

Uczenie maszynowe w bioinformatyce

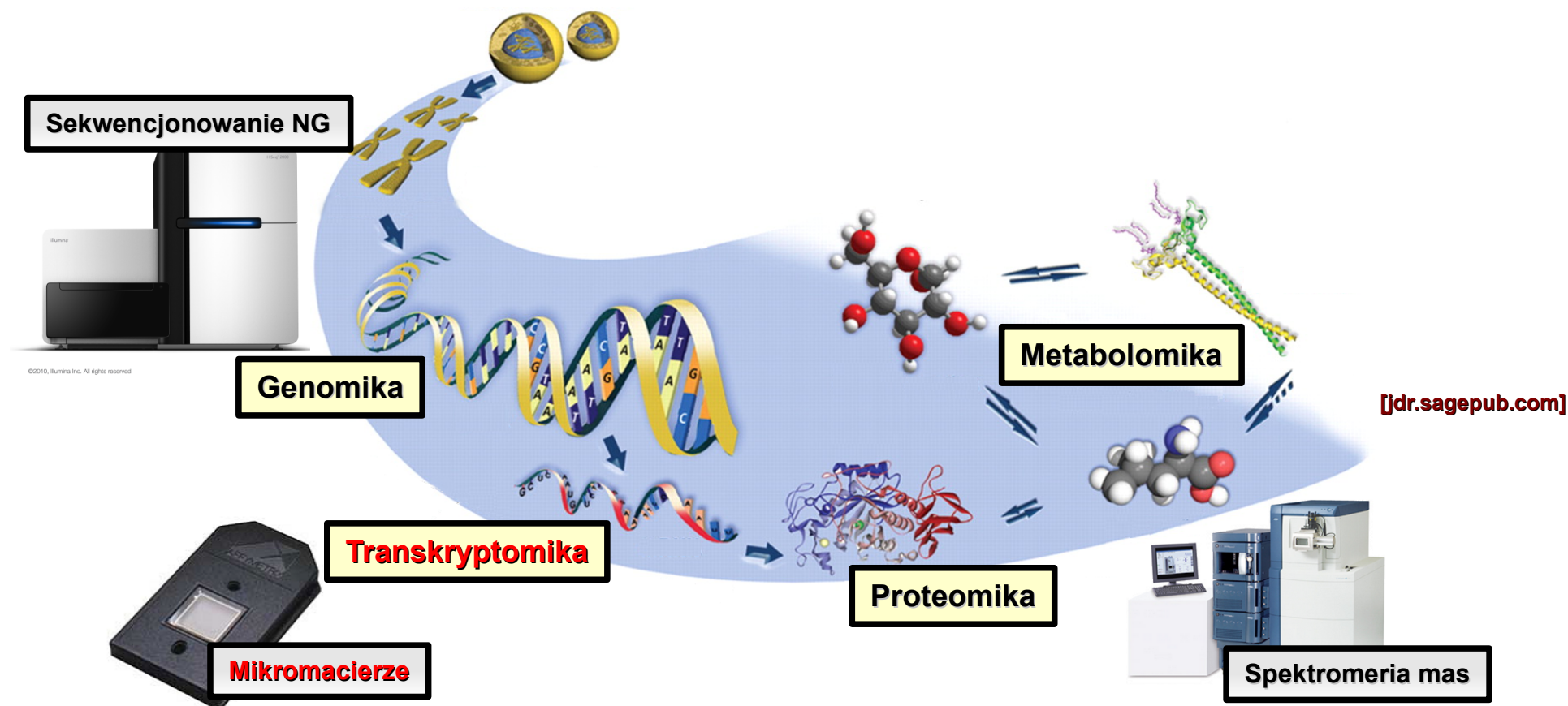
Wykład 5: transkryptomika

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Transkryptom jest to zbiór wszystkich transkryptów (cząsteczek mRNA będących efektem transkrypcji genów) występujących w komórce. W przeciwieństwie do genomu, transkryptom jest dynamiczny. Jego skład różni się pomiędzy komórkami tego samego organizmu. Zmienia się on również w czasie, w zależności od fazy rozwoju i stanu organizmu oraz wpływu czynników środowiskowych.

Transktypromika zajmuje się badaniem transkryptomów w sposób jakościowy (które geny ulegają ekspresji) i ilościowy (jak zmienia się ekspresja genów).



Uczenie maszynowe w bioinformatyce

Wykład 5: mikromacierze

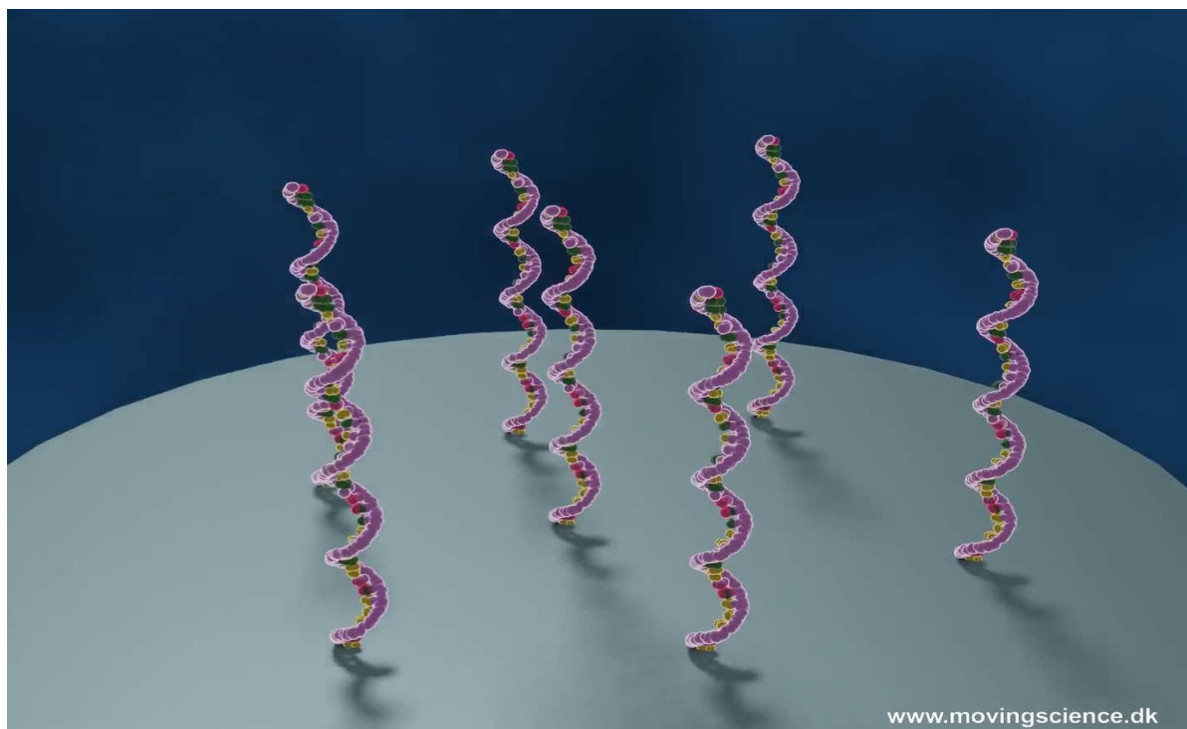
Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Mikromacierze DNA

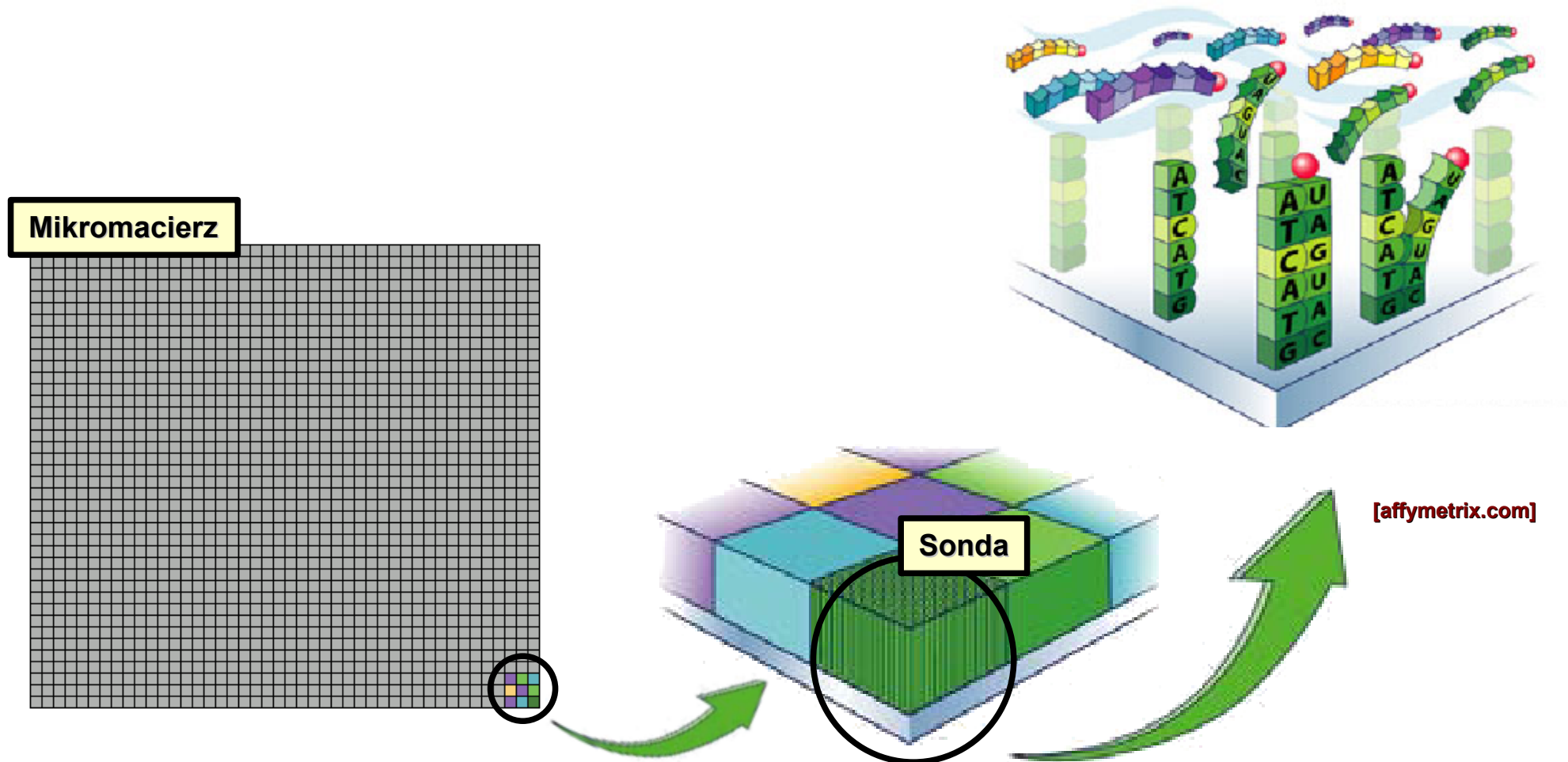
Mikromacierz DNA (DNA microarray) jest to płytka, na powierzchni której w sposób uporządkowany rozmieszczono dużą liczbę (od kilkuset do milionów) sond DNA.

Sonda DNA (DNA probe) jest niewielkim obszarem zawierającym wiele identycznych cząstek jednoniciowego DNA przytwierdzonych na stałe do podłoża. Sekwencja nukleotydów DNA z sondy jest komplementarna do fragmentu lub całości sekwencji cząstki mRNA powstającej podczas transkrypcji wybranego genu. W odpowiednich warunkach transkrypty tego genu mogą więc połączyć się (ulec **hybrydyzacji**) z fragmentami DNA znajdującymi się w sondzie.



Mikromacierze DNA

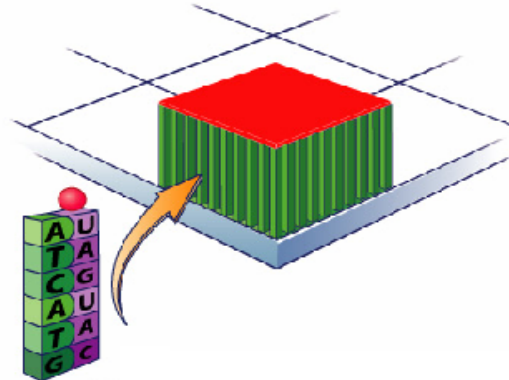
Każda z sond zawiera DNA o innej sekwencji. Zbiór wszystkich sond tworzy macierz, której każdy punkt jest „czuły” tylko na jeden rodzaj cząsteczek mRNA.



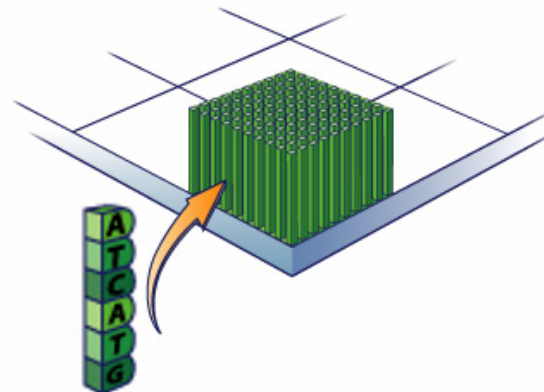
Mikromacierze DNA

Po wprowadzeniu mRNA z komórek na mikromacierz poszczególne rodzaje cząstek hybrydują z sondami zawierającymi komplementarne wobec nich nici DNA:

- jeżeli pewien rodzaj mRNA występuje w badanych komórkach, to jego cząsteczki połączą się z odpowiednią sondą (pod warunkiem, że sonda reprezentująca ten rodzaj mRNA została umieszczona na mikromacierzy);



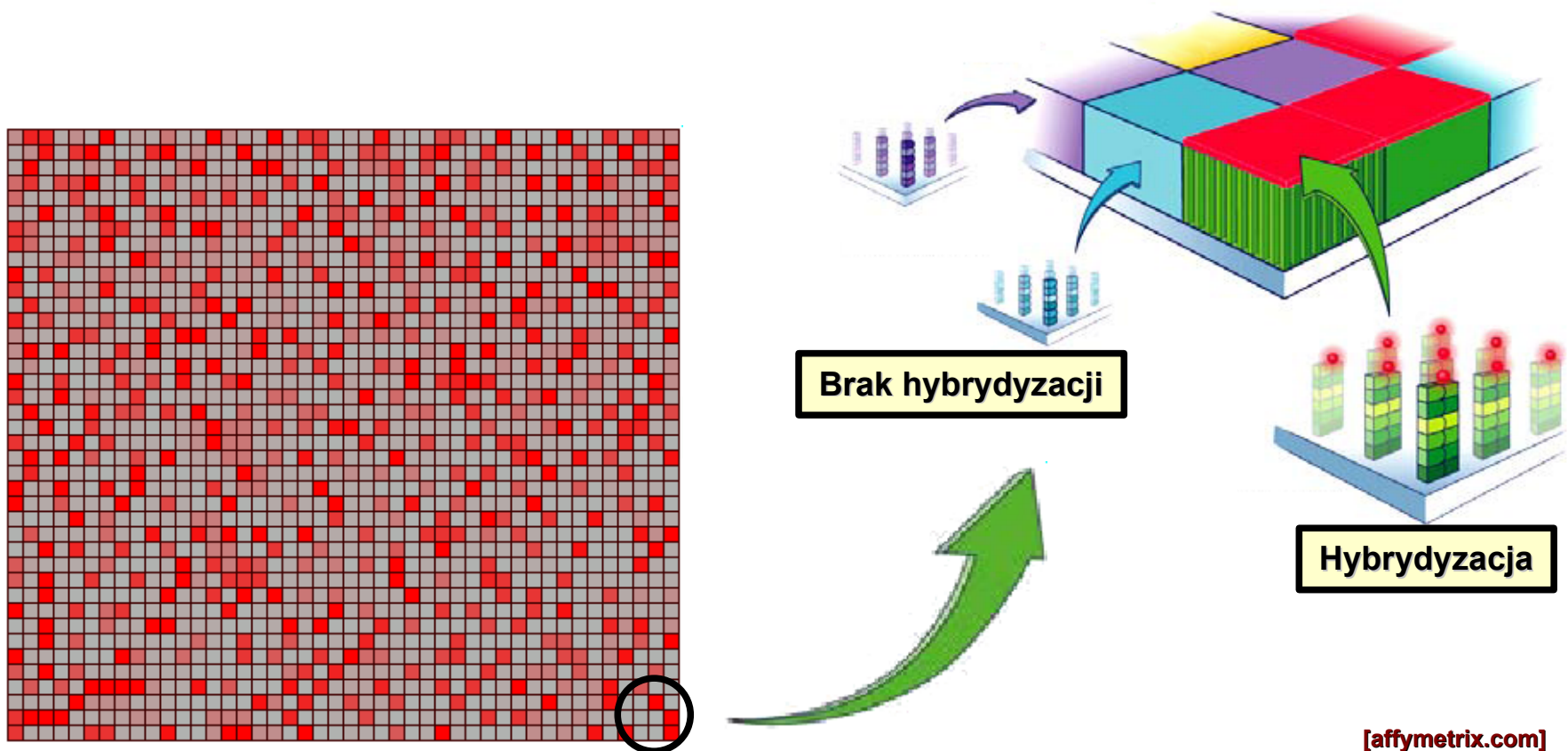
- jeżeli dany rodzaj mRNA nie występuje w próbkach, to reprezentująca go sonda pozostanie „pusta” (nie zhybrydują z nią żadne cząstki mRNA).



