Prof. IBB Szymon Kaczanowski **Warszawa 14.01.2020**

Recenzjarozprawydoktorskiej
magister Mai Małkowskiej **pt.**
„Opracowanie metody analizy sekwencji polimorficznych i funkcjonalności w regionach promotorowych"

Magister Maja Małkowska przygotowała rozprawę doktorska pod tytułem „*Opracowanie metody analizy sekwencji polimorficznych i funkcjonalności w regionach promotorowych*" pod kierunkiem profesora Lucjana Wyrwicza.

Celem przedstawionej rozprawy było stworzenie komputerowej metody przewidywania wpływu konkretnych mutacji na aktywność transkrypcyjną ludzkich genów. Z oczywistych względów tego typu metoda może mieć niezwykle istotne znaczenie dla rozwoju nauk medycznych.

W swojej pracy autorka wykorzystała uczenie maszynowe. W takim podejściu wykorzystuje się znane informacje o danym zjawisku do stworzenia modelu matematycznego, który opisuje to zjawisko. W epoce przed-komputerowej naukowcy tworzyli proste modele matematyczne. Dzisiaj, dzięki rozwojowi technologicznemu, przy pomocy komputerów można tworzyć bardzo złożone modele matematyczne. Uczenie maszynowe polega na tworzeniu modeli matematycznych opisujących dany proces przy pomocy metod komputerowych.

W rozprawie użyto uczenia maszynowego do klasyfikacji mutacji w promotorach:

i) mutacje, które mają wpływ na aktywność transkrypcyjną genu
ii) mutacje, które nie mają takiego wpływu.

W przeprowadzonych badaniach wykorzystano istniejącą wiedzę na temat regulacji genetycznej człowieka. Skorzystano tutaj z faktu, że w bazie danych HGMD (Human Gene Mutation Database) opisane są mutacje, które mają wpływ na regulacje aktywności genów. Doktorantka uznała natomiast, że mutacje polimorficzne w promotorach pojawiające się w populacji człowieka w częstości mniejszej niż 1% o których, że nie ma informacji, że powodują choroby genetyczne, nie zmieniają aktywności transkrypcyjnej genu.

Uczenie maszynowe zostało zastosowane do rozróżnienia tych dwóch typów mutacji i stworzenia klasyfikatora (czyli modelu matematycznego, który rozróżnia te dwa typy mutacji).

Podczas uczenia wykorzystano 227 różnych cech opisujących mutacje:

1. 52 cechy opisujące lokalną sekwencje DNA
2. 88 cechy opisujące przewidziany metodami komputerowymi lokalny kształt DNA

3. 8 cech określających zawartość par GC
4. 38 cech opisujących modyfikacje histonów obserwowane metodami eksperymentalnymi w okolicy analizowanych mutacji
5. 12 cech opisujących występowanie motywów wiążących czynniki transkrypcyjne w okolicy mutacji
6. 1 cechę wskazującą, czy analizowana mutacja ma bezpośredni destrukcyjny wpływ na motywy wiążące czynniki transkrypcyjne
7. 1 cechę wskazującą czy mutacja znajduje się w rejonie wrażliwym na działania DNazy 1 (czyli w regionie „otwartej chromatyny")
8. 10 cech określających konserwacje ewolucyjną danego motywu
9. 16 cech określających pary nukleotydów występujących w analizowanych krótkich motywach sekwencyjnych

Do stworzenia klasyfikatora użyto algorytm wzmacniania gradientowego drzew decyzyjnych (po angielsku Gradient Tree Boosting). Algorytm ten pozwala na wytrenowanie klasyfikatora używającego drzewa decyzyjne. Wcześniejsze badania wskazują, że algorytm ten często daje lepsze rezultaty inne algorytmy (np. sieci neuronowe). Jest to z pewnością, bardzo dobrze dobrana metodologia.

Przy pomocy algorytmu_MCFS (algorytm wyboru cech Monte Carlo) doktorantka uzyskała ranking ważności cech. Okazało się, że stosowany model matematyczny można istotnie uprościć używając w nim tylko najbardziej istotnych cech opisujących skład par GC, przewidziany komputerowo kształt DNA w okolicach mutacji i sekwencję. Co więcej taki uproszczony algorytm uzyskiwał lepsze wyniki, niż bardziej skomplikowane modele używające większą liczbę cech. Algorytm ten nazwano ShapeGTB. Algorytm ten jest też lepszy od starszych algorytmów używanych do przewidywania wpływu mutacji promotorowych na aktywność genów: FATHMM-MKL, CADD, DeepSEA.

Przedstawione w rozprawie doktorskiej wyniki stanowią więc istotny wkład do rozwoju medycznych metod bio-informatycznych. Wyniki przedstawione w rozprawie zostały opublikowane w specjalistycznym periodyku PeerJ.

Mam jednak jedną poważną krytyczną uwagę dotycząca przedstawionej metodologii.

