Tomasz GŁOWACKI, Adam KOZAK, Łukasz REK, Piotr FORMANOWICZ
Politechnika Poznańska

METODY ASEMBLACJI DŁUGICH ŁAŃCUCHÓW PEPTYDOWYCH I ICH ZŁOŻONOŚĆ OBLICZENIOWA

Streszczenie. Długie peptydy nazywane są białkami i pełnią wiele funkcji w ludzkim organizmie m.in są katalizatorami, transportują inne substancje, chronią przed antygenami. Peptydy zbudowane są z 20 typów aminokwasów połączonych w długie nierozgałęzione łańcuchy (sekwencje). Znajomość sekwencji to pierwszy krok do poznania funkcji białka. Znane metody pozwalają na określenie kolejności aminokwasów (sekwencjonowanie) jedynie krótkich łańcuchów. Długie peptydy są częściowo trawione do wielu krótkich sekwencji, które są następnie sekwencjonowane. W dalszej kolejności wykorzystuje się metody asemblacyjne do rekonstrukcji badanego białka. W pracy przedstawiono dwa problemy asemblacji. Zaproponowany został nowy model grafowy jako reprezentacja jednego z tych problemów.

PEPTIDES ASSEMBLY METHODS AND THEIR COMPUTATIONAL COMPLEXITY

Summary. Peptides are important chemical compounds. Long peptides are called proteins. They have many important functions in human body e.g. catalyzing reactions, transporting other substances, defending body against antigens. They are built of 20 types of amino acid arranged into long chain. Determining an order of aminoacids is an important step in determination of proteins 3D structure. Known methods allow to determine only short sequences. Long sequence is digested into short pieces. Each short piece is sequenced and the assembly method is used to discover the target sequence. In this paper two assembly problems are presented and the new graph model for one of the problems is also proposed.

1. Wprowadzenie

Peptydy to związki chemiczne składające się z aminokwasów tworzących długie nierozgałęzione łańcuchy [9]. Każde dwa sąsiednie aminokwasy połączone są specjalnym wiązaniem nazywanym wiązaniem peptydowym. W skład peptydów wchodzi 20 typów aminokwasów. Długie peptydy o masie przekraczającej 10000 Daltonów nazywane są białkami. Określenie kolejności aminokwasów w cząsteczce nosi nazwę sekwencjonowania i stanowi ważny krok na drodze do poznania przestrzennej budowy białka i jego funkcji. Można wyróżnić następujące poziomy opisu białek [2]:

- struktura pierwszorzędowa - kolejność aminokwasów
- struktura drugorzędowa - przestrzenne ułożenie łańcuchów aminokwasów

- struktura trzeciorzędowa - wzajemne położenie elementów struktury drugorzędowej
- struktura czwartorzędowa - określa sposób agregacji cząsteczek i ewentualnych struktur niebiałkowych

Na potrzeby dalszych rozważań należy wspomnieć, że łańcuchy peptydowe nie są symetryczne, z jednej strony zakończone są grupą $-COOH$ (C-koniec), a z drugiej grupą $-NH_2$ (N-koniec). Jedną z najbardziej popularnych metod sekwencjonowania jest degradacja Edmana. Jest to reakcja chemiczna, podczas której związek chemiczny PTH reaguje z N-końcowym aminokwasem. Wynikiem tej reakcji jest łańcuch skrócony o aminokwas N-końcowy oraz związek powstały przez połączenie PTH i odciętego aminokwasu. Do ustalenia aminokwasu wykorzystuje się chromatografię lub elektroforezę. Chromatografia to technika analityczna, w której wykonuje się detekcję składników bazując na wielkości cząsteczek albo ich zdolności do adsorpcji. Elektroforeza to rozpoznanie cząsteczki w oparciu o jej ładunek elektryczny: przebyta przez substancję droga w polu elektrycznym pozwala na identyfikację. Wykonanie degradacji Edmana umożliwia poznanie jednego aminokwasu, aby poznać kolejne wykonuje się degradację cyklicznie. Po każdej iteracji pochodna PTH jest poddawana identyfikacji. Po każdej iteracji metody w roztworze pozostają zanieczyszczenia - pochodne z poprzednich iteracji oraz łańcuchy w których nie nastąpiło odcięcie aminokwasu N-końcowego. W praktyce uniemożliwiają one sekwencjonowanie łańcuchów przekraczających 50 aminokwasów [9]. Inną metodą sekwencjonowania jest spektrometria masowa [1]. Obojętne cząsteczki białka ulegają w spektrometrze masowym jonizacji i fragmentacji tworząc tak zwane jony fragmentacyjne. Wynikiem eksperymentu przeprowadzonego za pomocą spektrometru masowego jest tzw. widmo masowe - wykres dwuwymiarowy, który przedstawia na osi X stosunek masy do ładunku (m/z) każdego z uzyskanych jonów fragmentacyjnych oraz ich względne stężenie w mieszaninie na osi Y. Widmo masowe można traktować jako "odcisk palca" związku chemicznego i jego analiza pozwala na rozpoznanie tej substancji. W praktyce analiza widma jest złożonym procesem wymagającym często zaangażowania skomplikowanych metod kombinatorycznych i może być przeprowadzona z powodzeniem tylko dla krótkich peptydów. Obie wspomniane metody posiadają ograniczenie co do długości sekwencjonowanego łańcucha. Rzeczywiste białka często przekraczają kilkaset aminokwasów, stąd metody te są niewystarczające. Naturalnym rozwiązaniem jest podzielenie analizowanego białka na wiele krótkich łańcuchów, ustalenie ich sekwencji a następnie asemblacja tych fragmentów w celu rekonstrukcji oryginalnego łańcucha. W niniejszej pracy zestawiono znane problemy asemblacji oraz kilka metod ich rozwiązania. Proponuje się również nowy model grafowy dla jednego z przedstawionych problemów.

