

Uczenie maszynowe w bioinformatyce

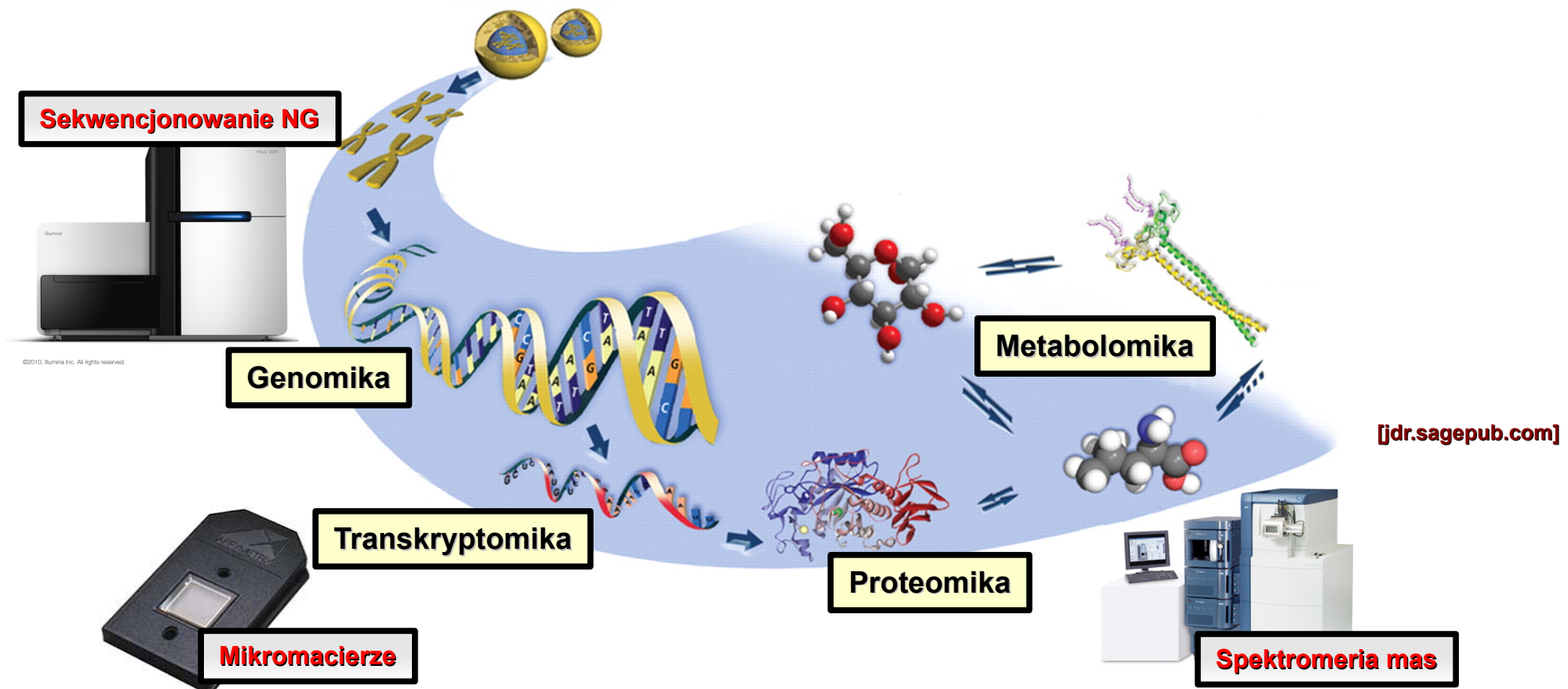
Wykład 3: wielkoskalowe metody pomiarowe jako źródła danych dla analiz bioinformatycznych

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Wielkoskalowe metody pomiarowe stosowane w biologii molekularnej

Każda z nauk „omicznych” posługuje się wieloma technikami analitycznymi, jednak podstawowe znaczenie dla ich rozwoju miało opracowanie i rozpowszechnienie się **wielkoskalowych (wysokowydajnych, *high-throughput*) metod pomiarowych**, w tym przede wszystkim: **technik szybkiego sekwencjonowania DNA, mikromacierzy oraz spektrometrii mas.**



Uczenie maszynowe w bioinformatyce

Wykład 3: wielkoskalowe metody pomiarowe (genomika: sekwnecjonowanie NG)

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Sekwencjonowanie DNA: metoda Sangera

Podstawowym narzędziem genomiki są metody **sekwencjonowania cząstek DNA**, czyli ustalania kolejności wchodzących w ich skład nukleotydów.

Klasyczną techniką sekwencjonowania jest **metoda terminacji łańcucha (metoda Sangera)**, dostarczająca w pojedynczym pomiarze informacji o 500-700 nukleotydach (z dokładnością 99,9%). Znalazła ona zastosowanie nie tylko w biologii molekularnej, ale też w medycynie i kryminalistyce. Umożliwiła również przeprowadzenie **Human Genome Project**, w którym ustalona została sekwencja ludzkiego genomu (w 2000 roku opublikowano wstępną wersję, sukcesywnie uzupełnianą w kolejnych latach).



Frederick Sanger

Sekwencjonowanie DNA: techniki NGS

Sekwencjonowanie Sangera jest metodą drogą i powolną, dlatego w ostatnich latach obserwuje się gwałtowny rozwój technik o większej wydajności, określanych ogólną nazwą **sekwencjonowania następnej generacji (NGS – Next Generation Sequencing)**.

Ich główną cechą jest to, że **pomiary odbywają się równolegle** dla milionów, a nawet miliardów odcinków DNA **o długości kilkudziesięciu nukleotydów**. Oznacza to, że dla dłuższych fragmentów DNA możliwe jest tylko określanie sekwencji początku i końca (są to tzw. **pary odczytanych końców**).

Solexa (Illumina)



Ion Torrent PGM (Thermo)



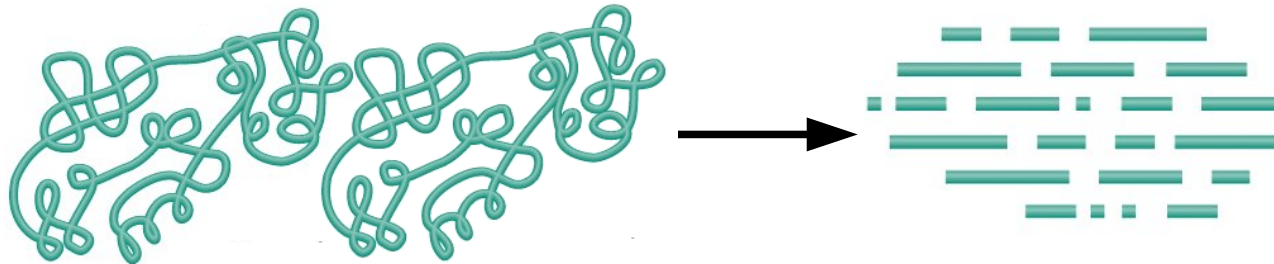
454 sequencing (Roche)



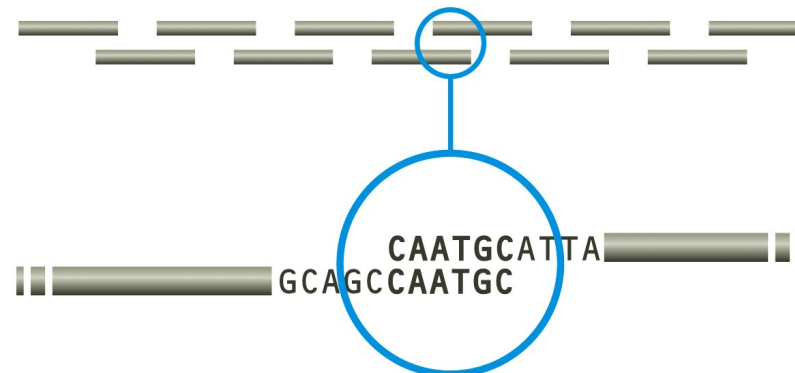
Sekwencjonowanie DNA: badanie genomu (strategia shotgun)

Sekwencjonowanie całego genomu wymaga użycia **strategii shotgun**, w której pełna sekwencja składana jest ze zbioru krótkich odcinków. W jej ogólnym schemacie można wyróżnić trzy podstawowe etapy:

