Adam KOZAK, Tomasz GŁOWACKI, Piotr FORMANOWICZ
Politechnika Poznańska

KLASYFIKACJA PROBLEMÓW ASEMBLACJI I SEKWENCJONOWANIA PEPTYDÓW

Streszczenie.
Sekwencjonowanie peptydów polega na ustaleniu kolejności aminokwasów (sekwencji) w cząsteczce. Bezpośrednie metody chemii analitycznej pozwalają na określenie jedynie krótkich sekwencji. Alternatywą dla tych metod jest spektrometria masowa. Widmo masowe powstałe w wyniku przeprowadzenia eksperymentu za pomocą spektrometru wymaga dodatkowej analizy. Z punktu widzenia nauk obliczeniowych analiza takiego widma jest źródłem ciekawych problemów. Ta metoda sekwencjonowania posiada swoje ograniczenia co do długości sekwencji. Rodzi to naturalną potrzebę projektowania metod asemblacyjnych, które pozwolą połączyć wiele krótkich łańcuchów w oryginalną cząsteczkę. W pracy tej przedstawiono problemy sekwencjonowania i asemblacji oraz zaproponowano ich klasyfikację. Przedstawiono również wybrane metody rozwiązujące te problemy.

SEQUENCING AND ASSEMBLY PROBLEMS CLASSIFICATION

Summary. Peptides sequencing is a determination of amino acid order (sequence) in a molecule. Direct analitycal methods allow to recognize only short sequences. The alternative method is based on mass spectrometry. As a result of a mass spectrometry based experiment a mass spectrum is obtained. An analysis of such a spectrum is a source of interesting computational problems. This approach has sequence length limitation. Because of that there is a need for assembly methods that allow to bring many short pieces together. In this paper peptide sequencing and assembling problems are described and their classification is also proposed. Selected methods for the problems were also discussed.

1. Wprowadzenie

Peptydy to związki chemiczne składające się z wielu połączonych ze sobą aminokwasów [5]. Istnieje 20 typów aminokwasów wchodzących w skład białek. Każde dwa kolejne aminokwasy są ze sobą połączone specjalnym wiązaniem zwanym wiązaniem peptydowym. W reakcji biosyntezy peptydy tworzą jedynie proste nierozgałęzione łańcuchy. Kolejność aminokwasów w cząsteczce, czyli ich sekwencja jest nazywana strukturą pierwszorzędową. Długie peptydy o masie powyżej 10000 Daltonów nazywane są białkami. Związki te pełnią wiele ważnych funkcji w organizmie m.in.:

- katalizują reakcje,

- transportują inne cząsteczki w organizmie,
- regulują procesy transkrypcji i ekspresji poszczególnych genów,
- są odpowiedzialne za skurcze mięśni.

Ze względu na skalę przestrzenną, można wyróżnić cztery poziomy opisu białek:

- struktura pierwszorzędowa,
- struktura drugorzędowa - przestrzenne ułożenie łańcuchów,
- struktura trzeciorzędowa - wzajemne położenie elementów struktury drugorzędowej,
- struktura czwartorzędowa - wzajemne położenie łańcuchów i ewentualnych struktur niebiałkowych.

Poznanie sekwencji aminokwasów badanego peptydu to pierwszy krok do ustalenia jego przestrzennej budowy, a co za tym idzie określenia własności chemicznych oraz jego funkcji. Dodatkowo budowa organizmów, ich fizjologia czy nawet zachowanie są wynikiem występowania w komórkach odpowiednich białek. Brak bezpośrednich metod ustalania struktury pierwszorzędowej długich białek oraz duży wolumen i skomplikowanie danych pochodzących ze spektrometru masowego i tych wykorzystywanych w procesie asemblacji, w naturalny sposób angażują w te badania specjalistów z biologii obliczeniowej i dziedzin pokrewnych. Wyjaśnia to wagę i motywację omawianych problemów.