Mikromacierze DNA

Informacje o liczbie cząstek, które uległy hybrydyzacji z poszczególnymi sondami można uzyskać dzięki związanemu z mRNA **zaczniowi fluorescencyjnemu**.

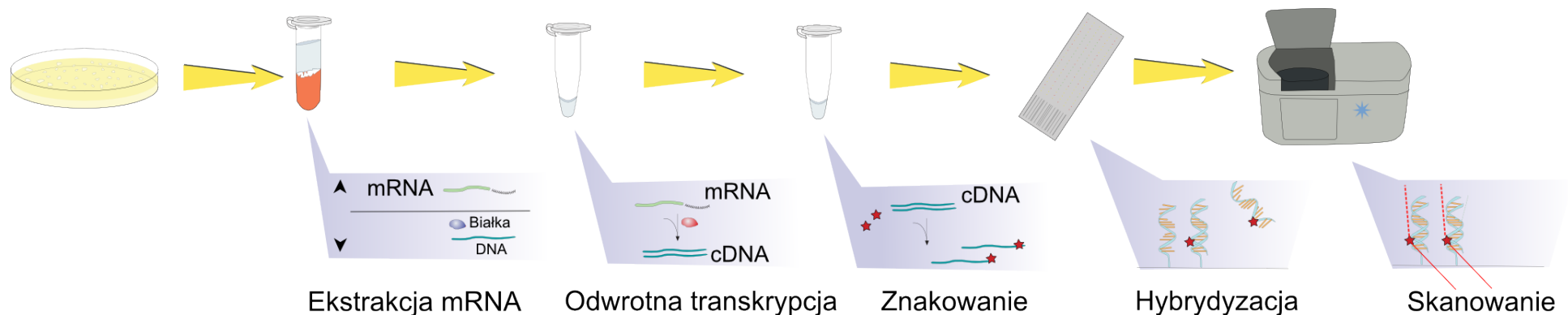
Odczyt mikromacierzy odbywa się poprzez wzbudzenie znacznika światłem lasera i rejestrację fluorescencji. **Bezpośrednim wynikiem pomiaru jest obraz**, w którym wartość każdego z pikseli jest proporcjonalna do natężenia fluorescencji z pewnego obszaru mikromacierzy.



Mikromacierze DNA

Ogólny schemat pomiarów przy użyciu mikromacierzy:

- **ekstrakcja mRNA** z badanej populacji komórek;
- **odwrotna transkrypcja mRNA**: synteza cząstek komplementarnego DNA (cDNA) na podstawie cząstek mRNA;
- denaturacja i **znakowanie fluorescencyjne** cDNA;
- **hybrydyzacja** cDNA z mikromacierzą;
- wypłukanie resztek niezhybrydyzowanego cDNA;
- odczyt mikromacierzy i **rejestracja obrazów fluorescencji**.

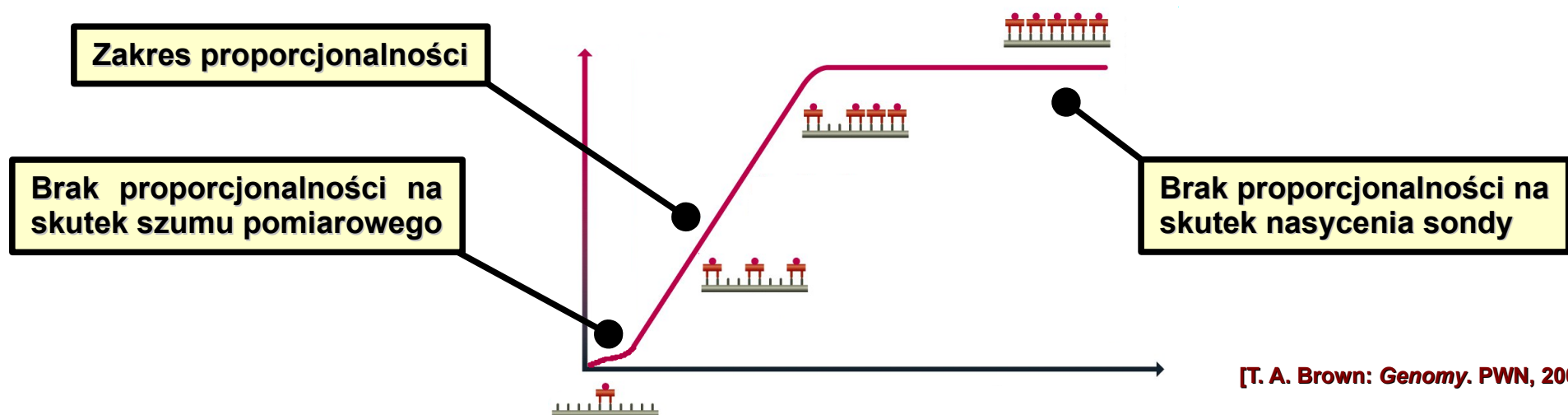


Schemat ten może ulegać pewnym – mniejszym bądź większym – modyfikacjom w zależności od celu eksperymentu i rodzaju stosowanych w nim mikromacierzy.

Ekspresyjne mikromacierze DNA

Celem pomiarów mikromacierzowych najczęściej jest badanie różnic w aktywności (w poziomach ekspresji) genów pomiędzy populacjami komórek, które reprezentują różne gatunki, tkanki, schorzenia, itd. Służące do tego mikromacierze nazywane są **ekspresyjnymi**. Ich działanie opiera się na następujących założeniach:

- fragmenty DNA w danej sondzie w możliwie jednoznaczny sposób reprezentują pojedynczy gen. Oznacza to, że powinny one **hybrydyzować specyficzn**ie jedynie z transkryptami wybranego genu;
- natężenie fluorescencji sondy reprezentującej gen jest, przynajmniej w pewnym zakresie, proporcjonalne do liczby kopii odpowiadającego mu mRNA w komórkach.

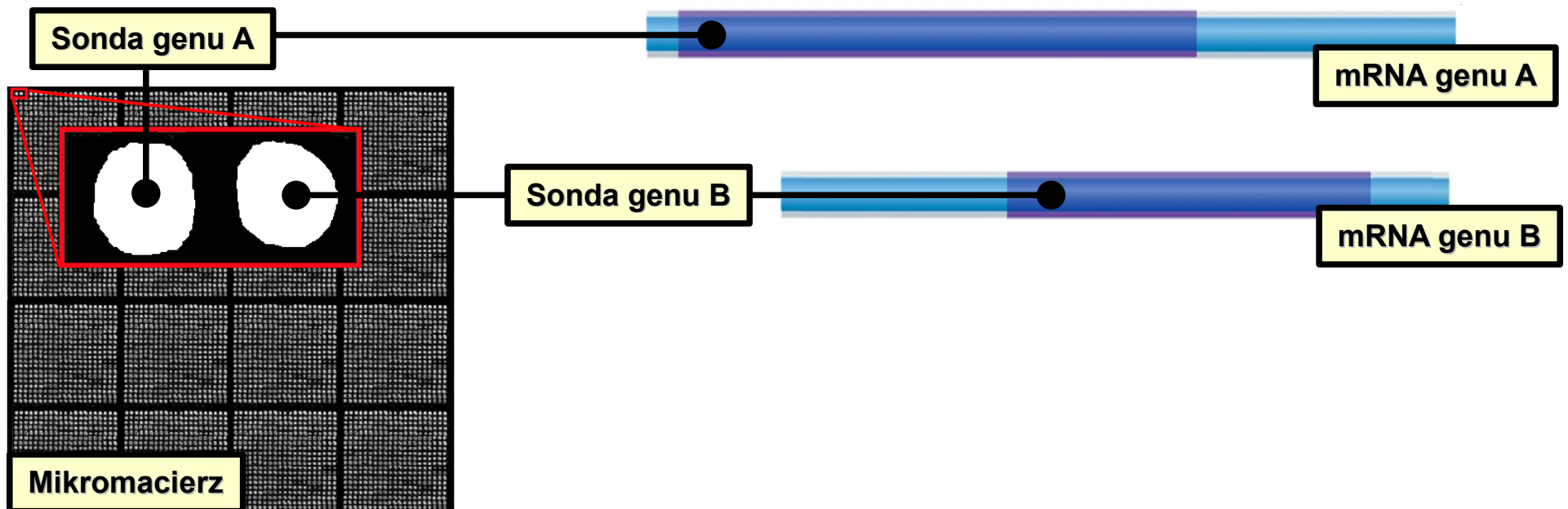


[T. A. Brown: *Genomy*. PWN, 2009]

Ekspresyjne mikromacierze cDNA

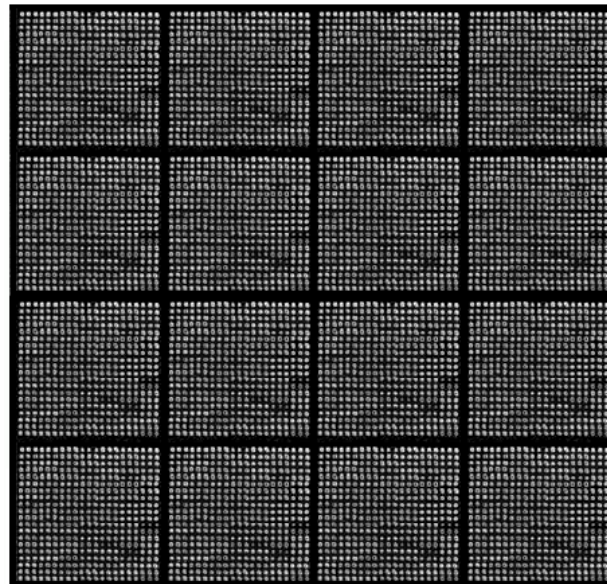
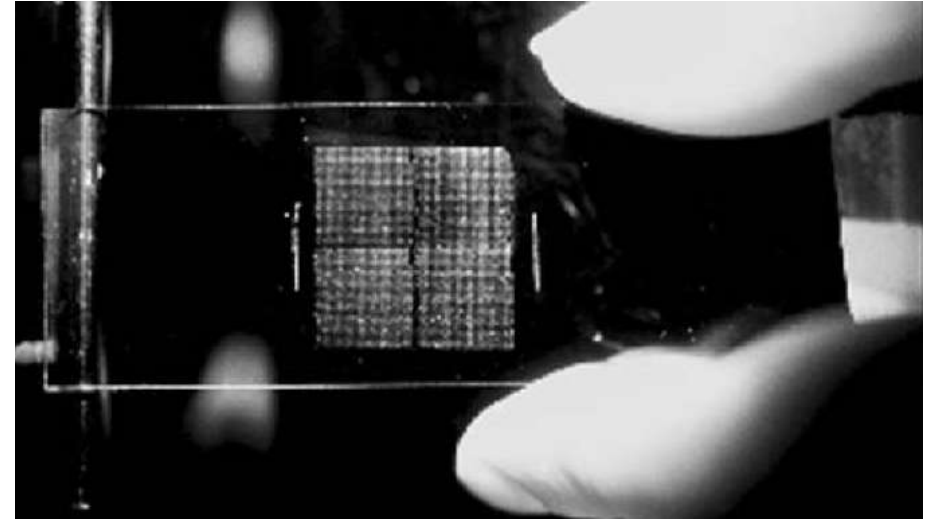
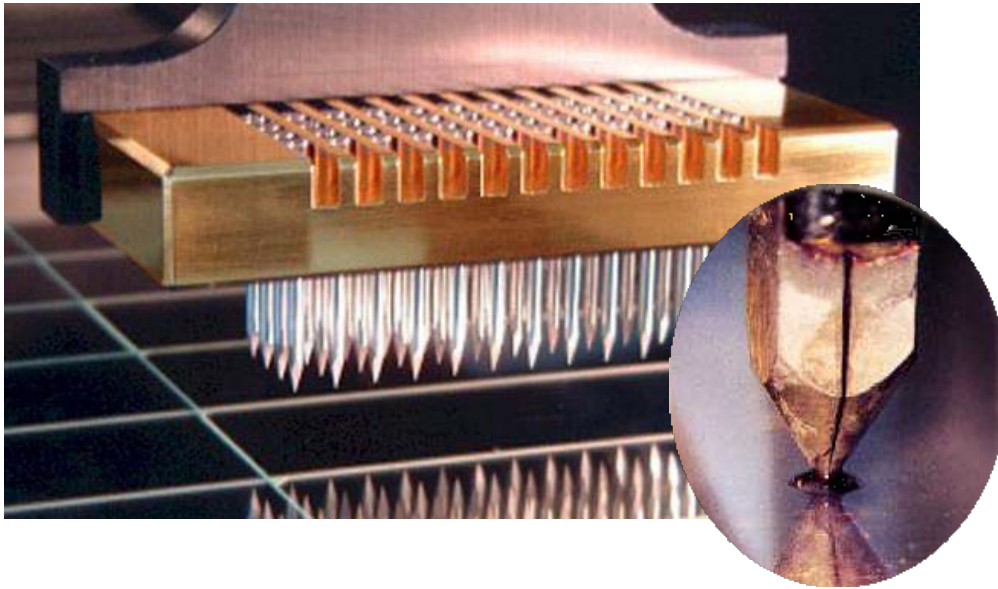
Charakterystyka **ekspresyjnych mikromacierzy cDNA**:

- w sondach znajdują się fragmenty cDNA pobrane z biblioteki klonów i powielone za pomocą PCR. Sekwencje cDNA w sondach mają długość 500-2000 nukleotydów, co oznacza, że obejmują znaczną część (lub nawet całość) sekwencji odpowiadających im genów. Wynikająca z tego duża specyficzność wiązania powoduje, że gen może być jednoznacznie reprezentowany przez pojedynczą sondę;
- typowa mikromacierz cDNA zawiera do kilku tysięcy sond rozmieszczonych na podłożu w postaci regularnej siatki. Obszary sond mają kształt w przybliżeniu kołisty i średnicę około 100µm.



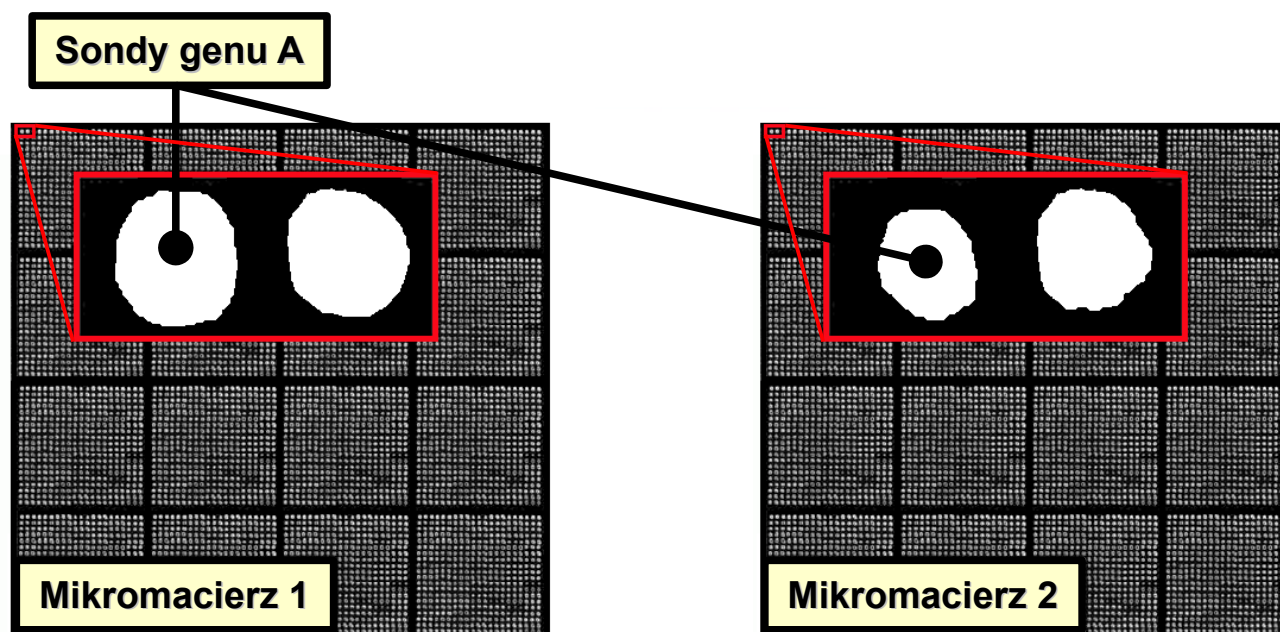
Ekspresyjne mikromacierze cDNA: budowa

Mikromacierze cDNA wytwarzane są poprzez maszynowe nakroplenie sond na podłoże, którym najczęściej jest szkło mikroskopowe o powierzchni pokrytej substancją wiążącą DNA.



