prof. dr hab. inż. Krzysztof Giaro
Politechnika Gdańska
Wydział Elektroniki, Telekomunikacji i Informatyki

Gdańsk, 24.10.2022

Recenzja

rozprawy doktorskiej mgr Agnieszki Mykowieckiej
pt. *Inference of Credible Associations between Genes and Genomes*
(tytuł polski: *Rekonstrukcja wiarygodnych relacji między genami a genomami*)

1. Przedmiot rozprawy

Jednym z głównych działów bioinformatyki jest filogenetyka komputerowa, dziedzina stawiająca sobie za cel wypracowanie metod obliczeniowych umożliwiających rekonstrukcję historii ewolucji zbiorów współczesnych gatunków w formie ich genealogii, zwanej drzewem filogenetycznym. Korzeniem drzewa powinien być wspólny przodek całej analizowanej grupy. Jednak przyjmowane modele ewolucji często nie odróżniają kierunku upływu czasu wzdłuż krawędzi grafu, dlatego w rzeczywistości rezultatem analiz zwykle są drzewa filogenetyczne nieukorzenione, przedstawiające strukturę relacji pokrewieństwa gatunków, lecz pozbawione informacji o położeniu ich wspólnego przodka. Próba umiejscowienia najstarszego punktu tak reprezentowanej historii (ukorzenianie drzewa) stanowi wówczas kolejny etap analizy filogenetycznej. Informację o historii rozchodzenia się linii gatunkowych pozyskujemy zazwyczaj, porównując sekwencje biologiczne (genów lub protein) obecne u badanych organizmów współczesnych. Pomimo rozwoju metod filogenetyki wiadomo, że rezultaty badań mogą być obarczone błędami, skutkującymi uzyskiwaniem drzew o zaburzonej topologii. Badania filogenetyczne muszą bowiem mierzyć się z szeregiem niekorzystnych warunków, m.in. z niedoskonałymi zestawieniami wejściowych sekwencji biologicznych, uproszczonymi modelami ich ewolucji, olbrzymimi rozmiarami dziedzin zagadnień optymalizacyjnych (zazwyczaj NP-trudnych), do których sprowadzane jest zagadnienie filogenetyczne, wreszcie – z faktem, że historia nawet najbardziej prawdopodobna w świetle współcześnie obserwowanych danych nie musiała wystąpić w przeszłości. Skutkuje to potrzebą opracowania metod szacowania wiarygodności otrzymywanych wyników, a więc całych drzew filogenetycznych lub choćby fragmentów ich topologii, w szczególności pojedynczych kladów (krawędzi grafu) w nich wyodrębnianych. Istnieje szereg takich testów: *bootstrap test*, *jackknife test*, testy oparte na podejściach dystansowych (np. estymowanie długość krawędzi i jej wariancji), wsparcie bayesowskie itp. Pierwszy z wymienionych, mimo dyskusyjnych podstaw teoretycznych, cieszy się największą popularnością i doczekał się implementacji w licznych pakietach narzędziowego oprogramowania bioinformatycznego. Polega na wygenerowaniu wielu sztucznych zestawów wejściowych danych sekwencyjnych (poprzez losowy wybór kolumn z zestawień rzeczywistych sekwencji biologicznych, będących podstawą uzyskania ocenianej filogenezy),

a następnie powtórzeniu procesu rekonstrukcji drzewa dla każdego z nich. Częstość wystąpień kladu drzewa oryginalnego w populacji filogenez odtworzonych na podstawie takich losowych transformacji danych wejściowych (drzewa próbkowe), zwykle wyrażaną w procentach, określa się jako poziom wsparcia bootstrapowego tego kladu. Im bliższa 100% jest ta wartość, tym wyższym zaufaniem obdarzamy oceniany fragment struktury filogenezy. Powyższy schemat postępowania odnosi się do podstawowego wariantu zagadnienia filogenetycznego, w którym współczesne gatunki/organizmy opisywane są nieco odmiennymi, zmutowanymi wersjami wspólnego genu, a ich ewolucyjną przeszłość utożsamiamy z historią samych organizmów. Jednak genomy podlegają transformacjom na drodze bardziej złożonych procesów. Głównym celem deklarowanym w rozprawie jest przeniesienie schematu postępowania metody *bootstrap* do oceny wiarygodności rezultatów analiz innego typu.