2. Eksperymenty chemiczne

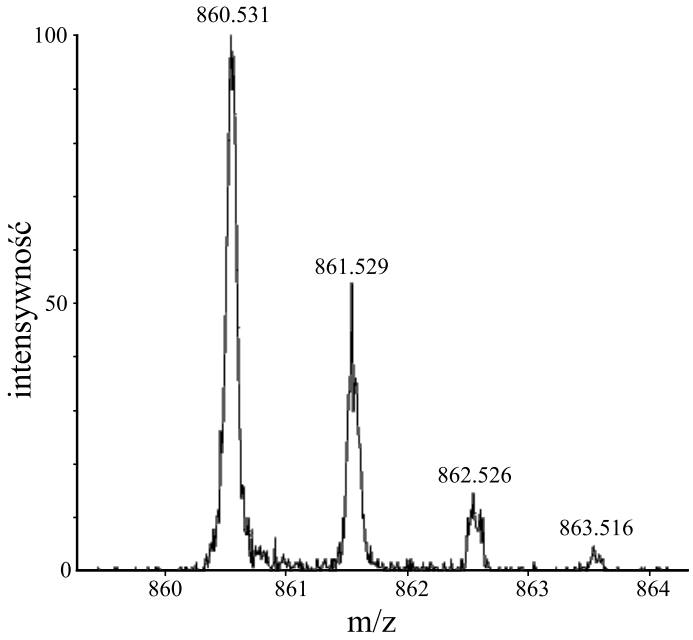
2.1. Sekwencjonowanie

Jak wspomniano, nie istnieją bezpośrednie metody pozwalające na poznanie sekwencji dowolnie długich białek. Jedną z metod chemii analitycznej wykorzystywaną do sekwencjonowania jest degradacja Edmana. W praktyce metoda ta może być wykorzystana do poznania sekwencji nieprzekraczających 50 aminokwasów [4]. Warto również zaznaczyć, że metoda ta nie generuje danych, które wymagałyby interpretacji specjalistów z dziedziny biologii obliczeniowej i została przytoczona tutaj jedynie jako alternatywa dla sekwencjonowania przy użyciu spektrometrii masowej.

W eksperymencie w spektrometrze masowym badana cząsteczka ulega jonizacji i fragmentacji [6]. Następnie jony fragmentaryczne zostają rozdzielone ze względu na stosunek ich masy do ładunku. Informacje pochodzące ze spektrometru przedstawia się w postaci widma masowego tj. wykresu na którym na osi rzędnych przedstawia się stosunek masy do ładunku zaobserwowanego jonu (m/z), a na osi odciętych jego względne stężenie/intensywność (Rys.1) .

Widmo masowe jest unikalne dla danego związku chemicznego i w uproszczeniu może być traktowane jako "odcisk palca" tej substancji. Omawiany eksperyment może być źródłem różnych błędów:

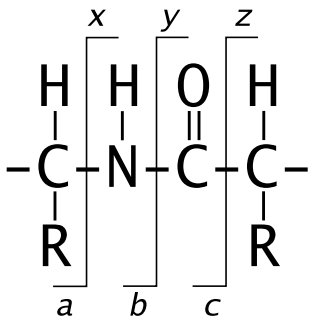
- błąd pozytywny to sytuacja, gdy na widmie pojawi się jon, który nie odpowiada rzeczywistemu jonowi fragmentacyjnemu cząsteczki; taki jon może pochodzić z jonizacji zanieczyszczenia,



Rys. 1. Przykładowe widmo masowe

- błąd negatywny to brak w widmie masowym jonu, który powinien pojawić się w procesie fragmentacji.

