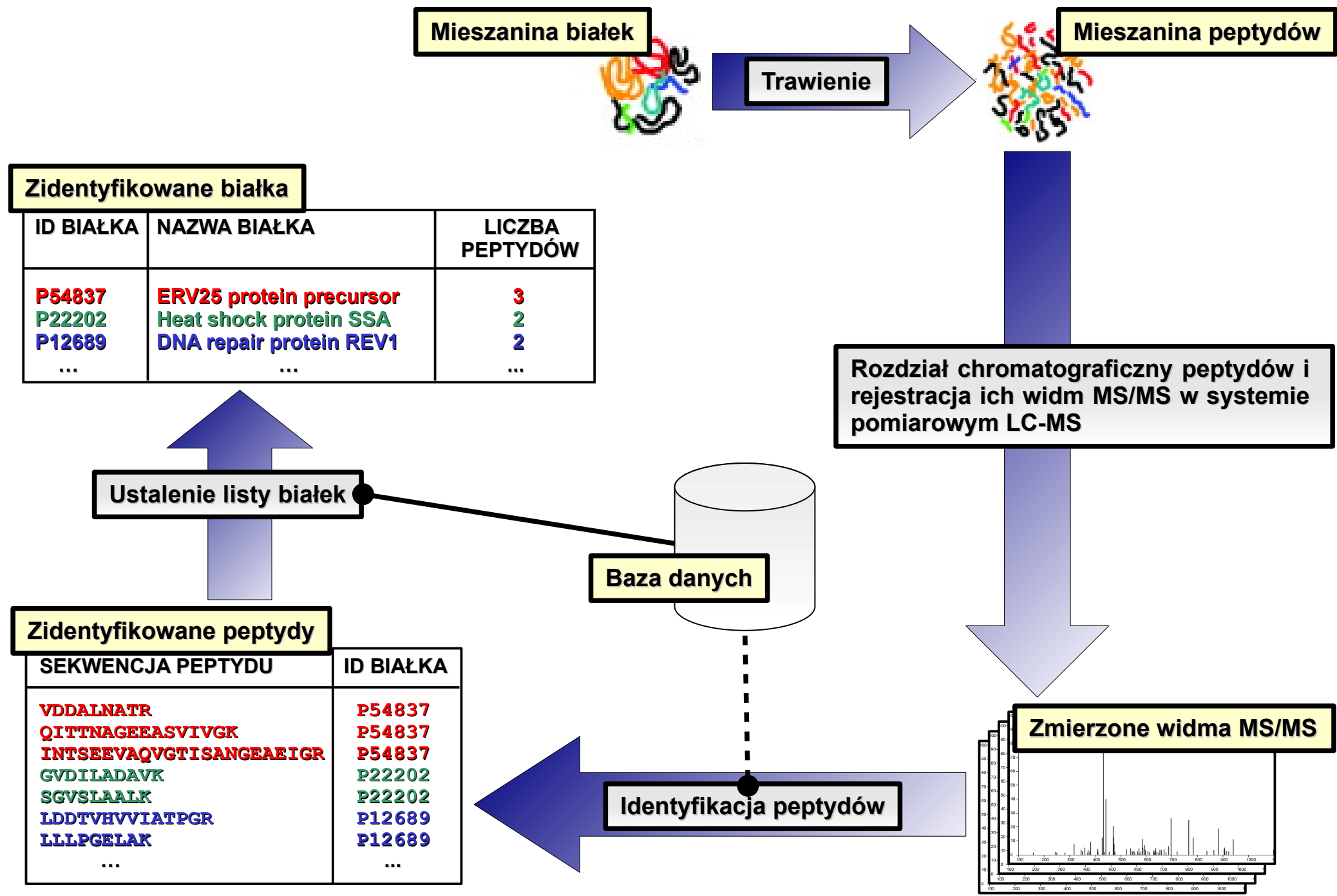
- **Identyfikacja białka odbywa się pośrednio, na podstawie detekcji pochodzących z niego peptydów.** Zakłada się przy tym, że znajomość sekwencji białka (jest ona zapisana w bazie danych) oraz specyficzności działania enzymu proteolitycznego pozwolą zidentyfikować białko nawet gdy okaże się, że możliwa była detekcja jedynie ograniczonej liczby jego peptydów.



Podstawową zaletą takiego podejścia do identyfikacji (określanego jako **bottom-up proteomics**) jest fakt, że peptydy mają relatywnie małe masy, co znacząco ułatwia jednoznaczne określenie ich sekwencji w tandemowym spektrometrze mas.

Innymi słowy: **bardzo trudne zadanie bezpośredniej identyfikacji białka rozbijane jest na szereg „podzadań” polegających na identyfikacji sekwencji jego peptydów.**

MS w proteomice: identyfikacja białek

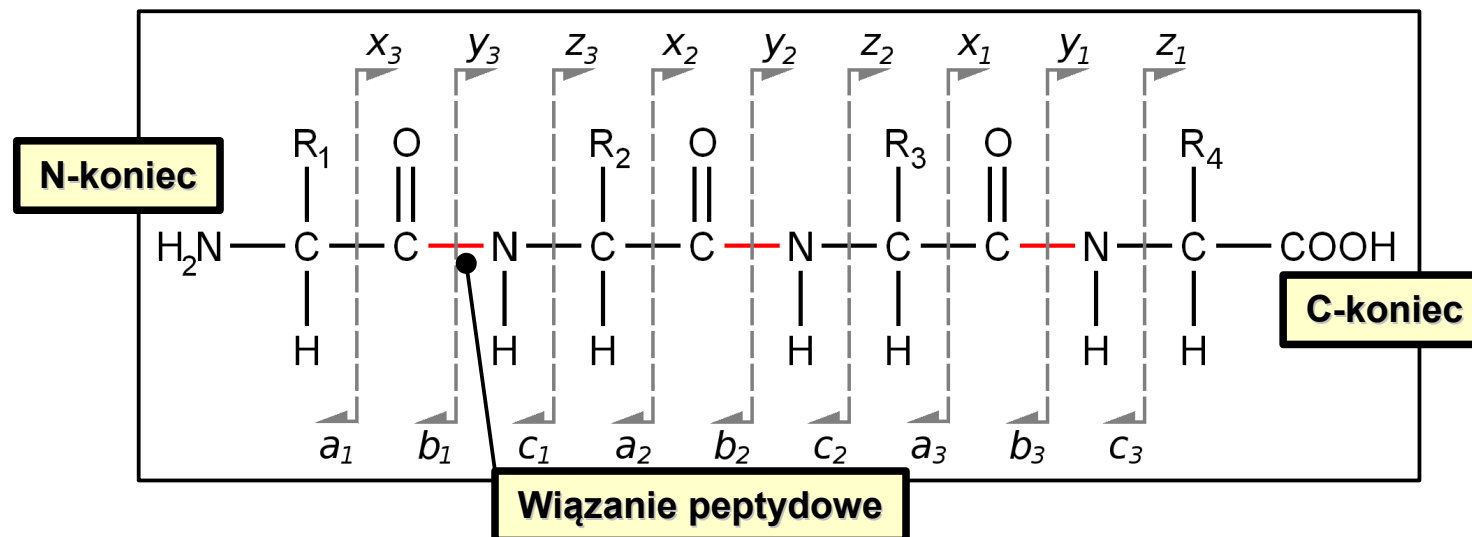


MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

Identyfikacja sekwencji peptydów odbywa się na podstawie widm fragmentacyjnych zmierzonych za pomocą tandemowej spektrometrii mas. Jest to możliwe dzięki znajomości zasad fragmentacji jonów peptydowych podczas zderzeń z cząsteczkami gazu obojętnego w komorze kolizyjnej spektrometru.

Przy typowo stosowanych energiach kolizji (do 100eV) dysocjacji ulegają najczęściej pojedyncze wiązania łańcucha głównego peptydu. Po rozerwaniu wiązania powstają dwa fragmenty, z których wykryty przez spektrometr może być ten, przy którym pozostanie ładunek.

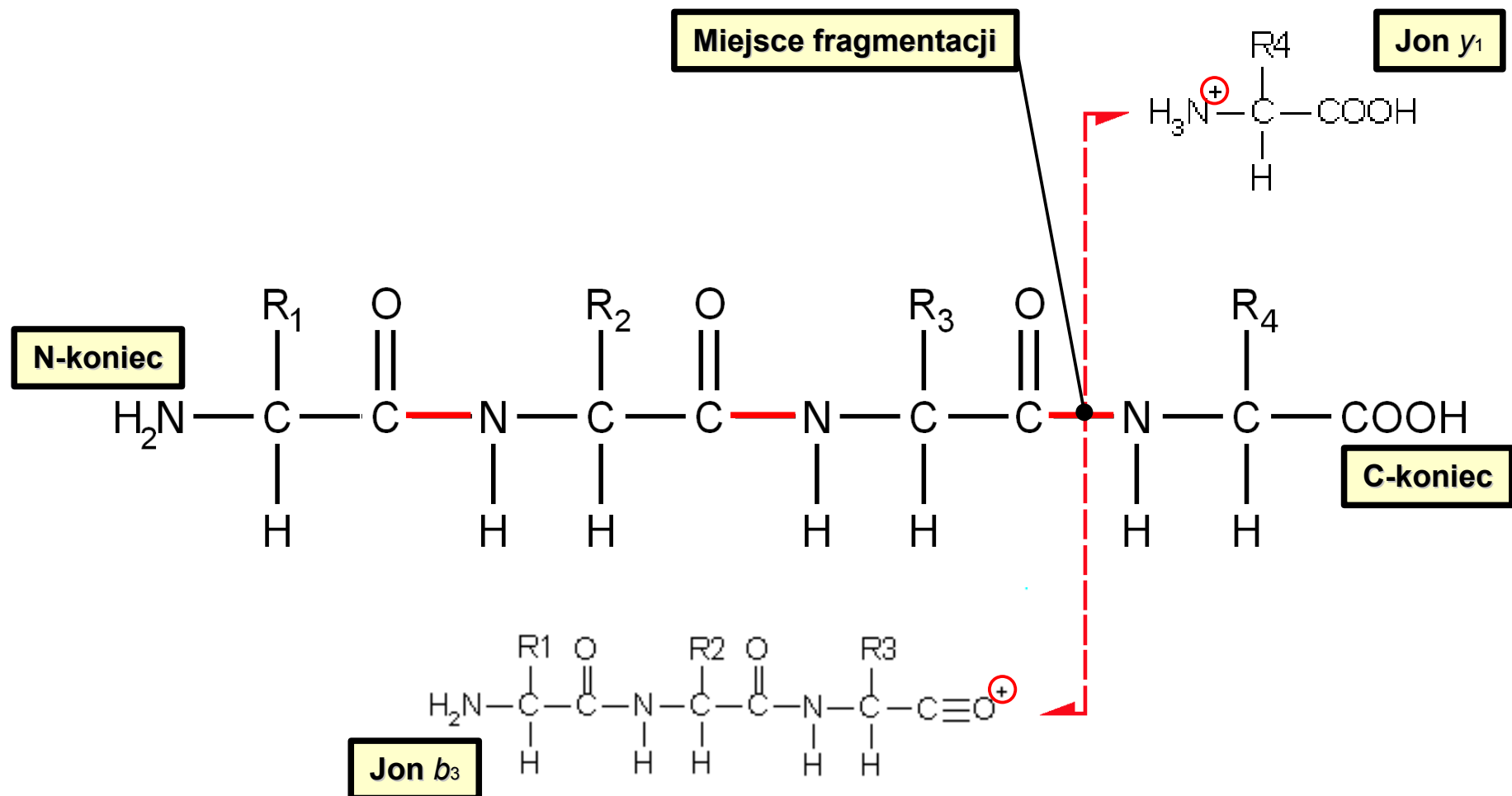
W zależności od miejsca dysocjacji, jony potomne nazywane są ***a*, *b*, *c*** jeżeli zawierają N-koniec peptydu (koniec łańcucha z wolną grupą aminową) oraz ***x*, *y*, *z*** jeżeli zawierają C-koniec (z grupą karboksylową). Indeksy przy nazwach jonów określają pozycję aminokwasu, którego wiązanie uległo przerwaniu. Indeksy jonów *a*, *b*, *c* liczone są od N-końca, podczas gdy jonów *x*, *y*, *z* od C-końca



MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

Prawdopodobieństwa powstania poszczególnych rodzajów jonów nie są jednakowe: dysocjacja najczęściej w miejscu wiązania peptydowego (pomiędzy atomami węgla i azotu), co prowadzi do rejestracji jonu typu b_i lub y_{L-i} .

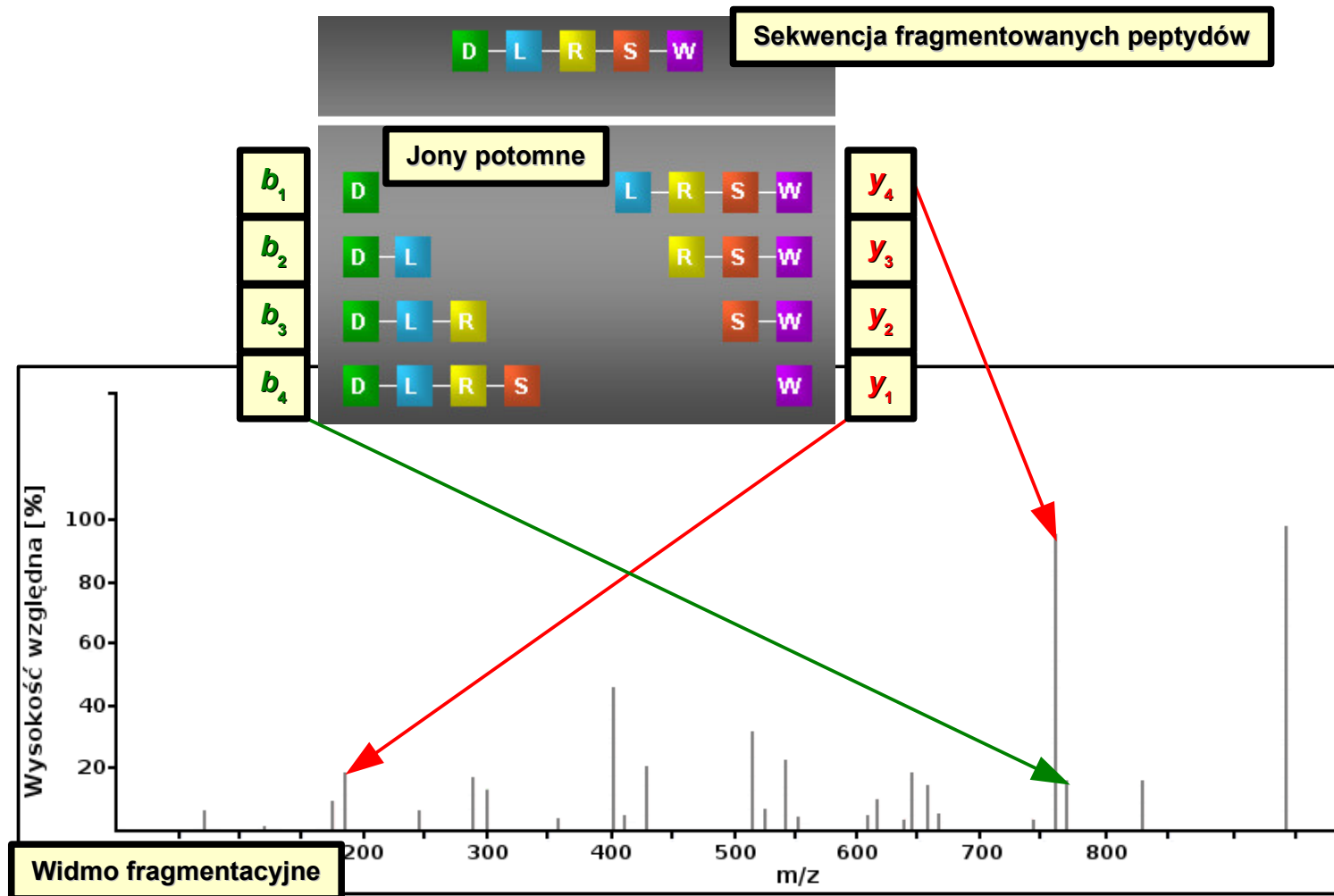
Przykładowo: jeżeli długość peptydu wynosi 4, to z dysocjacji trzeciego od N-końca wiązania peptydowego towarzyszyć będzie detekcja jonu b_3 lub y_1 , w zależności od tego, po której stronie miejsca fragmentacji pozostanie ładunek.



MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

W uproszczeniu można założyć, że pojedynczy jon peptydowy ulega fragmentacji w jednym miejscu. Gdy fragmentacji ulega jednocześnie cała populacja jednakowych peptydów, wówczas obserwowane będą całe rodziny jonów potomnych b i y , odpowiadających dysocjacji różnych wiązań peptydowych.

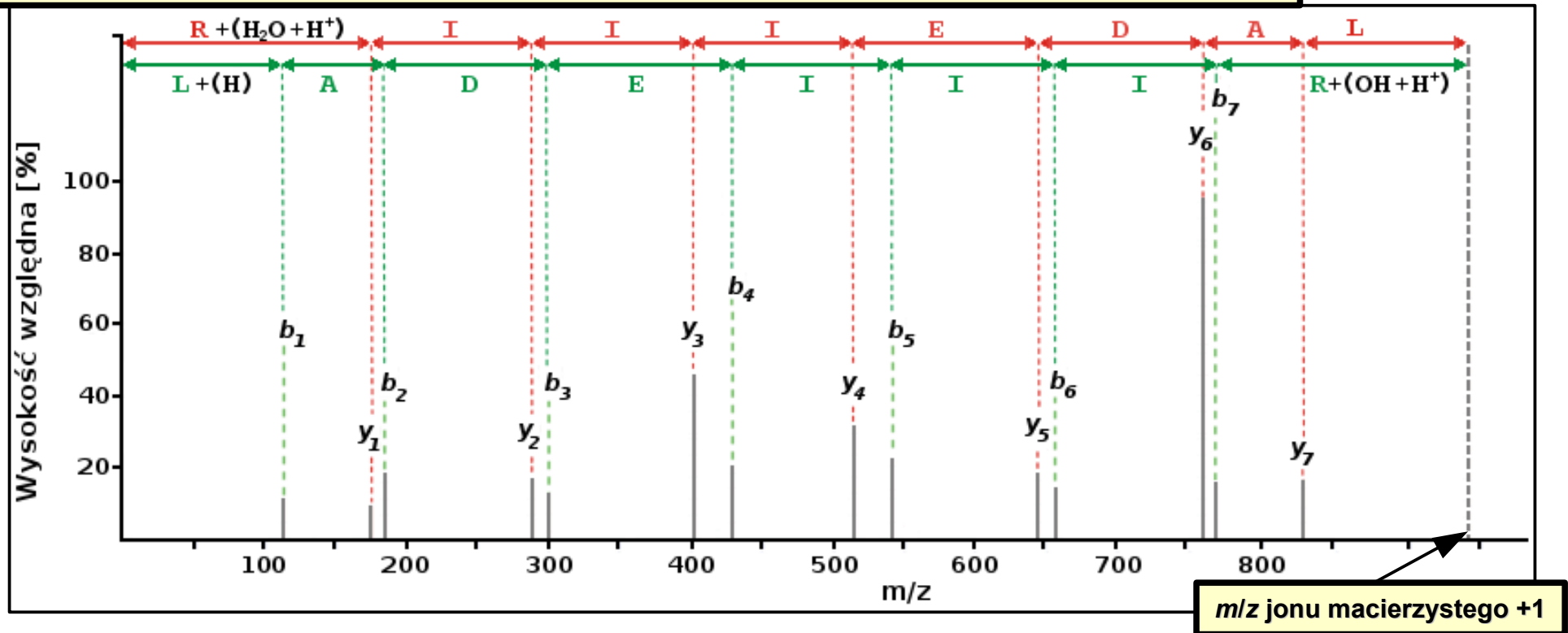
W widmie fragmenacyjnym jony potomne rejestrowane są jako piki o wartościach m/z wynikających z ich masy i stopnia naładowania.



MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

Jeżeli sekwencje dwóch jonów tego samego rodzaju i o jednakowej liczbie ładunków różnią się od siebie o jedną resztę aminokwasową (np. dwa kolejne jony b_i i b_{i+1}), to różnica wartości m/z odpowiadających im pików będzie równa masie tej reszty podzielonej przez stopień naładowania. W efekcie każdy rodzaj jonów generuje serię pików o wzajemnych położeniach zależnych od mas kolejnych aminokwasów w sekwencji peptydu, a tym samym niosących informacje o tej sekwencji.

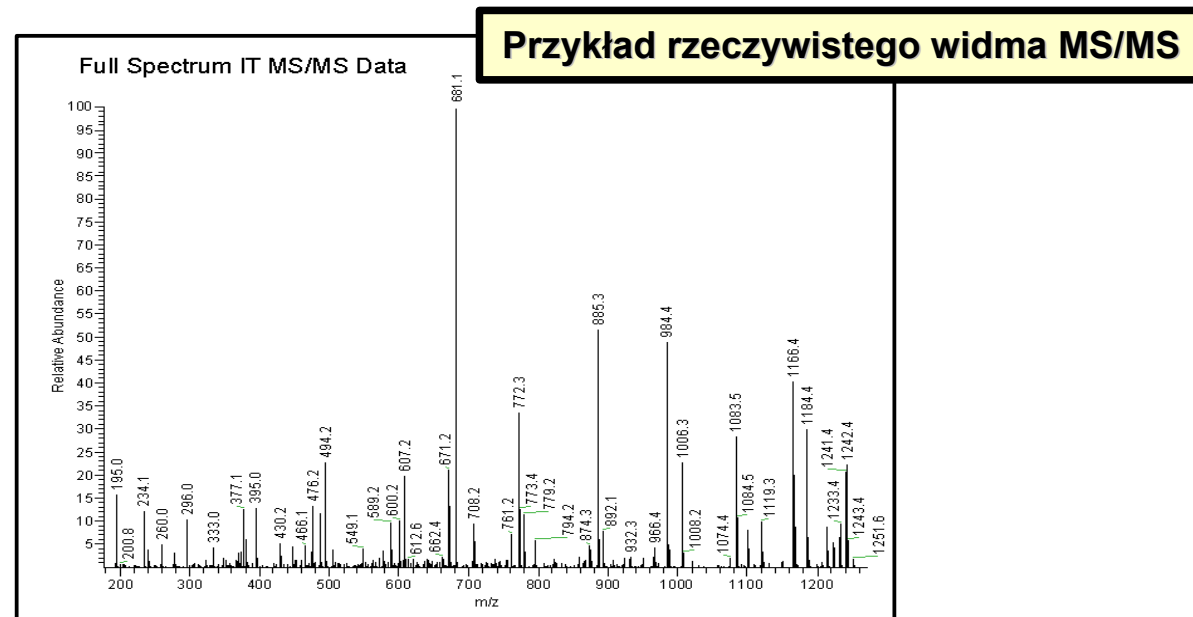
Fragmentacyjne widmo mas peptydu LADEIIIR (tylko piki y i b o stopniu naładowania +1)



MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

Określenie nieznanej sekwencji aminokwasów peptydu na podstawie widma MS/MS może odbywać się w różny sposób. Możliwe są m.in.:

- bezpośrednia („ręczna” lub automatyczna) interpretacja pików widma, nazywana **sekwencjonowaniem *de novo***. Problemem w tym przypadku jest duża złożoność widm MS/MS. Oprócz pików serii *b* i *y* mogą w nich występować również inne rodzaje pików, np. pochodzących od jonów pozostałych serii, jonów powstałych w wyniku wielokrotnej dysocjacji wiązań łańcucha głównego peptydu (jonów wewnętrznych) lub fragmentacji łańcuchów bocznych aminokwasów. Dodatkowym utrudnieniem może być również niski stosunek sygnału do szumu;

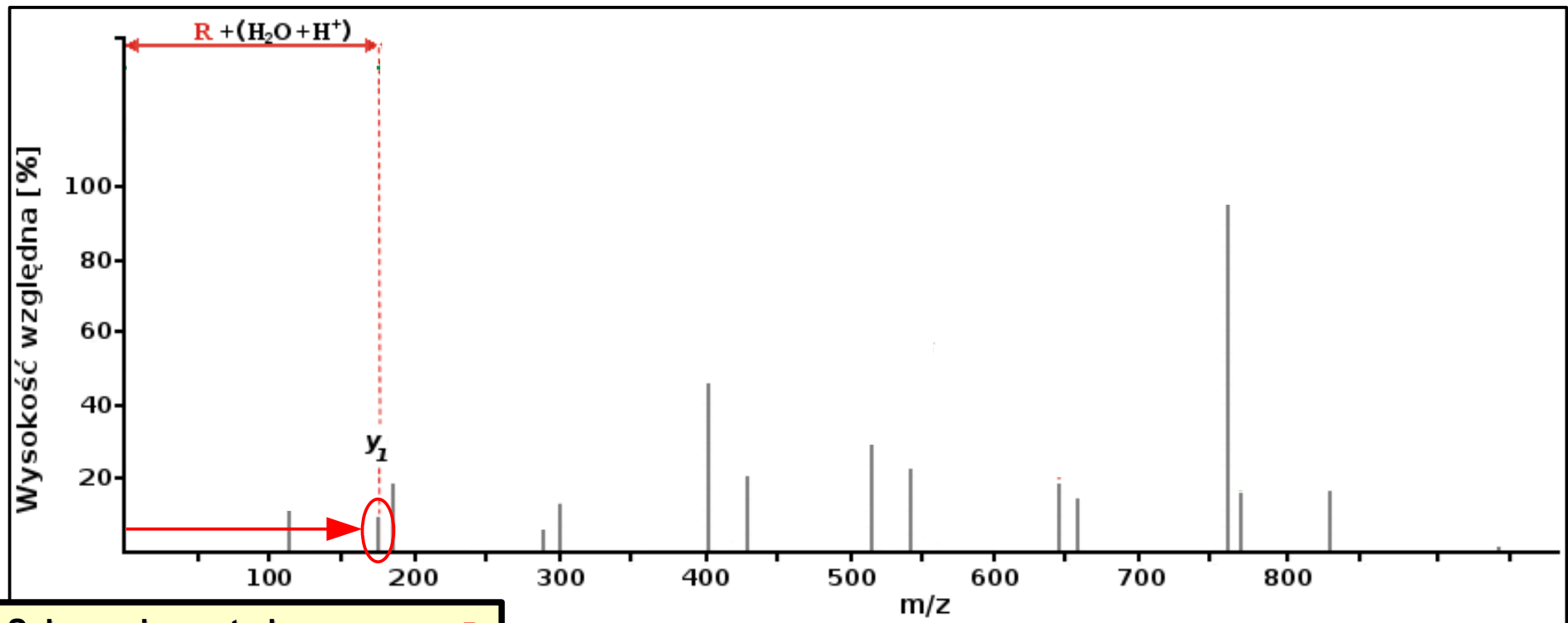


- **użycie algorytmu wykorzystującego bazę danych sekwencji znanych białek**, co jest rozwiązaniem mniej złożonym i obecnie zdecydowanie częściej stosowanym.

MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

„Ręczną” interpretację widma można zacząć od znalezienia pików jonu y_1 o stopniu naładowania +1, jako takiego, którego m/z jest równe masie jednej z 20 reszt aminokwasowych, powiększonej o masę wody i protonu (pozwala to określić C-końcowy aminokwas).

Następnie, idąc w kierunku rosnących wartości m/z , wybiera się pik odległy o masę jednej z reszt aminokwasowych od pików y_1 i zakłada się, że odpowiada on jonowi y_2 . Proces ten jest kontynuowany aż do osiągnięcia masy jonu macierzystego.

