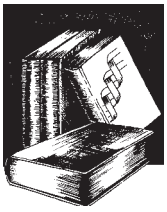

Mikromacierze DNA – zasady projektowania sond

Piotr Formanowicz[1,2,*], Radosław Urbaniak[1], Luiza Handschuh[2,3], Dorota Formanowicz[4], Marek Figlerowicz[2]

[1] Instytut Informatyki, Politechnika Poznańska, Poznań
[2] Instytut Chemii Bioorganicznej, Polska Akademia Nauk, Poznań
[3] Katedra i Klinika Hematologii i Chorób Rozrostowych Układu Krwiotwórczego, Uniwersytet Medyczny im. K. Marcinkowskiego, Poznań
[4] Katedra Chemii i Biochemii Klinicznej, Uniwersytet Medyczny im. K. Marcinkowskiego, Poznań

DNA microarray probe design

Summary

DNA microarrays are widely used in many areas of biological research. They are an efficient tool for gene expression analysis due to a high level of parallelism, what means that they allow for simultaneous measuring of the transcriptional activity of all genes present in the studied genome. The quality of the results obtained using microarrays depends among other factors on the proper design of probes. Two general features which should characterize each probe are sensitivity and specificity. Since designing a set of probes having both of these properties is usually a complex task, many algorithms supporting this process have been developed and implemented. However, the designing method should be carefully chosen such that the results will match the requirements following from the nature of the biological problem to be solved. In this paper the criteria used for DNA microarray design are described and some computer based approaches are presented.

Key words:
DNA micrarrays, probe selection, probe features, computer based methods.

Adres do korespondencji

Piotr Formanowicz,
Instytut Informatyki,
Politechnika Poznańska,
ul. Piotrowo 2,
60-965 Poznań;
e-mail:
piotr@cs.put.poznan.pl

biotechnologia
4 (83) 54–67 2008

1. Wstęp

Jednym z najwiekszych wyzwań przed jakimi stoi obecnie biologia molekularna i obliczeniowa jest dokładne poznanie struktury genomów oraz mechanizmów kontrolujących sposób

ich funkcjonowania. W tym celu tworzone są coraz bardziej doskonałe narzędzia umożliwiające precyzyjną analizę aktywności transkrypcyjnej genomu oraz śledzenie zachodzących w nim zmian. Jednym z takich narzędzi są mikromacierze DNA.

Mikromacierze są miniaturowymi układami hybrydyzacyjnymi składającymi się z sond specyficznie rozpoznających fragmenty poszczególnych genów lub transkryptów. Mogą one służyć zarówno do analizy strukturalnej jak i funkcjonalnej genomu, stąd znajdują liczne zastosowania w wielu dziedzinach biologii i medycyny.

Podstawowym problemem, jaki należy rozwiązać przed przystąpieniem do właściwych badań jest odpowiednie zaplanowanie całego eksperymentu. Jeżeli nie korzystamy z macierzy komercyjnej głównym zadaniem staje się zaprojektowanie zestawu sond, które będą następnie umieszczone na macierzy. Warunkiem niezbędnym uzyskania wiarygodnych wyników w eksperymencie mikromacierzowym jest wysoka czułość oraz specyficzność sond. Oznacza to, że każda z nich musi specyficznie rozpoznawać fragment genomu lub transkryptomu powstałego podczas ekspresji informacji genetycznej. Ze względu na złożoność problemu projektowania mikromacierzy do jego rozwiązania stosowane są metody informatyczne (1). W pracy przedstawione zostaną podstawowe kryteria stosowane przy doborze sond oraz przykłady wykorzystywanych w praktyce algorytmów.

2. Projektowanie sond

Dwie zasadnicze cechy, jakie powinny posiadać sondy, z których zbudowana jest mikromacierz to wysoka czułość oraz specyficzność. Pierwsza z tych cech oznacza, że dana sonda z wysokim prawdopodobieństwem hybrydyzuje z określonym fragmentem DNA lub RNA, którego obecność w badanej próbie ma wykrywać. Jednym z podstawowych warunków, jakie musi spełnić sonda jest zatem pełna komplementarność do wybranego fragmentu sekwencji docelowej. Z kolei specyficzność sondy oznacza minimalizację prawdopodobieństwa jej hybrydyzacji do sekwencji innej niż docelowa. Poszukiwana sonda powinna zatem charakteryzować się jak najmniejszym stopniem komplementarności do wszystkich sekwencji, które mogą znaleźć się w badanej próbie, z wyjątkiem sekwencji docelowej. Z algorytmicznego punktu widzenia zapewnienie wysokiej czułości jest zadaniem stosunkowo prostym – należy dla każdej sekwencji, która ma być wykrywana za pomocą mikromacierzy zaprojektować w pełni komplementarną sondę – przy czym zarówno sonda, jak i komplementarny do niej fragment sekwencji docelowej samodzielnie nie powinny tworzyć stabilnych struktur drugorzędowych. O wiele bardziej skomplikowanym problemem jest zapewnienie wysokiej specyficzności sond. Oczywiście, każda z nich musi posiadać obie właściwości jednocześnie, dlatego jako sondę należy wybrać taki oligonukleotyd, który jest w pełni komplementarny do fragmentu wybranego genu lub mRNA, a jednocześnie jak najmniej komplementarny do wszystkich innych genów lub mRNA z badanej próby. Jest to zatem zadanie minimalizacji war-

tości pewnego kryterium. Może ono jednak zostać sformułowane w inny sposób, tj. mogą być poszukiwane sekwencje, których stopień komplementarności do sekwencji innych niż docelowe nie przekracza pewnego progu, poniżej którego prawdopodobieństwo hybrydyzacji jest wystarczająco małe.

