dr hab. Krzysztof Topolski
Instytut Matematyczny
Uniwersytet Wrocławski
pl. Grunwaldzki 2/4
50-384 Wrocław

Wrocław, 20 czerwca 2014

Recenzja rozprawy doktorskiej mgr. Macieja Sobczyńskiego zatytułowanej

Wykorzystanie sztucznych sieci neuronowych i algorytmów genetycznych do badania proteomów

A. Uwagi ogólne.

Obecnie istniejące bazy danych biologicznych, oferujące łatwy dostęp do całkowicie zsekwencuionowanych genomów prokariotycznych i eukariotycznych oraz stała ich aktualizacja o kolejne zsekwencionowane genomy sprawiły, że w centrum zainteresowań biologii znalazła się możliwość integrowania tego typu danych. Jednocześnie ogromna i w szybkim tempie rosnąca ilość informacji genomowych wymaga zastosowania specjalistycznych narzędzi informatycznych, obliczeniowych i statystycznych. Statystyka wspomagała badania prowadzone w obszarze genetyki od samego początku. Gwałtowny postęp i intensyfikacja badań w genetyce i biotechnologii, jaką obserwujemy w ostatnich latach, nie tylko umocnił znaczenie statystyki jako nauki wspomagającej, ale również przyczynił się do rozwoju metod analizy danych i statystyki, czy wręcz do wyodrębnienia się nowych dyscyplin naukowych, za jaką można uznać bioinformatykę. Recenzowana rozprawa doktorska wpisuje się w ten nurt rozwoju współczesnej biologii. Rozprawa doktorska magistra Macieja Sobczyńskiego została napisana pod kierunkiem dr hab. Pawła Mackiewicza. Jest napisana w języku polskim i liczy 164 strony. Składa się ze wstępu, wydzielonego rozdziału, w którym autor przedstawia cel pracy, dwóch zasadniczych rozdziałów omawiających materiał badawczy, zastosowane metody, otrzymane wyniki oraz ich dyskusję. W ostatnim rozdziale autor przedstawia swoje wnioski z przeprowadzonych badań. Praca zaopatrzona jest dodatkowo w zawierającą 107 pozycji bibliografii, 33 stronicowy dodatek oraz wykaz publikacji autora oraz zgłoszeń konferencyjnych związanych z rozprawą doktorską. Całość pracy poprzedzona jest krótkim streszczeniem zarówno w języku polskim jak i angielskim. Praca została zredagowana w staranny sposób i tylko nieliczne jej fragmenty wymagają chwili zastanowienia w celu prawidłowego odczytania intencji autora. Napisana jest w sposób przejrzysty i przemyślany.

B. Przedmiot rozprawy.

Problematyka rozprawy doktorskiej, zgodnie z tytułem, dotyczy wykorzystania sieci neuronowej, a dokładnie sieci Kohonena, do przeanalizowania związku cech środowiskowych ze strukturą proteomu. Analizę tego zagadnienia autor przeprowadził na sekwencjach aminokwasowych potencjalnych genów, kodujących białka w 194 kompletnie zsekwencjonowanych genomach prokariotycznych bakterii i archeonów pobranych z bazy GenBank. Poddane analizie gatunki bakterii i archeonów autor pogrupował ze względu na 5 kategorii cech środowiskowych; temperaturę wzrostu,

wymagania tlenowe, preferowane zasolenie, stosunek do komórki gospodarza oraz środowisko życia, co dało 19 grup ekologicznych. Przypisując każdemu analizowanemu gatunkowi dokładnie jedną grupę ekologiczną w obrębie każdej z 5 kategorii środowiskowych autor oparł się na źródłach literaturowych oraz informacjach zamieszczonych na stronach NCBI. Autor za cel pracy postawił sobie ambitne zadanie, a mianowicie nie tylko wykazanie istnienia związku wyszczególnionych cech ekologicznych ze składem aminokwasowym proteomów, ale również zbadanie siły tego związku, wskazanie grupy białek, dla których ten związek jest najsilniejszy, odpowiedź na pytanie czy związek ten dotyczy całego białka, czy tylko niektórych jego elementów strukturalnych. Dodatkowo autor stara się rozstrzygnć na ile podobieństwo w składzie aminokwasowym proteomów można wytłumaczyć pokrewieństwem ewolucyjnym organizmów, a ile z tego podobieństwa należy przypisać presji mutacyjnej i selekcji.

