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

Kamil KWARCIAK, Marcin RADOM, Piotr FORMANOWICZ
Politechnika Poznańska

SEKWENCJONOWANIE DNA Z BŁĘDAMI NEGATYWNYMI ORAZ INFORMACJĄ O POWTÓRZENIACH

Streszczenie. W pracy przedstawiony jest algorytm typu podziału i ograniczeń dla problemu sekwencjonowania przez hybrydyzację z błędami negatywnymi oraz częściową informacją o powtórzeniach. Uwzględnienie tego rodzaju informacji możliwe jest ze względu na rozwój technologii stosowanej w eksperymencie hybrydyzacyjnym. Prowadzi ono do rozszerzenia standardowej wersji metody sekwencjonowania przez hybrydyzację oraz do polepszenia jakości uzyskiwanych rozwiązań.

DNA SEQUENCING WITH NEGATIVE ERRORS AND INFORMATION ABOUT REPETITIONS

Summary. In this paper a branch and bound algorithm for sequencing by hybridization problem with negative errors and partial information about repetitions is presented. Taking into account information of this type is possible because of the development of the technology used in the hybridization experiment. It leads to an extension of the standard sequencing by hybridization method and to an improvement of the quality of the obtained results.

1. Wstęp

Jedną z metod sekwencjonowania DNA jest sekwencjonowanie przez hybrydyzację [1, 5]. Metoda ta składa się z dwóch faz: biochemicznej i obliczeniowej. W pierwszej z nich przeprowadzany jest eksperyment hybrydyzacyjny z pełną biblioteką oligonukleotydów o pewnej ustalonej długości l, nazywanych l-merami. W wyniku przeprowadzenia tego eksperymentu otrzymuje się zbiór, nazywany *spektrum*, który zawiera wszystkie podciągi o długości l możliwe do wyróżnienia w analizowanej sekwencji DNA. Tak jest w przypadku idealnym, który jednak nie musi zachodzić w praktyce. W rzeczywistym eksperymencie hybrydyzacyjnym mogą bowiem wystąpić błędy dwojakiego rodzaju: *negatywne* i *pozytywne*. Spowodowane są one niedoskonałością procesu hybrydyzacji. W przypadku błędów negatywnych spektrum nie zawiera pewnych l-merów, mimo że występują one w badanej sekwencji, natomiast błędy pozytywne oznaczają, że spektrum zawiera l-mery, których w badanej sekwencji nie ma. Istnieje jeszcze jeden rodzaj błędów – są to również błędy negatywne, lecz wynikające z powtórzeń pewnych podciągów o długości co najmniej l w analizowanej sekwencji DNA. Występowanie tych błędów

wynika z ograniczeń technologicznych, które sprawiają, że w eksperymencie hybrydyzacyjnym nie można odczytać, ile razy dany fragment o długości l wystąpił w badanej cząsteczce DNA. W fazie obliczeniowej metody sekwencjonowania przez hybrydyzację na podstawie elementów spektrum należy zrekonstruować analizowaną sekwencję.

Rozwój technologii chipów (mikromacierzy) DNA sprawił, że staje się możliwe uzyskanie przynajmniej częściowej informacji o powtórzeniach. Informacja ta w praktyce jest nieprecyzyjna, pozwala jednak polepszyć jakość uzyskiwanych rozwiązań [3, 4]. W niniejszej pracy przedstawiony zostanie algorytm rozwiązujący problem sekwencjonowania przez hybrydyzację z błędami negatywnymi obu wspomnianych powyżej rodzajów oraz informacją o powtórzeniach. Przyjęty zostanie najprostszy model tego rodzaju informacji, zgodnie z którym wiadomo tylko, czy dany l-mer wystąpił raz w analizowanej sekwencji DNA, czy też wystąpił w niej więcej razy. Rozważany w niniejszej pracy problem został opisany w pracach [3, 4], gdzie sformułowano różne warianty problemu sekwencjonowania z informacją o powtórzeniach oraz przeanalizowano ich złożoność obliczeniową, natomiast przedstawiony tu algorytm jest pierwszym algorytmem dla tego rodzaju problemów sekwencjonowania.

