Colloquium Biometricum 38
2008, 231–246

PRZEGLĄD ANALIZ STOSOWANYCH W EKSPERYMENCIE MIKROMACIERZOWYM

Idzi Siatkowski, Joanna Zyprych

Katedra Metod Matematycznych i Statystycznych
Uniwersytet Przyrodniczy
ul. Wojska Polskiego 28, 60-637 Poznań
e-mails: idzi@au.poznan.pl, zjoanna@au.poznan.pl

Streszczenie

Zaawansowane metody stosowane obecnie w technologii mikromacierzowej mają ogromny wpływ na rozwiązywanie problemów w wielu dziedzinach życia. Artykuł ten dotyczy najbardziej powszechnych, dostępnych w literaturze narzędzi do analizy danych pochodzących z mikrotablic. Główne tematy tej pracy to: analiza zeskanowanego obrazu, normalizacja, klasteryzacja, analiza statystyczna. Na specjalną uwagę zasługuje część pracy przedstawiająca przykładowe, istniejące oprogramowania służące do przeprowadzenia powyższych analiz. Praca ta przedstawia podstawowe etapy analiz cDNA mikromacierzy. Ponadto omawia przykładowe metody interpretacji ogromnej ilości danych.

Słowa kluczowe: mikromacierze cDNA, analiza statystyczna, normalizacja, klasteryzacja

Klasyfikacja AMS 2000: 62-07, 62-09

1. Wstęp

W 1966 roku do badań naukowych wprowadzono mikromacierze (Krzemiński, 2003). Są to płytki szklane lub plastikowe z regularnie naniesionymi sekwencjami fragmentów DNA. Istnieją różne kryteria klasyfikacji mikromacierzy. Najbardziej podstawowym jest podział ze względu na budowę sond, które mogą różnić się długością, pochodzeniem, liczbą kopii czy

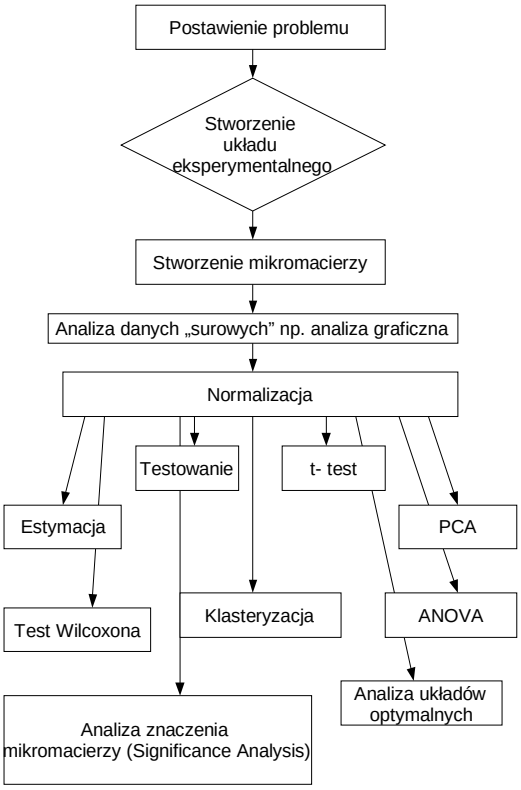
różnić się długością, pochodzeniem, liczbą kopii czy sekwencją docelową. Wyróżniamy mikromacierze oligonukleotydowe (na ogół 25-70 nukleotydowych sekwencji sond) oraz mikromacierze cDNA (sondy znacznie dłuższe, w praktyce zazwyczaj posiadają kompletną sekwencję wariantu genu). W mikromacierzach cDNA hybrydyzuje się dwie próby: badaną i kontrolną. Można także dokonać podziału na mikromacierze do badania ekspresji genów (na mikromacierzy umieszczamy np. zarówno próbki genów pochodzących z organizmów zdrowych, jak i chorych, a otrzymany wzór ekspresji jest porównywany ze wzorem genu odpowiedzialnego za schorzenie) oraz mikromacierze stosowane do analizy mutacji (mikromacierze SNP - odróżnianie jednonukleotydowych polimorfizmów DNA). Do najważniejszych zastosowań zaliczyć można przede wszystkim diagnozowanie chorób – polegające na analizie zależności pomiędzy aktywnością genów pacjenta, który jest „chory na określoną chorobę" o podłożu genetycznym, a stanem rozumianym jako „zdrowy". Dzięki tej technologii można rozszerzać swoją wiedzę o chorobach, wyróżniać jej typy i dostosowywać odpowiednie leczenie do konkretnego typu danej choroby. Dla badaczy bardzo ważnym zastosowaniem czujników DNA jest także identyfikacja nowych genów, odkrywanie wiedzy dotyczącej ich funkcjonowania i poziomie ekspresji w zależności od różnych warunków. Pozwala to na bardziej indywidualne dopasowywanie leczenia do danego typu choroby, na odkrywanie i produkcję skuteczniejszych leków, dostosowanych bezpośrednio do pacjenta, które redukowałyby skutki uboczne. Kolejnymi wykorzystaniami tej techniki mogą być badania własności czynników toksycznych i negatywnych skutków ich oddziaływania na organizm oraz wykrywanie mutacji (SNP).

