Mikromacierze DNA – analiza danych

Piotr Stępniak[1], Luiza Handschuh[1,2], Marek Figlerowicz[1]

[1] Centrum Doskonałości CENAT, Instytut Chemii Bioorganicznej, Polska Akademia Nauk, Poznań
[2] Katedra i Klinika Hematologii i Chorób Rozrostowych Krwi, Uniwersytet Medyczny im. K. Marcinkowskiego, Poznań

DNA microarray data anlysis

Summary

The paper gives an overview of common methods applied in microarray data analysis. High density oligonucleotide and low density home made microarray types are being considered. Presented exploration procedures follow preprocessing and higher analysis steps, including example methods. Describing higher analysis algorithms we focus on implementation of pattern search and machine learning approaches.

Key words:
DNA microarrays, microarray data analysis, background correction, normalization, summarization, filtration, clustering, support vector machines.

1. Wstęp

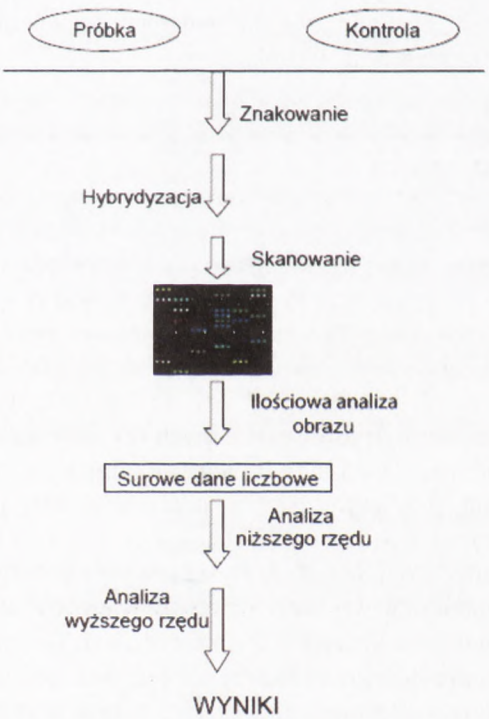
Mikromacierze stanowią szczególnie interesujące narzędzie współczesnej biologii molekularnej, nie tylko ze względu na szerokie spektrum zastosowań (analiza struktury genomu, profilu ekspresji genów, genotypowanie, sekwencjonowanie), ale i z uwagi na możliwość badania dużej liczby obiektów w jednym eksperymencie. Jednakże wyłonienie istotnych informacji z ogromnej ilości danych uzyskiwanych przy użyciu mikromacierzy wymaga zastosowania wyrafinowanych metod bioinformatycznych. W artykule zaprezentowano próbę przybliżenia podstaw tego zagadnienia i omówiono wybrane metody analityczne.

Adres do korespondencji

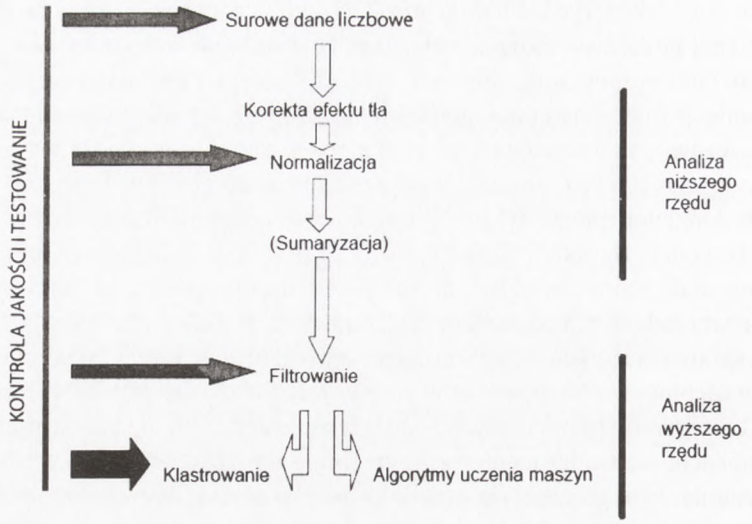
Piotr Stępniak,
Instytut Chemii Bioorganicznej,
Polska Akademia Nauk,
ul. Noskowskiego 12/14,
61-704 Poznań;
e-mail:
piotrek.stepniak@gmail.com

biotechnologia
4 (83) 68–87 2008

Główne etapy eksperymentu mikromacierzowego zaprezentowane zostały na rysunku 1. Na tej podstawie można stwierdzić, że wynikiem końcowym tzw. „mokrej" części eksperymentu jest mikromacierz, z którą związana została wyznakowana fluorescencyjnie próba. Następnie płytkę taką poddaje się skanowaniu za pomocą czytnika laserowego i uzyskuje obraz ukazujący, z jaką intensywnością świecą punkty (ang. *spots*) zawierające sondy specyficzne dla poszczególnych genów. W kolejnym etapie każdemu punktowi przyporządkowana zostaje liczba określająca natężenie fluorescencji. Uzyskane w ten sposób „surowe" dane liczbowe poddawane są najpierw normalizacji lokalnej (w obrębie pojedynczej płytki), a następnie globalnej (w obrębie wszystkich mikromacierzy składających się na eksperyment). Na każdym z etapów sprawdza się jakość mikromacierzy wykluczając takie, które mają poważne defekty techniczne. Następny etap zwany jest filtrowaniem genów i ma na celu wybranie tzw. genów różnicujących, których ekspresja zmienia się istotnie w badanych warunkach, oraz odrzucenie tych, które nie dają żadnego sygnału. Na podstawie tak zredukowanego zestawu genów prowadzi się analizy wyższego rzędu, poszukując grup genów o podobnym/odmiennym profilu ekspresji. Na zakończenie uzyskane wyniki poddawane są interpretacji biologicznej, polegającej na powiąza-



Rys. 1. Schematyczny opis eksperymentu mikromacierzowego.



Rys. 2. Schematyczny opis procesu analizy danych.

niu obserwowanych zmian w poziomie ekspresji genów z fizjologicznymi bądź patologicznymi procesami zachodzącymi w badanym organizmie. Przebieg analizy przedstawiono schematycznie na rysunku 2.

2. Ilościowa analiza obrazu

W przypadku produkowanych komercyjnie mikromacierzy, np. wysokiej gęstości chipów firmy Affymetrix, etapowi odczytu obrazu poświęca się mało uwagi, gdyż sprowadza się on do zliczenia intensywności obserwowanych punktów w sektorze zawierającym konkretną sondę. Proces ten przebiega w pełni automatycznie. Trochę więcej uwagi wymagają mikromacierze drukowane, gdyż często konieczne jest dopasowanie siatki rozmieszczenia punktów do ich rzeczywistego położenia na szkiełku. Istnieje kilka algorytmów stosowanych do odczytu obrazu. Podstawowe algorytmy to: niezmiennego koła (ang. *fixed circle*), dopasowanego koła (ang. *adaptive circle*) i histogramu, przy czym ostatnie dwa uwzględniają różnice w wielkości punktów. Interesującym rozwiązaniem jest też algorytm *seed region growing*, który obrysowuje każdy punkt z osobna. Niestety żaden nie jest odporny na błędy, które trzeba usuwać manualnie. Każdej kropce przypisywany jest status (ang. *good, bad, absent, found, not found*) zakodowany w formie cyfrowej. Ten proces nosi nazwę flagowania. Po zakończeniu ilościowej analizy obrazu (ang. *quantitation*) otrzymujemy plik tekstowy z szeregiem informacji zawartych w tabeli, w której w wierszach znajdują się informacje dla poszczególnych sond, a w kolumnach odpowiednie wartości.