Z punktu widzenia analizy danych spektrometrycznych interesujące są modyfikacje potranslacyjne (zwane również potranslacyjnymi). Są to zmiany, które występują po procesie translacji i mają wpływ na masę oraz własności fizykochemiczne analizowanych białek. Należy również wziąć pod uwagę tzw. wyjątki - sytuacje nietypowe, które również wpływają na masę cząsteczki, takie jak wydzielenie wody lub amoniaku. W procesie fragmentacji peptydu, pęknięcie cząsteczki może nastąpić na dowolnym wiązaniu peptydowym, w jednym z trzech miejsc tego wiązania a jonizacji może ulec N-końcowa (lewa) lub C-końcowa (prawa) część cząsteczki. W związku z tym, że do pęknięcia może dojść w jednym z trzech miejsc wiązania, a jonizacji ulega jeden z dwóch fragmentów cząsteczki, można wyróżnić 6 typów jonów. Dodatkowo należy wyróżnić dwie sytuacje:



Rys. 2. Typy cięć podczas rozpadu wiązania peptydowego

- gdy pęknięciu i jonizacji zawsze ulega cała obojętna cząsteczka białka, czyli powstały jon odpowiada prefiksowi lub sufiksowi cząsteczki,
- gdy pęknięciu i jonizacji mogą ulegać zjonizowane fragmenty cząsteczki, czyli

wynikiem procesu mogą być tzw. jony wewnętrzne odpowiadające wewnętrznym fragmentom cząsteczki.

2.2. Asemblacja

Wynikiem sekwencjonowania jest poznanie krótkiej sekwencji aminokwasowej. Aby ustalić kolejność aminokwasów w przypadku dłuższych sekwencji, naturalnym rozwiązaniem jest podzielenie długiej sekwencji na wiele krótkich. Do pocięcia (trawienia) łańcucha wykorzystuje się endopeptydazy. Są to związki, które tną sekwencje wewnątrz, zawsze po wystąpieniu jednego z kilku określonych dla tej endopeptydazy aminokwasów. Przykładowo, trypsyna tnie łańcuch po argininie i lizynie. Problemem pozostaje ułożenie tych krótkich sekwencji w odpowiedniej kolejności. W tym celu w literaturze [1] zaproponowano podział materiału biologicznego do dwóch naczyń i przeprowadzenie reakcji w każdym z nich przy użyciu innej endopeptydazy. W wyniku tej procedury otrzymuje się dwie mieszaniny krótkich peptydów. Mieszaninę rozdziela się, a krótkie łańcuchy sekwencjonuje dowolną metodą. Należy zauważyć, że dzięki pocięciu łańcuchów po różnych aminokwasach, fragmenty pokrywające ten sam kawałek oryginalnej cząsteczki częściowo się pokrywają. Analiza tych pokryć umożliwia rekonstrukcję szukanej sekwencji. Przedstawiony eksperyment może być źródłem następujących błędów:

- błąd pozytywny to sytuacja, gdy w spektrum krótkich peptydów pojawia się łańcuch, który nie pochodzi z asemblowanej cząsteczki; taka sytuacja może mieć miejsce w wyniku przekłamania w procesie sekwencjonowania lub w wyniku odczytu zanieczyszczeń,
- błąd negatywny to sytuacja, gdy w spektrum nie pojawia się łańcuch, który powinien się pojawić; źródłem błędów negatywnych może być zagubienie krótkiej sekwencji lub brak informacji o powótrzeniach takich sekwencji, gdy sposób analizy mieszaniny nie pozwala na uzyskanie takiej informacji,
- błąd procesu trawienia jest to sytuacja gdy w pewnych miejscach cząsteczki nie zachodzą cięcia, które wynikałyby z mechanizmu działania endopeptydazy.

Opcjonalnie istnieje możliwość poznania rozkładu aminokwasów w oryginalnej sekwencji, co pozwala na dodatkową weryfikację potencjalnych rozwiązań. W celu otrzymania informacji o rozkładzie, przeprowadza się pełne trawienie białka a następnie mierzy się stężenie wszystkich aminokwasów w roztworze.