Przebieg badania ekspresji genów dla mikromacierzy cDNA można opisać w nastepujący sposób. Mikromacierz to płytka na której znajdują się mikroskopowej wielkości pola zawierające fragmenty DNA (sondy). Eksperyment mikromacierzowy rozpoczyna się pobraniem dwóch próbek komórek: badanej i kontrolnej. Cząsteczki cDNA lub mRNA, znajdujące się w badanej próbce (np. fragmencie tkanki) podlegają oznakowaniu za pomocą substancji fluorescencyjnych. Najczęściej stosuje się dla znacznika Cyanine 5 (Cy5) barwnik czerwony, który np. przeznaczony jest do cDNA z komórek badanych, a dla znacznika Cyanine 3 (Cy3) – barwnik zielony z przeznaczeniem do cDNA z komórek kontrolnych. Znaczone cząsteczki cDNA nanoszone są na czujnik. Jak wiadomo, cząstki cDNA lub mRNA, o komplementarnej sekwencji nukleotydów w stosunku do badanej sondy, wiążą się z nią (hybrydyzują). Oznacza to, że gen, który połączył się z DNA na czujniku, był w próbce aktywny (uległ ekspresji). Dzięki znajomości sekwencji sondy i komplementarności zasad azotowych możemy wywnioskować, jaka jest sekwencja nukleotydów. Umie-szczamy czujnik w czytniku (skanerze). Komputer wyznacza dla każdego miejsca stosunek barwnika czerwonego do zielonego (aby określić zmiany w aktywności genów

wywołane działaniem leku) i tworzy barwny, złożony z punktów obraz o określonym znaczeniu poszczególnych kolorów. Punkt czerwony świadczy o ekspresji genu w próbie badanej i jej braku w próbie kontrolnej, a punkt zielony odwrotnie - świadczy o ekspresji genu w próbie kontrolnej i jej braku w próbie zielonej. Natomiast punkt żółty oznacza ekspresję w obu próbach, a punkt czarny o braku ekspresji w obu próbach.

2. Opis eksperymentu z mikromacierzami

W ciągu ostatnich dziesięciu lat, w badaniach opartych na technice mikromacierzowej, zastosowano wiele nowych metod analizy obrazu powstałego z czujnika, opublikowano nowe metody analiz graficznych, nowe sposoby testowania, zasady klasteryzacji i normalizacji danych. Przykładowy schemat eksperymentu mikromacierzowego przedstawia rys. 1.



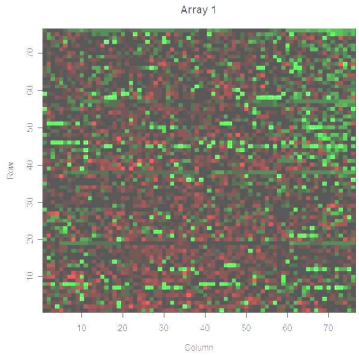
Rys. 1. Schemat przebiegu eksperymentu mikromacierzowego

3. Analiza zeskanowanego obrazu

Do zaprezentowania poszczególnych etapów analizy danych mikromacierzowych wykorzystane zostały dane *kidney* z CAMDA (Critical Assessment of Microarray Data Analysis), które znajdują się na http://projetos.inpa.gov.br/i3geo/pacotes/r/win/library/survival/html/kidney.html lub http://www.bioconductor.org/packages/2.1/data/experiment/vignettes/kidpack/inst/doc/kidpack.pdf oraz dane *ApoAI*, które znajdują się na http://www.bioconductor.org. Dane *kidney* zawarte są w 24 dwukolorowych tablicach. Do analizy tych danych użyto pakietu **maanova** (MicroArray ANalysis Of VAriance)**,** który wchodzi w skład Bioconductora. Natomiast dane *ApoAI* (Dudoit i in. 2000) dotyczą doświadczenia w którym porównywano 8 myszy (u których wyłączony był gen ApoAI - apolipoproteina) z 8 dzikimi myszami (kontrolnymi - C57BL/6. Dla każdej z 16 myszy docelowe mRNA otrzymano z wątroby i oznaczono barwnikiem Cy5. RNA z każdej myszy uległo hybrydyzacji na oddzielnych matrycach. Referencyjne RNA zostało oznaczone barwnikiem Cy3 i użyte dla każdej mikromacierzy. Referencyjne RNA otrzymano przez zsumowanie RNA wyekstrahowanych z 8 myszy kontrolnych. Do analizy tych danych zastosowano pakiet **limma** (Linear Models for Microarray Data).

Po stworzeniu mikromacierzy, dane *kidney* zostały zeskanowane za pomocą odpowiedniego skanera, a następnie z pomocą pakietu **maanova** w **R** (opis pakietu R – Rozdział 7 lub http://www.r-project.org) wygenerowano dwukolorowy obraz – patrz rys. 2.