Intensywność sygnału jest zwykle podawana jako mediana lub średnia z punktów dla danej sondy. Wśród pozostałych danych znajdują się m.in. położenie i identyfikator (nazwa sondy), ID punktu, intensywność tła (w postaci mediany i/lub średniej), wartości odchyleń standardowych intensywności punktów w obrębie sondy i status kropki. Po wykonaniu korekty tła można przystąpić do analizy niższego rzędu, czyli normalizacji danych. Część algorytmów normalizacyjnych dodatkowo wstępnie dokonuje korekty tła korzystając z surowych danych (ang. *probe-level data*).

3. Wstępna obróbka danych

Normalizacja danych umożliwia porównanie wyników zarówno w obrębie jednej mikromacierzy, jaki i pomiędzy mikromacierzami. Głównym celem normalizacji jest niwelacja tej części sygnału, która jest efektem niedoskonałości technicznych takich jak nierównomierne odmycie poszczególnych regionów macierzy, różnice w natężeniu sygnału emitowanego przez zastosowane barwniki fluorescencyjne czy różnice w wydajności znakowania kolejnych próbek. Wykrywanie błędów technicznych ułatwia sama konstrukcja mikromacierzy, zakładająca umieszczenie kilku–kilkunastu powtórzeń tej samej sondy w różnych miejscach oraz zastosowanie zestawu specjalnych sond kontrolnych (ang. *spikes*), które nie powinny hybrydyzować z badanym materiałem, a jedynie z komplementarnymi sekwencjami dodanymi w odpowiedniej ilości na etapie znakowania (kontrole zewnętrzne). Sondy te są bardzo użyteczne podczas normalizacji w obrębie wielu mikromacierzy, mając przewagę nad sondami specyficznymi dla genów o ekspresji konstytutywnej (ang. *housekeeping genes*), które jak dowiedziono nie wykazują absolutnie stałego poziomu transkrypcji (1).

Zabiegi przygotowujące dane z mikromacierzy do właściwej analizy bioinformatycznej, nazywa się wstępną obróbką danych (ang. *preprocessing*). Jej przebieg zależy w głównej mierze od sposobu, w jaki analizowana macierz została skonstruowana, tzn. czy jest to macierz o wysokiej gęstości czy drukowana.

3.1. Wstępna obróbka danych uzyskiwanych przy zastosowaniu oligonukleotydowych mikromacierzy DNA o wysokiej gęstości

Ze względu na fakt, że mikromacierze wysokiej gęstości cieszą się wciąż największą popularnością, dostępnych jest wiele gotowych programów przeznaczonych do obróbki wstępnej danych. Zwykle składa się ona z trzech etapów: korekty tła (ang. *background adjustment*), normalizacji i sumaryzacji (ang. *summarisation*). Korzystając z poszczególnych metod pamiętać należy, że często w znaczący sposób wpływają one na ostateczny rezultat analizy.

Mikromacierze firmy Affymetrix (2) złożone są z krótkich, 25-merowowych sond oligonukleotydowych, zorganizowanych w zespoły 11-20 par komplementarnych do

różnych regionów tego samego transkryptu, co zapewnia wyższą czułość detekcji sygnału. Każda para składa się z sondy w pełni komplementarnej (PM, ang. *perfect match*) oraz posiadającej 1 niekomplementarny nukleotyd w pozycji 13 (MM, ang. *mismatch*), co ma służyć podwyższeniu specyficzności sygnału po hybrydyzacji oraz określeniu wartości tła. Ze względu na gęste upakowanie sond na chipie bezpośredni pomiar intensywności tła płytki jest niemożliwy. Teoretycznie właściwy sygnał powinno się zatem otrzymać po odjęciu wartości sygnału MM od wartości PM. Jednak już przy wprowadzaniu korekty tła pojawia się problem, ponieważ w praktyce ok. 30% sond MM ma wyższą wartość sygnału niż odpowiadająca im sonda PM (3). W efekcie generowane są ujemne wartości intensywności sygnału, co nie tylko nie ma sensu merytorycznego, ale również uniemożliwia stosowanie funkcji logarytmicznych w dalszej części analizy.

Do najpopularniejszych metod korekty tła należą MAS 5.0 (nazwa pochodzi od skrótu programu ją implementującego – *Microarray Suite 5.0*, Affymetrix, 2002) oraz RMA (ang. *Robust Multi-array Analysis*, (4)).

W przypadku MAS 5.0 chip dzielony jest na k regionów (domyślnie 16), po czym dla każdego z nich 2% najniższej intensywności sygnału jest używane do wyliczenia tła. Następnie do korekty sygnału sondy używa się średniej ważonej wszystkich wartości tła, gdzie waga zależy od odległości sondy od centroidu regionu. Dodatkowo algorytm zapobiega powstaniu ujemnych wartości intensywności oraz w regionach o niskiej intensywności zmniejsza jej negatywny efekt w stosunku do całej mikromacierzy.

W metodzie RMA korekta tła odbywa się jedynie na podstawie wartości intensywności przypisanych sondom PM, na bazie globalnego modelu rozkładu ich intensywności. Wartość intensywności sygnału jest modelowana jako suma składnika tła (o przyjętym rozkładzie wg krzywej Gaussa), z uwzględnieniem jego średniej i odchylenia standardowego oraz składnika sygnału w funkcji ekspotencjalnej z uwzględnieniem jego średniej, a także z wykorzystaniem funkcji rozkładu normalnego i gęstości. Aby uniknąć ujemnych wartości wykorzystywana jest tylko dodatnia część rozkładu normalnego.

Kolejnym etapem wstępnej obróbki danych jest normalizacja, która polega na takim przekształceniu danych, aby można było porównywać ich wartości pomiędzy eksperymentami (mikromacierzami). Podstawową metodą proponowaną przez Affymetrix jest skalowanie. Spośród wszystkich mikromacierzy wybiera się jedną, która posłuży jako podstawa normalizacji (wzorzec). W przypadku pozostałych macierzy intensywność wszystkich sygnałów zostaje proporcjonalnie zwiększona/zmniejszona tak, aby jej średnia wartość była identyczna z obliczoną dla wzorca. W modyfikacji metody podczas obliczania średniej odrzuca się po 2% najsilniejszych i najsłabszych sygnałów. Affymetrix zaleca przeprowadzać skalowanie po obliczeniu wartości ekspresji dla zestawu sond specyficznych dla danego genu (ang. *expression values*), czyli po etapie sumaryzacji, ale można tę procedurę zastosować również na surowych danych (ang. *probe-level data*).

Innym popularnym algorytmem normalizacji jest normalizacja kwantylowa (ang. *quantile normalization*) (5). W tej metodzie głównym celem jest zachowanie rozkładu intensywności sygnałów na każdej macierzy. Transformacji dokonuje się na bazie uzyskanych w doświadczeniu funkcji rozkładu sygnałów na mikromacierzy oraz rozkładu średnich odległości między tymi sygnałami, tzw. kwantyli (ang. *quantiles*).

Interesującą metodą, łączącą w jeden etap korektę tła i normalizację, jest vsn (ang. *Variance stabilization and calibration for microarray data*) (6). Jej zaletą jest to, że oblicza wartości korekty tła na podstawie informacji z wielu macierzy, a nie tylko z jednej. Podczas normalizacji stosuje się tzw. uogólniony lub wyciszony logarytm (ang. *glog*) (7), dzięki któremu przy zastosowaniu odpowiednich parametrów skalowania macierzy oraz korekty tła można dopasować do siebie mikromacierze zachowując przy tym niezależność wariancji powtórzeń od ich średniej.