3. Klasyfikacja problemów sekwencjonowania

W tym rozdziale zostaną omówione problemy związane z sekwencjonowaniem za pomocą spektrometru masowego, gdyż jak wspomniano, degradacja Edmana nie stanowi inspiracji do wykorzystania metod kombinatorycznych. Można wyróżnić następujące podejścia do analizy widma masowego:

- przeszukiwanie baz danych,
- markery sekwencji,
- sekwencjonowanie de novo,
- podejście mieszane.

Przeszukiwanie baz danych polega na znalezieniu widma, które najbardziej przypomina to uzyskane w spektrometrze masowym. Należy tutaj mieć na względzie, że warunki eksperymentu oraz wykorzystany sprzęt mogą mieć wpływ na uzyskane widmo. Spośród znanych rozwiązań warto wymienić Sequest [10], Mascot [9], Tandem [11]. Oczywistym ograniczeniem jest fakt, że podejście to umożliwia rozpoznanie tylko tych związków, których widma znajdują się w przeszukiwanych bazach danych.

Markery sekwencji (sequence tagging) to istotne piki z widma masowego, które pozwalają na rozpoznanie danego białka. W tym podejściu bazy danych przeszukuje się w celu znalezienia widma posiadającego te wybrane piki. Zasadniczą zaletą tego podejścia jest możliwość znalezienia białek, które zostały poddane modyfikacjom potranslacyjnym. W tym wypadku białko zostaje rozpoznane po kilku pikach, na które modyfikacje nie miały wpływu. Jest to ważna przewaga tego podejścia w porównaniu do standardowego przeszukiwania baz danych. Najbardziej znane metody to: OpenSea [14], GutenTag [12] oraz SPIDER[13]. Markery sekwencji pozwalają na rozpoznanie białek poddanych modyfikacjom potranslacyjnym, pod warunkiem że oryginalne białko (pozbawione modyfikacji) znajduje się w bazie danych.

Podejście, w którym do rozpoznania sekwencji aminokwasowej wykorzystuje się jedynie informacje zawarte w widmie masowym, nazywane jest metodą sekwencjonowania de novo.

Na potrzeby dalszych rozważań zdefiniujmy pojedynczy pik z widma masowego jako uporządkowaną dwójkę (m, n), gdzie m to wartość na osi X dla tego (m/z), a n to wartość na osi Y (intensywność) dla tego piku. Matematycznie widmo masowe można przedstawić jako zbiór takich uporządkowanych dwójek.

Widmo, które zawiera wszystkie możliwe do uzyskania piki pochodzące z pęknięcia i jonizacji jednie obojętnych cząsteczek białka (czyli odpowiadające jedynie prefiksom lub postfiksom cząsteczki) zwane jest widmem idealnym i oznaczone $W^{a,b,c,x,y,z}$. Należy wyróżnić podzbiory widma idealnego, które zawierają wszystkie piki pochodzące od cięć jednego lub kilku wybranych typów i nie zawierają żadnych innych pików np. W^a zawiera wszystkie i tylko te piki, które pochodzą z jonów typu „a", natomiast $W^{a,x,z}$ zawiera wszystkie piki pochodzące od jonów „a", „x" oraz „z" i nie zawiera innych pików, np $W^{a,x,z} = W^a \cup W^x \cup W^z$.

Widmo, które zawiera wszystkie możliwe do uzyskania piki, również te odpowiadające wewnętrznym fragmentom cząsteczki zwane jest widmem pełnym i oznaczone W^*. Należy zauważyć, że $W^{a,b,c,x,y,z} \subseteq W^*$.

W poniższych problemach parametr widma W oznacza widmo pochodzące z eksperymentu (w zależności od rozważanej sytuacji mogą występować błędy), natomiast widma poszczególnych typów cięć ($W^a, W^{a,b,c}$ etc.) oznaczają widma, które byłyby uzyskane z bezbłędnego eksperymentu. Ogólny problem sekwencjonowania de novo można zdefiniować następująco:

Problem 1. Instancja: Uporządkowana trójka (W, m, f), gdzie W to widmo masowe, m jest masą analizowanego peptydu, a f to funkcja oceniająca trafność znalezionej sekwencji do widma masowego W.
Odpowiedź: Sekwencja aminokwasowa o masie m maksymalizująca wartość funkcji f.