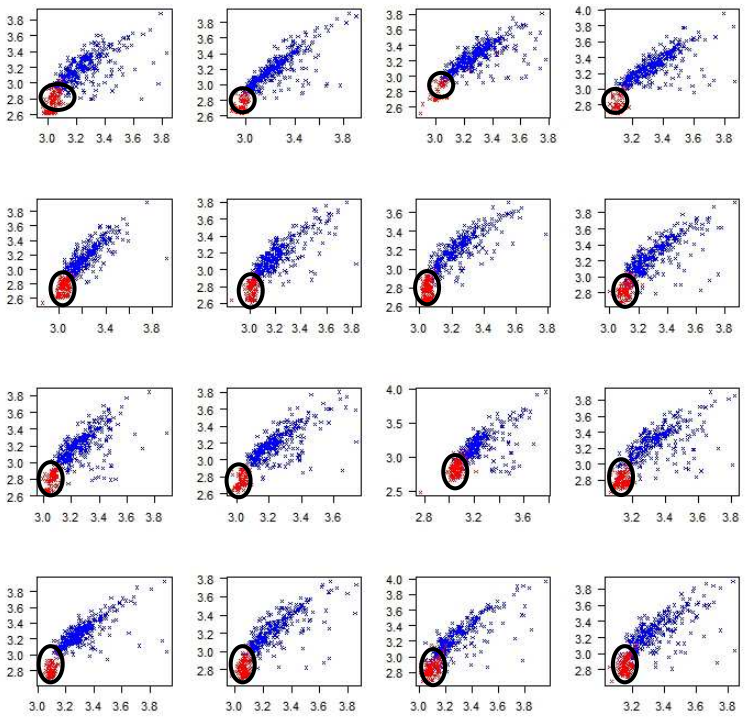
```
> library(maanova)
> data(kidney)
> arrayview(kidney.raw,zlim=c(-7,7))
```



Rys. 2. Zeskanowany obraz dla pierwszej macierzy dla danych *kidney*

W pracy Yang i in. (2001) przedstawiono metodę „wyciągania” informacji z zeskanowanego obrazu z mikromacierzy. Metoda ta składa się z trzech etapów: automatyczne adresowanie każdego punktu poprzez jego współrzędne, segmentacje (klasyfikacja punktu jako „sygnał” bądź też „tło”), obliczenie intensywności sygnału dla każdego punktu, obliczenie intensywności tła oraz jakości pomiarów dla każdego barwnika. W R graficzne sprawdzanie jakości danych można wykonać następująco:

> *library(maanova)*
> *data(kidney)*
> *gridcheck(kidney.raw)*



Rys. 3. Wykres prezentujący rozmieszczenie punktów dla pierwszej macierzy dla danych *kidney* (elipsy wskazują położenie punktów w kolorze czerwonym)

Obrazem jest wykres rozrzutu dla każdego układu punktów w każdej macierzy. Czerwone punkty oznaczają „oflagowane sondy". W przypadku, gdy niektóre wykresy wyglądają na nieuporządkowane możemy domyślać się o wystąpieniu błędów w hybrydyzacji bądź błędów w rozmieszczeniu sond na mikromacierzy. Rys. 3. przedstawia wykres rozmieszczenia danych *kidney* dla pierwszej macierzy.

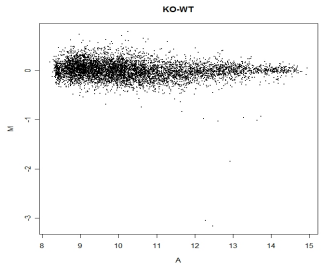
Przy obliczaniu intensywności sygnału dla każdej sondy wyróżniamy metody: okręgu stałego (fixed circle), histogramu i segmentacji dostosowawczej (adaptive segmentation), co zostało opisane w pracach Ahmeda i in. (2004) czy Bocianowskiego i in. (2002). Dyskusja na temat obróbki obrazu została przedstawiona również w pracach Buckley'a (2000), Yang'a i in. (2001), Rueda i Li (2005), przy czym problem odjęcia sygnału tła (background correction of the data) został dokładnie omówiony w pracach: Ritchie i in. (2007), Edwards'a (2003), Kooperberg'a i in. (2002), czy Yin'a i in. (2005).

Przed właściwą analizą danych konieczne jest ujednolicenie danych, gdyż na skutek różnic w wykonywanym eksperymencie (tj. różnice sprzętowe, między barwnikami, między procedurami znakowania) powstają różnice w pomiarze poziomu ekspresji genów, co może fałszować wyniki eksperymentu i utrudniać analizę. W związku z tym konieczne jest przeprowadzenie normalizacji danych.

