Prof. dr hab. inż. Franciszek Seredyński
Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie
Wydział Matematyczno-Przyrodniczy.SNŚ
Instytut Informatyki
f.seredynski@uksw.edu.pl

Warszawa, 18.11.2019

Recenzja rozprawy doktorskiej

mgr inż. Tomasza Prejzendanca

zatytułowanej:

Wykorzystanie kontrolowanych języków naturalnych do modelowania systemów dynamicznych w bioinformatyce

1. Problem badawczy i jego znaczenie

Pojawienie się w latach 70 – 80-tych ubiegłego stulecia nowej dyscypliny nauki określanej jako bioinformatyka było wynikiem rosnącego zainteresowania fundamentalnymi problemami biologii, związanymi w szczególności z procesem odczytu informacji zawartych w genomie żywego organizmu jak też z poszukiwaniem nowych możliwości rozwiązywania tych problemów z użyciem matematycznych i informatycznymi metod i narzędzi. Jedną z takich metod jest modelowanie systemów dynamicznych w biologii, które związane jest z tworzeniem matematycznych modeli zjawisk biologicznych oraz ich badanie z użyciem metod komputerowych, w tym metod symulacji komputerowej. Zagadnienie poszukiwania nowych innowacyjnych metod modelowania i symulacji komputerowej jest przedmiotem recenzowanej rozprawy doktorskiej.

Aktualnie istnieje wiele specjalizowanych metod modelowania systemów dynamicznych w bioinformatyce. Doktorant przeanalizował te metody z punktu widzenia zarówno ich zalet jak też wad. Jedną z głównych wad istniejących narzędzi są ograniczone możliwości korzystania z nich przez biologów ze względu na konieczność znajomości często zaawansowanych elementów wiedzy matematycznej i informatycznej. Spowalnia to w znacznym stopniu proces prowadzenia badań eksperymentalnych i korzystania z zaawansowanych metod matematyczno-informatycznych.

W celu pokonania tych trudności Doktorant zaproponował nowe podejście do tworzenia modeli biologicznych, które redukuje wskazane trudności, a jednocześnie ma zaletę kompatybilności z innymi aktualnie stosowanymi narzędziami. Zaproponowane podejście zakłada opracowanie nowego języka służącego do modelowania biologicznych systemów biologicznych. Zaproponowany język modelowania jest zbliżony do języka naturalnego, co znacznie redukuje potrzeby poznawania zaawansowanych elementów wiedzy matematycznej i umożliwia znacznie bardziej efektywne prowadzenie badań. Wraz z językiem zostało opracowane przez Doktoranta środowisko ułatwiające korzystanie z opracowanego narzędzia informatycznego oraz innych znanych i używanych narzędzi informatycznych.

2. Wkład autora

Doktorant w swojej rozprawie poświęconej opracowaniu innowacyjnej metodologii modelowania uzyskał szereg oryginalnych i znaczących dla współczesnej nauki wyników tworzących logiczną całość. Jako najważniejsze z nich chciałbym wskazać:

- Opracowanie nowego języka ModeLang przenaczonego do modelowania systemów dynamicznych w biologii, mającego charakter kontrolowanego języka naturalnego (Rozdz. 5),
- Eksperymentalna weryfikacja funkcjonalności języka ModeLang stosowanego do modelowania systemów dynamicznych w bioinformatyce opisywanych równaniami różniczkowymi (Rozdz. 6 oraz Rozdz. 7),
- Stworzenie i weryfikacja środowiska umożliwiającego analizę oraz konwersję zapisów w języku ModeLang do popularnego w oprogramowaniu bioinformatycznym formatu SBML (Rozdz. 7),
- Przegląd literaturowy i analiza aktualnie stosowanych formatów reprezentacji modeli biologicznych (Rozdz. 4),
- Stworzenie bazy modeli infekcji wirusowych VirDB (Rozdz. 8).

Uzyskane wyniki mają oryginalny charakter wnoszący nową jakość w obszarze badań dotyczących modelowania systemów dynamicznych w bioinformatyce. Zostały one zauważone i docenione przez międzynarodową społeczność, o czym świadczą publikacje w renomowanych czasopismach indeksowanych w bazie *JCR*, takich jak *Current Bioinformatics* oraz *Comput Math Methods Med.*

Przyglądając się opisom badań zwróciło moją uwagę kilka następujących kwestii, które warte byłyby doprecyzowania:

- Rozwiązywanie równań różniczkowych odbywa się zwykle z użyciem stosownych metod numerycznych zaimplementowanych w postaci specjalizowanych solwerów lub z użyciem odpowiednich bibliotek znajdujących się na wyposażeniu różnych języków programowania. ModeLang został zaimplementowany w C++/Pythonie. W jaki sposób kwestia solwera/biblioteki została rozwiązana w przypadku języka ModeLang ?
- Doktorant zaproponował interfejs dostępu do solwerów/bibliotek równań różniczkowych, umożliwiający traktowanie problemu biologicznego w terminach systemu wieloagentowego i opis tego problemu w postaci reguł. Pytanie jakie się pojawia jest następujące: Co się dzieje dalej ? W jaki sposób reguły/zbiory reguł opisujące problem są mapowane w odpowiedni solwer równań różniczkowych ?
- W jaki sposób odbywało się poszukiwanie docelowego zbioru reguł, w jaki sposób było to weryfikowane i co było podstawą ustalenia, że 10 przyjętych reguł spełnia oczekiwania dotyczące modelowania problemów ?
- Z użyciem 10 reguł można stworzyć bardzo dużą liczbę scenariuszy eksperymentu i tylko kilka z nich będzie odpowiadało badanym modelom. Czy opis danego modelu jest związany ściśle z określoną liczbą określnych reguł ? W jaki sposób jest to kontrolowane ?
- Wieloagentowa intepretacja problemów biologicznych opisywanych równaniami różniczkowymi daje potencjalną możliwość modelowania i rozwiązywania tych problemów w formie rozproszonej, w postaci interakcji autonomicznych agentów, zamiast używania zcentralizowanych metod oferowanych przez solwery równań różniczkowych. Czy Doktorant rozważał możliwość takiego rozwiązania ?

3. Poprawność

Rozprawa doktorska ma charakter teoretyczno-eksperymentalny i jej zasadnicza treść związana jest z językami programowania, modelowaniem matematycznym oraz badaniami eksperymentalnymi. Doktorant przeprowadził niezbędne badania eksperymentalne potwierdzające poprawność głównych treści rozprawy, pozostawiając jednocześnie pewne kwestie niedomknięte.

Wyniki badań eksperymentalnych mających na celu weryfikację opracowanego języka ModeLang zostały przeprowadzone w Rozdz. 6 i 7.

Rozdz. 6 zawiera opis dwóch eksperymentów. Celem pierwszego eksperymentu było sprawdzenie na pewnej liczbie potencjalnych użytkowników języka ModeLang czy język ten umożliwia, po wstępnym przeszkoleniu, w prosty sposób intuicyjne zapisywanie poprawnych modeli. Doktorant opisuje ogólnie plan eksperymentu, w którym uczestniczy grupa użytkowników języka, opisuje jego przebieg oraz przedstawia w Tabeli 5 ogólne statystyki eksperymentu. Doktorant stwierdza, że uzyskane wyniki umożliwiły wniesienie poprawek do języka ModeLang, ale samych wyników nie przedstawia. Niedosyt związany z tym eksperymentem jest spowodowany brakiem bardziej szczegółowej informacji opisowej i statystycznej o kolejnych fazach eksperymentu. Byłoby ciekawe prześledzić jak wyglądały typowe scenariusze rozwiązań problemu przez użytkowników w kolejnych iteracjach procesu tworzenia docelowego modelu, jaka była liczba iteracji niezbędna do uzyskania poprawnego modelu, jakie były typowe błędy i jakie było tempo ich usuwania. Wydaje się również, że po wprowadzeniu poprawek do języka celowe byłoby powtórzenie tego eksperymentu po to, aby przekonać się w jakim stopniu trudności ze stosowaniem języka zostały wyeliminowane.

Celem drugiego eksperymentu opisanego w Rozdz. 6 było sprawdzenie czy po poprawnym zapisie modelu wyniki symulacji będą poprawne. Doktorant nie podaje tu szczegółów eksperymentu jak też ogranicza się do ogólnej konkluzji stwierdzając, że „Eksperymenty pozwoliły potwierdzić poprawność procesu budowania modeli matematycznych z wykorzystaniem języka ModeLang".

W związku z kwestią poprawności budowania modeli matematycznych z wykorzystaniem języka ModeLang wydaje się, że celowe byłoby przedstawienie w pracy dodatkowych wyników. Rozdz. 6 zawiera opis 10 modeli biologicznych w postaci równań różniczkowych używanych do modelowania w języku ModeLang, ale powinien również zawierać ich opisy w postaci zaproponowanych reguł opisu agentowego, jak też przebiegi symulacyjne tych modeli potwierdzające poprawność odpowiadających im zbiorom reguł. Czy takie badania były przeprowadzone ?