Wspomniana pełna komplementarność lub jej brak ma zapewnić zachodzenie hybrydyzacji z jak największym lub jak najmniejszym prawdopodobieństwem. Jego wielkość zależeć będzie przede wszystkim od energii wiązań wodorowych, jakie mogą się utworzyć między cząsteczkami DNA lub RNA. Użytecznym parametrem będącym miarą tej energii jest temperatura topnienia dupleksu (T_m) DNA/DNA, DNA/RNA lub RNA/RNA. T_m definiuje się jako temperaturę, w której równo połowa dupleksów ulega rozpleceniu (przejściu z formy dwuniciowej do jednoniciowej) (2-4). W rezultacie podstawowym kryterium stosowanym przy projektowaniu sond nie jest komplementarność sekwencji (komplementarność dwóch ciągów znaków), lecz temperatura topnienia dupleksu tworzonego przez te sekwencje. Oczywiście istnieje zależność pomiędzy ciągami znaków reprezentującymi sekwencje a temperaturą topnienia, ale związek ten nie jest do końca jasny. Precyzyjne obliczenie temperatury topnienia dupleksu jest złożonym zagadnieniem termodynamicznym, które nie doczekało się dotąd dokładnego rozwiązania. Niemniej jednak wiele badań, w których dążono do określenia zależności temperatury topnienia od sekwencji nukleotydowej cząsteczek tworzących dupleks zostało przeprowadzonych i pewne modele z nich wynikające są z powodzeniem stosowane w praktyce. Różnią się one oczywiście dokładnością wyznaczanej temperatury i złożonością obliczeń koniecznych do przeprowadzenia. Według najprostszego z nich każda para A-T wnosi 2°C do temperatury topnienia dupleksu, a para C-G wnosi 4°C (2,3). Choć bardzo uproszczony, model ten jest często stosowany do projektowania starterów do reakcji PCR. Najbardziej zbliżone do rzeczywistości wyniki daje metoda najbliższego sąsiada, w której, przynajmniej do pewnego stopnia, uwzględniana jest nie tylko liczba poszczególnych nukleotydów w cząsteczkach tworzących dupleks, ale również ich sekwencje (4-6). Wadą standardowej wersji tej metody jest to, że daje ona stosunkowo dokładne wyniki dla sekwencji całkowicie komplementarnych, natomiast zawodzi w przypadku występowania różnego typu niedopasowań. Dlatego model został rozszerzany przez wprowadzanie dodatkowych parametrów odpowiadających różnego rodzaju niedopasowaniom (7-12). Niezależnie jednak od metody, jaka została wykorzystana do wyznaczenia temperatury topnienia, sondy należy zaprojektować w taki sposób, by dupleksy jakie tworzą one z sekwencjami docelowymi charakteryzowały się identycznymi lub zbliżonymi wartościami T_m. Jednakże dupleksy tworzone z sekwencjami innymi niż docelowe powinny mieć temperatury topnienia na tyle niskie, by nie dochodziło do ich utworzenia.

Przedstawiliśmy jedynie ogólny zarys metody projektowania macierzy DNA, w której zasadniczym kryterium przydatności sond jest temperatura topnienia dupleksu. W praktyce sondy wybiera się na podstawie szeregu łatwych do sprawdzenia kryteriów cząstkowych, których suma stanowi przybliżenie kryterium dokładnej temperatury topnienia. Najczęściej stosowane są następujące reguły (13-17):

– zawartość danego nukleotydu w sondzie nie może stanowić więcej niż 50% sekwencji,

– fragmenty składające się z nukleotydów jednego rodzaju nie powinny przekraczać 25% długości sondy,

– zawartość nukleotydów GC powinna mieścić się w granicach od 30 do 70%,

– oligonukleotydy będące sondami oraz komplementarne do nich sekwencje docelowe nie powinny tworzyć stabilnych struktur drugorzędowych,

– długość ciągłego fragmentu sondy (podciągu) komplementarnego do sekwencji nie będącej sekwencją docelową nie powinna przekraczać 15 nukleotydów,

– stopień komplementarności do sekwencji nie będącej sekwencją docelową nie powinien przekraczać 75%.

Kryteria te zostały m. in. wykorzystane przy tworzeniu programu PICKY (18), którego ciekawą właściwością jest brak konieczności określenia dokładnej długości projektowanych sond oraz temperatury topnienia. Użytkownik podaje jedynie pewien zakres długości sond oraz minimalną różnicę między temperaturami topnienia dupleksów tworzonych z sekwencjami docelowymi i pozostałymi sekwencjami. Biorąc pod uwagę te ograniczenia program PICKY dobiera sondy tak, by wykazywały one maksymalną czułość i specyficzność.

W jednej z metod projektowania mikromacierzy genomowych unikatowość sond sprawdzana jest na podstawie odległości Levensteina (18). Zgodnie z definicją odległość Levensteina między sekwencjami s i t, oznaczona przez $L(s,t)$, równa jest najmniejszej liczbie elementarnych operacji edycyjnych niezbędnych do przekształcenia s w t (lub odwrotnie) (19). Wspomnianymi operacjami jest zamiana, wstawienie i usunięcie pojedynczego znaku. Autorzy metody przyjmują, że oligonukleotyd s jest unikatowy, jeżeli nie istnieje (w zbiorze rozważanych sekwencji) oligonukleotyd t taki, że $L(s,t)\leq k$, gdzie k jest przyjętym progiem, a ponadto wystąpienia s i t w analizowanych sekwencjach nie nakładają się na siebie. Autorzy przyjęli długość sond równą 25 nukleotydów, natomiast wartość progu k ustalona została na 4.