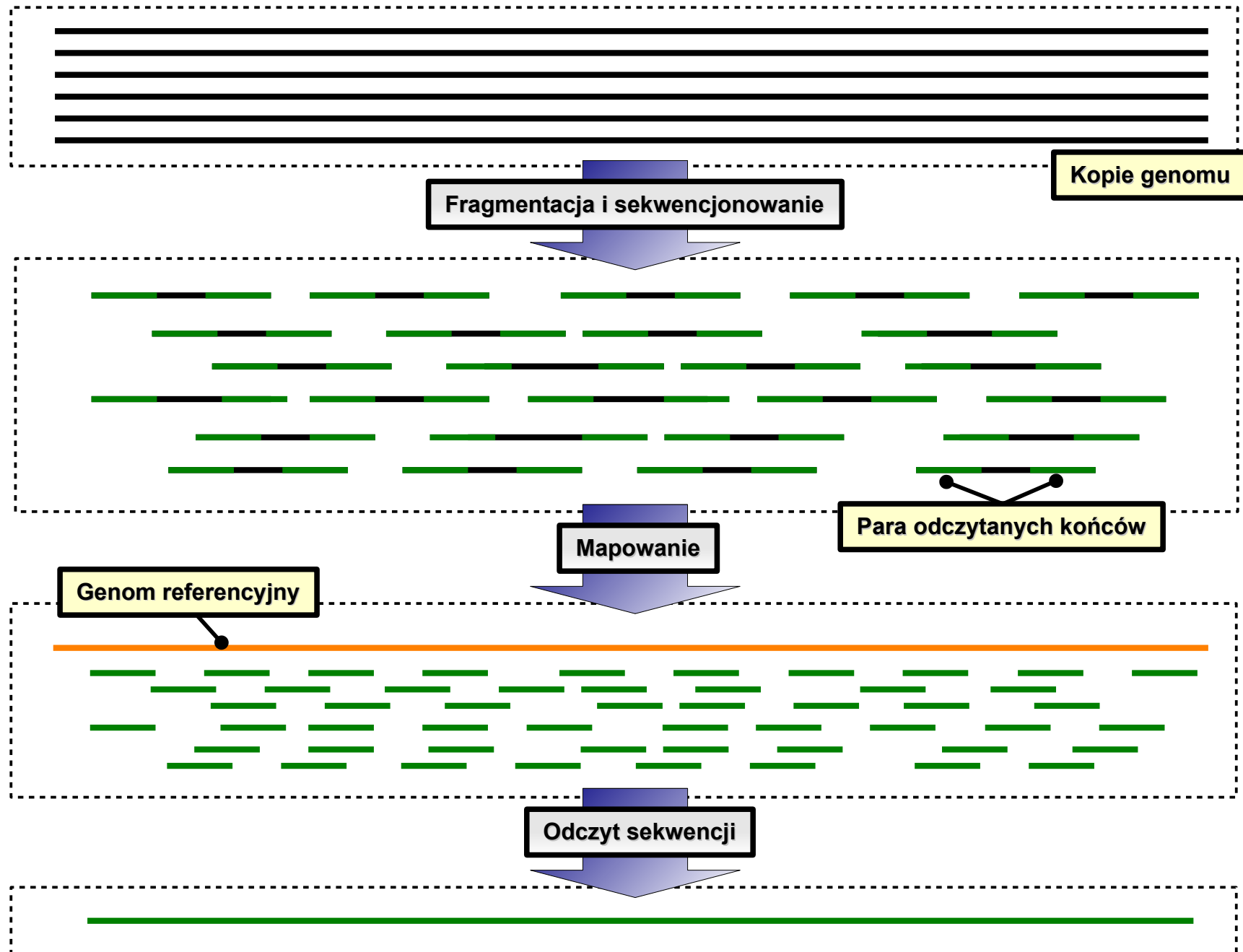
- **fragmentację wielu kopii całego genomu lub osobnych chromosomów na odcinki o losowej długości** poprzez sonikację (podział za pomocą ultradźwięków).



- **sekwencjonowanie** poszczególnych fragmentów;
- **składanie sekwencji genomu (*genome sequence assembly*)** ze zbioru zachodzących na siebie sekwencji fragmentów. Korzysta się przy tym z założenia, że sekwencje o jednakowej kolejności nukleotydów pochodzą z tego samego miejsca w genomie.

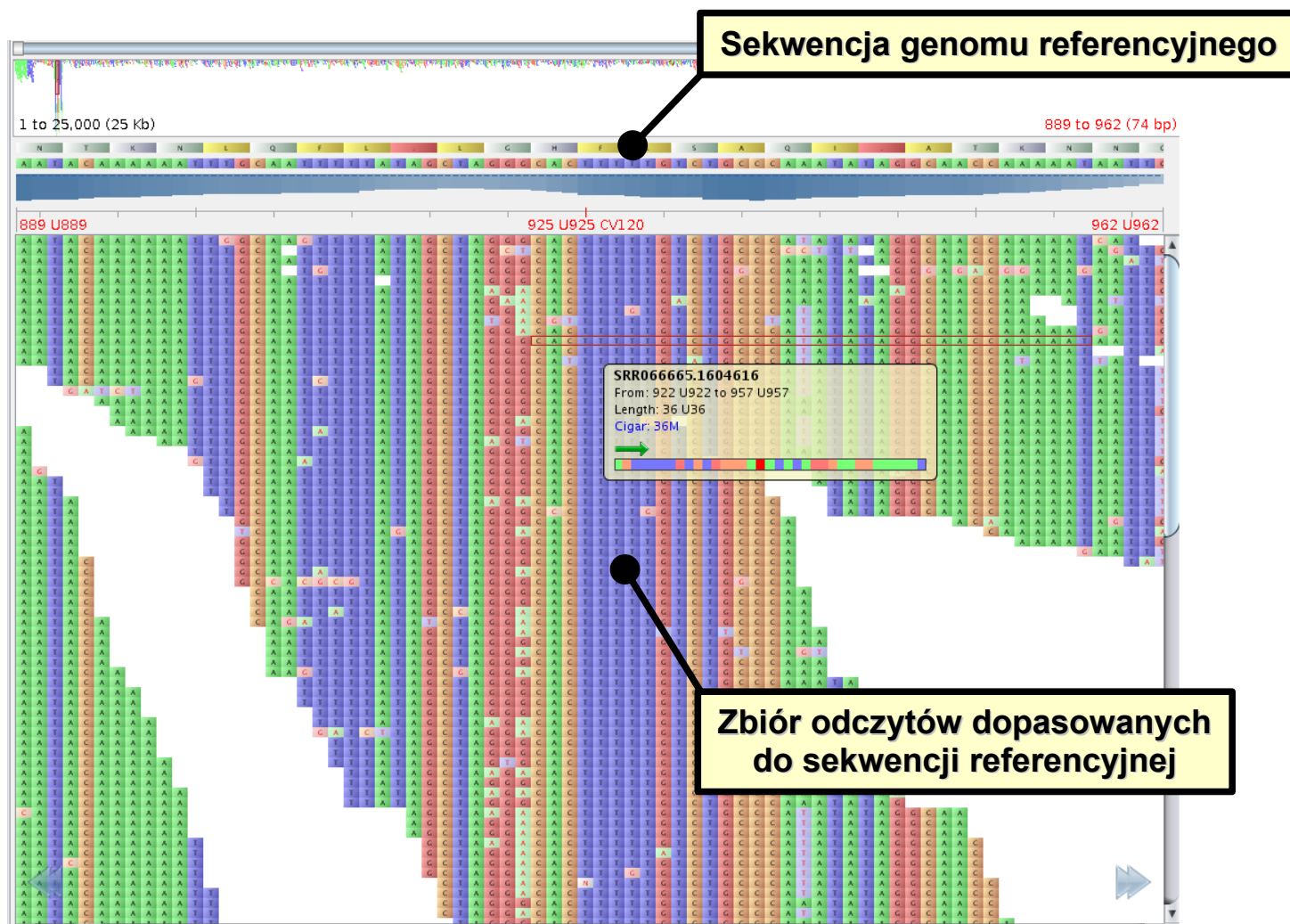


Sekwencjonowanie DNA: badanie genomu (strategia shotgun)



Sekwencjonowanie DNA: składanie sekwencji genomu

Składanie sekwencji zwykle odbywa się poprzez **dopasowywanie (mapowanie) do genomu referencyjnego**, które polega na określeniu pozycji odczytów w odniesieniu do znanej sekwencji referencyjnej.



Uczenie maszynowe w bioinformatyce

Wykład 3: wielkoskalowe metody pomiarowe (transkryptomika: mikromacierze DNA)

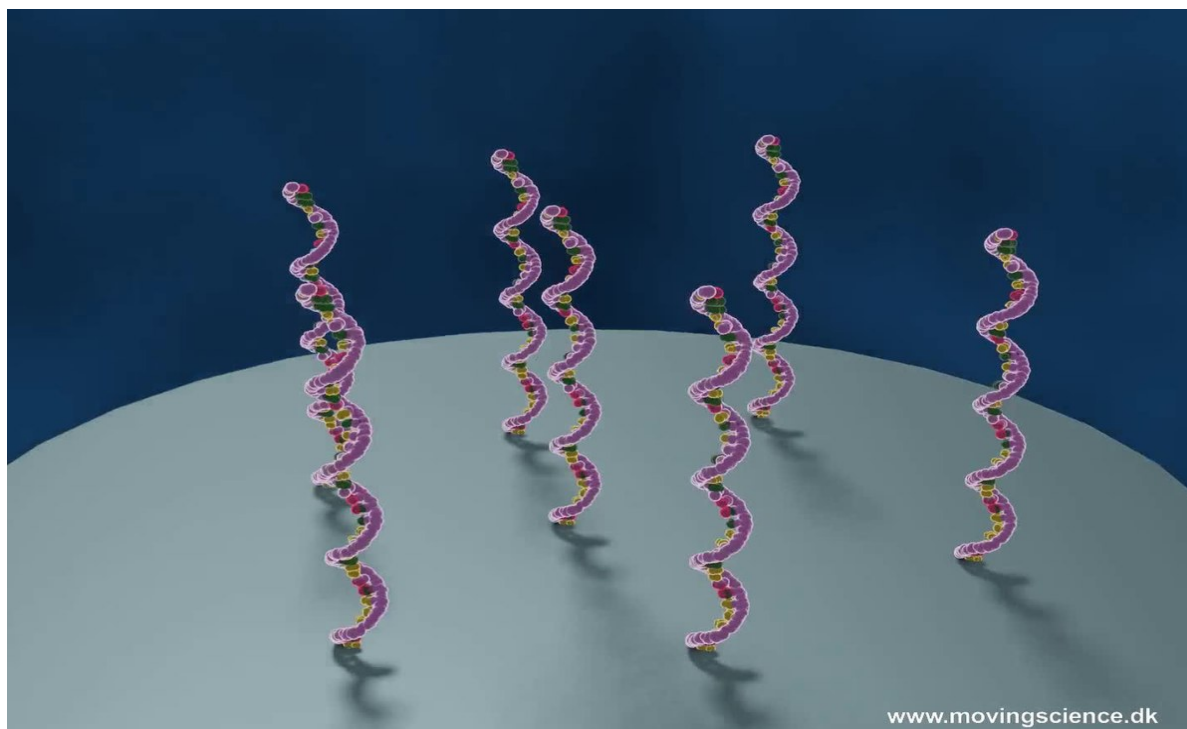
Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Mikromacierze DNA