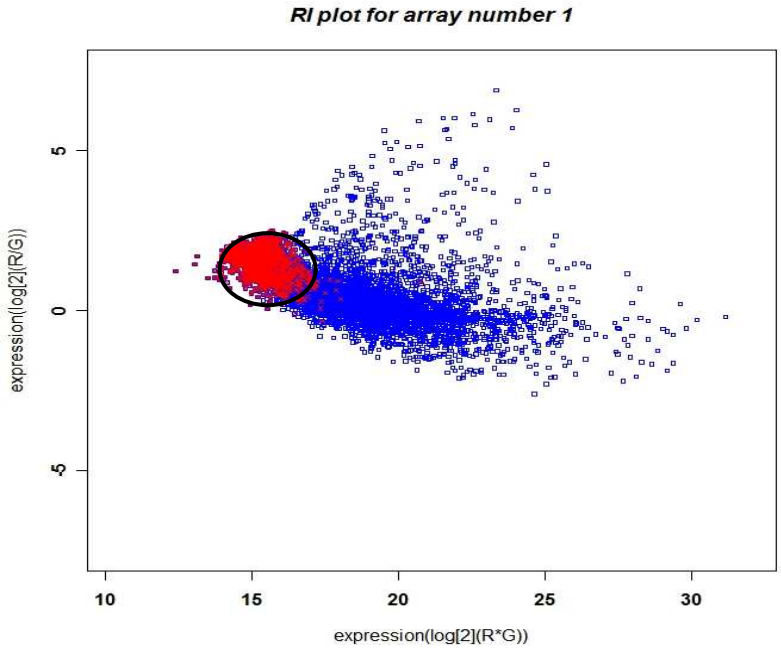
4. Normalizacja

Przed przystąpieniem do dalszej analizy danych otrzymanych w eksperymencie mikromacierzowym, konieczna jest normalizacja, czyli wyeliminowanie „zakłóceń" powstałych poprzez zastosowanie różnych płytek, materiałów znakujących oraz urządzeń. Problem normalizacji omawiają m. in. Venables i in. (1999), Bolstad i in. (2003), Quaackenbush (2002), Yang i in. (2002) oraz Hoffmann i in. (2002). Natomiast Reimers (2005) przedstawia normalizacje polegającą na wymnożeniu sygnałów dla każdego genu przez wspólny czynnik. Po operacji wymnożenia totalna intensywność sygnału dla każdego koloru wynosi 1, podczas gdy intensywność sygnału dla pojedynczego genu jest wielkością małą. Jedną z podstawowych metod normalizacji jest globalna normalizacja, która polega na wyznaczeniu sumy sygnałów z wielu obrazów i wykorzystania jej do wyrównania sygnału - użycie globalnego współczynnika normalizacji przedstawia Quaackenbush (2001). Innymi metodami normalizacji są: normalizacja względem genów o stałej ekspresji (housekeeping normalization)

oraz kontrola iglicowa (spiking control). Przykładowo DeRisi i in. (1996) używają normalizacji względem genów o stałej ekspresji stosując 90 specjalnie wyselekcjonowanych genów. Natomiast kontrola iglicowa polega na wprowadzeniu do mikromacierzy nieznacznych ilości próbek kontrolnych, służących do normalizacji danych pomiędzy szkiełkami oraz pomiędzy barwnikami.



Rys. 4. Wykres MA dla danych *ApoAI*



Rys. 5. Wykres RI dla pierwszej macierzy danych *kidney* (elipsa wskazuje położenie punktów w kolorze czerwonym)

Często stosowaną metodą normalizacji jest dopasowanie funkcji Lowess (Cleveland 1979 oraz Smyth i in. 2003) i wykonanie wykresu typu MA, gdzie M= $\log_2(R/G)$ oraz A= $\log_2\sqrt{R*G} = \log_2(R*G)/2$, przy czym współczynnik R oznacza efekt barwnika czerwonego, natomiast G - barwnika zielonego. Metoda ta jest metodą nieliniową i staje się coraz bardziej popularna dla normalizacji danych z mikromacierzy, gdyż w większości przypadków zależność pomiędzy dwoma tablicami jest nieliniowa (Yang i in., 2002). Odkąd wiadomo, że efekt barwnika jest zależny od intensywności, metoda ta jest bardziej odpowiednia niż globalna metoda normalizacji. Rys. 4 przedstawia wykres MA dla danych *ApoAI*.

Innym wykresem jest wykres RI (ratio intensity), który wygląda bardzo podobnie do MA i może być także używany w celu pokazania efektów normalizacji. Wykres RI (rys. 5) pokazuje specyfikę intensywności dla każdego genu dla pierwszej macierzy w danych *kidney*, jako funkcję indywidualnych intensywności (dla barwnika czerwonego i zielonego).

Inną metodą normalizacji bazującą na metodach liniowej regresji jest estymacja K_i (K = median (log2 (R / G))) dla każdego genu (Quaackenbush, 2001 oraz Yang i in. (2001).