Drugi rozdział niniejszej pracy składa się z dwóch części. W pierwszej z nich znajduje się sformułowanie analizowanego problemu, natomiast w drugiej zamieszczono opis zaproponowanego algorytmu typu podziału i ograniczeń. Rozdział trzeci zawiera wyniki eksperymentu obliczeniowego. Pracę kończy podsumowanie w rozdziale czwartym.

2. Algorytm

W rozdziale tym przedstawione zostanie sformułowanie problemu sekwencjonowania przy założeniu występowania błędów negatywnych oraz informacji na temat l-merów (oligonukleotydów, sond chipu DNA), które hybrydyzowały więcej niż z jednym fragmentem analizowanej sekwencji. Następnie opisany zostanie algorytm rozwiązujący tak postawiony problem.

2.1. Sformułowanie problemu

Opisany we wstępie niniejszego artykułu problem kombinatoryczny można formalnie opisać w następujący sposób (por. [3, 4]). Mając dane *spektrum* (będące zbiorem l-merów, które hybrydyzowały do analizowanej sekwencji S oraz długość n sekwencji Q, należy zrekonstruować poszukiwaną sekwencję. W omawianym przypadku spektrum może zawierać tylko błędy negatywne wynikające zarówno z powtórzeń, jak i błędów odczytu chipu DNA (błędy hybrydyzacji). Dodatkowo jednak znana jest informacja o l-merach, do których sekwencja Q hybrydyzowała więcej niż raz (tj. powtórzeniach tychże l-merów w Q). Dla każdego l-meru w idealnym przypadku posiadamy więc informację, czy występuje on więcej razy, czy dokładnie jeden raz w sekwencji Q. Informacja ta może jednak być zniekształcona, co oznacza, że l-mer, z którym związana jest informacja o jego jednokrotnym wystąpieniu w analizowanej sekwencji, wystąpił w niej w rzeczywistości więcej razy (zniekształcenie to spowodowane jest przez błąd hybrydyzacji).

2.2. Opis algorytmu

Powyżej sformułowany problem może być rozważany jako wariant *problemu komiwojażera*. W problemie tym mamy do czynienia z grafem skierowanym lub nieskierowanym, przy czym do każdego łuku (krawędzi) przydzielony jest koszt jego przejścia. Zadaniem algorytmu jest odwiedzenie każdego wierzchołka dokładnie jeden raz oraz powrót do wierzchołka startowego przy równoczesnej minimalizacji kosztów przejścia całego cyklu. Przy modyfikacji powyższego problemu, która znosi ograniczenie pełnego cyklu, tj. wierzchołek początkowy nie jest równy wierzchołkowi końcowemu (czyli poszukiwana jest najkrótsza ścieżka Hamiltona), oraz zakładając, że graf wejściowy jest skierowany, tak postawiony nowy problem stanowi odpowiednik problemu sekwencjonowania DNA. Biologiczny problem sekwencjonowania DNA odpowiada kombinatorycznemu problemowi poszukiwania najkrótszej ścieżki odwiedzającej każdy wierzchołek w grafie skierowanym, w którym wierzchołki odpowiadają elementom spektrum analizowanej sekwencji DNA (por. [2]).

W grafie tym każdy łuk pomiędzy dwoma wierzchołkami posiada koszt odpowiadający nałożeniu się dwóch l-merów. Na przykład, koszt przejścia pomiędzy wierzchołkami CGCTTA oraz GCTTAT wynosi 1, ponieważ nakładają się one dzięki podsekwencji GCTTA o 1 symbol mniejszej niż długość każdego z l-merów. Tenże koszt przejścia między wierzchołkami CGCTTA oraz ATTCCC wynosi 5, jeśli przyjąć możliwe nałożenie przez nukleotyd A.