Mianowicie drzewo historii genów obecnych w genomach grupy gatunków może znacząco różnić się od historii tejże grupy w wyniku zjawisk takich jak:

a) powielanie się genów na nici DNA (duplikacje – uważane za główny mechanizm wykształcania się nowych genów, zwłaszcza w ewolucji organizmów eukariotycznych), generujące osobne, niezależnie ewoluujące kopie materiału genetycznego;

b) delecja (utrata) odcinków DNA, często zawierających wiele genów;

c) transfery horyzontalne, przenoszące geny od gatunku dawcy do współczesnego mu, lecz bezpośrednio niespokrewnionego biorcy, inkorporującego do własnego DNA cudze geny (zjawisko rzadkie na obecnym etapie ewolucji zwierząt, lecz powszechne wśród drobnoustrojów, a niekiedy także roślin).

Finalnie u organizmów należących do badanej grupy może wystąpić kilka lub wręcz nie wystąpić żaden spośród rodziny historycznie spokrewnionych genów, a rekonstruowane na podstawie genów filogenezy mogą uzyskać odmienne topologie. W grafie historii genów, obok standardowych węzłów specjacyjnych (rozdzielenie się linii ewolucyjnych), występują wierzchołki duplikacji (rozdzielanie się linii historii obu nowych kopii genu), a transfery horyzontalne – reprezentowane przez dodatkowe krawędzie – wymuszają rezygnację z drzewiastej topologii grafu historii gatunków, co prowadzi do rozmaitych modeli sieci filogenetycznych, w różnych sformułowaniach proponowanych w aktualnej literaturze dziedziny.

Zagadnienia rozważane w przedłożonej rozprawie można najogólniej przedstawić następująco: dysponując znanym ukorzenionym drzewem dla zbiorów gatunków oraz drugim drzewem filogenetycznym zawartych w nich genów, a w pewnych ujęciach – także częściowymi informacjami o możliwych umiejscowieniach transferów horyzontalnych, należy odtworzyć przebieg wspólnej historii, odpowiednio „nawlekając" linie ewolucyjne drugiego drzewa poprzez pierwsze. Formalnie poprawnych przebiegów takiego uzgodnienia drzew (ang. *tree reconciliation*) jest zwykle wiele. W pracy obrano podejście parsymoniczne, umożliwiając przypisanie duplikacjom, delecjom i transferom wybranych dodatnich kosztów (są to stałe parametry modelu), a następnie przyjmując jako najbardziej wiarygodną taką historię, dla której sumaryczny koszt postulowanych transformacji jest najmniejszy. Uzgadnianie historii gatunków i ich genów zostaje tu sprowadzone do problemu optymalizacji dyskretnej. Doktorantka rozważa kilka wybranych zagadnień takiej postaci. Dla każdego z nich przedstawiła efektywne (wielomianowej złożoności) algorytmy optymalne oparte na paradygmacie programowania dynamicznego. Zostały one zaimplementowane i przetestowane w szeregu eksperymentów obliczeniowych, prowadzonych na rzeczywistych danych biologicznych lub sekwencjach wygenerowanych przez komputerową symulację procesów ewolucyjnych. Niski koszt obliczeniowy stworzonych algorytmów pozwala w praktyce na ich wielokrotne, iteracyjne wykonywanie, podobnie jak się to czyni w klasycznym teście *bootstrap*. Doktorantka zaproponowała więc jego odpowiedniki, definiując poziomy wsparcia duplikacji i transferów horyzontalnych postulowanych przez scenariusz optymalny (czyli o najniższym koszcie). Praca zdaje się koncentrować na tychże nowych wskaźnikach wiarygodności, choć oczywiście przedstawione algorytmy szybkiego uzgadniania historii genów i gatunków są wartościowe same w sobie, a zakres ich zastosowań jest szerszy.

2. Zawartość pracy, ocena ogólna

Dysertację spisano w języku angielskim, na 126 stronach, wydzielając 6 rozdziałów i podsumowanie. Całość zamyka adekwatnie dobrana, ponad 110-pozycyjna bibliografia.

Rozdział 1 to krótkie, poglądowe omówienie pierwszych, jeszcze starożytnych filozoficznych koncepcji ewolucji świata ożywionego, nowożytnych odkryć genetyki i biologii molekularnej, procesu wewnątrzkomórkowego przepływu informacji genetycznej, mikrobiologicznych podstaw ewolucji, wreszcie – metod odtwarzania jej historii i oceniania wiarygodności rezultatów.