2. Eksperyment chemiczny

W celu przygotowania danych wejściowych dla metody asemblacyjnej, wykorzystuje się endopeptydazy [9]. Są to enzymy, które tną białko wewnątrz cząsteczki, po rozpoznaniu odpowiedniego aminokwasu. Przykładowe endopeptydazy to:

- trypsyna, która tnie łańcuch aminokwasów po argininie oraz lizynie,
- chymotrypsyna, która rozpoznaje natomiast tryptofan, fenyloanilinę i tyrozynę.

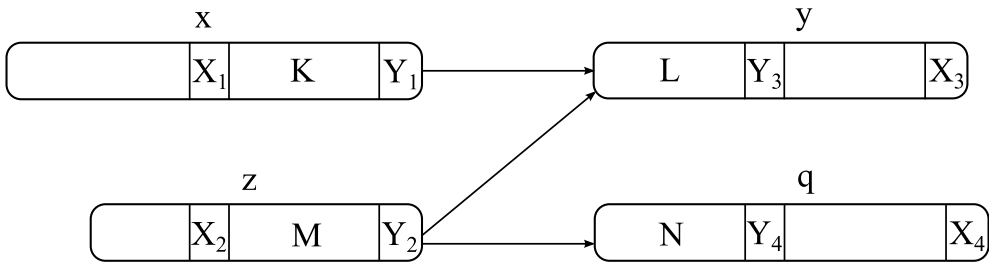
W eksperymencie dzieli się roztwór zawierający badane białka na dwa naczynia i przeprowadza w każdym z naczyń trawienie (cięcie) inną endopeptydazą. Wynikiem eksperymentu są dwa zestawy krótkich łańcuchów. Eksperyment może być źródłem błędów. Błędem pozytywnym określa się sytuację, gdy w zbiorze krótkich sekwencji pojawia się łańcuch, który nie jest rzeczywistym wynikiem eksperymentu chemicznego. Może być to na przykład zanieczyszczenie zewnętrzne albo fragment endopeptydazy, która również jest peptydem. Błędem negatywnym jest sytuacja, gdy nie zaobserwowano sekwencji, które powinny się pojawić. Taka sytuacja zdarza się, gdy łańcuch zostanie zagubiony. Błędy negatywne mogą również wynikać z braku informacji o powtórzeniach krótkich łańcuchów (gdy taka informacja jest gubiona w procesie analizy próbki). Błędem jest również sytuacja, gdy nie zachodzą wszystkie cięcia, których można się spodziewać biorąc pod uwagę mechanizm działania danej endopeptydazy. Należy zauważyć, że klasyfikacja tego błędu nie jest oczywista. W wyniku braku cięcia w spektrum pojawia się dodatkowy łańcuch, który pochodzi jednak z asemblowanego łańcucha, nie jest więc błędem pozytywnym. Trzeba pamiętać, że w roztworze znajduje się wiele takich samych cząsteczek białka. Brak cięcia w jednej cząsteczce, a co za tym idzie brak pewnego krótkiego fragmentu pochodzącego z trawienia tej cząsteczki, nie oznacza, że w innej cząsteczce to cięcie nie zaszło, czyli krótki fragment pomimo wszystko może pojawić się w spektrum. Brak cięcia nie gwarantuje więc zagubienia krótkiego fragmentu, a co za tym idzie nie jest błędem negatywnym.