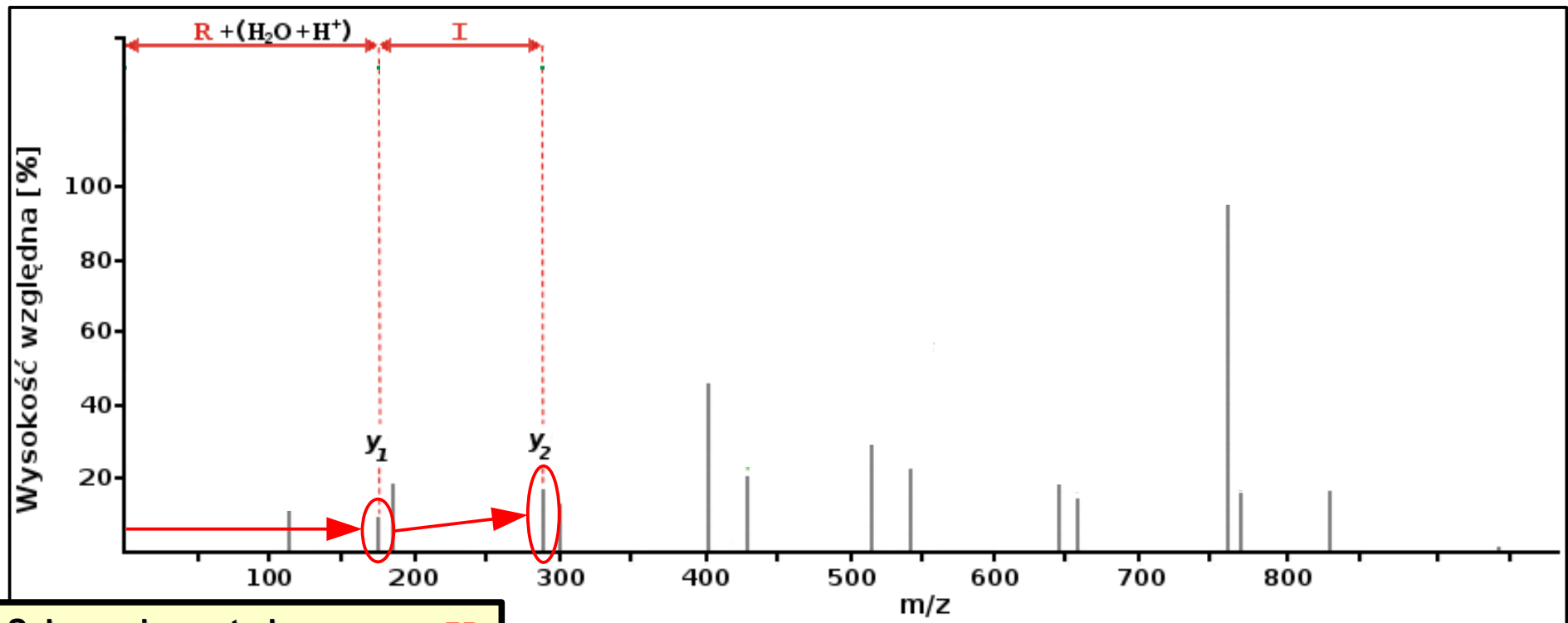


Sekwencja peptydu: **R**

MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

„Ręczną” interpretację widma można zacząć od znalezienia pików jonu y_1 o stopniu naładowania $+1$, jako takiego, którego m/z jest równe masie jednej z 20 reszt aminokwasowych, powiększonej o masę wody i protonu (pozwala to określić C-końcowy aminokwas).

Następnie, idąc w kierunku rosnących wartości m/z , wybiera się pik oddalony o masę jednej z reszt aminokwasowych od pików y_1 i zakłada się, że odpowiada on jonowi y_2 . Proces ten jest kontynuowany aż do osiągnięcia masy jonu macierzystego.

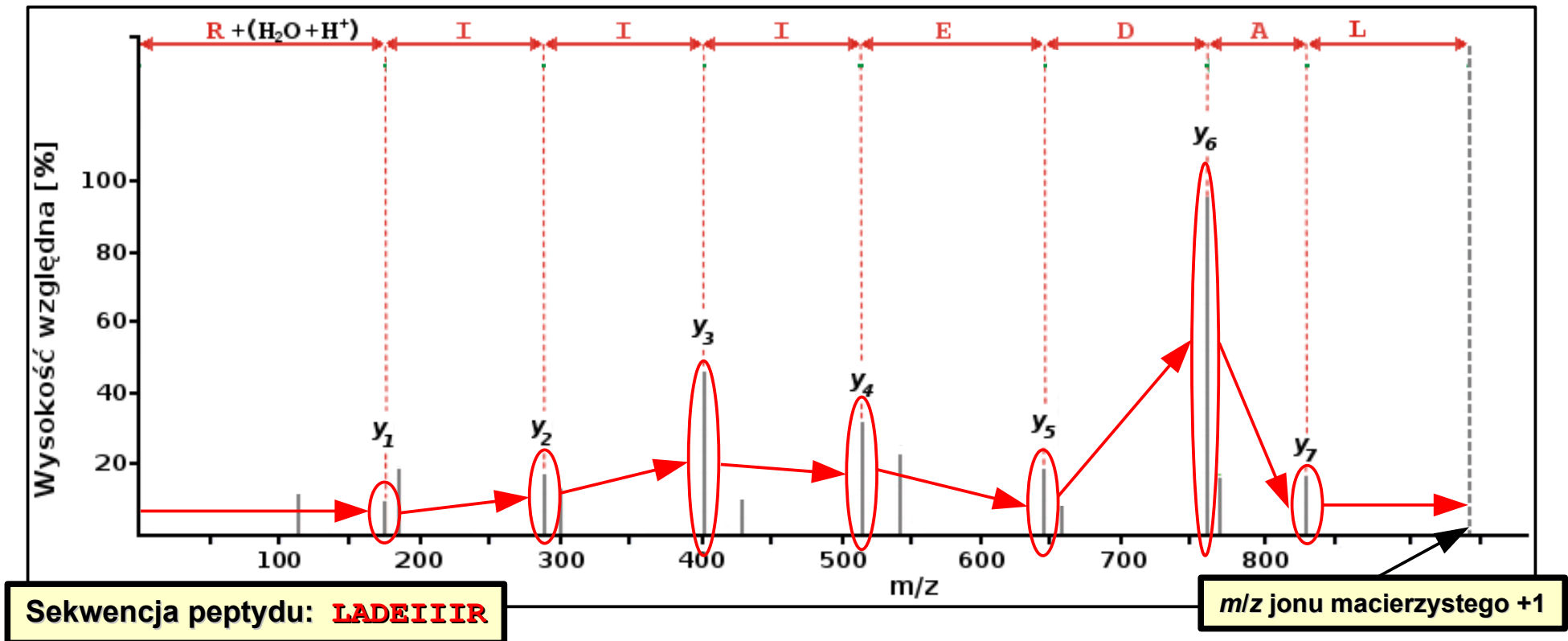


Sekwencja peptydu: **IR**

MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

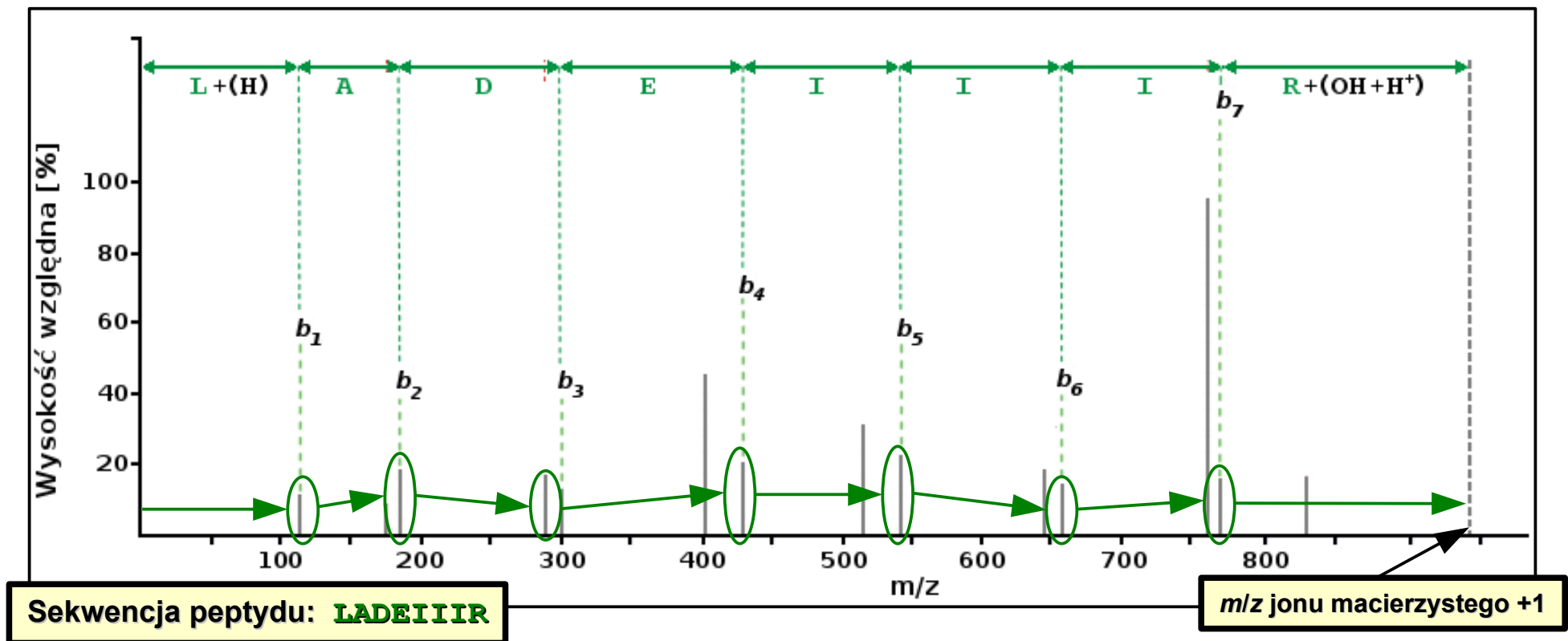
„Ręczną” interpretację widma można zacząć od znalezienia pików jonu y_1 o stopniu naładowania $+1$, jako takiego, którego m/z jest równe masie jednej z 20 reszt aminokwasowych, powiększonej o masę wody i protonu (pozwala to określić C-końcowy aminokwas).