W literaturze zaproponowano szereg algorytmów wielomianowych dla konkretnych warunków fizykochemicznych eksperymentu i urządzeń, gdyż od tych parametrów

zależy charakterystyka widma. Poniżej przedstawiono kilka ciekawych problemów sekwencjonowania de novo (1-7) dla których istnieją rozwiązania dokładne o złożoności wielomianowej. Odpowiedzi dla poniższych problemów są sformułowane identycznie jak w problemie 1.

Problem 2. Instancja: Uporządkowana trójka (W, m, f), gdzie

$$W \in \{W^a, W^b, W^c, W^x, W^y, W^z\}$$

oraz m jest masą analizowanego peptydu, a f to funkcja oceniająca trafność znalezionej sekwencji do widma masowego W.

Problem 2 przedstawia sytuację, gdy w drodze fragmentacji z każdej cząsteczki tworzy się jon tego samego typu (a, b, c, x, y lub z) i jon nie ulega dalszej fragmentacji. Dodatkowo zakłada się, że zaobserwowano wszystkie możliwe jony fragmentacyjne i brak zanieczyszczeń. Rozwiązanie można znaleźć obliczając różnicę mas między każdymi kolejnymi dwoma pikami i różnica ta determinuje kolejny aminokwas w sekwencji.

Problem 3. Instancja: Uporządkowana trójka (W, m, f), gdzie

$$W \in \{W^{a,b,c}, W^{a,b}, W^{a,c}, W^{b,c}, W^a, W^b, W^c\}$$

oraz m jest masą analizowanego peptydu, a f to funkcja oceniająca trafność znalezionej sekwencji do widma masowego W.

Problem 3 przedstawia sytuację, gdy w drodze fragmentacji tworzą się jony tylko z jednego końca cząsteczki. Dodatkowo zakłada się, że uzyskano wszystkie piki i brak jest zanieczyszczeń. Rozwiązanie można znaleźć przy wykorzystaniu podobnej metody jak dla problemu nr 2, rozpoznając typy jonów po specyficznej różnicy mas pomiędzy pikami (przykładowo różnica mas pomiędzy jonem typu „c" a jonem typu „b" pochodzącymi z przerwania tego samego wiązania peptydowego jest stała).

W pracy [8] rozważano przypadki sekwencjonowania zbliżone do sytuacji rzeczywistej, gdy dostarczając niewielką energię do spektrometru uzyskuje się w nim wszystkie jony typu „c" oraz „y". Zaproponowano algorytmy wielomianowe bazujące na założeniach programowania dynamicznego dla sytuacji idealnej (problem 4), sytuacji z błędami negatywnymi (problem 5) oraz dla sytuacji z błędami pozytywnymi (problem 6). Pokazano, że w wypadku występowania jednej modyfikacji potranslacyjnej problem jest również łatwy obliczeniowo (problem 7).

Problem 4. Instancja: Uporządkowana trójka (W, m, f), gdzie $W = W^{c,y}$, m jest masą analizowanego peptydu, a f to funkcja oceniająca trafność znalezionej sekwencji do widma masowego W.

Problem 5. Instancja: Uporządkowana trójka (W, m, f), gdzie $W \subseteq W^{c,y}$, m jest masą analizowanego peptydu, a f to funkcja oceniająca trafność znalezionej sekwencji do widma masowego W.

Problem 6. Instancja: Uporządkowana trójka (W, m, f), gdzie $W^{c,y} \subseteq W$, m jest masą analizowanego peptydu, a f to funkcja oceniająca trafność znalezionej sekwencji do widma masowego W.