5. Ocena ekspresji genów

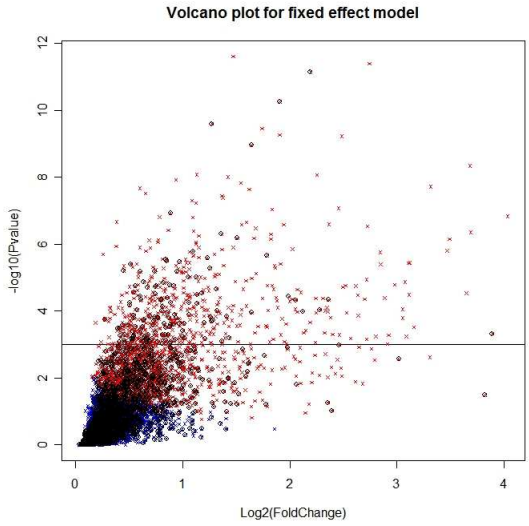
Po wykonaniu normalizacji można przystąpić do właściwej analizy danych. Wynikowy zbiór danych poddaje się różnym analizom. Ważną oceną jest określenie znaczenia zmian w poziomie ekspresji genów, tj. analiza zmian ekspresji w czasie, identyfikacja różnic w ekspresji genów oraz badanie, czy pewne geny wykazują ekspresję charakterystyczną np. dla pewnych chorób. Jedną z metod oceny znaczenia zmian wielkości poziomu ekspresji jest metoda bazująca na teście t-Studenta. W proponowanej metodzie (Dudoit i in. 2002) stosuje się do porównania wartość średnią oraz odchylenia dwóch grup prób – kontrolnej i badanej. Dane z ekspresji genów mogą byż zapisane w macierzy X – macierz $\log_2(R/G)$ z k-wierszami odpowiadającymi danym genom oraz n = n_1+n_2 kolumnami odpowiadającym n_1-próbkom kontrolnym oraz n_2-badanym. Hipoteza zerowa zakłada równość wartości średniej poziomu ekspresji genów dla obu tych prób, przy czym poziom ekspresji opisywany jest jako $\log_2$ (R/G). Autorzy podają interpretację wartości statystyki *t* porównującej ekspresję tych genów. Duża wartość statystyki wskazuje na to, ze odpowiednie geny mają różny po-

ziom ekspresji w grupie obiektów kontrolnych i badanych. Wadą tego testu jest to, że w większości eksperymentów n_1 i n_2 mają wartości małe.

Innym testem używanym w analizie danych mikromacierzowych jest test rang Wilcoxona. Ten nieparametryczny test stosowany jest ponieważ dane z ekspresji genów nie podlegają rozkładowi normalnemu.

Do analizy danych mikromacierzowych odbiegających od rozkładu normalnego stosuje się także metody oparte na testach permutacji, co zostało przedstawione u Tushera i in. (2001) oraz Reimers'a (2005).

Wizualną metodą oceny ekspresji genów jest np. wykres volcano – graficzna metoda prezentacji danych dużych rozmiarów wykorzystująca wyniki testu *t* (Ghanem M.M. 2005) - patrz rys. 6.



Rys. 6. Wykres volcano dla danych *kidney*

Gdy mamy do porównania więcej niż dwie populacje możemy wykonać to za pomocą analizy wariancji (ANOVA), co zostało przedstawione u Kerr i Churchill'a (2001). W proponowanych przez nich modelach uwzględniona jest zmienność między barwnikami fluorescencyjnymi, powtórzeniami genów wewnątrz mikrotablic oraz między tablicami.

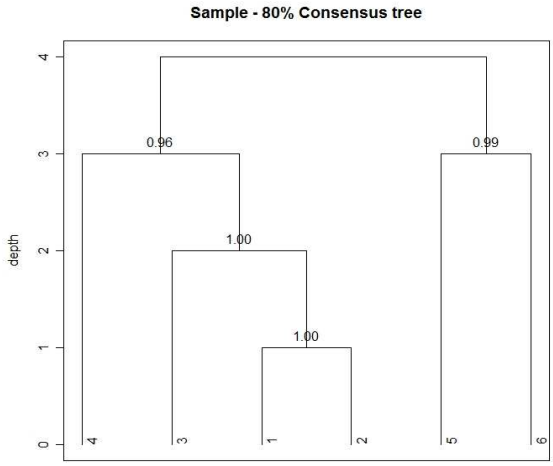
Jeśli chodzi o analizę układów optymalnych, została ona opisana w pracy Wun'a i Wang'a (2006). Autorzy proponują metody dla konstrukcji optymalnych i efektywnych układów dla ekspresji genów w mikromacierzach.

Natomiast identyfikację genów w przestrzeni trójwymiarowej z wykorzystaniem analizy składowych głównych (PCA) opisali Wall i in. (2001) oraz Yeung i Ruzzo (2001).

