Wydział Matematyki
i Nauk Informacyjnych

Warszawa, 12 grudnia 2018

Dr hab. inż. Przemysław Biecek
prof. nadzw. Politechniki Warszawskiej
Wydział Matematyki i Nauk Informacyjnych
Koszykowa 75/507
00-662 Warszawa

Recenzja rozprawy doktorskiej „Przewidywanie właściwości sekwencji biologicznych w oparciu o analizę n-gramów" mgr Michała Burdukiewicza. promotor: prof. dr hab. Paweł Mackiewicz, promotor pomocniczy: dr Paweł Błażej.

Informacje ogólne

Praca doktorska pana Michała Burdukiewicza jest poświęcona metodom oraz narzędziom bioinformatycznym wspomagającym przewidywanie właściwości białek i kwasów nukleidowych. Metody opisane w pracy są oparte o techniki analizy statystycznej i techniki uczenia maszynowego, głównie lasy losowe i modele semi- markowskie z ukrytymi stanami. Narzędzia te są dostosowane do analizy danych sekwencyjnych, w szczególności sekwencji nukleotydowej oraz aminokwasowej. Autor w rozprawie przytoczył trzy przykłady wykorzystania tych narzędzi, do predykcji amyloidogenności, predykcji peptydów sygnałowych i predykcji optymalnych warunków hodowlanych dla metanogenów. Tematyka pracy wpisuje się w ważny i popularny trend badania własności białek na podstawie sekwencji.

Politechnika Warszawska
ul. Koszykowa 75
00-662 Warszawa
www.mini.pw.edu.pl

Do pracy załączone są dwie publikacje:

- Praca „*Amyloidogenic motifs revealed by n-gram analysis*" opublikowana w 2017 roku w Scientific Reports (40 pkt na liście MNiSW). Praca ma 6 autorów, pan Michał Burdukiewicz jest pierwszym autorem tej pracy. Praca jest związana z trzecim rozdziałem pracy doktorskiej, opisuje model predykcyjny dla identyfikacji amyloidów na podstawie właściwości sekwencji aminokwasowej,
- Praca „*PhyMet2: a database and toolkit for phylogenetic and metabolic analyses of methanogens"* opublikowana w 2018 roku w Environmental Microbiology Reports (35 pkt na listach MNiSW). Praca ma 7 autorów, pan Michał Burdukiewicz jest pierwszym autorem tej pracy. Praca jest związana z czwartym rozdziałem rozprawy doktorskiej. Praca opisuje bazę danych PhyMet2 oraz narzędzie do przewidywania optymalnych parametrów hodowli metanogenów.

W rozdziale 8 rozprawy doktorskiej zamieszczona jest również lista dziesięciu innych prac, wszystkich opublikowanych w czasopismach z listy A, których pan Burdukiewicz jest autorem lub współautorem. W tym dwukrotnie występuje czasopismo Bioinformatics. Większość z tych prac jest poświęcona narzędziom informatycznym wspierającym analizy bioinformatyczne. Prace te powstawały we współpracy z silnymi zespołami badaczy z Polski, Danii i Niemiec. Z pewnością możliwości prowadzenia tak różnorodnych badań miały wpływ na poszerzenie doświadczenia badawczego i metodologicznego doktoranta.

W pracy autor odwołuje się do trzech opracowanych przez siebie webserwisów:

- Narzędzia do identyfikacji sekwencji sygnałowej, które jest oparte o modele semi- markowskie z ukrytymi stanami. Webserwis o nazwie **signalHsmm** jest dostępny pod adresem http://smorfland.uni.wroc.pl/signalhsmm;
- Narzędzia do predykcji sekwencji amyloidogennych **AmyloGram.** Webserwis dostępny pod adresem http://www.smorfland.uni.wroc.pl/shiny/AmyloGram/.
- Narzędzia do predykcji optymalnych warunków hodowlanych metylogenów **MethanoGram**. Webserwis dostępny pod adresem http://smorfland.uni.wroc.pl/shiny/MethanoGram.

Politechnika
Warszawska

ul. Koszykowa 75
00-662 Warszawa
www.mini.pw.edu.pl

Powyższe webserwisy są wykonane w oparciu na bibliotekę shiny dla programu R. Każdy z nich jest związany z pakietem lub skryptami R pozwalającymi na pełną reprodukowalność wyników.

