SPEKTROMETRIA

Piotr Formanowicz

Sekwencjonowanie DNA jest jednym z kluczowych etapów wielu badań prowadzonych na gruncie różnych gałęzi nauk biologicznych. Mimo że sekwencja DNA określa sekwencję aminokwasów odpowiadającego jest fragmentu białka, w wielu przypadkach bardzo ważna jest możliwość bezpośredniego ustalenia tejże sekwencji aminokwasów. Można to zrobić za pomocą metod **sekwencjonowania peptydów**. Metody te znacznie różnią się od metod sekwencjonowania DNA i często oparte są na **spektrometrii masowej**.

Za pomocą **spektrometru masowego** można podzielić peptyd (fragment białka, krótką sekwencję aminokwasów) na fragmenty i zmierzyć ich masy otrzymując w ten sposób **widmo masowe (spektrum)**.

Problem sekwencjonowania peptydów polega na odtworzeniu sekwencji danego peptydu na podstawie jego widma masowego.

W idelanym przypadku peptyd jest dzielony na fragmenty między każdą parą sąsiednich aminokwasów.

Dla procesu fragmentacji zachodzącego w sposób idealny i dla idealnego spektrometru problem sekwencjonowania peptydów jest obliczeniowo łatwy. Jednak w praktyce proces fragmentacji nie jest idealny i problem ten staje się obliczeniowo trudny.

Zamiast sekwencjonowania *de novo* można stosować podejście oparte na przeszukiwaniu baz danych.

Spektrum uzyskane za pomocą spektrometru (czyli spektrum eksperymentalne) porównuje się ze spektrami teoretycznymi zgromadzonymi w bazach danych. Peptyd, którego spektrum teoretyczne w największym stopniu odpowiada spektrum uzyskanemu za pomocą spektrometru prawdopodobnie jest badanym peptydem (tj. jego sekwencja aminokwasów jest sekwencją badanego peptydu).

Jednakże badane peptydy mogą być wynikiem wielu mutacji oraz modyfikacji i dlatego w bazach danych może nie być dokładnych odpowiedników ich spektr. Stanowi to poważne ograniczenie metod opartych na przeszukiwaniu baz danych.

Ponadto, metody te oczywiście nie mogą być stosowane przy analizie nieznanych wcześniej peptydów. Stąd algorytmy sekwencjonowania *de novo* są bardzo istotne w przypadku identyfikacji nieznanych białek, ale są najbardziej użyteczne wtedy, gdy ich działanie można oprzeć na spektrach wysokiej jakości.

Białka mogą podlegać bardzo wielu modyfikacjom. Prawie wszystkie sekwencje aminokwasów podlegają modyfikacjom potranslacyjnym. Znanych jest ok. 200 rodzajów modyfikacji reszt aminokwasowych. Ponieważ obecnie modyfikacje potranslacyjne nie mogą być wykrywane na podstawie sekwencji DNA, ich wykrywanie pozostaje nadal bardzo ważnym otwartym problemem.

Problem określenia sekwencji aminokwasów na podstawie widma masowego można rozwiązać za pomocą opisanej dalej metody, przedstawionej m. in. w [1, 2].

Niech A będzie **zbiorem aminokwasów** i dla każdego $p \in A$ niech $m(p)$ będzie **masą aminokwasu** p.

Peptyd $P = p_1, p_2, \ldots, p_n$ jest sekwencją aminokwasów.

Masa $m(P)$ **peptydu** P jest równa $m(P) = \sum_{i=1}^{n} m(p_i)$.

Częściowy peptyd P' jest podciągiem $p_i, \ldots, p_j$ peptydu P o masie $\sum_{i \le t \le j} m(p_t)$.

Fragmentacja peptydu w tandemowym spektrometrze masowym może być scharakteryzowana przez zbiór liczb $\Delta = \{\delta_1, \delta_2, \ldots, \delta_k\}$ odpowiadających **rodzajom jonów**.

Jon δ częściowego peptydu $P' \subset P$ jest modyfikacją peptydu P' taką, że jego masa wynosi $m(P') - \delta$.

Dla tandemowej spektrometrii masowej **teoretyczne spektrum** peptydu P może być obliczone przez odjęcie wszystkich możliwych rodzajów jonów $\delta_1, \delta_2, \ldots, \delta_k$ od mas wszystkich częściowych peptydów peptydu P. Oznacza to, że każdy częściowy peptyd wnosi k mas do teoretycznego spektrum.

