dr hab. inż. Piotr Formanowicz, prof. PP
Instytut Informatyki
Politechnika Poznańska
ul. Piotrowo 2, 60-965 Poznań

Poznań, 2015.03.15

RECENZJA
rozprawy doktorskiej mgr. Indrajita Saha pt.
„Knowledge Discovery in Biological Data"

Promotor: dr hab. Dariusz Plewczyński

I. Problematyka naukowa oraz przedmiot rozprawy

Bioinformatyka jest bardzo dynamicznie rozwijającą się od połowy lat 90. XX wieku, na pograniczu biologii, informatyki oraz matematyki, gałęzią nauki. Rozwój ten zapoczątkowany został m. in. przez pojawienie się nowych technik umożliwiających odczytywanie sekwencji biologicznych (tj. sekwencji nukleotydowych i aminokwasowych) w znacznie bardziej efektywny sposób niż to miało miejsce kiedykolwiek wcześniej. Szczególne znaczenie miało związane z nim powstawanie i realizacja projektów sekwencjonowania genomów kolejnych organizmów, z genomem człowieka na czele. Pojawienie się rosnących ilości danych biologicznych (w szczególności sekwencji DNA) spowodowało, iż coraz bardziej jasne stawało się, że świat ożywiony, wbrew temu, co mogłoby się wcześniej wydawać, rządzony jest pewnymi ścisłymi regułami oraz że jednym (być może podstawowym) ze sposobów poznania tych reguł jest analiza informacji zawartych w sekwencjach kwasów nukleinowych i białek. Od początku jasne było też, że ani poznanie tych reguł, ani skuteczna analiza wspomnianych sekwencji nie będą w praktyce możliwe bez zastosowania odpowiednich metod matematycznych oraz opartych na nich algorytmów. Stąd, nastąpił gwałtowny rozwój badań w dziedzinach określanych jako biologia obliczeniowa oraz bioinformatyka, których celem jest badanie obiektów i zjawisk biologicznych za pomocą metod i narzędzi matematycznych oraz algorytmicznych, przy czym badania prowadzone na ich gruncie nie ograniczają się do stosowania znanych już metod, ale koncentrują się w dużym stopniu (zwłaszcza w przypadku biologii obliczeniowej) na opracowaniu nowych, lepiej odpowiadających naturze analizowanych zjawisk biologicznych, metod. Bardzo istotnym nurtem badawczym biologii obliczeniowej i bioinformatyki jest analiza dużych zbiorów danych biologicznych. Pojawienie się w naukach biologicznych technik wysokoprzepustowych postawiło szeroko rozumianą dziedzinę analizy danych przed nowymi wyzwaniami związanymi z jednej strony z rzadko spotykanymi wcześniej ilościami danych, a z drugiej z ich naturą oraz stopniem złożoności zapisanej w nich informacji. Stąd, podejmując tego rodzaju wyzwania trzeba się liczyć z możliwością, że znane wcześniej metody okażą się niewystarczająco skuteczne w obliczu dużych zbiorów danych biologicznych i konieczne będzie opracowanie nowych podejść.

Tego rodzaju wyzwanie podjął w swojej rozprawie doktorskiej mgr Indrajit Saha, który w prowadzonych przez siebie badaniach koncentruje się na wykorzystaniu metod statystycznej analizy danych do rozwiązywania szerokiego spektrum problemów biologicznych.

II. Analiza treści rozprawy oraz uzyskanych wyników

1. Treść rozprawy

Rozprawa doktorska mgr. Indrajita Saha składa się z ośmiu rozdziałów, obszernego spisu literatury, streszczenia oraz spisów rysunków i tabel. Doktorant nie formułuje w swojej rozprawie tezy, która byłaby następnie dowodzona, ani celu rozprawy, natomiast w rozdziale pierwszym w wyczerpujący sposób określa jej zakres, który obejmuje opracowanie i zastosowanie metod analizy skupień oraz klasyfikacji do rozwiązywania wybranych problemów biologii oraz diagnostyki medycznej. Oprócz określenia zakresu rozprawy w rozdziale pierwszym znalazło się też ogólne wprowadzenie do jej tematyki, bardzo zwięzłe omówienie wybranych zagadnień biologicznych oraz przegląd metod stosowanych do eksploracji danych.

