Algorytm oceny jakości i selekcji widm MS/MS

W czasie pojedynczego eksperymentu MS generowane są zbiory danych złożone z kilkudziesięciu, nawet kilkuset tysięcy widm MS/MS, z których jednak tylko część charakteryzuje się jakością wystarczającą do przeprowadzenia wiarygodnej identyfikacji sekwencji peptydu. Dlatego też istotnym elementem przetwarzania danych spektrometrycznych jest krok wstępnej selekcji widm MS/MS. Usunięcie z analizy widm o niewielkiej wartości poznawczej pozwala po pierwsze ograniczyć jej koszt obliczeniowy oraz po drugie zmniejsza ryzyko występowania nieprawidłowych dopasowań do sekwencji aminokwasowych.

Niska jakość zgromadzonych widm fragmentacyjnych może być następstwem wielu przyczyn. Po pierwsze, peptydy występujące w badanej próbce charakteryzują się różnym stężeniem i w przypadku tych zarejestrowanych w pobliżu progu detekcji spektrometru uwidacznia się wpływ fluktuacji statystycznych, co znacząco pogarsza stosunek sygnału do szumu. Oprócz tego jakość generowanych widm MS/MS jest silnie skorelowana z dokładnością oraz rozdzielczością oferowaną przez dane urządzenie spektrometryczne. Ponadto, fragmentacji ulegają również inne związki niebędące peptydami, będące zanieczyszczeniami z punktu widzenia pomiaru, nie podlegające specyficznym zasadom rozpadu takim jak peptydy i w związku z tym charakteryzującymi się widmami niskiej jakości.

W literaturze opisano szereg metod dokonujących wstępnej selekcji widm fragmentacyjnych [1-5]. Podstawowym założeniem podczas konstrukcji takiego klasyfikatora jest fakt istnienia cech widm MS/MS, umożliwiających ich rozróżnienie biorąc pod uwagę kryterium jakości. Dla przykładu, każde widmo fragmentacyjne charakteryzuje się pewną ustaloną liczbą zarejestrowanych jonów, przy czym im liczba ta jest mniejsza tym mniejsza jest szansa na przypisanie do widma jednoznacznej sekwencji aminokwasowej. Z drugiej strony, zbyt duża liczba jonów może świadczyć o silnych zakłóceniach lub innej nietypowej sytuacji w czasie pomiaru widma jak nakładanie się widm pochodzących od różnych peptydów o bardzo zbliżonych masach. Z powyższej obserwacji wynika, że poszukiwana zależność między liczbą jonów i klasą widm ma charakter złożony i nieliniowy. Liczne badania, prowadzone w ostatnich latach [1-5] umożliwiły określenie bogatego zbioru cech charakteryzujących widma MS/MS. Wśród nich oprócz podstawowych wielkości opisujących każde widmo takich jak wspomniana już liczba zmierzonych jonów, można wyróżnić cechy,

których definicja wymaga dodatkowych, czasem skomplikowanych obliczeń jak np. wartość średnia różnic mas kolejnych par jonów albo całkowita liczba zliczeń par jonów, których różnica mas odpowiada masie jednego z dwudziestu aminokwasów.

Typowe metody wykorzystujące do oceny widm MS/MS ich wyspecyfikowane cechy, opierają się na obliczeniu prostej, liniowej funkcji [1, 2], opisanej w następujący sposób:

$$S = \sum_{i=1}^{N} s_i \cdot c_i, \tag{1.1}$$

gdzie s_i jest *i-tą* cechą widma, c_i jest ustaloną wagą odpowiadającą *i-tej* cesze oraz N jest liczbą cech. W zależności od implementacji liczba używanych do wyznaczenia oceny cech waha się od kilku do kilkunastu. Pewnym rozwinięciem wyżej opisanej funkcji oceny jest zastosowanie modelu kwadratowego [4]. Nie zmienia to jednak sytuacji, że w obydwu przypadkach założona jest bardzo konkretna postać odwzorowywanej funkcji, w której jedynymi parametrami są współczynniki wagowe. Wartości tych współczynników wyznaczane są przy pomocy jednej z wielu metod, począwszy od liniowej funkcji dyskryminacyjnej, po metody oparte na uczeniu maszynowym takie jak maszyna wektorów podpierających (SVM) lub algorytmy genetyczne.

Pewnym wyjątkiem wśród tego typu metod jest podejście zaproponowane przez twórców programu Spequal [5], w którym to użyto ściśle określonej, nieliniowej funkcji oceny widma MS/MS, nie wykorzystującej dodatkowych parametrów:

$$S = CSD \cdot (TIC + STN). \tag{1.2}$$

Ostatecznie zastosowana postać funkcji oceny była efektem długich badań prowadzonych przez autorów przytoczonej pracy i opierała się na trzech cechach widma MS/MS określających poprawność wyznaczenia ładunku peptydu (*CSD*), całkowitą liczbę zliczeń w widmie (*TIC*) oraz stosunek sygnału do szumu (*STN*). Wydaje się, że największą wadą tego typu podejścia jest jego naturalnie ograniczony zasięg poznawczy, wynikający z braku kontroli człowieka nad zbyt rozbudowanym, ręcznie kalibrowanym modelem.

Odmienne podejście stosowane w procesie selekcji widm polega na bezpośredniej klasyfikacji wektora cech. Dla przykładu, w pracy [3] klasyfikacja widm odbywa się przy pomocy kwadratowej analizy dyskryminacyjnej (QDA – Quadratic Discriminant Analyse), na podstawie siedmiu wybranych cech widm MS/MS. Ogromną zaletą metody QDA jest możliwość łatwego i szybkiego obliczenia parametrów modelu, co w praktyce sprowadza się do wyznaczenia estymatorów macierzy kowariancji i wektorów wartości średnich. Niestety jest to okupione założeniem, że rozkłady cech można przybliżać wielowymiarowym rozkładem Gaussa, co w rzeczywistości może okazać się zbyt silnym uproszczeniem.

W mniejszości w stosunku do metod selekcji wykorzystujących wiele cech widm fragmentacyjnych pozostają metody poświęcone szczegółowej analizie jednej wybranej cechy. Przykład taki można odnaleźć w pracy [3], w której to autorzy metody opartej o analizę QDA skonstruowali drugi klasyfikator w oparciu o zależność pomiędzy klasą widma MS/MS oraz zliczeniami par jonów charakteryzujących się określoną różnicą masy. W pierwszym kroku wszystkie widma MS/MS przekształcono w określonej długości wektory, tak aby następnie przeprowadzić ich klasyfikację przy pomocy SVM. W wspomnianej pracy przeprowadzono symulacje dla dwóch typów histogramów umożliwiających pomiar różnicy mas odpowiednio w zakresie do 187 *Da* oraz 384 *Da*, przy czym mniejsza z wartości odpowiada największej masie aminokwasu. Otrzymane wyniki były bardzo zbliżone do tych jakie osiągnięto przy zastosowaniu metody QDA.

Możliwości oferowane przez wymienione algorytmy selekcji widm MS/MS są na bardzo zbliżonym poziomie. Dokładny sposób opisu wszystkich przytoczonych metod uwzględnia wybrany do klasyfikacji zbiór cech, regułę kierującą procesem selekcji widm, a w wybranych przypadkach konkretne wartości współczynników opisujących model [1, 4]. Należy podkreślić, że nawet w sytuacji ujawnienia przez autorów uzyskanych w trakcie optymalizacji wartości parametrów, zasadność ich stosowania ogranicza się wyłącznie do wykorzystanego przez nich zbioru danych. Wynika to z faktu, że ostateczna postać rejestrowanego widma uzależniona jest od wielu czynników takich jak wspomniana już dokładność i rozdzielczość spektrometru, ale także zależy od rodzaju zastosowanego urządzenia do fragmentacji oraz ustalonych warunków procesów rozpadu. W efekcie, w celu poprawnej selekcji widm MS/MS, konieczne jest niezależne określenie optymalnych współczynników modelu dla konkretnych typów układów spektrometrycznych. W tym świetle, o użyteczności metody selekcji widm MS/MS świadczy łatwość jej adaptacji do innego systemu pomiarowego. Pomimo, że większość spośród przedstawionych algorytmów umożliwia taką adaptację to niestety nie są one wolne od innych problemów.