Budowanie webserwisów jest częste w obecnej literaturze. Konkurencją dla AmyloGram są np. narzędzia RFAmyloid (2018) czy pakiet appnn (2015).

Już w tym miejscu należy zauważyć, że:

- Dorobek publikacyjny pana Michała Burdukiewicza jest bardzo obszerny, zdecydowanie powyżej średniego. Parametry bibliograficzne są imponujące jak na pracę doktorską.
- Pan Michał Burdukiewicz pracuje z dużymi i silnymi, międzynarodowymi zespołami. Dowodzi to wysokich umiejętności dotyczących nawiązywania współpracy oraz tworzenia i publikowania wyników.
- Pod kątem inżynierskim, opracowane narzędzia bioinformatyczne są bardzo wysokiej jakości. Kod jest odpowiednio zdekomponowany, podzielone na logiczne moduły. Webserwisy są opisane i wyposażone w czytelne przykłady.

Przechodząc do rozprawy doktorskiej.

Praca doktorska pana Michała Burdukiewicza składa się z 62 stron (pomijając załączniki, wstęp i spis treści). Podzielona jest na cztery główne rozdziały. Pierwszy rozdział opisuje narzędzia statystyczne i bioinformatyczne opracowane i wykorzystywane przez autora. Drugi rozdział przedstawia aplikacje tych narzędzi do problemu predykcji peptydów sygnałowych. Trzeci rozdział przedstawia zastosowanie do predykcji peptydów aminokwasowych. Czwarty do problemu predykcji warunków hodowlanych metanogenów.

Poniżej omówię każdy z tych rozdziałów.

Rozdział 1

Pierwszy rozdział opisuje narzędzia statystyczne opracowane i zaimplementowane w pakiecie **biogram** dla programu R. Rozdział składa się z 6 stron. Wprowadza pojęcie n-gramu, opisuje procedurę weryfikacji istotności n-gramu oraz sposób konstrukcji alfabetów aminokwasowych.

Przedstawione w tym rozdziale pojęcia są później stosowane do rozwiązywania różnych problemów biologicznych. Prawdopodobnie ta różnorodność jest przyczyną kilku nieścisłości. Jedną, którą dobrze byłoby wyjaśnić, to definicja n-gramu, centralnego bytu w wielu stosowanych rozwiązaniach. Definicja ze strony 16 dopuszcza przerwy w n-gramach. Ale wyliczenia na liczbę n-gramów przestawione na stronie 12 (u do potęgi n, gdzie u to liczba liter w alfabecie) są poprawne tylko przy założeniu, że przerw nie ma.

W rozdziale 1 autor przedstawia procedurę testowania statystycznego, którą nazywa QuiPT (Quick Permutation Test). Z tą procedurą wiąże się kilka problemów:

1. Procedura QuiPT jest przedstawiana jako test statystyczny. Aby jednak opis testu był kompletny należy określić hipotezę zerową i hipotezę alternatywną oraz statystykę testową. Tych elementów zabrakło, przez co nie wiadomo do końca czy autorowi chodzi o dwustronną hipotezę alternatywną czy jednostronną.
2. Określenie „szybki test permutacyjny" jest mylące. Skonstruowana procedura nie jest oparta o permutacje, nie jest więc testem permutacyjnym. Dla testu niezależności (bo o taki test chodzi autorowi) jednym z rozwiązań na wyznaczenie rozkładu statystyki testowej jest skorzystanie z testów permutacyjnych. Procedura opisana jako QuiPT wyznacza rozkład pewnej statystyki testowej w sposób dokładny, nie korzystając z permutacji.
3. Opisana procedura dla tablic 2x2 jest identyczna z procedurą znaną od lat jako dokładny test Fishera.
4. Nie jest wprost napisane czy mamy do czynienia z testem jednostronnym czy dwustronnym. Komentarz z dołu strony 18 sugeruje, że to jednak test dwustronny. W takim przypadku warto sprawdzić czy wrzucenie do jednego worka identyfikacji nadreprezentowanych n-gramów i niedoreprezentowanych n-gramów jest lepszym pomysłem niż zastosowanie testu jednostronnego. W pewnych zagadnieniach, szczególnie z wieloma

etykietami, lepsze wyniki uzyskuje się wykorzystując tylko nadreprezentowane markery.