Taki pełny eksperyment weryfikacyjny został przeprowadzony tylko dla jednego modelu biologicznego w związku z modelowaniem mechanizmu CRISPR/Cas9 i jego wyniki (łącznie z praktycznie jedynym w pracy wynikiem w formie wykresu) przedstawiono w Rozdz. 7.

Podsumowując, wydaje się, że doktorant ograniczył się w rozprawie w kwestii poprawności do wybranych fragmentarycznych aspektów, mających jednak zasadnicze znaczenie dla stwierdzenia poprawności działania zaproponowanego języka, pozostawiając dalsze szczegółowe badania na przyszłość.

4. Wiedza kandydata

Doktorant w swojej rozprawie wielokrotnie potwierdza swoją głęboką wiedzę w zakresie bioinformatyki, informatyki i matematyki. Cytowana w rozprawie literatura obejmująca 148 pozycji bibliograficznych jest obszerna i wyczerpująca. Pierwsze cztery rozdziały rozprawy mają w dużym stopniu charakter przeglądu literaturowego pod kątem różnych aspektów związanych z tematyką rozprawy.

W Rozdz. 1 Doktorant dokonuje wstępnego literaturowego przeglądu historycznego dotyczącego technik programistycznych stosowanych przy modelowaniu systemów dynamicznych oraz uczenia maszynowego. W Rozdz. 2 prezentuje na podstawie literatury podstawowe pojęcia biologii systemowej, związane z wirusami, bakteriami, mechanizmami odpornościowymi oraz koncepcją ekologii gatunków. W Rozdz. 3 Doktorant przedstawia podstawowe koncepcje matematyki i informatyki, takie jak modelowanie matematyczne, w szczególności z użyciem równań różniczkowych, języki i gramatyki, systemy wieloagentowe oraz chmury obliczeniowe. Rozdz. 4 poświęcony jest w całości przeglądowi formatów reprezentacji modeli biologicznych.

Tak więc, wiedza zaprezentowana przez Doktoranta w rozprawie jest obszerna, wyczerpująca i zasługuje na uznanie.

5. Inne uwagi[1]

Rozprawa posiada logiczną strukturę rozdziałów oraz sekcji, a struktury językowe są precyzyjne i nie budzą zastrzeżeń. Duża liczba rysunków ułatwia zrozumienie przekazywanych treści. Tworzy to razem estetyczną formę i dzięki temu przyswojenie treści rozprawy nie sprawia trudności, mimo tego, że dotyka ona szerokiego spektrum wiedzy obejmującej biologię, matematykę i informatykę.

Pewne zastrzeżenia dotyczą dodatków. Zostały one bardzo skrótowo zasygnalizowane w końcu Rozdz. 1, bez dalszych odniesień w tekście rozprawy. Należałoby je poprzedzić krótkim omówieniem, a prawdopodobnie w przypadku niektórych z nich, np. dodatku C.1.2, celowe byłoby szersze przedstawienie w tekście rozprawy, a w przypadku dodatku D odniesienie się do niego w Rodz. 6.

W tekście dostrzegam drobne językowo-edytorskie usterki:

- Str. 19: „najbardziej optymalne podejścia",
- Str. 23: Po zdaniu „Pełna lista reguł obsługiwana przez język ModeLand jest następująca:" zamiast listy reguł mamy informację o sposobie ich wykorzystania,
- Str. 42: „Praca kończy się czterema dodatkami." – dodatków jest pięć
- Str. 45 – zamiast „Rysunek 1" powinno być raczej „Rysunek 1.1"; to samo dotyczy kolejnych rysunków,
- Str. 89: „Autor…zdefiniował…i dla każdej z nich, wyszukiwałem",
- Str. 94: „składa się z najwyższej liczby",
- Str. 126: „ W efekcie powstał kompletny proces, który obejmował",
- Str. 144: „udało się odnaleźć wiele pomysłów",
- Str. 145: „poprawność naukowa działania infrastruktury"
- Str. 149: „został zaprojektowany pod wpływem doświadczeń",
- Str. 157: „w skutek",
- Str. 170: „Pewne możliwości otwierają się każdego dnia".

[1] Opcjonalnie

6. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późniejszymi zmianami)[2] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego? (wybierz jedną opcję stawiając znak **X**)

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka ?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

Mając na uwadze wskazane wyżej kryteria jak też oryginalność uzyskanych wyników potwierdzona publikacjami w renomowanych czasopismach wnoszę o dopuszczenie rozprawy doktorskiej mgr inż. Tomasza Prejzendanca do dalszych etapów przewodu doktorskiego.



Podpis

[2] http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf