ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2008
Seria: AUTOMATYKA z. 151 Nr kol. 1797

Tomasz GŁOWACKI, Adam KOZAK, Piotr FORMANOWICZ
Politechnika Poznańska

ASEMBLACJA DŁUGICH ŁAŃCUCHÓW PEPTYDOWYCH PRZY WYKORZYSTANIU METAHEURYSTYKI GRASP

Streszczenie. Ustalenie kolejności aminokwasów w cząsteczce białka nosi nazwę sekwencjonowania. Brak bezpośrednich metod sekwencjonowania długich peptydów powoduje, że potrzebne są dedykowane metody asemblacyjne, które odpowiednio poskładają krótkie łańcuchy w jeden długi łańcuch aminokwasów. W pracy tej został zaproponowany algorytm asemblacyjny typu GRASP. Przedstawiony algorytm został zaimplementowany i przetestowany dla zbioru rzeczywistych peptydów, a uzyskane rozwiązanie zostało przedyskutowane.

ASSEMBLING LONG PEPTIDES USING GRASP METAHEURISTIC

Summary. Determining an order of amino acids in peptide structure is called sequencing method. Lack of direct sequencing methods for long peptides causes that assembling methods to combine many short peptides into one long structure are necessary. In this paper assembling algorithm based on GRASP method was proposed. The algorithm was implemented and tested on real peptides set and the obtained results was discussed.

1. Wprowadzenie

Nowoczesne, rozwijające się nauki przyrodnicze generują ogromną ilość informacji do przetworzenia. Zaawansowane algorytmy pozwalają na nowe spojrzenie na dane dostarczone z chemicznych i biologicznych doświadczeń. Jednym z najbardziej spektakularnych osiągnięć bioinformatyki jest zsekwencjonowanie genomu ludzkiego, czyli odczytanie za pomocą dedykowanych algorytmów sekwencji zasad w cząsteczce DNA człowieka. Znajomość sekwencji genomu ludzkiego otwiera możliwości analizy informacji w nim zapisanej, co prowadzi do wielu ważnych i interesujących problemów biologii molekularnej oraz obliczeniowej.

Innymi ważnymi dla żywych organizmów związkami są białka, nazywane także polipeptydami. Białka są wielocząsteczkowymi związkami składającymi się z 20 rodzajów aminokwasów, połączonych w łańcuch specjalnymi wiązaniami między kolejnymi aminokwasami, nazywanymi wiązaniami peptydowymi. Kolejność aminokwasów w cząsteczce nosi nazwę struktury pierwszorzędowej. Białka pełnią w organizmie funkcje budulcowe, a także katalizują wiele biochemicznych reakcji. Funkcja białka jest zależna od jego przestrzennej budowy. Określenie przestrzennej budowy białka, nazywanej

strukturą trzeciorzędową, jest jednym z największych wyzwań współczesnej biologii obliczeniowej. Struktura ta silnie zależy od jego struktury pierwszorzędowej. W poniższej pracy zaproponowano algorytm oparty na metodzie GRASP służący do ustalenia budowy pierwszorzędowej struktury białka. W rozdziale 2 omówiono chemiczne aspekty asemblacji łańcuchów peptydowych. W rozdziale 3 przedstawiono sformułowanie problemu asemblacji jako problemu kombinatorycznego. Rozdział 4 zawiera opis proponowanego algorytmu, a rozdział 5 wyniki eksperymentu obliczeniowego. Pracę kończą wnioski zamieszczone w rozdziale 6.

2. Chemiczne aspekty asemblacji

Brak jednoznacznych, chemicznych metod służących do określenia pierwszorzędowej budowy cząsteczki peptydowej czyni tę dziedzinę niezwykle atrakcyjną dla informatyków. Odpowiednie połączenie mechanizmu chemicznego z aparatem matematycznym i algorytmicznym pozwala na osiągnięcie interesujących, także dla chemików, wyników. Metody chemiczne pozwalają jedynie na określenie sekwencji krótkich łańcuchów peptydowych.

Wykorzystując metodę Edmana lub spektrometrię masową, można ustalić sekwencję jedynie krótkich łańcuchów o długości do 50 aminokwasów [6]. Białka posiadają jednak łańcuchy o długości nawet do 10000 aminokwasów. Do ustalenia pierwszorzędowej struktury białek stosuje się metody asemblacyjne. Asemblacja jest to składanie krótkich łańcuchów peptydowych w jeden długi łańcuch. Asemblacja umożliwia więc rozpoznanie budowy długich sekwencji przez składanie krótszych, odczytanych za pomocą metody Edmana lub spektrometru, w jedną całość.