Wyselekcjonowane w ten sposób sondy poddawane są dalszej analizie, w której zmierza się do usunięcia zarówno tych, które mogą hybrydyzować same ze sobą jak i tych, które tworzą z sekwencjami docelowymi dupleksy o zbyt niskiej temperaturze topnienia. W tym celu zastosowano następujące kryteria:

– oligonukleotyd może zawierać najwyżej 12 nukleotydów A, 12 nukleotydów T, 10 nukleotydów C i 10 nukleotydów G,

– żaden podciąg o długości 8 nukleotydów nie może zawierać więcej niż 6 nukleotydów A, 6 T, 4 C i 4 G,

– sonda może zawierać najwyżej 6 kolejnych nukleotydów A, 6 nukleotydów T, 5 nukleotydów C i 5 nukleotydów G,

– końce sondy nie powinny być wzajemnie komplementarne.

Warto zwrócić uwagę na fakt, że wymienione kryteria są jedynym warunkiem mającym zapewnić odpowiednią temperaturę topnienia dupleksów tworzonych przez sondy z sekwencjami z badanej próby.

Na podobnych zasadach oparty został program YODA (20). W tym przypadku proces projektowania ma zapewnić odpowiednią czułość, specyficzność oraz spójność sond. Założona czułość sond osiągana jest poprzez eliminację oligonukleotydów, które mogą tworzyć stabilne struktury drugorzędowe bądź homodimery. Sondy posiadające takie właściwości miałyby ograniczoną zdolność do hybrydyzacji z sekwencją docelową. Specyficzność zapewniana jest przez eliminację z początkowego zbioru oligonukleotydów tych jego elementów, które wykazują więcej niż 75% komplementarności do sekwencji innej niż docelowa oraz tych, które zawierają podciąg dłuższy niż 15 nukleotydów całkowicie komplementarny do sekwencji różnej od docelowej. Ponadto eliminowane są oligonukleotydy zawierające długie podciągi złożone z nukleotydów jednego rodzaju. Spójność zapewniana jest poprzez dobór oligonukleotydów o zbliżonej temperaturze topnienia oraz takich, które są komplementarne do określonego obszaru sekwencji docelowej, np. blisko końca 3' lub 5', bądź blisko środka sekwencji – w zależności od sposobu przygotowania badanej próby.

Do określenia temperatury topnienia stosowany jest model najbliższego sąsiada z parametrami podanymi przez SantaLucię (4). Najpierw wyznaczana jest średnia temperatura topnienia dupleksów tworzonych przez wszystkie oligonukleotydy o podanej długości, a następnie użytkownik podaje dopuszczalny zakres temperatur. Na wstępie sprawdza się czy w obrębie oligonukleotydów o zadanej długości występują wcześniej zdefiniowane przez użytkownika tzw. sekwencje zabronione, np. podciągi składające się z nukleotydów jednego rodzaju. Oligonukleotydy zawierające takie sekwencje są eliminowane ze zbioru potencjalnych sond. Następnie sprawdzana jest temperatura topnienia dupleksów tworzonych przez oligonukleotydy, które przeszły pozytywnie poprzedni test. Jeśli nie mieści się ona we wcześniej zdefiniowanym przedziale wartości, sonda jest odrzucana. W dalszej kolejności sprawdzana jest możliwość tworzenia przez oligonukleotydy stabilnych struktur drugorzędowych. W badaniu tym nie jest stosowane podejście termodynamiczne, gdyż jego celem nie jest znalezienie najbardziej stabilnej struktury, lecz sprawdzenie, czy powstanie jakiejkolwiek struktury tego typu jest prawdopodobne.

Na tym etapie weryfikacji oligonukleotydów każdej sekwencji docelowej można przypisać wiele potencjalnych sond (może się jednak również zdarzyć, że pewnej sekwencji nie będzie można przypisać żadnej sondy). W celu zidentyfikowania najlepszych sond, dla każdej z sekwencji docelowych przeprowadzana jest dodatkowa selekcja. Podstawowym jej celem jest wybór odpowiedniego podzbioru sond, którego elementy będą wykazywać jakąś charakterystyczną cechę np. równomierny rozkład wzdłuż sekwencji docelowej. Innym kryterium selekcji może być położenie sond blisko jednego z końców lub środka sekwencji docelowej. Możliwe jest też zażądanie, by sondy nie nakładały się na siebie.

Końcowa analiza potencjalnych sond, które przeszły przez wszystkie poprzednie etapy polega na sprawdzeniu możliwości dimeryzacji oraz określeniu komplementarności do sekwencji innych niż docelowe. Domyślny próg komplementarności, powyżej którego oligonukleotydy są odrzucane wynosi 80%.

