wpł. 24.04.2024 r.

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

Poznań, 22 kwietnia 2024 r.

RECENZJA
osiągnięcia naukowego pt. „Algorytmy analizy sekwencji nukleotydowych i aminokwasowych" stanowiącego cykl publikacji oraz pozostałego dorobku naukowego, dydaktycznego, organizacyjnego oraz w zakresie popularyzacji nauki dr. inż. Adama Gudysia.

Niniejsza recenzja została przygotowana w odpowiedzi na pismo RDITT.530.4.2024 z dnia 19 lutego 2024 r. prof. dr. hab. inż. Andrzeja Polańskiego, Przewodniczącego Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej, w związku z powołaniem mnie na recenzenta w postępowaniu habilitacyjnym dr. inż. Adama Gudysia.

1. Opinia o osiągnięciu naukowym

Przedstawiony przez dr. inż. Adama Gudysia cykl publikacji będący habilitacyjnym osiągnięciem naukowym zatytułowanym „Algorytmy analizy sekwencji nukleotydowych i aminokwasowych" składa się z następujących siedmiu artykułów:

1. S. Deorowicz, A. Debudaj-Grabysz, A. Gudyś. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Scientific Reports*, 2016, 6, 33964.

2. A. Gudyś, S. Deorowicz. QuickProbs 2: towards rapid construction of high-quality alignments of large protein families. *Scientific Reports*, 2017, 7, 41553.

3. S. Deorowicz, A. Debudaj-Grabysz, A. Gudyś, S. Grabowski. Whisper: read sorting allows robust mapping of DNA sequencing data. *Bioinformatics*, 2019, 35, 2043–2050.

4. S. Deorowicz, A. Gudyś, M. Długosz, M. Kokot, A. Danek. Kmer-db: instant evolutionary distance estimation. *Bioinformatics*, 2019, 35, 133–136.

5. A. Zieleziński, S. Deorowicz, A. Gudyś. PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences, *Bioinformatics*, 2022, 38, 1447–1449.

6. M. Kokot, A. Gudyś, H. Li, S. Deorowicz. CoLoRd: Compressing long reads. *Nature Methods*, 2022, 19, 441–444.

7. L. Santus, E. Garriga, S. Deorowicz, A. Gudyś, C. Notredame. Towards the accurate alignment of over a million protein sequences: Current state of the art. *Current Opinion in Structural Biology*, 2023, 80, 102577.

Problemy związane z analizą sekwencji nukleotydowych i aminokwasowych są jednymi z centralnych zagadnień biologii obliczeniowej i bioinformatyki. Pojawienie się dużych zbiorów tego typu sekwencji, które było związane z szybkim rozwojem metod sekwencjonowania w połowie lat 90. XX wieku było też bezpośrednią przyczyną dynamicznego rozwoju obu wspomnianych dziedzin.

Poznanie sekwencji DNA jest często niezbędnym punktem wyjściowym dalszych analiz prowadzonych na różnych obszarach nauk biologicznych. Jest to zrozumiałe zważywszy na fakt, że sekwencje te zawierają informacje, które w dużym stopniu określają budowę i funkcjonowanie organizmów. Mimo iż samo odczytanie sekwencji DNA niewiele mówi o zawartej w niej informacji, to jest ono niezbędne dla dalszych analiz zmierzających do zrozumienia tej informacji. Pomimo iż, jak zostało wspomniane wcześniej, odczytanie sekwencji DNA jest „tylko" punktem wyjściowym do prowadzenia szerszych analiz, to sam proces sekwencjonowania stanowi duże wyzwanie. Jest tak dlatego, że dane pochodzące z sekwenatorów nie mają postaci sekwencji docelowej, którą należy odczytać, lecz zbioru bardzo dużej liczby krótkich sekwencji, na podstawie których tę sekwencję docelową należy skonstruować lub dopasować je do sekwencji referencyjnej. Oczywiście, w tym celu należy zastosować odpowiednie algorytmy. Problemy, które za ich pomocą są rozwiązywane nie są łatwe z co najmniej dwóch powodów, tj. niedoskonałości danych wejściowych oraz dużych ich rozmiarów. Ze względu na ten ostatni czynnik bardzo duże znaczenie ma złożoność obliczeniowa algorytmów.