Spektrum eksperymentalne $S = \{s_1, s_2, \ldots, s_m\}$ jest zbiorem mas jonów.

Dopasowanie między spektrum S i peptydem P jest liczbą elementów wspólnych spektrum teoretycznego i spektrum eksperymentalnego.

PROBLEM SEKWENCJONOWANIA PEPTYDÓW
INSTANCJA: Spektrum S, zbiór rodzajów jonów Δ, masa m.
ODPOWIEDŹ: Peptyd P o masie m taki, który ma największe dopasowanie do S.

Oznaczmy **częściowy N-końcowy peptyd** $p_1, p_2, \ldots, p_i$ jako P_i oraz **częściowy C-końcowy peptyd** $p_{i+1}, p_{i+2}, \ldots, p_n$ jako P_i^-, $i = 1, 2, \ldots, n$.

W praktyce spektrum uzyskane za pomocą tandemowego spektrometru masowego zawiera głównie δ-jony częściowych N-końcowych i C-końcowych peptydów. Stąd, teoretyczne spektrum zawiera tylko jony N-końcowych i C-końcowych częściowych peptydów.

Przykładowo, najczęściej pojawiające się N-końcowe jony to b-jony (b_i odpowiada P_i z $\delta = -1$), a najczęściej pojawiające się C-końcowe jony to y-jony (y_i odpowiada P_i^- z $\delta = 19$).

Algorytmy sekwencjonowania peptydów oparte są na ogół na wyczerpującym przeszukiwaniu lub na **grafie spektrum**.

Podejście oparte na wyczerpującym przeszukiwaniu polega na wygenerowaniu wszystkich sekwencji aminokwasowych oraz odpowiadających im spektr teoretycznych i odnalezieniu tych sekwencji, których spektra teoretyczne w największym stopniu pasują do spektrum eksperymentalnego.

Niestety, liczba sekwencji rośnie wykładniczo wraz z ich długością, dlatego muszą być stosowane różne metody ograniczania przestrzeni przeszukiwania.

Odcinanie prefiksów ogranicza przestrzeń rozwiązań do sekwencji, których prefiksy pasują do spektrum eksperymentalnego.

Pojawia się jednak w tym przypadku problem z poprawnymi sekwencjami, których prefiksy nie są wystarczająco dobrze odwzorowane w spektrum – mogą one zostać odrzucone.

Inny poważny problem polega na tym, że informacja o spektrum jest używana dopiero po wygenerowaniu sekwencji aminokwasowych.

Podejścia oparte na **grafach spektrum** są na ogół bardziej efektywne, ponieważ informacja zawarta w spektrum jest wykorzystywana zanim potencjalne sekwencje aminokwasowe zostaną ocenione.

W podejściu tym piki ze spektrum (widma) są przekształcone w graf spektrum. Każdemu pikowi odpowiadają wierzchołki w grafie, natomiast łuki łączą wierzchołki odpowiadające częściowym peptydom, których masy różnią się o masę jednego aminokwasu.

Każdy pik w widmie jest przekształcany w pewną liczbę wierzchołków w grafie – każdy z tych wierzchołków odpowiada możliwemu typowi jonu, który można przyporządkować danemu pikowi.

Problem sekwencjonowania peptydów odpowiada problemowi poszukiwania najdłuższej ścieżki w tak otrzymanym skierowanym grafie acyklicznym.

Niech spektrum $S = \{s_1, s_2, \ldots, s_m\}$. Załóżmy, że spektrum zawiara tylko N-końcowe jony. Konstruujemy na jego podstawie graf spektrum $G_\Delta(S)$. Wierzchołki tego grafu odpowiadają liczbom całkowitym $s_i + \delta_j$ reprezentującym potencjalne masy częściowych peptydów.

Każdemu elementowi spektrum $s \in S$ odpowiada k wierzchołków $V(s) = \{s + \delta_1, s + \delta_2, \ldots, s + \delta_k\}$. Zatem zbiór wierzchołków konstruowanego grafu to $V = \{s_{pocz}\} \cup V(s_1) \cup V(s_2) \cup \ldots \cup V(s_m) \cup \{s_{konc}\}$, gdzie $s_{pocz} = 0$ oraz $s_{konc} = m(P)$.