W innej metodzie projektowania sond wykorzystuje się drzewa sufiksowe oraz programowanie dynamiczne (21). Metoda ta rozpoczyna działanie od konstrukcji uogólnionego drzewa sufiksowego na podstawie sekwencji komplementarnych do sekwencji docelowych. Drzewo sufiksowe jest strukturą danych umożliwiającą szybkie wyszukiwanie powtarzających się podsekwencji. Właściwość ta jest wykorzystana do identyfikacji niespecyficznych oligonukleotydów, które są usuwane ze zbioru sond. Dla wszystkich dupleksów tworzonych przez kandydatów na sondy i podciągi sekwencji docelowych obliczana jest ich temperatura topnienia. Jest ona wyznaczana za pomocą rozszerzonego modelu najbliższego sąsiada, w którym oprócz par Watsona-Cricka uwzględnione są również inne pary zasad, a także pozycje niesparowane. Ze względu na fakt, że genomowe DNA zawierają dużo powtarzających się podsekwencji, efektywność algorytmu może zostać zwiększona przez unikanie wielokrotnego obliczania temperatury topnienia dla pewnych fragmentów sond oraz sekwencji docelowych. W tym celu już w początkowej fazie działania algorytmu potencjalne sondy są zapisane w uogólnionym drzewie sufiksowym. W interesujący sposób rozwiązany został problem wyznaczania temperatur topnienia za pomocą programowania dynamicznego. Ponieważ oprócz dupleksów całkowicie dopasowanych należy wziąć pod uwagę również takie, w których występują niedopasowania, stąd najpierw należy wyznaczyć optymalne dopasowania sond do sekwencji niedocelowych. Miarą służącą do wyznaczenia tych optymalnych dopasowań jest temperatura topnienia. Zatem oba problemy, tj. znalezienie optymalnego dopasowania sond z sekwencjami niedocelowymi oraz wyznaczenie temperatur topnienia odpowiadających takim dopasowaniom sekwencji są ze sobą ściśle powiązane. Autorzy rozwiązują oba te problemy jednocześnie za pomocą programowania dynamicznego wyznaczającego dopasowanie termodynamiczne, tj. takie, któremu odpowiada najwyższa temperatura topnienia.

Inną interesującą metodę doboru sond opracował Hu i wsp. (22). W odróżnieniu od wielu innych nie polega ona na sprawdzaniu kolejno wybranych kryteriów cząstkowych i eliminowaniu oligonukleotydów ich niespełniających, lecz na sprawdzaniu kryterium zbiorczego, utworzonego ze standardowych kryteriów cząstkowych, którym przypisano odpowiednie wagi. Tymi kryteriami są: specyficzność, złożoność sekwencji, temperatura topnienia oraz prawdopodobieństwo utworzenia struktury drugorzędowej. Klasyczne podejście oparte na sekwencyjnym sprawdzaniu kryteriów cząstkowych w niektórych przypadkach może prowadzić do wyboru sond o niskiej jakości, zwłaszcza gdy projektowana mikromacierz ma zostać wykorzystana do badania genomu, w którym występuje dużo powtórzonych podsekwencji lub występuje duża zmienność zawartości nukleotydów GC. Oligonukleotydy, które pomyślnie przeszły przez etap sprawdzania danego kryterium są następnie weryfikowane pod kątem kolejnego z nich, jednak bez uwzględniania rezultatów poprzednich testów. Innymi słowy, na kolejnych etapach filtrowania (kryterium cząstkowe działa jak filtr) wszystkie oligonukleotydy traktowane są jednakowo. Wady tej nie posiada metoda, w której stosowane jest tylko jedno kryterium, składające się z kryteriów cząstkowych.

Interesujące podejście do projektowania dłuższych sond (ok. 50 nukleotydów i więcej) zaimplementowane zostało w programie GoArrays (23). W opisanych tradycyjnych metodach zakłada się pełną komplementarność sondy z sekwencją docelową na całej długości. Dodatkowo, niemal we wszystkich stosuje się następujące ograniczenia:

– komplementarność do sekwencji nie będącej sekwencją docelową nie powinna przekraczać 75%,

– długość ciągłego fragmentu sondy (podciągu) komplementarnego do sekwencji nie będącej sekwencją docelową nie powinna przekraczać 15 nukleotydów.

Jednakże nie zawsze możliwe jest zaprojektowanie specyficznych sond z uwzględnieniem tych ograniczeń. Przykładowo, dla drożdży *Saccharomyces cerevisiae* 253 rejony kodujące (4,5% wszystkich tego typu rejonów) nie mogą być reprezentowane przez specyficzne sekwencje. Program OligoArray 2.0, wykorzystujący model termodynamiczny najbliższego sąsiada, nie znajduje specyficznych sond dla 7% rejonów kodujących *Arabidopsis thaliana*. Sytuacja przedstawia się jeszcze gorzej w przypadku *Encephalitozoon cuniculi*, gdzie dla tradycyjnej metody projektowania sekwencji o długości 50 nukleotydów, utworzyć można sondy specyficzne zaledwie dla ok. 40% rejonów kodujących.

Autorzy programu GoArrays próbują rozwiązać problem specyficzności poprzez zastosowanie nieco innego podejścia. Zamiast jednej specyficznej sondy program wyszukuje dwa krótsze podciągi specyficzne (o długości np. 25 nukleotydów), oddalone od siebie o zadaną liczbę nukleotydów (liczba ta powinna mieścić się w określonym przedziale). Oba podciągi muszą być w pełni komplementarne do fragmentów rozpatrywanego obszaru. Następnie łączone są one krótkim, losowo wygenerowanym ciągiem nukleotydów (zwykle 3-6 nukleotydów). Utworzona w ten sposób sonda nie jest komplementarna do sekwencji docelowej na całej długości. Sekwencja docelowa po przyłączeniu tworzy pętlę, której długość równa jest odległości między znalezionymi podciągami. Specyficzność skonstruowanej w ten sposób sondy sprawdza się ponownie za pomocą opisanych testów, ponieważ mogła ona zostać zaburzona przez wstawienie losowego łącznika. W ostatnim etapie eliminowane są sondy, które:

– nie mieszczą się w dopuszczalnym przedziale temperatury topnienia, obliczanej za pomocą modelu najbliższego sąsiada,

– tworzą stabilne struktury drugorzędowe (jest to sprawdzane za pomocą programu Mfold),

– zawierają zdefiniowane przez użytkownika sekwencje zabronione.

W przypadku gdy sonda zostanie odrzucona, analizowany obszar sekwencji docelowej zostaje przesunięty na kolejną pozycję, a cały proces trwa tak długo, aż poprawny oligonukleotyd zostanie znaleziony.

