RDITT - wpł. 20.05.2024
M. Skarg

dr hab. inż. Tomasz Gambin, prof. uczelni
Instytut Informatyki
Politechnika Warszawska
ul. Nowowiejska 15/19, 00-665 Warszawa
tomasz.gambin@pw.edu.pl

Warszawa, 13.05.2024

Recenzja osiągnięcia oraz istotnej aktywności naukowej dr. inż. Adama Gudysia w związku z postępowaniem w sprawie nadania stopnia naukowego doktora habilitowanego nauk inżynieryjno-technicznych w dyscyplinie informatyka techniczna i telekomunikacja

Niniejsza recenzja została przygotowana w związku z postępowaniem habilitacyjnym dr. inż. Adama Gudysia prowadzonym przez Rade Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej, w odpowiedzi na pismo przewodniczącego Rady prof. dr hab. inż. Andrzeja Polańskiego. Recenzja będzie zgodna z przepisami ustawy z dn. 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. 2020 r. poz 85. z poźn. zm.).

Informacje ogólne

Adam Gudyś ukończył studia wyższe uzyskując tytuł zawodowy Magistra Inżyniera Informatyki na Wydziale Automatyki, Elektroniki i Informatyki Politechniki Śląskiej w 2009 roku. W roku 2014 kandydat obronił rozprawę doktorską pt. "Serial and parallel algorithms for multiple sequence alignment problem and some of its variants" otrzymując stopień Doktora w dyscyplinie Informatyka. Aktualnie dr inż. Adam Gudyś jest zatrudniony na stanowisku adiunkta na Politechnice Śląskiej oraz na stanowisku młodszego specjalisty-architekta systemów informatycznych w Instytucie Technik Innowacyjnych EMAG działającego w ramach Sieci Badawczej Łukasiewicz. Z informacji zawartych we wniosku nie wynika, żeby kandydat ubiegał się wcześniej o stopień doktora habilitowanego.

Zawartość osiągnięcia naukowego

Osiągnięcie naukowe pt. "Algorytmy analizy sekwencji nukleotydowych i aminokwasowych" obejmuje cykl siedmiu powiązanych tematycznie artykułów opublikowanych w czasopismach z listy JCR. W skład osiągnięcia wchodzą następujące publikacje:

H1) Deorowicz, S., Debudaj-Grabysz, A., **Gudyś, A.** (2016) FAMSA: Fast and accurate multiple sequence alignment of huge protein families. Scientific Reports, 6: 33964.
Punkty MEiN (2016/2023): 40 / 140, dwuletni IF (2016/2022): 4.259 / 4.997

H2) **Gudyś, A.**, Deorowicz, S. (2017) QuickProbs 2: towards rapid construction of high-quality alignments of large protein families. Scientific Reports, 7: 41553. Punkty MEiN (2017/2023): 40 / 140, dwuletni IF (2017/2022): 4.122 / 4.997

H3) Deorowicz, S., Debudaj-Grabysz, A., **Gudyś, A.**, Grabowski, S. (2019) Whisper: read sorting allows robust mapping of DNA sequencing data. Bioinformatics, 35(12): 2043–2050.
Punkty MEiN (2019/2023): 200 / 200, dwuletni IF (2019/2022): 5.610 / 6.931

H4) Deorowicz, S., **Gudyś, A.**, Długosz, M., Kokot, M., Danek, A. (2019) Kmer-db: instant evolutionary distance estimation. Bioinformatics, 35(1): 133–136.
Punkty MEiN (2019/2023): 200 / 200, dwuletni IF (2019/2022): 5.610 / 6.931

H5) Zieleziński, A., Deorowicz, S., **Gudyś, A.** (2022) PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences, Bioinformatics, 38(5): 1447–1449.
Punkty MEiN (2022/2023): 200 / 200, dwuletni IF (2022): 6.931