Podstawowym ograniczeniem opisanych modeli jest przyjęcie przez ich autorów bardzo konkretnej postaci funkcji oceny widm fragmentacyjnych. Należy podkreślić, że związek pomiędzy cechami oraz jakością widm nie jest znany i tym samym nie może zostać wyrażony przy pomocy ściśle określonej formuły. W związku z powyższym w niniejszej pracy zaproponowano nową metodę oceny umożliwiającą selekcję widm MS/MS, wykorzystującą wielowarstwową sieć neuronową. Podejście to stanowi rozwinięcie modelu liniowego opisanego równaniem (1.1) charakteryzującym sieć neuronową, ale zbudowaną wyłącznie z jednego neuronu. W równaniu tym wartościom c_i odpowiadają wagi neuronu

natomiast wartościom s_i zmienne wejściowe sieci. Sieć neuronowa jest w stanie odwzorować z dowolną precyzją każdą ciągłą funkcję zmiennych wejściowych pod warunkiem zastosowania w warstwie ukrytej lub warstwach ukrytych neuronów o nieliniowej funkcji aktywacji [6]. W wyniku tego sieć neuronowa umożliwia konstrukcję dowolnej, nieliniowej funkcji, bez zakładania z góry jej konkretnej postaci tak jak to miało miejsce w zaprezentowanych wcześniej rozwiązaniach.

W proponowanej strukturze sieci wagi neuronów określane są przy użyciu metody wstecznej propagacji błędów, której zadaniem jest minimalizacja funkcji celu określonej jako wartość średnia kwadratu różnic pomiędzy wartościami oczekiwanymi na wyjściu sieci i wyjściem sieci. Technika ta jest jedną z podstawowych metod opartych o estymację gradientu funkcji celu, przy czym należy jednocześnie do metod najskuteczniejszych. Wykorzystywana w trakcie nauki informacja o jakości widma MS/MS uzyskana jest przy pomocy programu Mascot [7] firmy MatrixScience i opiera się na porównaniu najwyższej oceny sekwencji $score_{MAX}$ przypisanej do widma MS/MS, z ustalonymi progami istotności. Dla każdego widma MS/MS Mascot definiuje dwa progi istotności: Mascot Identity Threshold (*MIT*) oraz Mascot Homology Threshold (*MHT*). Dla potrzeb niniejszej pracy, informacja o jakości widma fragmentacyjnego została zapisana w postaci binarnej, przy czym do klasy widm niskiej jakości zaliczane są te, dla których $score_{MAX}$ jest mniejszy niż *min(MIT, MHT),* natomiast do klasy widm wysokiej jakości, na podstawie których przypuszczalnie możliwe jest odnalezienie właściwej sekwencji peptydu, zaliczane są pozostałe widma fragmentacyjne.

W celu oceny skuteczności sieci posłużono się dwoma wielkościami: czułością oznaczaną *Se* i specyficznością oznaczaną *Sp* [8]. Wielkości te zdefiniowane są w następujący sposób:

$$Se = \frac{N_{TP}}{N_{TP} + N_{FN}}, \tag{1.3}$$

$$Sp = \frac{N_{TN}}{N_{TN} + N_{FP}}, \tag{1.4}$$

gdzie N_{TP} jest liczbą poprawnie sklasyfikowanych widm wysokiej jakości, N_{FP} jest liczbą widm niskiej jakości błędnie rozpoznanych jako widma wysokiej jakości, N_{TN} jest liczbą poprawnie sklasyfikowanych widm niskiej jakości klasyfikacji oraz N_{FN} jest liczbą widm wysokiej jakości błędnie rozpoznanych jako widma niskiej jakości. Selekcja widm fragmentacyjnych polega na odrzuceniu jak największej liczby widm MS/MS niskiej jakości przy zachowaniu jak największej liczby widm wysokiej jakości. Typowo, w czasie eksperymentu wartość czułości (*Se*) ustalana jest na poziomie 0,95, co oznacza, że odrzucenie

z dalszego procesu widm o niskiej wartości poznawczej odbywa się kosztem utraty maksymalnie 5% widm wysokiej jakości. Dla tak ustalonej wartości *Se*, najlepszy klasyfikator będzie charakteryzował się najwyższą wartością specyficzności (*Sp*) lub inaczej najniższą wartością *1-Sp*. W dalszej części pracy analizie poddany będzie podlegający minimalizacji parametr *1-Sp*.