Jak już napisałem doktorantka uznała, że mutacje polimorficzne w promotorach pojawiające się w populacji człowieka w frekwencji mniejszej niż 1% o których, że nie ma informacji, że powodują choroby genetyczne, nie zmieniają aktywności transkrypcyjnej genu. Moim zdaniem nie jest to najlepiej wybrana grupa. Mutacje pojawiające się rzadziej w populacji często mają jednak negatywny wpływ na dostosowanie i są przedmiotem tak zwanego doboru oczyszczającego. Także w przypadku człowieka, choroby genetyczne są rzadkie w populacji. Fakt, że nie wiemy, że dana mutacja ma fenotypowe konsekwencje nie oznacza, że takich nie ma. Dlatego uważam, że lepiej było użyć mutacje polimorficzne pojawiające się w populacji człowieka częściej niż w 10% populacji. Przypuszczam, że tego typu zbiór zawierałby mniej szumu i umożliwiłby uzyskanie lepszych wyników.

Mam też sugestie dotyczące dodatkowych analiz, których wykonanie mogłoby podnieść wartość naukową wyników.

W moim przekonaniu też ciekawe byłoby pokazanie na wykresie zależności między wartościami cech (np. zawartością GC, zmianą zawartości GC), a prawdopodobieństwem, że mutacja spowoduje zmiana transkrypcji genu. Tego typu analizy pozwoliłyby lepiej zrozumieć badane zależności.

Szczególnie interesujące mogą być prostszy model oparty wyłącznie na zmiennej opisującej zmianę składu par GC spowodowanych przez mutacje. Rycina ósma rozprawy doktorskiej pokazuje, że cecha ta ma wpływ na inne cechy użyte w klasyfikatorze. Co więcej suplement II wskazuje, że najbardziej istotną cechą klasyfikatora jest właśnie ten parametr.

Z recenzenckiego obowiązku muszę ocenić samą formę rozprawy doktorskiej.

Rozprawa doktorska jest napisana w języku polskim i zawiera 91 stron. Nie potrafiłem doszukać się w niej poważnych błędów literowych, co świadczy o starannej edycji. Warto zwrócić uwagę na to, że rozprawa zawiera bardzo interesujący i wyczerpujący wstęp, w którym opisane są stosowane metody.

Wadą rozprawy doktorskiej jest niska precyzja. Czytelnikowi trudno często zrozumieć szczegóły. Nawet przeczytanie publikacji (PeerJ [redacted]), w oparciu o która przygotowano rozprawę doktorską nie do końca umożliwia ich zrozumienie.

Na przykład autorka pisze na stronie 39 o 52 zmiennych opisujących lokalną sekwencje DNA (dokładniej 9 nukleotydową lokalną sekwencje,), które są kodowane przy użyciu 4 bitowego kodowania binarnego. Czytelnikowi trudno jest zgadnąć o jakie zmienne tutaj chodzi. Szczególnie, że w dalszej części czytelnik dowiaduje się, że skład GC, nie jest jedną z tych 52 cech.

Autorka pokazuje w rozprawie wyniki dotyczące wszystkich cech opisujących zawartość GC i cechy opisującej różnice zawartości ilości GC spowodowanej przez mutacje. W efekcie czytelnikowi sprawia dużą trudność zrozumienie, jakie analizy są omawiane.

Problem ten szczególnie jest wyraźny w przypadku opisu ryciny numer 7 ze strony 46 zatytułowanej „Średnia wartość względnego znaczenia 5 najwyżej ocenianych cech". Rysunek ten jest inny, niż pokazany w publikacji. i zatytułowany "Mean importance of five best scoring features in each feature group". Według zamieszczonego opisu największe znaczenie dla konstrukcji poprawnych klasyfikatorów mają parametry opisujące następujące cechy: zawartość GC, lokalny kształt DNA i sekwencja. Tak wynika też z angielskiego obrazka. Natomiast polska rycina sugeruje, że najbardziej istotnymi cechami są: „zawartość GC, lokalny kształt DNA i nadwrażliwość na DNazę 1. Jednocześnie w rozprawie autorka odsyła czytelnika do suplementu II, w którym opisane są szczegółowe wartości opisujące znaczenia różnych cech. Jak już wspomniałem z suplementu wynika, że najbardziej istotną cechą jest różnica w zwartości GC spowodowana przez mutacje. Jest to jednak tylko jedna z ośmiu cech, które opisują zawartość GC.

W moim przekonaniu tego typu drobne niedociągnięcia wynikają z tego, że przedmiot rozprawy doktorskiej jest skomplikowany. Nie umniejszają one mojej pozytywnej oceny samej rozprawy.

Literatura jest starannie dobrana.

Podsumowując opisane w rozprawie badania są niezwykle ciekawe, dotyczą bardzo istotnego zjawiska z punktu widzenia rozwoju nauki. Doktorantka włożyła dużą ilość pracy w wykonanie badań. Natomiast redakcja i sposób napisania samej rozprawy zawiera pewne drobne niedociągnięcia, które jednak podaję z recenzenckiego obowiązku.

Stwierdzam więc, że oceniana rozprawa doktorska spełnia wszystkie wymagania zwyczajowe oraz stawiane przez ustawodawcę. Wnoszę o dopuszczenie doktorantki do dalszych etapów obrony doktoratu.

Ponadto ponieważ wyniki uzyskane w rozprawie doktorskiej są opublikowane i doktorantka posiada już dorobek publikacyjny składający się z pięciu publikacji to wnoszę do rady naukowej Centrum Instytutu Onkologii imienia Marii Skłodkowskiej Curie o rozważenie przyznania stosownego wyróżnienia zgodnego z panującym zwyczajem.

Szymon Kaczanowski

Szymon Kaczanowski