H6) Kokot, M., **Gudyś, A.**, Li, H., Deorowicz, S. (2022) CoLoRd: Compressing long reads. Nature Methods, 19, 441–444.
Punkty MEiN (2022/2023): 200/200, dwuletni IF (2022): 47.99

H7) Santus, L., Garriga, E., Deorowicz, S., **Gudyś, A.**, Notredame, C. (2023) Towards the accurate alignment of over a million protein sequences: Current state of the art. Current Opinion in Structural Biology, 80:102577.
Punkty MEiN (2023): 140, dwuletni IF (2022): 7.786

Najważniejsze dokonania opisane w artykułach wchodzących w skład osiągnięcia naukowego

Prace wchodzące w skład osiągnięcia naukowego dotyczą obszaru bioinformatyki i w szczególności obejmują zagadnienia związane z: (i) dopasowaniem wielu sekwencji aminokwasowych, (ii) mapowaniem odczytów z sekwencjonowania następnej generacji na genom referencyjny, (iii) metodami analizy danych genomowych niewymagających mapowania oraz (iv) kompresji odczytów z sekwencjonowania trzeciej generacji.

Główne dokonania Habilitanta w tematyce dopasowania wielu sekwencji aminokwasowych stanowi opracowanie nowych algorytmów FAMSA oraz QuickProbs2.

W przypadku pierwszego z algorytmów, zaprezentowanego w publikacji **H1,** udało się wykazać istotną poprawę jakości w zadaniu dopasowania dużych zbiorów sekwencji aminokwasowych w stosunku do istniejących rozwiązań, przy zachowaniu niskiej złożoności pamięciowej i obliczeniowej. Jednocześnie, dzięki przystosowaniu algorytmu do działania na procesorach graficznych udało się zapewnić wysoki poziom zrównoleglenia obliczeń. Dla dużych zbiorów danych poskutkowało to przyspieszeniem obliczeń od kilku(set) krotnym w stosunku do wiodących rozwiązań konkurencyjnych. Habilitant posiada główny wkład w opracowanie i implementację algorytmu budowy drzewa dopasowania, w tym w przygotowanie zrównoleglonych wersji algorytmów: poszukiwania najdłuższej wspólnej podsekwencji oraz grupowania typu single-linkage. Na potrzeby przetestowania działania narzędzia na większych zbiorach sekwencji Habilitant opracował nowy zbiór benchmarkowy extHomFam. Habilitant jest również autorem kolejnej wersji algorytmu, która nie została jeszcze opublikowana, ale wstępne wyniki świadczą o dalszym przyspieszeniu działania budowania drzewa dopasowania. O wysokiej popularności i światowym uznaniu opracowanego narzędzia świadczy 50 cytowań publikacji **H1**, ponad 130 gwiazdek i 26 fork'ów w serwisie github i prawie 32 tysiące pobrań pakietu famsa z Bioconda. Za olbrzymie osiągnięcie można też uznać fakt, że narzędzie to zostało wykorzystane w procesach przygotowania danych w projektach takich jak Pfam i AlphaFold.

Drugi z opracowanych przez Habilitanta algorytmów - QuickProbs2, opisany w pracy **H2**, jest przeznaczony do analizy mniejszych zbiorów danych sekwencji aminokwasowych i zapewnieniu wysokiej jakości dopasowania. W celu poprawy działania istniejących algorytmów Habilitant

opracował i zaimplementował dwie nowe metody, czyli iteracyjną poprawę dopasowania wykorzystującą kolumny oraz mechanizm selektywnej spójności. W pierwszej metodzie Habilitant zmodyfikował tradycyjne podejście do iteracyjnej poprawy dopasowania poprzez zastąpienie losowego podziału dopasowania na dwie części, podziałem uwzględniającym występowanie przerw w wybranej losowo kolumnie. Druga zmodyfikowana przez kandydata metoda miała na celu poprawę działania mechanizmu spójności, który przy znajdowaniu dopasowania pary sekwencji uwzględnia informacje z innych par dopasowań zawierających te sekwencje. Habilitant zaproponował podejście selektywnej spójności, które zawęża zbiór sekwencji, które są brane przy dopasowaniu danej pary do najbardziej informatywnych, czyli położonych relatywnie blisko w drzewie dopasowania. Zastosowanie obydwu metod zapewniło nie tylko poprawę jakości ale również zmniejszenie złożoności obliczeniowej. Wkładem Habilitanta było także wykonanie implementacji algorytmu z wykorzystaniem biblioteki OpenCL co umożliwia uruchomienie narzędzia na procesorach graficznych.