Rozdział drugi dotyczy metody analizy danych mikromacierzowych opartej na ulepszonym algorytmie ewolucji różnicowej. Po krótkim wprowadzeniu bardzo zwięźle przedstawiono w nim postać danych mikromacierzowych oraz omówiono algorytm ewolucji różnicowej, a także opracowaną jego modyfikację. W dalszej części rozdziału omawiana jest zaproponowana przez Autora metoda analizy skupień oparta na ulepszonym algorytmie ewolucji różnicowej. Omówione też zostało połączenie wspomnianej metody z metodą maszyn wektorów wspierających. Rozdział kończy się przedstawieniem i omówieniem wyników eksperymentu obliczeniowego, w którym zaproponowane metody zostały wykorzystane, wraz z innymi algorytmami, do analizy kilku zbiorów danych mikromacierzowych oraz krótkim podsumowaniem.

Rozdział trzeci poświęcony jest zastosowaniu zaproponowanej przez Autora metody analizy skupień opartej na wielokryterialnym algorytmie ewolucji różnicowej do analizy obrazów mózgu otrzymanych za pomocą obrazowania metodą rezonansu magnetycznego. W rozdziale tym, po krótkim wprowadzeniu, przedstawione są bardzo zwięźle wybrane podstawowe zagadnienia optymalizacji wielokryterialnej, po czym przedstawiony jest zaproponowany przez Autora algorytm analizy skupień oparty na wielokryterialnym algorytmie ewolucji różnicowej. W dalszej części rozdziału przedstawione są wyniki zastosowania zaproponowanego algorytmu oraz kilku innych metod do analizy obrazów mózgu otrzymanych za pomocą obrazowania metodą rezonansu magnetycznego. Rozdział kończy się krótkim podsumowaniem.

Rozdział czwarty dotyczy analizy skupień danych pochodzących z bazy AAindex, zawierającej liczbowe indeksy charakteryzujące fizykochemiczne i biologiczne własności aminokwasów. W rozdziale tym opisane zostało podejście do analizy tego typu danych polegające na zastosowaniu kilku metod analizy skupień i konstrukcji rozwiązania na podstawie skupień wygenerowanych przez te metody. Podejście takie zastosowano zarówno dla przypadku, w którym liczba skupień znana była z góry, jak i dla przypadku, w którym nie była ona znana. Także i w tym rozdziale przedstawiono wyniki eksperymentu obliczeniowego, w którym testowana była zaproponowana metoda. Rozdział kończy się krótkim podsumowaniem.

W rozdziale piątym omawiana jest oparta na sieciach neuronowych metoda przewidywania miejsc modyfikacji potranslacyjnych białek. Po krótkim wprowadzeniu Autor opisał wspomnianą metodę, a następnie zaprezentowane zostały wyniki eksperymentu obliczeniowego, w którym została ona przetestowana. Krótkie podsumowanie kończy rozdział.

Rozdział szósty poświęcony jest przewidywaniu przyłączania peptydów do białek głównego układu zgodności tkankowej klasy II. Opisana w nim została bardzo zwięźle rola tego rodzaju białek, po czym przedstawione zostało zaproponowane przez Autora oparte na metodzie maszyn wektorów wspierających oraz metodzie analizy głównych składowych podejście do identyfikacji peptydów przyłączających się do wspomnianych białek. Jak każdy z zasadniczych rozdziałów rozprawy, także i ten kończy się prezentacją wyników eksperymentu obliczeniowego i krótkim

podsumowaniem.

Rozdział siódmy dotyczy oddziaływań białko-białko. Omówione w nim zostało zastosowanie pięciu znanych metod uczenia maszynowego do przewidywania tego rodzaju oddziaływań. Po krótkim wprowadzeniu w rozdziale tym zwięźle przedstawione zostały bazy danych, z których skorzystano w celu przeprowadzenia eksperymentu obliczeniowego, sposób przygotowania danych oraz metody uczenia maszynowego, które wykorzystano w eksperymencie. Rozdział kończy się przedstawieniem wyników tego eksperymentu i krótkim podsumowaniem.

Ostatni, ósmy rozdział, stanowi podsumowanie rozprawy, w którym m. in. wskazano kierunki dalszych badań.