5. Przy proponowaniu testu statystycznego należy porównać moc nowo-zaproponowanego testu z obecnie stosowanymi rozwiązaniami. W przypadku testu QuiPT autor pisze, że test można go oprzeć o rozkład wzajemnej informacji lub inną statystykę. Różne statystyki mogą prowadzić do testów o różnej mocy. Stąd pytanie, czy użycie wzajemnej informacji, a nie innej statystyki poprawia lub choć nie pogarsza wyników.
6. Nie jest prawdziwe stwierdzenie, że użycie dokładnego rozkładu statystyki testowej pozwala na uzyskiwanie dowolnie małych p-wartości. Problemem jest tutaj dyskretny rozkład statystyki testowej (opartej o tabele zliczeń), który powoduje, że można otrzymać jedynie skończoną liczbę różnych p-wartości. Paradoksalnie przewagą testu permutacyjnego w tym przypadku jest jego zrandomizowanie.

Na usprawiedliwienie tych usterek, trzeba wziąć pod uwagę dziedzinę doktoratu. Nie jest to doktorat ze statystyki. Warto jednak w przyszłości konsultować takie rozwiązania ze statystykami. Nie jest to w żaden sposób znaczące uchybienie doktoranta, raczej bolączka całej bioinformatyki, która pewne rozwiązania wynajduje na nowo.

Pomimo wyżej wymienionych usterek, **otrzymujemy rozwiązanie działające szybko i dobrze, o czym świadczą wyniki walidacji.**

Druga część rozdziału 1 jest poświęcona uproszczonym alfabetom aminokwasowym. Przedstawiona jest ogólna koncepcja, brakuje jednak zarówno szczegółowego opisu jak i dyskusji przyjętych rozwiązań. Ostatnie zdanie ze strony 20 sugeruje klastrowanie hierarchiczne z algorytmem łączenia Warda. Ale brakuje informacji o zastosowanej funkcji odległości oraz o zastosowanej normalizacji. Obie te rzeczy mają silny wpływ na wyniki. W kolejnych rozdziałach autor przedstawia wybrane przykłady zredukowanych alfabetów, co trochę ułatwia zrozumienie podjętych wyborów.

Dla mnie, jako statystyka, oglądanie wyników dotyczących zredukowanych alfabetów bardzo silnie przywołuje na myśl dwa problemy – regularyzacji (mniejszy alfabet, mniejsze ryzyko przeuczenia) i – reprezentacji cech (lepsza reprezentacja cech, łatwiejsze uczenie). Szczególnie ten drugi punkt wydaje się być istotny. W obliczu spektakularnych wyników, uzyskiwanych przez modele głębokiego uczenia sieci neuronowych (też

ostatni wynik algorytmu AlfaFold), warto autor pracy, już jako młody doktor przyjrzał się możliwościom jakie dają techniki głębokiego uczenia.

Rozdział 2

W rozdziale drugim przedstawiony jest model do identyfikacji peptydów sygnałowych oparty o modele semi- markowskie. Rozdział składa się z 20 stron, na których opisana jest zarówno metodologia tworzenia modelu predykcyjnego jak i opis weryfikacji wyników.

Rozdział rozpoczyna omówienie problemu identyfikacji peptydów sygnałowych. Omówione jest rozwiązanie oparte o klasyczne modele markowskie oraz wskazane są słabe punkty tego podejścia. Następnie wprowadzone jest rozszerzenie oparte o modele semi- markowskie oraz o zredukowany alfabet aminokwasów. Zaprezentowane tutaj wyniki dotyczą różnych możliwych alfabetów (rysunek 8 w rozprawie). W szczegółach porównany jest alfabet skutkujący największą czułością z alfabetem o największej specyficzności. Wybieranie alfabetu osobno o największej specyficzności i czułości jest jednak zaskakujące, ponieważ te dwa współczynniki są ze sobą związane. Zazwyczaj kosztem specyficzności można zwiększyć czułość i odwrotnie. Odpowiednio zmieniając próg odcięcia można więc każdy alfabet zamienić na taki o największej czułości. Obserwacje tę uzasadnia też rysunek 8, ponieważ tutaj zazwyczaj alfabety o wyższej czułości mają mniejszą specyficzność. Bezpieczniejszą miarą do porównań jest AUC. Szczęśliwie, sądząc po tabeli 8, wybrane alfabety mają również wysokie wartości AUC.