Końcową fazę wstępnej obróbki danych z chipów genomowych stanowi sumowanie wartości intensywności sygnałów pochodzących od zestawu sond (ang. *probe-level data*) specyficznych dla pojedynczego transkryptu w celu obliczenia wartości ekspresji (ang. *expression value*). Metody sumowania można podzielić na dwie grupy: prowadzone w obrębie jednej mikromacierzy oraz prowadzone w oparciu na ich zestawie. W obrębie jednej mikromacierzy może to być średnia arytmetyczna logarytmów o podstawie 2 z wartości sygnałów dla zestawu sond, logarytm naturalny średniej wartości sygnałów, mediana wartości przedstawionych w skali logarytmicznej, logarytm naturalny mediany obliczonej dla całego zestawu sygnałów czy logarytm o podstawie 2 z wartości drugiego najmocniejszego sygnału. W metodach obejmujących wiele mikromacierzy dominuje podejście polegające na tworzeniu modeli opartych na skali logarytmicznej liniowej lub na algorytmie „wygładzania" mediany (ang. *median polish*) (8).

Warto zaznaczyć, że dostępne narzędzia służące analizie danych mikromacierzowych przeważnie łączą wymienione trzy etapy obróbki wstępnej. Wspomniana popularna procedura RMA obejmuje korektę tła, normalizację kwantylową i oparte na zestawie macierzy sumowanie z wykorzystaniem algorytmu wygładzania mediany. Ponieważ sekwencje sond na chipie są znane zaproponowano również metodę uwzględniającą wpływ sekwencji sondy na powstawanie szumu tła. Metoda GCRMA (9) stanowiąca rozwinięcie metody RMA, wykorzystuje podczas korekty tła obliczoną na podstawie sekwencji sond wartość powinowactwa, a następnie stosuje algorytmy normalizacji i sumowania typowe dla RMA.

3.2. Wstępna obróbka danych uzyskiwanych przy zastosowaniu mikromacierzy drukowanych

Mikromacierze drukowane składają się z sond cDNA lub długich sond oligonukleotydowych (najczęściej 50-70 nt), a gęstość ich upakowania na powierzchni płytki zależy od możliwości drukarki oraz właściwości fizycznych buforów użytych do drukowania. Eksperyment z zastosowaniem takiej mikromacierzy zakłada jednocze-

sną hybrydyzację dwóch próbek, badanej i kontrolnej. Każda z prób wyznakowana jest innym barwnikiem fluorescencyjnym, np. cyjaniną 3 i cyjaniną 5 (Cy3 i Cy5).

Dzięki większym odległościom pomiędzy sondami możliwy jest bezpośredni pomiar tła. Korekta tła dla każdej sondy odbywa się poprzez odjęcie wartości intensywności tła od sygnału. Ten etap jest wykonywany automatycznie w przypadku większości algorytmów normalizujących jako jedna ze wstępnych czynności.

Możemy wyróżnić dwa podejścia do normalizacji danych z dwukolorowych mikromacierzy: normalizację dwukanałową (ang. *two-channel normalization*) i normalizację jednokanałową (ang. *separate-channel*). Algorytmy normalizujące dane z dwukolorowych mikromacierzy wykorzystują jako dane wejściowe zintegrowane wartości intensywności z dwóch kanałów jednocześnie, najczęściej na jeden z dwóch sposobów:

[1] $$M = \log_2\left(\frac{Cy5}{Cy3}\right) = \log_2(Cy5) - \log_2(Cy3)$$

lub

[2] $$A = \log_2\sqrt{Cy5 \times Cy3} = \frac{1}{2}(\log_2(Cy5) + \log_2(Cy3))$$

gdzie Cy5 i Cy3 oznaczają intensywność świecenia sondy zmierzoną odpowiednio dla cyjaniny 5 i cyjaniny 3. M jest skrótem od angielskiego słowa *minus*, oznacza bowiem różnicę logarytmów intensywności pochodzących od dwóch kanałów, podczas gdy A pochodzi od słowa *add* (dodać), bowiem w tym przypadku oblicza się połowę z sumy intensywności dla obu barwników (10).

Normalizację dwukanałową można podzielić na dwa etapy – pierwszy uwzględnia położenie, drugi natomiast skalę (11). W niektórych podejściach pomija się faktyczną lokalizację sondy na płytce, np. w metodzie mediany globalnej (ang. *global median normalization*), gdzie lokalizacja jest uznawana za tożsamą w obrębie całej macierzy. Czasem też sondy będące w bliskim sąsiedztwie na mikromacierzy lub drukowane tą samą igłą grupuje się w tzw. zespoły normalizacyjne. Z sytuacją taką mamy do czynienia w podejściu opartym na wartości *A* [2], gdzie funkcją wygładzającą (ang. *smooth function*) *A* jest lokalnie ważona regresja liniowa (LOESS albo LOWESS, ang. *locally weighted scatterplot smoothing*). Tę samą funkcję zastosować można w obrębie grup sond drukowanych tą samą igłą (ang. *print-tip A-dependent loess normalization*), w celu zatarcia różnic wynikających z nierównomiernego drukowania.

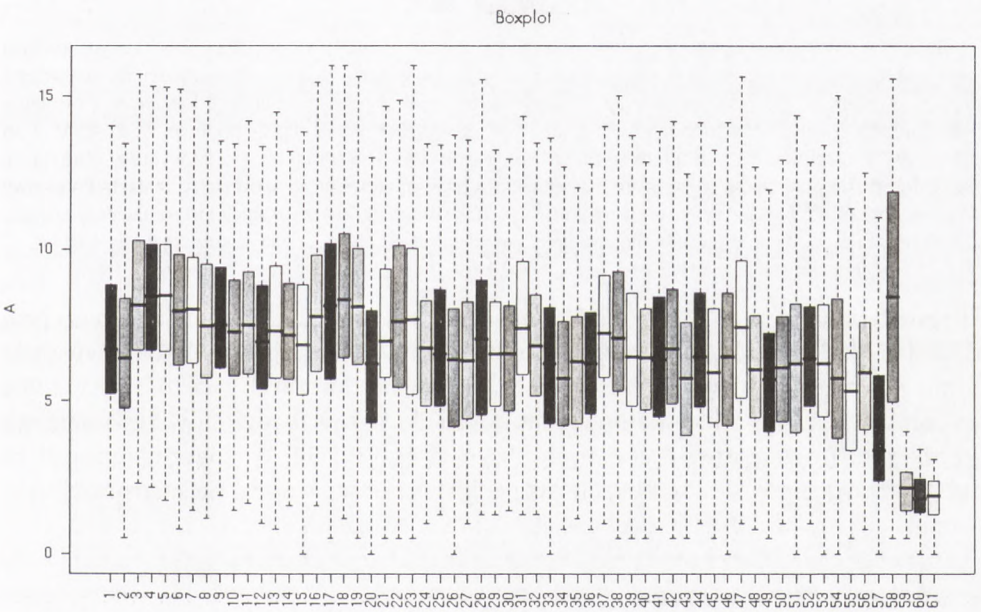
Zastosowanie parametru lokalizacji powoduje rozmieszczenie wartości A w funkcji M w okolicy zera, nie niweluje to jednak różnic w skali pomiędzy mikromacierzami, co jest możliwe dzięki użyciu parametru skali. Kiedy jednak różnice w skali są nieduże, wtedy korzyść ze skalowania jest niewielka w porównaniu z dodatkową zmiennością wprowadzaną przez samą procedurę.