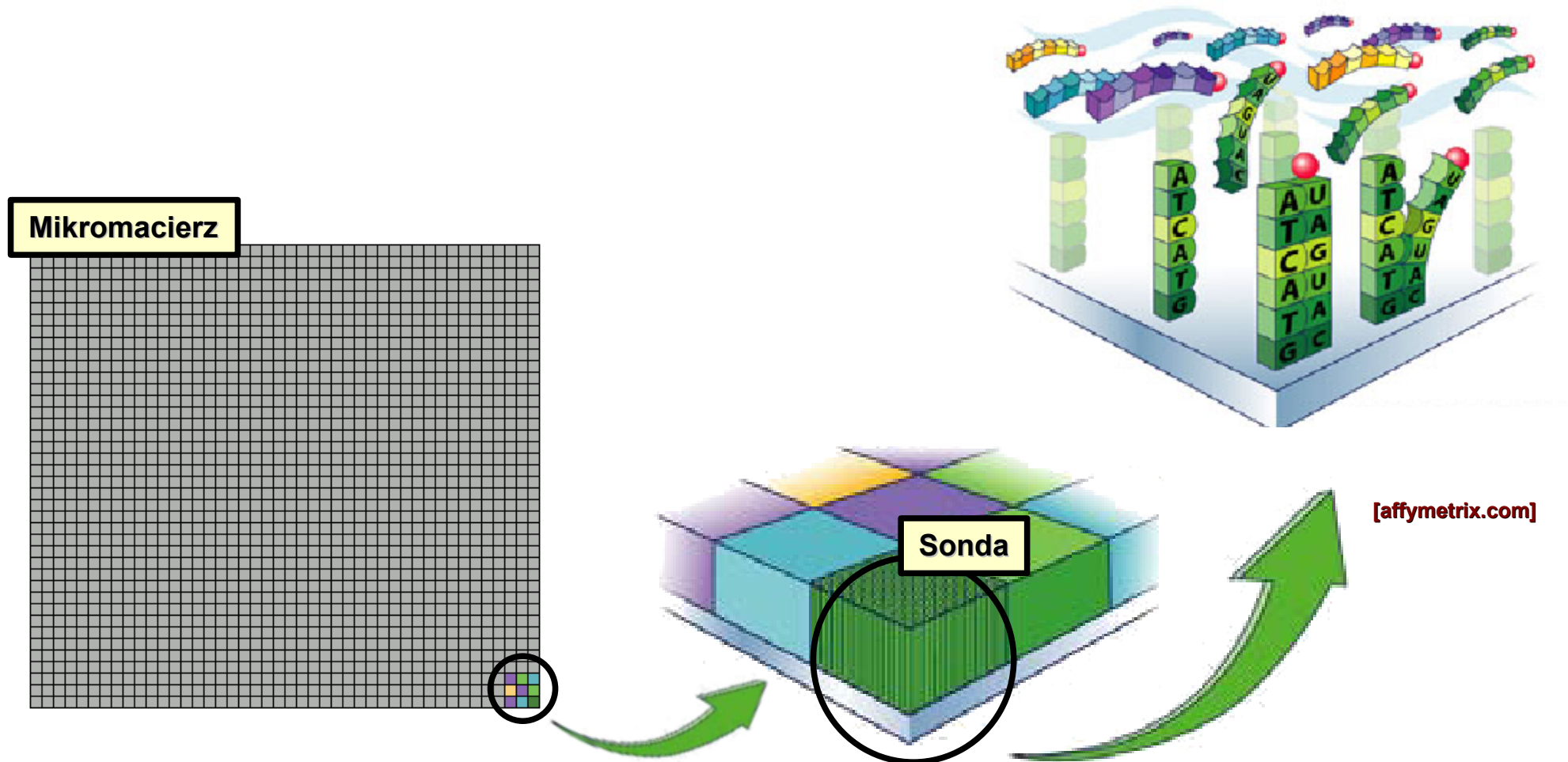
Mikromacierz DNA (*DNA microarray*) jest to płytka, na powierzchni której w sposób uporządkowany rozmieszczono dużą liczbę (od kilkuset do milionów) sond DNA.

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

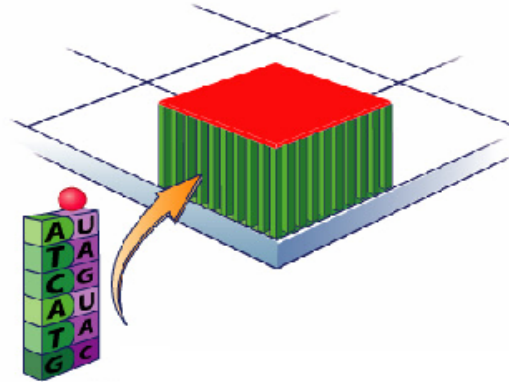
Każda z sond zawiera DNA o innej sekwencji. Zbiór wszystkich sond tworzy macierz, której każdy punkt jest „czuły” tylko na jeden rodzaj cząsteczek mRNA.



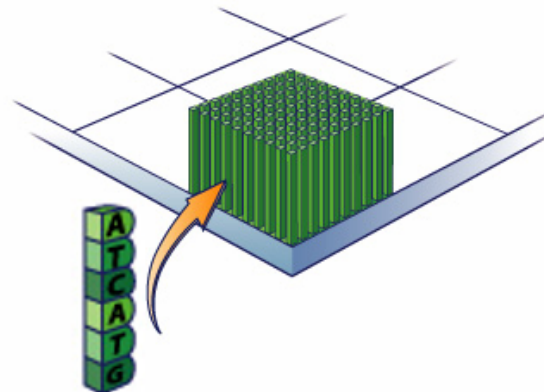
Mikromacierze DNA

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);



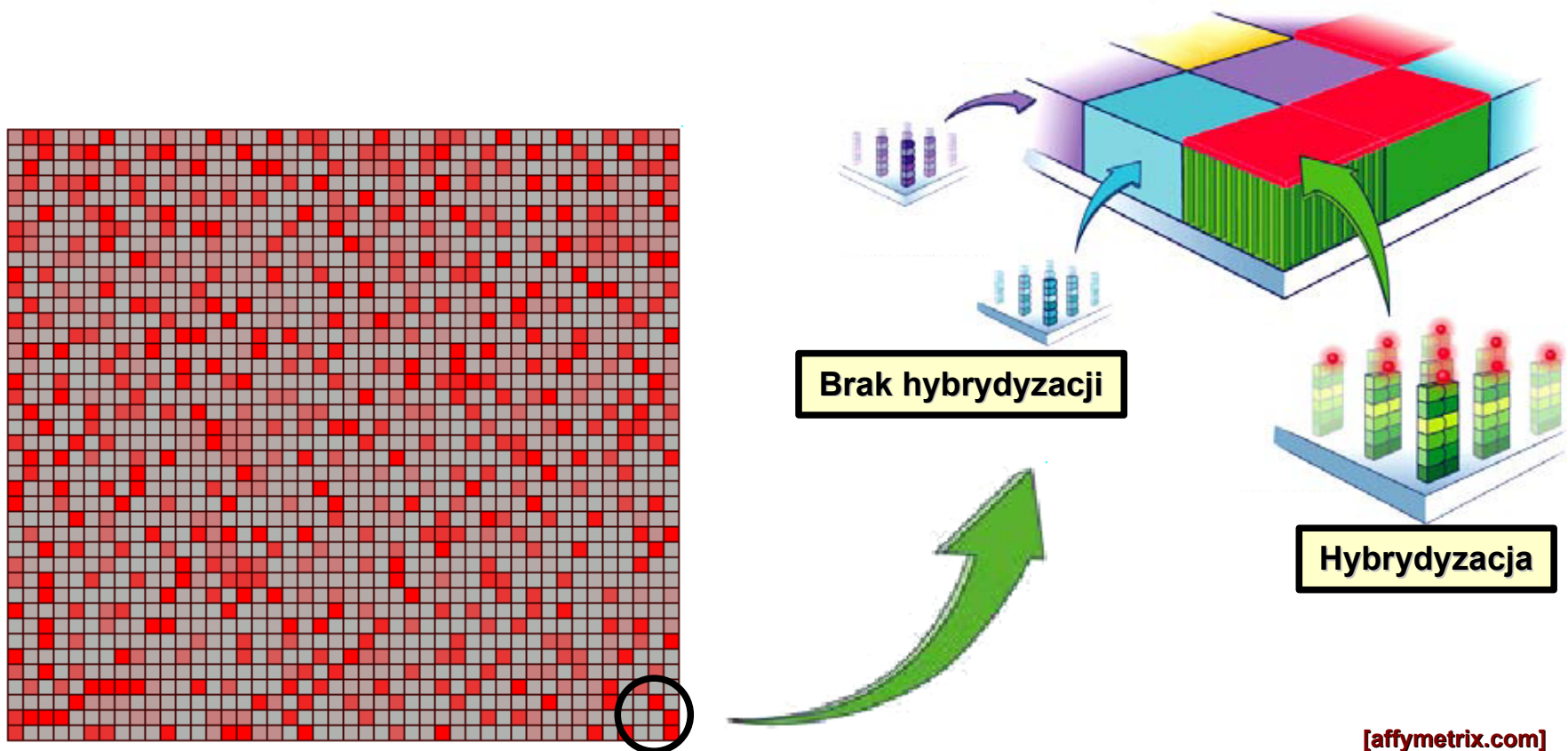
- jeżeli dany rodzaj mRNA nie występuje w próbkach, to reprezentująca go sonda pozostanie „pusta” (nie zhybrydują z nią żadne cząstki mRNA).