6. Klasteryzacja

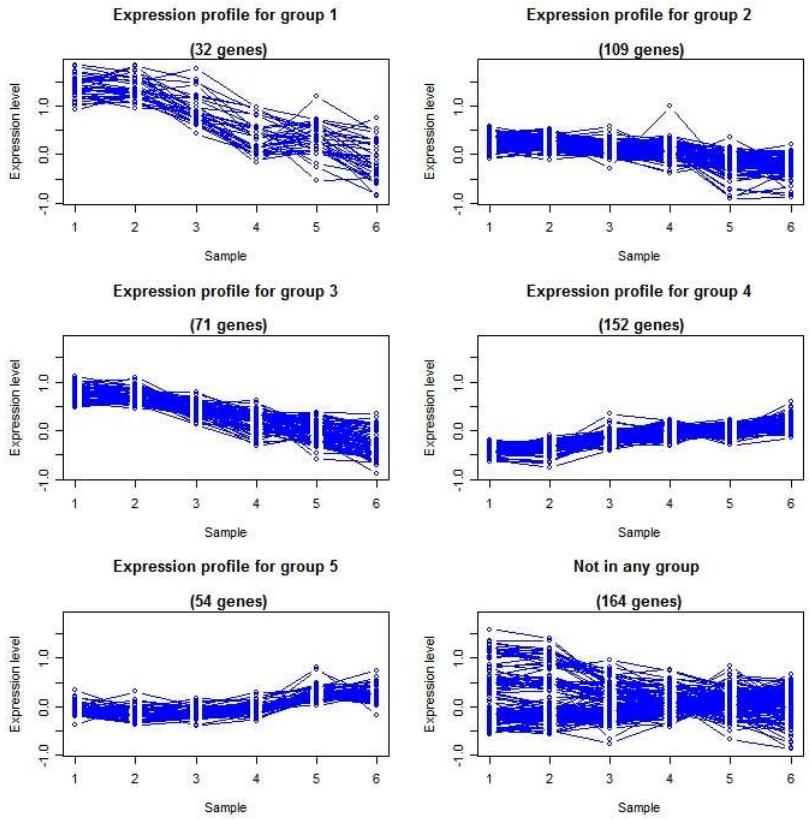
Kolejnym sposobem analizy ekspresji genów jest połączenie genów w klastry (grupy) o podobnym profilu ekspresji genów. Wyróżniamy klasyfikacje i grupowanie genów ze względu na profil ekspresji próbek (grupowanie względem wierszy) oraz klasyfikacje i grupowanie próbek ze względu na profil ekspresji genów (grupowanie względem kolumn). Standardowe podejście do klasteryzacji opisują Shannon i in. (2003).

Główne typy klasteryzacji to nieukierunkowana i ukierunkowana analiza skupień. Nieukierunkowana (nienadzorowana) analiza skupień (unsupervised clustering) to nastepujące metody: grupowanie hierarchiczne, metoda k-średnich oraz SOM (samoorganizujące mapy Kohonena). Metoda grupowania hierarchicznego jest najbardziej powszechna dla wyznaczania grup genów blisko spokrewnionych. Klastrowanie hierarchiczne używa w swojej pracy Eisen i in. (1998). Do tego algorytmu możemy stosować różne definicje odległości: minimalną odległość (Quaackenbush, 2001), odległość najdalszych punktów z obu klastrów (Quaackenbush 2001 oraz Maciejewski i in. 2005), odległość średnią czy też odległość Ward'a (Ward, 1963). Przykład klasteryzacji hierarchicznej dla pojedynczych próbek przedstawia rys. 7.



Rys. 7. Klasteryzacja hierarchiczna dla danych *kidney*

Najczęściej stosowaną metodą jest metoda k-średnich (k-means) (Quaackenbush, 2001). Metoda ta dzieli obiekty na k – klastrów w taki sposób, że klastry są wewnętrznie podobne, ale zewnętrznie różne, co można zaobserwować na rys. 8.



Rys. 8. Wykresy profilów ekspresji danych *kidney* po zastosowaniu metody k- średnich

Samoorganizujące mapy Kohonena (SOM) są metodą polegającą na sieciach, które "rozpinają się" wokół zbiorów danych (komputerowego modelu skanowanego obiektu), dopasowując do nich swoją strukturę. W metodzie tej przydziela się geny do serii podziałów na podstawie podobieństwa w ekspresji wektorów do wektora źródłowego - referencyjnego (Kohonen 1992 i Tamayo i in. 1999).

Metody nienadzorowane są bardzo często niedokładne stąd też potrzeba stosowania klasteryzacji ukierunkowanej (supervised clustering) zwanej też grupowaniem dyskryminacyjnym. Używa ona istniejących informacji biologicznych

dotyczących specyfiki genów oraz przypisania im odpowiedniego algorytmu klasteryzacji. Quaackenbush (2001) opisuje techniki ukierunkowanej analizy skupień, natomiast Zhao i in. (2004) opisują jej algorytmy. Wśród metod nadzorowanych wyróżniamy: drzewa decyzyjne, klasyfikację najbliższych sąsiadów (Singh i in. 2001), SVM (support vector machines) (Finley i in. 2005, Brown i in. 2000, Fujarewicz i in. 2003), okna Parzen'a (Khan i in. 2001) oraz liniowy dyskryminator Fishera (Welling, 2005).

7. Narzędzia do analizy danych mikromacierzowych

Najbardziej znanym komercyjnym narzędziem do analizy danych mikromacierzowych jest SAS Microarray Solution (Wolfiningera i in. 2004). Innymi przykładowymi komercyjnymi oprogramowaniami są: arrayScout, GeneSpring, Spotfire DecisionSite, GeneTraffic.