W tabeli 5 przedstawione jest odchylenie standardowe dla różnych współczynników, nie jest jednak określone jakim sposobem to odchylenie standardowe był mierzone. Jest to istotny element, bez niego nie wiadomo czy przedstawione alfabety się istotnie od siebie różnią. Mam nadzieję, że ten punkt będzie wyjaśniony podczas obrony, ponieważ odchylenia standardowe są zaskakująco małe, biorąc pod uwagę niewielką liczbę danych i przyjętą technikę weryfikacji krzyżowej.

Na stronie 40 autor odwołuje się do interpretowalności oraz podaje charakterystyki regionów c i h oparte o hydrofobowość. Brakuje jednak odniesienia do wyników, tabeli, rysunków, z których pochodzą te wnioski.

Na rysunku 10 brakuje informacji jaką ilość wariancji wyjaśniają składowe na panelach A i B.

Rozdział 3

Rozdział trzeci poświęcony jest predykcji amyloidegenności peptydów. Rozdział składa się z 20 stron na których przedstawiona jest metoda konstrukcji zbioru uczącego. Sposób konstrukcji modeli predykcyjnych oraz opis procedury weryfikacji eksperymentalnej uzyskanych wyników.

O ile sam proces modelowania nie jest tutaj zaskakujący i jest związany ze standardowymi dobrymi praktykami, o tyle interesujący jest sposób walidacji wyników przez weryfikacje eksperymentalną dla wybranych ośmiu heksapeptydów wskazanych przez model jako te o największym prawdopodobieństwie amyloidogenności.

Weryfikacja eksperymentalna potwierdziła dużą skuteczność modelu opracowanego przez autora, co jest silniejszym argumentem niż klasyczna weryfikacja przez miary typu AUC.

Pan Michał Burdukiewicz w kilku miejscach podkreśla wagę interpretowalności zbudowanego modelu. W tym kontekście bardzo ciekawym wynikiem jest rysunek 15, który wskazuje n-gramy najsilniej różnicujące grupę amyloidów i nieamyloidów. Ciekawym rozszerzeniem tego wykresu byłoby zaprezentowanie na osi OX nie częstości ale log-szansy (log-odds), lub wprost ilorazu szans dla obu grup (log odds ratio). Łatwiej byłoby odczytać nie tylko na ile dany motyw jest częsty w grupie amyloidów lub nieamyloidów, ale też ilokrotnie jest częstszy w jednej grupie względem drugiej grupy. Wykonanie tego kroku mogłoby prowadzić do konstrukcji w pełni interpretowalnego modelu opartego np. o uogólnione modele liniowe.

Rozdział 4

Ostatni rozdział pracy doktorskiej jest poświęcony predykcji optymalnych warunków hodowlanych dla metanogenów. Rozdział składa się z 10 stron.

Z punktu widzenia statystycznego jest to szalenie interesujący model w którym predykcja dotyczy czterech różnych parametrów (czasu podwojenia populacji, optymalnej temperatury, optymalnego stężenia NaCl i optymalnego pH). Problem (nazywany multi-label classification / multivariate regresion) daje duże pole popisu dla badania zależności pomiędzy przewidywanymi cechami.

Niestety autor buduje niezależne modele dla każdego z parametrów, większy nacisk kładąc na automatyczne strojenie parametrów dla poszczególnych modeli. Być może kwestia jednoczesnego modelowania, potencjalnie skorelowanych cech, będzie przedmiotem dalszych badań.

Uwagi redakcyjne

Praca zawiera pewne potknięcia redakcyjne. Nie wpływają one na ocenę merytoryczną, ale obniżają komfort czytania pracy. Najpoważniejsze usterki to:

1. W rozprawie doktorskiej umieszczane są spolszczone wersje wykresów z załączonych artykułów naukowych. Nie mają one jednak referencji do oryginalnych prac, a powinny zawierać przypisy lub odniesienia do źródłowych wykresów. Przykład: Rysunek 13 i 15 w rozprawie doktorskiej to Figure 2 i 3 z pracy o Amyloidach. Rysunek 19 w rozprawie doktorskiej to odpowiednio Figure S2 z pracy PhyMet2.
2. Brakuje kilku istotnych cytowań. Wszystkie opracowane przez autora narzędzia są zaimplementowane w programie R jednak sam program R nie jest jednak cytowany. Powinien być jako*: R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. Podobnie wszystkie webserwisy oparte są o bibliotekę **shiny**. Ona również nie jest cytowana. Wykorzystywany pakiet **ranger** ma cytowanie pracy z serwisu *arxiv*, choć zgodnie z informacjami umieszczonymi w pakiecie powinna być cytowana praca z *Journal of Statistical Software* z roku 2017.
3. W rozprawie doktorskiej jest wiele pojedynczych znaków na końcach linii (tzw. zawieszki) oraz pojedynczych linii na końcach lub początkach stron (tzw. wdowy i sieroty). Te usterki powinien usunąć skład.