Sieć neuronowa użyta do badań ma strukturę statyczną, tzn. niezmienną w trakcie procesu uczenia. Wstępne badania pokazały, że najlepsze wyniki uzyskiwane są dla sieci neuronowej zbudowanej z jednej warstwy ukrytej i w związku z tym tylko sieć z jedną warstwą ukrytą rozważana jest w dalszej części pracy. Odmienne warianty sieci, różniące się liczbą neuronów w warstwie ukrytej były niezależnie analizowane i ostatecznie wyłoniono spośród nich strukturę charakteryzującą się najniższym poziomem błędu (*1-Sp*).

Bardzo ważną zaletą sieci neuronowej jest możliwość wykorzystania dowolnej liczby zmiennych wejściowych, z których każde jednoznacznie określa konkretną cechę widma fragmentacyjnego. W znanych autorowi doniesieniach literaturowych analizowane zestawy cech różniły się w nieznaczny sposób, co pozwoliło na wyłonienie najczęściej powtarzającego się i jednocześnie najsilniej różnicującego zbioru cech widm MS/MS. Powstały w ten sposób, przedstawiony poniżej, ostateczny zestaw cech został ograniczony do 19-tu elementów:

- całkowita liczba zliczeń par jonów, których różnica mas odpowiada masie jednego z dwudziestu aminokwasów;
- całkowita liczba zliczeń par jonów, których łączna masa odpowiada masie jonu macierzystego;
- całkowita liczba zliczeń par jonów, których różnica mas odpowiada masie wody, amoniaku lub tlenku węgla;
- wartość średnia różnic mas kolejnych par jonów;
- wariancja różnic mas kolejnych par jonów;
- stosunek liczby zmierzonych jonów do masy jonu macierzystego;
- maksymalny stosunek liczby zmierzonych jonów do masy jonu macierzystego mierzony w 50 rozłącznych, równomiernych przedziałach *m/z*;
- średnia długość ścieżki utworzonej z jonów oddalonych od siebie o masę aminokwasu każdy;
- całkowita liczba zarejestrowanych jonów;
- całkowita liczba zliczeń w widmie;
- masa jonu macierzystego;

- najmniejszy zakres *m/z* zawierający 95% wszystkich zliczeń w widmie;
- najmniejszy zakres *m/z* zawierający 50% wszystkich zliczeń w widmie;
- frakcja jonów o intensywności >= 1% największego jonu;
- frakcja jonów o intensywności >= 20% największego jonu;
- różnica intensywności pomiędzy dwoma najintensywniejszymi jonami;
- wartość średnia intensywności jonów;
- wariancja intensywności jonów;
- całkowita intensywność jonów tworzących grupy ze względu na występowanie różnych odmian izotopowych.

1.1. Literatura

1. Moore R. E., Young M. K., Lee T. D.: Method for Screening Peptide Fragment Ion Mass Spectra Prior to Database Searching. J. Am. Soc. Mass Spectrom. 2000, 11, 422–426.
2. Nesvizhskii A. I., Roos F. F., Grossmann J., Vogelzang M., Eddes J. S., Gruissem W., Baginsky S., Aebersold R.: Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data. Molecular & Cellular Proteomics, 2006, 5, 652–670.
3. Bern M., Goldberg D., McDonald W. H., Yates J. R.: Automatic Quality Assessment of Peptide Tandem Mass Spectra. Bioinformatics, 2004, 20, Suppl. 1, I49 –I54.
4. Xu M., Geer L. Y., Bryant S. H., Roth J. S., Kowalak J. A., Maynard D. M., Markey S. P.: Assessing Data Quality of Peptide Mass Spectra Obtained by Quadrupole Ion Trap Mass Spectrometry. J. Proteome Res. 2005, 4, 300–305.
5. Purvine S., Kolker N., Kolker E.: Spectral Quality Assessment for High-Throughput Tandem Mass Spectrometry Proteomics. OMICS, 2004, 8, 255–265.
6. Poggio T., Girosi F.: Network for approximation and learning. Proceedings of the IEEE, 1990, 78, 1481–1497.
7. Perkins D. N. et al.: Probability-based protein identification by searching sequence database using mass spectrometry data. Electrophoresis, 1999, 20, 3551–3567.
8. Green D. M., Swets J. M.: Signal detection theory and psychophysics. New York: John Wiley and Sons Inc. 1966.