Problem 7. Instancja: Uporządkowana trójka (W, m, f), gdzie

$$\forall_{(m,n)\in W}\ (m,n) \in W^{c,y} \quad \vee \quad \exists_{t\in Q^+}\ \forall_{(m,n)\in W}\ (m+t,n) \in W^{c,y}$$

oraz t oznacza masę modyfikacji potranslacyjnej, Q^+ oznacza zbiór dodatnich liczb wymiernych, m jest masą analizowanego peptydu, a f to funkcja oceniająca trafność znalezionej sekwencji do widma masowego W.

Zostało pokazane, że przypadek sekwencjonowania, gdy w widmie występują piki odpowiadające wewnętrznym fragmentom cząsteczki, czyli w sytuacji gdy zaobserwowano tzw. jony wewnętrzne (problem 8), jest trudny obliczeniowo [15].

Problem 8. Instancja: Uporządkowana trójka (W, m, f), gdzie

$$W \subseteq W^* \quad \wedge \quad W \not\subset W^{a,b,c,x,y,z}$$

oraz m jest masą analizowanego peptydu, a f to funkcja oceniająca trafność znalezionej sekwencji do widma masowego W.

Podejście mieszane łączy ze sobą trzy przedstawione sposoby rekonstrukcji cząsteczki oryginalnej.

4. Klasyfikacja problemów asemblacji

Wynikiem eksperymentu chemicznego w procesie asemblacji są dwa multizbiory krótkich peptydów. Multizbiór wszystkich krótkich sekwencji otrzymanych w eksperymencie z dwoma peptydazami nazywany jest spektrum.

W przypadku, gdy zachodzą wszystkie cięcia, których można oczekiwać z mechanizmu działania endopeptydaz, powstałe spektrum nosi nazwę *spektrum* idealnego i jest oznaczone P^i.

Należy rozpatrzyć sytuację, gdy nie wszystkie cięcia zachodzą. Załóżmy, że cząsteczka posiada k aminokwasów, po których może nastąpić cięcie. Podejmując dla każdego z aminokwasów decyzję, czy nastąpi po nim cięcie czy nie, otrzymuje się pewien multizbiór krótkich fragmentów tego łańcucha. Należy zauważyć, że istnieje 2^k różnych procesów decyzyjnych prowadzących do różnego pocięcia sekwencji i powstania innych krótkich sekwencji. Załóżmy dodatkowo, że na każdej cząsteczce w roztworze przeprowadzono inny proces decyzyjny (uzyskano inne fragmenty z jej pocięcia) a w roztworze znajduje się na tyle dużo cząsteczek, aby przeprowadzić wszystkie możliwe procesy decyzyjne (co najmniej 2^k cząsteczek). Wynikiem takiego trawienia jest uzyskanie *pełnego widma*, czyli widma które zawiera wszystkie możliwe do uzyskania krótkie sekwencje. Pełne widmo oznaczone zostanie P^p.

Każdy łańcuch peptydowy może być przedstawiony jako słowo nad 20-literowym alfabetem Σ. Każdemu spektrum odpowiada zatem pewien multizbiór słów. Ze spektrum P^i zostanie powiązany odpowiadający mu zbiór słów S^i, natomiast P^p odpowiada zbiór słów S^p.

Jak wspomniano, występuje częściowe nakładanie się tych peptydów. W przypadku, gdy zachodzą wszystkie cięcia za pomocą endopeptydaz, rezultatem eksperymentu jest widmo idealne i wyniki można przedstawić w postaci modelu grafowego, gdzie każdy krótki peptyd odpowiada wierzchołkowi grafu [1]. Każdemu aminokwasowi zostaje przypisana odpowiadająca mu litera, wierzchołki zostają zaetykietowane