Mikromacierze DNA

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

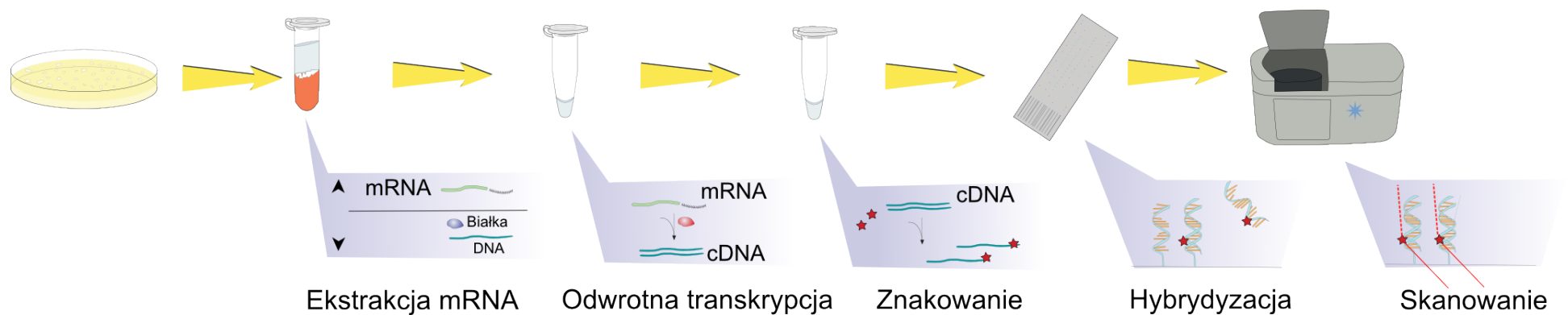
Odczyt mikromacierzy odbywa się poprzez wzbudzenie znacznika światłem lasera i rejestrację fluorescencji. **Bezpośrednim wynikiem pomiaru jest obraz**, w którym wartość każdego z pikseli jest proporcjonalna do natężenia fluorescencji z pewnego obszaru mikromacierzy.



Mikromacierze DNA

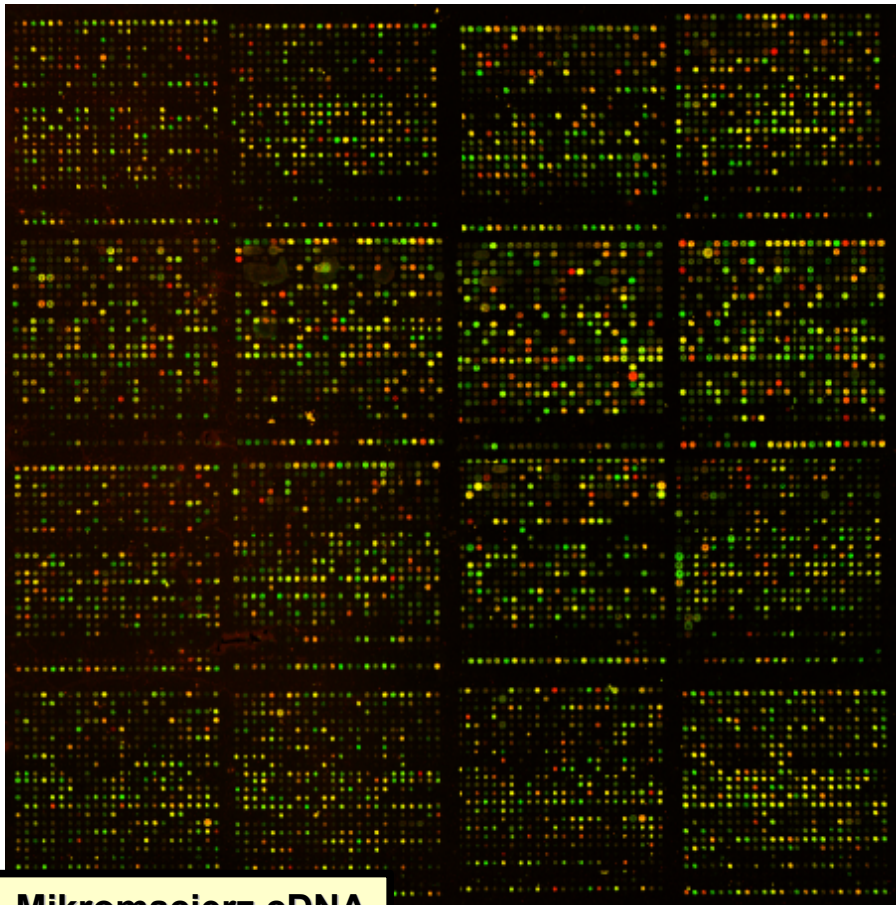
Ogólny schemat pomiarów przy użyciu mikromacierzy:

- **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**.



Mikromacierze DNA: postać danych

Bezpośrednim wynikiem pomiaru jest obraz o formacie zależnym od producenta i rodzaju mikromacierzy.



Mikromacierz cDNA



Mikromacierz oligonukleotydowa

Wielkoskalowe metody biologii molekularnej

Wykład 3: wielkoskalowe metody pomiarowe (proteomika: spektrometria mas)

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

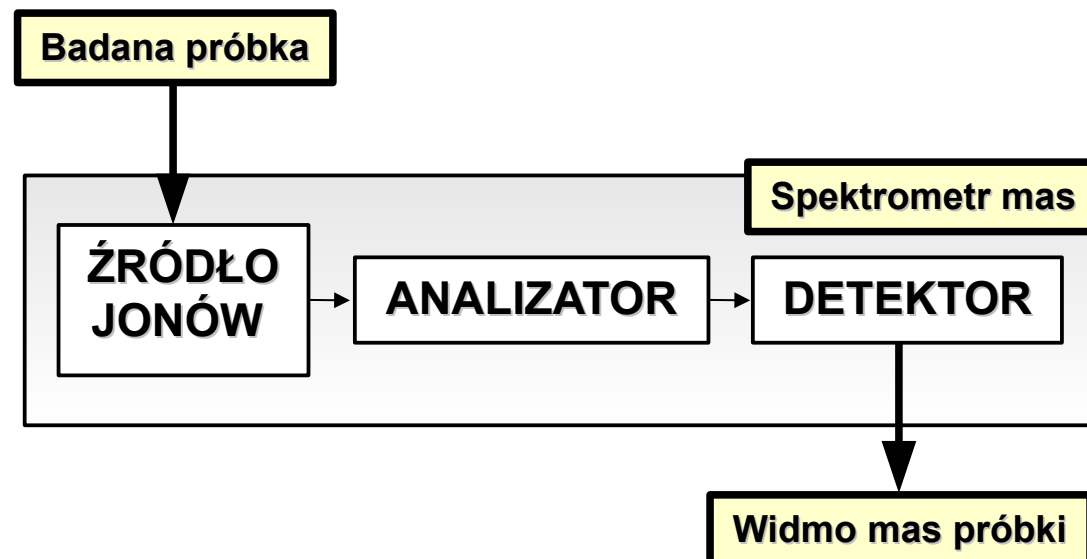
Spektrometria mas

Podstawowym narzędziem proteomiki jest obecnie **spektrometria mas (MS – Mass Spectrometry)**, technika analityczna umożliwiającą pomiar mas cząstek po jonizacji.

W spektrometrii mas badane są jony, więc pomiar nie dotyczy bezpośrednio masy: mierzona jest wartość m/z , czyli wartość **stosunku masy jonu do jego ładunku**.

W dużym uproszczeniu można powiedzieć, że spektrometr mas składa się z:

- **źródła jonów**, w którym obojętne cząstki zamieniane są na jony;
- **analizatora**, odpowiedzialnego za rozdzielanie jonów o różnych wartościach m/z ;
- **detektora**, zliczającego liczby jonów o poszczególnych wartościach m/z .



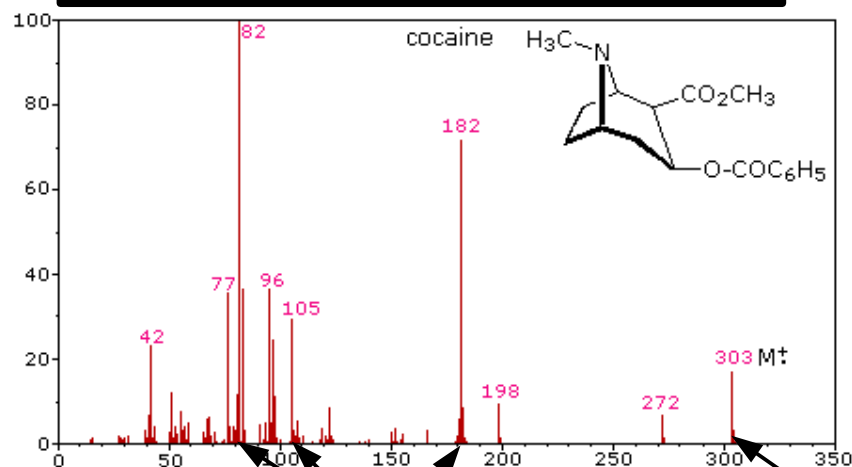
Spektrometria mas: widmo mas

Rozdzielone w analizatorze jony o różnych wartościach m/z są zliczane w detektorze, a ich liczby zapamiętywane są w postaci **widma mas**.

Widmo mas jest zwykle prezentowane w postaci wykresu liczby zarejestrowanych jonów o poszczególnych wartościach m/z . Często wykres jest znormalizowany wobec najwyższego piku.

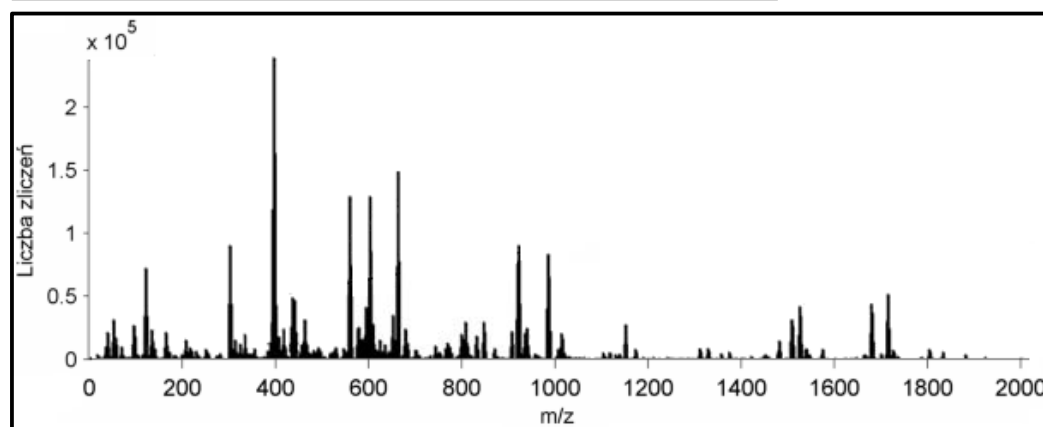
Występujące w widmie piki mogą odpowiadać całym jonom molekularnym związków występujących w badanej próbce lub jonom ich mniejszych fragmentów.