Rozważany graf jest multigrafem, w którym każda para wierzchołków połączona jest co najmniej dwoma łukami przeciwnie skierowanymi. Łuków tych może być więcej, jeżeli odpowiadające danym wierzchołkom sekwencje nakładają się na siebie na wiele sposobów. Zauważmy, że w przypadku sekwencji, które nie nakładają się na siebie w opisanym powyżej sensie, między odpowiadającymi im wierzchołkami istnieje łuk o koszcie równym długości sekwencji. Dodatkowa infomacja na temat wierzchołków, które muszą zostać odwiedzone więcej niż jeden raz, pozwala na dodatkową weryfikację niejednoznacznego zrekonstruowania sekwencji spełniających podstawowe założenia problemu (odwiedzono wszystkie wierzchołki grafu oraz długość zrekonstruowanej sekwencji wynosi n).

Ogólna zasada działania algorytmu może być opisana w następujący sposób:

1. Zbuduj graf skierowany na podstawie spektrum analizowanej sekwencji DNA.
2. Wykorzystując algorytm typu podziału i ograniczeń dla problemu najkrótszej ścieżki Hamiltona, znajdź każdą taką ścieżkę i dodaj ją do zbioru rozwiązań Z.
3. Dla wszystkich sekwencji ze zbioru Z o długości mniejszej niż n zastosuj rozciąganie sekwencji.
4. Dla wszystkich sekwencji o długości n ze zbioru Z powstałego po wykonaniu punktów 2. i 3. przeprowadź weryfikację rozwiązań, wykorzystując informację o wierzchołkach, dla których wymagane jest ich wielokrotne odwiedzenie.

Zastosowany algorytm przeszukuje wszystkie możliwe rozwiązania; zaczynając od pierwszego znanego wierzchołka (l-meru), przechodzi przez wszystkie wierzchołki grafu. Jeśli zostały odwiedzone wszystkie wierzchołki grafu oraz koszt rekonstrukcji jest mniejszy lub równy n, dana sekwencja jest dodawana do zbioru rozwiązań. Jeśli koszt przejścia do następnego wierzchołka przekroczy ustalony koszt dopuszczalny, dalsze przeszukiwanie aktualną drogą jest niemożliwe. Jeśli aktualny wierzchołek nie

przekracza kosztu maksymalnego, ale pójście nim nie da akceptowalnego rozwiązania, taki wierzchołek nie jest brany pod uwagę w przeszukiwaniu. Weryfikacja następuje na podstawie określania minimalnego kosztu dołączenia nieodwiedzonych wierzchołków do aktualnego rozwiązania. Następnicy danego wierzchołka są ustalani według niemalejącego kosztu. Jeśli dodanie jednego z nich nie daje akceptowalnego rozwiązania z powodu przekroczenia maksymalnego kosztu ścieżki, kolejni następnicy muszą także zostać odrzuceni.

Po pełnym przeszukaniu ścieżek w grafie (przy zastosowaniu opisanych ograniczeń) algorytm dysponuje zbiorem sekwencji o koszcie mniejszym lub równym n, zbudowanych ze wszystkich l-merów ze *spektrum*. Następnie zbiór ten jest poddawany rozciąganiu wszystkich sekwencji o koszcie mniejszym od n. Dla każdego łuku łączącego dwa wierzchołki v_1 oraz v_2 o koszcie równym 1 mogą istnieć także łuki prowadzące z v_1 do v_2 o kosztach od 2 do l. Obliczana jest różnica między długością rozciąganej ścieżki a długością n badanej sekwencji. Następnie przeszukiwana jest cała przestrzeń możliwych rozciągnięć, na bieżąco weryfikując długość rozciąganej sekwencji. W momencie osiągnięcia żądanej długości dalsze rozciąganie sekwencji nie jest kontynuowane, nawet jeśli jest wciąż możliwe.