W przypadku normalizacji traktującej każdy kanał mikromacierzy (wyznakowany innym barwnikiem) oddzielnie wyróżnia się dwa etapy procedury: normalizację w obrębie płytki i normalizację zestawu płytek. Algorytmy wykonujące ten typ normalizacji działają przeważnie na surowych danych, wykorzystując metody znane

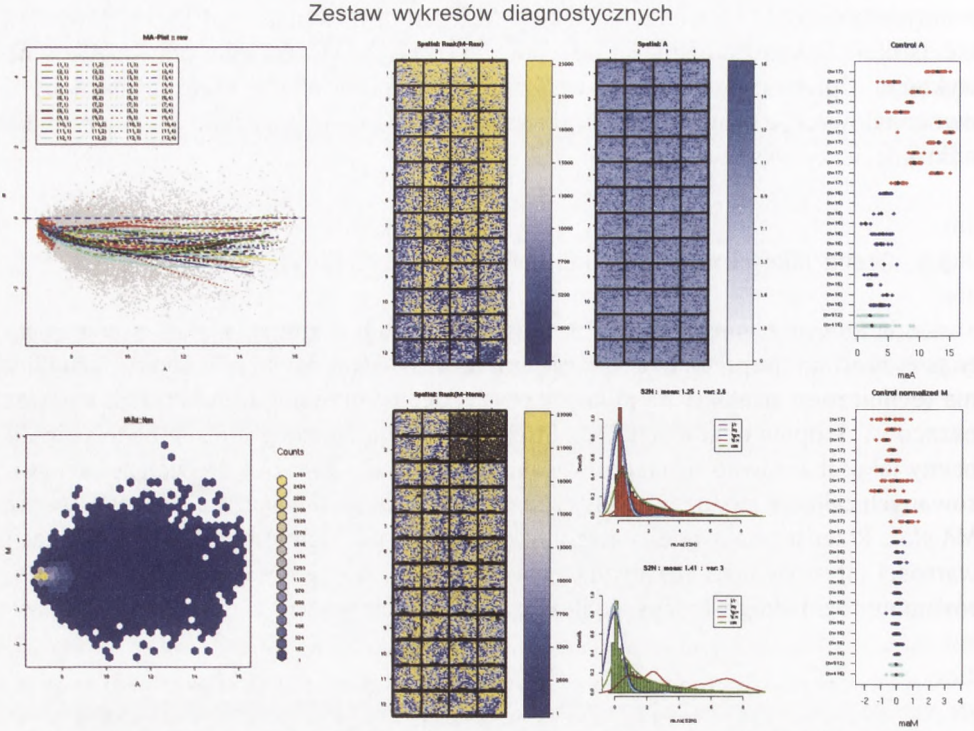
z normalizacji mikromacierzy wysokiej gęstości, np. normalizację kwantylową czy *vsn*. Istnieje również możliwość zastosowania podejścia łączonego, polegającego na wykonaniu pierwszego etapu normalizacji (w obrębie płytki) metodą dla danych dwukanałowych, a drugiego (w obrębie zestawu płytek) jedną z metod typowych dla mikromacierzy wysokiej gęstości.

3.3. Ocena jakości w eksperymentach mikromacierzowych

Dodatkowym elementem analizy niższego rzędu jest kontrola jakości (ang. *quality assesment*) np. poprzez wizualizację graficzną. Można dzięki niej wykryć uchybienia techniczne i artefakty na mikromacierzy, ocenić przebieg normalizacji, a nawet oszacować stopień degradacji RNA. Stosowane tu narzędzia graficzne nadają się do oceny jakości zarówno macierzy wysokiej jak i niskiej gęstości. Do najczęściej stosowanych zaliczyć można wykresy typu *log-scale plot, spatial-plot, boxplot* (rys. 3) oraz *MA-plot*. Przedstawiają one obraz mikromacierzy po logarytmicznej transformacji wartości intensywności (co uwidacznia artefakty i problemy techniczne) oraz porównanie rozkładu gęstości sygnałów pomiędzy mikromacierzami. Zestawienie wy-



Rys. 3. Przykładowy wykres boxplot surowych danych z eksperymentów na mikromacierzach dwukolorowych. Każdy słupek odpowiada jednej mikromacierzy w zestawie. Pojedynczy słupek składa się z kreski – wartości minimalnej, dolnego kwartylu (odcina dolne 25% punktów), mediany, górnego kwartylu, wartości maksymalnej, oraz obserwacji odstających (ang. *outliers*) (w postaci kółek – w tym przypadku brak). Zamalowany obszar znajduje się pomiędzy górnym i dolnym kwartylem.



Rys. 4. Przykładowy zestaw wykresów diagnostycznych przed i po normalizacji. W lewym górnym rogu wykres *scatterplot* typu M A dla surowych danych. Linie oznaczają trendy wartości intensywności punktów w danym bloku mikromacierzy. Poniżej wykres typu M A po normalizacji loess, kolor obrazuje ilość punktów o danych wartościach. W środku na górze wykresy *spatial plot* – po lewej wartości M, a po prawej A dla surowych danych. Poniżej *spatial plot* wartości M po normalizacji oraz wykresy obrazujące rozkład gęstości punktów o różnych intensywnościach (tzw. *density plot*) w osobnych kanałach. Po prawej wykresy obrazujące rozkład wartości sond kontrolnych w stosunku do pozostałych mikromacierzy w analizowanym zbiorze. Wykresy wykonano dla tych samych danych co wcześniej przedstawiony *boxplot*.

kresów typów *boxplot* i *MA* przed i po normalizacji obrazuje efektywność tego procesu (rys. 4). Analiza stopnia degradacji RNA na mikromacierzach wysokiej gęstości firmy Affymetrix opiera się na porównaniu logarytmicznych wartości sygnałów w obrębie zestawu sond, komplementarnych do różnych regionów tego samego transkryptu (fragmentów położonych blisko końca 3' i 5'). Wyznacznikiem jest tu stosunek sygnałów 3'/5' w obrębie wszystkich zestawów sond na mikromacierzy.

4. Analiza wyższego rzędu

Analiza wyższego rzędu skupia się na najważniejszym z biologicznego punktu widzenia problemie, a mianowicie poszukiwaniu biologicznych zależności w danych uzyskanych z eksperymentu mikromacierzowego. Większość metod wykorzystywa-

nych na tym etapie opiera się na klasyfikacji bądź grupowaniu zarówno genów jak i próbek w zespoły, wykazujące wspólne cechy, np. profil ekspresji, regulacja przez ten sam czynnik, związek z procesem chorobowym, itd. Pierwszym etapem analizy wyższego rzędu jest filtracja, czyli wykluczanie genów, których poziom ekspresji lub jego zmianę uznamy za zbyt niski. Ułatwia to dalszą analizę, gdyż prowadzi się ją na zredukowanym zestawie danych, a także chroni przed nadmiarem wyników fałszywie pozytywnych. Następnie za pomocą specjalnych funkcji matematycznych określa się bliskość lub podobieństwo danych genów i próbek. Ważnym elementem analizy wyższego rzędu jest konfrontacja otrzymanych wyników z informacjami zdeponowanymi w internetowych bazach danych. Na przykład sprawdzenie ontologii genów (GO, ang. *Gene Ontology*, www.geneontology.org) może potwierdzić, że otrzymane zależności mają swoje źródło biologiczne, a nie są wyłącznie pochodzenia matematyczno-statystycznego.

4.1. Filtracja

Przeciętna mikromacierz może zawierać sondy identyfikujące od kilkuset do kilkudziesięciu tysięcy transkryptów. Ponieważ nie wszystkie geny ulegają w danym momencie transkrypcji od części sond nie otrzymamy żadnego sygnału, a dla części sygnały bardzo słabe. Jednym z zadań filtracji jest odsianie takich właśnie sond. Ponadto porównując próbkę badaną z kontrolną, stwierdzić można, że zmianę poziomu ekspresji wykazuje jeszcze mniejsza liczba genów. Aby „wyłowić" interesujące geny z całej puli stosuje się wiele metod filtracji i oceny statystycznej uzyskanych danych. Najprostszy sposób selekcji sond, stosowany w mikromacierzach dwukolorowych, polega na wprowadzeniu progów zmiany poziomu ekspresji. Ocenia się wówczas stosunek intensywności sygnału próbki względem kontroli przedstawiony w skali logarytmicznej. Wartości zazwyczaj powyżej 1,75 lub 2 (oznaczające przyrost ekspresji 1,75 lub 2-krotny) oraz poniżej 0,5 lub 0,75 (oznaczające spadek ekspresji o połowę lub ¾) pozostają jako istotne, a reszta sond jest wykluczana z dalszej analizy. Takie podejście, choć dość skuteczne, nie jest odporne na błędy, zarówno typu I jak i II. Błędy typu I wiążą się z ryzykiem uzyskania wyników fałszywie pozytywnych (ang. *flase positives*) poprzez pozostawienie genów, które faktycznie nie wykazują zmienionej ekspresji, natomiast błędy typu II generują wyniki fałszywie negatywne (ang. *false negative*) poprzez wykluczenie z analizy genów wykazujących w rzeczywistości zmianę poziomu ekspresji. Aby rozwiązać ten problem w obu typach mikromacierzy (jedno- i dwukolorowych) stosuje się szereg testów statystycznych. Najpopularniejszy test T wylicza tzw. „wartość p" (ang. *p-value*), która ocenia prawdopodobieństwo uzyskania takiego samego lub „lepszego" wyniku losowego. Zwykle za podstawowy punkt odcięcia przyjmuje się $p < 0,05$, co oznacza, że istnieje mniej niż 5% szans, że obserwowany wynik może być przypadkowy. Niestety w analizie 10 000 genów oznacza to, że istnieje możliwość błędnej klasyfi-

kacji aż 500 z nich. Dlatego coraz częściej stosuje się zaawansowane metody statystyczne: analizę wariancji (ANOVA, ang. *analysis of variance*), wieloczynnikową analizę wariancji (MANOVA, ang. *multifactor analysis of variance*) czy wieloetapowe procedury testowania (MTP, ang. *Multiple Testing Procedures*) (12,13).

4.2. Obliczanie odległości

Wartości ekspresji genów są podawane, albo w bezwzględnej wartości intensywności (zlogarytmizowanej), albo relatywnego poziomu ekspresji względem próbki kontrolnej (iloraz logarytmów). Bezwzględna wartość intensywności obliczana jest dla danych pochodzących z mikromacierzy typu Affymetrix, gdzie tylko jedna próbka jest hybrydyzowana z mikromacierzą. Natężenie sygnału jest wówczas uważane za liniowo zależne od ilości mRNA. Okazuje się jednak, że dwa różne geny bardzo często mają różny rozkład sygnału, innymi słowy dwa różne mRNA występujące w tym samym stężeniu mogą dawać inny poziom intensywności sygnału, co sprawia, że porównanie pomiędzy genami w obrębie jednej próbki jest problematyczne. Nie przeszkadza to natomiast w porównywaniu poziomu ekspresji danego genu pomiędzy próbkami. W przypadku macierzy dwukolorowej określanie relatywnego poziomu ekspresji względem kontroli pozwala porównywać zmiany ekspresji różnych genów w obrębie pojedynczej mikromacierzy. Nie można natomiast określić bezwzględnego poziomu ekspresji genów, a tym samym oszacować rzeczywistej ilości mRNA dla danego genu. Oczywiście w przypadku mikromacierzy typu Affymetrix można także uzyskać relatywną wartość poziomu ekspresji genów – poprzez porównanie poszczególnych macierzy z jedną mikromacierzą, do której hybyrdyzowano próbkę kontrolną. Nie jest to jednak pomiar bezpośredni jak w mikromacierzach dwukolorowych, gdzie zarówno próba jak i kontrola hybrydyzują razem na tej samej płytce w identycznych warunkach.

Użyta skala definiująca zakres poziomu ekspresji ma bezpośredni wpływ na wartość funkcji określających parametr zwany odległością między genami i próbkami. Funkcja określająca odległość to taka, która spełnia warunki nieujemności, symetrii i identyfikacji (warunki 1-3, tab.). Funkcję spełniającą również warunki 4 i 5 nazywa się funkcją metryczną (ang. *metric*). W niektórych przypadkach mówi się również o funkcji podobieństwa (ang. *similarity*), która spełnia jedynie warunki nieujemności i symetrii, a jej wartość rośnie wraz ze wzrostem podobieństwa dwóch genów/próbek.

Tabela

Warunki spełniane przez funkcje odległości pomiędzy genami i/lub próbkami w eksperymencie mikromacierzowym

Warunek	Wzór
nieujemność	$d(x, y) \geq 0$
symetria	$d(x, y) = d(y, x)$
identyfikacja	$d(x, x) = 0$
jednoznaczność	$d(x, y) = 0 \Leftrightarrow x = y$
warunek trójkąta	$d(x, y) + d(y, z) \geq d(x, z)$

Funkcja spełniająca wszystkie 5 warunków nazywana jest funkcją metryczną (*metric*), warunki 1-3 – funkcją odległości (*distance*), funkcje podobieństwa (*similarity*) i niepodobieństwa (*dissimilarity*) spełniają jedynie warunki 1-2.

W eksperymentach mikromacierzowych mamy do czynienia z sytuacją, w której mierzymy wartości *m* cech (genów) w *n* przypadkach (próbkach, mikromacierzach). Dlatego potrzebujemy metod pozwalających określić podobieństwo zarówno pomiędzy genami (wykazującymi wspólny profil ekspresji) jak i pomiędzy próbkami (np. pacjentami o takiej samej chorobie). Istnieją dwa główne sposoby mierzenia odległości w eksperymentach mikromacierzowych. Pierwszy, znacznie częściej stosowany, przedstawia wartości ekspresji z różnych próbek dla dwóch wybranych genów jako wektory w przestrzeni wielowymiarowej i oblicza ich odległość parami w obrębie pojedynczej próbki. Drugi natomiast traktuje dane wartości jako obserwowane przypadki z funkcji losowych dystrybucji i gęstości rozkładu pomiarów ekspresji.

Funkcje wyliczające odległość parami mają ogólną postać:

[3] $$d(x, y) = F[d_1(x_1, y_1), \ldots, d_m(x_m, y_m)]$$

Odległość *d* pomiędzy genami *x* i *y* (traktowana jako wektor) jest ogólną miarą odległości pomiędzy nimi, wyliczoną na podstawie wartości ekspresji (czyli współrzędnych wektora) w każdej z *m* próbek, z zastrzeżeniem, że poszczególne d_k nie muszą być traktowane jednakowo, a mogą na przykład posiadać różne wagi. Dwie podstawowe funkcje stosowane do obliczenia odległości wykorzystują w obliczeniach wartość bezwzględną różnicy wartości ekspresji genów w poszczególnych próbkach. Są to odległość euklidesowa i Manhattan, które przedstawiono we wzorach:

[4] $$d_{eku}(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

[5] $$d_{man}(x,y) = \sum_{i=1}^{m} |x_i - y_i|$$

Inne mierniki odległości korzystają z pojęć opisujących siłę zależności pomiędzy wektorami, czyli korelację. Należą do nich:

– miernik odległości oparty na współczynniku korelacji Pearsona:

[6] $$d_{cor}(x,y) = 1 - r(x,y) = 1 - \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

– miernik odległości oparty na współczynniku korelacji próbki Spearmana:

[7] $$d_{spear}(x,y) = 1 - \frac{\sum_{i=1}^{m}(x'_i - \bar{x}')(y'_i - \bar{y}')}{\sqrt{\sum_{i=1}^{m}(x'_i - \bar{x}')^2 \sum_{i=1}^{m}(y'_i - \bar{y}')^2}}$$

gdzie x'_i = ranga (x_i) i y'_i = ranga (y_i)

– miernik odległości oparty na współczynniku korelacji próbki tau Kendalla:

[8] $$d_{tau}(x,y) = 1 - \tau(x,y) = 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} C_{x_{ij}} C_{y_{ij}}}{m(m-1)}$$

gdzie $C_{x_{ij}} = \text{sign}(x_i - x_j)$ i $C_{y_{ij}} = \text{sign}(y_i - y_j)$

Wszystkie funkcje oparte na korelacji przedstawione są w postaci *1-c*, gdzie *c* to odpowiedni współczynnik korelacji. Dzięki takiemu przekształceniu dane o wysokim współczynniku korelacji znajdą się blisko siebie. Dodatkowo funkcje te wykazują brak wrażliwości na przekształcenia związane z położeniem i skalą oraz tendencje do grupowania razem genów, których profile ekspresji są liniowo zależne.