Znormalizowane widmo mas związku chemicznego ($C_{17}H_{21}NO_4$, $M = 303.147$)



Piki fragmentów jonu molekularnego

Widmo mas próbki biologicznej złożonej z dużej liczby białek



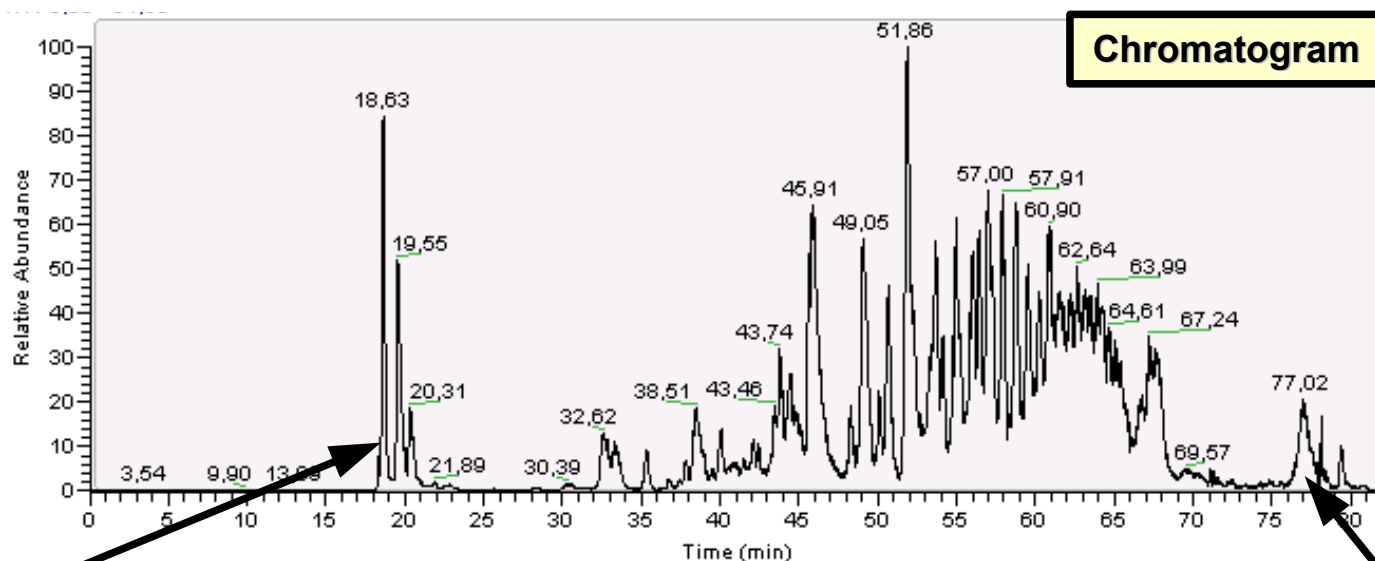
Pik jonu molekularnego

LC-MS: MS połączone z chromatografią cieczową

W przypadku badania złożonych mieszanin znaczną poprawę czułości analizy można uzyskać poprzez połączenie spektrometru z systemem chromatograficznym.

Chromatografia jest metodą rozdzielania składników mieszanin w wyniku ich różnego podziału pomiędzy fazę ruchomą i stacjonarną. Przemieszczanie się wzdłuż układu chromatograficznego jest możliwe tylko w fazie ruchomej, dlatego też dłużej będą w nim przebywać składniki mieszaniny silniej oddziałujące z fazą stacjonarną.

Składnikom opuszczającym układ chromatograficzny w różnym czasie odpowiadać będą osobne piki w rejestrowanym **chromatogramie**. Czas od wprowadzenia badanej mieszaniny do detekcji maksimum pików nazywany jest **czasem retencji**.



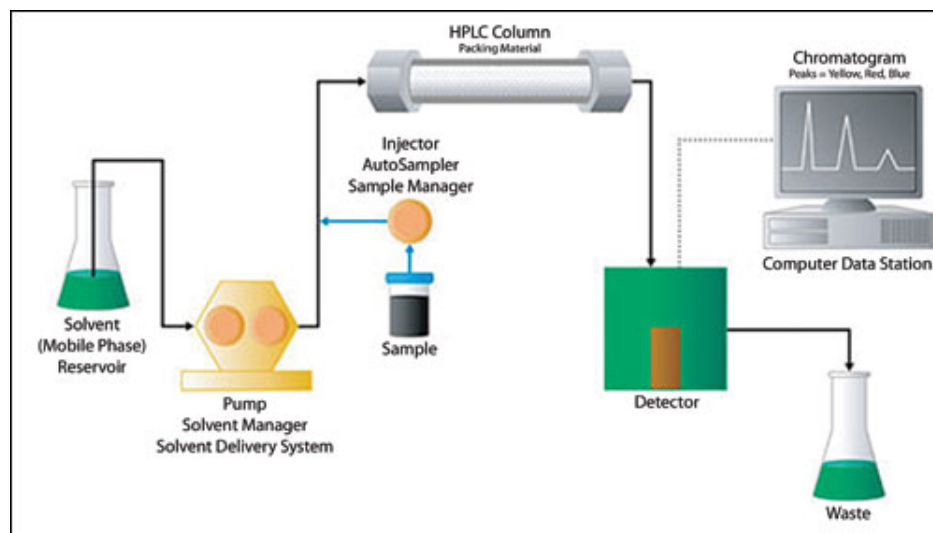
Piki odpowiadające cząstkom słabo oddziałującym z fazą stacjonarną

Piki odpowiadające cząstkom silnie oddziałującym z fazą stacjonarną

LC-MS: MS połączone z chromatografią cieczową

W zależności od stosowanej techniki chromatograficznej, fazą ruchomą może być gaz, ciecz lub fluid, a fazą stacjonarną ciało stałe lub ciecz. W proteomice używa się techniki **LC-MS (*Liquid Chromatography – Mass Spectrometry*)**, w której rozdzielanie składników następuje w sprzężonym ze spektrometrem systemie **wysokosprawnej chromatografii cieczowej (HPLC – *High Performance Liquid Chromatography*)**.

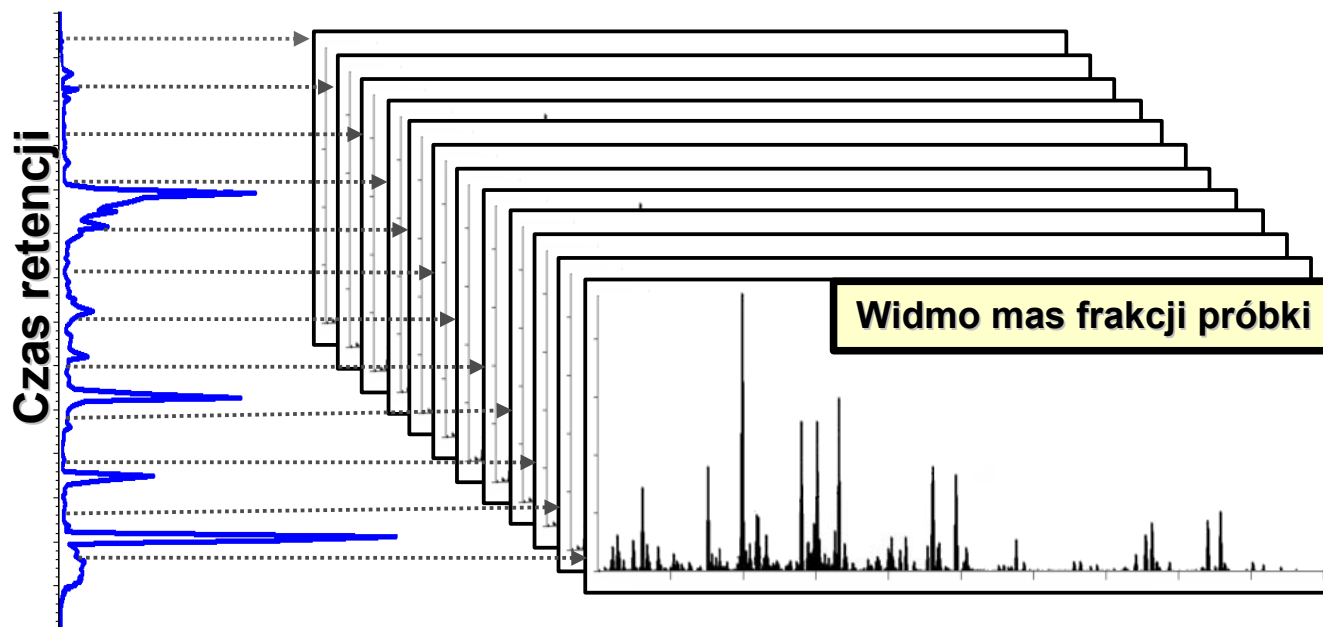
Podstawą rozdzielania w HPLC są oddziaływania pomiędzy analitem, niepolarną fazą stacjonarną i polarnym rozpuszczalnikiem. Częstki hydrofobowe (niepolarne) silniej oddziałują z fazą stacjonarną, zaś cząstki hydrofilowe (polarne) spędzać będą więcej czasu w fazie ruchomej. Oznacza to, że rozdzielanie mieszaniny następuje zgodnie z hydrofobowością jej składników.