Zwieńczeniem badań nad dopasowaniem zbiorów sekwencji aminokwasowych stanowi artykuł przeglądowy **H7.** Na potrzeby publikacji Habilitant opracował nowe zestawy zbiorów testowych w oparciu o największe rodziny białkowe z repozytorium Pfam, oraz przeprowadził badania eksperymentalne, które umożliwiły kompleksowe porównanie istniejących rozwiązań w kontekście analizy wielkich wolumenów danych sekwencji aminokwasowych.

Kolejny wątek podejmowany w badaniach Habilitanta dotyczy mapowania odczytów z sekwencjonowania na genom referencyjny. Habilitant jest współautorem narzędzi Whisper, zaprezentowanego w pracy **H3**, służącego do mapowania sparowanych odczytów z sekwencjonowania następnej generacji. Opracowany algorytm wykorzystuje strategię polegającą na jednoczesnym sortowaniu sufiksów genomu referencyjnego oraz odczytów sekwencyjnych. Głównym wkładem Habilitanta było opracowanie i implementacja trzeciej fazy algorytmu, której celem jest wskazanie ostatecznych pozycji dla par odczytów. W tej fazie uwzględniana jest oczekiwana odległość pomiędzy odczytami (TLEN). Ponaddto Habilitant zaimplementował zrównoleglona wersje algorytmu wyznaczania odległości edycyjnej i zaproponował sposób wyznaczania wartości jakości mapowania odczytów (MAPQ). Eksperymenty przeprowadzone na danych z projektu Genome in the Bottle wykazały wysoką czułość i precyzje algorytmu Whisper (porównywalną z wiodącymi algorytmami takimi jak BWA-mem) przy jednoczesnym zmniejszeniu czasów obliczeń. Warto nadmienić, że po publikacji **H3**, algorytm był aktywnie rozwijany i jego kolejna wersja - Whisper2 została opublikowana w czasopiśmie SoftwareX. Artykuł **H3** uzyskał 7 cytowań, a repozytorium Github zawiera 24 gwiazdki oraz 4 forki.

Następne dokonanie Habilitanta znajduje zastosowania w obszarze metagenomiki i szybkiego porównywania całych genomów. Opracowane narzędzie Kmer-db, zaprezentowane w pracy **H4,** umożliwia obliczenie dystansu pomiędzy dziesiątkami tysięcy genomów w czasie kilkudziesięciokrotnie mniejszym od najlepszego konkurencyjnego narzędzia (Mash). Aby ograniczyć złożoność obliczeniową i pamięciową przygotowana została dedykowana hierarchiczna struktura, która zawiera informacje o mapowaniu każdego zidentyfikowanego k-meru na listę próbek/genomów w których dany k-mer występuje. Habilitant był odpowiedzialny za projekt i implementację struktury danych i zrównoleglonych algorytmów wykorzystywanych do utworzenia struktury. Tak jak w przypadku poprzednich prac, projekt po publikacji jest aktywnie rozwijany. O popularności narzędzia świadczą 22 cytowania artykułu **H4**, 74 gwiazdki i 16 fork'ów w repozytorium Github oraz ponad 10 tysięcy pobrań pakietu kamer-db z Bioconda.

