Silesian University of Technology in Gliwice, Poland
Automatic Control, Electronics, and Computer Science Department

Title: Classification of white blood cells based on single-cell sequencing data for biodosimetry purposes

Author: Katarzyna Sieradzka

Supervisor: prof. dr hab. inż. Joanna Polańska
Advisor: dr Christophe Badie

Acknowledgments: This work has been supported by European Union under the European Social Fund grant AIDA – POWR.03.02.00-00-I029.

Abstrakt

Sekwencjonowanie RNA pojedynczych komórek (scRNA-seq) jest coraz szerzej stosowaną technologią do analizy transkryptomu wielu pojedynczych komórek. Dzięki zastosowaniu sekwencjonowania genomu pojedynczych komórek można uniknąć problemu generalizacji danych występującego w technologiach sekwencjonowania, które nie koncentrują się na pojedynczych komórkach. W wyniku wykorzystania tej technologii generowane są dane wielowymiarowe, co wymaga coraz większych zasobów obliczeniowych do właściwej analizy. Technologia scRNA-seq jest niezbędna do badania heterogeniczności między komórkami w analizie wpływu określonych czynników, takich jak odpowiedź komórkowa na promieniowanie jonizujące. Zarówno nieświadoma, jak i świadoma ekspozycja na promieniowanie jonizujące wywołuje zmiany w odpowiedziach komórkowych spowodowane modyfikacjami ekspresji wielu genów regulujących życie komórek. Analiza takich modyfikacji może ujawnić geny najbardziej zaangażowane w odpowiedź na promieniowanie. Taka analiza może również wykazać ścieżki komunikacji genów, które mogą dać wgląd w to, jakie zmiany zachodzą w złożonym systemie komórkowym. Integrując wiedzę o stopniu zmian indukowanych promieniowaniem i dostępnymi technologiami sekwencjonowania, możemy przeprowadzić odpowiednie kroki analityczne, które pozwolą nam poznać geny reagujące na promieniowanie.

Przedstawiona rozprawa doktorska ma dwa główne cele. Jeden z nich ma podłoże biologiczne, a drugi związany jest z inżynierią. Celem biologicznym niniejszej pracy jest poszukiwanie znanych i nowych genów odpowiedzi na promieniowanie w oparciu o dane z technik sekwencjonowania RNA pojedynczych komórek. W ten sposób zostaną określone różnice w sygnaturze genowej normalnych komórek i tych poddanych działaniu promieniowania jonizującego w środowisku ex vivo. Podstawowym celem pracy w zakresie inżynierii jest stworzenie odpowiedniego schematu pracy analizy bioinformatycznej, aby częściowo zautomatyzować kolejne etapy pracy z wielowymiarowymi danymi pochodzącymi z eksperymentów sekwencjonowania pojedynczych komórek. Główne aspekty zawarte w proponowanej metodzie opierają się na procedurach selekcji cech oraz problemie klasyfikacji komórek. Jest to duże wyzwanie, zwłaszcza biorąc pod uwagę bardzo istotną złożoność i wymiarowość analizowanych danych, ale także oczekiwania nastawione na uzyskanie zadowalających wyników w zakresie jakości klasyfikacji komórek napromienionych. Oczekiwany wynik stworzonego narzędzia związany jest przede wszystkim z biologicznym celem badań, tj. rozpoznaniem pełnego profilu genetycznego komórek napromienionych w środowisku ex vivo.

Pierwszy etap prac koncentruje się na kontroli jakości danych. W tym celu przetestowano dwie próbki ex vivo, będące technicznymi powtórzeniami tego samego eksperymentu. Opracowano kilka ścieżek statystycznych i wizualizacyjnych, aby umożliwić szczegółową analizę jakości zarówno genów, jak i komórek poddanych analizie. Zastosowana metodologia, a zwłaszcza nienadzorowane podejście klasyfikacyjne zastosowane do celów wizualizacji, pozwala na wyciągnięcie jednoznacznego wniosku