Innym spośród podstawowych zagadnień dotyczących analizy sekwencji biologicznych jest problem dopasowania. Dotyczy on zarówno sekwencji nukleotydowych, jak i aminokwasowych. Dopasowanie sekwencji jest jednym ze sposobów określenia ich podobieństwa, a to z kolei jest często podstawą przy określaniu ich funkcji, konstrukcji drzew filogenetycznych, identyfikacji mikroorganizmów oraz przy prowadzeniu wielu innego rodzaju analiz na gruncie biologii molekularnej. Również i w tym przypadku efektywność algorytmów jest bardzo istotna.

Warto też dodać, że ze względu na podstawowy charakter, zarówno problemy sekwencjonowania, jak i problemy dopasowania sekwencji od lat przyciągały i przyciągają uwagę wielu badaczy, a opracowane przez nich algorytmy są bardzo wysokiej jakości.

Tego rodzaju zagadnieniami w swojej pracy badawczej zajął się doktor Adam Gudyś, co doprowadziło do przygotowania przez niego przedstawionego habilitacyjnego osiągnięcia naukowego.

Problemy, których dotyczą publikacje dr. Gudysia składające się na wspomniane osągnięcie, można podzielić na cztery grupy:

1) dopasowywanie wielu sekwencji aminokwasowych (publikacje [1, 2, 7]);
2) mapowanie odczytów sekwencjonowania do genomów referencyjnych (publikacja [3]);
3) metody analizy genomów niewymagające dopasowywania/asemblacji (publikacje [4, 5]);
4) kompresja odczytów sekwencjonowania trzeciej generacji (publikacja [6]).

Jak było wspomniane wcześniej, dopasowanie wielu sekwencji, w szczególności aminokwasowych, jest jednym z podstawowych problemów biologii obliczeniowej i bioinformatyki, a także jest podstawą analiz prowadzonych na różnych obszarach nauk biologicznych. Wraz ze stale rosnącymi rozmiarami zbiorów sekwencji wydajność algorytmów do dopasowywania sekwencji staje się zagadnieniem niezwykle istotnym, a wąskim gardłem wielu stosowanych metod jest konieczność wyznaczenia podobieństwa dla wszystkich par dopasowywanych sekwencji. Stąd opracowywane są metody, w których obliczanie tego podobieństwa nie jest konieczne. Metody te charakteryzują się mniejszą złożonością obliczeniową, ale ich dokładność okazuje się być w wielu przypadkach niewystarczająca.

W pracy [1] z cyklu publikacji będącego osiągnięciem naukowym Habilitanta zaproponowany został algorytm FAMSA, który bazuje na wyznaczaniu podobieństwa między wszystkimi parami sekwencji, a mimo to jest bardzo efektywny czasowo. Efektywność ta została osiągnięta m. in. ze

względu na zastosowanie bitowo-równoległego algorytmu wyznaczania najdłuższej wspólnej podsekwencji oraz wielowątkowości, a także konstrukcji drzewa dopasowania za pomocą algorytmu single-linkage.

Zaproponowany algorytm został przetestowany w eksperymencie obliczeniowym, na potrzeby którego opracowany został zestaw testowy o nazwie extHomFam, zawierający setki tysięcy sekwencji. Przeprowadzone testy wykazały, że zaproponowany algorytm jest wielokrotnie szybszy (o rzędy wielkości) od najlepszych algorytmów do dopasowania wielu sekwencji, dając przy tym wyniki o lepszej jakości.