Rozdział 2 jest wprowadzeniem formalnym, definiującym stosowane w rozprawie podstawowe pojęcia drzew filogenetycznych i ich elementów, najprostszych scenariuszy uzgadniania drzew, a także funkcji kosztu tych scenariuszy. Zaprezentowano też klasyczne sformułowanie metody testu *bootstrap*.

W rozdziale 3 Doktorantka rozważa model uzgadniania, w którym ukorzenione drzewo gatunków S i drzewo genów G są dane, bierzemy zaś pod uwagę zjawiska duplikacji i delecji genów. Jeśli G, podobnie jak S, ma korzeń, wiadomo, że scenariuszem o najmniejszej liczbie takich zdarzeń (koszt DL) jest proste mapowanie LCA (ang. *least common ancestor*), odwzorowujące węzły drzewa genów G w najniższe biologicznie sensowne odpowiedniki w filogenezie gatunków. Określa ono jednoznacznie, które węzły wewnętrzne w G były specjacjami, a które – duplikacjami. Doktorantka proponuje zaadaptowanie metody *bootstrap* do wyznaczania poziomów wsparcia wierzchołków w G razem z określeniem ich typów – sprawdzamy, jak często w drzewach próbkowych pojawia się dany klad z przypisanym mu tym samym typem (specjacji lub duplikacji) co w ocenianym drzewie genów. Główne wyniki rozdziału 3 idą jednak dalej i dotyczą przypadku nieukorzenionego drzewa genów G. Opierają się na wcześniejszej pracy Góreckiego i Tiuryna opublikowanej w „Bioinformatics" (2007), charakteryzującej ukorzenienia G o minimalnym koszcie DL. Okazuje się, że zbiór krawędzi będących optymalnymi punktami ukorzenienia można wyznaczyć w czasie liniowym, a każdy z takich scenariuszy jednakowo identyfikuje węzły G jako specjacje bądź duplikacje. Można więc znów zdefiniować typ węzła drzewa G, niezależnie od wyboru optymalnego ukorzenienia, a później także wyznaczyć poziom wsparcia bootstrapowego tej klasyfikacji. W pracy zaproponowano efektywny (liniowej złożoności) algorytm określający specjacyjny/duplikacyjny typ węzłów wraz z odpowiednimi wartościami testu *bootstrap*. Algorytm ten odwołuje się do wyników wspomnianej pracy promotora Doktorantki, a także do techniki zbliżonej do stosowanej w algorytmie szybkiego wyznaczania dystansu Robinsona–Fouldsa między drzewami. Zaproponowano też dwie wsparte eksperymentami obliczeniowymi propozycje zastosowania nowych testów, np. do ukorzeniania drzewa genów. Ciekawe, dlaczego tylko w rozdziale 3 przyjęto prosty model optymalizacji nieważonego kosztu DL, po prostu zliczającego wystąpienia duplikacji i delecji, bez możliwości nadania obu rodzajom mutacji zróżnicowanych dodatnich wag. Wydaje się, że uogólnienie przedstawionych wyników na wariant ważony kryterium powinno być łatwe.

W rozdziale 4 Doktorantka poszukuje optymalnych (o minimalnym ważonym koszcie łącznym duplikacji, delecji i transferów) scenariuszy uzgodnienia binarnego ukorzenionego drzewa genów G z grafem gatunków S, czyli ukorzenioną filogenezą gatunków uzupełnioną o dodatkowe skierowane krawędzie transferowe, wzdłuż których mogą przepływać transfery horyzontalne. Przedstawiono rozwiązanie w formie efektywnego algorytmu o złożoności $O(|G||S|)$. Doktorantka deklaruje, że inspiracją była praca opublikowana przez Scornavaccę i in. w „Journal of Theoretical Biology" (2017), jednak przyjęty tam formalizm sieci filogenetycznej gatunków różni się od używanego w rozprawie. W dalszej części rozdziału znów określono odpowiedniki poziomów wsparcia bootstrapowego, tym razem poszczególnych krawędzi transferowych w wejściowym grafie gatunków S (jest to ułamek scenariuszy uzgodnienia używających tej krawędzi spośród wszystkich uzgodnień optymalnych). Do wyznaczania grafu S zaproponowano heurystykę przypominającą schemat *branch and bound*, rozbudowującą wejściową filogenezę gatunków o nowe łuki transferowe. Procedura wprowadza je na próbę i usuwa w razie stwierdzenia nieak-

ceptowalnie niskiego wsparcia. Działanie tych metod sprawdzono w eksperymentach obliczeniowych.

Jeszcze inny model przyjęto w rozdziale 5: zarówno historia genów, jak i historia gatunków są tu podane w formie drzew ukorzenionych, a wykrycie, które krawędzie pierwszego z nich są transferami horyzontalnymi, stanowi część procesu wyznaczania najbardziej wiarygodnego (najtańszego w sensie sumy wag) scenariusza uzgadniania. W literaturze przedmiotu już wcześniej rozstrzygnięto, że choć w ogólnym ujęciu problem ten ma złożoność wielomianową, to optymalne rozwiązania mogą być pozbawione fizycznego sensu. Dzieje się tak, gdy otrzymywana historia transferów formuje cykle na osi czasu. Wprowadzenie wykluczającego takie patologiczne wyniki warunku acykliczności scenariusza skutkuje NP-trudnością problemu. Poza sformułowaniem ogólnym opisywano także podprzypadek wielomianowy wariantu acyklicznego, w którym węzły wejściowego drzewa gatunków S są dodatkowo etykietowane swoistymi stemplami czasowymi, określającymi kolejność wystąpień specjacji w historii badanej grupy. W rozdziale 5 Doktorantka zdefiniowała nieco ogólniejszy cel badań: zakładamy, że jedynie o części współczesnych genów z G wiadomo, z których gatunków pochodzą, a właściwe przyporządkowanie pozostałych zamierzamy odczytać z wyznaczanego optymalnego scenariusza uzgodnienia. Jest to problem występujący przy analizach metagenomowych zsekwencjonowanej puli genów pobranych np. z próbki osadów wodnych, kiedy to nie mamy informacji o przynależności sekwencji do organizmów żyjących w środowisku. Przedstawione w rozprawie rozwiązania to wielomianowe algorytmy nieograniczające się do binarnych drzew gatunków – dopuszczalne są stopnie wyjściowe wierzchołków większe niż 2. Cechują je niskie złożoności $O(|G||S|\Delta S)$ dla wariantu ogólnego oraz $O(|G||S|^2\Delta S)$ przy węzłach S uporządkowanych w czasie. W rozprawie zadeklarowano możliwość dalszej redukcji drugiej z wymienionych złożoności do $O(|G||S|\Delta S \log|S|)$. Kończące rozdział eksperymenty na danych biologicznych rzeczywistych i symulowanych pokazują, jak przez losowe próbkowanie wyznaczonych scenariuszy optymalnych można estymować prawdopodobieństwa przynależności elementów puli genów do gatunków obecnych w środowisku. Chociaż uwzględnienie niebinarnych drzew gatunków jest ciekawe algorytmicznie i znacząco komplikuje opis formalny problemu, w analizie wyników Doktorantka stwierdza, że obecność wierzchołków wysokiego stopnia prowadzi do przeszacowania wystąpień postulowanych zdarzeń utrat genów, obniżając wiarygodność wnioskowania. Nie jest to zaskakujące, wszak węzły niebinarne w filogenezie są *de facto* przyznaniem naszej niewiedzy odnośnie do konkretnego fragmentu historii. Jeśli jest ich wiele – nie wiemy prawie nic, jeśli zaś mało – właściwsze byłoby zapewne powtórzenie testu osobno dla każdego sensownego rozwinięcia binarnego drzewa S. Wydaje się, że ciekawsze (i raczej łatwe do zaimplementowania) byłoby wyposażenie liści wejściowego drzewa G w listy dopuszczalnych dla nich gatunków – nawet jeśli nie wiemy dokładnie, jaki organizm jest posiadaczem znalezionego genu, sama jego sekwencja może dostarczyć częściowej wiedzy na ten temat.

Rozdział 6, odróżniający się od pozostałych, opisuje prowadzone przy współpracy z Leiden University Medical Center (LUMC) badania próbek sekwencyjnych pochodzących z biopsji pacjentów cierpiących na chłoniaka pęcherzykowego, a więc nowotwór, w którym intensywne rearanżacje i transfery materiału genetycznego w obrębie receptorów limfocytów B uniemożliwiają zaprezentowanie tego procesu w formie struktury jakkolwiek przypominającej drzewo ewolucji. Autorzy analizują zachodzące zmiany chorobowe w modelu gęstej sieci filogenetycznej o strukturze wnioskowanej głównie na podstawie dystansów międzysekwencyjnych; uzyskiwaną sieć poddają klasteryzacji i wizualizują w czytelniejszej formie. Chociaż dotychczasowe wyniki współpracy badawczej z LUMC zostały opublikowane w najwyżej punktowanym artykule Doktorantki („Blood” – Impact Factor Web of Science, IF WoS: 25.476, Ministerstwo Edukacji i Nauki: 200 pkt) i można mieć nadzieję, że przyczynią się do głębszego zrozumienia niektórych procesów nowotworowych, a w dalszej perspektywie – do opracowania nowych metod diagnostycznych, to jednak metody zastosowane w rozdziale 6 oddalają się od problematyki algorytmicznej w kierunku rozwiązań inżynierii analizy danych oraz ich prezentacji.

Rozprawę sformatowano starannie, z dbałością o estetykę prezentacji. Szczególne wrażenie robią skomplikowane, wielobarwne ilustracje, przedstawiające przykłady definiowanych pojęć i skrupulatnie raportujące wyniki eksperymentów obliczeniowych.

3. Słabe strony rozprawy

Nie mam istotnych uwag krytycznych do pracy.

Faktem jest, że przedstawione procedury w swoim jądrze eksploatują ten sam, wspólny paradygmat programowania dynamicznego i (cytat ze s. 65 i 87) „standard backtracking method". Niemniej modele uzgadniania historii ewolucji i zliczanie ich kosztów, zwłaszcza z uwzględnieniem transferów horyzontalnych, cechują się niewdzięcznym opisem formalnym, a szczegółowa analiza ich dopuszczalnych struktur musiała być skomplikowanym i wymagającym zadaniem.

Miejscami można odnieść wrażenie, iż Autorka nadmiernie opiera się na tekstach własnych publikacji. Nie zaszkodziłoby skorzystać z formuły dysertacji do szerszego rozwinięcia wątków jedynie wzmiankowanych oraz zadbać o ujednolicenie założeń i symboliki.

Uchybienia, uwagi polemiczne

- Zawężona definicja paralogów na s. 16 – zaliczamy do nich także geny obecne genomach innych gatunków, lecz wywodzące się z duplikacji.
- Pomijane założenia o binarności filogenez. Sekcja 2.1 definiuje podstawowe pojęcia drzew gatunków i genów w wersji ogólnej, tj. niekoniecznie binarnej, ale:
    - jeszcze w tym samym akapicie bliźniaka wierzchołka (ang. *sibling*) określono jedynie w drzewie binarnym (co nie jest wystarczające w sekcji 5.1.2);
    - drugi wiersz definicji M(g) ze s. 23 milcząco zakłada (jak bodajże cała praca) binarność G;
    - formuła na koszt L(T,S) (s. 24) jest poprawna dla drzew binarnych, choć w rozdziale 2 nie wspomniano o tym założeniu (faktycznie, nie obowiązuje ono w rozdziale 5);
    - czy gdziekolwiek zaznaczono wprost, że drzewa gatunków rozważane w rozdziale 3 i drzewa genów w rozdziale 4 są binarne?
- Podstawowe mapowanie LCA geny→gatunki ze s. 23 jest źle opisane. Oprócz wspomnianego założenia o binarności nie podano także, czym jest s. Akapit wyżej inkluzja między zbiorami liści $\mathcal{L}_G \subseteq \mathcal{L}_S$ jest niepotrzebnym ograniczeniem, wszak zaledwie 2 strony wcześniej zadeklarowano, że przypisanie genów do właścicieli (gatunków) podaje *leaf labelling* $\Lambda_G: \mathcal{L}_G \to \mathcal{L}_S$. Dalej w pracy oznaczenia zbiorów liści L i $\mathcal{L}$ bywają używane zamiennie.
- Nadmiarowe zdanie pod dowodem lematu 1. Już na s. 24 założono, że wierzchołkiem tym jest korzeń, więc zapewne miało to obowiązywać w całej pracy (choć wyjątkiem wydaje się rozdział 5).
- Zbyt krótki dowód twierdzenia 2. Fakt, że mapowanie M trzech sąsiadów węzła nie zmienia się podczas przemieszczania korzenia, nie wystarcza, gdyż na typ wpływa także zmieniająca się para dzieci wierzchołka. Należało raczej odpowiednio uzupełnić dowód lematu 1.
- Lemat 3 w obecnym sformułowaniu nie jest prawdziwy, np. klastrów typu (3) może w ogóle nie być, gdy drzewa genów i gatunków są identyczne i całe G jest *plateau*. Natomiast w przykładzie nad lematem 3 warto by wyjaśnić, że chodzi o klastry niewystępujące w ukorzenieniach z obszaru *plateau*.

- W eksperymencie z sekcji 3.2.2 rozwiązuje się trudny obliczeniowo problem superdrzewa dla kosztu DL za pomocą iteracyjnej heurystyki *fasturec*. Czy mamy pewność, że faktycznie znajduje ona globalne minimum, a może raczej uznano to za nieistotne?
- Na s. 45 nie wiadomo, czy odległość od wartości optymalnej o 100 i 400 to dużo, czy mało – nie podano bowiem, jaki rząd wielkości przyjmuje to optimum.
- Czy algorytm z twierdzenia 7 można łatwo (nie zwiększając asymptotycznej złożoności) wzbogacić o zliczanie rozwiązań optymalnych, bez ich jawnego zapamiętywania? Jeśli nie, to czy może ich być (wykładniczo) wiele? Jeśli tak, warto było jawnie odnotować ten fakt już w sekcji 4.4, gdyż definicja 2 poziomu wsparcia krawędzi transferowej wymaga znajomości mocy zbiorów scenariuszy optymalnych oraz optymalnych wykorzystujących tę krawędź. To samo można by napisać o procedurze z sekcji 5.1.2, choć faktycznie na s. 93 można znaleźć wzmiankę o losowaniu rozwiązań optymalnych z rozkładem równomiernym.
- Jeśli drugi z istotnych wyników algorytmicznych z rozdziału 5 (sekcja 5.2.1) można usprawnić, redukując złożoność o czynnik $|\Delta S|/\log |S|$, należało opisać tę implementację, zamiast odsyłać czytelnika do literatury. Jeszcze oszczędniej potraktowano interesujące możliwe uogólnienia pierwszej procedury (sekcja 5.3). Opis przypadku z nieukorzenionym drzewem genów sugeruje dodatkową pętlę przebiegającą jego krawędzie, a więc i wzrost złożoności o czynnik $|E_G|$. Skoro jednak złożoność nie rośnie, najprawdopodobniej zastosowano wyliczanie wartości pośrednich dla wszystkich ukorzenień naraz (podobnie jak w algorytmie 1), co raczej nie kwalifikuje się na – tu cytat – „omit easy details for brevity". Głębszych modyfikacji wymaga zapewne wariant problemu z niebinarnym drzewem genów, o którym jednak nie napisano nic, ograniczając się do podania złożoności.
- Pomyłki, niedoróbki redakcyjne, inne drobiazgi.
    - Oznaczenia:
        - na s. 33 ukorzeniamy na krawędzi nieskierowanej, a więc $G_{\{v,w\}}$; to samo pod koniec dowodu lematu 5;
        - punkt 1 algorytmu 1 – raczej *neighbors* zamiast *siblings* w drzewie bez korzenia, a wyżej winno być $\Lambda_P(L_P)$;
        - na s. 61 zapewne $s^*:=c$, a niżej, w H2: $m^{-1}(s^*)\subseteq V_{S'}$ lub ewentualnie w skrócie $m^{-1}(s^*)\subseteq S'$;
        - s. 62, wiersz -4 – winno być *g* zamiast *q* i już tu przydałby się przypis ze s. 81;
        - zdanie nad definicją 2 – na rys. 4.2 nie ma transferu δ, a podpis stwierdza, że jednak chodzi o κ;
        - czy akronim DP (sekcje 4.6–4.7.2) oznacza *dynamic programming*?
        - zbędne *c* przed zbiorami liści w definicji 3;
        - definicja D3 ze s. 84 – zamiast $\delta(g,s)$ raczej $\delta(g,v)$, a w drugim podpunkcie zamieniono argumenty s,v;
        - dowód twierdzenia 8: winno być *p* zamiast π (s. 85, tu również zdanie „We omit easy verification of cases D2, D3 and D4" – czy właśnie ich nie sprawdzono?), a stronę dalej: $\delta^{\Delta}(g,s)\rightarrow\delta^{\Delta}(g,q)$ w końcowym wzorze;
        - w cytowanej definicji *acyclicity condition* (z podwójną numeracją, s. 87) zgubiono *v* [powinno być: $M(\xi(v))$], a pod nią – zwykła nierówność $\tau(s) > \tau(s')$ między liczbami;
        - nieokreślone *s*' (raczej *s*) w nowej wersji (L3) na s. 88.
    - Podpisy i odsyłacze:
        - na rys. 2.3 nie ma oznaczeń krawędzi E1, E2, E3;
        - czy rys. 3.4 pokazuje wyniki dla S*/DL/…/400, czy S*/D/…/100?
        - na s. 32 winno znaleźć się odwołanie do twierdzenia 2 zamiast twierdzenia 3;
        - odwołania do (I)–(IV) w dowodzie twierdzenia 8 dotyczą D1–D4.

- Bibliografia:
  - zdublowane pozycje (przypuszczalnie podczas scalania spisów literatury zaczerpniętych z artykułów nie usunięto powtórzeń): Górecki, P. (2004a) = Górecki, P. (2004b); Ma, B., Li, M. and Zhang, L. (2000a) = Ma, B., Li, M. and Zhang, L. (2000b); być może także Rambaut, A. and Grass, N. C. (1997) = Rambaut, A. and Grassly, N. C. (1997);
  - autor Felenstein, J. (2004) wygląda, jakby zmutował podczas duplikacji;
  - brak danych o Maddison, W. (1997);
  - bibliografii nie uwzględniono w spisie treści.

4. Podsumowanie

Doktorantka jest zaangażowana w aktualny obszar badań, obecny we współczesnej literaturze bioinformatycznej. Metody rekonstrukcji historii ewolucji z uwzględnieniem zjawisk duplikacji, rozmaitych rearanżacji i transferu genów są wciąż rozwijane i nawet ich opis formalny nie przybrał jeszcze ostatecznego, powszechnie przyjmowanego kształtu. Dysertacja proponuje szereg nowych – a nawiązujących do testu bootstrapowego – miar oceny wiarygodności takich analiz wraz z efektywnymi algorytmami ich obliczania. Odwołują się one do opartych na paradygmacie programowania dynamicznego procedur wyznaczania parsymonicznie optymalnych scenariuszy uzgadniania historii genów i gatunków. Zakres ich zastosowań wykracza poza samo określenie poziomu wsparcia. Proponowane algorytmy zostały zaimplementowane i przetestowane w różnorodnych eksperymentach obliczeniowych, prowadzonych zarówno na rzeczywistych danych biologicznych, jak i danych otrzymanych poprzez symulacje. Zadania te wymagały również biegłego rozeznania w dostępnych pakietach narzędziowego oprogramowania bioinformatycznego. Stwierdzam, że *dysertacja spełnia wymogi ustawy o stopniach naukowych i tytule naukowym*, dlatego *wnioskuję o jej dopuszczenie do dalszych etapów przewodu doktorskiego*.

Należy także docenić wyróżniającą się aktywność publikacyjną Autorki. Wyniki przedstawione w rozprawie opublikowano w dwóch artykułach w międzynarodowym periodyku bioinformatycznym „IEEE/ACM Transactions on Computational Biology and Bioinformatics" (IF WoS: 3.702), materiałach dwóch konferencji międzynarodowych (AlCoB 2016 – LNIB 2016, vol. 9702; BIBM 2018), wreszcie – w prestiżowym piśmie medycznym „Blood" (IF WoS: 25.476). Prace Doktorantki z obszaru filogenetyki nie ograniczają się ściśle do treści dysertacji, jest ona także współautorką czterech kolejnych tekstów o międzynarodowym zasięgu (dwa czasopisma indeksowane na WoS i dwie konferencje zagraniczne). W związku z powyższym proponuję *przyznanie wyróżnienia* rozprawie.

Krzysztof Giaro