**Cel projektu, badania realizowane w projekcie oraz powody podjęcia danej tematyki badawczej**

Dzięki współpracy z Polską Akademią Nauk i udziale w projekcie dotyczącym ostrej białaczki szpikowej podczas analiz danych mikromacierzowych pojawiła się potrzeba selekcji takich genów, które bardzo precyzyjnie potrafią opisywać dane klasy chorobowe (AML, ALL). Budowanie klasyfikatorów nie może być oparte na zbyt dużej ilości genów gdyż to prowadzi do przeuczenia klasyfikatora. Z tego względu zaczęto się zastanawiać nad tym, jak ograniczać geny brane pod uwagę. Okazało się, że uwzględnianie korelacji pomiędzy wartościami ekpresji genów ma znaczący wpływ na jakość predykcji jednostek chorobowych do danych klas. Dlatego głównym celem projektu stała się konstrukcja testu statystycznego dzięki któremu badacz znajdowałby takie interesujące go geny, które dawałyby najlepszą efektywność predykcji jednostek do danej klasy (np. zdrowy - chory) z wykorzystaniem informacji dotyczącej wartości ekspresji dla genów oraz korelacji pomiędzy nimi. Geny, które są ze sobą w wysokiej korelacji, należą do tej samej, bądź pełniącej podobne funkcje w organiźmie ścieżki, w procesie klasyfikacji chcemy uniknąć selekcji tych właśnie wzajemnie wysoko skorelowanych genów. Geny, których wzajemna korelacja jest wysoka nie wnoszą zbyt dużo dodatkowej informacji w odniesieniu do późniejszej klasyfikacji i generują podobne błędy podczas zastosowania metody walidacji krzyżowej. Według wiedzy aplikanta nie istnieją w literaturze metody, które uwzględniałyby w ten właśnie sposób zarówno wielkość korelacji pomiędzy poziomami ekspresji, jak i wartości samych poziomów ekpresji.  Cel drugi polega na przedstawieniu zastosowania zaproponowanej metody dla danych pochodzących z mikromacierzy i RNA-seq, sprawdzeniu efektywności wprowadzonej metody oraz porównaniu efektywności metod już istniejących z zaproponowanym testem statystycznym. Na podstawie wstępnych badań możemy stwierdzić, że uwzględnienie efektu korelacji (preferowane będą te geny, które mają jak największe różnice w poziomach ekspresji pomiędzy klasami i równocześnie najmniejsze współczynniki korelacji pomiędzy poziomami ekspresji) pomiędzy wartościami ekspresji genów ma znaczący wpływ na analizę dyskryminacyjną. Chcielibyśmy sprawdzić, czy uwzględnienie korelacji w każdym przypadku (dla różnych typów danych) polepszy predykcje. Badając korelację pomiędzy genami uwzględnione zostaną różne współczynniki wykorzystywane przy budowie sieci koekspresji (tj. informacja wzajemna, korelacja rang Spearmana, odległość Euklidesowa, korelacja cząstkowa, analiza regresji, współczynnik Giniego oraz wspólczynnik Kendalla).

W projekcie wykorzystane zostaną metody symulacyjne w celu stworzenia danych o określonych strukturach korelacji oraz znanym podziale na klasy.  Zaproponowany zostanie model statystyczny wykorzystany w późniejszej  analizie danych mikromacierzowych i RNA-seq.  Zaproponowana zostanie konstrukcja testu statystycznego bazującego zarówno na wartościach ekspresji, jak i wartościach korelacji (sprawdzone zostaną różne miary badające korelację) pomiędzy tymi poziomami. W celu walidacji wyników zastosowane zostaną nastepujące metody uczenia maszyn: klasyfkator naiwny Baysowski, sieci neuronowe, klasyfikator najbliższych sąsiadów, mechanizm wektorów wspierających i lasy losowe. Zastosowana zostanie walidacja krzyżowa typu LOOCV oraz k-krotna walidacja. Na zakończenie porównane zostaną istniejące statystyki testowe zarówno te, które uwzględniają korelację, jak i te tradycyjne. Porównanie przeprowadzone będzie na różnych płaszczyznach. Wyznaczone zostaną błędy predykcji dla zastosowania każdej metody, krzywe ROC, wartości AUC oraz wykresy diagnostyczne. Na wszystkich etapach analiz jako narzędzie będzie wykorzystywany sprzęt komputerowy o wysokiej wydajności oraz platforma statystyczna R oraz Bioconductor.