SEKWENCJONOWANIE CZ. 3

Piotr Formanowicz

Sekwencjonowanie wieloetapowe

Wykonywana jest seria eksperymentów hybrydyzacyjnych – w kolejnych eksperymantach wzrasta długość elementów biblioteki oligonucleotydów.

Podejście takie powinno zmniejszyć liczbę błędów wynikających z powtórzeń l-merów w badanej sekwencji DNA.

Prawdopodobieństwo powtórzenia się podciągu o długości l w losowej sekwencji maleje wraz ze wzrostem l. Innymi słowy, im większa jest długość elementów biblioteki oligonukleotydów, tym lepsza powinna być jakość uzyskanej sekwencji.

Może to prowadzić do wniosku, że należałoby korzystać z bibliotek zawierających jak najdłuższe oligonukleotydy.

W praktyce jest to jednak niemożliwe ze względu na ograniczenia technologiczne.

W wieloetapowym SBH pierwszy eksperyment hybrydyzacyjny wykonywany jest za pomocą klasycznej pełnej biblioteki oligonukleotydów o długości l. W ten sposób otrzymuje się informację o wszystkich rodzajach podciągów o długości l występujących w badanej sekwencji.

Wszystkie podciągi o długości $2l$ występujące w badanej sekwencji muszą być konkatenacjami pewnych podciągów o długości l, które również w tej sekwencji występują, a zatem zostały wykryte w pierwszym eksperymencie hybrydyzacyjnym.

Celem drugiego eksperymentu hybrydyzacyjnego jest wykrycie wszystkich podciągów o długości $2l$.

Jest on przeprowadzany za pomocą biblioteki składającej się ze wszystkich oligonukleotydów o długości $2l$ będących konkatenacjami ciągów o długości l wykrytych w pierwszym eksperymencie (zamiast pełej biblioteki $2l$-merów).

W ten sposób uzyskuje się zmniejszenie liczby elementów biblioteki z 4^{2l} do $(n-l+1)^2$ (w najgorszym przypadku).

Możliwe jest dalsze zmniejszenie wielkości biblioteki.

Wiele z $2l$-merów uzyskanych przez konkatenację ciągów otrzymanych w pierwszym eksperymencie nie występuje w badanej sekwencji.

Łatwo zauważyć, że każdy podciąg o długości $l < l' < 2l$ występujący w badanej sekwencji musi zawierać dwa podciągi o długości l nakładające się na siebie na $2l - l'$ pozycjach.

Stąd, zamiast bibliotekę składającą się ze skonkatenowanych par l-merów wykrytych w pierwszym eksperymencie, można zastosować bibliotekę składającą się z częściowo nakładających się tego typu par (będzie ich mniej).

Oczywiście, w ten sposób długość elementów biblioteki oligonukleotydów nie jest podwajana w kolejnych iteracjach.

Występowanie powtórzeń podciągów w badanej sekwencji nie wpływa na jakość otrzymywanego rozwiązania (wystarczy wykrycie jednego wystąpienia danego podciągu, by biblioteka stosowana w kolejnym etapie zawiarała odpowiednie podciągi).

Błędy negatywne wynikające z niedokładności hybrydyzacji mają istotny wpływ na jakość rozwiązań, gdyż brak któregoś z oligonukleotydów w bibliotece na danym etapie będzie się propogował przy konstruowaniu bibliotek wykorzystywanych na kolejnych etapach.

Błędy pozytywne nie stanowią dużego problemu (nie uniemożliwiają wykrycia wszystkich podciągów badanej sekwencji o danej długości), ale mogą powodować nieporządane zwiększenie wielkości biblioteki.

Metoda wieloetapowego SBH minimalizuje wpływ błędów wynikających z powtórzeń, ale jest czuła na błędy wynikające z niedoskonałości hybrydyzacji.

W celu wyeliminowania tego ograniczenia można ją połączyć z metodą, która zmniejsza prawdopodobieństwo wystąpienia błędów tego drugiego rodzaju (ale być może jest czuła na błędy wynikające z powtórzeń).

Metodą taką jest izotermiczne SBH.

Trzeba jednak udowodnić, że w opisany wcześniej sposób można uzyskiwać biblioteki izotermiczne i zaprojektować algorytmy konstrukcji odpowiednich bibliotek, co jest możliwe do przeprowadzenia.

Informacja o powtórzeniach

W klasycznym SBH oraz wielu jego odmianach nie jest brana pod uwagę informacja o powtórzeniach l-merów.

Informacja odczytana z chipu DNA jest binarna. Spektrum jest zbiorem, a nie multizbiorem.

Na obecnym etapie rozwoju technologii chipów DNA

możliwe jest odczytanie częściowej (przybliżonej) informacji o powtórzeniach.

Jest ona związana z intensywnością sygnału, podobnie jak w przypadku mikromacierzy DNA służących do badania ekspresji genów.

Nie można jednak dokładnie związać liczby powtórzeń z intensywnością sygnału, dlatego możliwe jest pozyskanie tylko przybliżonej informacji o liczbie powtórzeń poszczególnych podciągów.

Przyjmuje się dwa podstawowe rodzaje informacji o powtórzeniach, tj. "jeden i wiele" oraz "jeden, dwa i wiele".

Przy rozważaniu problemów sekwencjonowania związanych z informacją o powtórzeniach pojawia sią konieczność zdefiniowania kilku rodzajów spektr.

$S(Q)$ – spektrum sekwencji Q.

$S^{(is)}(Q)$ – idealne spektrum sekwencji Q. Spektrum takie zawiera wszystkie rodzaje l-merów występujących w sekwencji Q, ale nie wszystkie takie l-mery (nie jest multizbiorem).

$S^{(im)}(Q)$ – idealne multispektrum sekwencji Q. Zawiera wszystkie l-mery występujące w sekwencji Q.

$S^{(m)}(Q)$ – multispektrum sekwencji Q. Multispektrum, w przeciwieństwie do idealnego multispektrum, może zawierać błędy (podobnie jak spektrum). Ponadto, wielokrotność elementów multispektrum może być ograniczona w wielu sformułowaniach problemów SBH z informacją o powtórzeniach.

Ponadto, z każdą sekwencją $s_i \in S(Q)$ związny jest parametr m_i, który jest równy liczbie wystąpień s_i w $S^{(m)}(Q)$.

Istnieje ścisły związek między wymienionymi rodzajami spektr.

Idealne multispektrum jest multizbiorem zawiarającym wszystkie l-mery występujące w badanej sekwencji DNA.

Multispektrum jest multizbiorem, który może zawiarać tylko część l-merów będących elementami idealnego multispektrum. Ponadto, może ono zawierać l-mery, które nie należą do idealnego multispektrum. Stąd, multispektrum odpowiada wynikowi rzeczywistego eksperymentu hybrydyzacyjnego, w którym można uzyskać częściową informację o powtórzeniach.

Spektrum jest zbiorem, który można uzyskać z multispektrum poprzez pominięcie powtórzeń l-merów. Idealne spektrum można uzyskać w ten sam sposób z idealnego multispektrum.

Decyzyjne wersje problemów SBH z informacją o powtórzeniach należą do klasy **P**, jednak wersje przeszukiwania wielu z nich są **NP**-trudne.

Informacja typu "jeden i wiele"

W przypadku problemów z informacją typu "jeden i wiele" każdy oligonukleotyd może wystąpić w multispektrum 0, 1 lub 2 razy. Stąd, dla każdej sekwencji s_j, $m_i \in \{1, 2\}$, przy czym $m_i = 2$ interpretowane jest jako co najmniej dwa wystąpienia s_i w badanej sekwencji w przypadku, gdy nie ma błędów pozytywnych lub co najmniej jedno wystąpienie s_i oraz jeden błąd pozytywny związany z s_i w przypadku, gdy takie błędy występują.

Informacja typu "jeden i wiele" odpowiada sytuacji, w której z chipu można jedynie odczytać informację, że dany l-mer w ogóle nie wystąpił w badanej sekwencji DNA, wystąpił w niej jeden raz, bądź wystapił on wielokrotnie.

Problem bez błędów

$$S(Q) = S^{(is)}(Q) = S^{(m)}(Q) = S^{(im)}(Q)$$

INSTANCJA: zbiór $S(Q) = S^{(is)}(Q)$, długość n sekwencji Q.
ODPOWIEDŹ: sekwencja Q' o długości n zawierająca wszystkie i tylko te l-mery, które należą do $S(Q)$.

Zauważmy, że brak błędów oznacza, że w badanej sekwencji nie mogą występować powtórzenia o długości co najmniej l. Gdyby takie powtórzenia występowały, parametr m_i miałby wartość 2 dla niektórych s_i, co jednak nie mogłoby być jednoznacznie zinterpretowane jako konkretna liczba powtórzeń i byłoby źródłem błędów.

Problem ten jest równoważny klasycznemu problemowi SBH bez błędów, stąd może być rozwiązany za pomocą algorytmu Pevznera.

Problem z błędami negatywnymi wynikającymi z powtórzeń

$$S(Q) = S^{(is)}(Q) \subset S^{(m)}(Q) \subseteq S^{(im)}(Q)$$

INSTANCJA: zbiór $S(Q) \subset S^{(im)}(Q)$, długość n sekwencji Q, parametr $m_i \in \{1, 2\}$ dla każdej sekwencji $s_i \in S(Q)$.
ODPOWIEDŹ: sekwencja Q' o długości n zawierająca wszystkie i tylko te l-mery, które są elementami $S(Q)$, gdzie sekwencja $s_i \in S(Q)$ występuje raz w Q', jeżeli $m_i = 1$ oraz występuje ona w Q' co najmniej dwa razy, jeżeli $m_i = 2$.

Zauważmy, że jeżeli $\forall_{s_i \in S(Q)} m_i = 1$, to problem ten redukuje się do problemu bez powtórzeń i może być rozwiązany w czasie wielomianowym.

Problem z błędami negatywnymi dowolnego rodzaju

$S(Q) \subseteq S^{(is)} \subseteq S^{(im)}(Q)$ oraz $S(Q) \subseteq S^{(m)}(Q) \subseteq S^{(im)}(Q)$.

INSTANCJA: zbiór $S(Q) \subseteq S^{(im)}(Q)$, długość n sekwencji Q, parametr $m_i \in \{1, 2\}$ dla każdego $s_i \in S(Q)$.
ODPOWIEDŹ: sekwencja Q' o długości n zawierająca wszystkie elementy z $S(Q)$, gdzie $s_i \in S(Q)$ występuje w Q' raz, jeżeli $m_i = 1$ oraz występuje w Q' co najmniej dwa razy, jeżeli $m_i = 2$. Ponadto, Q' może zawierać pewne l-mery, które nie są elementami $S(Q)$.

W przypadku, gdy $\forall_{s_i \in S(Q)} m_i = 1$, problem ten redukuje się do problemu z błędami negatywnymi wynikającymi z hybrydyzacji, stąd jest on silnie **NP**-trudny.

Informacja typu "jeden, dwa i wiele"

W przypadku problemów z informacją typu "jeden, dwa i wiele" każdy l-mer może wystąpić w multispektrum 0, 1, 2 lub 3 razy (czyli dla dowolnej sekwencji s_i, $m_i \in \{1, 2, 3\}$).

Przypadek $m_i = 3$ oznacza, że s_i pojawia się w badanej sekwencji co najmniej trzy razy, jeżeli nie ma błędów pozytywnych lub co najmniej dwa razy i dodatkowo pojawia się jeden błąd pozytywny z nią związany (w przypadku problemów z błędami pozytywnymi nie wiadomo, która z tych dwóch możliwości zachodzi).

Jeżeli $m_i = 2$, oznacza to, że badana sekwencja zawiera dokładnie dwa wystąpienia s_i, w przypadku bez błędów pozytywnych. Jeżeli błędy pozytywne występują, $m_i = 2$ oznacza, że badana sekwencja zawiera dwa wystąpienia s_i lub jedno wystąpienie s_i i jeden błąd pozytywny (te dwa przypadki są nierozróżnialne).

Jeżeli $m_i = 1$, to s_i występuje raz w badanej sekwencji, jeżeli nie ma błędów pozytywnych. Jeżeli błędy pozytywne występują, $m_i = 1$ oznacza, że s_i występuje raz w badanej sekwencji lub w niej nie występuje, ale w danych wejściowych jest obecny błąd pozytyny z nią związany (przypadki te są nierozróżnialne).

Literatura

[1] Błażewicz J., Formanowicz P., Multistage isothermic sequencing by hybridization, *Computational Biology and Chemistry*, 2005, 29, 69–77.
[2] Formanowicz P., DNA sequencing by hybridization with additional information available, *Computational Methods in Science and Technology*, 2005, 11, 21–29.
[3] Kruglyak S., Multistage sequencing by hybridization, *Journal of Computational Biology*, 1998, 5, 165–171.

v. 1.0.0