Kolejny program, OligoArray, służy do projektowania sond o stałej długości (24). Może on także działać przy założeniu kilku stałych parametrów, np. możliwe jest przyjęcie: stałej liczby sond dla jednej sekwencji docelowej, maksymalnej odległości

od końca sekwencji docelowej, zakresu temperatur topnienia, progu temperatury topnienia struktur drugorzędowych, modyfikacji (chemicznych) końca 5' i/lub 3' sondy oraz zbioru sekwencji zabronionych.

Każda z sekwencji, dla których mają być zaprojektowane sondy przeglądana jest od końca 3' za pomocą przesuwającego się okna o długości projektowanych sond. Zawartość tego okna jest w pierwszej kolejności porównywana ze zbiorem sekwencji zabronionych. Następnie sprawdzana jest unikatowość wskazywanych przez okno oligonukleotydów przez porównanie ze zbiorem wszystkich transkrybowanych sekwencji organizmu, dla którego projektowana jest macierz. W metodzie tej próg specyficzności jest wyższy niż często stosowany próg zaproponowany przez Kane'a i wsp. (13). Dla fragmentów o długości większej niż 50 nukleotydów stopień identyczności powinien być mniejszy niż 50%. Dla fragmentów o długości od 36 do 50 nukleotydów powinien on być mniejszy niż 60%, a dla fragmentów o długości od 15 do 35 nukleotydów – mniejszy niż 70%. Podciągi o długości mniejszej niż 15 nukleotydów w pełni komplementarne do fragmentów sekwencji niedocelowych są akceptowane.

Sekwencje, które przejdą test specyficzności są sprawdzane pod względem możliwości tworzenia struktur drugorzędowych. Temperatura topnienia poszczególnych struktur drugorzędowych obliczana jest za pomocą programu Mfold (25). Oligonukleotyd jest akceptowany, jeżeli nie tworzy struktury o temperaturze topnienia przekraczającej określony przez użytkownika próg. Jeżeli znajdujący się w oknie oligonukleotyd nie spełnia postawionych kryteriów, jest ono iteracyjnie przesuwane o 10 nukleotydów w kierunku końca 5', dopóki odpowiedni oligonukleotyd nie zostanie znaleziony lub nie zostanie osiągnięta minimalna dopuszczalna odległość końca 5' oligonukleotydu od końca analizowanej sekwencji.

W programie OligoPicker zaimplementowana jest metoda projektowania sond dla rejonów kodujących (26). Metoda ta polega na sekwencyjnym przeprowadzaniu testów weryfikujących określone właściwości potencjalnych sond. Podstawowym testem jest sprawdzenie, czy oligonukleotyd zawiera odpowiednio długi ciągły fragment komplementarny do jakiejkolwiek z analizowanych sekwencji. Badania przeprowadzone przez autorów metody są zgodne z wcześniejszymi obserwacjami, że odrzucane powinny być sondy zawierające 15-nukleotydowe fragmenty komplementarne do innych sekwencji niż docelowa. Ponadto, eliminowane są oligonukleotydy zawierające ciągi identycznych nukleotydów oraz tworzące struktury drugorzędowe. Jednak według autorów dwa ostatnie testy w nieznacznym tylko stopniu zmniejszają liczność zbioru potencjalnych sond. Sprawdzana jest również temperatura topnienia oligonukleotydów, obliczana wg wzoru:

$$T_m \quad 64{,}9 \quad 41\frac{g}{l} \quad \frac{600}{l}$$

gdzie g oznacza liczbę nukleotydów C i G w oligonukleotydzie, a l jest jego długością. Ponieważ RNA inne niż mRNA mogą zaburzać eksperyment hybrydyzacyjny, do-

datkowo odrzucane są również oligonukleotydy o sekwencjach podobnych do rRNA lub snRNA.

W pracy Suzuki i wsp. przedstawiono z kolei wyniki badania wpływu długości sond na ich specyficzność, a konkretnie na ich zdolność do wykrywania jednonukleotydowych niedopasowań (27). W tym celu autorzy zaprojektowali sztuczne 25-mery o losowych sekwencjach, a następnie dla tych sekwencji zostały zaprojektowane sondy całkowicie z nimi komplementarne o długościach od 14 do 25 nukleotydów oraz sondy zawierające po jednym niedopasowanym nukleotydzie występującym kolejno we wszystkich możliwych pozycjach. Na podstawie wyników eksperymentu hybrydyzacyjnego przeprowadzonego za pomocą stworzonej w ten sposób mikromacierzy wskazuje się, że optymalna długość sond ze względu na specyficzność wynosi od 19 do 21 nukleotydów. Warto zauważyć, że długość ta jest mniejsza niż stosowana w standardowych mikromacierzach o dużej gęstości. Ponadto w eksperymencie tym potwierdzono, że specyficzność sond maleje, jeżeli niedopasowany nukleotyd znajduje się blisko jednego z końców sondy.

Interesująca metoda selekcji sond zaproponowana została do projektowania mikromacierzy przeznaczonych do wykrywania organizmów zmodyfikowanych genetycznie (GMO) (28). Autorzy przyjęli założenie, że pojawienie się sygnału na mikromacierzy oznacza, iż badana próbka zawiera materiał pochodzący z GMO. Wszystkie sondy mają jednakową długość l. Metoda rozpoczyna działanie na zbiorze wszystkich oligonukleotydów o długości l i za pomocą pewnych biologicznych i kombinatorycznych zasad eliminacji zmniejsza liczbę oligonukleotydów do takiej, która odpowiada technicznym możliwościom konstrukcji mikromacierzy. Reguły eliminacji podzielone są na trzy grupy:

1. Usuwanie sond odpowiadających obu niciom genomu odniesienia, którego obecność jest spodziewana w badanej próbce (w przeciwieństwie do GMO, który jest nieznany). Celem zastosowania reguł z tej grupy jest ograniczenie liczby błędów pozytywnych.