Ostatnim krokiem algorytmu jest weryfikacja znalezionych rozwiązań w zbiorze Z o długości n. Ponieważ w założeniach omawianego problemu była mowa o l-merach ze *spektrum*, co do których posiadamy wiedzę o konieczności ich wystąpienia w rozwiązaniu więcej niż raz, dzięki tym danym możliwe jest znaczne ograniczenie zbioru rozwiązań dopuszczalnych. Do finalnego zbioru rozwiązań znajdowanego przez algorytm brane są tylko sekwencje, które spełniały warunki odpowiedniej liczby powtórzeń pewnych elementów spektrum.

3. Eksperyment obliczeniowy

Opisany algorytm był testowany na sekwencjach DNA uzyskanych z GenBanku (NCBI). Użyto sekwencji ludzkiego DNA, które podzielono na odpowiednie fragmenty od 107 do 507 nukleotydów.

Dane do eksperymentu zostały przygotowane przez pocięcie sekwencji na fragmenty odpowiedniej długości, po czym zasymulowano eksperyment hybrydyzacyjny czytając oligonukleotydy o długości 8. Długość taka gwarantowała odnalezienie sekwencji o odpowiedniej liczbie błędów negatywnych wynikających z powtórzeń dla ustalonej długości sekwencji. Dodatkowo po otrzymaniu spektrum z powtórzeniami usuwano z niego odpowiedni procent oligonukleotydów celem zasymulowania błędów negatywnych wynikających z błędnego odczytu mikromacierzy w prawdziwych eksperymentach hybrydyzacyjnych.

Do eksperymentów wybrano sekwencje o ustalonej długości oraz o dokładnie określonej liczbie błędów negatywnych wynikających z powtórzeń. W tabelach zamieszczonych poniżej przyjęto następujący sposób opisu kolumn. Zapis 200-10% oznacza sekwencję, w której od idealnego spektrum wynoszącego 200 oligonukleotydów (dla sekwencji o długości 207nt oraz l-merów długości 8) odjęto pewien procent oligonukleotydów jako błędy negatywne wynikające z odczytu, zależny od liczby błędów wynikających z powtórzeń, tak aby sumaryczny procent błędów wynosił 10%. Komórka znajdująca się w kolumnie oznaczonej 2% oraz wierszu 200-10% odpowiada spektrum,

Tabela 1

Średni czas obliczeń dla danej wielkości sekwencji oraz 10% błędów negatywnych wynikających z usunięcia informacji ze spektrum (czas podano w sekundach)

Spektrum				
100-10%	200-10%	300-10%	400-10%	500-10%
0.03	0.28	2.07	20.82	249.67

Tabela 2

Średnia liczba rozwiązań znalezionych przez algorytm dla instancji zawierających 10% błędów negatywnych bez powtórzeń l-merów

Spektrum				
100-10%	200-10%	300-10%	400-10%	500-10%
1.00	1.00	2.40	3.40	9.33

w którym 2% błędów wynikało z powtórzeń oraz 8% z symulacji błędnego odczytu danych z mikromacierzy.

W tabeli 1 przedstawiono czasy obliczeń dla instancji, które zawierały 10% błędów negatywnych.

Wyniki zamieszczone w tej tabeli wyraźnie wskazują, że czas obliczeń jest silnie zależny od długości badanej sekwencji. W powyższym eksperymencie użyto sekwencji o długości od 107 do 507 nukleotydów, aby dla oligonukleotydów o długości l równej 8 w idealnym przypadku spektrum wynosiło od 100 do 500 l-merów. Sekwencje użyte w eksperymencie zostały sprawdzone pod kątem błędów wynikających z powtarzania się l-merów, tak by błędy te nie występowały. Czasy podane w powyższej tabeli są wartościami średnimi dla 10 instancji dla każdej długości sekwencji, z wyjątkiem sekwencji o długości 507nt, dla których ze względu na przyjęty limit czasu obliczeń (10 min) wyliczonych zostało 6 instancji.

W tabeli 2 przedstawiono średnią liczbę rozwiązań znajdowanych przez algorytm w zależności od długości analizowanej sekwencji dla instancji zawierających 10% błędów, które nie wynikały z powtórzeń l-merów.

Jak widać, algorytm był w stanie podać jednoznaczne rozwiązanie tylko dla sekwencji do długości 207nt. Przy dłuższych sekwencjach zauważamy wzrost liczby niejednoznacznych rozwiązań, których liczba dla badanych instacji o długości 507nt wahała się od 2 do 19. Podobnie jak w przypadku tabeli 1, wyniki zamieszczone w tabeli 2 są wynikami średnimi dla 10 instancji, z wyjątkiem sekwencji o długości 507nt, dla których obliczenia zakończono dla 6 instancji.

Wyniki te mogą się wydawać nieco zaskakujące, są jednak zgodne z przewidywaniami teoretycznymi (por. np. [6]). Należy również zauważyć, że liczby rozwiązań dla poszczególnych instancji nie wynikają ze sposobu działania algorytmu, lecz z natury analizowanych sekwencji. Omawiany algorytm, jako metoda dokładna, znajduje wszystkie optymalne rozwiązania kombinatorycznego problemu sekwencjonowania, ale tylko

Tabela 3

Średni czas obliczeń w zależności od liczby błędów wynikających z powtórzeń (czas podano w sekundach)

	Błędy negatywne w spektrum wynikające z powtórzeń				
Spektrum	1% błędów	2% błędów	3% błędów	4% błędów	5% błędów
100-10%	0.04	0.07	0.12	1.40	44.86
200-10%	4.03	35.57	208.38	301.11	–

Tabela 4

Liczba instancji rozwiązanych w ciągu 10 min stanowiących podstawę do obliczenia średnich czasów zamieszczonych w tabeli 3

	Błędy negatywne w spektrum wynikające z powtórzeń				
Spektrum	1% błędów	2% błędów	3% błędów	4% błędów	5% błędów
100-10%	10	10	10	10	10
200-10%	10	10	8	2	0

jedno z nich odpowiada rozwiązaniu problemu biologicznego. Do zidentyfikowania tego rozwiązania byłby potrzebny jednak dodatkowy eksperyment biochemiczny. Liczbę rozwiązań problemu kombinatorycznego problemu sekwencjonowania można zmniejszyć, zwiększając długość l-merów, jednak ze względu na ograniczenia technologiczne obecnie możliwe jest zastosowanie w eksperymencie hybrydyzacyjnym pełnych bibliotek oligonukleotydów zawierających l-mery o długości co najwyżej 10nt.

W tabeli 3 zamieszczono średnie czasy obliczeń dla instancji odpowiadających sekwencjom o długości 107 i 207 nukleotydów zawierających 10% błędów negatywnych, przy czym liczba błędów wynikających z powtórzeń waha się od 1% do 5%.

Wyniki te wyraźnie wskazują, że czas pracy algorytmu rośnie bardzo szybko wraz ze wzrostem liczby błędów wynikających z powtórzeń. O ile dla 5% błędów tego typu oraz dodatkowych 5% błędów hybrydyzacji średni czas pracy wynosił około 14 sekund, o tyle dla 4% błędów wynikających z powtórzeń i 6% błędów hybrydyzacji, dla sekwencji o długościach 207nt czas pracy algorytmu wynosił już ok. 5 minut. Czasy zamieszczone w tabeli 3 są średnimi czasami obliczeń. Liczby instancji stanowiących podstawę do obliczenia średnich czasów zamieszczone są w tabeli 4.

W tabeli 5 przedstawiono wpływ informacji o powtórzeniach l-merów w badanych sekwencjach na liczbę znalezionych rozwiązań.