ciągami znaków opisującymi związane z nimi krótkie sekwencje. Jeśli prefiks pewnego wierzchołka jest równy sufiksowi drugiego, a dodatkowo wierzchołki reprezentują sekwencje pochodzące z reakcji z różnymi endopeptydazami, to istnieje łuk prowadzący od tego pierwszego wierzchołka do drugiego. Rozwiązaniem problemu asemblacji jest znalezienie w tym grafie ścieżki Hamiltona. Zostało pokazane [1], że tak zdefiniowany graf jest grafem sprzężonym. Aby znaleźć cykl Hamiltona w grafie sprzężonym wystarczy znaleźć obwód Eulera w grafie oryginalnym tego grafu [16]. Jako, że istnieją wielomianowe algorytmy znajdujące obwód Eulera, to omawiany problem asemblacji jest łatwy obliczeniowo. Formalnie powyższy problem asemblacji można sformułować następująco:

Problem 9. Instancja: Multizbiór słów S_i nad alfabetem Σ.
Odpowiedź: Superciąg dla zbioru słów S.

W przypadku, gdy część cięć oczekiwanych przy działaniu endopeptydazy nie zachodzi, wyniki eksperymentu chemicznego mogą zostać przedstawione jako multigraf. W tym wypadku każdej krótkiej sekwencji również odpowiada wierzchołek tego grafu, zasadnicza różnica polega na tym, że należy rozważyć różne możliwe nałożenia dwóch etykiet. Można pokazać, że problem asemblacji bez wszystkich cięć jest trudny obliczeniowo [3], w przypadku gdy znany jest rozkład aminokwasów (problem 10) oraz w sytuacji gdy nie ma tej dodatkowej informacji (problem 11).

Problem 10. Instancja: Multizbiór słów S_p nad alfabetem Σ oraz rozkład D symboli z alfabetu Σ tj. zbiór par (x, i) dla wszystkich symboli x z alfabetu Σ, gdzie i jest nieujemną liczbą całkowitą.
Odpowiedź: Superciąg dla zbioru słów S spełniający rozkład D.

Problem 11. Instancja: Multizbiór słów S_p nad alfabetem Σ.
Odpowiedź: Superciąg dla zbioru słów S.

Ciekawa jest sytuacja pośrednia pomiędzy omówionymi powyżej dwoma przypadkami, gdy nie ma wszystkich cięć, jednak spektrum zawiera wszystkie krótkie sekwencje odpowiadające pełnemu trawieniu (problem 12). Spektrum jest w tym wypadku nadzbiorem spektrum uzyskanego w przypadku idealnym. Można wykazać, że problem ten jest łatwy obliczeniowo[3]. Aby znaleźć rozwiązanie problemu w czasie wielomianowym, należy wskazać w zbiorze słów wszystkie te słowa, które odpowiadają fragmentom pochodzącym z pełnego trawienia, a następnie wykorzystać algorytm zaproponowany dla przypadku idealnego. Słowa odpowiadające fragmentom pochodzącym z procesu pełnego trawienia można rozpoznać po tym, że zawierają co najwyżej jedną literę odpowiadającą aminokwasowi po którym następuje cięcie. Jest to ostatnia litera w tych słowach. Wybranie odpowiednich słów można wykonać zatem w czasie liniowym.

Problem 12. Instancja: Multizbiór słów S nad alfabetem Σ, taki że $S \subseteq S_p$ oraz $S_i \subseteq S$.
Odpowiedź: Superciąg dla zbioru słów S.