4. Wzory matematyczne nie są numerowane, przez co trudno się do nich odwołać. Notacja matematyczna nie jest spójna (np. na stronie 17 i 18). Miejscami dwa dywizy skleiły się w półpauzę (np. strona 16).

Uwagi końcowe

Bogata liczba artykułów naukowych, których pan Michał Burdukiewicz jest autorem lub współautorem, potwierdza głęboką wiedzę, szerokie zainteresowania i dużą umiejętność wchodzenia we współprace. To bardzo ważne cechy dla naukowca. Z tego bogatego wachlarza wyników autor wybrał trzy problemy, które złożył w rozprawę doktorską. Każdy z tych ważnych problemów biologicznych został rozwiązany w inny sposób, z wykorzystaniem stosownego aparatu narzędzi statystycznych lub uczenia maszynowego.

Dla problemu predykcji peptydów sygnałowych autor opracował model na danych dla przedstawicieli *Plasmodiidae*, który uzyskuje również wysoką skuteczność dla przedstawicieli innych organizmów. W dzisiejszych czasach zbudowanie modelu jest proste, ale zbudowane stabilnego modelu proste wcale nie jest. Autorowi udało się tego dokonać.

Dla problemu peptydów amyloidogennych autor nie tylko zbudował model, ale też wyłuskał z niego fizykochemiczne właściwości markerów. Model zweryfikował eksperymentalnie a przy jego udziale współpracujący zespół odkrył nowe białko amyloidalne.

Autor nie ogranicza się do zaproponowania koncepcji rozwiązania. Dla każdego problemu zbudował metodologię, opracował model a zbudowany model przekształcił w narzędzie dostępne online jako webserwis. Rozwiązania te są wysokiej jakości inżynierskiej, nie powstydziłby się ich doktorant statystyki czy informatyki.

Obok zbudowanych webserwisów autor w wybranych przypadkach wykonał walidacje eksperymentalną. Pomimo stosowania złożonych modeli predykcyjnych zadbał o interpretowalność wyników.

Co ważne, narzędzia opracowane przez pana Michała Burdukiewicza nie kurzą się na półce. Autor aktywnie dba o ich promocje na konferencjach i wyjazdach badawczych. Skutkuje to ciekawym wykorzystaniem.

Przykładowo pakiet AmyloGram przysłużył się od identyfikacji nowego ameloidu, jak opisano w pracy „*The Sheaths of Methanospirillum Are Made of a New Type of Amyloid Protein*" przez zespół Christensen et al. w roku 2018.

Lektura pracy doktorskiej pana Michała Burdukiewicza i współtworzonych przez niego artykułów naukowych, pokazuje zarówno głęboką znajomość narzędzi wykorzystywanych w uczeniu maszynowy (lasy losowe, modele łańcuchów Markowa) jak i głęboką znajomość istotnych problemów i wyzwań biologicznych. Te dwie silne strony tworzą unikalny zestaw kompetencji. Te cechy, w połączeniu z dużymi umiejętnościami organizacyjnymi oraz olbrzymią aktywnością pana Michała Burdukiewicza, wróżą wiele dalszych, spektakularnych sukcesów w nauce.

Biorąc pod uwagę powyższe, stwierdzam, że rozprawa mgr Michała Burdukiewicza spełnia warunki ustawowe stawiane pracom doktorskim i wnioskuję o dopuszczenie autora do dalszego toku przewodu doktorskiego. Uwzględniając wysoki poziom merytoryczny rozprawy, oraz dorobek publikacyjny doktoranta wnoszę o wyróżnienie rozprawy.

Przemysław Biecek

Politechnika
Warszawska

ul. Koszykowa 75
00-662 Warszawa
www.mini.pw.edu.pl