2. Najważniejsze wyniki przedstawione w rozprawie

Do najważniejszych wyników przedstawionych w rozprawie zaliczyć można m. in.:

1. Opracowanie ulepszonej wersji algorytmu ewolucji różnicowej.
2. Opracowanie metody analizy skupień opartej na ulepszonym algorytmie ewolucji różnicowej.
3. Zdefiniowanie trzech zbiorów wysokiej jakości indeksów aminokwasowych (HQI8, HQI24, HQI40).
4. Opracowanie systemu AutoMotif Server 4.0 służącego do przewidywania miejsc modyfikacji potranslacyjnych białek.

3. Uwagi merytoryczne i redakcyjne

Rozprawa napisana została starannie. Jej zasadnicza część składa się z rozdziałów od drugiego do siódmego. Każdy z tych rozdziałów, jak to zostało opisane w p. 1, zawiera omówienie metod rozwiązywania pewnego problemu biologicznego, bądź problemu związanego z diagnostyką medyczną. Wszystkie one mają zbliżoną strukturę, bowiem zaczynają się od wprowadzenia, po czym przedstawiona jest dana metoda, wyniki eksperymentu obliczeniowego, a kończą się krótkim podsumowaniem. Taka struktura rozprawy pozytywnie wpływa na jej czytelność. Niestety, choć Autor zastosował właściwą strukturę poszczególnych rozdziałów nie wykorzystał jej w pełni, brakuje bowiem w nich jasnego i wyczerpującego wytłumaczenia na czym polegają rozwiązywane problemy biologiczne. W każdym z tych rozdziałów znalazło się bardzo krótkie opisanie danego problemu (a właściwie tylko jego zasygnalizowanie), ale jest ono dalece niewystarczające do tego, by z jednej strony zrozumieć na czym ten problem polega i jakie znaczenie będzie miało zastosowanie do jego rozwiązania zaproponowanej metody (lub metod), a z drugiej, żaden z tych problemów nie został formalnie zdefiniowany w sposób, który jasno przedstawiałby jaką konkretnie postać mają dane wejściowe i jaką postać będą miały wyniki oraz jaka jest interpretacja biologiczna zarówno danych, jak i wyników. Przykładowo, pisząc o analizie danych mikromacierzowych należałoby jasno omówić jaką mają one postać i w jaki sposób są uzyskiwane, a także dlaczego właściwie przeprowadza się w ich przypadku analizę skupień i dlaczego problem ten nie jest prosty. W rozprawie jest na rys. 2.1 przedstawiona macierz ekspresji genów, ale nie wiadomo, jak należy ją interpretować. Czy kolumny tej macierzy odpowiadają danym uzyskanym z różnych mikromacierzy (prawdopodobnie tak, ale dla czytelnika nie musi to być oczywiste)? Jakie znaczenie ma znakowanie mRNA dwoma kolorami? Czy w tego rodzaju eksperymencie mikromacierzowym, którego wyniki poddawane są analizom za pomocą zaproponowanej przez Doktoranta metody jeden kolor nie byłby wystarczający (albo nawet bardziej odpowiedni)? Z kolei w przypadku analizy obrazów uzyskanych za pomocą metody obrazowania metodą rezonansu magnetycznego natura i postać tych obrazów opisana jest w dziesięciu liniach tekstu i czytelnik jest pozostawiony sam na sam z informacją, że np. „the images are available in three bands: T1-