5. Podsumowanie

W pracy przedstawiono klasyfikację problemów sekwencjonowania oraz asemblacji peptydów. W przypadku sekwencjonowania peptydów zaprezentowano dwa typowe podejścia: degradację Edmana oraz spektrometrię masową. Pierwsze podejście nie wymaga wykorzystania metod kombinatorycznych. W przypadku drugiego podejścia wynikiem eksperymentu jest widmo masowe, które następnie w całości lub częściowo wyszukuje się w bazie danych lub stosuje podejście de novo - ustalenie struktury pierwszorzędowej bazując tylko na informacjach dostępnych w widmie. Dla podejścia de novo zaprezentowano w literaturze wiele algorytmów, w zależności od urządzenia i warunków przeprowadzenia eksperymentu. W literaturze udowodniono, że problem sekwencjonowania de novo z jonami wewnętrznymi jest trudny obliczeniowo oraz zaprezentowano wiele algorytmów wielomianowych w przypadku, gdy na widmie dostępne są tylko jony odpowiadające prefiksom i/lub sufiksom cząsteczki. W przypadku problemów asemblacji, zostało pokazane w literaturze, że w przypadku idealnym, gdy zachodzą wszystkie cięcia, problem jest łatwy obliczeniowo. W przypadku, gdy część oczekiwanych cięć nie zachodzi, problem jest obliczeniowo trudny. Dla problemu trudnego obliczeniowo zaproponowano kilka rozwiązań przybliżonych w literaturze [17,18]. Rozważono również sytu-ację, gdy uzyskane widmo jest nadzbiorem widma idealnego. W tym wypadku problem nadal jest łatwy obliczeniowo.

LITERATURA

1. Błażewicz J., Borowski M., Formanowicz P., Głowacki T. On graph theoretical models for peptide sequence assembly, Foundations of Computing and Decision Sciences 30 (2005) p. 183–191.
2. Formanowicz P. Selected Combinatorial Aspects of Biological Sequence Analysis, Poznań, Publishing House of Poznań University of Technology 2005.
3. Gallant J. K. The complexity of the overlap method for sequencing biopolymers. Journal of Theoretical Biology 101 (1983) p. 1–17.
4. Stryer L. Biochemistry, 4th edition, New York, W.H. Freeman and Company, 1995.
5. Doonan S. Peptides and Proteins. Royal Society of Chemistry, 2002.
6. Johnstone R. A. W. Mass spectrometry for organic chemists. Cambridge University Press, 1972.
7. Kraj A., Silberring J. Proteomika. EJB, Kraków, 2004.
8. Chen T., Kao M. Y., Tepel M., Rush J., Church G. M. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. Journal of Computational Biology, 8:325–337, 2001.
9. Perkins D. N., Pappin D. J., Creasy D. M., Cottrell J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20(18):3551–3567, 1994.
10. Eng J. K., McCormack A. L., Yates J. R. An approach to correlate tandem mass

spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry, 5:976–989, 1994.

11. Craig R., Beavis R.C. A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commununications in mass spectrometry: RCM, 17:2310–2316, 2003.
12. Tabb D. L., Saraf A., Yates J. R. GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model. Analytical Chemistry, 75:6415–6421, 2003.
13. Han Y., Ma B., Zhang K. SPIDER: Software for protein identification from sequence tags with de novo sequencing error. Journal of Bioinformatics and Computational Biology, 3:697–716, 2005.
14. Searle B. C., Dasari S., Turner M., Reddy A. P., Choi D., Wilmarth P. A., McCormack A. L., David L. L., Nagalla S. R. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. Analytical Chemistry, 76:2220—2230, 2004.
15. Xu C., Ma B. Complexity and scoring function of MS/MS peptide de novo sequencing. Computational Systems Bioinformatics / Life Sciences Society. Computational Systems Bioinformatics Conference, 361–369, 2006.
16. Błażewicz J., Hertz A., Kobler D., de Werra D. On some properties of DNA graphs. Discrete Applied Mathematics, 98:1–19, 1999.
17. Głowacki T., Kozak A., Formanowicz P.: Asemblacja długich łańcuchów peptydowych przy wykorzystaniu metaheurystyki GRASP, Zeszyty Naukowe Politechniki Śląskiej z. 150, 2008, p. 203–209.
18. Błażewicz J., Borowski M., Formanowicz P., Stobiecki M.: Tabu Search Method for Determining Sequences of Amino Acids in Long Polypeptides, Lecture Notes in Computer Science 3449 (2005) p. 22–32.