Podobnie jak **H4**, również praca **H5** lokuje się w obszarze metagenomiki. Celem opisanego w pracy narzędzia PHIST jest identyfikacja bakterii-gospodarzy dla zadanych sekwencji

wirusowych. Narzędzie PHIST wykorzystuje Kmer-db do wyznaczenia liczby wspólnych k-merów między genomami wirusów a genomami bakterii, czyli potencjalnych gospodarzy. Ponadto, algorytm znajduje wszystkie dokładne dopasowania >=k oraz ocenia istotność statystyczną powiązania wirusa do gospodarza. Przeprowadzone eksperymenty wykazały jakość porównywalną do innych narzędzi bazujących na dopasowaniu (takich jak BLASTN) przy kilkuset krotnej redukcji czasu wykonania analizy. Habilitant był odpowiedzialny za przystosowanie narzędzia Kmer-db do realizacji kroków algorytmu PHIST oraz opracowanie algorytmu poszukującego wszystkie dokładne dopasowania o długości >=k. Mimo, że artykuł ukazał się dopiero w 2022 roku posiada on już 26 cytowań, a liczba gwiazdek w serwisie github wynosi 24. Jednocześnie pakiet phist został pobrany ponad 1200 razy z Bioconda.

Ostatnie przedstawione we wniosku zagadnienie badawcze, którym zajmował się Habilitant dotyczy problemu kompresji danych z sekwencjonowania trzeciej generacji. Główny pomysł, opisany w pracy **H6**, polegał na zastosowaniu kompresji różnicowej wykorzystującej podobieństwo na poziomie k-merów do określenia sąsiadujących odczytów i użycia go do zmniejszenia rozmiaru strumienia sekwencji DNA. Ponadto, została zaproponowana kompresja stratna strumienia poziomów jakości nukleotydów, która nie wpływa na czułość i precyzję w późniejszym procesie identyfikacji wariantów. Co ważne zaproponowana metoda nie wymaga dostępu do genomu referencyjnego. Habilitant brał udział w opracowaniu ogólnej koncepcji algorytmu realizującego kompresję odczytów oraz przeprowadził analizę eksperymentalną opracowanego algorytmu. Uzyskane wyniki wykazały istotną przewagę algorytmu CoLoRd w poziomie kompresji w stosunku do algorytmów nie wymagających dostępu do referencji (takich jak gzip). Co więcej udało się poprawić poziom kompresji w stosunku do algorytmów wykorzystujących referencję takich jak CRAM. Ważną cechą algorytmu jest fakt, że wraz ze wzrostem pokrycia rośnie wydajność poziomu kompresji, co może mieć duże znaczenie w wielkoskalowych projektach wymagających zwiększonej głębokości pokrycia odczytami. O wysokiej randze powyższego osiągnięcia świadczą: opublikowanie pracy w prestiżowym czasopiśmie Nature Methods, współautorstwo prof. Heng'a Li (twórcy m.in. SAMtools i bwa), 22 cytowania, 46 gwiazdek i 11 forków w serwisie Github, oraz prawie 4000 pobrań z Bioconda.

Ocena osiągnięcia naukowego

Dr inż. Adam Gudyś od wielu lat prowadzi konsekwentne badania w zakresie analizy sekwencji aminokwasowych i nukleotydowych. Jego prace koncentrują się na tworzeniu algorytmów i narzędzi informatycznych pozwalających na dokładną, wydajną i skalowalną analizę sekwencji mającą zastosowania w filogenetyce, wykrywaniu wariantów z danych sekwencyjnych, metagenomice. W mojej opinii, do najważniejszego dorobku Habilitanta (wchodzącego w skład przedstawionego osiągnięcia naukowego), który przyczynił się do rozwoju dziedziny (informatyki/bioinformatyki) należy zaliczyć:

- Opracowanie i implementacja nowych, dokładnych, wydajnych i przystosowanych do działania na procesorach graficznych algorytmów dopasowania zbioru sekwencji aminokwasów (FAMSA i QuickProbs2).
- Opracowanie i implementacja trzeciej fazy wysoko-wydajnego algorytmu Whisper służącego do mapowania odczytów na genom referencyjny.
- Opracowanie i implementacja algorytmu Kmer-db pozwalającego na szybkie określenie podobieństwa pomiędzy genomami.
- Dostosowanie narzędzia Kmer-db na potrzeby realizacji zadań w ramach algorytmu PHIST służącego do identyfikacji bakteri gospodarzy dla genomów wirusowych.
- Udział w opracowaniu koncepcji algorytmu CoLoRd oraz jego przetestowanie.