Z kolei w artykule [2] opisany został algorytm QuickProbs2 (będący udoskonaloną wersją algorytmu zamieszczonego w rozprawie doktorskiej Habilitanta) dla problemu dopasowania wielu sekwencji aminokwasowych dla mniejszych zbiorów danych (tj. nieprzekraczających kilkuset sekwencji) jednak przy zachowaniu maksymalnej dokładności uzyskiwanych wyników. W pracy tej zbadano wpływ liczby dopasowywanych sekwencji na skuteczność dwóch powszechnie stosowanych do poprawy jakości dopasowań podejść, tj. iteracyjnej poprawy dopasowania i spójności, a ponieważ wyniki okazały się być negatywne, opracowane zostały nowe ich odmiany. Ich zastosowanie w zaproponowanym algorytmie spowodowało, że charakteryzuje się on dobrą skalowalnością wraz ze wzrostem liczby sekwencji, a jakość uzyskiwanych wyników jest znacznie lepsza w porównaniu do wyników generowanych przez konkurencyjne algorytmy, zwłaszcza dla większych zbiorów sekwencji. W algorytmie wykorzystano równoległość, co przyczyniło się do uzyskiwania niewielkich czasów obliczeń przy bardzo dobrej jakości wyników.

W pracy [7] omówiony został problem dopasowania wielu sekwencji aminokwasowych w kontekście stale rosnących zbiorów sekwencji podlegających analizie. Przedstawione w nim zostały algorytmy stosowane do rozwiązywania tego problemu, ze wskazaniem ich wad i zalet, omówione zostały możliwe kierunki dalszego rozwoju, a także przeprowadzono testy porównawcze omawianych algorytmów na zaproponowanym w pracy [1] zbiorze extHomFam oraz dwóch nowych zbiorach testowych, przygotowanych na potrzeby omawianej publikacji (zawierające 1,8 i 3,5 miliona sekwencji). W pracy wskazano też najistotniejsze wyzwania dotyczące metod dopasowania wielu sekwencji związane z przewidywanym pojawianiem się coraz większych ich zbiorów.

Kolejnym zagadnieniem, którym zajmował się dr Adam Gudyś, jest mapowanie odczytów do genomów referencyjnych. Jest ono jednym z podstawowych podejść do sekwencjonowania DNA, polegającym na dopasowaniu krótkich sekwencji otrzymywanych z sekwenatora do genomu odniesienia. Ze względu na konieczność wykorzystania tego genomu nie pozwala ono na sekwencjonowanie dowolnego genomu od podstaw, jak ma to miejsce w przypadku sekwencjonowania *de novo*, jednak ze względu na fakt, ze obecnie dostępnych jest już bardzo wiele zsekwencjonowanych genomów różnych organizmów, zwłaszcza człowieka, jest ono niezwykle istotne, gdyż pozwala na wykrywanie różnic między badanym genomem, a genomem odniesienia. Ma to duże znaczenie m. in. w medycynie, ponieważ umożliwia wykrywanie wariantów genów (np. mutacji lub zmian innego rodzaju) mogących być przyczyną różnych chorób.

W tym nurcie badań zaproponowany został algorytm Whisper, opisany w publikacji [3]. Umożliwia on mapowanie odczytów pochodzących z sekwenatorów drugiej generacji do genomu referencyjnego. Algorytm ten wykorzystuje podejście odmienne niż to, które jest stosowane w wielu znanych algorytmach mapowania, tzn. oparty jest na leksykograficznym sortowaniu k-merów genomu referencyjnego i odczytów z sekwenatora oraz łączeniu wyników. W algorytmie tym zostały wykorzystane m. in. wielowątkowość i wektorowość procesora, co przyczyniło się do jego bardzo dobrej wydajności – jest on 6,5-krotnie szybszy od jednych z najlepszych algorytmów służących do mapowania, tj. Bowtie2 i BWA-MEM oraz o ok. 20-30% szybszy od jednych z najszybszych algorytmów, tj. GEM3 oraz Kart, charakteryzując się przy tym zbliżoną precyzją oraz czułością. Algorytm ten był dalej rozwijany, co zaowocowało jego drugą wersją, opisaną w jednej z publikacji Habilitanta, które nie weszły w skład cyklu stanowiącego osiągnięcie naukowe.