2. Usuwanie sekwencji, które najprawdopodobniej nie są genetycznie funkcjonalne (np. hiperzmienne motywy mikrosatelitarne, długie fragmenty składające się z nukleotydów jednego rodzaju lub powtórzenia krótkich sekwencji). Sekwencje takie zazwyczaj nie są wynikiem zamierzonych modyfikacji genetycznych.

3. Usuwanie oligonukleotydów, które tworzą dupleksy o małej sile wiązania.

Jeżeli przez A, B, i C oznaczone zostaną zbiory sond określonych przez reguły 1, 2 i 3, to zbiorem sond wybranych przez opisywaną metodę jest $A \quad B \quad C$, gdzie przestrzenią, w której określone są te zbiory jest zbiór wszystkich sekwencji o długości l.

Reguły z pierwszej grupy oprócz eliminacji sekwencji dokładnie dopasowanych do genomu odniesienia usuwają ze zbioru potencjalnych sond również te oligonukleotydy, które mają pewną, określoną jako parametr, liczbę niedopasowań w stosunku do tego genomu. Jest to wskazane z kilku powodów, m. in. dlatego że sekwencje takie również mogą tworzyć dupleksy, a ponadto, ze względu na naturalną

różnorodność, nie wszystkie cząsteczki DNA pochodzące z organizmu odniesienia muszą mieć dokładnie taką samą sekwencję nukleotydową. Wreszcie, kompensowane są w ten sposób, przynajmniej do pewnego stopnia, błędy sekwencjonowania.

Reguły z drugiej grupy eliminują oligonukleotydy, które zawierają więcej niż 50% nukleotydów jednego rodzaju lub więcej niż trzy kolejne jednakowe dinukleotydy, bądź więcej niż 33% identycznych dinukleotydów w całej sekwencji.

Reguły z trzeciej grupy oparte są na empirycznie wyprowadzonych heurystykach podanych przez Affymetrix dla sond o długości 20 nukleotydów (11). Zgodnie z tymi regułami eliminowane są oligonukleotydy, które zawierają:

– więcej niż 9 nukleotydów A, 9 nukleotydów T, 9 nukleotydów C lub 9 nukleotydów G,

– w dowolnym podciągu o długości 8 nukleotydów więcej niż 7 nukleotydów A lub 7 nukleotydów T,

– w dowolnym podciągu o długości 8 nukleotydów więcej niż 6 nukleotydów C lub 6 nukleotydów G,

– podciąg o długości 6 nukleotydów składający się z nukleotydów C i G,

– podciąg o długości 7 nukleotydów składający się z nukleotydów A i T.

Ponadto, eliminowane są również oligonukleotydy, dla których połowa maksymalnej liczby komplementarnych par zasad między sekwencjami 5'-3' i 3'-5' jest większa od 6. Wyznaczana jest również temperatura topnienia potencjalnych sond, która dla całego projektowanego zbioru powinna znajdować się w jak najwęższym zakresie. Temperatura ta wyznaczana jest za pomocą modelu najbliższego sąsiada.

Wiele interesujących wyników biologicznych uzyskano w eksperymentach, w których wykorzystano mikromacierze zaprojektowane za pomocą programu ArrayOligoSelector (29,30). Program ten generuje zbiór sond dla wszystkich otwartych ramek odczytu i wymaga podania pełnej sekwencji genomowej badanego organizmu oraz sekwencji otwartych ramek odczytu, dla których mają zostać zaprojektowane sondy.

W pierwszym etapie w programie wykorzystuje się metodę BLAST lub BLAT do sprawdzenia lokalizacji poszczególnych ramek względem całego genomu. Algorytm BLAST jest bardziej dokładny, przez co generuje większy zbiór danych i w konsekwencji program działa wolniej. Metoda BLAT jest szybsza, jednak mniej dokładna, gdyż istnieje ryzyko pominięcia niektórych dopasowań.

W kolejnym etapie identyfikowane są oligonukleotydy o największej specyficzności. W tym celu dla każdej ramki znajdowane są sekwencje wykazujące najmniejszą specyficzność w obrębie pozostałej części genomu. Dla wszystkich potencjalnych rejonów hybrydyzacji, znalezionych wcześniej algorytmem BLAST lub BLAT, obliczana jest energia wiązania za pomocą metody najbliższego sąsiada, z uwzględnieniem niedopasowań dupleksów.

Następnie sekwencje sprawdzane są pod kątem tworzenia struktur drugorzędowych. Ze względu na długi czas obliczeń nie wykorzystano programu Mfold, lecz szybszą metodę bazującą na algorytmie Smitha-Watermana.

Kolejny etap to sprawdzanie zawartości par G-C, która jest głównym czynnikiem mającym wpływ na temperaturę topnienia dupleksu. Wykorzystywany jest tutaj próg określony przez użytkownika.

W ostatnim etapie dokonywany jest wybór najlepszego oligonukleotydu w obrębie danej ramki. Dla każdej z nich wykonywane są następujące kroki:

– wybierane są te oligonukleotydy, których energia wiązania jest najbliższa wartości zdefiniowanej przez użytkownika i nie przekracza progu odcięcia,

– opcjonalnie wystąpić może eliminacja zdefiniowanych przez użytkownika sekwencji niepożądanych, np. zawierających zbyt dużą liczbę par A-T,

– wybierane są oligonukleotydy, dla których złożoność sekwencji ma wartość mniejszą niż próg odcięcia oraz wynik badania możliwości powstania struktury drugorzędowej jest zadowalający; jeśli wszystkie sekwencje w ramach analizowanej ramki zostały odrzucone, progi odcięcia ulegają obniżeniu, a jeśli nadal żadna sekwencja nie zostanie wybrana obniżony zostaje próg zawartości par G-C poniżej wartości zdefiniowanej przez użytkownika,

– ostatnim parametrem jest bliskość sąsiedztwa końca 3' – wybierany jest oligonukleotyd leżący najbliżej końca 3' ramki i ten oligonukleotyd jest uznawany za najlepszy.