Następnie, idąc w kierunku rosnących wartości m/z , wybiera się pik oddalony o masę jednej z reszt aminokwasowych od pików y_1 i zakłada się, że odpowiada on jonowi y_2 . Proces ten jest kontynuowany aż do osiągnięcia masy jonu macierzystego.



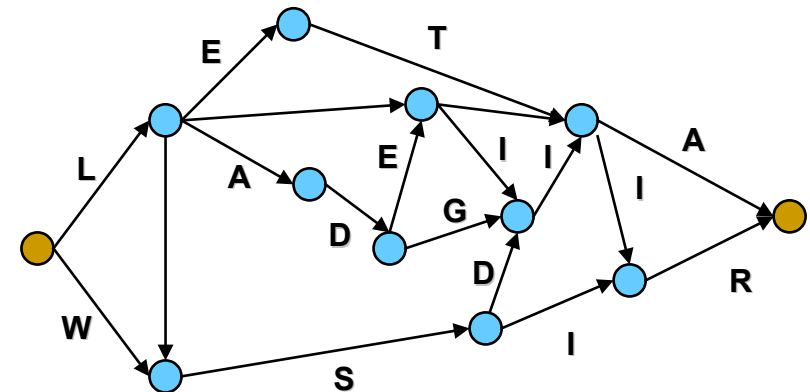
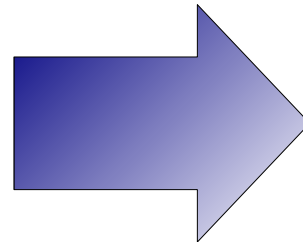
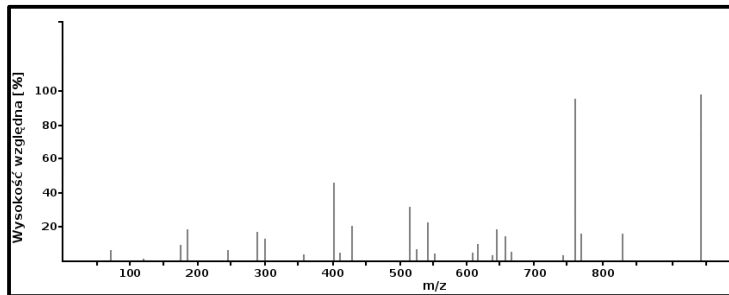
MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

W podobny sposób można wykorzystać piki odpowiadające pikom *b*. Otrzymana w ten sposób sekwencja powinna być identyczna jak ta określona na podstawie pików serii *y*, tyle że jej odczytywanie przebiega w odwrotnym kierunku (pierwszy zostanie zidentyfikowany aminokwas N-końcowy).

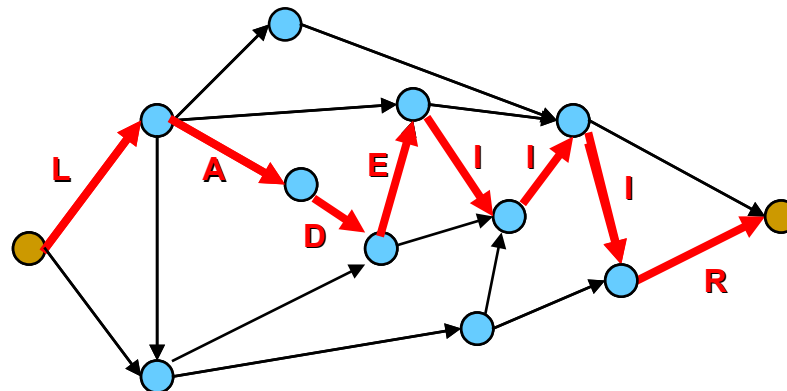


MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

Istnieją algorytmy **automatycznego sekwencjonowania *de novo***. Często bazują one na przekształceniu widma MS/MS w **graf** odzwierciedlający relacje pomiędzy występującymi w nim pikami. Wierzchołkami grafu są piki widma, a połączenia między nimi tworzone są wtedy, gdy różnica mas między pikami odpowiada masie jednej z reszt aminokwasowych.



Przy zastosowaniu takiej reprezentacji problem określenia sekwencji sprowadzony zostaje do znalezienia w grafie ścieżki o najniższym koszcie.

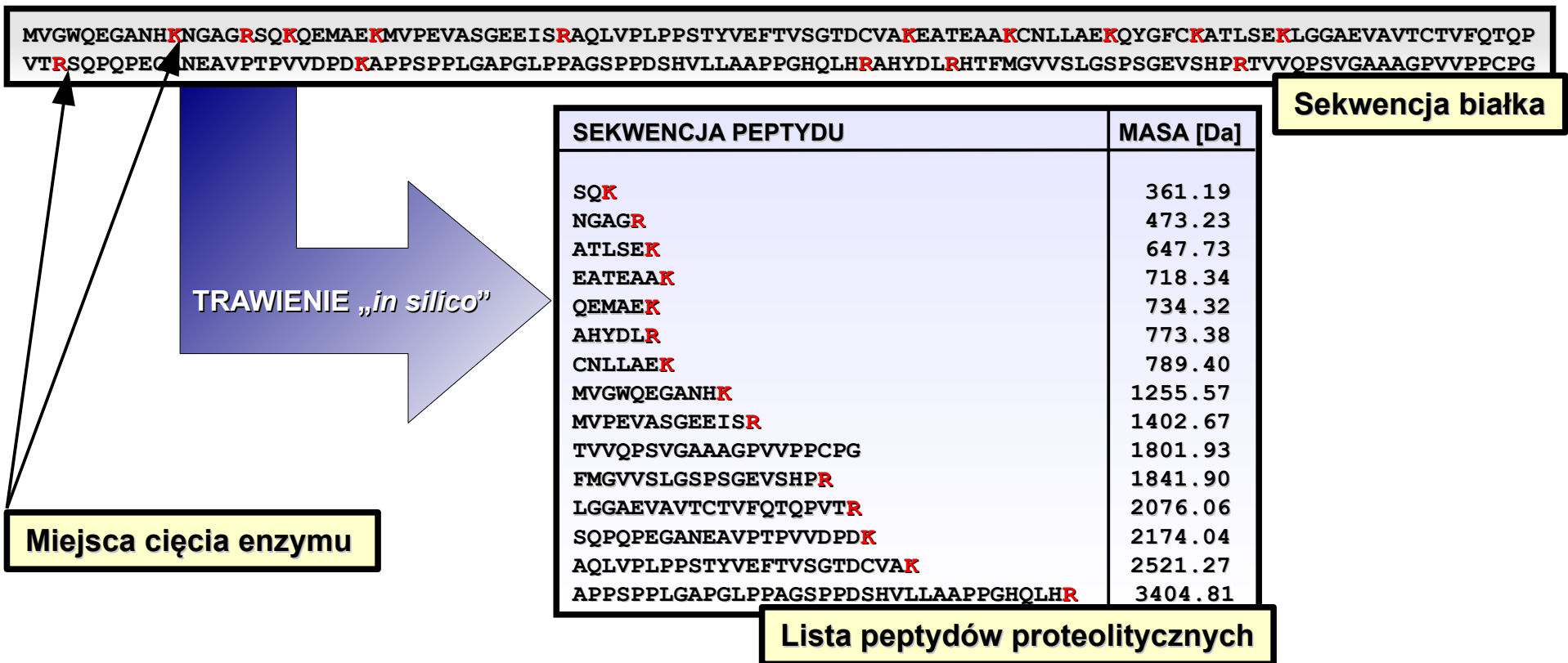


MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

Skuteczność algorytmów sekwencjonowania *de novo* jest silnie uzależniona od jakości danych pomiarowych. Algorytmy te są również kosztowne obliczeniowo i mogą prowadzić do wygenerowania sekwencji pozbawionych sensu biologicznego.