Ekspresyjne mikromacierze cDNA: pomiar

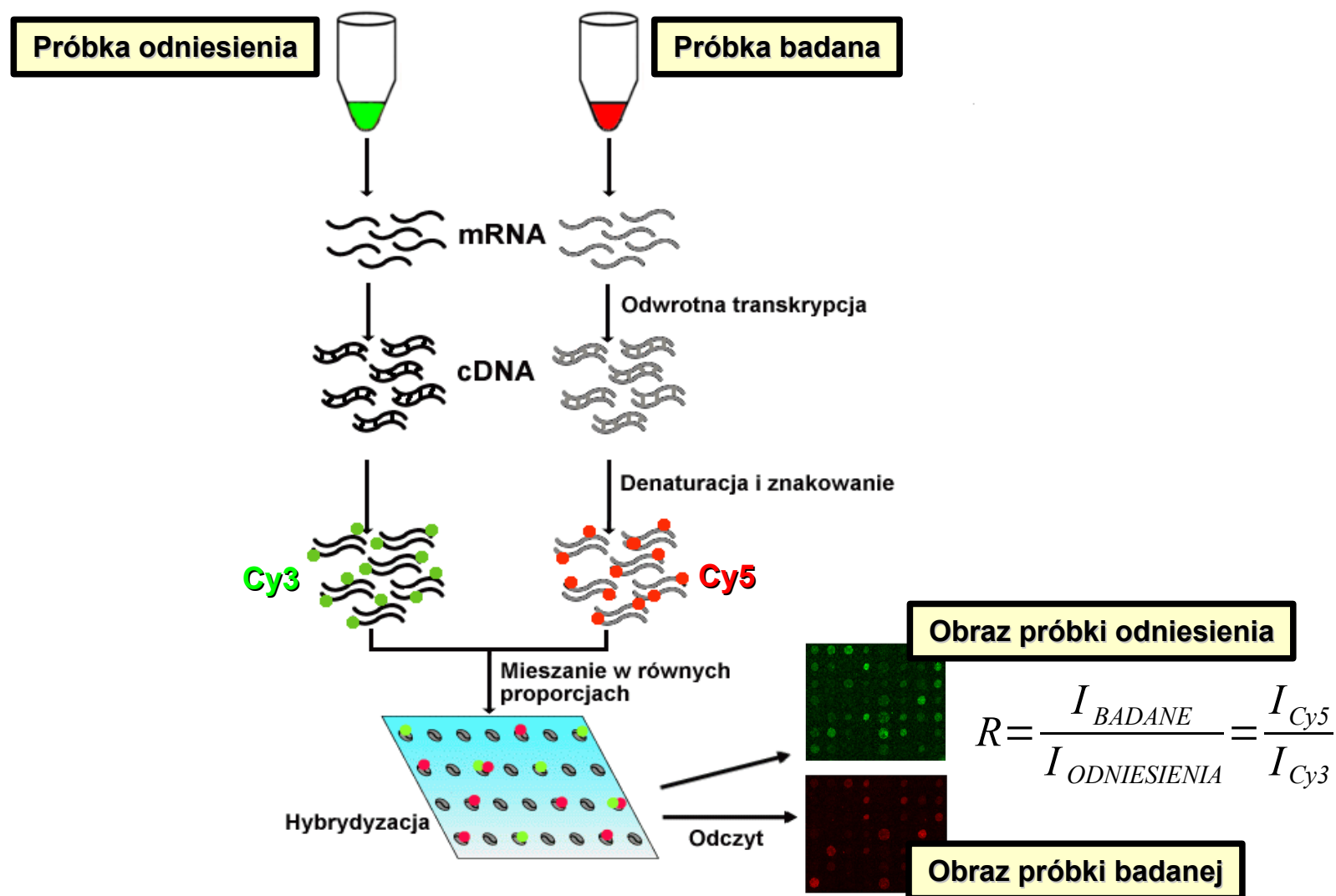
Ze względu na słabą powtarzalność procesu wytwarzania, sondy reprezentujące ten sam gen na różnych mikromacierzach cDNA mogą nie mieć jednakowych rozmiarów. Fakt ten znacząco utrudnia porównywanie wyników osobnych pomiarów.



Problem ten można ominąć stosując **pomiar różnicowy**, w którym z jedną mikromacierzą hybrydyzowane są dwie próbki mRNA: badana oraz referencyjna. Obie próbki znakowane są fluoroforami o różnych długościach fali emitowanego światła, np. **Cy3** ($\lambda_{wzb} \approx 550 \text{ nm}$, $\lambda_{em} \approx 570 \text{ nm}$) i **Cy5** ($\lambda_{wzb} \approx 650 \text{ nm}$, $\lambda_{em} \approx 670 \text{ nm}$).

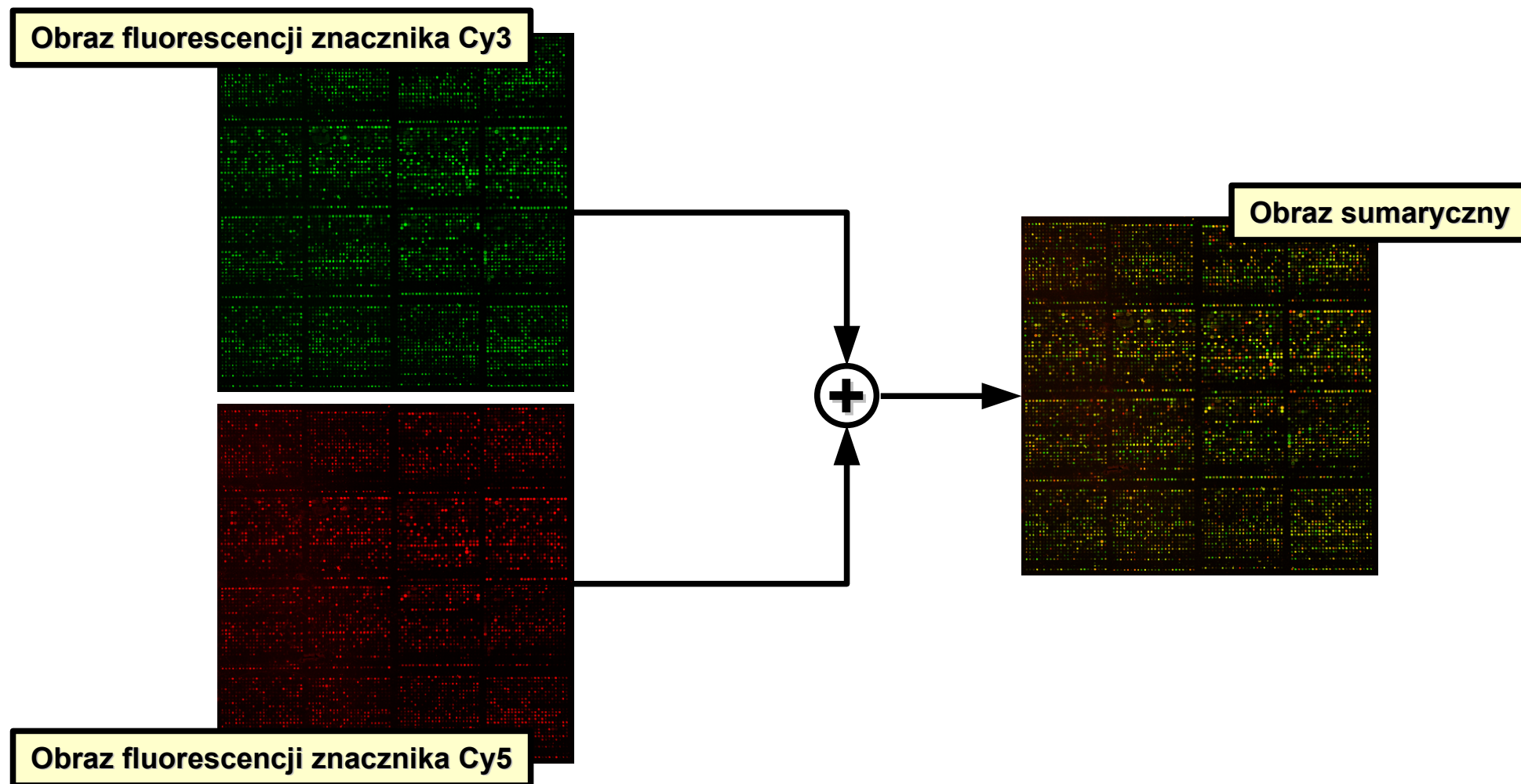
Ekspresyjne mikromacierze cDNA: pomiar

Odczyt odbywa się poprzez niezależne wzbudzenie znaczników i rejestrację ich fluorescencji. Celem pomiaru jest wyznaczenie dla każdej z sond stosunku sygnałów pochodzących od obu znaczników. Oznacza to, że ekspresja genów w próbce badanej określana jest w sposób względny do ich ekspresji w próbce referencyjnej (jednakowej dla wszystkich pomiarów w ramach eksperymentu).



Ekspresyjne mikromacierze cDNA: postać danych

Bezpośrednimi danymi pomiarowymi są dwa obrazy powstające w wyniku rejestracji fluorescencji znaczników próbki badanej i referencyjnej. Przy typowej rozdzielczości skanerów obrazy mają rozmiary $\sim 2000 \times 2000$ pikseli (sondy o średnicy ~ 10 pikseli). Obrazy zwykle zapisywane są w formacie TIFF z 16-bitową głębią koloru. Często przedstawiane są jako obraz sumaryczny o skali barwnej zgodnej z kolorami znaczników, ale przetwarzanie odbywa się osobno dla obu kanałów.



Zalety mikromacierzy cDNA:

- relatywnie niskie koszty wytwarzania;
- łatwość dostosowania zawartości sond do potrzeb konkretnego eksperymentu;
- zaprojektowanie mikromacierzy nie wymaga znajomości sekwencji genów.

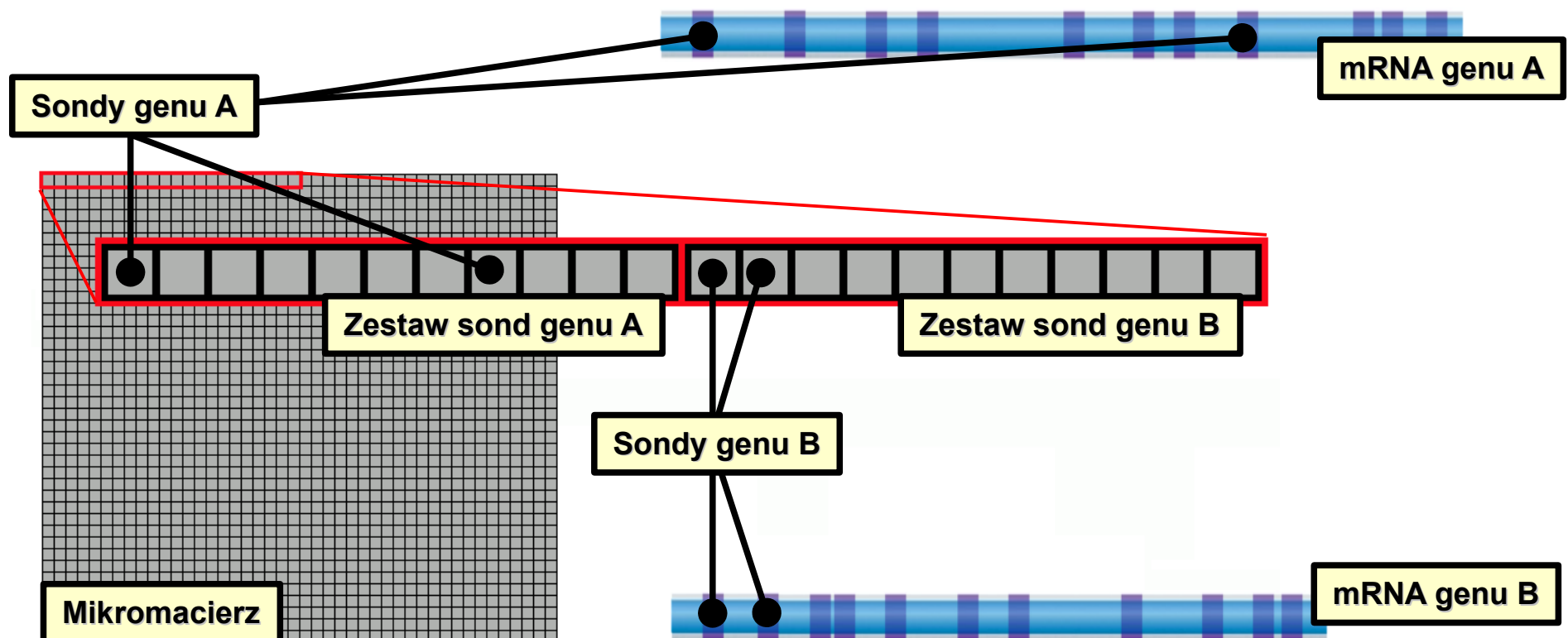
Wady mikromacierzy cDNA:

- mała gęstość i wynikająca z niej niewielka liczba sond;
- słaba powtarzalność procesu wytwarzania: konieczność wykonywania pomiarów różnicowych, większa złożoność procesu ekstrakcji wartości liczbowych z obrazów będących wynikiem pomiaru.

Oligonukleotydowe mikromacierze ekspresyjne

Charakterystyka **ekspresyjnych mikromacierzy oligonukleotydowych**:

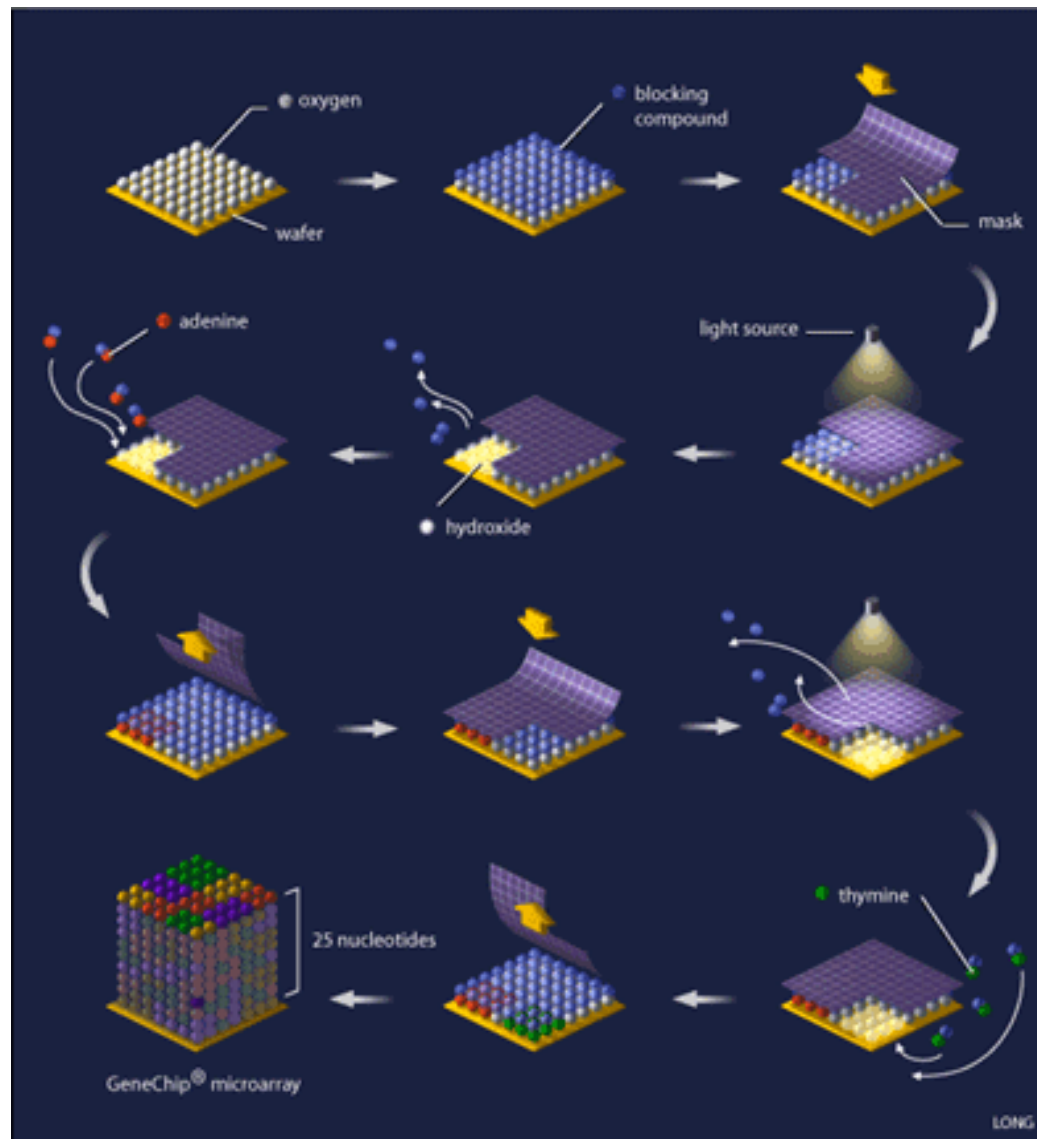
- sondy zawierają **oligonukleotydy** o długości kilkudziesięciu nukleotydów. Niewielka specyficzność wiązania do krótkich sekwencji powoduje, że pojedynczy gen musi być reprezentowany przez **zestaw sond (*probe set*)** zawierających oligonukleotydy komplementarne do różnych fragmentów jego sekwencji;
- oligonukleotydy syntetyzowane są bezpośrednio na podłożu, dzięki czemu sondy są silnie upakowane (może być ich nawet ponad milion) i mają regularne kształty.