Innym nurtem badawczym, którym zajmował się dr Gudyś, są algorytmy analizy genomów, które nie wymagają dopasowania sekwencji. Rozwój tego typu metod jest niezwykle istotny ze względu na szybko rosnące zbiory sekwencji i konieczność ich analizy. Tradycyjne metody analizy sekwencji oparte na ich dopasowaniu są w wielu przypadkach zbyt wolne. Stąd konieczność poszukiwania alternatywnych podejść do problemu analizy genomów. Wśród nich algorytmy oparte na k-merach są szczególnie obiecujące. Podciągi takie (tj. k-mery) mogą być fragmentami zarówno genomów zdeponowanych w bazach danych, jak i fragmentami odczytów z sekwenatorów. Porównywanie sekwencji, wykonywane m. in. na bazie k-merów, może być podstawą analiz, w przypadku których istotne jest podobieństwo między organizmami, np. konstrukcji drzew filogenetycznych czy identyfikacji bakterii.

W ramach prac badawczych, które zaowocowały przygotowaniem habilitacyjnego osiągnięcia naukowego dr Gudyś opracował narzędzie Kmer-db, opisane w publikacji [4]. Jego podstawą, dającą mu przewagę nad narzędziami konkurencyjnymi, jest zastosowanie nowej, efektywnej reprezentacji k-merów w postaci skompresowanej oraz równoległości. We wspomnianej publikacji opisane zostały wyniki eksperymentu obliczeniowego, w którym narzędzie Kmer-db wykorzystane zostało do analizy 40715 genomów bakteryjnych. Możliwe było przeprowadzenie tej analizy w ciągu mniej niż 7 minut, co oznacza, że zaproponowana metoda jest 26 razy szybsza od jej głównego konkurenta, tj. algorytmu Mash. Ponadto, Kmer-db umożliwił analizę wszystkich k-merów z badanych genomów bakteryjnych w czasie, w którym algorytm Mash był w stanie przeanalizować zbiór k-merów o wielkości stanowiącej zaledwie 0,2% wielkości zbioru analizowanego przez Kmer-db. Świadczy to o istotnej przewadze zaproponowanego przez Habilitanta podejścia nad najlepszymi algorytmami w tej dziedzinie.

Z kolei w pracy [5] przedstawione zostało rozszerzenie narzędzia Kmer-db o nazwie PHIST służące do identyfikacji bakterii będących gospodarzami dla bakteriofagów. Znajdowanie gospodarzy dla bakteriofagów oparte jest na spostrzeżeniu, że genomy fagów i ich bakteryjnych gospodarzy często zawierają fragmenty o wysokim wzajemnym podobieństwie. Stąd, wiele metod służących do rozwiązywania tego problemu opartych jest na dopasowaniu sekwencji, ale i w tym przypadku może ono stanowić istotną barierę przy dużych zbiorach sekwencji. Jest to powodem, dla którego konstruowane są również metody, które nie są oparte na dopasowywaniu. Charakteryzują się one większą szybkością, ale z drugiej strony, ich dokładność jest znacznie gorsza niż w przypadku algorytmów opartych na dopasowywaniu sekwencji. Opisane w pracy [5] podejście do identyfikacji bakteryjnych gospodarzy dla fagów oparte jest na wyznaczaniu liczby wspólnych dla fagów i ich potencjalnych gospodarzy k-merów.

Wyniki przeprowadzonych eksperymentów pokazały, że zaproponowane narzędzie znajduje gospodarzy bakteriofagów z dokładnością nieco przewyższającą dokładność metod opartych na dopasowywaniu sekwencji oraz o kilkanaście procent większą niż metody, które nie bazują na dopasowywaniu, przy czym PHIST jest od metod obu rodzajów wielokrotnie szybszy (nawet 300-krotnie szybszy niż metody oparte na dopasowywaniu).

Kolejny problem rozważany przez Habilitanta w ramach badań, których wyniki złożyły się na zaprezentowane osiągnięcie naukowe, jest kompresja danych pochodzących z sekwenatorów trzeciej generacji. Jest to bardzo istotne zagadnienie, gdyż z jednej strony, sekwenatory te generują bardzo duże zbiory danych i minimalizacja kosztów ich przechowywania jest koniecznością, a z drugiej strony, stosowane dotąd algorytmy kompresji danych pochodzących z sekwenatorów tylko w niewielkim stopniu przewyższają algorytmy kompresji ogólnego przeznaczenia. Stąd, istniała potrzeba opracowania znacznie bardziej efektywnej metody kompresji danych sekwencyjnych. Metoda taka, o nazwie CoLoRd została opisana w pracy [6] cyklu publikacji stanowiącego habilitacyjne osiągnięcie naukowe dr. Gudysia. Jest ona przeznaczona do kompresji długich odczytów pochodzących z sekwenatorów trzeciej generacji i jest oparta m. in. na tzw. lekkiej asemblacji oraz zastosowaniu stratnego algorytmu kompresji strumienia jakości. Zaproponowany algorytm wykorzystuje też wielowątkowość. Został on przetestowany w eksperymencie