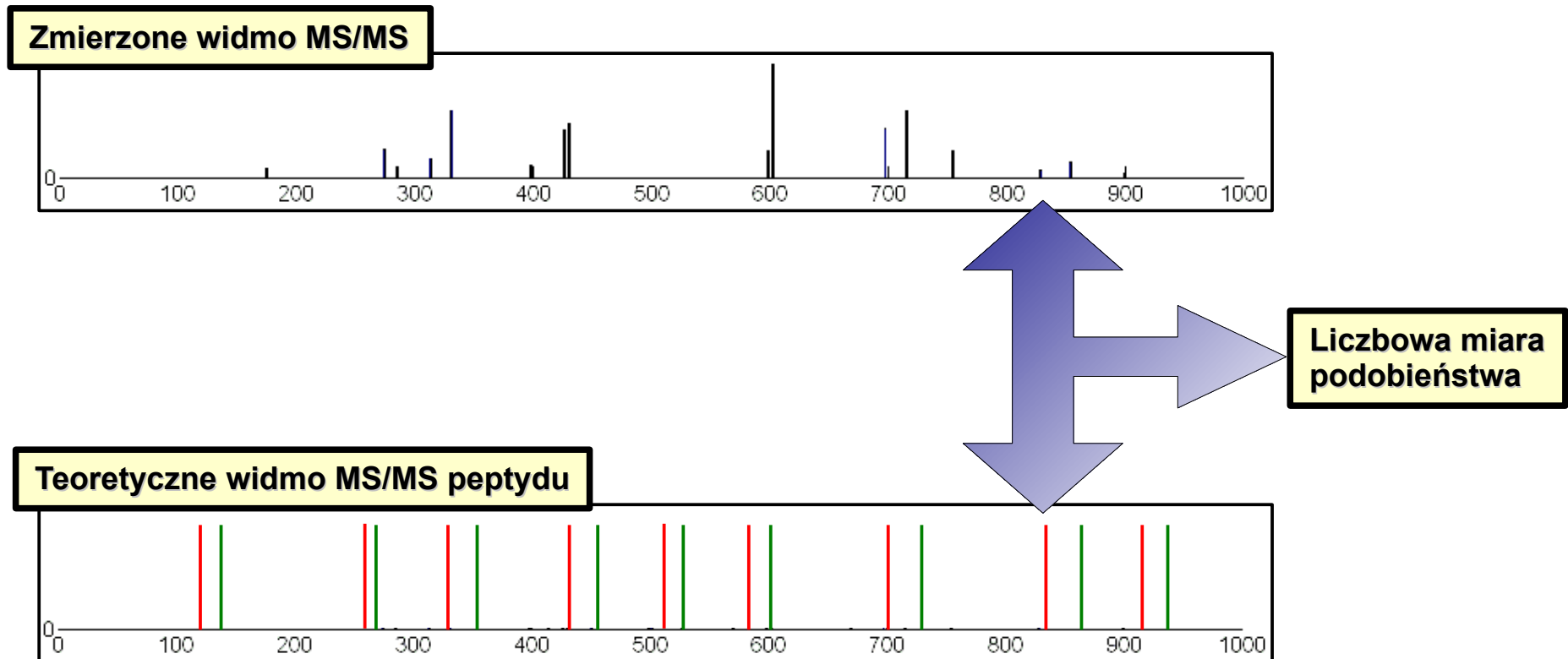
Dlatego też **podczas identyfikacji peptydów w badaniach proteomicznych zwykle używa się z baz danych sekwencji białek**. Korzysta się przy tym z dwóch założeń:

- znajomość miejsc cięcia enzymu proteolitycznego pozwala określić **teoretyczną listę peptydów powstających podczas trawienia** dowolnego białka o znanej sekwencji;



MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

- znajomość reguł fragmentacji pozwala wyznaczyć **teoretyczny model widma MS/MS peptydu** o znanej sekwencji. Tym samym istnieje możliwość wyznaczenia wartości **miary podobieństwa modelu z widmem pomiarowym**.

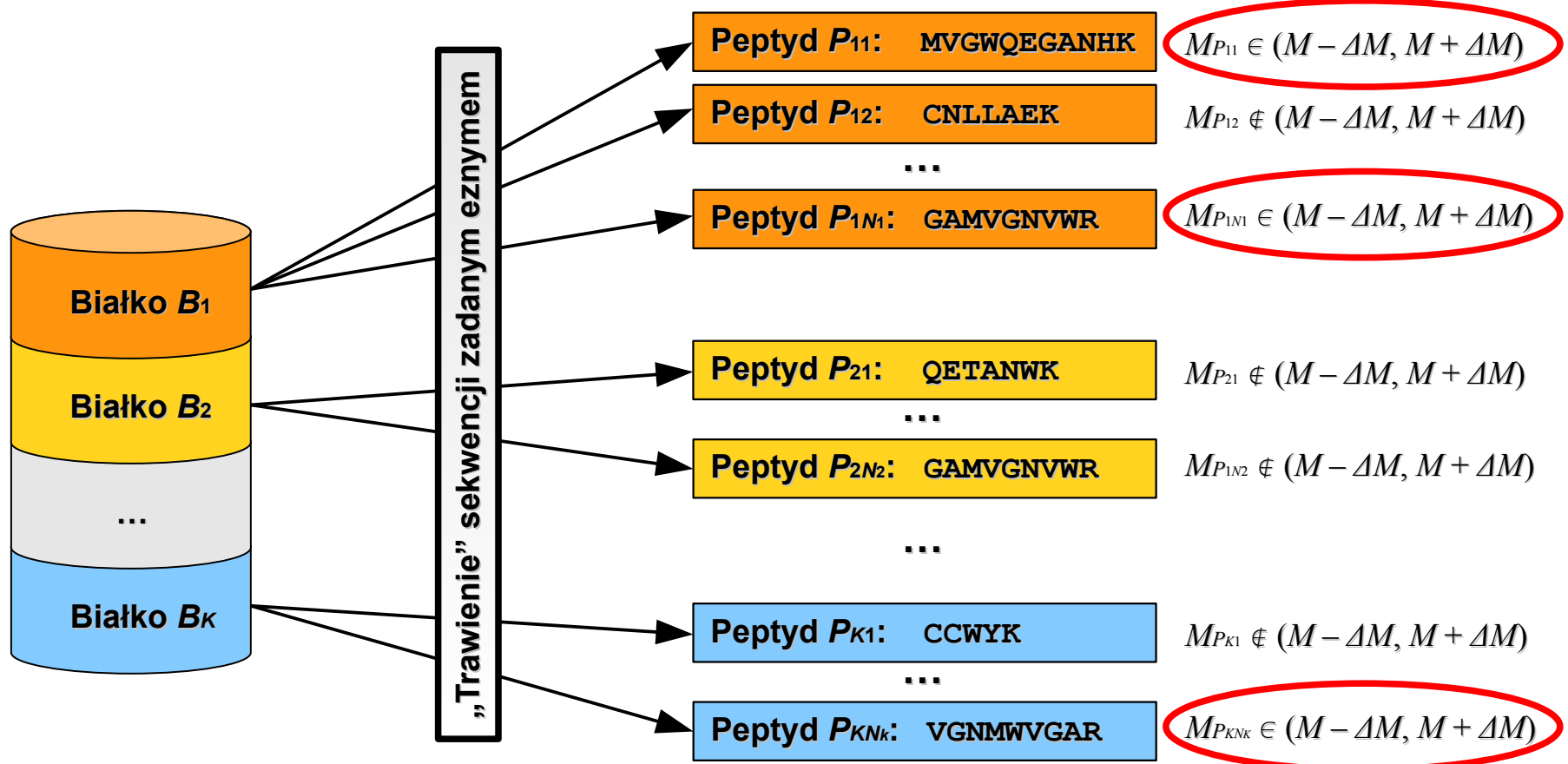


Użycie bazy danych pozwala zastąpić problem sekwencjonowania *de novo* znacznie prostszym zadaniem, polegającym na wyszukaniu spośród peptydów ze znanych białek takiego, którego teoretyczne widmo fragmentacyjne jest w jak największym stopniu zgodne (w sensie wartości wybranej liczbowej miary podobieństwa) z zarejestrowanym widmem eksperymentalnym.

MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

Ogólny schemat procesu przypisywania sekwencji peptydu do zmierzonego widma fragmentacyjnego (powtarzany dla każdego z widm w zbiorze danych pomiarowych):

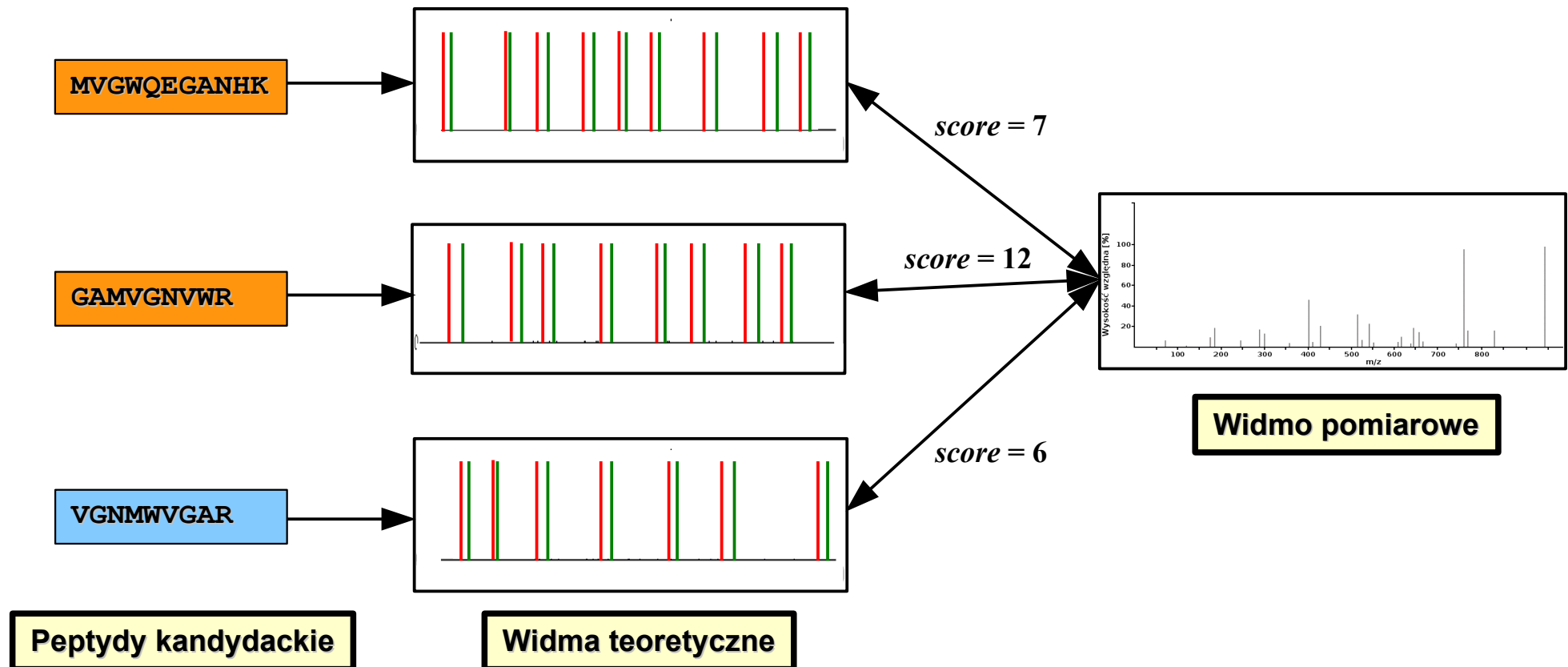
- Spośród peptydów, które mogą powstać w wyniku teoretycznego trawienia sekwencji wszystkich białek z bazy danych przez zadany enzym proteolityczny, wybierany jest zbiór **peptydów kandydackich**, o masach mieszczących się w tolerancji ΔM (jest ona parametrem zadany przez użytkownika) wokół zmierzonej masy jonu macierzystego M (jest ona znana z pomiaru MS).



MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

Ogólny schemat procesu przypisywania sekwencji peptydu do zmierzonego widma fragmentacyjnego (powtarzany dla każdego z widm w zbiorze danych pomiarowych):

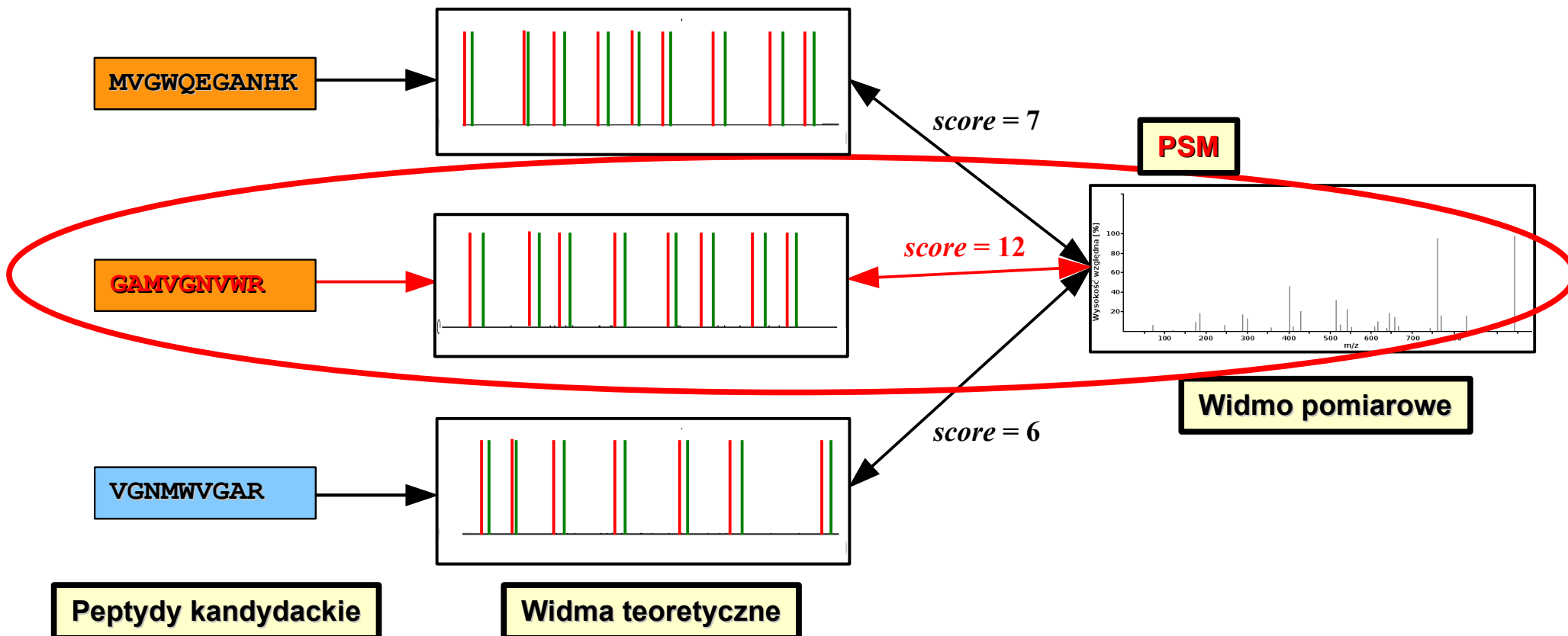
- Dla wszystkich peptydów kandydackich wyznaczane są **teoretyczne modele widm fragmentacyjnych**. Są one następnie porównywane z widmem eksperymentalnym, czemu towarzyszy wyznaczenie **miary dopasowania** (często nazywanej **score**).



MS w proteomice: identyfikacja białek (określanie sekwencji peptydów)

Ogólny schemat procesu przypisywania sekwencji peptydu do zmierzonego widma fragmentacyjnego (powtarzany dla każdego z widm w zbiorze danych pomiarowych):

- Widmu eksperymentalnemu przypisywana jest sekwencja peptydu, którego model widma charakteryzuje się największym dopasowaniem. Powstała w ten sposób para złożona z widma i sekwencji nazywana jest **Peptide-Spectrum Match (PSM)**.



MS w proteomice: identyfikacja białek

Po wyznaczeniu PSM dla każdego widma, lista białek występujących w próbce jest tworzona na podstawie listy peptydów uznanych za wiarygodnie zidentyfikowane (mających możliwie „pewne” sekwencje). W najprostszym przypadku uwzględnia się wszystkie peptydy o wartości score powyżej pewnego progu.

Lista zidentyfikowanych peptydów

#	PEPTYD	BIAŁKO	SCORE
1	ELLIIGGVAAR	Q92499	97.56
2	IALTDNALIAR	P18124	94.07
3	ALVDILSEVSK	Q9BQG0	90.88
4	SSSGLLEWESK	P14866	89.15
5	IDEPLEGSEDR	P61978	88.07
6	EQISDIDDAVR	P53999	87.73
7	ELLIQIFSTPR	Q8TDN6	81.21
8	IFGLLMGTLQK	Q9H307	81.11
9	GVTFLFPIQAK	Q9NR30	79.04
10	HIEIFTDLSSR	P13010	76.55
...	...	...	...
5039	DLEEFFSTVGK	Q14498	59.12
5040	AVFPSIVGRPR	A5A3E0; P60709; F	40.27
5041	TAAENDFVTLK	P35908	40.15
5042	FSPAGPVLSIR	Q13310	39.5
5043	EITALAPSTMK	P60709; P62736; P	38.91
5044	SEIEYYAMLAK	P62888	37.18
5045	SLQELFLAHIL	P09874	36.14
5046	SVAGGFVYTYK	Q15393	35.91
5047	AGGSAALSPSK	Q92522	34.12
5048	SEMEVQDAELK	Q9UQ80	33.32
5049	YPEETLSLMTK	P78527	32.34
5050	HILSPWGAEVK	P09874	31.75
5051	VPVHDVTDASK	P21333	24.07

Próg score

Lista zidentyfikowanych białek

#	ID	NAME	PEPTIDES #
1	P09874	Poly [ADP-ribose] polymerase 1	76
2	Q09666	Neuroblast differentiation-associated protein AHNAK	57
3	P16403	Histone H1.2	39
4	P09651	Heterogeneous nuclear ribonucleoprotein A1	35
5	P10412	Histone H1.4	35
6	Q9NR30	Nucleolar RNA helicase 2	32
7	P0C7M2	Putative heterogeneous nuclear ribonucleoprotein A1-like	31
8	P16402	Histone H1.3	30
9	P35579	Myosin-9	28
10	P22626	Heterogeneous nuclear ribonucleoproteins A2/B1	27
11	Q32P51	Heterogeneous nuclear ribonucleoprotein A1-like protein	26
12	P19338	Nucleolin	26
13	Q92499	ATP-dependent RNA helicase DDX1	25
14	P78527	DNA-dependent protein kinase catalytic subunit	25
15	P23246	Splicing factor, proline- and glutamine-rich	24
16	O00571	ATP-dependent RNA helicase DDX3X	24
17	O43390	Heterogeneous nuclear ribonucleoprotein R	22
...	...	...	...
190	P61978	Heterogeneous nuclear ribonucleoprotein K	2
191	P14866	Heterogeneous nuclear ribonucleoprotein L	1
192	Q92841	Probable ATP-dependent RNA helicase DDX17	1
193	P63261	Actin, cytoplasmic 2	1
194	P60709	Actin, cytoplasmic 1	1
195	Q08211	ATP-dependent RNA helicase A	1
196	Q15233	Non-POU domain-containing octamer-binding protein	1
197	P26599	Polypyrimidine tract-binding protein 1	1
198	Q9Y3I0	UPF0027 protein C22orf28	1
199	O15523	ATP-dependent RNA helicase DDX3Y	1

MS w proteomice: systemy bazodanowe

Przedstawiony powyżej schemat implementowany jest przez szereg **bazodanowych systemów identyfikacji**, zarówno komercyjnych (np. Mascot, Sequest, Peaks DB), jak i typu *open source* (np. X!Tandem, MS-GF+, OMSSA, Andromeda).