4.3. Analiza skupień

Najprostszą i zarazem najpopularniejszą metodą poszukiwania trendów w danych mikromacierzowych jest łączenie genów i próbek w grupy o wspólnym profilu, tzw. klastry. Reprezentacją grupy staje się wtedy pojedynczy profil, który jest uśrednieniem wszystkich elementów klastra, lub jednym z jego elementów, tzw. medoidem lub centroidem. Podstawą działania każdego algorytmu grupującego jest utworzenie „macierzy niepodobieństwa" (ang. *dissimilarity matrix*), według której możliwe będzie uporządkowanie elementów od najbardziej do najmniej podobnych względem siebie. Do obliczeń służą przeważnie omówione funkcje odległości. Algorytm grupujący wymaga również podania liczby poszukiwanych klastrów (wersje pa-

rametryczne) lub co najmniej sposobu jej ustalenia (w przypadku podejścia nieparametrycznego). Bardzo ważnym elementem jest odpowiednie dobranie kryterium optymalizowanego przez algorytm. Kryterium jest funkcją znaczników klastrów, według której określa się stopień podobieństwa elementów w obrębie grupy lub stopień zróżnicowania pomiędzy elementami należącymi do różnych grup.

Algorytmy grupujące możemy podzielić według strategii ich działania na dwie główne klasy, tj. strategie podziału na części (ang. *partitioning*) oraz hierarchiczne. Istnieją również metody łączące oba podejścia.

Strategia podziału na części zakłada podział przestrzeni punktów (genów) na obszary (grupy) niezależne od siebie. Do algorytmów tej klasy zalicza się klastrowanie k-średnich (ang. *k-means clustering*), podział wokół medoidów (PAMs, ang. *partitioning around medoids*) oraz samoorganizujące się mapy (SOMs, ang. *self-organizing maps*). Cechą wspólną tych algorytmów jest to, że wyniki obliczeń dla danego klastra nie są używane w obliczeniach prowadzonych dla pozostałych klastrów.

Algorytmy hierarchiczne z kolei polegają na tworzeniu drzewa, w którym korzeń to klaster zawierający wszystkie geny, a liście to pojedyncze geny. Drzewa te mają zazwyczaj charakter binarny – każdy węzeł rozdziela się na dwie gałęzie bądź liście. Końcowym efektem grupowania hierarchicznego jest uszeregowanie elementów w taki sposób, aby te wykazujące największe podobieństwo były najbliżej siebie (na tej samej gałęzi).

Są dwa sposoby budowy drzewa: podziałowa (ang. *divisive*) i aglomeratywna (ang. *agglomerative*). W pierwszym wychodzimy od „korzenia" (klastra zawierającego wszystkie elementy) drzewa i dzielimy go na coraz mniejsze klastry, aż do pojedynczych genów („liści"). Metoda aglomeratywna rozpoczyna budowę drzewa od „liści", kolejno łącząc elementy najbardziej podobne do siebie. W tej metodzie uwzględnia się kryterium powinowactwa klastrów (ang. *linkage*), które jest jednocześnie miarą odległości pomiędzy nimi. Średnie powinowactwo (ang. *average linkage*) to odległość liczona pomiędzy średnimi obu grup. Pojedyncze powinowactwo (ang. *single linkage*) to odległość pomiędzy najbliższymi sobie elementami dwóch zbiorów (najbliższymi sąsiadami), czyli minimalna odległość dwóch grup. Niekiedy używa się również kryterium całkowitego powinowactwa (ang. *complete linkage*), które mierzy odległość pomiędzy najbardziej oddalonymi od siebie elementami dwóch klastrów, czyli stanowi maksymalną odległość dwóch grup.

Określając liczbę klastrów możemy posłużyć się podejściem parametrycznym, czyli z góry wybrać liczbę docelowych grup na podstawie hipotezy własnej lub oczekiwanego wyniku (np. chcąc podzielić próbki pochodzące od pacjentów cierpiących na różne odmiany choroby). Często jednak, szczególnie grupując geny a nie próbki, trudno dobrać odgórnie docelową liczbę klastrów, w związku z czym opracowano szereg metod nieparametrycznych służących do określenia ich liczby. Metody te dzielą się na bezpośrednie oraz testujące. Te ostatnie nastawione są na testowanie hipotezy wobec hipotezy zerowej i zazwyczaj opierają się na ponownym próbkowaniu (ang. *resampling*) danych. Są więc bardzo kosztowne jeśli chodzi o moc oblicze-

niową komputerów, a przez to cieszą się małą popularnością. Wśród metod bezpośrednich, które działają na zasadzie optymalizacji wybranej funkcji/kryterium, wyróżnić można metodę „sylwetki średniej" (ang. *average silhouette*) (14), której zaletą jest to, że stosować ją można zarówno do podziałowych jak i hierarchicznych metod klastrowania, bez względu na wybraną funkcję obliczania podobieństwa. Jej minusem, jak i większości metod wyboru liczby klastrów, jest to, że bada ona jedynie globalną strukturę klastrów, a nie osobno każdy poziom drzewa czy dużego klastra. Interesujące rozwiązanie tego problemu znaleźć można w metodzie „mediany podziału sylwetki" (MSS, ang. *median split silhouette*) (15). Polega ona na poszukiwaniu małych grup w obrębie większych klastrów, wychodząc z założenia, że przeważnie taka struktura jest odzwierciedleniem zależności biologicznych. Głównym elementem algorytmu jest ocena jak dobrze elementy klastra pasują do siebie. W tym celu wybrany algorytm ogranicza się do grupowania tylko tych elementów, ignorując pozostałe zewnętrzne klastry, a następnie ocenia się jednorodność badanego klastra z wykorzystaniem oceny średniej sylwetki.

Zasadę działania algorytmu MSS przedstawić można następująco. Dla każdego genu *j* obliczamy wartość funkcji niepodobieństwa do innych genów a_j w danym klastrze. Dodatkowo dla każdego genu *j* oraz klastra *l*, do którego ten gen nie należy, oblicza się wartość b_{jl} funkcji niepodobieństwa do genów w klastrze *l*. Przyjmijmy, że wtedy sylwetkę genu można opisać jako wzór:

[9] $$S_j = \frac{(b_j - a_j)}{\max(a_j, b_j)} \qquad b_j = \min_l b_{jl}$$

Funkcja ta stanowi miarę dopasowania genu w obrębie danej grupy w porównaniu z jego dopasowaniem do najbliższej sąsiedniej grupy. „Sylwetka" przyjmuje wartość 1, gdy gen wykazuje identyczność z pozostałymi członkami swojego klastra; 0, gdy gen leży równo pomiędzy dwoma sąsiednimi klastrami; oraz -1, gdy w rzeczywistości powinien znajdować się w sąsiednim klastrze.

Przeprowadzając grupowanie na *k* klastrów, algorytm dzieli każdy klaster na dwa lub więcej mniejszych klastrów (kryterium wyboru może być, np. maksymalizacja średniej sylwetki). Każdy gen będzie miał teraz nową wartość sylwetki, obliczoną tylko wobec genów, z którymi dzieli klaster macierzysty. Sylwetka podziału MSS to mediana wartości dla każdego wyjściowego klastra:

MSS (k) = median (SS_1, ..., SS_k). Wartość ta ukazuje jak jednorodny był wyjściowy klaster i jest niska, jeśli nie należało go dzielić. Wybierając liczbę klastrów do grupowania należy starać się minimalizować MSS. W celu zwiększenia czułości i zmniejszenia intensywności obliczeń można zamiast mediany zastosować średnią.