Oligonukleotydowe mikromacierze ekspresyjne GeneChip: budowa

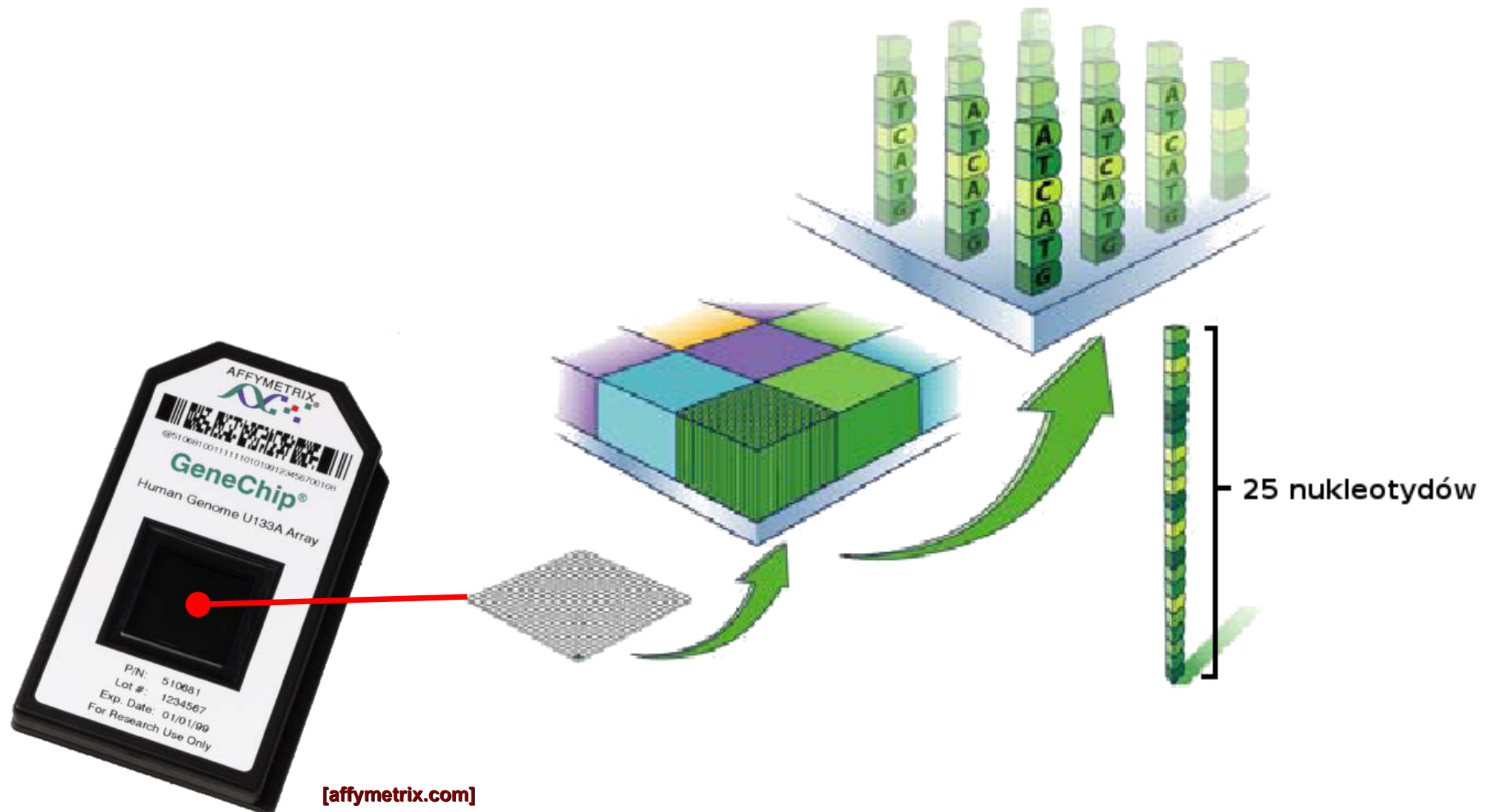
Mikromacierze oligonukleotydowe **GeneChip** firmy Affymetrix (obecnie Thermo):

- sondy zawierające oligonukleotydy o długości 25 nukleotydów syntetyzowane są na podłożu krzemowym w procesie wykorzystującym fotolitografię;



Oligonukleotydowe mikromacierze ekspresyjne GeneChip: budowa

- sondy mają kształt kwadratów o boku długości 5-24 μm (zależnie od modelu) rozmieszczone na obszarze o wymiarach 1,28 x 1,28 cm. Liczba sond może przekraczać milion, a w każdej z sond znajduje się około 10^7 oligonukleotydów. Cała mikromacierz zamknięta jest w standardowej kasetce ochronnej;



Oligonukleotydowe mikromacierze ekspresyjne GeneChip: budowa

- zestawy sond złożone są z 11-22 par sond (w zależności od modelu mikromacierzy);
- liczba i zawartość zestawów sond są zależne od organizmu, do badania którego przeznaczony jest dany model mikromacierzy oraz od technologii wykonania;

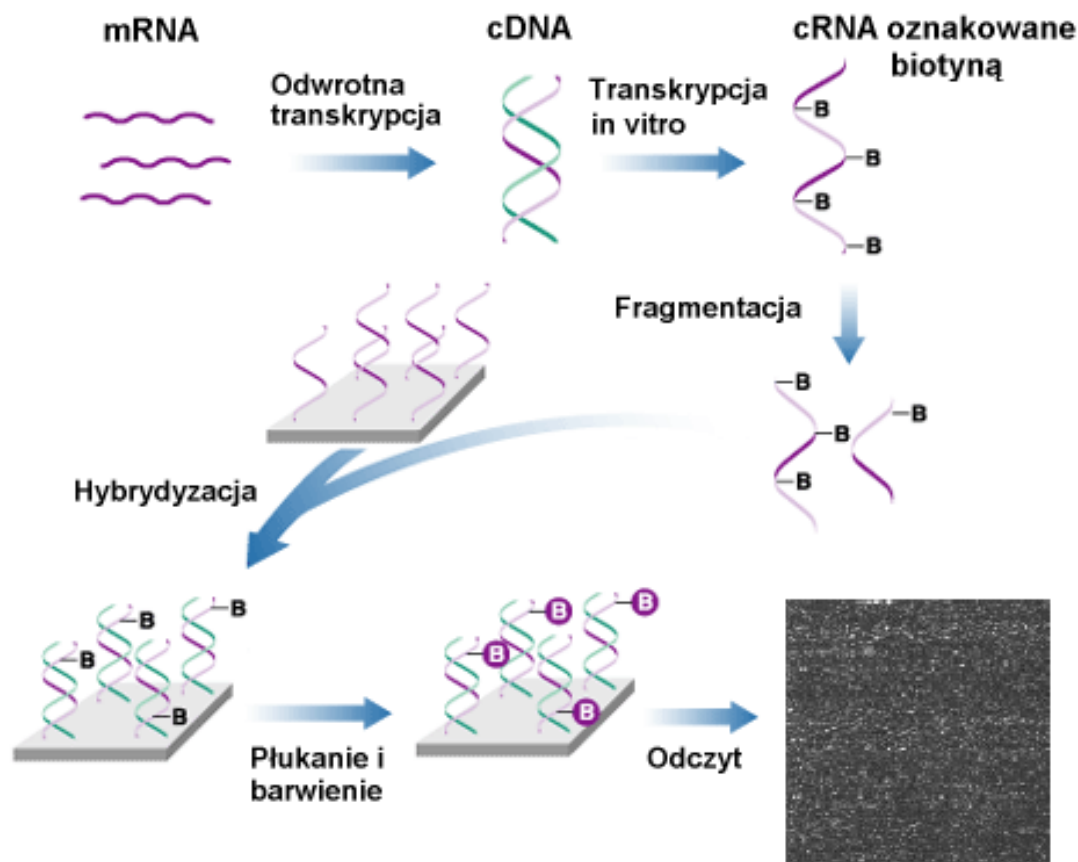
	Human Genome U95A 2.0	Human Genome U133 plus 2.0	Rat Genome 230 2.0
Rozmiar sondy	20µm	11µm	11µm
Liczba sond w zestawie	16	11	11
Liczba zestawów sond	12625	54675	31042
Liczba genów i ich wariantów	> 9600	> 38000	> 28000

- znaczna część genów reprezentowana jest przez więcej niż jeden zestaw sond (mogą być również uwzględnione różne warianty transkryptu, wynikające np. z alternatywnego składania mRNA);
- oprócz zestawów sond reprezentujących geny, na mikromacierzach znajdują się również zestawy sondy pozwalające zweryfikować jakość hybrydyzacji i stopień degradacji mRNA oraz sondy ułatwiające przetwarzanie obrazów fluorescencji.

Oligonukleotydowe mikromacierze ekspresyjne GeneChip: pomiar

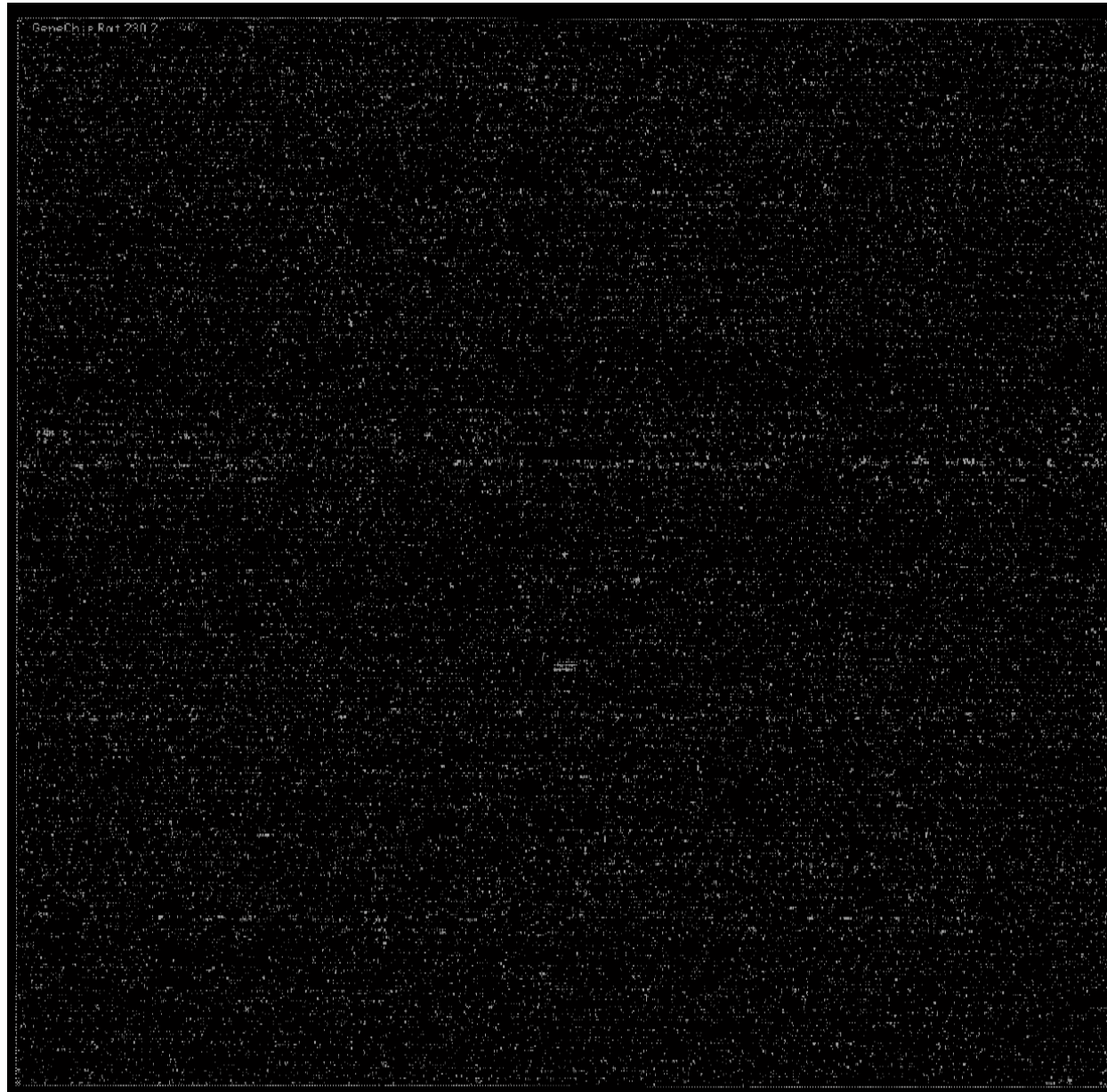
Różnice wobec ogólnego schematu przebiegu pomiarów mikromacierzowych:

- **transkrypcja *in vitro*, czyli synteza cRNA na matrycy cDNA** (sekwencje w sondach odpowiadają tylko nici kodującej w DNA genu);
- **fragmentacja cRNA na krótkie odcinki** (bo w sondach są krótkie oligonukleotydy);
- **cRNA łączy się z biotyną**, która po hybrydyzacji jest znakowana fluorescencyjnie.



Oligonukleotydowe mikromacierze ekspresyjne GeneChip: postać danych

Bezpośrednim wynikiem pomiaru jest 16-bitowy obraz zapisywany produkcyjnym formacie DAT, w którym pojedyncza sonda ma rozmiar 5x5 pikseli.



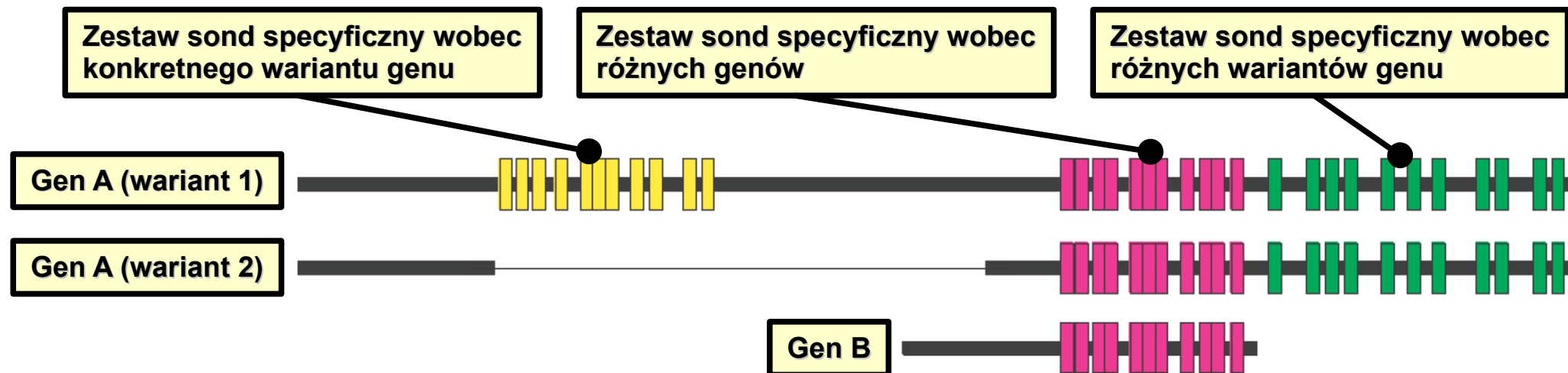
Oligonukleotydowe mikromacierze ekspresyjne GeneChip: zalety i wady

Zalety mikromacierzy oligonukleotydowych:

- duża liczba sond, reprezentujących większość genów badanych organizmów;
- ściśle kontrolowany proces wytwarzania zapewnia dobrą powtarzalność pomiarów.

Wady mikromacierzy oligonukleotydowych:

- stosunkowo wysokie koszty;
- producent może dostarczyć mikromacierze jedynie dla wybranych gatunków gdyż zaprojektowanie mikromacierzy wymaga znajomości sekwencji genów;
- nie wszystkie zestawy sond są w pełni specyficzne wobec genu, który reprezentują.

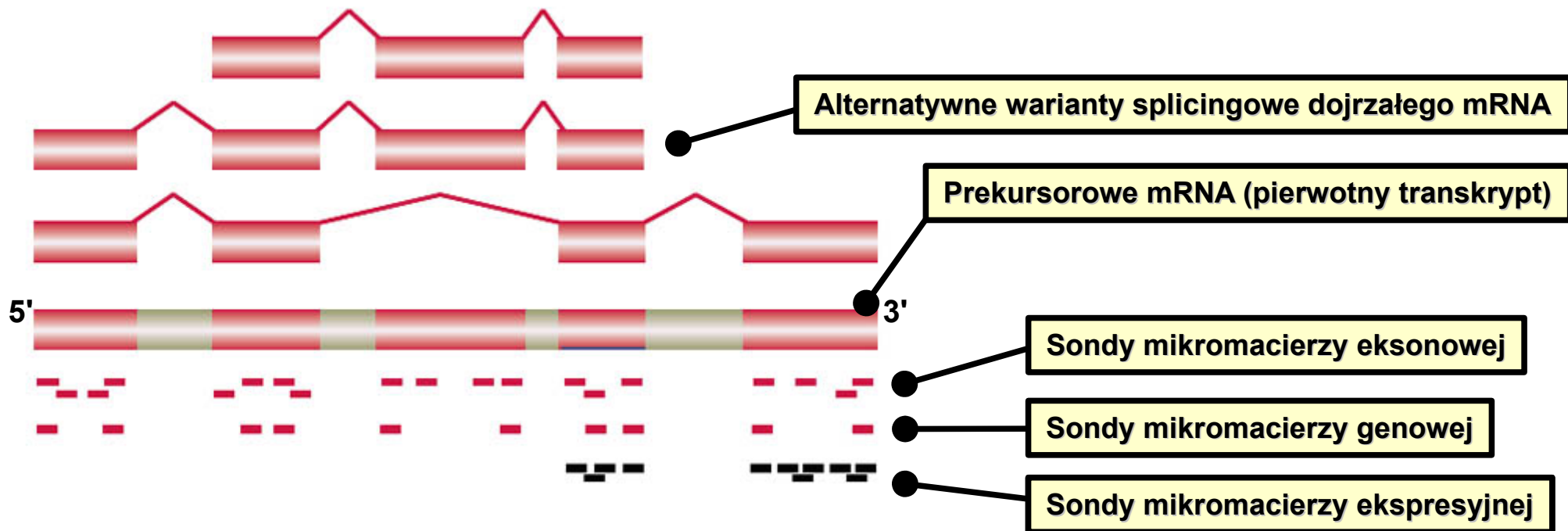


Oligonukleotydowe mikromacierze GeneChip: nie tylko badanie ekspresji

Jakkolwiek w ramach wykładu omawiane będą typowe mikromacierze ekspresyjne, należy pamiętać, że istnieją również ich specjalizowane odmiany.