Łatwo zauważyć, że znajomość sekwencji wielu krótkich (do 50 aminokwasów) peptydów to za mało, aby zrekonstruować budowę analizowanego białka. Aby umożliwić zrekonstruowanie łańcucha białkowego, stosuje się wiele specjalistycznych zabiegów utrzymania kontekstu (informacji) o kolejności zsekwencjonowanych łańcuchów w szukanym białku. Na potrzeby poniższych badań zaproponowano wykorzystanie endopeptydaz. Są to enzymy z grupy proteaz, które katalizują rozkład cząsteczki białka na kilka krótszych łańcuchów białkowych. Na potrzeby opisywanego doświadczenia-symulacji wykorzystano chymotrypsynę i trypsynę. Enzymy te działają kontekstowo. Poddawane ich działaniu białko zawsze ulega podziałowi w tych samych miejscach - trypsyna katalizuje rozkład wiązań, w których grupy karbonylowe należą do lizyny albo argininy [6]. Chymotrypsyna rozkłada białko w miejscach, gdzie grupy karbonylowe wiązania należą do tryptofanu, fenyloaniliny lub tyrozyny. Upraszczając: trypsyna tnie białko na wiązaniu peptydowym po wystąpieniu lizyny lub argininy, natomiast chymotrypsyna tnie białko na pierwszym wiązaniu peptydowym po wystąpieniu tryptofanu, fenyloaniliny lub tyrozyny. Dodatkowo taki dobór enzymów powoduje, że powstałe łańcuchy prawie nigdy nie przekraczają 50 aminokwasów, mogą więc być zsekwencjonowane za pomocą metody Edmana. W przeprowadzonym doświadczeniu materiał białkowy jest rozdzielony do dwóch naczyń. Następnie w jednym z nich przeprowadza się trawienie enzymatyczne, wykorzystując trypsynę, a w drugim chymotrypsynę. Powstają dwa zbiory krótkich peptydów, które są sekwencjonowane. Dzięki zjawisku elektroforezy materiał białkowy zostaje podzielony na możliwie jednakowe frakcje, co ułatwia pobranie materiału do sekwencjonowania metodą Edmana. Dzięki podziałowi materiału

biologicznego na dwie części i wyborze enzymów tnących cząsteczkę białka w różnych miejscach udaje się zachować kontekst. Ciąg aminokwasów, którym kończy się pewien krótki peptyd, jest taki sam jak ciąg aminokwasów, którym rozpoczyna się kolejny krótki peptyd w rekonstruowanej cząsteczce [3].

3. Definicja problemu

W pracy Gallanta [1] pokazano, że wersja omawianego problemu asemblacji o znanym rozkładzie aminokwasów i bez wszystkich cięć wynikających z działania na cząsteczkę endopeptydazą jest NP-trudna.

Z punktu widzenia teorii grafów wersja asemblacji bez wszystkich cięć indukuje multigraf (por. [3, 4]). Każdy krótki peptyd, wynik doświadczenia z endopeptydazą jest zaprezentowany jako wierzchołek w tym grafie. Do zaetykietowania wierzchołków wykorzystano 20-literowy alfabet, gdzie każda litera alfabetu odpowiada pewnemu aminokwasowi. Wierzchołki grafu zostały zaetykietowane ciągami znaków, które odpowiadają korespondującym z nimi krótkim peptydom. Istnienie łuku pomiędzy dwoma dowolnymi wierzchołkami determinuje nakładanie się dwóch związanych z nimi peptydów. Łuk między dwoma wierzchołkami jest zdefiniowany następująco:

$$w : A \rightarrow N, w(v_i, v_j) = \{p : \exists_{k \in \{1,2,\ldots,|s_i|\}} \forall_{q \in \{1,2,\ldots,p\}} s_i(k-1+q) = s_j(q)\}$$

gdzie s_i oznacza etykietę wierzchołka v_i. Można zauważyć, że jedynie nałożenia między peptydami pochodzącymi z różnych wyników doświadczeń z endopeptydazą determinują właściwe nałożenie w szukanej cząsteczce, ponieważ peptydy z tego samego doświadczenia nie nakładają się, są rozłącznymi, poprzecinanymi łańcuchami. Obserwacja ta prowadzi do usunięcia łuków między wierzchołkami, które związane są z peptydami trawionymi tą samą peptydazą, co w ostateczności prowadzi do utworzenia dwudzielnego grafu, gdzie każdy zbiór wierzchołków związany jest z peptydami powstałymi w trawieniu inną peptydazą. Dodatkowo z takiego grafu można usunąć wierzchołki, których etykiety zawierają się w etykietach wierzchołków z drugiego zbioru. Dzięki temu zabiegowi różnica wierzchołków w poszczególnych zbiorach wynosi:

$$|V_i - V_j| = 1$$

Dodatkowo do grafu dodano łuki o wadze 0, między wszystkimi możliwymi parami wierzchołków, zachowując własność grafu dwudzielnego. W tak zbudowanym grafie znalezienie dowolnej ścieżki przechodzącej przez wszystkie wierzchołki jest łatwe, co pozwoli na wielomianowe znalezienie rozwiązania początkowego. Rozwiązaniem problemu asemblacji w takim grafie jest znalezienie ścieżki przechodzącej przez wszystkie wierzchołki, dla której wartość poniższej funkcji celu wynosi 0:

$$f = \sum_{i=1}^{20} |Z_i - O_i|$$

gdzie: O_i i Z_i to kolejno liczby aminokwasów typu i w rozwiązaniu optymalnym i rozwiązaniu znalezionym, co oznacza, że szukany peptyd zawiera wszystkie krótkie łańcuchy aminokwasowe, a jego rozkład jest równy zadanemu.

4. Algorytm

Dla zdefiniowanego problemu zaproponowano algorytm GRASP (akronim Greedy Randomized Adaptive Search Procedure), który jest metaheurystyką bazującą na znajdowaniu dobrego rozwiązania początkowego [5]. Główną ideą tej metody jest stworzenie dobrego rozwiązania początkowego, a następnie jego lokalna optymalizacja. Do budowy rozwiązania początkowego używa się w GRASP specjalnej listy RLC (Restricted Candidate List - ograniczona lista kandydatów). RLC nie zawiera wszystkich elementów, których dołożenie w danym kroku do rozwiązania częściowego jest możliwe, lecz jedynie elementy, których dodanie do rozwiązania powoduje największy przyrost wartości funkcji oceny heurystycznej rozwiązania f. Funkcją oceny heurystycznej jest funkcja obliczająca długość bieżącej ścieżki od wierzchołka początkowego do wierzchołka końcowego. Na każdym etapie budowania rozwiązania lista RLC jest uaktualniana, w zależności od istniejącego cząstkowego rozwiązania początkowego. Następnie z listy RLC jest losowo wybierany jeden element i dodawany do częściowego rozwiązania początkowego.

Dla dowolnego wierzchołka v_i, który może być dodany do częściowego rozwiązania, na każdym kroku algorytmu sprawdzane są wszystkie możliwe sposoby dołączenia tego wierzchołka na początku lub na końcu istniejącego rozwiązania. Eksperymentalnie przyjęto, że lista RLC zawiera rozwiązania nie gorsze o więcej niż 35% od rozwiązania optymalnego w danym kroku algorytmu, jednak nie więcej niż 30% wszystkich rozwiązań.

Uzyskane rozwiązanie początkowe jest następnie optymalizowane. Jako sąsiedztwo danego rozwiązania X zdefiniowano wszystkie rozwiązania, które mogą powstać po zamianie miejscami dwóch wierzchołków w rozwiązaniu X oraz wybraniu dowolnego nałożenia na sąsiadów (dowolna wartość łuku w multigrafie). Przeszukując przestrzeń rozwiązań, tworzy się listę wszystkich możliwych rozwiązań polepszających funkcję celu g i wybiera losowo jedno z nich. Maksymalizowana funkcja celu g jest zdefiniowana jako odległość w metryce taksówkowej otrzymanego rozwiązania od rozwiązania optymalnego, zdefiniowanego jako 20-wymiarowy wektor określający liczbę każdego aminokwasu w rozwiązaniu optymalnym:

$$g = \frac{1}{\sum_{i=1}^{20} |Z_i - O_i|}$$

Prawdopodobieństwo wyboru danego rozwiązania jest wprost proporcjonalne do wartości polepszenia funkcji celu:

$$P(X_j) = \frac{g(X_j) - g(X_i)}{\sum_{\forall X_k\, g(X_k) > g(X_i)} (g(X_k) - g(X_i))}$$

5. Wyniki eksperymentu

Przedstawiony algorytm został zaimplementowany w języku Java 1.5 i przetestowany na komputerze klasy PC z procesorem Intel 2xXenon 3.6 GHz z 4 GB RAM. Sekwencje peptydowe wykorzystane w doświadczeniu zostały pobrane ze strony http://www.clcbio.com/. Przygotowano 15 podzbiorów peptydów, które zróżnicowano