Ciekawym przykładem kombinacji strategii klastrowania jest algorytm HOPACH (ang. *Hierarchical Ordered Partitioning and Collapsing Hybrid)* (16). Na jego podstawie budowane jest drzewo hierarchiczne, ale na każdym poziomie klastry układane są według niepodobieństwa (ang. *dissimilarity*) rozpatrywanego parami. Punktem startu jest strategia podziału – cały zbiór genów dzielony jest na dwa lub więcej kla-

strów, ale na każdym poziomie budowanego drzewa hierarchicznego stosuje się dodatkowy etap aglomeratywny w celu sprawdzenia czy podział na klastry jest poprawny. Jeśli algorytm wykryje, że nastąpił błąd, łączy dwa najbliższe sobie klastry. Wybór liczby klastrów na każdym etapie oraz decyzja czy któreś z nich należy połączyć podejmowane są na podstawie algorytmu MSS.

Inną interesującą strategią analizy skupień jest grupowanie próbek poprzez grupowanie genów. Do grupowania próbek wykorzystuje się wtedy profile ekspresji (medoid lub średnią) otrzymane dla klastrów genów (17), co znacznie redukuje ilość wymiarów danych (liczbę parametrów dla każdej próbki), a przez to upraszcza procedurę grupowania. Dodatkową zaletą jest to, że profile ekspresji klastrów są stabilne, stąd wynik grupowania próbek będzie również bardziej wiarygodny. Obecność kilku źle dopasowanych genów w obrębie klastrów nie będzie miała wpływu na rezultat analizy.

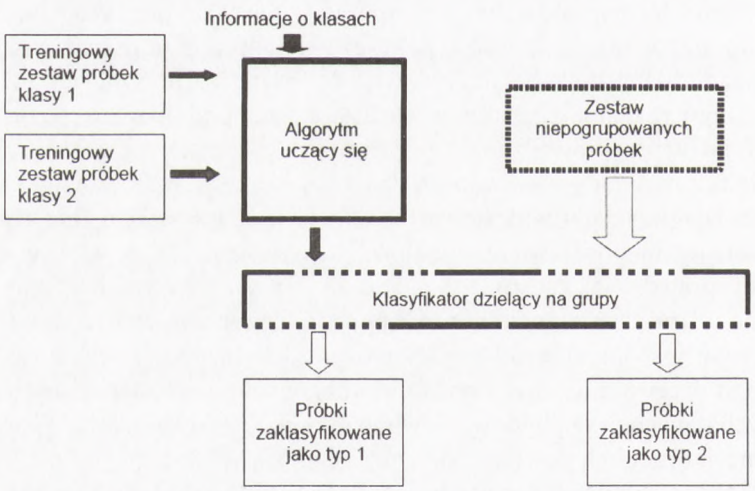
Na podstawie technik grupowania genów można identyfikować bardzo ciekawe zależności pomiędzy danymi, ale niestety dość często brakuje im oparcia statystycznego. Dodatkowo zaobserwowano, że w większości algorytmów grupowania hierarchicznego ostateczny wynik jest w dużej mierze zależny od początkowego ułożenia danych. Problem wynika przede wszystkim z wielowymiarowej struktury danych (bardzo dużo genów) pozyskiwanych z relatywnie małej ilości próbek. Powoduje to, że obserwujemy eksperymentalny rozkład wyników nie znając rzeczywistego ich rozkładu, przy czym jest bardzo mało prawdopodobne, żeby dane mikromacierzowe prezentowały rozkład normalny. Dlatego trudno jest go testować i z tego też powodu powstało wiele metod za pomocą których próbuje się rozwiązać ten problem. Najpopularniejsze z nich to metoda *jacknife* (18) oraz *bootstrap* (19,20).

Obie metody generują przypuszczalny rzeczywisty rozkład danych i na jego podstawie sprawdzają na ile otrzymany przez nas wynik jest losowy. Podstawą tego podejścia jest założenie, że uzyskane dane to tylko przykładowe warianty z niewiadomego rozkładu. Na ich podstawie możemy wygenerować pozostałe, teoretycznie możliwe do uzyskania obserwacje poprzez wielokrotne ponowne próbkowanie i przemieszanie danych eksperymentalnych. Na bazie tak otrzymanego zbioru wyznacza się jego rozkład, a następnie określa prawdopodobieństwo uzyskania danych takich jak w przeprowadzonym eksperymencie. Dodatkowo w metodzie *jacknife* zmienia się też rozmiar generowanych danych, ponieważ ma on również wpływ na ich rozkład. Dla przykładu podczas testowania drzewa grupowania hierarchicznego metodą *bootstrap* na podstawie oryginalnych danych generuje się n losowych drzew, a następnie sprawdza jak często uzyskane przez nas w eksperymencie pary klastrów znajdują się wśród tych drzew. Im częściej dane ułożenie powtarza się w losowych wynikach, tym jest mniej istotne statystycznie. Innym podejściem do testowania tego samego drzewa może być ponowne próbkowanie i podmienianie tylko części oryginalnych danych przed budową nowego drzewa. Tym razem im częściej w nowo generowanych drzewach powtarza się dany układ klastrów, tym bardziej prawdopodobne, że jest on wynikiem zależności opartej na większej części danych (w domyśle – o podłożu biologicznym), a nie o lokalną wariancję czy szum tła.

4.4. Algorytmy uczące się

Różnego rodzaju metody grupowania genów najczęściej używane są w badaniach mechanizmów odpowiedzialnych za regulację ekspresji genów, szlaków metabolicznych, odpowiedzi na stres lub terapię. Całkowicie innego podejścia wymagają eksperymenty zmierzające do identyfikacji tzw. genów markerowych stosowanych w diagnostyce medycznej. W takich sytuacjach zwykle wykorzystywane są algorytmy uczące się. Pod pojęciem „uczenia maszyn" (ang. *machine learning*) kryje się szereg metod, których zadaniem jest utworzenie klasyfikatorów przyporządkowujących nieznane im próbki do określonych grup. Metody te dzielą się na nadzorowane (ang. *supervised*) i nienadzorowane (ang. *unsupervised*). W podejściu nadzorowanym część próbek (tzw. zestaw treningowy) wykorzystywana jest jako materiał, na podstawie którego algorytm ma nauczyć się rozpoznawać zadane klasy (np. pacjenci chorzy na chorobę A, chorobę B i zdrowi). W metodach nienadzorowanych algorytm sam dokonuje grupowania na klasy w obrębie zestawu treningowego próbek, a następnie buduje na ich podstawie klasyfikator do segregacji kolejnych próbek (tzw. zestawu testowego). Ze względów praktycznych najczęściej stosowane jest podejście nadzorowane. Większość eksperymentów nastawiona jest na ustalenie jak najlepszej metody rozpoznania danego typu schorzenia i zazwyczaj dysponuje się wiedzą *a priori* w tym zakresie.

Schemat działania większości nadzorowanych algorytmów uczących się został przedstawiony na rysunku 5. W pierwszym etapie zawsze określany jest zestaw klas oraz oznaczenie przynależności każdej próbki do danej klasy. Następnie algorytm uczy się rozpoznawać próbki należące do różnych klas. Proces uczenia się przebie-



Rys. 5. Ogólny schemat działania algorytmów uczenia maszyn. Ciemne strzałki przedstawiają proces uczenia maszyny, białe strzałki symulują proces klasyfikacji na bazie ustalonego wzorca.

ga dwustopniowo: najpierw należy wybrać cechy (geny), na podstawie których tworzony będzie klasyfikator, a następnie za ich pomocą opisać etykietę każdej klasy, stanowiącą czynnik selekcyjny w procesie rozróżniania próbek. Elementem krytycznym jest wybór genów służących za podstawę klasyfikatora.

