Nr rejestracyjny: 2020/37/N/ST6/04070; Kierownik projektu: mgr Mateusz Krystian Staniak

Identyfikacja i kwantyfikacja białek: podejście regularyzacyjne

Mateusz Staniak

Proteomika jest dziedziną badań zajmującą się charakteryzacją całego składu białkowego komórki, tkanki lub organizmu. Badanie białek pozwala uzyskać wiedzę, której nie da się uzyskać z badania samych genów lub transkryptów, bo to białka, a nie geny, determinują fenotyp komórek ze względu na procesy takie jak modyfikacje potransalcyjne (PTM). Główną technologią w proteomice, zdolną do identyfikacji i kwantyfikacji białek w próbkach biologicznych, jest spektrometria masowa (MS). Kwantyfikacja pozwala na analizę porównawczą ekspresji białek, która jest niezbędna dod wykrywania biomarkerów, tzn. znajdowania białek, których ekspresja jest związana z chorobą.

W oddolnym podejściu do MS to peptydy - mniejsze fragmenty białek - są analizowane w spektrometrze masowym i wobec tego pomiary są dokonywane na poziomie peptydu. Niektóre peptydy można powiązać z wieloma białkami. Tego problemu nie da się rozwiązać poprzez postęp technologii, bo jest spowodowany przez homologię sekwencji, często związaną z rodzinami białek (grupami powiązanych ewolucyjnie białek) lub wariantami białek (podobnymi białkami pochodzącymi z jednego genu lub rodziny genów). Takie peptydy nazywany współdzielonymi lub zdegenerowanymi peptydami. Obecność takich peptydów, wraz z białkami, które są identyfikowane tylko przez pojedynczy peptyd, utrudnia zarówno wiarygodne określenie, które białka są obecne w próbce (*identyfikacja białek*), jak i estymację ich ilości (*kwantyfikacja białek*).

Aktualnie najczęściej współdzielone peptydy są ignorowane. Powoduje to kilka problemów w kwantyfikacji białek. Po pierwsze, identyfikowana jest mniejsza liczba białek, ponieważ białka zidentyfikowane tylko przez współdzielone peptydy są usuwane z analizy wraz z odrzuconymi peptydami. Po drugie, wyestymowane ilości białek są zaburzone przez usunięcie współdzielonych peptydów (lub grupowanie białek, kiedy to ich ilość jest estymowana łącznie). Ponadto, dokładność estymacji ilości białek jest zmnniejszona przez obniżony rozmiar próby. Pierwszym celem tego projektu jest rozwój nowej metodologii statystycznej zdolnej do włączenia informacji ze współdzielonych peptydów do kwantyfikacji białek. Ścisla matematyczna analiza połączona z implementacją w R, poskutkuje lepszym zrozumieniem wpływu współdzielonych peptydów na wyniki kwantyfikacji i zwiększoną precyzję w analizie porównawczej i bezwzględnej kwantyfikacji. Nasza metodologia nie będzie usuwac informacji ze współdzielonych peptydów ani grupować niepotrzebnie białek dzięki użyciu dodatkowych informacji: niepewności z poziomu peptydów i białek, wykrywalności peptydów i otrzymanych in silico powiązań między peptydami i białkami.

Relacja pomiędzy zidentyfikowanymi peptydami i potencjalnymi białkami może być reprezentowana za pomocą grafu dwudzielnego. Wobec tego zaproponujemy nowe podejście uczenia statystycznego do analizy grafów dwudzielnych oparte na regularyzacji. Używając profili ilościowych peptydów jako danych, będziemy równocześnie estymować ilości białek or wybierać białka, które faktycznie znajdują się w próbce poprzez usuwanie krawędzi między peptydami i białkami tak, by kontrolować frakcję fałszywych odkryć wśród zidentyfikowanych białek bez poświęcania mocy statystycznej.