Sondy klasycznych mikromacierzy ekspresyjnych GeneChip wybierane są jedynie z obszaru końca 3' sekwencji genu. Jest to wystarczające do pomiaru ekspresji, jednak nie pozwala rozróżnić odmian splicingowych transkryptów.

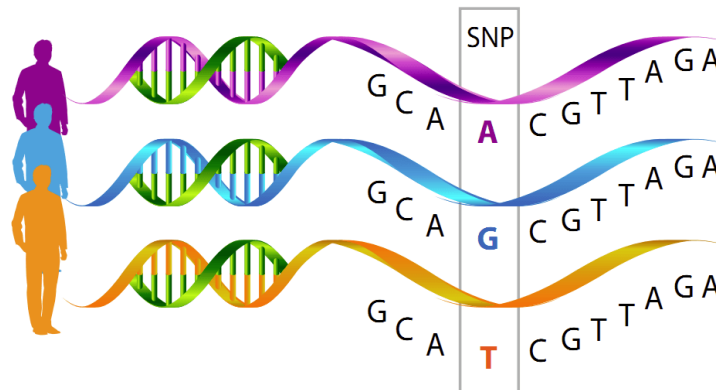
Do badania mechanizmu alternatywnego składania służą **mikromacierze eksonowe** (**exon microarrays**), w których każdy z eksonów genu reprezentowany jest przez równą liczbę sond (zwykle po 4). Produkowane są też ich prostsze wersje o mniejszym pokryciu sekwencji eksonów (**mikromacierze genowe, gene microarrays**).



Oligonukleotydowe mikromacierze GeneChip: nie tylko badanie ekspresji

Mikromacierze GeneChip znajdują również zastosowanie w badaniach nad budową i funkcjami genomu (czyli poza zakresem transkryptomiki). Przykładami mogą być:

- **mikromacierze SNP** umożliwiają badanie **polimorfizmów pojedynczych nukleotydów** (*Single Nucleotide Polimorphism – SNP*), czyli występujących w ramach populacji różnych wariantów danej pozycji genomu. Zawierają do miliona sond o sekwencjach pochodzących ze znanych miejsc występowania SNP w genomie;



[neuroendoimmune.wordpress.com]

- **mikromacierze tilingowe** (*tiling microarrays*) zawierają sondy pokrywające kolejne fragmenty ciągłego obszaru genomu. Mogą służyć m.in. do poszukiwania obszarów genomu ulegających ekspresji oraz sekwencji regulatorowych (np. miejsc wiązania się czynników transkrypcyjnych z DNA).



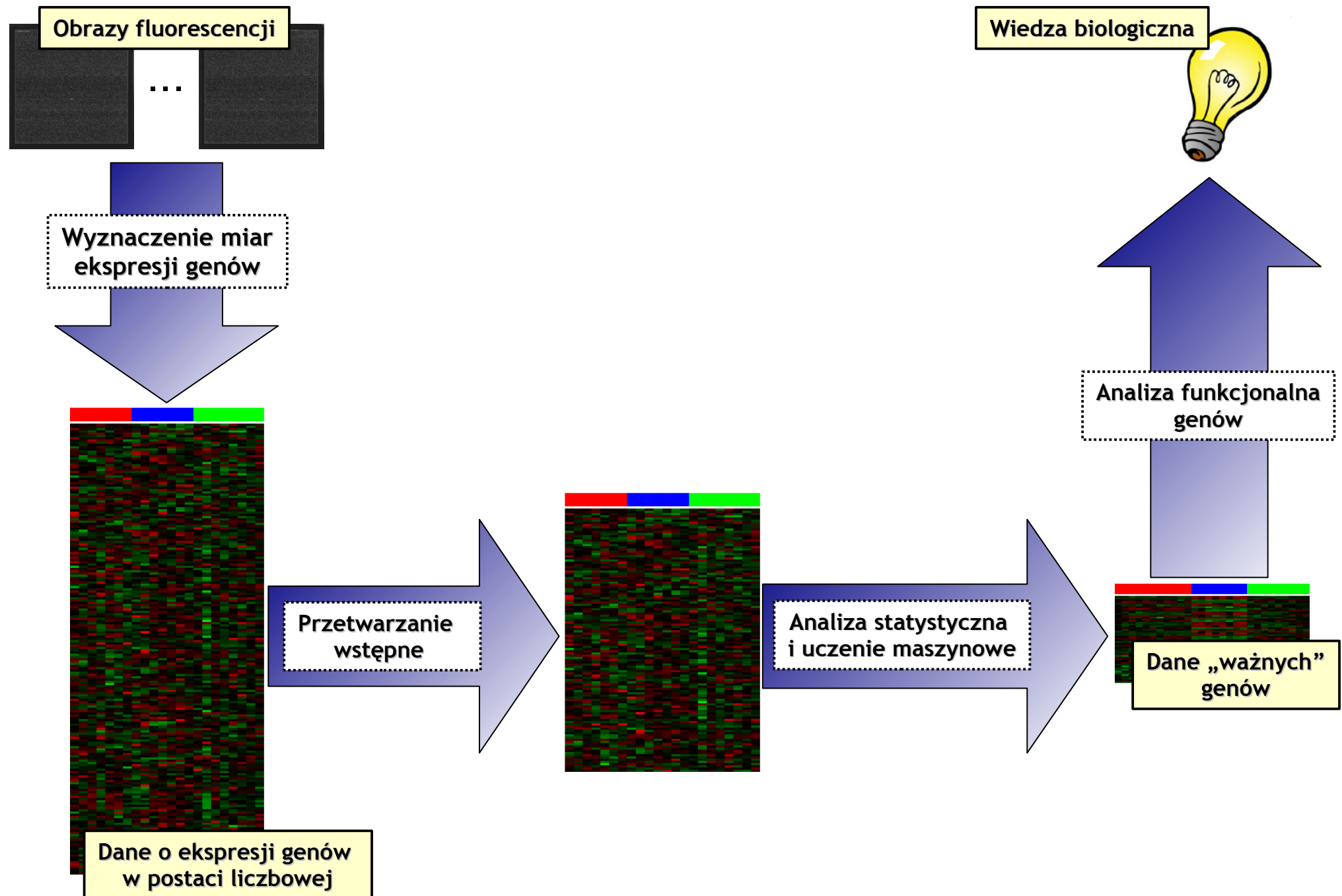
Uczenie maszynowe w bioinformatyce

Wykład 5: mikromacierze (ogólny schemat analizy danych)

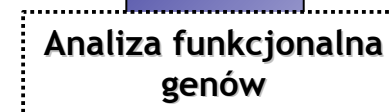
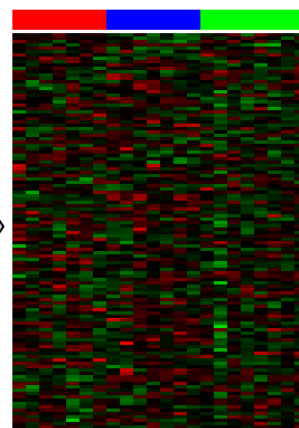
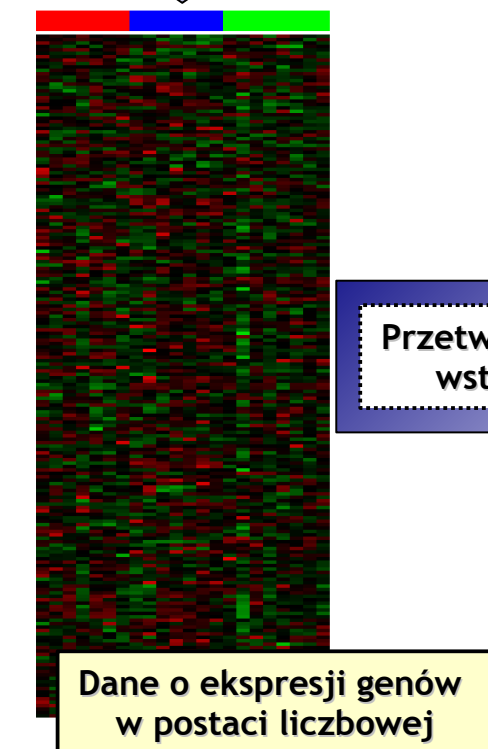
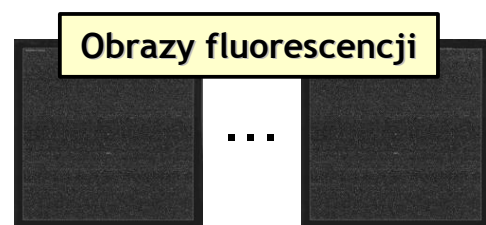
Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Przetwarzanie danych z mikromacierzy ekspresyjnych



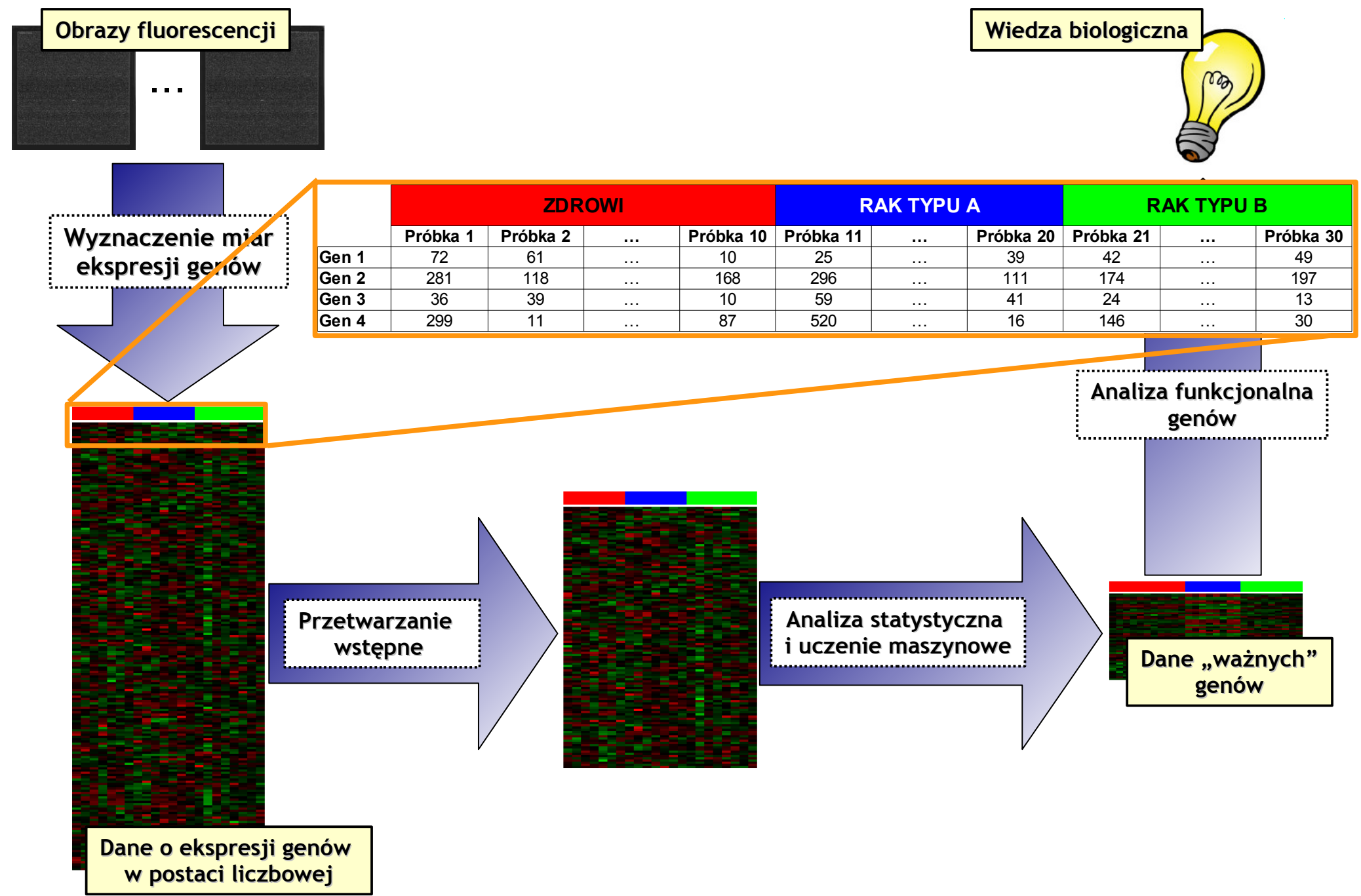
Przetwarzanie danych z mikromacierzy ekspresyjnych



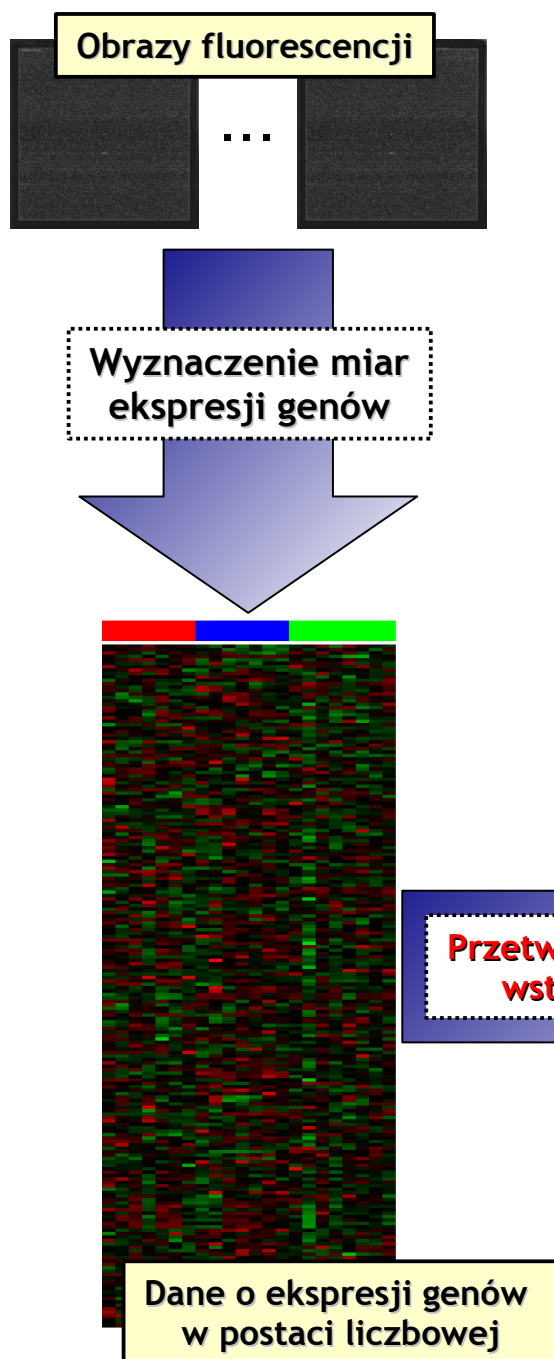
Celem **przetwarzania niskiego poziomu** jest określenie wartości liczbowych miar aktywności (poziomu ekspresji) genów we wszystkich próbkach.

Czynności wykonywane na etapie przetwarzania niskiego poziomu są ściśle uzależnione od typu stosowanych mikromacierzy (często są realizowane automatycznie przez oprogramowanie producenta systemu pomiarowego, co szczególnie dotyczy etapu obróbki obrazów). Nie zależą one natomiast od biologicznego celu eksperymentu.

Przetwarzanie danych z mikromacierzy ekspresyjnych



Przetwarzanie danych z mikromacierzy ekspresyjnych



Przetwarzaniem wstępnym określa się ogół czynności mających na celu poprawę jakości danych (eliminację artefaktów pomiarowych oraz wpływu czynników o niebiologicznym pochodzeniu) i doprowadzenie ich do postaci ułatwiającej dalszą analizę. Przykładowo są to:

- eliminacja brakujących i odstających wartości;
- ograniczenie zakresu wartości i zmiana ich skali;
- poprawa stosunku sygnału do szumu danych;
- normalizacja;
- usunięcie cech nieistotnych z punktu widzenia analizy.

Kroki przetwarzania wstępnego i stosowane metody są zazwyczaj niezależne od celu dalszej analizy danych. Zwykle też na tym etapie nie wykorzystuje się wiedzy o przynależności próbek do grup badanych.