W eksperymencie możliwy jest do uzyskania rozkład wszystkich aminokwasów w cząsteczce. W tym celu przeprowadza się całkowite trawienie cząsteczki do aminokwasów i mierzy ich stężenie w roztworze.

3. Problemy kombinatoryczne

W idealnym przypadku wyniki eksperymentu nie zawierają błędów pozytywnych ani negatywnych, a w procesie chemicznym zaszły wszystkie możliwe cięcia. Problem asemblacji w takim przypadku może być reprezentowany przez szczególny model grafowy [3]. Wszystkie wierzchołki grafu odpowiadają krótkim sekwencjom pochodzącym z trawienia obiema endopeptydazami i są zaetykietowane przez odpowiadające tym sekwencjom łańcuchy znaków nad 20 literowym alfabetem Σ. W grafie tym istnieje łuk pomiędzy dwoma wybranymi wierzchołkami, gdy zachodzi niezerowe nałożenie końcowych liter etykiety poprzednika z początkowymi literami etykiety następnika. Można zauważyć, że tego rodzaju graf jest grafem dwudzielnym, gdyż takie nałożenie może istnieć tylko między łańcuchami pochodzącymi z trawienia różnymi endopeptydazami. W przypadku, gdy zakładamy wystąpienie wszystkich cięć, można zauważyć również, że jedynie najdłuższe możliwe nałożenie jest poprawne, wiec taki graf jest 1-grafem. Rozwiązanie problemu asemblacji stanowi ścieżka Hamiltona w tym grafie, która odpowiada trawionemu łańcuchowi.

Graf uzyskany z takiego eksperymentu w przypadku idealnym nazywany jest *grafem peptydowym*. Można pokazać, że graf peptydowy jest grafem sprzężonym. Graf sprzężony można traktować jako pewną transformację innego grafu (grafu oryginalnego), która polega na zamianie zbioru łuków w zbiór wierzchołków i poprowadzenie łuków między takimi wierzchołkami, dla których w grafie oryginalnym istnieje możliwość bezpośredniego przejścia przez wspólny wierzchołek:



Rys. 1. Para wierzchołków x, z w grafie peptydowym, dla której zbiory następników nie są rozłączne tj. $N^+(x) \cap N^+(z) = \{y\}$

Definicja 1. [10] 1-graf skierowany $H = (A, U)$ jest grafem sprzężonym grafu $G = (V, A)$ o zbiorze wierzchołków A i takim zbiorze łuków U, że między wierzchołkami $x, y \in A$ w grafie H występuje łuk (x, y) wtedy i tylko wtedy, gdy wierzchołek końcowy łuku x w grafie G jest wierzchołkiem początkowym łuku y w grafie G.

Poniższe twierdzenie przedstawia własność, która pozwala na rozpoznanie grafu sprzężonego.

Twierdzenie 1. *[10] Niech $H = (V, A)$ będzie 1-grafem. H jest grafem sprzężonym wtedy i tylko wtedy, gdy:*

$$\forall_{x,y \in V} N^+(x) \cap N^+(y) \neq \emptyset \Rightarrow N^+(x) = N^+(y)$$

Twierdzenie 2. *Graf peptydowy jest grafem sprzężonym.*

Dowód. Aby udowodnić, że graf peptydowy jest grafem sprzężonym wystarczy udowodnić własność Twierdzenia 1, że dla dowolnych dwóch wierzchołków zbiory ich następników są rozłączne lub równe. W tym celu wystarczy rozważyć parę wierzchołków w takim grafie, dla której zbiory następników nie są rozłączne i wykazać, że w takim przypadku zbiory te muszą być równe.