Wyniki zamieszczone w tej tabeli wskazują na istotną redukcję liczby rozwiązań optymalnych dla problemu, w którym dostępna jest częściowa informacja o powtórzeniach l-merów w stosunku do liczby takich rozwiązań dla analogicznego problemu bez tego rodzaju informacji. Wpływ ten wynika z faktu, że rozwiązanie problemu, którego instancje nie zawierają informacji o powtarzających się l-merach, może zawierać powtórzenia dowolnych (prawie) l-merów (oczywiście, ograniczone ich sekwencją nukleotydów, długością badanej sekwencji DNA oraz licznością spektrum), natomiast rozwiązanie problemu, w którego instancjach informacja o powtórzeniach występuje, musi zawierać powtórzenia ściśle określonych l-merów. Zatem zbiór rozwiązań problemu z

Tabela 5

Porównanie liczby otrzymanych rozwiązań dla sekwencji o długości 107bp przed oraz po weryfikacji na podstawie danych o powtarzających się l-merach

Błędy negatywne w spektrum wynikające z powtórzeń					
weryfikacja	1% błędów	2% błędów	3% błędów	4% błędów	5% błędów
przed	1.1	3.8	8.6	38.9	148.7
po	1.0	1.4	1.6	10.6	19.8

informacją o wielokrotnościach jest podzbiorem zbioru rozwiązań odpowiadającego mu problemu bez takiej informacji (choć nie zawsze musi to być podzbiór właściwy).

4. Podsumowanie

W niniejszej pracy przedstawiono wyniki wstępnej fazy badań problemów sekwencjonowania przez hybrydyzację w przypadku, gdy dostępna jest częściowa informacja o powtórzeniach występujących w analizowanej sekwencji DNA. Rozpatrywano tu najprostszy, ale jednocześnie najbardziej realistyczny model tego typu informacji, zgodnie z którym w instancji problemu z każdym elementem spektrum skojarzona jest informacja o tym, czy dany l-mer występuje w badanej sekwencji raz, czy więcej razy. Jest to, oczywiście, informacja niezbyt precyzyjna, lecz nawet ona pozwala na zmniejszenie liczby roziązań, co ma duże znaczenie dla ewentualnych praktycznych zastosowań metody sekwencjonowania przez hybrydyzację. Można, oczywiście, rozważać problemy, w których informacja o powtórzeniach jest bardziej dokładna [3, 4], co stanowi przedmiot kolejnych planowanych badań. Warto jednak mieć na uwadze ograniczenia technologiczne obecnie stosowanej technologii chipów DNA, które sprawiają, że rozważanie problemów, w których dostępna jest dokładna informacja o powtórzeniach, nie ma na razie bezpośredniego przełożenia na zastosowania praktyczne (co, oczywiście, nie oznacza, że jest bezwartościowe).

BIBLIOGRAFIA

1. Bains W., Smith G.C.: A novel method for nucleic acid sequence determination. Journal of Theoretical Biology, 135, 1988, p. 303–307.
2. Błażewicz J., Formanowicz P., Kasprzak M., Markiewicz W.T., Węglarz J.: DNA sequencing with positive and negative errors. Journal of Computational Biology, 1999, 6, p. 113–123.
3. Formanowicz P.: DNA sequencing by hybridization with additional information available. Computational Methods in Science and Technology, 11, 2005, p. 21–29.
4. Formanowicz P.: Selected combinatorial aspects of biological seqeuence analysis. Wydawnictwo Politechniki Poznańskiej, Poznań 2005.
5. Lysov Yu.P., Florentiev V.L., Khorlin A.A., Khrapko K.R., Shik V.V., Mirzabekov A.D.: Determination of the nucleotide sequence of DNA using hybridization

with oligonucleotides. A new method. Doklady Akademii Nauk SSSR, 303, 1988, p. 1508–1511.

6. Pevzner P.A.: Computational molecular biology. An algorithmic approach. The MIT Press, Cambridge, Massachusetts, 2000.

Recenzent: Dr inż. Krzysztof Fujarewicz

Abstract

In this paper an algorithm for some variant of sequencing by hybridization method is presented. In the standard version of the method information about repetitions is not available. In the paper it is assumed that partial information of this type is a part of the problem instance. Here a simple but realistic model of this information is assumed, i.e. it is known if any element of spectrum appears in the target sequence once or more that once. The proposed algorithm of branch and bound type solves the variant of the problem with negative errors. Results of computational experiment are reported which, among others, confirm that the additional information leads to improvement of the obtained solutions.