Przetwarzanie danych z mikromacierzy ekspresyjnych

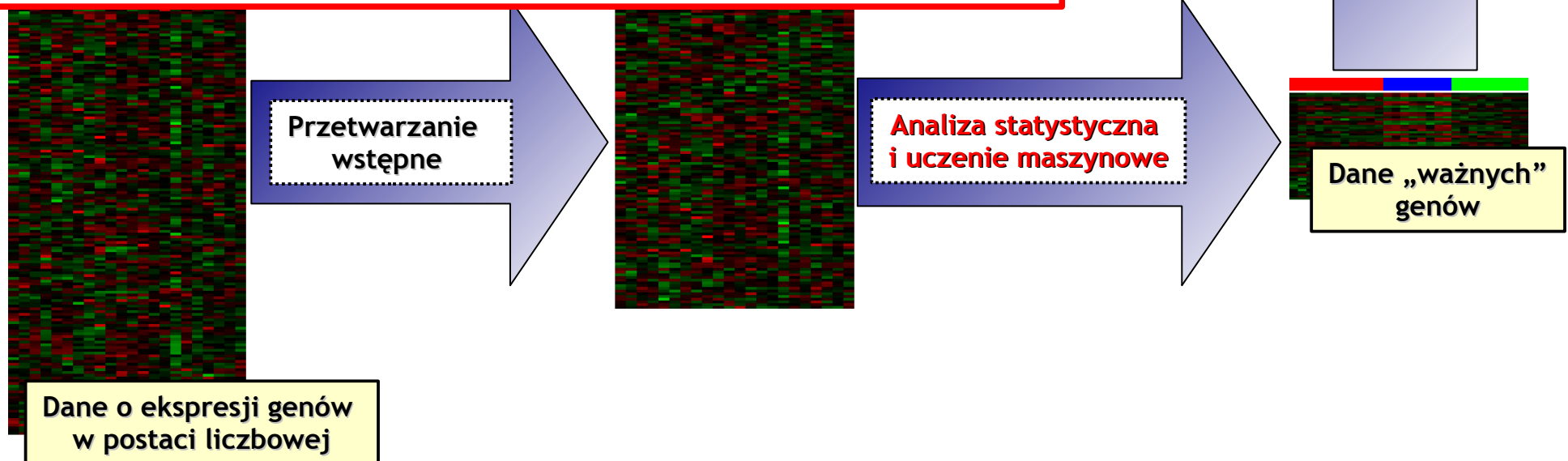
Techniki stosowane na etapie analizy danych są ściśle uzależnione od celu badań. W ogólności można je podzielić na dwie grupy:

Metody działające bez nadzoru - używane do wykrycia zależności pomiędzy próbkami lub genami jedynie w oparciu o dane, bez korzystania z ich opisu (w szczególności, nie uwzględniając sposobu przypisania badanych obiektów do zadanych z góry klas). Do metod nienadzorowanych należą m.in.:

- klasteryzacja;
- techniki redukcji wymiarowości.

Metody nadzorowane - działające w oparciu o znane przypisanie obiektów do jednej z wcześniej ustalonych klas. Przykładami zagadnień, w których używa się metod nadzorowanych są:

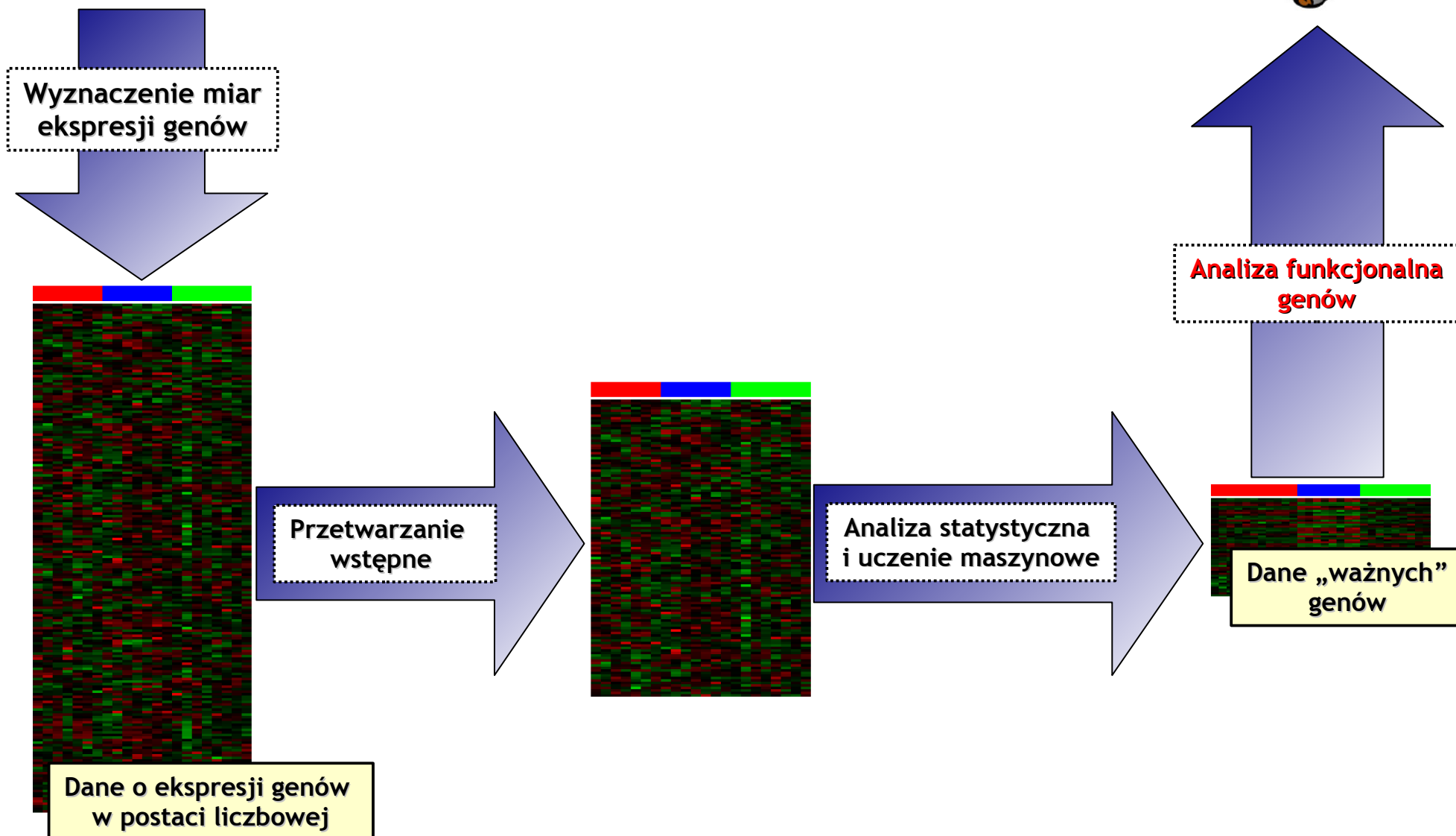
- klasyfikacja;
- selekcja cech różnicujących grupy próbek.



Przetwarzanie danych z mikromacierzy ekspresyjnych

Określenie funkcji biologicznej zbioru genów wybranych na podstawie wcześniejszych etapów analizy danych z pomiarów mikromacierzowych.

Odbyna się ona przy wykorzystaniu baz danych gromadzących informacje o genach i białkach oraz w oparciu o dane literaturowe i wiedzę ekspercką.



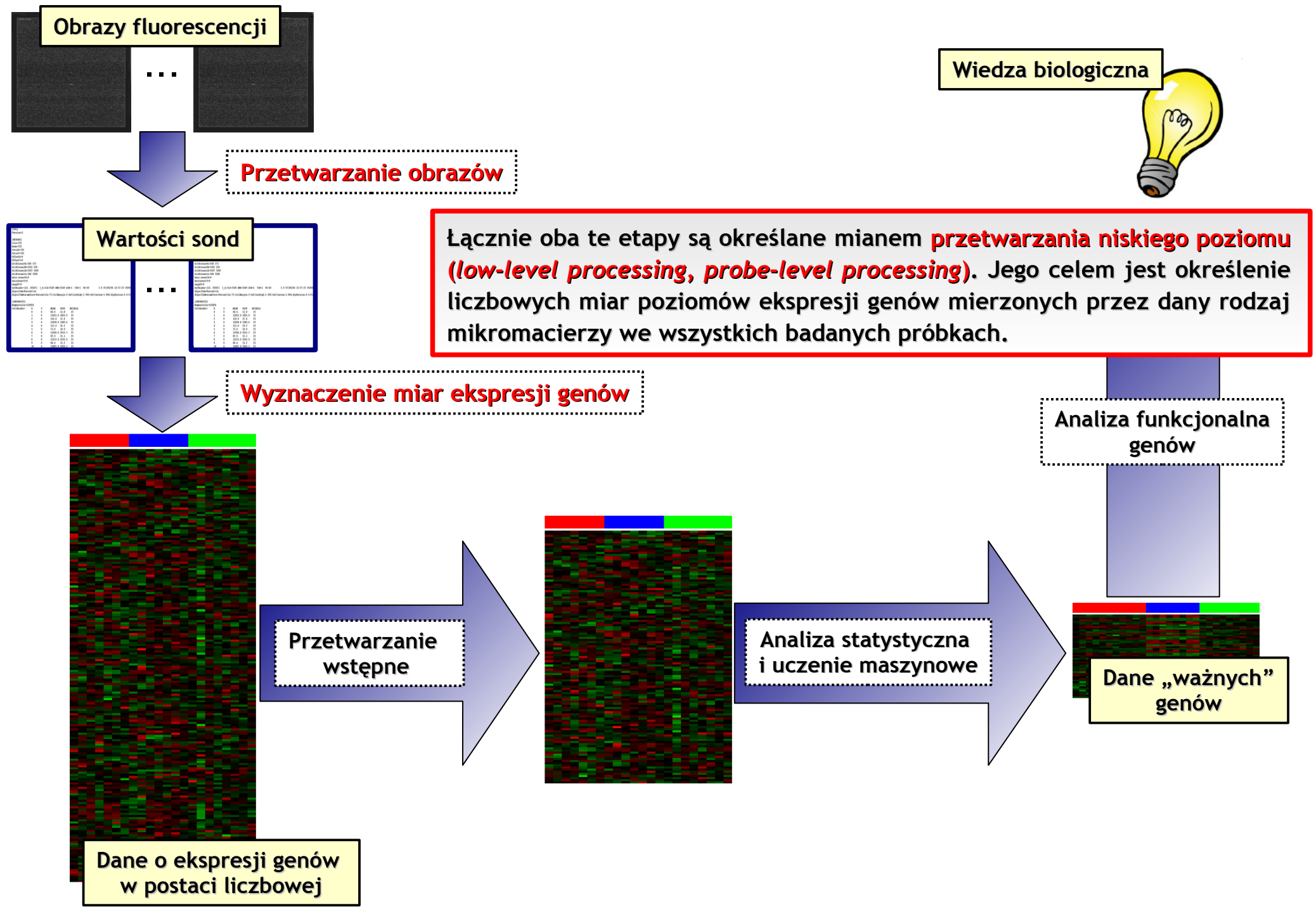
Uczenie maszynowe w bioinformatyce

Wykład 5: mikromacierze (przetwarzanie niskopoziomowe)

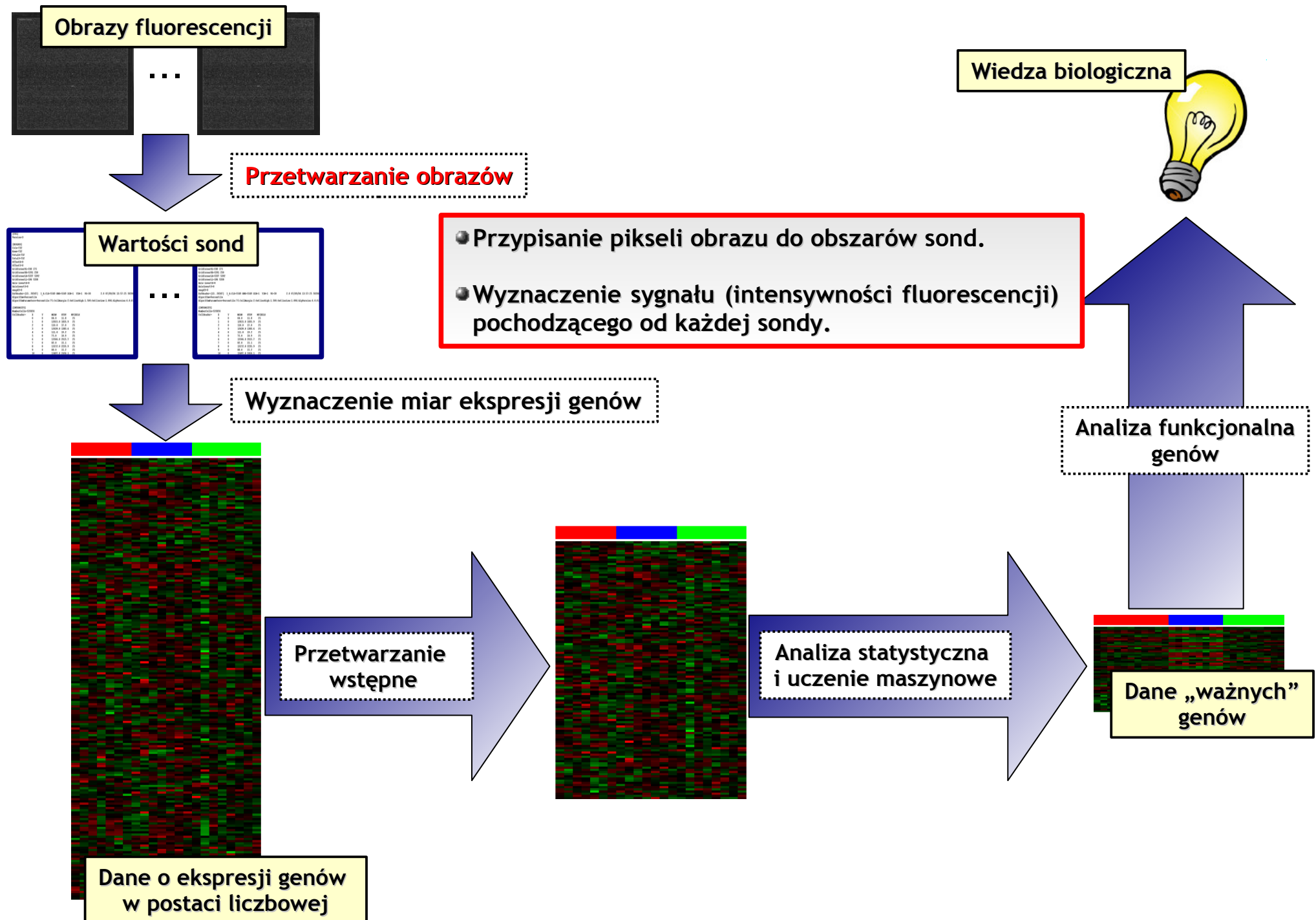
Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Przetwarzanie niskopoziomowe danych z mikromacierzy ekspresyjnych



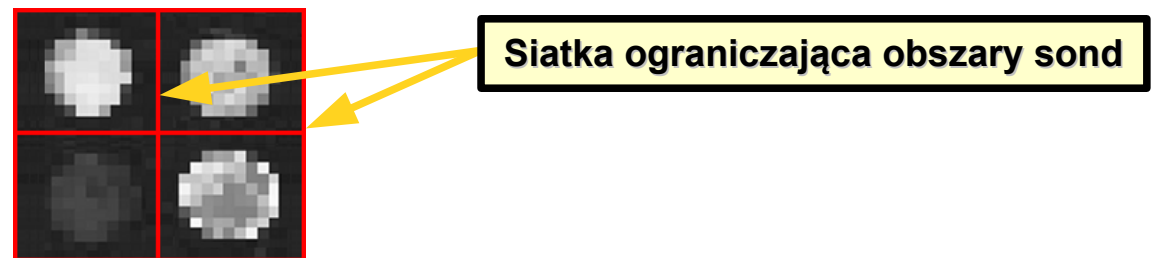
Przetwarzanie niskopoziomowe danych z mikromacierzy ekspresyjnych



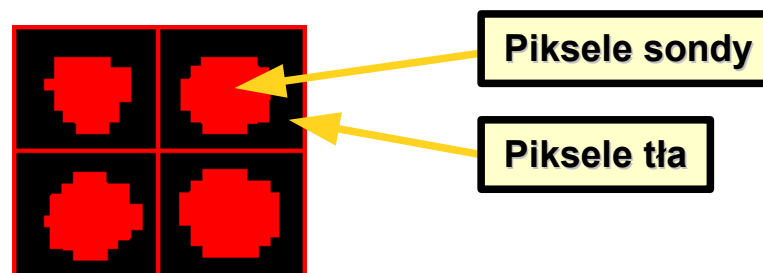
Przetwarzanie obrazów z mikromacierzy cDNA

Dla mikromacierzy cDNA przetwarzanie odbywa się niezależnie dla obrazów obu znaczników fluorescencyjnych (połączenie uzyskanych z nich informacji następuje dopiero na dalszym etapie). Dla obu kanałów proces przetwarzania obejmuje:

- adresowanie sond (nałożenie siatki wyznaczającej pozycję każdej z sond);



- segmentację (przypisanie pikseli do obszarów sond i podłoża);

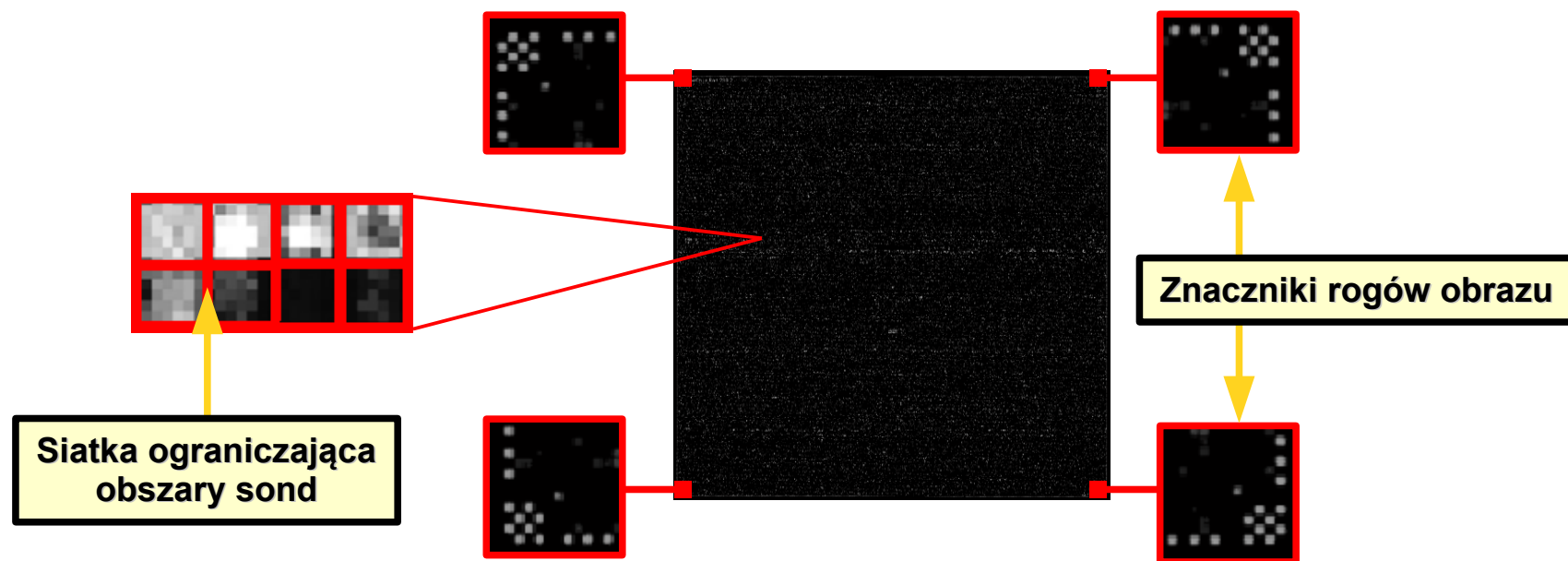


- wyznaczenie sygnału fluorescencji sondy, np. jako sumę intensywności pikseli przypisanych w obrazie do jej obszaru.