Przedstawiony dorobek świadczy o tym, że Habilitant bardzo sprawnie wykorzystuje istniejące narzędzia informatyczne i twórczo rozwija nowe algorytmy na potrzeby rozwiązywania rzeczywistych problemów bioinformatycznych i biologicznych. Jego prace mają charakter interdyscyplinarny, wymagały od Habilitanta nie tylko sprawnego posługiwania się warsztatem informatycznym ale również zdobycia wiedzy z zakresu biologii molekularnej i metod sekwencjonowania. Podejmowana przez Habilitanta tematyka jest bardzo ważna i ma duże znaczenie praktyczne, zwłaszcza w biologii i medycynie, gdzie coraz większe koszty przechowywania i przetwarzania danych, a także czas obliczeń zaczynają stanowić główną barierę w dalszym rozwoju wielkoskalowych projektów genomicznych. Znaczący wpływ na rozwój dyscypliny potwierdza duża liczba cytowań artykułów wskazanych we wniosku oraz popularność powiązanych z nimi pakietów oprogramowania w serwisach Github i Bioconda.

Należy podkreślić, że artykuły **H1-H7** wchodzące w skład osiągnięcia naukowego w większości zostały opublikowane w wiodących i bardzo dobrych czasopismach bioinformatycznych, w tym w czasopiśmie Nature Methods (1 artykuł, **200** pkt MNiSW), Bioinformatics (3 artykuły, **200** pkt MNiSW), Scientific Reports (2 artykuły, **140** pkt MNiSW), Current Opinion in Structural Biology (1 artykuł, **140** pkt MNiSW), co świadczy o bardzo wysokim poziomie badań prowadzonych przez Habilitanta. Wszystkie prace są pracami współautorskimi. Habilitant jest pierwszym autorem (lub równorzędnym pierwszym autorem) w dwóch publikacjach, oraz w trzech autorem korespondencyjnym, co potwierdza jego wiodącą rolę w powstanie tych publikacji. O istotnym wkładzie Habilitanta w opracowanie oraz implementacje algorytmów potwierdza jego wysoka aktywność i wiodący udział w tworzeniu kodu i jego deponowaniu w serwisie Github. Sumaryczny Impact Factor prac zgłoszonych jako osiągnięcie wynosi 85,563, a liczba cytowań to 80 (Web of Science) i 169 (Google Scholar).

Opis osiągnięcia naukowego w autoreferacie jest bardzo czytelny. Jedynym drobnym błędem, który dostrzegłem jest niewłaściwe wskazanie źródła na Rys 2 (wskazano publikacje **H2** zamiast **H1**).

Podsumowując, opracowane przez Habilitanta narzędzia i publikacje stanowią znaczny wkład w rozwój dyscypliny informatyka techniczna i telekomunikacja. Dorobek Habilitanta uważam za bardzo pokaźny i wartościowy.

Ocena pozostałego dorobku naukowego

W skład "pozostałego dorobku" wchodzi 9 artykułów opublikowanych w czasopismach z listy JCR, z czego pięć przed uzyskaniem stopnia doktora. Ponadto, do "pozostałego dorobku" Habilitanta należy zaliczyć 9 rozdziałów w monografiach. Duża liczba publikacji oraz inne wskaźniki, takie jak indeks Hirscha (8 wg. Web of Science, 11 wg. Google Scholar), całkowita liczba cytowań uzyskanych przed (117 wg. Web of Science, 191 wg. Google Scholar) i po uzyskaniu stopnia doktora (125 wg. Web of Science, 256 wg. Google Scholar) potwierdzają duże zaangażowanie Habilitanta w działalność naukową.