o znacznej heterogeniczności badanych zbiorów danych. W związku z tym, podjęto próby ustalenia przyczyny takiej heterogeniczności komórek wykorzystując zarówno ogólnodostępne, jak i opracowane na potrzeby niniejszej rozprawy metody matematyczno-statystyczne. Ponadto, do rozpoznania subpopulacji komórek białych krwinek, wykorzystano również listę genów markerowych specyficznych dla określonych subpopulacji. Wykazano, że wybrana ścieżka badawcza pozwoliła na określenie przyczyny wewnętrznej heterogeniczności, w złożonych zbiorach danych, związanej z występowaniem wysoko zróżnicowanych podtypów komórek. Co więcej, w wyniku przeprowadzonych serii analiz udało się wykryć często występujące subpopulacje tej frakcji, w kontekście ilościowym, oraz rzadkie i małe subpopulacje białych krwinek. Główny etap prac ma na celu zbudowanie klasyfikatora opartego o metody regresji logistycznej. Zadaniem klasyfikatora jest rozróżnienie komórek kontrolnych i poddanych działaniu promieniowania jonizującego. Na tym etapie pracy uwzględniono jedynie subpopulację limfocytów T, gdyż stanowiła ona większość rozpoznanych podtypów analizowanych komórek białych krwinek. Co istotne, zastosowana procedura pozwoliła na usunięcie czynnika odpowiedzialnego za występowanie wykrytej heterogeniczności zbioru danych. Następnie, w celu ujednolicenia struktury analizowanego zbioru danych, przeprowadzono procedurę normalizacji. Kolejny etap selekcji cech został oparty na zbiorze komórek i genów przygotowany w przedstawiony sposób. Celem selekcji cech, samodzielnie opracowano i zaimplementowano schemat analizy, umożliwiający klasyfikację komórek normalnych i napromieniowanych, z wykorzystaniem odpowiednich miar jakości klasyfikacji. W wyniku zastosowania zaimplementowanego algorytmu uzyskano ostatecznie panel genów odpowiedzi na napromienienie. Co istotne, znaczna większość rozpoznanych genów odpowiedzi radiacyjnej odpowiada aktualnym doniesieniom literaturowym. Zmniejszenie wpływu heterogeniczności w zbiorze danych pozwoliło na poprawę jakości klasyfikacji i uzyskanie bardzo zadowalających wyników z wartością ważonej jakości klasyfikacji, opartej na testowym zbiorze danych, na poziomie powyżej 93%. Dodatkowo przeprowadzono szczegółową analizę z wykorzystaniem podejścia sieci neuronowych w celu porównania schematu pracy opartego o metody regresji logistycznej z inną, dobrze znaną metodą uczenia maszynowego. Utworzono drugi schemat analizy, który jest spójny z określonym celem rozprawy, czyli rozpoznaniem komórek napromienionych. Podejście to miało na celu przede wszystkim sprawdzenie i porównanie jakości klasyfikacji wynikających z zastosowania dwóch różnych technik selekcji cech. Wykorzystanie sieci neuronowych pozwoliło uzyskać obiecujące wyniki, z wartością ważonej jakości klasyfikacji na poziomie prawie 91%. Co więcej, takie wyniki uzyskano w znacznie krótszym czasie, porównując sieci neuronowe z podejściem opartym o metody regresji logistycznej. Z drugiej strony, co jeszcze ważniejsze przy przeprowadzaniu analiz zgodnie z zaproponowanym schematem analizy, możliwe było również porównanie profili genetycznych komórek napromienionych, rozpoznanych w wyniku zastosowania metod regresji logistycznej oraz sieci neuronowych. Okazało się, że 8 na 10 genów tworzących model oparty o sieci neuronowe jest spójnych z modelem opartym o regresję logistyczną. Te dobrze znane geny odpowiedzi na promieniowanie obejmują RPS19P1, BAX, DDB2, RPS27L, PHPT1, CCNG1, TNFSF8 i AEN.

Niniejsza rozprawa doktorska pokazuje, że wykorzystując dane pochodzące z precyzyjnej i szczegółowej technologii, takiej jak scRNA-seq, można określić specyficzną strukturę genów dla komórek poddanych działaniu promieniowania jonizującego. Przeprowadzone prace umożliwiły również porównanie dwóch technik uczenia maszynowego w kontekście selekcji cech. Kilka opracowanych metod bioinformatycznych, a przede wszystkim zaproponowany schemat analizy, mogą być w przyszłości wykorzystane jako wsparcie w medycynie, nauce i inżynierii. Opracowana metoda selekcji cech i klasyfikacji komórek napromienionych sprostała wyzwaniom postawionym w rozprawie z bardzo wysoką skutecznością. Badania te dokładnie opisują przebieg analizy danych wysokowymiarowych, pochodzących z eksperymentów sekwencjonowania pojedynczych komórek, takich jak: rozszerzona kontrola jakości, rozpoznanie genów odpowiedzi na promieniowanie, określenie sygnatury genowej komórek napromienionych, klasyfikację białych krwinek wraz z rozpoznawaniem określonych subpopulacji komórkowych, porównanie procedur uczenia maszynowego pod kątem analizy danych wysokowymiarowych i klasyfikacji obserwacji, a także biologiczną interpretację wyników. Niniejsza praca obejmuje, wraz ze szczegółowym opisem proponowanych etapów analizy oraz efektów w postaci wyników, wszystkie aspekty niezbędne do osiągnięcia założonych celów, łącząc je w logiczny schemat pracy wraz z odpowiednimi komentarzami i wnioskami zarówno od strony technicznej, inżynieryjnej, jak i wsparcia tych aspektów w postaci interpretacji biologicznej.