C. Opis wyników rozprawy.

We wstępie autor przedstawia tło badanych w rozprawie zagadnień, omawia aktualny stań badań, omawia najważniejsze dostępne w literaturze wyniki oraz wysuwane hipotezy. W pierwszym po wstępie rozdziale autor wyraźnie precyzuje cel swojej rozprawy.

W rozdziale 3 autor omawia charakterystykę analizowanego zbioru danych. Szczegółowo omawia sposób klasyfikacji badanych mikroorganizmów oraz ich przyporządkowania do poszczególnych grup ekologicznych. Następnie w zwięzły sposób przytacza podstawowe definicje i fakty dotyczące samoorganizujących się sieci neuronowych typu Kohonena oraz zasadnicze elementy użytej w rozprawie sieci Kohonena. Rozdział 3 kończy się omówieniem zagadnienia wyznaczenia, przy użyciu algorytmu ewolucyjnego, zbioru aminokwasów najlepiej dyskryminującego badane proteomy organizmów o danej cesze ekologicznej od proteomów organizmów, które tej cechy nie posiadają.

Rozdział 4 w całości poświecony jest szczegółowemu omówieniu, otrzymanych przy użuciu omówionej w rozdziale 3 sieci neuronowej, wyników dotyczących związku między cechami ekologicznymi badanych organizmów, a składem aminokwasowym ich proteomów. Pozostałą część rozdziału 4 autor poświęcił omówieniu otrzymanych przy użyciu algorytmu ewolucyjnego zbiorów aminokwasów najlepiej rozróżniających proteomy gatunków przypisanych do ustalonej kategorii środowiskowej, a pozostałymi. Autor w oddzielnych podrozdziałach specyfikuje zbiory aminokwasów dające największą odległość w przypadku hipertermofili, tlenowców, beztlenowców, psychrofili, wewnątrzkomórkowców oraz halofili.

Rozdział 5 zawiera porównanie otrzymanych w rozprawie wyników dotyczących związku składu aminokwasowego z cechami ekologicznymi z wynikami otrzymanymi przez innych autorów oraz stawianymi w literaturze hipotezami. Otrzymane przez autora wyniki sugerują między innymi, że hipotezy o unikaniu przez tlenowce aminokwasów wrażliwych na utlenianie oraz o unikaniu przez beztlenowce aminokwasów bogatych w tle są nieprawdziwe natomiast hipoteza o preferowaniu przez mikroorganizmy zimnolubne aminokwasów bez ładunku i unikaniu aminokwasów aromatycznych i zasadowych jest tylko częściowo prawdziwa. Wyniki autora potwierdzają natomiast doniesienia innych autorów, mówiące o najsilniejszym związku składu aminokwasowego z temperaturą, co powoduje zdecydowaną różnicę w używalności

aminakwasów pomiędzy hypertermofilami, a pozostałymi mikroorganizmami. Przeprowadzone przez autora badania wyraźnie sugerują, że różnic w składzie aminokwasów proteomów mikroorganizmów nie można wyjaśnić jedynie ich pokrewieństwem ewolucyjnym i presją mutacyjną, gdyż istotnym czynnikiem okazuje się być presja selekcyjna. Autor podaje argumenty świadczące o tym, że, wyeliminowanie doboru naturalnego jako czynnika kształtującego używalność aminokwasów w białkach prowadzi do nieobserwowanych w naturze wyników.

D. Ocena głównych wyników rozprawy.

Na sztuczne sieci neuronowe można patrzeć jak na statystyczny model o specjalnej architekturze umożliwiający wyszukiwanie, rozpoznawanie oraz klasyfikację wzorców. W obszarze bioinformatyki sieci neuronowe wykorzystywane były początkowo do przewidywania genów. Autor w swojej rozprawie użył samoorganizującej sieci Kohonena do zagadnień dyskryminacyjnych zwiazanych z proteomiką. Klasteryzacja przeprowadzona przy wykorzystaniu sieci Kohonena jest w gruncie rzeczy podobna do klasycznej metody $k-$średnich. Do rozpoznawania wzorców wykorzystuje się sieć neuronową. Rozpoczyna się od zdefiniowania pewnej liczby węzłów. Na początku losowo przypisuje się pomiary do poszczególnych węzłów, po czym wylicza się odległości pomiędzy wejściowymi pomiarami, a centroidami. Następnie sukcesywnie dopasowuje się pomiary do węzłów i ponownie wylicza się ich odległości od centroidów. Po pewnej liczbie iteracji zwykle osiąga się stabilny układ klastrów o minimalnych odległościach pomiarów od centroidów. Ponieważ, w przeciwieństwie do metody $k-$średnich, w wypadku sieci Kohonena przy wyznaczaniu centroidu bierze się pod uwagę nie tylko informacje z obrębu danego klastra, ale również z klastrów sąsiednich, kalsteryzacja metodą sieci Kohonene lepiej radzi sobie z szumem w danych. Dodatkowo w tej metodzie ostateczna liczba klastrów może być mniejsza niż początkowa liczba węzłów, co sprawia, ze metoda sieci Kohonena jest metodą mniej subiektywną niż metoda $k-$średnich. Biorąc pod uwagę wymienione własności, autor zdecydował się użyć sieci Kohonena jako narzędzie służące do uzyskania odpowiedzi na pytania dotyczące związku cech ekologicznych ze składem analizowanych aminokwasowym analizowanych w rozprawie proteomów mikroorganizmów. Poruszana w rozprawie tematyka jest żywa, szeroka i intensywnie badana. Autor w efektywny i twórczy sposób wykorzystał sieci Kohonena do odpowiedzi na postawione w rozprawie pytania naukowe. Uzyskane przez autora wyniki pozwalają zweryfikować wysuwane w literaturze hipotezy oraz wysłowić kilka nowych w miarę dobrze umotywowanych hipotez. Na dodatkową uwagę zasługuje podjęta przez autora próba weryfikacji statystycznej rezultatów otrzymanych przy użyciu sieci Kohonena. Zwykle samoorganizujące się mapy Kohonena używane są do graficznej wizualizacji danych i wstępnej obróbki wielowymiarowych danych. Autor w swojej rozprawie idzie o krok dalej, dodatkowo weryfikując statystyczną istotność dokonanej przy użyciu sieci klasteryzacji. Następnie, korzystając z algorytmu ewolucyjnego i odpowiedniej modyfikacji sieci, autor dokonuje wyboru aminokwasów najlepiej dyskryminujących badane proteomy ze względu na cechy środowiskowe.

Autor w rozdziale 6 w systematyczny sposób wylicza uzyskane w rozprawie wyniki, dlatego poprzestanę jedynie na wskazaniu rezultatów rozprawy, które moim zdaniem zasługują na szczególna uwagę.

(a) Uzyskane w rozprawie wyniki pokazują, że sieci neuronowe mogą być przydatne w rozpatrywaniu niektórych zagadnień związanych z proteomiką.

(b) Autorowi udało się, używając sieci Kohonena oraz obszerniejszego niż dotychczas zbioru danych, zweryfikować obecne w literaturze naukowej przedmiotu hipotezy dotyczące związku składu aminokwasowego z cechami ekologicznymi.

(c) Dokonana przez autora, przy użyciu sieci Kohonena, klasyfikacja sekwencji aminokwasowych potencjalnych genów kodujących białka oraz wynikające z tej klasyfikacji wnioski stanowią dobry materiał do dalszych obiecujących badań.

(d) Za wartościowy element rozprawy należy uznać próbę statystycznej weryfikacji otrzymanych przy użyciu sieci neuronowych wniosków, dotyczących

(e) Za dodatkową wartość rozprawy uważam sporządzone, na podstawie danych literaturowych, zestawienia analizowanych w rozprawie organizmów ze względu na warunki środowiskowe, w jakich żyją. związku składu aminokwasowego z poszczególnymi cechami ekologicznymi.