Wierzchołki v_i i v_j są połączone łukiem (v_i, v_j), jeżeli $v_j - v_i$ jest równe masie któregoś z aminokwasów i łuk taki jest zaetykieotwany nazwą tego aminokwasu.

Jeżeli wierzchołki odpowiadają potencjalnym N-końcowym częściowym peptydom, istnienie łuku (v_i, v_j) oznacza, że sekwencja aminokwasów odpowiadająca wierzchołkowi v_i może być rozszerzona do sekwencji odpowiadającej wierzchołkowi v_j przez dodanie jednego aminokwasu.

Spektrum S peptydu P nazywane jest **pełnym**, jeżeli zawiera ono przynajmniej jeden typ jonu odpowiadający P_i dla każdego $i = 1, 2, \ldots, n$.

Możliwość zastosowania grafu spektrum wynika z faktu, że dla pełnego spektrum istnieje w nim skierowana ścieżka o długości n z wierzchołka s_{pocz} do wierzchołka s_{konc} zaetykietowana peptydem P.

Niestety, rzeczywiste dane na ogół nie są idealne i otrzymane na ich podstawie spektrum nie jest pełne.

Ponadto, za pomocą spektrometrów różnych rodzajów otrzymuje się na ogół różne spektra dla tego samego peptydu. Różne metody jonizacji stosowane w spektrometrach różnych typów mają wpływ na powstawanie jonów różnych rodzajów. Stąd, algorytmy sekwencjonowania powinny być dostosowane do konkretnych rodzajów spektrometrów.

Jednym ze sposobów rozwiązania tego problemu jest automatyczne wykrywanie, jakiego rodzaju jony powstają w danym spektrometrze na podstawie spektr dla znanych sekwencji aminokwasów.

Podejście to oparte jest na **funkcji częstości przesunięcia**, za pomocą której można określić tendencję danego spektrometru do generowania jonów określonych rodzajów.

Jeżeli rodzaje jonów $\Delta = \{\delta_1, \delta_2, \ldots, \delta_k\}$ generowane przez dany spektrometr nie są znane, otrzymane spektrum nie może zostać poprawnie zinterpretowane.

Rodzaje jonów generowanych przez spektrometr

można określić na podstawie danych eksperymentalnych bez żadnej wiedzy na temat wzorców fragmentacji w następujący sposób.

Niech $S = \{s_1, s_2, \ldots, s_m\}$ będzie spektrum odpowiadającym peptydowi P. Dla częściowego peptydu P_i oraz piku s_j istnieje przesunięcie $x_{ij} = m(P_i) - s_j$.

Zmienną x_{ij} można traktować jako zmienną losową.

Ponieważ prawdopodobieństwa przesunięć odpowiadających rzeczywistym jonom fragmentów są dużo większe niż prawdopodobieństwa przesunięć losowych, piki w empirycznym rozkładzie przesunięć wskazują na uzyskiwane jony fragmentów.

Statystyka przesunięć dla wszystkich jonów i wszystkich częściowych peptydów jest podstawą algorytmu określania typów jonów.

Mając spektrum S, przesunięcie x oraz dokładność ϵ niech $H(x, S)$ będzie liczbą par (P_i, s_j), $i = 1, 2, \ldots, n-1$, $j = 1, 2, \ldots, m$ mających przesunięcie $x_{ij} = m(P_i) - s_j$ odległe nie bardziej niż o ϵ od x (tj. $|x_{ij}| \leq |x - \epsilon|$).

Funkcja częstości przesunięcia jest zdefiniowana jako $H(x) = \sum_S H(x, S)$, gdzie sumowanie przebiega po wszystkich spektrach z danych uczących.

W celu wykrycia rodzajów jonów C-końcowych należy zastosować tę samą procdurę dla par (P_i^-, s_j).

Przesunięcia $\Delta = \{\delta_1, \delta_2, \ldots, \delta_k\}$ odpowiadające pikom $H(x)$ reprezentują rodzaje jonów generowane przez dany spektrometr.

Literatura

[1] Pavel A. Pevzner. *Computational Molecular Biology. An Algorithmic Approach.* The MIT Press, Cambridge, Massachusetts 2000.

[2] Pavel A. Pevzner, Vlado Dančík, Chris L. Tang. Mutation-Tolerant Protein Identi cation by Mass Spectrometry. *Journal of Computational Biology* 7, 2000, 777–787.

v. 1.0.0