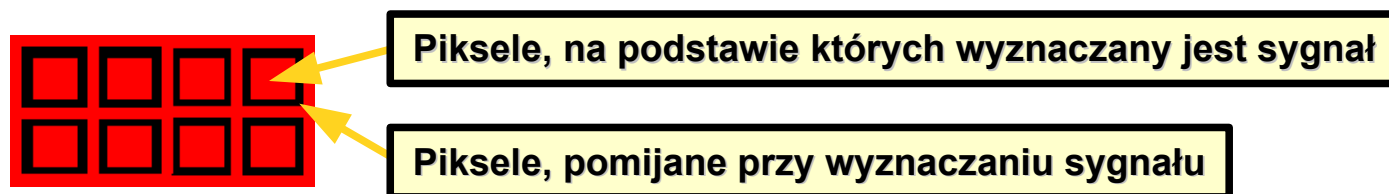
Przetwarzanie obrazów z mikromacierzy GeneChip

Przetwarzanie obrazów w przypadku mikromacierzy GeneChip ma w ogólności podobny przebieg, jest jednak znacząco ułatwione przez regularną geometrię sond:

- pozycje sond określone są podstawie znaczników rogów obrazu, którymi są specjalne sondy o dużej intensywności;

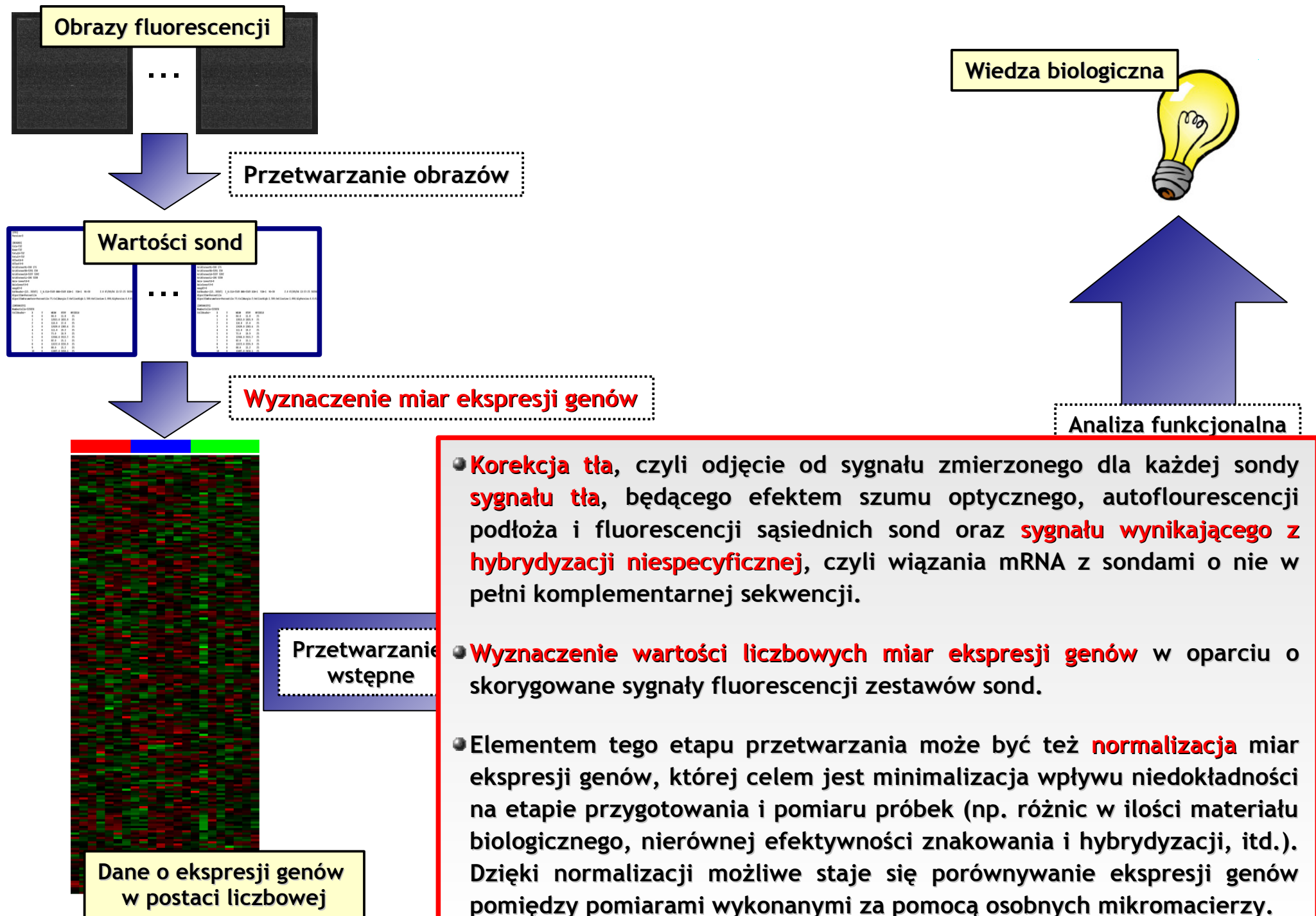


- segmentacja ogranicza się do odrzucenia skrajnych pikseli z obszarów sondy;



- sygnał wyznaczany jest jako 75 percentyl wartości pozostałych pikseli.

Przetwarzanie niskopoziomowe danych z mikromacierzy ekspresyjnych

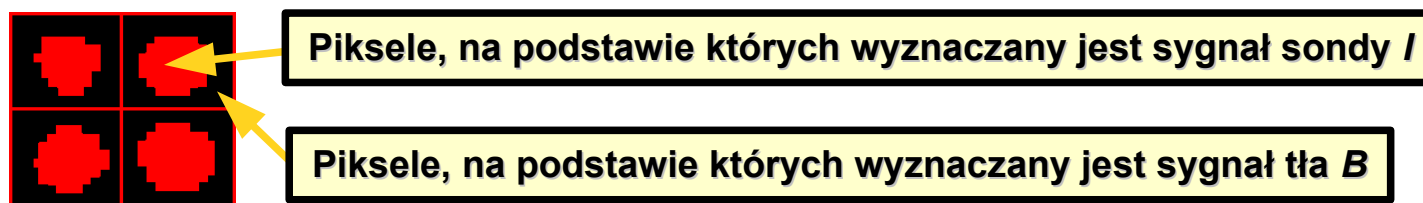


Wyznaczanie miar ekspresji dla mikromacierzy cDNA

Celem korekcji tła jest oszacowanie wartości sygnału zakłócającego B , tak aby możliwe było ich odjęcie od sygnału zmierzonego.



W przypadku mikromacierzy cDNA lokalny poziom tła B może być oszacowany w oparciu o sygnał zarejestrowany w otoczeniu sondy. Korekcja tła polega na odjęciu tak wyznaczonej wartości B od sygnału I , pochodzącego z obszaru sondy.



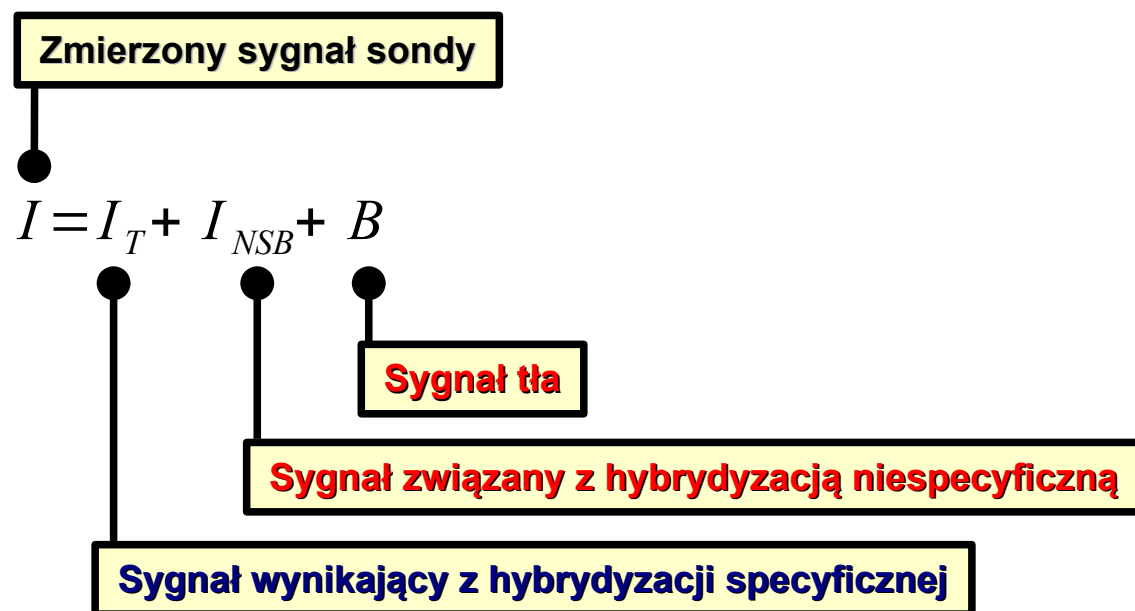
Miarą względnej ekspresji i -tego genu staje się stosunek sygnałów jego sond w obu obrazach (po korekcji tła), najczęściej wyrażany w skali logarytmicznej:

$$x_i = \log_2 \left(\frac{I_{\text{RED}i} - B_{\text{RED}i}}{I_{\text{GREEN}i} - B_{\text{GREEN}i}} \right)$$

Wyznaczanie miar ekspresji dla mikromacierzy GeneChip

Sondy mikromacierzy oligonukleotydowych przylegają bezpośrednio do siebie i nie są rozdzielone obszarami, które mogłyby posłużyć do oceny sygnału tła. Dlatego też korekcja tła musi odbywać się w oparciu o sygnał zmierzony przez sondy. Ponadto, ze względu na małą specyficzność wiązania krótkich sekwencji duże znacznie ma korekcja sygnału sond PM pod kątem niespecyficznego hybrydowania.

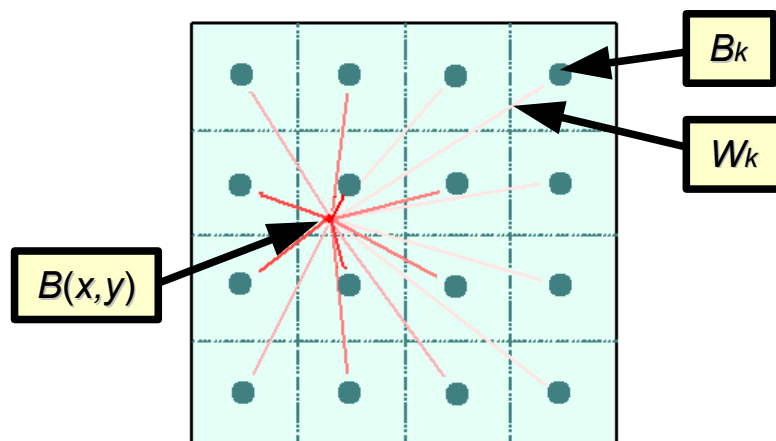
Zakłada się, że dla każdej sondy zmierzony sygnał jest sumą trzech składników:



Zadaniem algorytmów przetwarzania niskopoziomowego jest oszacowanie wartości sygnału B (związanego z tłem) i oraz I_{NSB} (wynikającego z niespecyficznego hybrydowania), tak aby możliwe było ich odjęcie od sygnału zmierzonego. Istnieje wiele różnych algorytmów realizujących ten etap obróbki – jako przykład opisany zostanie **algorytm zaimplementowany w oprogramowaniu MAS 5.0 firmy Affymetrix**.

Procedura korekcji tła:

- mikromacierz dzielona jest na K segmentów (zwykle $K = 16$)
- dla każdego segmentu wybieranych jest 2% sond o najmniejszych wartościach zarejestrowanego sygnału. Poziom tła B_k związany z segmentem określany jest jako średnią arytmetyczną wartości sygnałów zmierzonego dla tych sond
- poziom tła $B(x, y)$ dla sondy o pozycji (x, y) określany jest jako średnia ważona poziomów tła B_k poszczególnych segmentów mikromacierzy, przy czym wagami $w_k(x, y)$ są odwrotności kwadratów odległości Euklidesa pomiędzy pozycją sondy (x, y) a pozycjami środków segmentów (x_k, y_k)



$$B(x, y) = \frac{\sum_1^K B_k w_k(x, y)}{\sum_1^K w_k(x, y)}$$

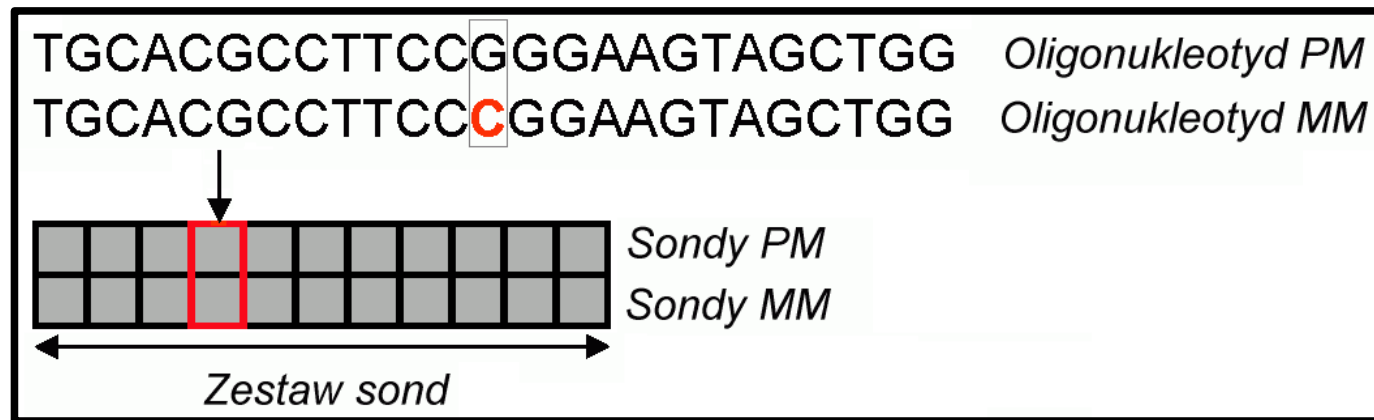
$$w_k(x, y) = \frac{1}{d_k^2(x, y) + 100} = \frac{1}{((x_k - x)^2 + (y_k - y)^2) + 100}$$

- poziom tła $B(x, y)$ odejmowany jest od zmierzonej wartości dla sondy $I(x, y)$:

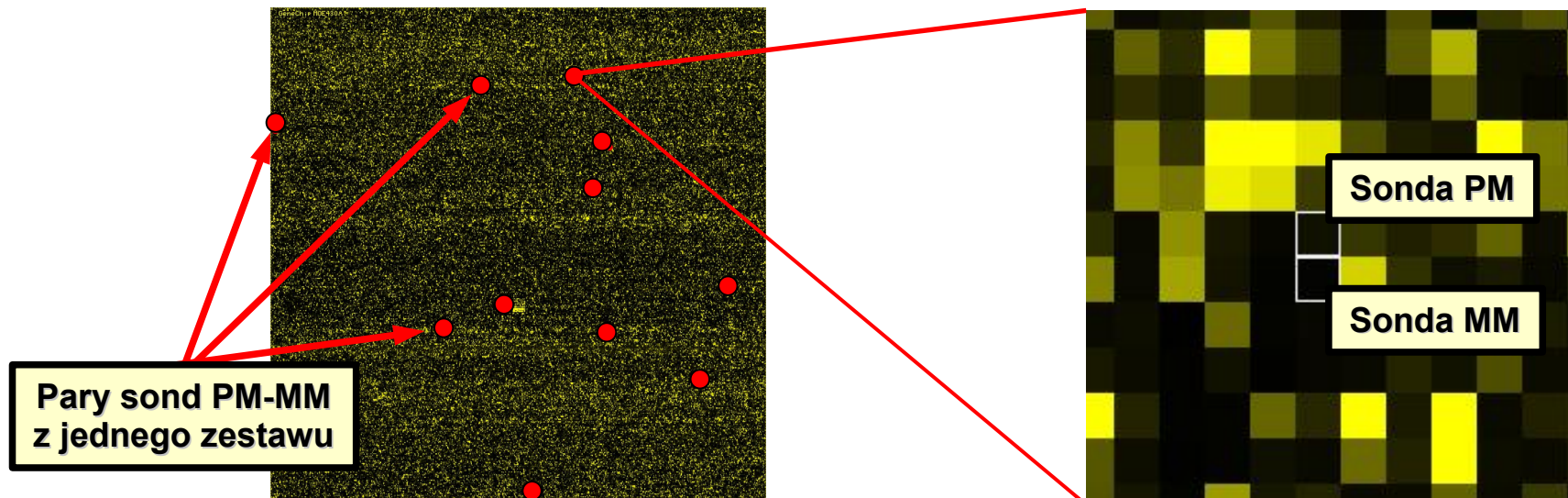
$$I^*(x, y) = I(x, y) - B(x, y)$$

Wyznaczanie miar ekspresji dla mikromacierzy GeneChip: alg. MAS 5.0

Zestawy sond (*probe sets*) złożone są z 11-22 par sond dwóch rodzajów: **Perfect Match (PM)** oraz **Mismatch (MM)**. Sekwencje sondy PM są zgodne z sekwencjami odpowiednich fragmentów nici kodującej genu. Sondy MM różnią się od sond PM nukleotydem na środkowej pozycji (jest on zastąpiony nukleotydem komplementarnym) i **służą do szacowania poziomu hybrydyzacji niespecyficznej**.