obliczeniowym, który wykazał jego istotne zalety w porównaniu ze stosowanymi wcześniej metodami, m. in. daje on kilkukrotnie lepszą kompresję niż gzip.

Opisane w przedstawionym cyklu publikacji badania są bardzo interesujące. Opracowane w ich toku algorytmy dają bardzo dobre wyniki w porównaniu do znanych algorytmów, zarówno pod względem czasu obliczeń, jak i jakości uzyskiwanych rozwiązań. Na uwagę zasługuje też wykorzystanie równoległości w zaproponowanych algorytmach. Stanowią one bardzo istotny wkład do biologii obliczeniowej i bioinformatyki, zwłaszcza że dotyczą metod intensywnie wykorzystywanych w obu tych dziedzinach oraz biologii molekularnej. Należy też podkreślić, że wyniki omawianych badań opublikowane zostały w bardzo dobrych czasopismach z listy JCR, tj. *Bioinformatics*, *Nature Methods*, *Scientific Reports* oraz *Current Opinion in Structural Biology* i były wielokrotnie (tj. 80 razy wg bazy Web of Science) cytowane, co świadczy o uznaniu ich dużej wartości przez środowisko naukowe. W przypadku dwóch spośród publikacji stanowiących habilitacyjne osiągnięcie naukowe dr. Gudysia jest on pierwszym autorem. W dwóch artykułach jest autorem do korespondencji. Warto też dodać, że opracowane narzędzia zostały udostępnione w serwisach GitHub i Bioconda, skąd były pobierane tysiące razy.

Podsumowując, należy stwierdzić, że dr inż. Adam Gudyś wniósł bardzo istotny wkład do dyscypliny informatyka techniczna i telekomunikacja, a przedstawione przez niego habilitacyjne osiągnięcie naukowe spełnia wymagania określone w obowiązujących przepisach.

2. Opinia o dorobku naukowym, dydaktycznym, organizacyjnym oraz w zakresie popularyzacji nauki

Doktor Adam Gudyś tytuł zawodowy magistra inżyniera uzyskał na Wydziale Automatyki, Elektroniki i Informatyki Politechniki Śląskiej w 2009 r. Na tym samym wydziale w roku 2014 uzyskał stopień naukowy doktora w dyscyplinie informatyka.

Działalność naukowa dr. inż. Adama Gudysia koncentruje się wokół dwóch grup zagadnień. Pierwszą z nich stanowią zagadnienia związane z konstrukcją algorytmów dla problemów analizy sekwencji nukleotydowych i aminokwasowych (jest to zatem grupa zagadnień, których dotyczy habilitacyjne osiągnięcie naukowe), natomiast do drugiej należą problemy związane z uczeniem maszynowym, a zwłaszcza z konstrukcją i wykorzystaniem modeli regułowych.

