Nr rejestracyjny: 2016/23/D/ST6/03613; Kierownik projektu: dr Neo Christopher Honghoon Chung

STRESZCZENIE POPULARNONAUKOWE

NOWE METODY STATYSTYCZNE DLA PROBLEMÓW KLASYFIKACJI I INTEGRACJI DANYCH MOLEKULARNYCH I GENOMOWYCH

Nowoczesne nauki przyrodnicze masowo generują wielowymiarowe dane, których analiza ma pomóc w zrozumieniu złożonych procesów molekularnych stanowiących podłoże wielu chorób. Wraz z rozwojem najnowszych technik pomiarowych ilość gromadzonych informacji narasta w wielkim tempie.

Jako przykład rozważmy pomiar poziomu mRNA. Zamiast projektować eksperyment dla jednego interesującego genu, biolog molekularny ma do dyspozycji mikromacierz i wysoko-przepustową technologię sekwencjonowania, które umożliwiają pomiar poziomu mRNA dla wszystkich genów aktywnych w danej komórce.

Podobnie możemy badać współcześnie różnorodność molekularną na wielu płaszczyznach takich jak zmienność sekwencji DNA czy poziomu metabolitów. Wysoko-przepustowe techniki pomiarowe, które umożliwiły rozwój takich dziedzin jak genomika funkcjonalna, czy medycyna genomowa, wymagają adekwatnych metod statystycznych integrujących dostępne dane, tak aby wykorzystać cały potencjał zawartej w nich informacji.

Bezpośrednią motywację dla tego projektu stanowią problemy jakie napotykamy analizując dane genomowe i metaboliczne. Postanowiliśmy odpowiedzieć na te wyzwania proponując nowe efektywne metody uczenia maszynowego, które mogą integrować różne źródła danych o procesach molekularnych.

W pierwszej kolejności zastanowimy się jak rozstrzygnąć, kiedy ogromne zbiory danych są w statystycznie istotny sposób różne (lub podobne) używając popularnych indeksów Tanimoto i Jaccarda. Analizowane przez nas dane obejmą: sekwencje DNA, profile obecności metabolitów oraz reakcje i cząsteczki chemiczne. Porównywanie tego rodzaju danych jest kluczowe przy identyfikacji ortologów, przeszukiwaniu baz molekularnych oraz w zadaniu klasyfikacji.

Następnie, zaproponujemy rozwiązania, które pozwolą ocenić jakość grupowania dużych zbiorów obserwacji. Mimo tego, że klastrowanie jest wszechobecne w genomice i metabolomice, nadal nierozwiązanym zagadnieniem jest statystyczna ewaluacja przypisania danego obiektu do grupy. Nasza metoda pozwoli oszacować na ile obiekty znalazły się w danej grupie w sposób losowy, a na ile taki podział jest stabilny. Zaproponowane podejście w oczywisty sposób może poprawić jakość wielu algorytmów klasyfikacji.

Ostatnim wyzwaniem jakie chcemy podjąć jest odpowiedź na bardzo konkretne pytanie: czy metabolizm tłuszczów może być odpowiedzialny za rozwój raka jajnika ? Badając to zagadnienie planujemy wykorzystać i rozszerzyć metodologię rozwijaną przez nas w pracy doktorskiej. Chcielibyśmy opracować metodę odkrywania wspólnej ukrytej struktury w heterogenicznych danych biologicznych. Wstępne badania wskazują na znaczącą rolę cząsteczek lipidowych w procesie rozwoju raka jajnika, na którego zachorowalność w Polsce jest bardzo wysoka. Mamy nadzieję, że zaproponowane przez nas metody statystyczne pozwolą na wyjaśnienie tej roli identyfikując kluczowe szlaki metaboliczne, które ulegają zaburzeniu w procesie nowotworowym.

Będąc gorącymi zwolennikami metod *open source*, zamierzamy wszystkie opracowane w ramach projektu algorytmy udostępnić jako pakiety w języku *R*, tak aby mogły posłużyć innym badaczom w podobnych zastosowaniach. *Last but not least*, projekt *Sonata*, który umożliwi zatrudnienie kierownika, będącego naukowcem pochodzącym z USA, na wiodącym polskim uniwersytecie, znacząco wzmocni interdyscyplinarną i międzynarodową współpracę pomiędzy biostatystykami, informatykami i biologami.