Po zapoznaniu się z rozprawą mogę stwierdzić, że jej autor swobodnie porusza się po tematyce sztucznych sieci neuronowych typu Kohonena, umie w poprawny sposób wykorzystać metody statystyczne do analizy danych empirycznych, zna i umie korzystać z pakietu statystycznego R. Dodatkowo autor potrafi przedstawić wyniki swoich badań oraz zagadnienia genetyki molekularnej w sposób zrozumiały dla niespecjalisty.

E. Uwagi szczegółowe.

1. **str. 38**, formaly opis przebiegu modyfikacji węzłów sieci w kolejnych iteracjach zawarty we wzórze (3) jest niepoprawny

2. **str. 39**, wzór (5) proponuję zapisać w postaci $h(BMU, k) = \exp\left(-\frac{l(\mathbf{x}_k, \mathbf{x}_{BMU})}{2\sigma^2}\right)$ lub $h_k^{BMU}(r) = \exp\left(-\frac{r^2}{2\sigma^2}\right)$

3. **str. 41, wiersz 12** zamiast $g_1 \times g_2 \in [2, 3, 4, 5, 6, 7, 8, 9]$ powinno być $g_1,\ g_2 \in [2, 3, 4, 5, 6, 7, 8, 9]$

4. **str. 43**, w opisie rysunku 4 zamiast $g_1 \times g_2 \in [2, 3, 4, 5, 6, 7, 8, 9]$ powinno być $g_1,\ g_2 \in [2, 3, 4, 5, 6, 7, 8, 9]$ lub $g_1 \times g_2 \in [4, ..., 400]$

5. **str. 44**, proponuję dodać kilka słów komentarza, dotyczącego zaproponowanej we wzorze (10) odległości pomiędzy dwoma grupami białek, tym bardziej, że na podstawie tej odległości autor wyciąga w dalszej części rozprawy wiele wniosków

6. **str. 45**, po wierszu 18 proponuję dodać kilka słów komentarza dotyczącego zaproponowanych wartości parametrów oraz zależności między wielkościami wybieranych parametrów

7. **str. 54**, używając zamiast miary Hamanna zdefiniowanej we wzorze (16) statystyki χ^2 do testowania hipotezy o niezależności pojawiania się par aminokwasów w zbiorze chromosomów zwycięskich można wykorzystać test χ^2-Pearsona.

8. **str. 56, wiersz 7**, zamiast słowa "własnoręcznie" zgrabniej jest użyć "samodzielnie" lub "osobiście"

9. **str. 57, wiersz 12**, zapis $p = 0,999$ jest bardziej przejrzysty i zgodny z przyjętą w rozprawie konwencją zapisu liczb zmiennoprzecinkowych

10. **str. 79**, na rysunku 19 pojawia się wyraźna grupa organizmów o stosunkowo dużej wartości odległości d, wyrażającej różnice w składzie aminokwasowym białek, przy małej liczbie podstawień. Powstaje naturalne pytanie, czy jest to efekt czysto losowy, czy też ma swoje biologiczne podłoże. Dodatkowo przedstawiona na rysunku 19 zależności pomiędzy liczbą podstawień nukleotydów, a odległością d jest na tyle regularna, że warto bardziej dokładnie zbadać charakter tej zależność.

11. **str. 126, wiersz 7**, biorąc pod uwagę charakter analizowanych danych bardziej właściwym wydaje się do testowania hipotezy o normalności skorzystanie z testu Shapiro-Wilka, zamiast użytego w rozprawie testu Andersona-Darlinga

F. Konkluzja.

Uważam, że rozprawa Pana magistra Macieja Sobczyńskiego spełnia wszystkie wymagania stawiane ustawowo rozprawie doktorskiej. Rozprawę oceniam pozytywnie i wnoszę o dopuszczenie Pana magistra Macieja Sobczyńskiego do dalszych etapów przewodu doktorskiego.