Dr hab. inż. Marta Szachniuk, prof. nadzw. IChB
Instytut Informatyki
Politechnika Poznańska
mszachniuk@cs.put.poznan.pl

Poznań, 02.07.2018

Recenzja Rozprawy Doktorskiej

Tytuł: Inferring genomic duplication events (Studium zdarzeń duplikacji w genomie)

Autor: mgr Jarosław Paszek

Promotor: dr hab. Paweł Górecki

Tematyka badawcza

Recenzowana rozprawa dotyczy istotnych problemów ewolucyjnej genomiki obliczeniowej, która jest poddziedziną genomiki obliczeniowej - jednej z najszybciej rozwijających się gałęzi bioinformatyki. Nauka o genomach to stosunkowo młody obszar nauk o życiu. Od momentu pojawienia się pierwszych technik sekwencjonowania, a potem sekwencjonowania następnej/nowej generacji (ang. NGS, Next Generation Sequencing) obserwujemy jej dynamiczny rozwój. Eksperymenty sekwencjonowania generują ogromne ilości danych, które przetwarzane są za pomocą efektywnych metod obliczeniowych. Do zadań takich algorytmów bioinformatycznych należą m.in. składanie zsekwencjonowanych fragmentów, indeksowanie i gromadzenie danych o genomach, rozpoznawanie sekwencji kodujących, porównywanie sekwencji, poszukiwanie fragmentów różniących się i podobnych, etc. Modele, metody i algorytmy powstające w odpowiedzi na zapotrzebowanie współczesnej genomiki definiują genomikę obliczeniową. Z kolei śledzenie zmian jakie następowały w genomach oraz odkrywanie zależności między nimi a ewolucją gatunków stanowi podstawowe zagadnienie badawcze genomiki ewolucyjnej. Również tutaj badania pomagające zrozumieć genetyczne podstawy różnic pomiędzy poszczególnymi gatunkami wspomagane są metodami komputerowymi. Metody te m.in. wyszukują zmian w genomach czy analizują relacje między mutacjami w genomie a powstawaniem nowych gatunków organizmów żywych. Na styku genomiki ewolucyjnej oraz informatyki leży ewolucyjna genomika obliczeniowa.

Ewolucja genomów jest wynikiem występowania mutacji, duplikacji sekwencji oraz rearanżacji chromosomowych. Badania wskazują, że jest ona procesem powolnym, przebiegającym w sposób ciągły. Jednak od czasu do czasu w procesie tym występują eksplozje, których wynikiem jest powstanie wielu nowych gatunków (specjacja). Jedna ze współcześnie stawianych hipotez mówi, iż przyczyną takich zjawisk są wielkoskalowe duplikacje, tzn. wielokrotne duplikacje genów lub duplikacje całych genomów (ang. WGD, whole-genome duplications). Ich wynikiem jest strata genów lub neofunkcjonalizacja, w ramach której jedna ze zduplikowanych kopii genu zachowuje dotychczasową funkcję, a druga kopia, w której kumulują się mutacje, zyskuje nową funkcję. Wykrywanie duplikacji genów pomaga odtworzyć ewolucję genomów. Nałożenie (uzgadnianie) drzewa genów na drzewo gatunków umożliwia z kolei zlokalizowanie zdarzeń duplikacji w drzewie gatunków. Dzięki zidentyfikowaniu momentów wystąpienia duplikacji oraz wpisaniu ich w drzewo gatunków możemy zrozumieć w jaki sposób ewoluują rodziny genów oraz całe genomy.

Uzgadnianie drzewa genów z drzewem gatunków jest jednym z wyzwań ewolucyjnej genomiki obliczeniowej. Jego celem jest takie wpisanie drzewa genów w drzewo gatunków, aby zostały zachowane relacje na linii rodzic-potomek a liczba duplikacji genów umożliwiająca nałożenie odpowiadających sobie poddrzew była jak najmniejsza. Proces uzgadniania polega na modelowaniu pojedynczych duplikacji oraz ich dopuszczalnych lokalizacji w drzewie genów. W kolejnym etapie często dokonuje się grupowania podzbiorów pojedynczych duplikacji w zdarzenia duplikacji wielokrotnej. Uzgadnianie można wykorzystać do rozwiązywania różnych klasycznych problemów biologii obliczeniowej, np. rekonsktrucji superdrzew, mapowania gen-genom w metagenomice, ukorzeniania drzewa genów, klastrowania pojedynczych duplikacji. Istniejące podejścia do problemu uzgadniania stosują różne scenariusze ewolucyjne modelowania pojedynczych duplikacji oraz różne reguły grupowania duplikacji. Autor recenzowanej rozprawy doktorskiej podejmuje temat uzgadniania drzew genów z drzewem gatunków, rozpatrując go zarówno w aspekcie modelowania pojedynczych duplikacji jak i interpretowania podzbiorów tych zdarzeń jako duplikacji wielokrotnych.

Zakres pracy i wkład autora

W pracy doktorskiej, mgr J. Paszek zajął się problemem duplikacji genomowych, który w ogólności sformułowano jako problem poszukiwania klastrowania wszystkich pojedynczych duplikacji genów o minimalnym rozmiarze. Autor rozprawy bardzo dokładnie przeanalizował

istniejące scenariusze ewolucyjne i przedstawił modele określające zbiór scenariuszy dopuszczalnych. Zaproponował własny model PG (Paszek-Górecki) dopuszczalnych scenariuszy dla drzewa genów i drzewa gatunków posiadających minimalną liczbę duplikacji genów oraz porównał go z innymi istniejącymi modelami. Model PG został opublikowany w 2016 r. w czasopiśmie *BMC Genomics* (czasopismo w II kwartylu; Impact Factor: 3,729; Punkty MNiSW: 35). W pracy zaprezentowano szereg zagadnień związanych z uzgadnianiem ukorzenionych i nieukorzenionych drzew genów. W szczególności przedstawione zostały problemy REC (Rooted Episode Clustering), RME (Rooted Minimum Episodes), UEC (Unrooted Episode Clustering) oraz UME (Unrooted Minimum Episodes) wraz z proponowanymi dla nich nowymi rozwiązaniami. Jednym z ważniejszych wyników badawczych autora pracy było opracowanie pierwszego liniowego algorytmu rozwiązującego problem REC oraz metod rozwiązujących otwarte problemy UME i UEC. Nowe metody zostały przetestowane na trzech zbiorach danych eksperymentalnych. Eksperymenty wykazały efektywność proponowanych metod we wnioskowaniu duplikacji genomowych. Wyniki badań dla drzew nieukorzenionych zostały opublikowane w czasopiśmie *BMC Genomics* (czasopismo w II kwartylu; Impact Factor: 3,729; Punkty MNiSW: 35) w artykule z roku 2018 oraz w materiałach po konferencji RECOMB-CG'17 (RECOMB Comparative Genomics) opublikowanych w 2017 roku w serii wydawniczej *Lecture Notes in Computer Science*.

Ocena strony merytorycznej

Rozprawa doktorska zawiera 115 stron maszynopisu. Składa się z siedmiu rozdziałów i bibliografii zawierającej 300 pozycji literaturowych. Dodatkowo autor dołączył krótkie streszczenie w języku angielskim i polskim, listę najważniejszych publikacji związanych z tematyką pracy doktorskiej oraz informację o projektach badawczych realizowanych w czasie pracy doktorskiej i źródłach finansowania badań. Rozprawa została napisana w języku angielskim. Tekst opatrzony jest 33 kolorowymi ilustracjami oraz 4 tabelami.

Rozdział 1 wprowadza czytelnika w tematykę, w której ulokowana jest problematyka rozpatrywana w rozprawie doktorskiej. Autor ukazuje historię badań nad ewolucją, początkami życia, zjawiskiem dziedziczenia, ewolucji genomów oraz filogenetyki molekularnej. Zwraca uwagę bardzo szerokie podejście do tematu, a rozległa wiedza i świetne rozeznanie autora w literaturze światowej są imponujące.

Rozdział 2 przedstawia problematykę uzgadniania drzew. Wyjaśniono w nim na czym polega uzgadnianie drzewa genów i drzewa gatunków dla przypadku ukorzenionego i nieukorzenionego. Podane są podstawowe twierdzenia oraz definicje zilustrowane na przykładowych drzewach. Objaśniona zostaje idea scenariusza ewolucyjnego oraz ukazany jest model LCA (Least Common Ancestor) dla ukorzenionych drzew genów. W rozdziale znajdujemy również krótką charakterystykę badań nad rekonstrukcją zdarzeń ewolucyjnych wraz nawiązaniami do literatury światowej, w której badania te były opisywane.

W Rozdziale 3 opisana jest koncepcja duplikacji genów. Wprowadzone zostają podstawowe pojęcia i definicje z zakresu uzgadniania drzew oraz wykrywania duplikacji. Pokazane są istniejące modele uzgadniania drzew i scenariusze ewolucyjne. Autor przedstawia różne podejścia do interpretowania wielokrotnych duplikacji. W rozdziale znajdziemy obszerne studium problemu duplikacji genomowych oraz przegląd istniejących podejść proponowanych do rozwiązania tego problemu. W rozdziale zebrane zostały informacje na temat złożoności obliczeniowej algorytmów dla problemów REC, RME i GD w kontekście różnych modeli scenariuszy oraz klas złożoności tych problemów. Tabela jest jednak niekompletna i niespójna w treści: w niektórych komórkach podana jest złożoność obliczeniowa algorytmu oraz klasa złożoności problemu, w innych tylko klasa złożoności problemu, w jeszcze innych wyłącznie złożoność algorytmu. Dla problemu REC-FHS nie podano ani złożoności metody ani klasy złożoności problemu, choć wpis sugeruje, że algorytm jest liniowy a problem łatwy. Z kolei dla problemu RME-FHS autor podaje, że złożoność algorytmu jest wykładnicza – dla spójności lepiej byłoby tę złożoność zapisać w notacji O, podobnie jak w pozostałych przypadkach. Pozostałe treści w tym rozdziale nie budzą zastrzeżeń.

Rozdział 4 jest poświęcony wykrywaniu wielokrotnych duplikacji w ukorzenionych drzewach genów. Autor formułuje badany problem oraz przedstawia trzy algorytmy rozwiązujące go dla różnych modeli scenariuszy. Drugi opisany algorytm (Algorithm 2) działający w czasie liniowym został opracowany przez autora rozprawy. Algorytmy przedstawione zostały w postaci czytelnego i zrozumiałego pseudokodu. W rozdziale znajdują się również wyniki eksperymentu obliczeniowego, w którym porównano działanie przedstawionych algorytmów. Wyniki przedstawione są bardzo klarownie, zostały zilustrowane wykresami i rysunkami. Rozdział zawiera też wyczerpujący opis trzech zbiorów danych, które wykorzystano do przeprowadzenia eksperymentu obliczeniowego. Z opisem każdego zbioru danych skojarzony jest rysunek pokazujący topologię drzewa gatunków.

W Rozdziale 5 autor skupił się na problemie uzgadniania nieukorzenionych drzew genów z drzewem gatunków. Uzgadnianie w takim przypadku stanowi poważne wyzwanie dla bioinformatyka i znacznie podnosi złożoność problemu. Rozdział zawiera analizę problemu UEC (Unrooted Episode Clustering) w połączeniu z modelem PG dopuszczalnych scenariuszy. Wprowadzona zostaje formalna definicja problemu oraz opisane jego właściwości. Następnie pokazane są dwa nowe algorytmy dokładne rozwiązujące problem UEC. Rozdział zawiera również analizę złożoności proponowanych metod oraz wynik ich ewaluacji w eksperymencie obliczeniowym. Obydwa algorytmy zostały poddane testom na tych samych zbiorach danych, które posłużyły do przebadania efektywności algorytmów z Rozdziału 4.

Rozdział 6 zadedykowano problemowi UME (Unrooted Minimum Episodes), czyli wykrywaniu wielokrotnych duplikacji w nieukorzenionych drzewach genów, z grupowaniem metodą minimalnych epizodów oraz dopuszczalnymi scenariuszami, które zachowują minimalną liczbę pojedynczych duplikacji genomowych. Autor przedstawia obszerne studium problemu UME i prezentuje cztery algorytmy (dokładne oraz heurystyczne) adresowane do rozwiązania tego problemu. Podobnie jak w poprzednich przypadkach, algorytmy są zaprezentowane w formie czytelnych pseudokodów. Rozdział zawiera również dość szczegółową analizę działania zaimplementowanych metod. Do eksperymentalnej ewaluacji algorytmów dla problemu UME wykorzystano zbiory opisane w Rozdziale 4.

Ostatni rozdział, tj. Rozdział 7, zawiera krótkie podsumowanie wyników zaprezentowanych w rozprawie oraz propozycję przyszłych badań. Zarysowane plany badawcze są ambitne. Autor zamierza m.in. przeprowadzić analizę złożoności kilku problemów przedstawionych w rozprawie, wykonać eksperymenty obliczeniowe zaprezentowanych modeli oraz zastosować swoje algorytmy do wykrywania duplikacji całych genomów.

Głównym mankamentem rozprawy jest brak jasno sformułowanego celu badań oraz planu badawczego, które powinny zostać podane w rozdziale wstępnym. Przedstawienie na początku rozprawy celu badawczego oraz scenariusza zwięźle objaśniającego sposób realizacji poszczególnych zadań znacznie ułatwiłoby czytelnikowi zrozumienie co i dlaczego opisano w kolejnych rozdziałach. Brak takiego ogólnego wprowadzenia sprawia, że pierwsza połowa pracy jest dość trudna w odbiorze a kontekst badań niezupełnie jasny. Cel badawczy ujawniony zostaje dopiero w jednym krótkim zdaniu na początku Rozdziału 3 (str. 45). Z kolei motywacja przedstawiona jest w Rozdziale 1.4 (str. 26). Uważam, iż kolejność powinna tu zostać odwrócona (najpierw cel, a dopiero potem motywacja).

Drobny mankament stanowi również brak wyraźnego podkreślenia głównych wyników badawczych (co można było wypunktować w krótkim wstępie na początku rozprawy). Autor szczegółowo opisuje metody wykorzystywane przy identyfikowaniu duplikacji genomowych ale nie zawsze jest oczywistym, które z tych metod są wynikiem jego pracy doktorskiej, a które zostały zaproponowane wcześniej przez innych badaczy (jedyną wskazówką bywa odnośnik do literatury i nazwiska autorów cytowanych prac). Pomocnym jest jednak krótki wstępny akapit każdego rozdziału zapowiadający co znajdziemy w jego dalszej części.

Chciałabym też zwrócić uwagę na fakt, iż temat rozprawy w języku polskim jaki podaje w autor („Studium zdarzeń duplikacji w genomie") niezupełnie odpowiada angielskiej wersji tematu („Inferring genomic duplication events"). Dodatkowo w dokumentacji przesłanej przez Dziekana Wydziału Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego brzmi on inaczej („Rekonstrukcja zdarzeń duplikacji genomów").

Mimo wyżej wymienionych mankamentów, moja ocena strony merytorycznej niniejszej rozprawy jest bardzo wysoka. Autor pracy wykazał się doskonałą znajomością aktualnego stanu badań w dziedzinie genomiki obliczeniowej, genetyki i filogenetyki ewolucyjnej. Jest specjalistą w zakresie algorytmiki praktycznej. Swobodnie porusza się również w obszarze informatyki teoretycznej, co pokazują m.in. jego analizy złożoności obliczeniowej badanych problemów i algorytmów. Opracował nowe, efektywne algorytmy dla kilku wersji nietrywialnego problemu jakim jest wykrywanie duplikacji genomowych i przetestował je na rzeczywistych i symulowanych zbiorach danych biologicznych. Dowiódł, że zaproponowane i zaimplementowane przez niego metody są konkurencyjne w stosunku do już istniejących rozwiązań i otwierają nowe możliwości w badaniach nad ewolucją genomu.

Ocena strony redakcyjnej

Język recenzowanej rozprawy doktorskiej jest poprawny i zrozumiały. Ogólna struktura pracy nie budzi zastrzeżeń. Praca została przygotowana bardzo starannie, nad wyraz estetycznie, z dużą dbałością o szczegóły. Wszystkie ilustracje wykonane są wzorowo, świetnie uzupełniają treść poszczególnych rozdziałów, mają wysoką jakość i są wykonane bardzo estetycznie. Źródła informacji zostały prawidłowo wybrane i zacytowane. Pracę ogląda się z dużą przyjemnością. W trakcie czytania zauważyłam niewiele usterek i błędów. Większość z nich wymieniam poniżej:

1. Str. 9: "In the 4th century BC ancient Greek" → "In the 4th century BC, ancient Greek"

2. Str. 10: "Only in 1668 Francesco" → "Only in 1668, Francesco"
3. Str. 10: "Experiments on the Generation of Insects" – tytuł pracy powinien być podany w cudzysłowiu
4. Str. 10: "In XVIII century the origins" → "In XVIII century, the origins"
5. Str. 10: "Finally in 1859 Pasteur" → "Finally in 1859, Pasteur"
6. Str. 11: "They grew cultures (...) and then add T1" → "They grew cultures (...) and then added T1"
7. Str. 11: "In 1998 a model" → "In 1998, a model"
8. Str. 13: "Escherichia coli" – nazwy łacińskie powinny być pisane kursywą
9. Str. 15: "When TIM bind to PER" → "When TIM binds to PER"
10. Str. 16: "Haemophilus influenzae", "Saccharomyces cerevisiae", "Caenorhabditis elegans" – nazwy łacińskie powinny być pisane kursywą
11. Str. 21: "purine is replaced another purine" → "purine is replaced by another purine"
12. Str. 21: "98,5%" → "98.5%"
13. Str. 21: "Replication slippage is probably also responsible for Huntington's disease (...)" – brakuje referencji
14. Str. 22: "Transposition activity influence genes" → "Transposition activity influences genes"
15. Str. 22: "1,4 x 10^9" → prawdopodobnie miało być "1.4 x 10^9"
16. Str. 22: "0,541 x 10^9" → prawdopodobnie miało być "0.541 x 10^9"
17. Str. 23: "*Evolution by Gene Duplication*" – tytuły są pisane raz kursywą a innym razem zwyczajnym stylem, raz w cudzysłowiu, raz bez; dobrze byłoby to uspójnić
18. Str. 25: „1,73% (...) less than 1,5%" → „1.73% (...) less than 1.5%"
19. Str. 28: "faster in in murid" → "faster in murid"
20. Str. 28: "metabolic rate and effective" → "metabolic rate, and effective"
21. Str. 29: "in Africa 200,000 years", "in Africa some 200 000 years" – brak spójności w pisowni, separatorem powinien być przecinek
22. Str. 38: "A rooted gene tree is a rooted binary tree" → wg mnie definicja powinna być uzupełniona tak: „A rooted gene tree T is a rooted binary tree"
23. Str. 40: "is depicted on Figure" → "is depicted in Figure"
24. Str. 42: "to edges of *X* in unique" → "to edges of *X* is unique"
25. Str. 45: "we propose mathematical description" – matematyczny opis czego?

26. Str. 56: „Let chose minimum" → „Let us chose the minimum"
27. Str. 81: Tabela 5.1 - w języku angielskim separatorem dziesiętnym jest kropka a nie przecinek.
28. Str. 101: „led to results that are comparable" → "led to results that were comparable"

Powyższe usterki nie mają znaczącego wpływu na czytelność pracy i nie umniejszają jej wartości. Nie zmieniają również mojej ogólnej wysokiej oceny recenzowanej dysertacji.

Wnioski końcowe

W moim przekonaniu autor recenzowanej pracy wykazał się umiejętnością poprawnej i przekonywującej prezentacji wyników przeprowadzonych badań oraz trafnością wnioskowania. Dowiódł, iż w bardzo wysokim stopniu poznał dotychczasowy stan wiedzy o podejmowanym w pracy badawczej temacie (identyfikacja zdarzeń duplikacji w genomie oraz ich lokalizacja w uzgodnionym drzewie gatunków), przedstawiany w przedmiotowej literaturze krajowej i zagranicznej. Posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyki, Bioinformatyki, Genomiki i Genomiki Obliczeniowej oraz wykazuje się umiejętnością samodzielnego prowadzenia pracy naukowej i zastosowania wiedzy w praktyce.

Recenzowana praca zawiera oryginalne rozwiązanie problemu naukowego. Uzyskane przez autora wyniki badań zostały opublikowane w wiodących, wysoko punktowanych czasopismach z dziedziny. Były również prezentowane na najważniejszych konferencjach naukowych skupiających się na badaniach związanych z analizą danych genomowych (m.in. RECOMB-CG). Pracę oceniam bardzo wysoko. Ze względu na istotność uzyskanych wyników, szerokie podejście do rozwiązywanego problemu oraz bardzo dobre publikacje będące podstawą pracy, składam wniosek o jej wyróżnienie.

Stwierdzam, że praca pana mgra Jarosława Paszka pt. „Inferring genomic duplication events" spełnia wymagania stawiane rozprawom doktorskim określone w art. 13.1 Ustawy o stopniach naukowych i tytule naukowym z dnia 14.03.2003 oraz stanowi oryginalne rozwiązanie przez autora zagadnienia naukowego.

Marta Szachniuk

Dr hab. inż. Marta Szachniuk