LC-MS: postać danych

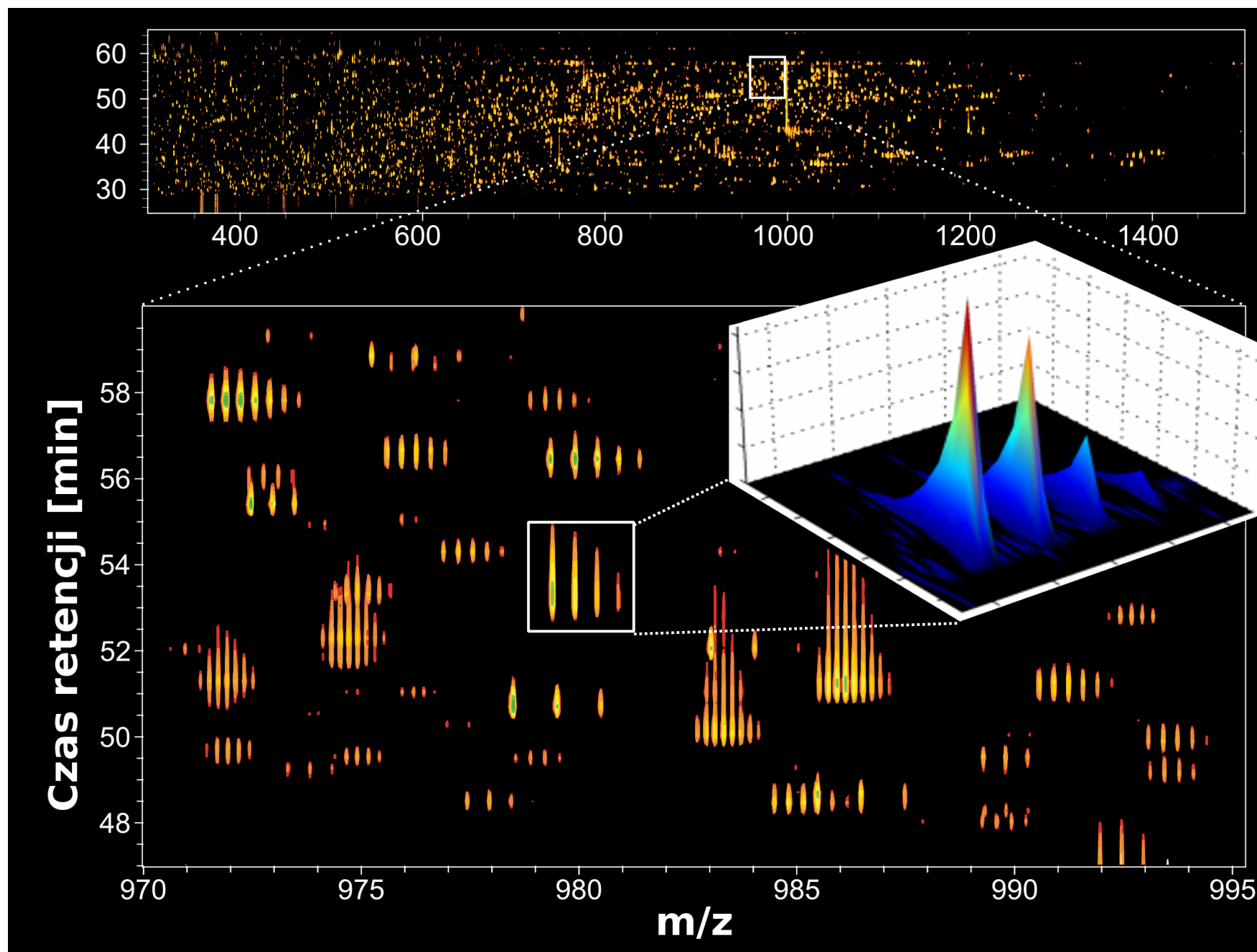
W czasie pomiaru techniką LC-MS próbka jest wstępnie rozdzielana w systemie HPLC, a kolejne jej frakcje po opuszczeniu układu chromatograficznego poddawane są jonizacji i wprowadzane do spektrometru. Pomiar widm mas następuje zwykle w regularnych odstępach czasu, zależnych od szybkości skanowania spektrometru.

Dane LC-MS mają postać zbioru widm mas kolejnych frakcji próbki. Każde widmo może zawierać nawet kilkaset tysięcy punktów, zaś liczba skanów może sięgać kilku tysięcy. W efekcie rozmiary plików z danymi LC-MS mierzone są w gigabajtach.



LC-MS: postać danych

Dane LC-MS można też przedstawić w postaci dwuwymiarowej mapy.



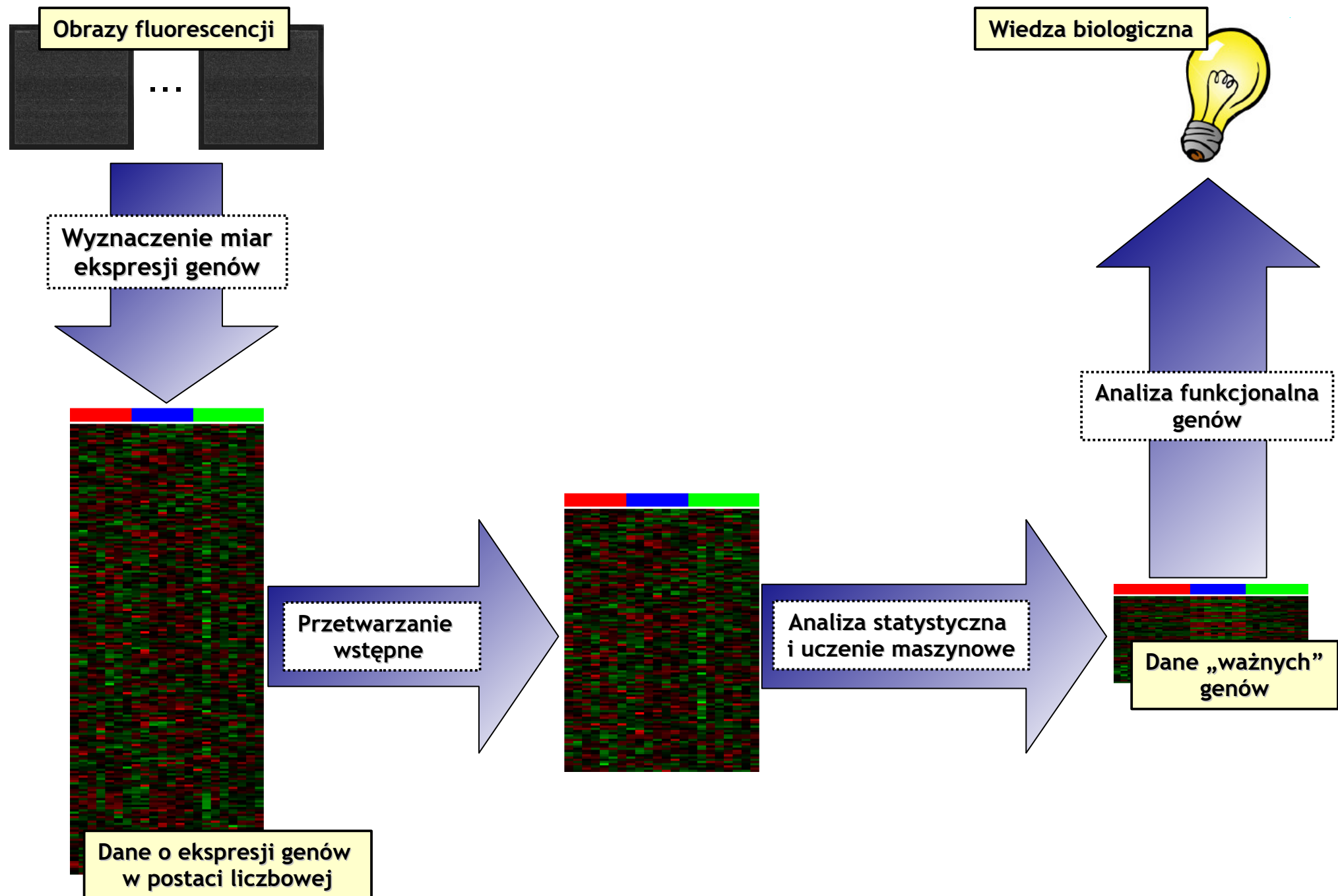
Uczenie maszynowe w bioinformatyce

Wykład 3: ogólny schemat analizy danych (na przykładzie mikromacierzy)

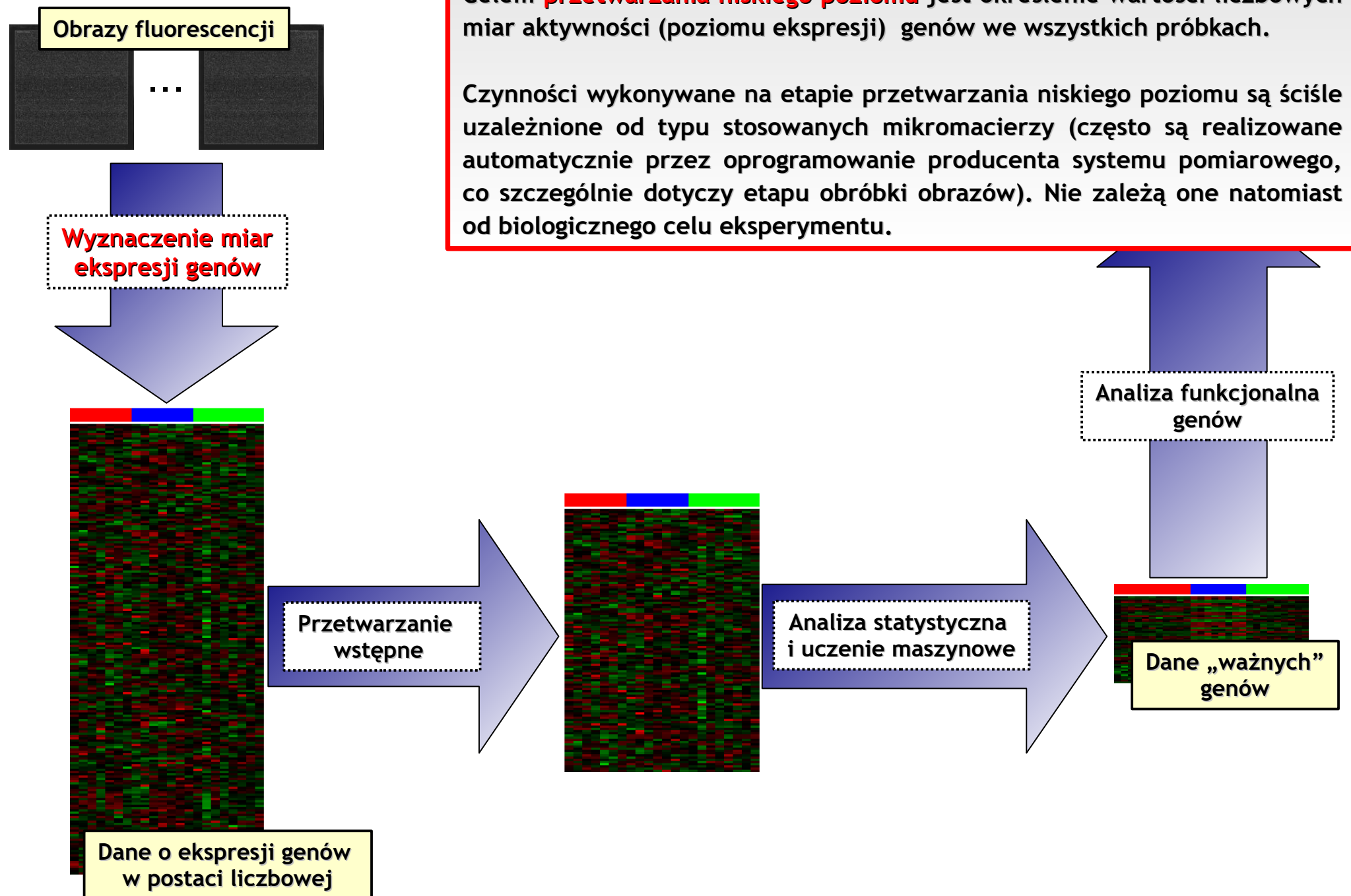
Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Przetwarzanie danych z mikromacierzy DNA