Całkowity dorobek naukowy Habilitanta obejmuje 16 artykułów opublikowanych w czaopismach z listy JCR, z czego 5 zostało opublikowanych przed uzyskaniem przez niego stopnia doktora, a 7 składa się na cykl będący przedstawionym osiągnięciem naukowym. Sumaryczny Impact Factor tych publikacji liczony wg roku publikacji wynosi 113,644, z czego 82,308 przypada na przedstawiony cykl, a 19,040 na pozostałe prace opublikowane po uzyskaniu stopnia doktora. Całkowita liczba punktów Ministerstwa (wg listy czasopism punktowanych z 2023 r.) wynosi 2440, z czego 1220 pkt. przypada na publikacje stanowiące osiągnięcie naukowe, a 700 na pozostałe artykuły opublikowane po uzyskaniu stopnia doktora. Wymienione publikacje wg bazy Web of Science były cytowane 242 razy (228 bez autocytowań), z czego 80 cytowań przypada na przedstawiony cykl. Ponadto, przed uzyskaniem stopnia doktora Habilitant opublikował 4 rozdziały w monografiach, a po uzyskaniu stopnia naukowego opublikował ich 5. Doktor Gudyś występował również na konferencjach naukowych – na czterech przed uzyskaniem stopnia doktora i na pięciu po jego uzyskaniu. Według bazy Web of Science indeks Hirscha Habilitanta wynosi 8. Wiele spośród publikacji dr. Gudysia ukazało się w bardzo dobrych czasopismach, m. in. *Bioinformatics*, *Nature Methods*, *Scientific Reports*, *SoftwareX*, *Current Opinion in Structural Biology*, *Knowledge-Based Systems*, *Plant and Cell Physiology*, *BMC Bioinformatics*.

Doktor Gudyś brał udział w realizacji 11 projektów badawczych, z których 7 rozpoczęło się po uzyskaniu przez niego stopnia doktora. W jednym z projektów rozpoczętych przed uzyskaniem stopnia naukowego i w jednym rozpoczętym po uzyskaniu stopnia Habilitant pełnił funkcję kierownika. Ponadto w jednym projekcie rozpoczętym po uzyskani stopnia doktora pełnił funkcję kierownika w

Politechnice Śląskiej. Oprócz wymienionych projektów bierze on aktualnie udział w realizacji dwóch kolejnych projektów. Wszystkie wymienione projekty były lub są finansowane przez NCN lub NCBiR.

W 2020 r. Habilitant odbył trzymiesięczny staż naukowy w Centre for Genomic Regulation w Barcelonie w Hiszpanii. Ponadto, w latach 2011-2017 współpracował z Polsko-Japońską Akademią Technik Komputerowych, a od 2014 r. współpracuje z Siecią Badawczą Łukasiewicz – Instytutem Technik Innowacyjnych EMAG (od 2020 r. jest tam zatrudniony). Doktor Gudyś współpracuje także ze Stanford University.

Doktor Gudyś pełni też funkcję recenzenta – recenzował artykuły m. in. dla czasopism *Bioinformatics*, *Nucleic Acid Research*, *NAR Genomics and Bioinformatics*, *Machine Learning Research*, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *Genes*, *Bulletin of the Polish Academy of Sciences: Technical Sciences* i *International Journal of Parallel Programming* oraz na konferencje *International Conference on Man-Machine Interactions* i *Asian Conference on Intelligent Information and Database Systems*. Był on również członkiem komitetu organizacyjnego trzech edycji konferencji *International Conference on Man-Machine Interactions*. Habilitant należy do *International Society for Computational Biology*.

Doktor Gudyś bierze także aktywny udział w kształceniu studentów. Prowadzi lub prowadził zajęcia z kilkunastu przedmiotów, przy czym w przypadku części z nich opracował cały lub część przedmiotu, a prowadzone przez niego zajęcia to zarówno wykłady, jak i laboratoria i projekty. Habilitant był promotorem 18 prac inżynierskich oraz 9 prac magisterskich. W dostarczonych materiałach nie ma informacji o osiągnięciach dotyczących popularyzacji nauki.

Podsumowując, uważam, że dorobek naukowy, organizacyjny i dydaktyczny dr. inż. Adama Gudysia spełnia wymagania stawiane przez obowiązujące przepisy osobom ubiegającym się o stopień naukowy doktora habilitowanego i świadczy o jego dużej aktywności naukowej, organizacyjnej i dydaktycznej.

3. Wniosek końcowy

Uważam, że habilitacyjne osiągnięcie naukowe oraz dorobek naukowy, dydaktyczny i organizacyjny dr. inż. Adama Gudysia spełniają wymagania dotyczące stopnia naukowego doktora habilitowanego określone w ustawie Prawo o szkolnictwie wyższym i nauce (Dz. U. 2022 poz. 574 z późn. zm.). Wnioskuję zatem o dopuszczenie dr. inż. Adama Gudysia do dalszych etapów postępowania habilitacyjnego.