Program umożliwia również generowanie więcej niż jednej sondy dla każdej z ramek.

Opisane dotąd metody projektowania mikromacierzy oparte są m. in. na założeniu, zgodnie z którym sondy powinny być specyficzne dla odpowiednich genów. Jest to założenie ze wszech miar słuszne, jednak w praktyce może okazać się trudne, bądź wręcz niemożliwe do spełnienia. Stąd prowadzone są intensywne badania nad metodami projektowania zbiorów sond mniej specyficznych, jednak wybranych w taki sposób, że możliwe jest za ich pomocą jednoznaczne zidentyfikowanie analizowanych genów (31). Problem znalezienia takiego zbioru sond dla danego zbioru sekwencji docelowych (genów) można sformułować następująco. Niech dana będzie macierz $H=[h_{ij}]$, nazywana macierzą incydencji. W macierzy tej wiersze odpowiadają sekwencjom docelowym, natomiast kolumny odpowiadają potencjalnym sondom. Element h_{ij} równy jest 1 wtedy i tylko wtedy, gdy sonda j hybrydyzuje z sekwencją i. W przeciwnym przypadku element ten równy jest 0.

Mając daną macierz incydencji należy wybrać zbiór sond o jak najmniejszej liczności, taki, by za jego pomocą możliwe było jednoznaczne zidentyfikowanie dowolnej z sekwencji docelowych reprezentowanych przez wiersze tej macierzy. Generalnie jest to interesujący i złożony problem matematyczny. Jego rozwiązanie, nawet przybliżone, może zostać wykorzystane do skonstruowania efektywnych pod względem skuteczności działania oraz kosztów mikromacierzy DNA.

W tabeli przedstawiona jest przykładowa macierz incydencji. Występują w niej 4 sekwencje docelowe ($t_1 - t_4$) oraz 7 potencjalnych sond ($p_1 - p_7$). Z macierzy tej wynika m. in., że sonda p_3 hybrydyzuje z sekwencją t_2 (jedynka w komórce (2,3)), natomiast sonda p_5 z tą sekwencją nie hybrydyzuje (zero w komórce (2,5)).

Tabela

Przykładowa macierz incydencji

	p_1	p_2	p_3	p_4	p_5	p_6	p_7
t_1	1	0	1	1	0	1	0
t_2	0	1	1	1	0	0	0
t_3	1	1	0	1	1	0	1
t_4	0	0	1	0	0	1	1

Łatwo można zauważyć, że gdyby w badanej próbie mogła znaleźć się tylko jedna z sekwencji docelowych t_1 – t_4, do ich wykrycia wystarczyłyby tylko trzy sondy p_1, p_2 i p_3.

Istotnie, hybrydyzacja z sondami p_1 i p_3 oznaczałaby obecność w badanej próbie sekwencji t_1, hybrydyzacja z p_2 i p_3 oznaczałaby wykrycie sekwencji t_2, obecność sekwencji t_3 wykryta byłaby poprzez hybrydyzację z sondami p_1 i p_2, natomiast hybrydyzacja wyłącznie z sondą p_3 oznaczałaby obecność w próbie sekwencji t_4.

W podobny sposób wykryć można obecność w badanej próbie par sekwencji docelowych. Przykładowo, hybrydyzacja z sondami p_1, p_2, p_4, p_6, p_7 oznacza obecność sekwencji t_1 i t_4. Jeśli hybrydyzacja zachodzi ze wszystkimi sondami, w próbie obecne są sekwencje t_3 i t_4. Za pomocą przedstawionego w tabeli zestawu sond nie można jednak badać prób, w których mogą wystąpić trójki sekwencji docelowych, np. wystąpienie trójki t_1, t_2, t_3 spowodowałoby hybrydyzację ze wszystkimi sondami, czyli wynik identyczny z uzyskanym w przypadku obecności w roztworze sekwencji t_3 i t_4.

W pracy Meneses i wsp. opisano algorytm przybliżony, rozwiązujący przedstawiony problem wyboru sekwencji niespecyficznych oraz wynik jego zastosowania do projektowania sond dla sekwencji genomowej ludzkiego wirusa upośledzenia odporności (HIV) (32).

3. Podsumowanie

Mikromacierze DNA są nowoczesnym i bardzo efektywnym narzędziem służącym m. in. do badania ekspresji genów. Ich główną zaletą w porównaniu z innymi metodami służącymi do tego rodzaju badań jest wysoki stopień równoległości umożliwiający analizę ekspresji wielu, niekiedy nawet kilkudziesięciu tysięcy genów jednocześnie. Należy jednak pamiętać, że jakość wyników uzyskiwanych za pomocą mikromacierzy jest silnie uzależniona od sposobu ich zaprojektowania. Dwoma głównymi kryteriami przy projektowaniu mikromacierzy powinny być czułość i specyficzność. Kryteria te nie zawsze jest łatwo ze sobą pogodzić, stąd projektowanie mikro-

macierzy jest skomplikowanym problemem kombinatorycznym, do rozwiązania którego stosuje się metody informatyczne. Ponadto, oprócz sond specyficznych dla badanych genów, należy także uwzględnić odpowiednie sondy kontrolne: 1) negatywne, które nie powinny hybrydyzować z żadną sekwencją obecną w próbie biologicznej, oraz 2) pozytywne, czyli specyficzne dla określonych sekwencji zewnętrznych, dodawanych do próbki w znanym stężeniu, jeszcze przed procesem znakowania (ang. *spike controls*). W celu lepszej kontroli warunków hybrydyzacji często stosuje się także sondy o stopniowo obniżającym się stopniu komplementarności do określonej sekwencji docelowej (np. sonda w pełni komplementarna, sonda komplmentarna w 90, 80, 70% itd.). Sondy kontrolne powinny spełniać te same kryteria (długości, składu nukleotydowego czy temperatury topnienia), co zestaw sond służący do badania interesujących nas sekwencji.

