Streszczenie rozprawy *Metody statystyczne dla danych proteomicznych ze strukturą wielokrotnej przynależności pozyskanych przy pomocy spektrometrii mas*

mgr Mateusz Staniak

Ta rozprawa rozwiązuje praktyczne problemy w statystycznej analizie danych proteomicznych ze spektrometrii mas, które wymagają zarówno rozwoju i oceny nowych statystycznych modeli, jak i projektu i implementacji programów do ich wykorzystania. Praca skupia się na problemach z kategorii danych ze strukturą wielokrotnej przynależności. W przeciwieństwie do standardowej analizy wariancji, gdzie każda obserwacja należy do dokładnie jednej grupy, wielokrotna przynależność pojawia się tam, gdzie obserwacje są związane jednocześnie z kilkoma grupami, wpływając na estymację wielu efektów równocześnie. Rozważamy szczegółowo dwa przykłady tej struktury.

W estymacji ilości białek w próbce, problem ten pojawia się przy próbie włączenia do analizy współdzielonych peptydów, to znaczy peptydów, które pochodzą z więcej niż jednego białka. Typowe podejścia często usuwają takie peptydy z analizy, zmniejszając moc statystyczną i wprowadzając obciążenie. Modelowanie przynależności takich peptydów do odpowiadających im białek generuje problem wielokrotnej przynależności: zaobserwowana intensywność peptydu wpływa na estymację ilości każdego odpowiadającego mu białka. Proponujemy model statystyczny, który rozszerza istniejące podejście do estymacji białek o nazwie MSstats poprzez wprowadzenie wag opisujących zgodność zaobserwowanych wzorców ilościowych na poziomie białek i nieznanych ilości białek, które stanowią główny obiekt zainteresowania analiz. Model dopasowujemy metodami optymalizacji dwuwypukłej. Jakość metody oceniamy na podstawie symulacji i analiz danych rzeczywistych, skupiając się na dwóch aspektach: precyzji estymacji względnej ilości białek pomiędzy stanami biologicznymi (np. chorzy i zdrowi pacjenci), i statystycznych własnościach powiązanych testów.

W oddolnych eksperymentach wymiany wodór–deuter nadzorowanej przy pomocy spektrometrii mas obserwacje stanowią widma na poziomie peptydó, podczas gdy celem badawczym jest opis wymiany na poziomie fragmentów białek, zwykle pojedynczych aminokwasów lub krótkich ciągów. Każdy zaobserwowany peptyd składa się z wielu aminokwasów, a większość aminokwasów jest pokryta przez kilka nakładających się peptydów. Wobec tego pojedynczy aminokwas wpływa na wymianę wielu peptydów, a każdy peptyd podsumowuje wymianę wielu aminokwasów. Tworzy to strukturę wielokrotnej przynależności: dane na poziomie peptydów należy modelować jako złożenie zachowania ich składowych aminokwasów, a parametry opisujące każdy aminokwas można estymować na podstawie wielu peptydów. Proponujemy model statystyczny, który reprezentuje pokrywające się segmenty sekwencji peptydów przy pomocy zmiennych losowych o rozkładach wielomianowych i modelujemy zaobserwowane widma w oparciu o sploty rozkładów tych zmiennych losowych. Omawiamy dopasowywanie tego modelu przy różnych założeniach i oceniamy go przy użyciu symulacji i analiz danych rzeczywistych z perspektywy zdolności do odtworzenia prawdopodobieństw wymiany na poziomie segmentów na podstawie rozkładów izotopowych opisujących peptydy w zaobserwowanych widmach.

Na koniec rozprawa opisuje nowe projekt i implementację `MSstats`, szeroko stosowanej rodziny programów statystycznych do analizy różnicowej ilości białek w oparciu o dane ze spektrometrii mas rozwiniętej w grupie prof. Olgi Vitek (Northeastern University, USA). Oceniamy zaproponowane zmiany z perspektywy złożoności struktury pakietów, łatwości ich rozszerzania i wydajności.

Opisujemy też praktyczne implementacje obu metod. Metoda statystyczna pozwalająca włączyć współdzielone peptydy do analizy opisana w tej rozprawie została zintegrowana z MSstats.