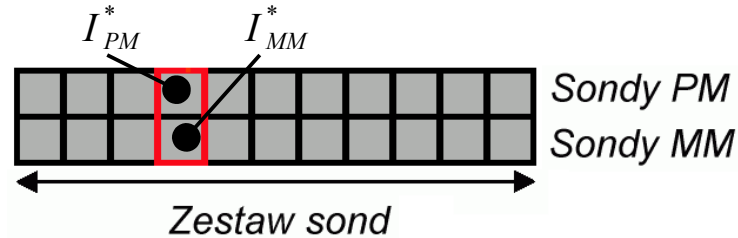


Pary sond PM i MM zawsze sąsiadują ze sobą, jednak poszczególne pary z tego samego zestawu mogą znajdować w różnych miejscach mikromacierzy.



Wyznaczanie miar ekspresji dla mikromacierzy GeneChip: alg. MAS 5.0

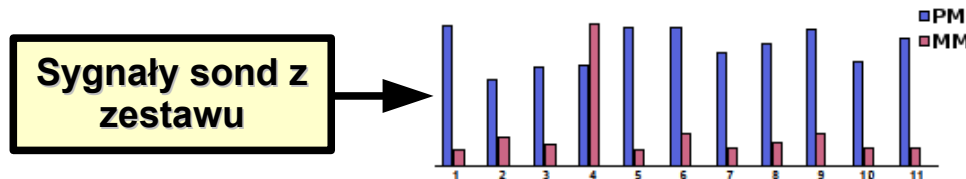
Wartość sygnału I_{NSB} , będącego efektem niespecyficznego hybrydowania określany jest na podstawie sygnału sondy MM (po korekcji tła) towarzyszącej danej sondzie PM.



$$I_{NSB} = I_{MM}^*$$

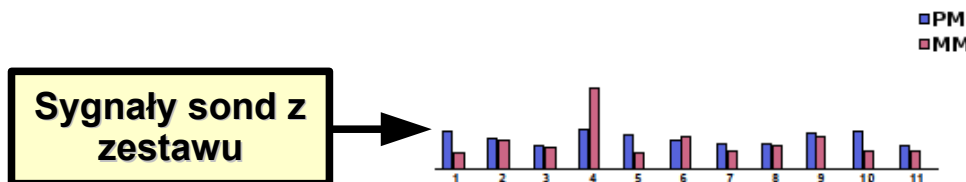
W szczególnym przypadku gdy $I_{MM}^* \geq I_{PM}^*$ przy szacowaniu poziomu niespecyficznego hybrydowania brana jest pod uwagę wartość średnia $\bar{R}$ stosunków I_{PM}^* / I_{MM}^* dla wszystkich par sond w zestawie:

▪ gdy $\bar{R}$ większe od zadanej wartości progowej



$$I_{NSB} = \frac{I_{PM}^*}{\bar{R}}$$

▪ gdy $\bar{R}$ nie większe od zadanej wartości progowej



$$I_{NSB} = \text{wartość nieznacznie mniejsza od } I_{PM}^*$$

Po wyznaczeniu I_{NSB} , wartość sondy PM wyznaczany jest jako: $I_T = I_{PM}^* - I_{NSB}$

Wyznaczanie miar ekspresji dla mikromacierzy GeneChip: alg. MAS 5.0

Wyznaczenie wartości E_i i -tego zestawu sond (miary ekspresji i -tego genu) odbywa się poprzez uśrednienie wartości I_T sond PM wchodzących w skład zestawu.

Po określeniu miar ekspresji wszystkich genów są one przeskalowywane liniowo tak aby ich **obcięta średnia** (**trimmed mean**, średnia arytmetyczna nie uwzględniająca 2% najmniejszych i największych wartości) była równa zadanej wartości S_c (zwykle wynosi ona 500). Krok ten ma sens bardzo **prostej procedury normalizacyjnej**, pozwalającej porównywać wartości pomiędzy osobnymi pomiarami.

Ostateczna wartość miary ekspresji wyrażana jest w skali logarytmicznej:

$$x_i = \log_2(S_f E_i) \quad \text{gdzie:} \quad S_f = \frac{S_c}{\text{TrimMean}(E_i, 0.02, 0.98)}$$

Uczenie maszynowe w bioinformatyce

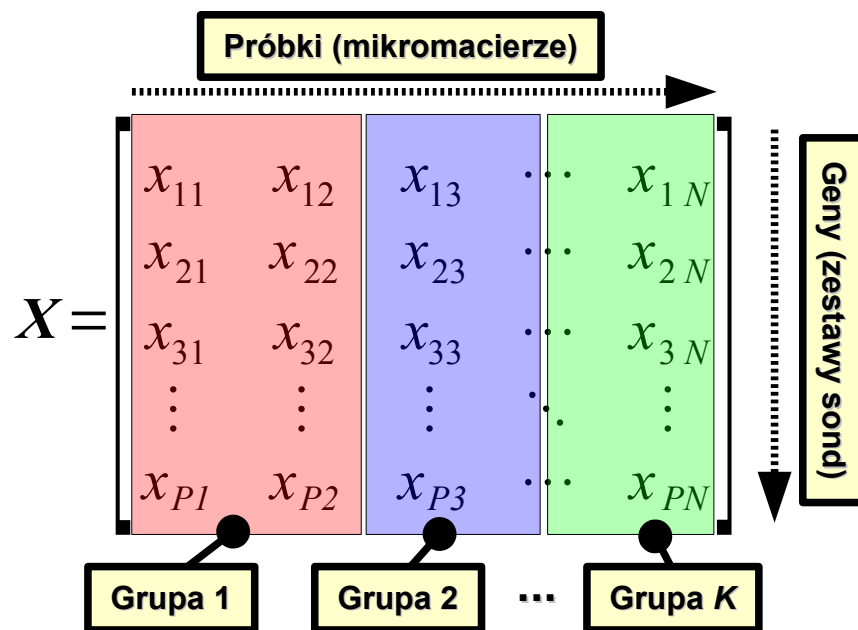
Wykład 5: mikromacierze (dane w postaci liczbowej)

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Dane mikromacierzowe w postaci liczbowej: macierz danych

Po obliczeniu miar ekspresji genów dla wszystkich mikromacierzy z eksperymentu, uzyskany zbiór danych można przedstawić w postaci macierzy X o wymiarze $P \times N$:



Kolumna x_j macierzy X reprezentuje j -tą próbkę (populację komórek, których poziom ekspresji genów został zmierzony za pomocą jednej mikromacierzy). Każda z próbek należy jest do jednej z **K grup badanych** (np. reprezentujących różne schorzenia).

Wiersz x_i macierzy X reprezentuje i -ty gen (zestaw sond). Zwykle liczba genów (do kilkudziesięciu tysięcy) znacznie przekracza liczbę próbek (do kilkuset), czyli **$P \gg N$** .

Element x_{ij} macierzy X jest względnym poziomem ekspresji i -tego genu, w j -tej populacji komórek, najczęściej wyrażonym w skali logarytmicznej.

Dane mikromacierzowe w postaci liczbowej: macierz danych

Przykład danych z mikromacierzy oligonukleotydowych GeneChip. Najczęściej mają one postać pliku tekstowego z polami rozdzielanymi znakami tabulacji:

	b016.cel	b132.cel	b134.cel	b136.cel	b138.cel	bp015.cel	bp131.cel	bp133.cel
1007_s_at	10.4454690589064	10.4071089205469	10.2622507944183	10.6156148106393	10.2338498301722	7.401127662552	10.3262090047782	10.2604337485715
1053_at	5.43775448150326	5.91264202068973	6.84581524959128	6.57701779840229	6.47865730868465	6.00121604451832	8.16300036318048	7.08864128764872
117_at	5.18826816640405	5.11585055845764	5.22177820180801	5.09862990637727	5.03661581557071	5.21854693261856	5.3116300036318048	5.0263
121_at	7.60218699875416	7.76094770088337	7.61950564233599	7.87464400134457	7.53122327806837	6.67169481504801	7.9116300036318048	7.043
1255_g_at	3.47491930187561	3.69226652086254	3.47540158090511	3.48599630641576	3.55175192484593	4.4218460663901	3.6116300036318048	3.1002
1294_at	7.81085837593124	7.94199755688827	7.56357321823726	7.64074266849863	7.6152361084739	6.72813884497327	7.20810664603654	7.09211089378672
1316_at	4.77726687187599	4.59110498069184	4.60567654247351	4.48655388533683	4.5083907419365	5.40779234279086	4.36658441039149	4.50646078840988
1320_at	4.60830807602174	4.8295167696508	5.04850141087965	5.01204559045464	4.8565501943956	4.92182439039217	4.89834501027322	4.86985576393017
1405_i_at	5.65419774738202	5.5962593976514	5.35405812007412	4.9113374260012	5.2208345127651	3.95754174878536	5.25443309056589	4.58970569289832
1431_at	4.21730842852391	4.3452683						
1438_at	6.1494121383617	7.1652063						
1487_at	8.22595623679706	8.0874501362488	7.68716237883348	8.68743383138783	7.68871432428338	3.63433332324233	7.36268436338848	7.33833333333333
1494_f_at	5.59183712884893	5.84062605828341	5.525953009365	5.72624120891342	5.65219240587031	6.04653636012996	5.82839771687024	5.6941316057524
1598_g_at								
160020_at								
1729_at	5.97331100196898	6.63283671842643	5.61078064069341	6.72596433523986	6.83605124400333	6.34119291833238	7.17931159369374	7.35303880024489
177_at	5.34328730963196	5.85060281395861	5.93255878413522	6.44676313981112	5.90343872840074	6.45231773522602	7.24733566229057	6.83106861854103
1773_at	8.53833216249888	8.7930309903341	8.4965200598029	8.66788275156115	8.62967395261811	8.98432202393434	8.98763316866932	8.79114463958511
179_at	5.52170076952969	6.3201809749932	7.17335438470933	6.63439525970733	6.74837052879373	6.19984406395378	6.64445387638719	6.46297786406849
1861_at	8.81629205065555	8.87706339169029	8.87100142066223	8.64618270769412	8.79243148711355	6.80192451220323	8.89886462204755	8.80206248933533
200000_s_at								
...	...	...	...	...	...	...	...	...

Nazwy próbek

Wartości zestawów sond w kolejnych próbkach (zwykle w skali log₂)

Identyfikatory zestawów sond (można je przetłumaczyć na nazwy genów)

Danym zwykle towarzyszy plik z opisem przynależności próbek do grup badanych:

	Rozpoznanie	Plec	Wiek	...
b016.cel	Rak	K	66	...
b132.cel	Rak	M	62	...
b134.cel	Rak	K	54	...
b136.cel	Rak	M	72	...
b138.cel	Rak	M	57	...
bp15.cel	Kontrola	K	60	...
bp131.cel	Kontrola	M	49	...
bp133.cel	Kontrola	K	71	...

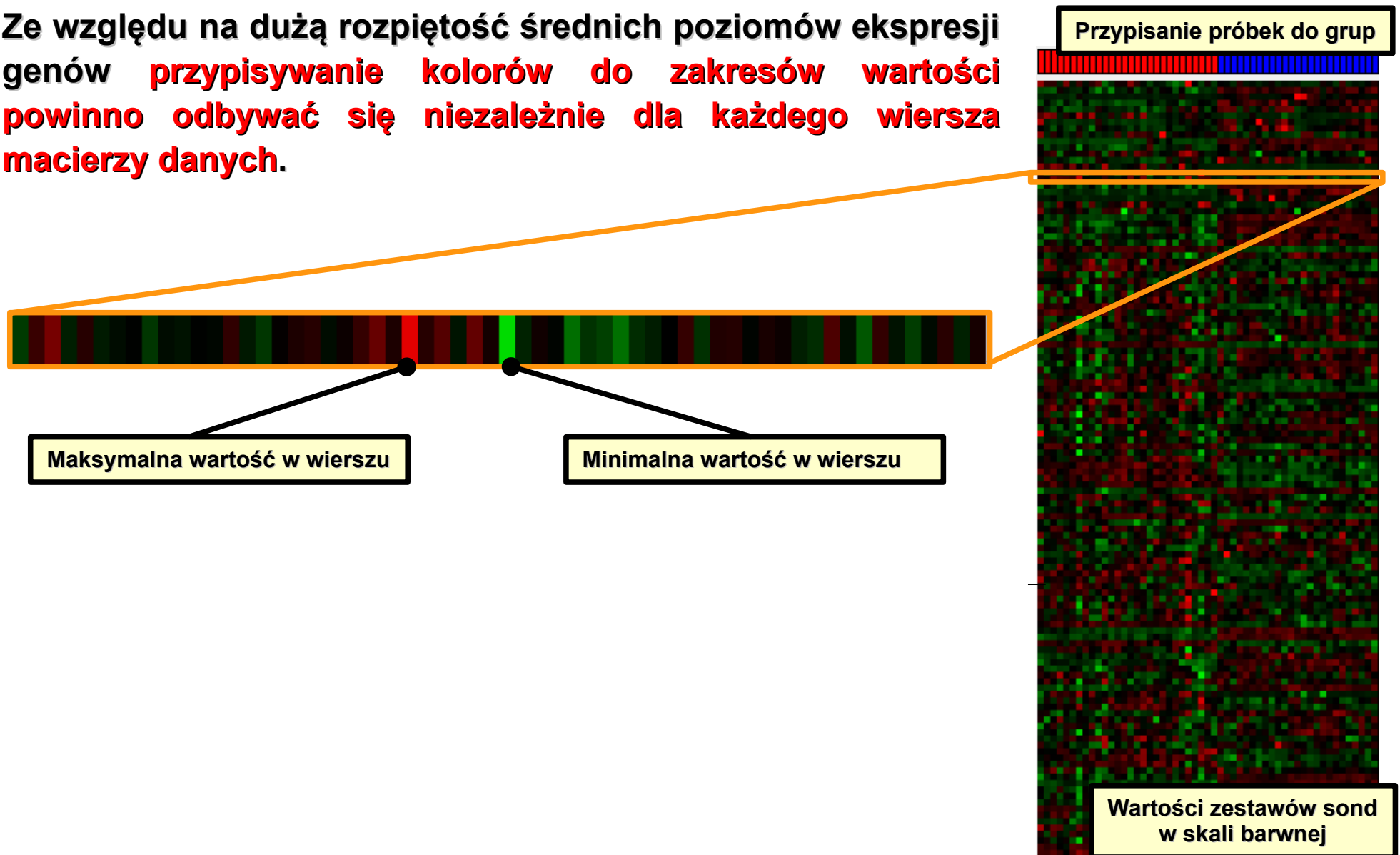
Nazwy próbek

Przypisanie próbek do grup

Dane mikromacierzowe w postaci liczbowej: wizualizacja typu *heat-map*

Duży wymiar macierzy X powoduje, że przedstawienie jej w postaci liczbowej jest mało czytelne. Dlatego też często jest ona prezentowana jako tzw. **heat-map**, czyli obraz, którego kolory odpowiadają zakresom wartości.

Ze względu na dużą rozpiętość średnich poziomów ekspresji genów **przypisywanie kolorów do zakresów wartości powinno odbywać się niezależnie dla każdego wiersza macierzy danych.**



Dane mikromacierzowe w postaci liczbowej: wizualizacja typu *heat-map*

Duży wymiar macierzy X powoduje, że przedstawienie jej w postaci liczbowej jest mało czytelne. Dlatego też często jest ona prezentowana jako tzw. **heat-map**, czyli obraz, którego kolory odpowiadają zakresom wartości.

Dobór palety kolorów jest dowolny, jednak często stosuje się **zielono**-czarno-**czewoną** skalę barwną.

Przykładowo, środek palety (czarny) może oznaczać wartość średnią wiersza, a skrajne kolory (**zielony** i **czewony**) to granice symetrycznego przedziału o szerokości równej wielokrotności odchylenia standardowego wartości wiersza.