Sytuacja ta jest przedstawiona na Rysunku 1. Wierzchołki x, z należą do grafu peptydowego $G = (V, A)$. Zgodnie z ilustracją $x, y, z, q \in V$ oraz $\{(x, y), (z, y), (z, q)\} \in A$, zbiory następników x i z nie są rozłączne tj. $N^+(x) \cap N^+(z) = \{y\}$. Symbole $X_1, \ldots, X_4$ reprezentują aminokwasy po których tnie pierwsza endopeptydaza, symbole $Y_1, \ldots, Y_4$ reprezentują aminokwasy po których tnie druga endopeptydaza, natomiast K, L, M, N są podłańcuchami dzielącymi wystąpienia tych aminokwasów. Można zauważyć, że istniejące łuki implikują równość $Y_1 = Y_2 = Y_3 = Y_4$ oraz zachodzą implikacje $(x, y) \in A \Rightarrow K = L$, $(z, y) \in A \Rightarrow M = L$ oraz $(z, q) \in A \Rightarrow M = N$. Implikacje te wynikają z faktu, że w oryginalnym łańcuchu przed fragmentami L i N musiał wystąpić jeden z aminokwasów, po którym tnie pierwsza endopeptydaza, więc prawidłowe jest tylko nałożenie z pełnymi fragmentami K/M. Wynika stąd, że $K = N$ co implikuje istnienie łuku (x, q) i tym samym równość zbiorów następników $N^+(x) = N^+(z)$. Rozumowanie prowadzi do własności opisanej w Twierdzeniu 1 charakteryzującej grafy sprzężone. □

Grafy sprzężone mają własność, że cykl/ścieżka Hamiltona może być znaleziona za pomocą obwodu/ścieżki Eulera w grafie oryginalnym [10]. Wynika z tego, że dla

przypadku idealnego problem asemblacji może być rozwiązany w czasie wielomianowym.

W przypadku braku niektórych cięć, gdy dodatkowo dany jest rozkład aminokwasów, sytuacja komplikuje się. Należy rozważyć wszystkie możliwe nałożenia dwóch etykiet, a nie tylko najdłuższe. Powstający w takim, problem asemblacji może zostać sformułowany następująco:

Instancja: Multizbiór S ciągów znaków nad alfabetem Σ (Σ-zbiór wszystkich symboli reprezentujących poszczególne aminokwasy), rozkład D symboli z alfabetu Σ tj. zbiór par (x, i) dla wszystkich symboli x z alfabetu Σ, gdzie i jest liczbą całkowitą dodatnią.

Odpowiedź: Superciąg dla multizbioru ciągów S spełniający rozkład D.

Zostało udowodnione, że problem ten jest silnie NP-trudny [8], co oznacza, że nie istnieją wielomianowe algorytmy rozwiązujące go, jeśli $P \neq NP$. W literaturze przebadano możliwość wykorzystania różnych metaheurystyk do rozwiązania tego problemu. Zaimplementowano i przetestowano algorytm ewolucyjny [4], metodę GRASP [5] oraz metodę Tabu [6].

Algorytmy przedstawione w tych trzech pracach zakładały wstępny preprocessing i zawężenie przestrzeni rozwiązań na której działają metaheurystyki. Jakość uzyskanych przez te metaheurystyki wyników zależała więc od sposobu wykonania preprocessingu. Poniżej zaproponowano model grafowy, który reprezentuje pełną przestrzeń rozwiązań dopuszczalnych. Dany jest graf $H = (V, A)$. Każdy wierzchołek $v \in V$ odpowiada pewnej krótkiej sekwencji, dodatkowo wierzchołki te są etykietowane przez odpowiadające tym sekwencjom łańcuchy znaków nad 20 literowym alfabetem Σ. Ponadto dla każdego wierzchołka $v_i \in V$ wprowadza się 20-wymiarowy wektor $\vec{w}(v_i)$. Kolejne wymiary tego wektora odpowiadają alfabetycznie kolejnym literom alfabetu Σ. Wartości w kolejnych wymiarach tego wektora to liczności odpowiednich liter w etykiecie. W grafie tym istnieje łuk pomiędzy dwoma wierzchołkami, gdy istnieje nałożenie się pomiędzy opisującymi je etykietami. W omawianym przypadku może istnieć wiele nałożeń dwóch etykiet, odpowiada to wielu łukom pomiędzy dwoma wierzchołkami. H jest więc multigrafem. Każdemu łukowi $a_i \in A$ przyporządkowuje się wagę $\vec{w}(a_i)$ będącą 20-wymiarowym wektorem, analogicznym jak w przypadku wierzchołków. Wartości tego wektora w kolejnych wymiarach dotyczą liczności odpowiednich liter we fragmencie etykiety, który determinuje istnienie łuku.

W analizowanej wersji problemu może istnieć wiele krótkich sekwencji pokrywających ten sam fragment asemblowanej cząsteczki. Jeśli istnieje wiele wzajemnie nakładających się krótkich sekwencji, to należy wyróżnić dwie najdłuższe. Pozostałe krótsze fragmenty będą się zawierać w tych najdłuższych i powinny zostać pominięte. Do odtworzenia kolejności krótkich sekwencji wystarczy, aby każda pozycja z oryginalnej cząsteczki była odzwierciedlona w wynikach eksperymentu co najwyżej dwa razy. W celu oznaczenia wierzchołków, które potencjalnie mogą zostać usunięte z rozważań, przydziela się im kolory:

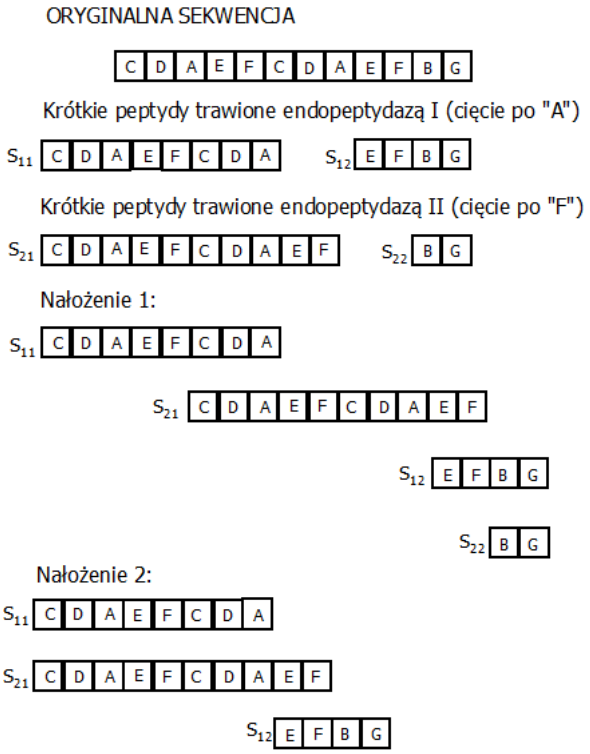
- każdemu wierzchołkowi, którego etykieta zawiera się w etykiecie innego wierzchołka i oba wierzchołki opisują sekwencje z doświadczeń z różnymi endopeptydazami, należy przydzielić czarny kolor
- pozostałym wierzchołkom przydziela się kolor biały.

Przykładową sekwencję dla problemu asemblacji przedstawiono na rysunku 2, a graf dla tej sekwencji zamieszczono na rysunku 3.

Dla każdego wierzchołka w kolorze czarnym należy rozważyć dwie sytuacje: gdy odpowiadający wierzchołkowi krótki peptyd jest rzeczywiście podsekwencją innego krótkiego peptydu lub gdy odpowiada innemu fragmentowi oryginalnej cząsteczki. Dla każdego czarnego wierzchołka należy więc podjąć decyzję, czy uwzględnić go w rozwiązaniu, czy pominąć. Jeśli graf H posiada k czarnych wierzchołków, to istnieje 2^k różnych wyników procesu decyzyjnego. Jeśli wszystkie k decyzji zostanie podjętych poprawnie, czyli pominięte zostaną tylko wierzchołki opisujące sekwencje, które rzeczywiście zawierają się w innych sekwencjach, to krótkie sekwencje odpowiadające pozostałym wierzchołkom pokrywają asemblowany łańcuch w każdym miejscu co najwyżej dwukrotnie. Rozwiązaniem problemu asemblacji jest znalezienie ścieżki przechodzącej przez wszystkie białe wierzchołki dokładnie raz i przez czarne wierzchołki co najwyżej raz. Dodatkowo sekwencja związana z tą ścieżką powinna posiadać uzyskany rozkład aminokwasów. Do obliczenia rozkładu aminokwasów (wektor $\vec{E}$) w uzyskanej sekwencji może posłużyć poniższy wzór:

$$\vec{E} = \sum_{v \in \alpha} \vec{w}(v) - \sum_{\vec{w}(a_i) \in R} \vec{w}(a_i)$$

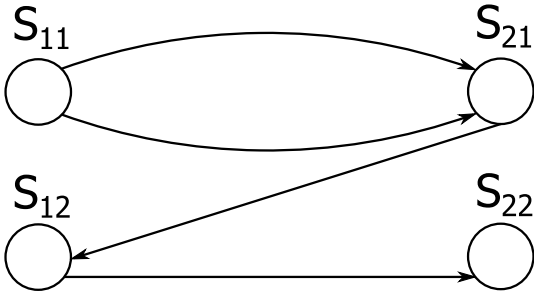
gdzie R to zbiór łuków należących do znalezionej ścieżki, $\alpha \subseteq V$ to zbiór wszystkich wierzchołków należących do tej ścieżki.



Rys. 2. Przykładowa sekwencja i jej asemblacja dla problemu bez niektórych cięć

Opis grafu dla przykładu:

[A,B,C,D,E,F,G,...]



Rys. 3. Graf H dla przykładowej sekwencji

v_1 - czarny, $\overrightarrow{w}(v_1) = [2,0,2,2,1,1,0,...]$
etykieta = "CDAEFCDA"

v_2 - biały, $\overrightarrow{w}(v_2) = [2,0,2,2,2,2,0,...]$
etykieta = "CDAEFCDAEF"

v_3 - biały, $\overrightarrow{w}(v_3) = [0,1,0,0,1,1,1,...]$
etykieta = "EFBG"

v_4 - czarny, $\overrightarrow{w}(v_4) = [0,1,0,0,0,0,1,...]$
etykieta = "BG"

$\overrightarrow{w}(a_1) = [2,0,2,2,1,1,0,...]$
$\overrightarrow{w}(a_2) = [1,0,1,1,0,0,0,...]$
$\overrightarrow{w}(a_3) = [0,1,0,0,1,1,0,...]$
$\overrightarrow{w}(a_4) = [0,1,0,0,0,0,1,...]$

4. Podsumowanie

W pracy przedstawiono różne metody asemblacji peptydów. Zaprezentowano model grafowy dla przypadku, gdy nie ma błędów i zachodzą wszystkie spodziewane cięcia. Udowodniono, że przedstawiony graf jest grafem sprzężonym. Rozwiązaniem tego problemu asemblacji jest ścieżka Hamiltona. Szukanie ścieżki Hamiltona w grafie sprzężonym można sprowadzić do szukania drogi Eulera w innym grafie. Znane są wielomianowe algorytmy znajdujące drogę Eulera, przedstawiony problem jest więc obliczeniowo łatwy.

W przypadku, gdy nie ma wszystkich cięć problem jest obliczeniowo trudny. W pracy przytoczono definicję tego problemu oraz zaproponowano model grafowy do jego zaprezentowania. Jest to pierwszy model matematyczny w literaturze, który umożliwia znalezienie dowolnego rozwiązania dopuszczalnego, dotychczas głównie bazowano na dość uproszczonych i przybliżonych modelach wykorzystywanych przez metaheurystyki. Przedstawiony model może zostać z powodzeniem wykorzystany do projektowania rozwiązań dokładnych. Rozwiązaniem problemu jest ścieżka Hamiltona w jednym z podgrafów indukowanych zaproponowanego grafu. Dodatkowo sekwencja związana z tą ścieżką musi spełniać założenie o rozkładzie aminokwasów.

LITERATURA

1. Johnstone R. A. W. Mass spectrometry for organic chemists. Cambridge University Press, 1972.
2. Kraj A., Silberring J. Proteomika. EJB, Kraków, 2004.
3. 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.
4. Błażewicz J., Borowski M., Formanowicz P, Głowacki T.: Genetic and tabu search algorithms for peptide assembly problem, RAIRO - Operations Research, 44 (2010) p. 153–166
5. Głowacki T., Kozak A., Formanowicz P.: Asemblacja długich łańcuchów peptydowych przy wykorzystaniu metaheurystyki GRASP, Zeszyty Naukowe Politechniki Śląskiej z. 150, 2008, p. 203–209.
6. 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.
7. Formanowicz P.: Selected Combinatorial Aspects of Biological Sequence Analysis, Poznań, Publishing House of Poznań University of Technology 2005.
8. Gallant J. K.: The complexity of the overlap method for sequencing biopolymers, Journal of Theoretical Biology 101 (1983) p. 1–17.
9. Stryer L., Biochemistry, 4th edition, New York, W.H. Freeman and Company, 1995.
10. Błażewicz J., Hertz A., Kobler D., de Werra D.: On some properties of DNA graphs. Discrete Applied Mathematics 98, 1999, p. 1–19.