Przetwarzanie danych z mikromacierzy DNA



Przetwarzanie danych z mikromacierzy DNA

Obrazy fluorescencji

...

Wiedza biologiczna



Wyznaczenie miar ekspresji genów

	ZDROWI				RAK TYPU A			RAK TYPU B		
	Próbka 1	Próbka 2	...	Próbka 10	Próbka 11	...	Próbka 20	Próbka 21	...	Próbka 30
Gen 1	72	61	...	10	25	...	39	42	...	49
Gen 2	281	118	...	168	296	...	111	174	...	197
Gen 3	36	39	...	10	59	...	41	24	...	13
Gen 4	299	11	...	87	520	...	16	146	...	30

Analiza funkcjonalna genów

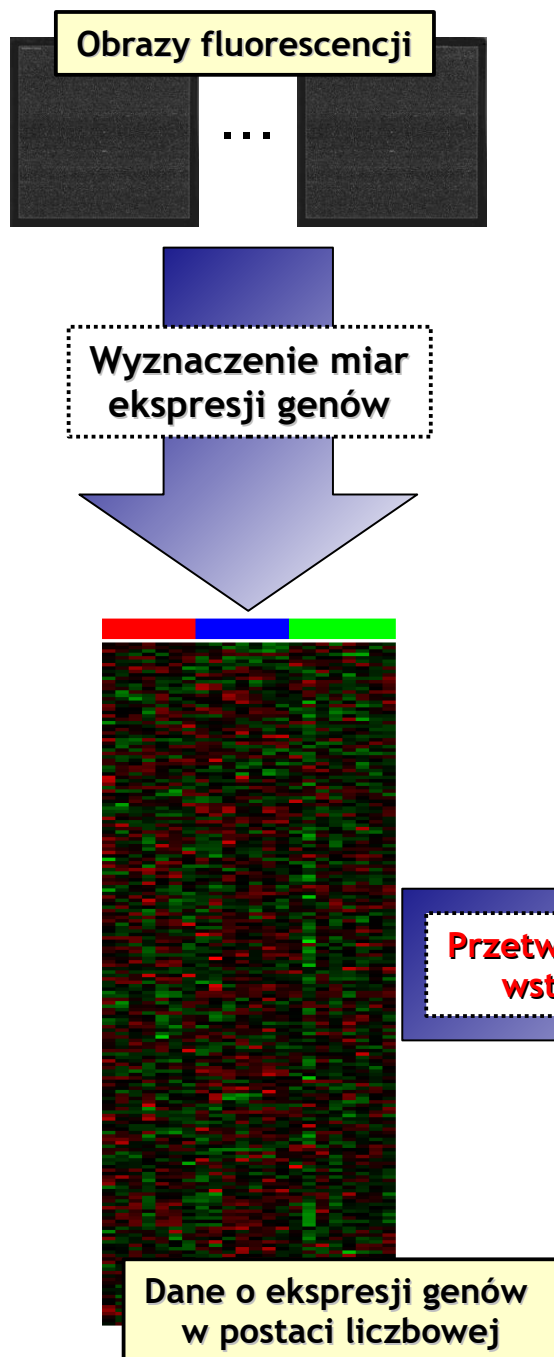
Przetwarzanie wstępne

Analiza statystyczna i uczenie maszynowe

Dane o ekspresji genów w postaci liczbowej

Dane „ważnych” genów

Przetwarzanie danych z mikromacierzy DNA



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 DNA

Wybór **metod statystycznych i uczenia się maszyn** używanych na etapie analizy danych jest ściśle uzależniony od celu badań. Stosowane tutaj techniki można w ogólności podzielić na **działające z nadzorem** lub **nienadzorowane**.

Metody działające bez nadzoru stosuje się do wykrywania zależności pomiędzy próbkami lub genami jedynie w oparciu o dane, bez korzystania z dodatkowych informacji (w tym też nie uwzględniając przynależności badanych obiektów do zadanych z góry klas). Do metod nienadzorowanych zaliczają się m. in.:

- klasteryzacja;
- techniki redukcji wymiarowości.

Metody nadzorowane działają w oparciu o znajomość przypisania obiektów do wcześniej ustalonych klas. Przykładami zagadnień, w których używa się metod nadzorowanych są:

- klasyfikacja;
- selekcja cech różnicujących grupy próbek.