W ciągu ostatnich lat powstało wiele pakietów oprogramowania wspomagających projektowanie mikromacierzy. Programy te rozwiązują (często w sposób przybliżony) problem projektowania odpowiedniego zestawu sond biorąc pod uwagę różne kryteria cząstkowe, których suma jest w praktyce przybliżeniem wspomnianych dwóch głównych kryteriów, tj. czułości i specyficzności. Ze względu na fakt, że problemy biologiczne rozwiązywane za pomocą mikromacierzy są bardzo różnorodne należy przy wyborze metody projektowania wziąć pod uwagę kryteria zawarte w tej metodzie i rozważyć, czy odpowiadają one rozwiązywanemu problemowi biologicznemu.

Opracowanie powstało w ramach realizacji projektu badawczego finansowanego przez Ministerstwo Nauki i Szkolnictwa Wyższego, nr PBZ-MNiI-2/1/2005.

Literatura

1. Formanowicz P., Handschuh L., Urbaniak R., Błażewicz J., Figlerowicz M., (2005), Na Pograniczu Chemii i Biologii, 12, 513-530.
2. Sambrook J., Russel D. W., (2001), *Molecular Cloning. A Laboratory Manual*, 3rd ed., 10.47-10.52, CSHL Press.
3. Suggs S. V., Hirose T., Miyake T., Kawashima E. H., Johnson M. J., Itakura K., Wallace R. B., (1981), *Developmental biology using purified genes*, Ed. Brown D. B., 683-693, Academic Press, New York.
4. SantaLucia Jr. J., (1998), Proc. Natl. Acad. Sci. USA, 95, 1460-1465.
5. Panjkovich A., Melo F., (2005), Bioinformatics, 21, 711-722.
6. SantaLucia Jr. J., Allawi H. T., Seneviratne P. A., (1996), Biochemistry, 35, 3555-3562.
7. Allawi H. T., SantaLucia Jr. J., (1997), Biochemistry, 36, 10581-10594.
8. Allawi H. T., SantaLucia Jr. J., (1998), Nucleic Acid Res., 26, 2694-2701.
9. Allawi H. T., SantaLucia Jr. J., (1998), Biochemistry, 37, 2170-2179.
10. Allawi H. T., SantaLucia Jr. J., (1998), Biochemistry, 37, 9435-9444.
11. Peyret N., Seneviratne P. A., Allawi H. T., SantaLucia Jr. J., (1999), Biochemistry, 38, 3468-3477.
12. Bommarito S., Peyret N., SantaLucia Jr. J., (2000), Nucleic Acid Res., 28, 1929-1934.
13. Shoemaker D. D., Linsley P. S., (2002), Curr. Opin. Microbiol., 5, 334-337.
14. Alon U., Barkai N., Notterman D. A., Gish K., Ybarra S., Mack D., Levine A. J., (1999), Proc. Natl. Acad. Sci. USA, 8, 96 (12), 6745-6750.

15. Zhu T., Wang X., (2000), Plant Physiol., 124, 1472-1476.
16. Li F., Stormo G. D., (2001), Bioinformatics, 17, 1067-1076.
17. Chou H.-H., Hsia A.-P., Mooney D. L., Schnable P. S., (2004), Bioinformatics, 20, 2893-2902.
18. Hyyrö H., Juhola M., Vihinen M., (2005), Nucleic Acid Res., 33, e115.
19. Levenstein V., (1966), Soviet Phys. Doklady, 10, 707-710.
20. Nordberg E. K., (2005), Bioinformatics, 21, 1365-1370.
21. Kaderali L., Schliep A., (2002), Bioinformatics, 18, 1340-1349.
22. Hu G., Llinás M., Li J., Preiser P. R., Bozdech Z., (2007), BMC Bioinformatics, 8, 350.
23. Rimour S., Hill D., Militon C., Peyret P., (2005), Bioinformatics, 21, 1094-1103.
24. Rouillard J.-M., Herbert C. J., Zuker M., (2002), Bioinformatics, 18, 486-487.
25. Zuker M., Mathews D. H., Turner D. H., (1999), *Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide*, NATO ASI Series, Kluwer, Dordrecht.
26. Wang X., Seed B., (2003), Bioinformatics, 19, 796-802.
27. Suzuki S., Ono N., Furusawa C., Kashiwagi A., Yomo T., (2007), BMC Genomics, 8, 373.
28. Nesvold H., Kristoffersen A. B., Holst-Jensen A., Berdal K. G., (2005), Bioinformatics, 21, 1917-1926.
29. Zhu J., (2006), *The application of functional genomics, systems biology and drug development to the study of infectious disease*, Ph.D. thesis, University of California San Francisco.
30. ArrayOligoSelector **http://derisilab.ucsf.edu/index.php?software=46**
31. Du D. H. Z., Hwang F. K., (2006), Pooling Designs and Nonadaptive Group Testing, World Scientific, Singapore.
32. Meneses C. N., Pardalos P. M., Ragle M. A., (2007), Ann. Biomed. Eng., 35, 651-658.