Publikacje wchodzące w skład pozostałego dorobku dotyczą wykorzystania metod uczenia maszynowego w analizach dużych zbiorów danych wielowymiarowych w tym zbiorów danych bioinformatycznych. Podobnie jak w przypadku publikacji wchodzących w skład osiągnięcia artykułom z pozostałego dorobku towarzyszą implementacje narzędzi zamieszczone w serwisie Github (np. https://github.com/adaa-polsl/RuleKit). Warto zwrócić uwagę na dbałość Habilitanta na aspekt reprodukowalności wyników, dostarczanie skryptów CI/CD, danych testowych oraz dokładnych instrukcji uruchomienia.

Dorobek był Habilitanta był czterokrotnie nagradzany na uczelni w ramach nagród Rektora oraz nagród za wysoko-punktowane publikacje w latach 2018-2023.

Podsumowując, pozostały dorobek naukowy dr. inż. Adama Gudysia oceniam bardzo pozytywnie.

Ocena osiągnięć naukowo-organizacyjnych i dydaktycznych

Habilitant brał udział w 13 krajowych projektach badawczych NCN i NCBiR, w tym w trzech w charakterze kierownika projektu. Jest on również aktywnym recenzentem w najbardziej prestiżowych czasopismach bioinformatycznych takich jak Bioinformatics, Nucleic Acids Research oraz na konferencjach międzynarodowych. Habilitant był również członkiem 3 komitetów konferencji ICMMI. Liczba wystąpień konferencyjnych po uzyskaniu stopnia doktora jest stosunkowo niewielka (5 wystąpień), jednak były to wystąpienia w większości na prestiżowych konferencjach bioinformatycznych (ISMB/ECCB oraz The Biology of Genomes w Cold Spring Harbor). Habilitant posiada duże doświadczenie dydaktyczne. W latach 2010-2023 prowadził zajęcia wykładowe, projektowe i laboratoryjne na przedmiotach związanych z programowaniem, bioinformatyką i sztuczną inteligencją.

Podsumowując, dorobek naukowo-organizacyjny i dydaktyczny dr. inż. Adama Gudysia oceniam pozytywnie.

Ocena aktywności naukowej kandydata realizowanej w więcej niż jednej uczelni, instytucji naukowej lub instytucji kultury, w szczególności zagranicznej

Poza Politechniką Śląską kandydat prowadził prace naukowe w innych ośrodkach badawczych w tym w Polsko Japońskiej Akademii Technik Komputerowych (umowa o dzieło), Sieci Badawczej Łukasiewicz (umowa o dzieło oraz umowa o pracę), a także w dwóch instytucjach zagranicznych: Centre for Genomic Regulation w Hiszpanii (dwa staże w latach 2020 i 2022) oraz Stanford University (umowę o dzieło). W ramach tej aktywności powstały dzieła wchodzące w skład głównego osiągnięcia (publikacja H7 przygotowana we współpracy z ośrodkiem w Barcelonie) a także publikacje z "pozostałego dorobku" kandydata.

Podsumowując, kandydat spełnia kryterium wykazywania się istotną aktywnością naukową w więcej niż jednej uczelni.

Wniosek końcowy

Habilitant posiada bogaty dorobek publikacyjny, organizacyjny i dydaktyczny. O jego znaczącej pozycji w środowisku naukowym świadczą m.in. wysoko-cytowane artykuły, opracowane przez niego narzędzia cieszące się dużą popularnością innych badaczy.

W mojej opinii recenzowane osiągnięcie naukowe "Algorytmy analizy sekwencji nukleotydowych i aminokwasowych" oraz przedłożony do opinii dorobek naukowy, dydaktyczny i organizacyjny dr. inż. Adama Gudysia spełniają wszelkie wymagania stawiane przy nadawaniu stopnia doktora habilitowanego i stawiam wniosek o dopuszczenie Habilitanta do dalszych etapów postępowania habilitacyjnego.

Tomasz Gambin