Tabela 1

Wyniki eksperymentu obliczeniowego

GRASP			
długość sekwencji	błędy	dopasowanie (%)	czas (s)
100	1	82,17	0,87
	2	87,56	0,921
	3	89,17	0,923
150	1	85,94	1,078
	2	90,17	1,007
	3	75,38	1,144
200	1	74,95	1,125
	2	81,02	1,117
	3	72,17	1,103
250	1	63,5136	1,435
	2	63,14	1,489
	3	63,95	1,397
300	1	65,18	1,642
	2	69,29	1,598
	3	60,93	1,572

w zależności od długości asemblowanej sekwencji (100, 150, 200, 250 lub 300 aminokwasów) oraz od liczby błędów wynikających z braku cięć (1, 2 lub 3 błędy).

Dla przygotowanych sekwencji zasymulowano opisane doświadczenie częściowego trawienia peptydów i sekwencjonowania krótkich łańcuchów, przy założeniu że błędy mogą pochodzić jedynie z braku cięć w eksperymencie trawienia przez endopeptydazy. Takie dane stały się danymi wejściowymi do przetestowania skuteczności algorytmu. Skuteczność metody zmierzono przez porównanie podobieństwa uzyskanej sekwencji do oryginalnej sekwencji za pomocą algorytmu Needlemana–Wunscha. Algorytm GRASP został 10-krotnie wykonany dla każdej instancji danych, a jego wyniki czasowe oraz jakościowe zostały dla każdej sekwencji uśrednione. Tabela 1 przedstawia wyniki eksperymentu.

6. Podsumowanie i wnioski

Zaproponowany algorytm został zaimplementowany i przetestowany dla 150 przykładowych instancji prawdziwych peptydów. Średnie dopasowanie uzyskanej cząsteczki do cząsteczki oryginalnej, zmierzone za pomocą algorytmu Needlemana-Wunscha, wynosi 74,97%. Zmierzono także średnie dopasowanie dla tego samego zbioru instancji wejściowych dla algorytmu Tabu zaprezentowanego w pracy [2]; dopasowanie algorytmu Tabu wynosi 61,36%. Wyniki świadczą o dużej preferencji dobrych rozwiązań początkowych dla omawianego problemu.

Preferowanie w grafie łuków o największych wagach wyróżnia ze zbioru peptydów te pary, które posiadają duże nałożenia między sobą. Dla przedstawionej definicji

problemu tylko niezerowe łuki determinują poprawne pary peptydów; dodatkowo duże nałożenia często w praktyce determinują poprawne pary peptydów, gdyż w rzeczywistości rzadko zdarzają się łuki o wysokiej wartości, które nie odzwierciedlają prawdziwego nakładania się tych dwóch peptydów w szukanej cząsteczce.

BIBLIOGRAFIA

1. Gallant J.K.: The complexity of the overlap method for sequencing biopolymers. Journal of Theoretical Biology, 101, 1983, p. 1–17.
2. 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, 2005, 3449, 22–32.
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. Formanowicz P.: Selected combinatorial aspects of biological seqeuence analysis. Wydawnictwo Politechniki Poznańskiej, Poznań 2005.
5. Resende M., Reibeiro C.: Greedy Randomized Adaptive Search Procedures. Handbook of Metaheuristics, Kluwer Academic Publishers, 2003, p. 219-–249.
6. Stryer L.: Biochemistry, 4th edition. W.H. Freeman and Company, New York 1995.

Recenzent: Prof. dr hab. inż. Andrzej Polański

Abstract

Peptide sequencing is a method of determining order of amino acids in peptide structure. Known chemical sequencig methods like Edman's method or mass spectrometry allow to discover only short peptide sequences up to 50 amino acids. Asembling methods give possibility to combine many short peptides into one long structure. This paper describes a method for peptide asembling using two endopeptidazes. Endopeptidaze is chemical molecule which cuts peptide in places where appriopriate amino acid occured. These short peptides may be sequenced by using traditional methods. In 1983 Galant proved that asembling problem for known distribution of aminoacids and without all cuts from endopeptidases is NP-hard. This paper introduces GRASP algorithm for resolving defined problem. Representation of this problem is based on labeled multigraph, where each vertex coresponds to a short peptide chain - result of endopeptidaze reaction. Additionally, the presented graph is a bipartial graph, where each set of vertices coresponds to result of endopeptidaze reaction. A path of a given length that contains all vertices is a solution of the problem. This algorhitm was implemented and tested on a set of real peptides. Results were presented and discussed.