Część algorytmów uczących się, takich jak sztuczne sieci neuronowe (ANN, ang. *artificial neural networks*) czy drzewa decyzyjne (ang. *decision trees*), same wybierają odpowiednie geny w trakcie tworzenia etykiety klasyfikacyjnej. Jednak w większości przypadków stosuje się poprzedzające etap uczenia metody filtrowania lub oceny użyteczności genów w tworzeniu klasyfikatora. Najpopularniejsze metody filtrowania opierają się na ocenie: 1) statystyki t (ang. *t-statistic*), 2) stosunku sygnału do tła (SNR, ang. *signal-to-noise ratio*) oraz 3) korelacji. Do oceny użyteczności genów w procesie klasyfikacji stosuje się metody genów dyskryminujących (IDG, ang. *individual discriminatory gene*, JDG, *jointly discriminatory gene*) oraz metody oparte na algorytmach genetycznych (GAs, ang. *genetic algorithms*).

Dotąd zaproponowano bardzo wiele algorytmów klasyfikujących próbki. Najprostsze z nich posługują się kombinacjami liniowych zależności cech (LDA, ang. *linear discriminant analysis*), modyfikacji regresji liniowej (ang. *logistic regression*) czy przypisywania wag (ang. *weighted voting*). Wśród bardziej skomplikowanych algorytmów znajdują się np. takie, które opierają się na sieci współpracujących, ułożonych warstwami elementów wykonujących obliczenia klasyfikujące, tzw. sieci neuronowe. Należą do nich np. algorytm MLPs (ang. *multilayer perceptrons*) czy oparty na strategii Bayesa PNN (ang. *probabilistic neural network*). Interesujący i dość popularny algorytm oparty na tzw. maszynach wektorowych (SVMs, ang. *support vector machines*), zamiast opracowywać skomplikowane funkcje w przestrzeni o wymiarach równych ilości genów w próbkach treningowych, tak jak to robią sztuczne sieci neuronowe, mapuje wprowadzone elementy do hipotetycznej przestrzeni o większej ilości wymiarów, a następnie rozdziela je używając prostych liniowych funkcji klasyfikacyjnych. Ponieważ SVMs opiera się na binarnym systemie decyzyjnym może jedynie dzielić próbki na te, które znajdą się w grupie oraz te, które się do niej nie klasyfikują. Aby móc dzielić próbki na więcej niż dwie grupy zastosowano podejście „jeden kontra wszyscy" (OVA, ang. *one-vs.-all*), polegające na utworzeniu dla podziału na k klasy k klasyfikatorów binarnych SVM konkurujących ze sobą o każdą próbkę. Ostatecznie próbka trafia do grupy reprezentowanej przez SVM o najwyższej wartości klasyfikacji (najdalej od granicy klasyfikacyjnej) (21).

Podobnie jak przy metodach klastrujących nie można jednoznacznie określić, który z algorytmów uczących się jest najlepszy. Wynik i skuteczność otrzymanych klasyfikatorów zależą głównie od jakości i ilości danych użytych do trenowania algorytmu (np. wielkości zestawu próbek) oraz parametrów zastosowanych podczas analizy. Jedyną różnicą jest możliwość bezpośredniej weryfikacji skuteczności algorytmów uczących się poprzez zastosowanie otrzymanych klasyfikatorów na niezależnym zbiorze próbek o znanych klasach.

5. Podsumowanie

Liczba narzędzi służących do uzyskiwania informacji z ogromu danych dostarczanych przez eksperyment mikromacierzowy wciąż rośnie, a te już opisane są bezustannie udoskonalane, zarówno poprzez stosowanie coraz bardziej wysublimowanych metod statystycznych, jak i poprzez wykorzystanie coraz bardziej skomplikowanych układów analitycznych. Metody służące do wstępnej obróbki danych otrzymywanych z klasycznych mikromacierzy ekspresyjnych są już całkiem dobrze opracowane. Niestety szczegółowe opisanie większości algorytmów przekracza możliwości tego opracowania. Warto jednak pamiętać, że każda z metod ma swoje zalety i wady, a ostateczny wynik analizy powinno się porównać z wynikiem uzyskanym inną metodą. Jeśli dana zależność znajduje potwierdzenie w wynikach uzyskanych różnymi metodami istnieje większe prawdopodobieństwo, że jej podłoże ma podstawy biologiczne. Równie ważne jest konfrontowanie otrzymanych wyników z informacjami zawartymi w literaturze i bazach danych, w tym *Gene Ontology*. Stosowanie wszelkiego rodzaju testów statystycznych oceniających wiarygodność stawianych hipotez ma ogromne znaczenie, szczególnie ze względu na specyfikę danych mikromacierzowych, w których ilość obserwowanych cech (genów) wielokrotnie przewyższa ilość próbek (obserwacji jednostkowych).

Planując eksperyment mikromacierzowy należy oprócz ustalenia platformy i rodzaju mikromacierzy odgórnie dokonać wyboru metod analizy, dostosowując je do problemu biologicznego, który chcemy rozwiązać. Wszystkie etapy analizy, począwszy od normalizacji, poprzez sposób przedstawienia wartości poziomu ekspresji genów, metody obliczania odległości, filtrowania, aż po analizę wyższego rzędu, mają wpływ na ostateczny wynik. Dlatego tak istotne jest dokładne zrozumienie poszczególnych elementów analizy. Warto również wiedzieć, że zmieniając nieco dobór parametrów statystycznych, np. zmniejszając selektywność stosowanych metod, można pozyskać dodatkowe, istotne biologicznie informacje.

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

Literatura

1. Kågedal B., Farnebäck M., et al., (2007), Clin. Chem. Lab. Med., 45, 1481-1487.
2. Affymetrix, (2002), Technical Report, 19, Santa Clara, CA.
3. Neaf F., Lim D. A., Patil N., Magnasco M. A., (2001), http://xxx.lanl.gov/abs/physics/0102010.
4. Irizarry R. A., Hobbs B., Collin F., et al., (2003), Biostatistics, 4, 249-264.
5. Bolstad B. M., Irizarry R. A., Astrand M., et al., (2003), Bioinformatics, 19,185-193.
6. Huber W., von Heydebreck A., Sultmann H., et al., (2003), Stat. Appl. Genet. Mol. Biol., 2.
7. Rocke D. M., Durbin B., (2003), Bioinformatics, 19, 966-972.
8. Emerson J. D., Hoaglin D. C., (1983), *Understanding robust and exploratory data analysis*, Eds. Hoaglin D. C., Mosteller F., Turkey J. W., John Wiley & Sons, Inc., Nowy Jork.

9. Wu Z., Irizarry R., Gentleman R., et al., (2004), Journal of the American Statistical Association, 468, 909-917.
10. Smyth G.. K., Speed T., (2003), Methods, 31, 265-273.
11. Yang Y. H., Paquet A. C., (2005), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Eds. Gentleman R., Carey V. J., Irizarry, et al., Springer, New York.
12. Dudoit S., van der Laan M. J., Brikner M. D., (2004), Technical Report, 166, Division of Biostatistics, University of California, Berkeley.
13. Dudoit S., van der Laan M. J., Pollard K. S., (2004), Stat. Appl. Genet. Mol. Biol., 3, 13.
14. Kaufman L., Rousseeuw P. J., (1990), Finding Groups in Data, Wiley.
15. Pollard K., van der Laan M., (2002), W SCI2002 Proceedings, vol. II, 318-325.
16. van der Laan M., Pollard K., (2003), Journal of Statistical Planning and Inference, 117, 275-303.
17. Pollard K., van der Laan M., (2002), Math. Biosci., 176, 99-121.
18. Yeung K. Y., Haynor D. R., Ruzzo W. L., (2001), Bioinformatics, 20, 226-234.
19. Kerr M., Churchill G. A., (2001), Proc. of Natl. Acad. Sci. USA, 98, 8961-8965.
20. van der Laan M., Bryan J., (2001), Biostatistics, 2, 445-461.
21. Ressom H. W., Varghese R. S., Zhang Z., et all, (2008), Front. Biosci., 13, 691-708.