weighted, T2-weighted and proton density (pd)-weighted" oraz że „the images of the Z planes Z10, Z60 and Z130 are considered". Co to znaczy? Czy są to informacje istotne dla dalej prowadzonych rozważań, czy też informacje te są czysto techniczne i nie mają większego znaczenia (jeżeli tak by było, to dlaczego zostały podane?)? Tego na podstawie lektury rozprawy nie wiadomo. Z kolei w przypadku danych pochodzących z bazy AAindex należałoby podać postać tych danych i dokładniej omówić ich sens, by jasne było jakie będą na nich przeprowadzane operacje i w jakim celu. Podobne przykłady niewystarczającego opisania rozwiązywanych problemów można by podać dla każdego z kolejnych rozdziałów. Należy pamiętać, że rozprawa dotyczy nauk technicznych i dyscypliny informatyka, a zatem jasne sformułowanie rozwiązywanych problemów powinno być podstawą dalej prowadzonych rozważań. Niestety, Autor rozprawy chyba o tym zapomniał, co powoduje, że czytelnik bez odpowiedniego przygotowania biologicznego może czuć się zupełnie zagubiony próbując wyobrazić sobie, jakie konkretnie problemy są rozwiązywane, a czytelnik posiadający takie przygotowanie musi, w miarę swoich możliwości, sam zdefiniować odpowiednie problemy, żeby dokładnie zrozumieć co konkretnie robią zaproponowane metody. Taki sposób napisania rozprawy budzi pewne zdziwienie również dlatego, że jest ona we wspomnianych problemach biologicznych osadzona bardzo głęboko, w tym sensie, że zaproponowane przez mgr. Indrajita Saha metody zdają się być ściśle dedykowane do rozwiązania tych konkretnych problemów – są one zatem bardzo istotnym składnikiem prowadzonych przez niego badań. W tym miejscu nasuwa się też pytanie o to w jak dużym stopniu zaproponowane przez Doktoranta metody są dedykowane do rozwiązywania problemów biologicznych, dla których zostały zaprojektowane, tzn. czy przynajmniej niektóre z nich można by z powodzeniem zastosować również do rozwiązywania innego rodzaju problemów (niekoniecznie biologicznych). Autor o tym w rozprawie nie wspomina, a być może warto byłoby się nad tym zastanowić.

Pewne wątpliwości budzi też zawartość rozdziału siódmego, bowiem Autor nie proponuje w nim żadnej nowej metody, lecz bada przydatność kilku standardowych metod do analizy oddziaływań białko-białko. Nasuwa się pytanie, czy uzyskane w ten sposób wyniki są rzeczywiście na tyle istotne, by umieszczać je w rozprawie doktorskiej dotyczącej informatyki.

Opisany w rozdziale trzecim algorytm MODEFC wymaga określenia z góry liczby skupień. W związku z tym pojawia się pytanie, czy tę liczbę zawsze można określić w przypadku zastosowania wspomnianego algorytmu do analizy obrazów mózgu uzyskanych za pomocą rezonansu magnetycznego?

Ponadto, wiele spośród zaproponowanych w rozprawie metod działa dla określonej z góry liczby iteracji (generacji). Czy nie można by uelastycznić ich w ten sposób, by warunek zatrzymania uzależniony był od ewentualnego uzyskiwania przez dany algorytm poprawy najlepszego znalezionego dotąd rozwiązania?

Można przypuszczać, że Autor włożył sporo wysiłku, by od strony redakcyjnej i językowej rozprawa była napisana poprawnie i w znacznym stopniu osiągnął ten cel, choć nie ustrzegł się pewnej liczby błędów tego rodzaju, które miejscami nieco utrudniają jej lekturę.

4. Podsumowanie

Wspomniane powyżej uwagi merytoryczne, a tym bardziej redakcyjne, nie mają zasadniczego wpływu na jakość przedstawionych w rozprawie wyników naukowych i nie wpływają na ogólną wysoką jej ocenę. Autor zaproponował w niej szereg metod statystycznej analizy danych do rozwiązywania wybranych istotnych problemów biologii oraz diagnostyki medycznej. Doktorant zastosował właściwe dla poruszanej problematyki metody badawcze, a dobór cytowanej literatury nie budzi zastrzeżeń i świadczy o głębokiej wiedzy w zakresie poruszanych przez niego zagadnień.

Można stwierdzić, że rozprawa doktorska mgr. Indrajita Saha stanowi istotny wkład w rozwój szeroko rozumianej bioinformatyki. Warto również podkreślić, że jego osiągnięcia

badawcze zostały już docenione przez środowisko naukowe, bowiem wyniki jego prac zostały opublikowane w renomowanych czasopismach z listy JCR, takich jak *Amino Acids*, *International Journal of Data Mining and Bioinformatics*, *Expert Systems with Applications*, *Fundamenta Informaticae*, *Immunogenetics*, *Molecular BioSystems*.

III. Konkluzja

Rozprawa doktorska mgr. Indrajita Saha zawiera oryginalne i interesujące wyniki naukowe dotyczące statystycznej analizy danych biologicznych będącej istotną gałęzią bioinformatyki. Uważam, że wymagania stawiane rozprawom doktorskim przez Ustawę o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki zostały spełnione. Wnoszę zatem o dopuszczenie wspomnianej rozprawy do publicznej obrony.