Mascot: www.matrixscience.com

X!Tandem: thegpm.org

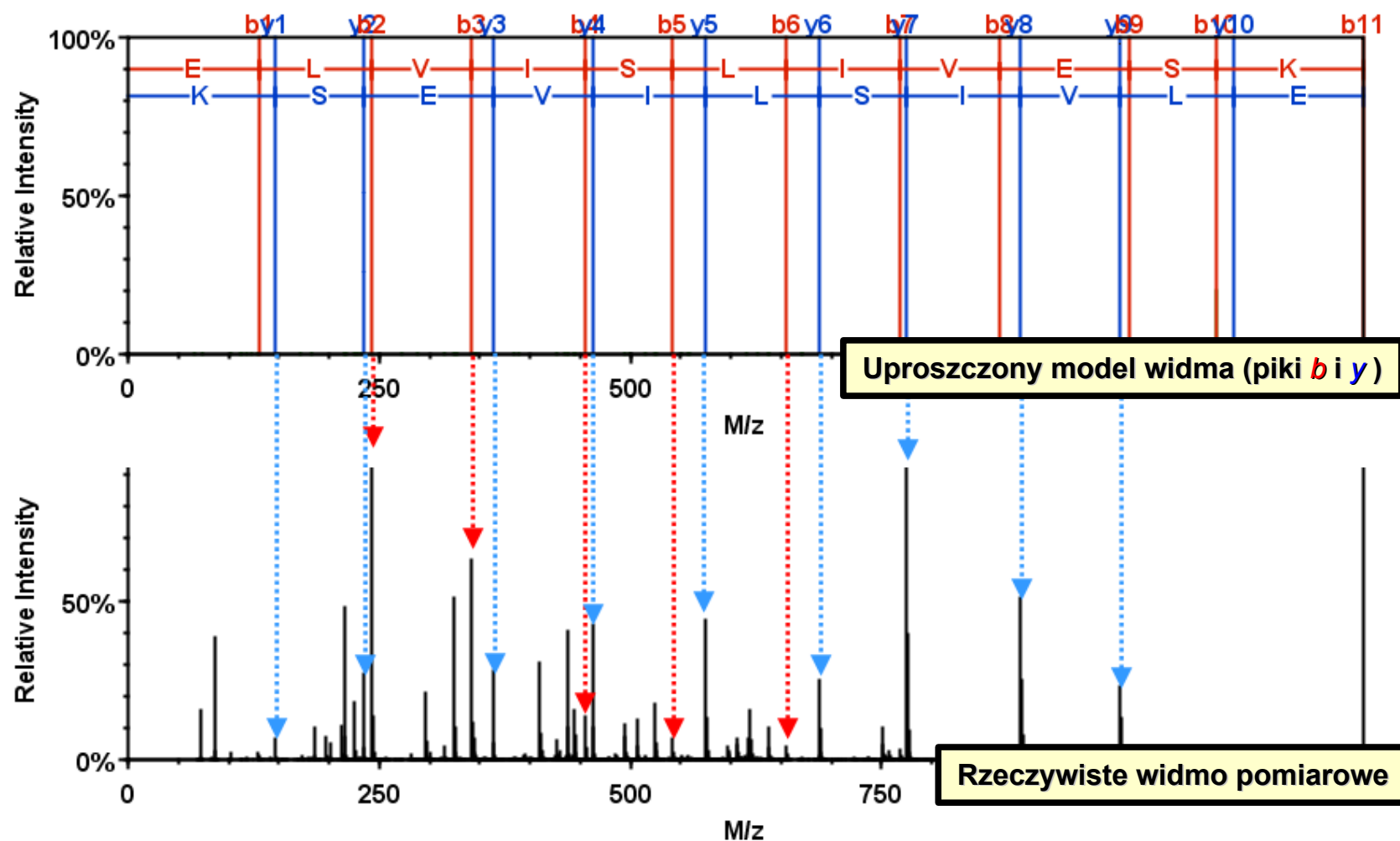
MS w proteomice: systemy bazodanowe

Podstawowa zasada działania wszystkich bazodanowych systemów identyfikacji jest podobna, jednak używane przez nie algorytmy mogą się znacząco różnić w zakresie:

- **sposobu wyznaczania teoretycznego modelu widma fragmentacyjnego;**
- **rodzaju miary dopasowania** zmierzonych widm MS/MS do modeli teoretycznych i interpretacji wartości score;
- **typu i znaczenia parametrów** decydujących o sposobie przeszukania bazy danych;
- **pre-processingu danych;**
- **określania istotności statystycznej przypisań sekwencji;**
- **formy prezentacji wyników i formacie ich zapisu.**

MS w proteomice: systemy bazodanowe (model widma)

Większość algorytmów stosuje **uproszczone modele widma**, uwzględniające jedynie wybrane serie jonów (np. jednokrotnie naładowane jony b i y). Wysokości pików zwykle ustalone są w sposób arbitralny (często są jednakowe dla wszystkich jonów).



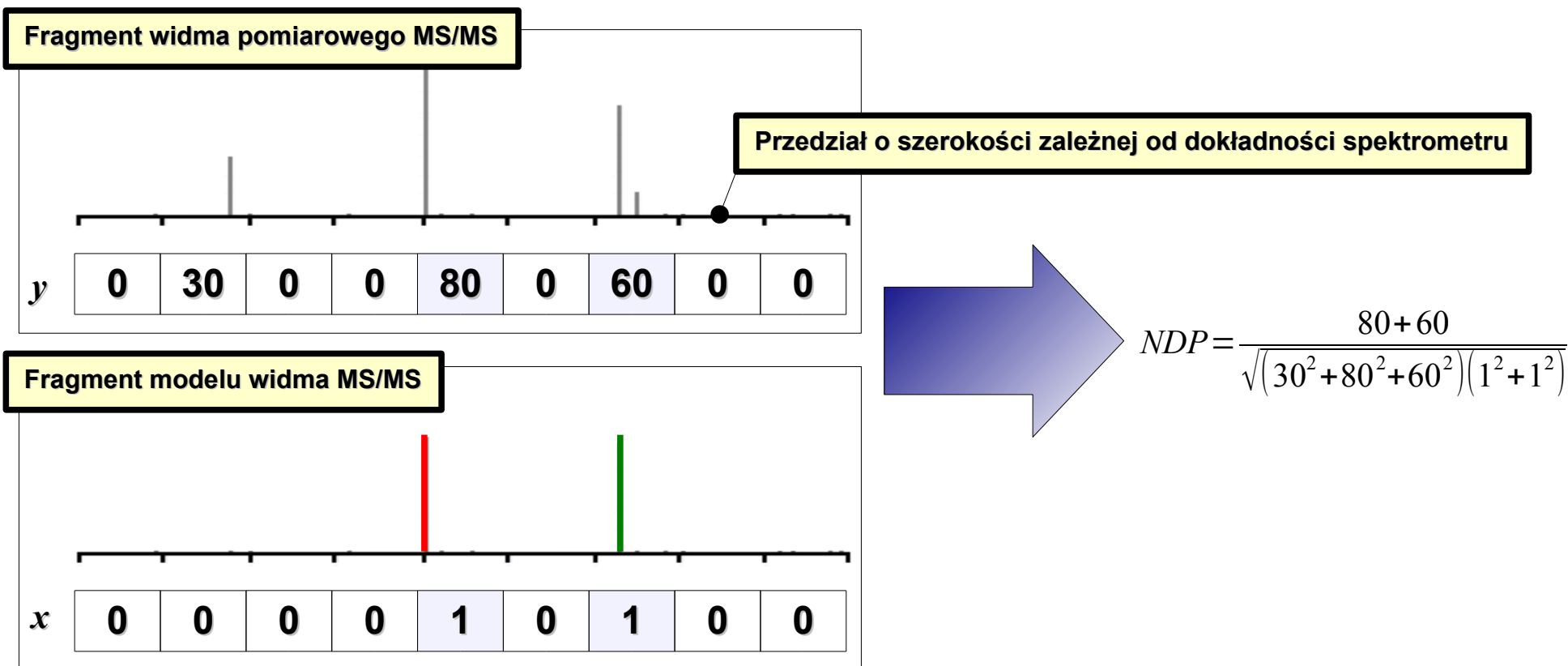
Zdecydowanie rzadziej spotykane są modele różnicujące wysokości pików w sposób zależny od empirycznie ustalonych prawdopodobieństw zajścia fragmentacji w różnych miejscach łańcucha głównego peptydu.

MS w proteomice: systemy bazodanowe (score)

Prostą miarą podobieństwa jest **znormalizowany iloczyn skalarny (normalized dot product)** wektorów reprezentujących widma teoretyczne (x) i eksperymentalne (y):

$$NDP = \frac{\sum_{i=1}^N x_i y_i}{\sqrt{\sum_{i=1}^N x_i^2 \sum_{i=1}^N y_i^2}}$$

Wektory x i y wyznaczone są po podzieleniu zakresu wartości m/z na N przedziałów o stałej szerokości (zależnej od dokładności pomiarowej stosowanego spektrometru mas) i wybraniu z każdego przedziału najwyższego pików.



Innym przykładem może być, stosowana w systemie SEQUEST miara dopasowania oparta na funkcji **korelacji wzajemnej (cross-correlation)**:

$$f(\tau) = \sum_i x_i y_{i+\tau}$$

gdzie τ jest przesunięciem w Da pomiędzy reprezentującymi widma wektorami x i y . Jeżeli widma są podobne, to $f(\tau)$ ma wyraźne maksimum dla $\tau=0$, w przeciwnym wypadku wartość $f(\tau=0)$ nie różni się znacząco od wartości dla innych przesunięć.

Ostateczna miara dopasowania widm (X_{corr}) obliczana jest jako różnica pomiędzy wartością funkcji $f(\tau=0)$ a średnią wartości $f(\tau)$ dla przesunięć $\tau \in [-75 \text{ Da}, 75 \text{ Da}]$.