Kolejnym bardzo ważnym narzędziem umożliwiającym analizę danych mikromacierzowych jest **Bioconductor** stworzony w ramach platformy obliczeniowej R. R jest projektem GNU opartym o licencje GPL GNU, czyli jest darmowy dla wszystkich zastosowań. Platforma R wyposażona jest w dobrą, ogólnie dostępną dokumentację. Bardzo ważnym elementem tego oprogramowania jest swoboda w tworzeniu i upowszechnianiu pakietów. Obecnie dostępnych jest ponad 1000 pakietów. Pakiet R pozwala także na wykonywanie procedur z bibliotek przygotowywanych w innych językach (np. C, C++, Fortran). Jego ważną zaletą jest możliwość generowania wykresów o wysokiej jakości. Bioconductor służy do analizy danych pochodzących z Affymetrix oraz z mikromacierzy cDNA. Złożoność oraz wielość pakietów ułatwiających analizę czyni to narzędzie jeszcze atrakcyjniejszym. Na szczególną uwagę zasługują, oprócz już cytowanych **limma** i **maanova**, następujące pakiety: **marray** (analiza statystyczna cDNA mikromacierzy), **cgh** (CGH analiza wykorzystująca algorytm Smith'a i Waterman'a), **FunCluster** (profilowanie danych pochodzących z analizy ekspresji genów), **pamr** (predykcyjna analiza mikromacierzy), **samr** (SAM-Significance Analysis of Microarrays), **sma** (analiza statystyczna mikromacierzy), **twslm** (dwukierunkowy semi–liniowy model dla normalizacji i analizy danych pochodzących z cDNA mikromacierzy). Pakiety te stanowią część możliwości tego programu - nie jest jednak celem tej pracy dokładne analizowanie ich zastosowań, dlatego też poprzestaniemy na wspomnieniu powyższych.

Oprócz pakietów komercyjnych oraz platformy R z Bioconduktorem istnieją inne oprogramowania. Na uwagę zasługuje TM4, który jest oprogramowaniem

open source. W jego skład wchodzą nastepujące aplikacje: Microarray Data Manager - MADAM (wczytywanie danych), TIGR_Spotfinder (analiza obrazu z mikromacierzy), Microarray Data Analysis System - MIDAS (normalizacja danych), oraz Multiexperiment Viewer - MeV (analiza znormalizowanych i przefiltrowanych danych).

Programem zawierającym możliwości zaawansowanych metod analiz jest Genowiz. Program ten umożliwia badaczom lepszą organizację zbiorów danych, szybszy i prostszy import danych oraz pozwala na prace z wieloma danymi w tym samym czasie. Dzięki niemu można łatwo wykonać: transformacje danych (log transformacje, wartość średnia, mediana oraz Z – transformacja), normalizacje (dla mikromacierzy cDNA i Affymetrix), filtrowanie danych oraz analizę danych (klasteryzacja hierarchiczna, SOM, PCA, metoda k-średnie, analiza dyskryminacyjna i SVM).

Jeszcze innym, ogólnodostępnym programem jest Scanalyzer – program do analizy obrazu z mikromacierzy. Jego główne moduły to: Cluster (klasteryzacja hierarchiczna, SOM, k- średnia klasteryzacja, PCA) oraz TreeView i MapleTree (graficzne prezentacje wyników klasteryzacji).

8. Podsumowanie

Technologia mikromacierzowa dostarcza nam ogromnego potencjału dla poprawy wiedzy dotyczącej badań biomedycznych, farmakogenomiki, odkrywania nowych genów i ich roli w organizmie, budowania modeli opisujących ekspresję genów. Dlatego też bardzo ważnym jest zarówno rozwój starych, jak i tworzenie nowych narzędzi stosowanych w tego rodzaju analizach. Niniejsza praca ma na celu przegląd i usystematyzowanie metod stosowanych w tej dziedzinie. Dostępność wielu metod analizy jest rezultatem prac wielu grup naukowców, co sprzyja lepszemu zrozumieniu wyników doświadczenia i lepszej interpretacji wyników badania.

Podziękowania

Autorzy pragną bardzo serdecznie podziękować nieznanemu recenzentowi za wnikliwą recenzję pracy.

Literatura cytowana

Ahmed A.A, Vias M., Gopalakrishna I.N., Caldas C., Brenton J.D. (2004). Microarray segmentation methods significantly influence data precision. *Nucleic Acids Res.*32(5), e50.

Bocianowski J., Gallavotti A., Krajewski P., Pe E. (2002). On methods of collecting data from DNA microarrays. *Colloquium Biometryczne* 32, 133-139.

Bolstad B.M., Irizarry R.A., Astrand M., Speed T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 22, 185-93.

Brown M.P.S., Grundy W.N., Lin D., Cristianini N., Sugnet Ch.W., Furey T.S., Haussler D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.* 97, 262-267.

Buckley M.J. (2000). *The Spot user's guide*. CSIRO Mathematical and Information Sciences.

Callow M.J., Dudoit S., Gong E.L., Speed T.P., Rubin E.M. (2000). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research* 10, 2022-2029.

Cleveland W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *J. Am. Stat. Assoc.* 74, 829-836.

DeRisi J., Penland L., Brown O., Bittner L., Meltzer P.S., Ray M., Chen Y., Su Y.A., Trent J.M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics* 14, 457-460.

Dudoit S., Yang Y.H., Speed T.P., Callow M.J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12, 111-139.

Edwards D. (2003). Non – linear normalization and background correction on onechannel cDNA microarray studies. *Bioinformatics* 19, 825-833.

Eisen M.B., Spellman P.T., Brown P.O., Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* 96, 10781-10786.

Finley T., Thorsten J. (2005). Supervised Clustering with Support Vector Machines, *International Conference Proceeding Series*. Vol. 119, Bonn, Niemcy.