Wiedza biologiczna



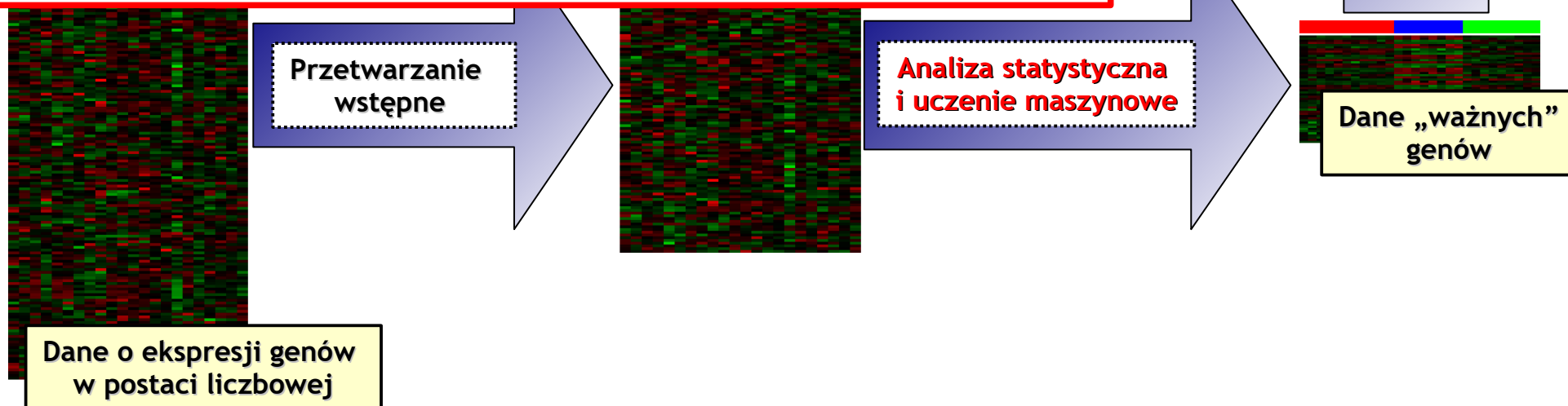
Analiza funkcjonalna genów

Analiza statystyczna i uczenie maszynowe

Dane „ważnych” genów

Przetwarzanie wstępne

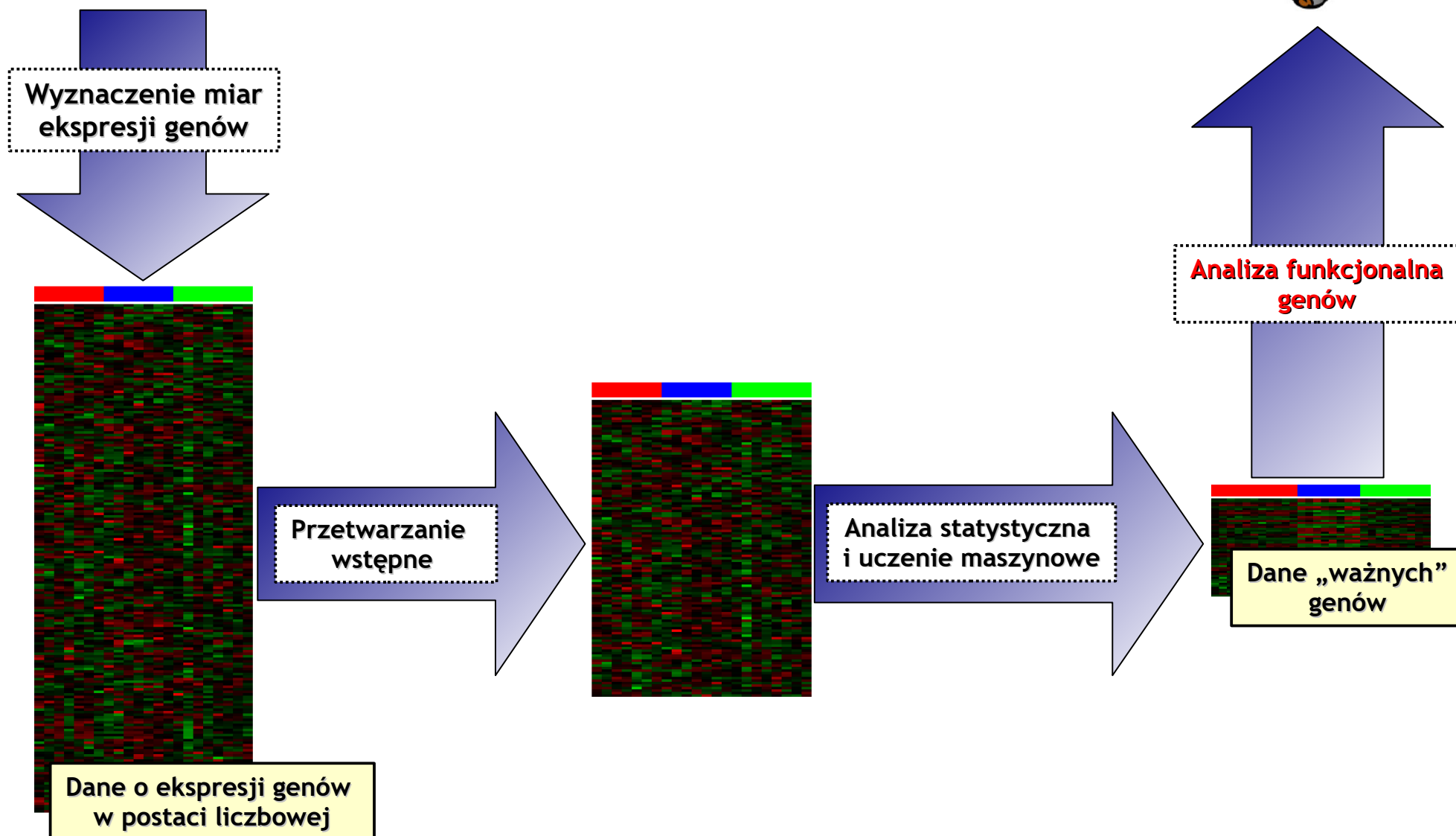
Dane o ekspresji genów w postaci liczbowej



Przetwarzanie danych z mikromacierzy DNA

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

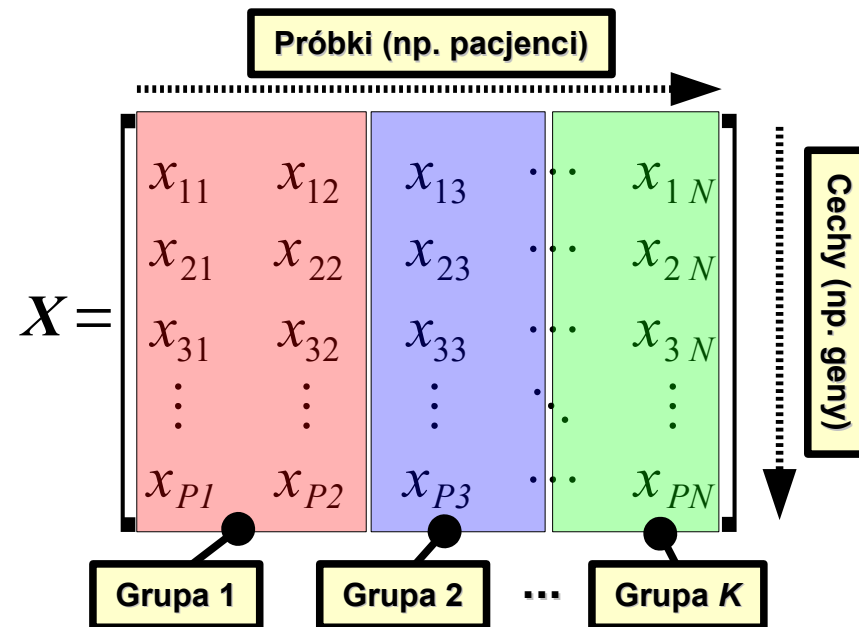
Wykład 3: dane w postaci liczbowej

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Dane z technik wielkoskalowych w postaci liczbowej: macierz danych

Po przetwarzaniu niskopoziomowym zbiorów danych uzyskany z wielu różnych technik eksperymentalnych (m. in. mikromacierzy DNA lub białkowych, RT-PCR, RNA-Seq, spektrometrii mas) można przedstawić w postaci macierzy X o wymiarze $P \times N$:



Kolumna x_j macierzy X reprezentuje j -tą próbkę (np. populację komórek, pacjenta). Każda z próbek należy do jednej z K **grup badanych** (oznaczających np. różne schorzenia lub układy eksperymentalne).

Wiersz x_i macierzy X reprezentuje i -tą cechę (np. gen, białko). Zwykle liczba cech znacznie przekracza liczbę próbek, czyli $P \gg N$.

Dane z technik wielkoskalowych 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	5.31111111111111	5.0263
121_at	7.60218699875416	7.76094770088337	7.61950564233599	7.87464400134457	7.53122327806837	6.67169481504801	7.91111111111111	7.043
1255_g_at	3.47491930187561	3.69226652086254	3.47540158090511	3.48599630641576	3.55175192484593	4.4218460663901	3.61111111111111	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	8.68743883138783	7.68871432428338	3.63433632324233	7.3626843636848	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								
179_at	8.53833216249888	8.7930309903341	8.4965200598029	8.66788275156115	8.62967395261811	8.98432202393434	8.98763316866932	8.79114463958511
1861_at	5.52170076952969	6.3201809749932	7.17335438470933	6.63439525970733	6.74837052879373	6.19984406395378	6.64445387638719	6.46297786406849
200000_s_at	8.81629205065555	8.87706339169029	8.87100142066223	8.64618270769412	8.79243148711355	6.80192451220323	8.89886462204755	8.80206248933533
...	...	...	...	...	...	...	...	...

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	...

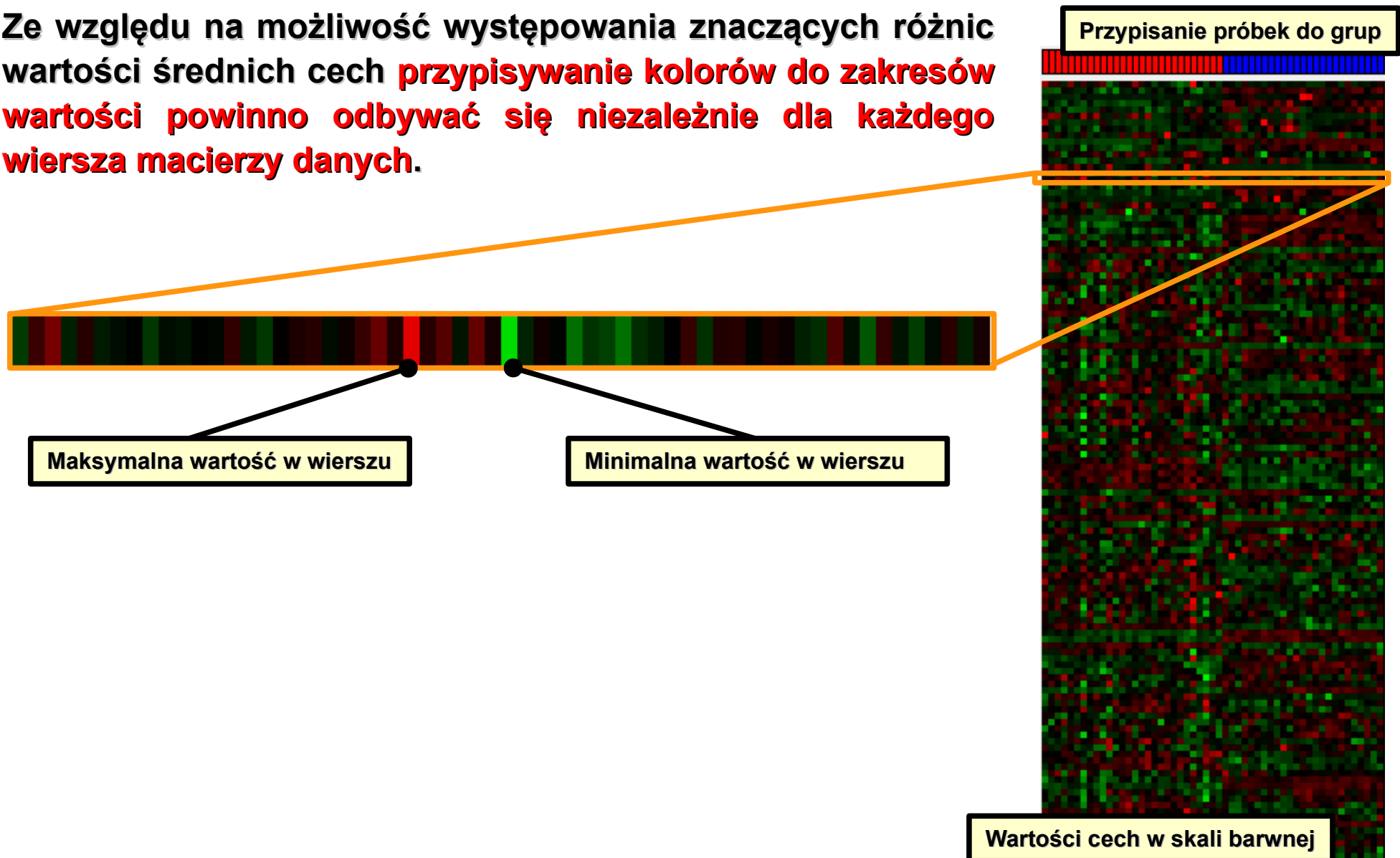
Nazwy próbek

Przypisanie próbek do grup

Dane z technik wielkoskalowych w postaci liczbowej: wizualizacja

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 możliwość występowania znaczących różnic wartości średnich cech **przypisywanie kolorów do zakresów wartości powinno odbywać się niezależnie dla każdego wiersza macierzy danych.**



Dane z technik wielkoskalowych w postaci liczbowej: wizualizacja

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.

