Recenzja osiągnięcia naukowego Dra inż. Adama Gudysia w związku z ubieganiem się o uzyskanie stopnia doktora habilitowanego w dyscyplinie: Informatyka Techniczna i telekomunikacja.

Tytuł osiągnięcia: Algorytmy analizy sekwencji nukleotydowych i aminokwasowych.

Jako osiągnięcie Kandydat podaje zestawienie 7-miu publikacji (zestawienie H). Temat osiągnięcia – jak podaje to Autor – dotyczy działu bioinformatyka i biologia obliczeniowa, a w szczególności omawia zagadnienia związane z analizą sekwencji molekularnych oraz aminokwasowych w białkach. Określenie „sekwencje molekularne" wydaje się nie oddawać tematyki publikacji. Sekwencje molekularne to np. może być błona, w której odpowiednia sekwencja molekuł dostarcza konstrukcji błony.
Tematem jest opracowanie narzędzi do analizy sekwencji nukleotydowych w materiale genetycznym oraz sekwencji aminokwasowych w układach po-translacyjnych. Potrzeba stałego rozwijania narzędzi zorientowanych na tę tematykę wynika z ciągle rosnących zasobów baz danych w tej dziedzinie.
Tematyka w szczególności może być określona następującymi pozycjami – jak formułuje to Kandydat:
1. Ocena stopnia podobieństwa sekwencji aminokwasowych dla układu wielu ciągów,
2. Identyfikacja odczytów sekwencjonowania w odniesieniu do genomu referencyjnego,
3. Opracowanie innych technik numerycznej oceny genomów nie wymagających dopasowania
4. Techniki kompresji sekwencjonowania trzeciej generacji.

Analiza porównawcza sekwencji nukleotydowej jak i aminokwasowej opiera się na wspólnym mechanizmie poszukiwania sposobu nałożenia ciągów znaków w celu identyfikacji najwyższego, optymalnego stopnia podobieństwa zapisów.
Zagadnienie to sprowadza się do identyfikacji najdłuższych odcinków o identycznej sekwencji, porządkowaniu ujawniającym najwyższy stopień dopasowania przy wprowadzeniu insercji/delecji oraz prezentacji zidentyfikowanej zmienności w formie drzewa ujawniającego rozwój i umożliwiającego śledzenie ewentualnych zmian ewolucyjnych.
Pozycje publikacji z listy H o numerach 1, 2 i 7 dotyczą poszukiwania optymalnej techniki dla identyfikacji nałożenia ciągów znaków dla sekwencji aminokwasowych. Problemem

numerycznym jest tutaj poszukiwanie możliwości nałożenia wielu ciągów sięgających liczby tysięcy czy nawet więcej.

W szczególności wprowadzono techniki oparte na rozszerzeniach wektorowych procesora takich jak SSE czy AVX umożliwiających identyfikację najdłuższych odcinków o identycznych sekwencjach. Ogólna ocena wykorzystana jest do tworzenia drzew genealogicznych. Możliwa staje się też ocena kosztów wprowadzenia insercji/delecji zależna od rozmiaru zbudowanego poprzednio fragmentu dopasowania.

Opracowane narzędzie poddano analizie porównawczej w relacji do innych dostępnych narzędzi dotyczących tego samego zagadnienia.

Udział Kandydata w tych pracach polega na aktywności typowo informatycznej. Narzędzie zaproponowane przez Kandydata charakteryzuje się najniższą konsumpcją czasu obliczeniowego. Serwer umożliwiający obliczenia okazuje się być wykorzystywany w znacznym stopniu przez użytkowników zewnętrznych.

Realizacja tej części osiągnięcia stanowi efekt działalności Kandydata w dwóch projektach badawczych w ramach NCN.

Istotne osiągniecie Kandydata dotyczy udziału w projektach zorientowanych na wielko-skalowe obliczenia wynikające z potrzeby analizy porównawczej sekwencji w skali milionów. Niezmiernie cenne jest też opracowani narzędzi do prezentacji graficznej.

Wprowadzono miary odległości między sekwencjami w oparciu o najdłuższe odcinki o identycznej sekwencji. Tutaj zastosowano technikę rozszerzeń wektorowych procesora o nazwie SSE oraz AVX. Budowę drzewa dopasowania skonstruowano w oparciu o single linkage bazującej na liniowej złożoności pamięciowej.

Cenne jest wprowadzenie miary kosztów insercji/delecji w relacji do długości porównywanych łańcuchów.

Przeprowadzono też analizę porównawczą efektywności nowo opracowanego narzędzia (FAMSA) wobec tradycyjnych narzędzi jak ClustalOmega, MAFFT, MUSCLE czy UPP oraz Kalign2.

Wyniki realizacji wyrażone w pkcie 2 z listy tematów prezentuje publikacja H2, gdzie opracowano modyfikację dotychczasowej wersji programu QuickProbs wprowadzając wersję QuickProbs2. Programy te dotyczą identyfikacji stopnia podobieństwa sekwencji aminokwasowych. Obie wersje tego programu są autorstwa Kandydata. Wielką zaletą zastosowanego algorytmu jest możliwość identyfikacji pod-sekwencji oraz relacji pomiędzy nimi. Stanowi to formę swoistego drzewa zmienności sekwencji. Pytanie

dotyczy zgodności tak uzyskanego drzewa z drzewami filogenetycznymi. Analiza porównawcza z innymi narzędziami dedykowanymi dla tego samego celu wskazuje poprawę efektywności obliczeń programu QuickProbs2 nad MSAProbs czy PicXAA poprzez skrócenie czasu obliczeń.

Praca z indeksem H7 prezentuje przegląd narzędzi dla porównań sekwencji aminokwasów w łańcuchach białek w skali milionowej liczby porównywanych sekwencji. To wyzwanie związane jest z nadchodzącym zapotrzebowaniem dla wielko-skalowych obliczeń tego typu. Rozważana jest możliwość zastosowania algorytmów głębokiego uczenia.

Pozycje z listy publikacji H4 i H5 zawierają wyniki opracowania tematu określonego jako 2 na liście osiągnięć naukowych Kandydata.

Zmiana operowaniem strukturą danych wprowadziła modyfikację programu Kmer-db, co skutkuje zwiększonym wykorzystaniem tego programu przez użytkowników w systemie otwartego dostępu tego oprogramowania.

Praca w pozycji H6 to realizacja zadań wymienionych w pkcie 4 listy osiągnięć.

W podsumowaniu dorobku należy stwierdzić, że dotyczy on rozwoju w dziedzinie BIG DATA. Ta dyscyplina o rosnących wykładniczo zasobach wymaga coraz lepszych szybszych i efektywniejszych narzędzi. Wszystkie narzędzia, w opracowaniu których Kandydat brał udział są dostępne w sieci. Liczba użytkowników jest znacząca.

Aktywność naukowa Kandydata wychodzi poza ramy podane jako osiągnięcie naukowe (punkty 1-4). Wyrażona została w 9-ciu rozdziałach w ramach monografii oraz w 9-ciu artykułach w czasopismach punktowanych z listy JCR.

Tematyka naukowa w ramach przygotowań do uzyskania stopnia naukowego doktora habilitowanego stanowi kontynuację tematyki opracowana w pracy doktorskiej. Oznacza to konsekwentne specjalizowanie się w określonej tematyce.

Liczne udziały w konferencjach zarówno przed uzyskaniem jak i po uzyskaniu stopnia doktora dowodzi aktywnego zaangażowania w wybranej dziedzinie. Trzykrotny udział w organizacji konferencji w roli członka Komitetu Naukowego stawia Kandydata w środowisku specjalistycznym.

Cztery projekty badawcze z rolą wykonawcy ale i też kierownika przed uzyskaniem stopnia doktora oraz siedmiu po uzyskaniu stopnia doktora dowodzi konsekwentnego utrzymywania

aktywności docenionej przez NCN. Cztery projekty przyznane przez Politechnikę Śląską prowadzone przez Kandydata wyraża lokalne uznanie fachowości i przygotowania do realizacji projektów.
Oceną dorobku wyrażona za pomocą indeksu Hirscha to wg Web of Science: 8 oraz wg Google Scholar: 11.
Ocena całościowego dorobku wyrażona za pomocą IF to 123,028 natomiast liczba cytowań to wg Web of Science 228 natomiast wg Google Scholar 433.

Kandydat legitymuje się też potwierdzeniem odbycia stażu zagranicznego w Centre for Genomic Regulation, Barcelona, Hiszpania. Trwający trzy miesiące staż zaowocował opracowaniem prototypowej implementacji protokołu umożliwiającego analizę milionów sekwencji na komputerze osobistym w przeciągu kilku godzin.
W dokumentacji dostępny jest też zestaw dokumentów potwierdzających znaczący udział Kandydata w opracowaniach kolejnych publikacji potwierdzony przez współautorów przedstawionych do oceny publikacji.
Kandydat nie zgłasza pozycji w dziedzinie aktywności dla środowiska.
Jako uwagę krytyczną muszę podać posługiwanie się Kandydata adresem .gmail. Od pełniącego rolę Autora korespondencyjnego oczekuje się adresu wiarygodnego.
Inna uwaga dotyczy określenia „sekwencja białkowa", które jest określeniem niejednoznacznym. Należy użyć określenia sekwencji aminokwasowej. Analogicznie sekwencje nukleotydowe nie są określane jako sekwencje nici DNA tylko sekwencje nukleotydowe.

Podsumowując ocenę dorobku naukowego i zrealizowanego projektu badawczego stanowiącego podstawę do ubiegania się o stopień doktora habilitowanego należy stwierdzić, że aktywność dra inż. Adama Gudysia wpisuje się w aktualne zapotrzebowania na nowe narzędzia informatyczne adresowane do zagadnień biologicznych w tym analiz sekwencji nukleotydowych i aminokwasowych w szczególności. Powiększające się dramatycznie szybko rozmiary baz danych w postaci ogromnych zasobów magazynujących wyniki analiz sekwencji oczekują coraz bardziej sprawnych narzędzi. Aktywność naukowa Kandydata wychodzi naprzeciw tym oczekiwaniom. Współpraca specjalisty w dziedzinie reprezentowanej przez dra inż. Adama Gudysia z zespołami eksperymentatorów o profilu biologicznym przyniesie znaczące osiągnięcia w genomice porównawczej.

Dorobek i osiągnięcie naukowe dra inż. Adama Gudysia oceniam jako w pełni reprezentujące do pełnienia samodzielnej aktywności naukowej w dziedzinie o niezmiernie istotnym znaczeniu dla gwałtownie rozwijających się dyscyplin eksperymentalnych dostarczających cenne dane. Potrzeba dostępu do narzędzi o zaawansowanych możliwościach obliczeniowych umożliwiających określanie ogólnych reguł w dziedzinie zmienności gatunkowej jak i znaczenia pojawienia się mutacji w tym w procesach ewolucyjnych w szczególności. Badania te poza aspektem naukowych mają istotne znaczenie praktyczne w formie analiz medycznych oraz projektowania terapii w tym – zapewne w przyszłości – terapii zindywidualizowanej.

Oceniam dorobek dra inż. Adama Gudysia jako spełniające wymogi stawiane w art. 219 ust. 1. pkt. 2b Ustawy i wnoszę do Rady Dyscypliny Informatyka Techniczna i Telekomunikacja o kontynuację procedury i wdrażanie kolejnych jej etapów w celu nadania stopnia doktora habilitowanego w dziedzinie **nauk inżynieryjno-technicznych** w dyscyplinie: **informatyka techniczna i telekomunikacja**

prof. dr hab. Irena Roterman-Konieczna

Kraków 18. marzec 2024