Fujarewicz K., Kimmel M., Rzeszowska-Wolny J., Wierniak A. (2003). A note on classification of gene expression data using support vector machines. *J. Biol. Sys.* 11, 43-56.

Ghanem M.M. (2005). Introduction to Bioinformatics. 6. Statistical Analysis of Gene Expression Matrices II. Course 341. Department of Computing Imperial College, London. (http://www.doc.ic.ac.uk/~mmg/341/Lecture6.ppt).

Hoffmann R., Seidl T., Dugas M. (2002). Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol.*14 (http://genomebiology.com/2002/3/7/research/0033).

Kerr M.K, Churchill G. (2001). Experimental design for gene expression microarrays, *Biostatistics* 2, 183-201.

Khan J., Wei J.S., Ringnér M., Saal L.H., Ladanyi M., Westermann F., Berthold F., Schwab M., Antonescu C.R., Peterson C., Meltzer P.S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artifical neutral netwoks. *Nature Medicine* 7, 673-79.

Kohonen T. (1992). Self-organized formation of topologically correct feature maps. *Biol. Cybernetics* 43, 59-69.

Kooperberg C., Fazzio T.G., Delrow J.J., Tsukiyama T. (2002). Improved background correction for spotted DNA microarrays. *J. Comput. Biol.* 9, 55-66.

Krzemiński Z. (2003). Postępy w diagnostyce mikrobiologicznej chorób zakaźnych, *Przegl. Epidemiol*. 57, 377-80.

Maciejewski H., Konarski Ł., Jasińska A., Drath M. (2005). Analiza danych z mikromacierzy DNA metody, narzędzia. *Journal Edited by Medical College – Jagiellonian University*, 129-132.

Quaackenbush J. (2001). Computational analysis of microarray data. *Nature* 2, 418-427.

Quaackenbush J. (2002). Microarray data normalization and transformation. *Nat. Genet.* 32 Suppl, 496-501.

Reimers M. (2005). Statistical Analysis of Microarray Data. *Addiction Biology* 10, 23-35.

Ritchie M.E., Silver J., Oshlack A., Holmes M., Diyagama D., Holloway A., Smyth G.K. (2007). A comparison of background correction methods for two-colour microarrays. Bioinformatics 23, 2700 – 2707.

Rueda L., Li Q. (2005). *A new method for DNA microarray image segmentation*. Lecture Notes in Computer Science. Springer.

Singh S., Haddon J., Markou M. (2001). Nearest – neighbour classifiers in natural scene analysis. *Pattern Recognition* 34, 1601-1612.

Shannon W., Culverhouse R., Duncan J. (2003). Analyzing microarray data using cluster analysis. *Pharmacogenomics* 4, 41-52.

Smyth G. K., Yang Y.-H., Speed T. P. (2003). Statistical issues in cDNA microarray data analysis. *Methods in Molecular Biology* 224, 111-136.

Tamayo P., Slonim D., Masirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S., Golub T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences USA* 96, 2907-2912.

Tusher V.G., Tibishirani R., Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98, 5116-21.

Wall M.E., Dyck P.A., Brettin T.S. (2001). SVDMAN – singular value decomposition analysis of microarray data. *Bioinformatics* 17, 566 – 568.

Ward J.H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236-244.

Welling M. (2005). *Fisher Linear Dicriminant Analysis*. Department of Computer Science, University of Toronto, Canada (http://www.ics.uci.edu/~welling/classnotes/papers_class/Fisher-LDA.pdf).

Wolfininger R., Czika W., Kleiss K., Kreiss W., Progress in SAS Scientific Discovery Solutions. *Genetics, Microarrays, and Proteomics*, Paper 213 – 29.

Wun-Yi Shu, Wang Y.Ch. (2006). *Optimal Design for Gene Expression Microarrays*. Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan.

Yang H., Buckley M.J., Dudoit S., Speed T.P. (2001). Image processing on cDNA microarray data (http://www.ipam.ucla.edu/publications/fgtut/fgtut_4045.ppt).

Yang Y.H., Dudoit S., Luu P., Lin D.M., Peng V., Ngai J., Speed T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* 30(4), e15.

Yeung K.Y., Ruzzo W.L. (2001). Principial component analysis for clustering gene expression data. *Bioinformatics* 17, 763-774.

Yin W., Chen T., Zhou S.X., Chakraborty A. (2005). Background correction for cDNA microarray images using the TV + L1 model. *Bioinformatics* 21, 2410-2416.

Zhao Z., Eick C.F., Zeidat N. (2004). Supervised clustering - algorithms and benefits (http://www2.cs.uh.edu/~ceick/kdd/EZZ04.pdf).

MICROARRAY DATA ANALYSIS - REVIEW

Summary

This article aims to distill the most useful practical results from the literature about microarray data analysis. The main topics are image analysis, normalization, cluster analysis and tests for differential expresion. Special attention is paid to the software analysis. This review leads through all steps in the microarray data analysis and gives a basic understanding of the challenges in interpreting large datasets.

Key words and phrases: Microarray, analysis, statistics, normalization, clustering

Classification AMS 2000: 62-07, 62-09