Gliwice 28 stycznia 2012 r.

Dr hab. inż. Joanna Polańska, prof. Pol. Śl.
Instytut Automatyki
Politechniki Śląskiej

Recenzja rozprawy doktorskiej

Tytuł rozprawy: Algorithms for modeling of the evolution of complex stochastic genetic systems

Autor rozprawy: Mgr inż. Tomasz Wojdyła

Promotor rozprawy: Prof. dr hab. inż. Marek Kimmel

1. Ogólna charakterystyka rozprawy

Recenzowana rozprawa doktorska jest napisana po angielsku, liczy 134 strony, składa się z siedmiu rozdziałów, spisu literatury obejmującego 188 pozycji, oraz kilku dodatków. Tematem pracy są modele matematyczne stosowane w genetyce statystycznej oraz w statystycznej teorii ewolucji, zwłaszcza algorytmiczne i numeryczne aspekty stosowania tych modeli i ich dopasowywania do danych eksperymentalnych.

Modelowanie matematyczne procesów i mechanizmów ewolucji, takich jak różnego rodzaju mutacje w replikacji DNA, dryf genetyczny, rekombinacja i selekcja mają historię badań obejmującą już wiele dziesięcioleci. Jednak w ostatnich latach dokonuje się ilościowa zmiana wynikająca z bardzo gwałtownego wzrostu ilości dostępnych danych eksperymentalnych. Zarówno liczebności badanych grup jak też dostępność pełnogenomowych sekwencji DNA podlega wielokrotnemu wzrostowi w krótkich odstępach czasu. Powoduje to z jednej strony możliwość coraz lepszych dopasowań (estymacji struktury i parametrów) stosowanych w genetyce statystycznej modeli matematycznych, a drugiej strony potrzebę stosowania coraz bardziej złożonych modeli. W modelach tych liczba analizowanych lokusów staje się coraz większa, a także coraz większe są np. rozmiary analizowanych genealogii. Dla analiz i rozwijania modeli matematycznych w genetyce statystycznej coraz istotniejsze stają się aspekty algorytmiczne i stosowane techniki obliczeniowe. Dlatego tematykę pracy należy uznać za bardzo ważną i współczesną.

1.1. Cel i Tezy rozprawy

Cel i tezy rozprawy sformułowane są na stronach 22 (cel) i 2 (tezy). Celem rozprawy jest stworzenie metod obliczeniowych dla efektywnego badania złożonych modeli ewolucyjnych. Trzy tezy rozprawy, sformułowane na stronie 2 można streścić następująco: Poprzez zastosowanie kombinacji pewnych metod analitycznych z numerycznymi formułami rekurencyjnymi możliwe jest 1) rozwiązanie problemu identyfikowalności rekombinacji (recombination identifiability) w modelu Morana, 2) stworzenie efektywnego iteracyjnego algorytmu wyliczanie rozkładów czasów do koalescencji w wielkich genealogiach, z dowolnymi modelami wzrostu populacji, 3) stworzenie efektywnego modelu ewolucji z interakcjami pomiędzy populacjami.

Wykazaniu powyżej sformułowanych tez poświęcona jest pozostała część rozprawy.

2. Zawartość rozprawy

Rozdział 1 jest pomyślany jako wstęp. Przedstawiona jest motywacja dla przeprowadzonych w ramach pracy studiów, przede wszystkim w aspekcie szerokiej dostępności danych z sekwencjonowania DNA i potrzeby ich analizy. Przedstawione są tezy pracy. Opisana jest także w skrócie struktura pracy. A także wymienione są oryginalne elementy rozprawy.

Rozdział 2 ma także charakter wprowadzający. Przedstawia się w nim mechanizmy ewolucyjne, mutację, dryf genetyczny, rekombinację, selekcję i migrację a także bardzo wszechstronny przegląd idei modeli matematycznych tych mechanizmów, model Fishera Wrighta, model Morana, modele procesów gałązkowych, modele koalescencji z czasem biegnącym zarówno wstecz jak i w przód, aproksymacje dyfuzyjne, modele Markowa, a także związane z nimi algorytmy symulacji stochastycznych.

W rozdziale 3 sformułowany jest cel pracy i jeszcze raz zapisane są tezy pracy. Każdej z tez poświęca się tu oddzielny podpunkt, w którym teza jest dokładnie omówiona, a także przedstawione są związane z nią wyniki w literaturze.

W rozdziale 4 sformułowany jest model Morana z dryfem genetycznym, rekombinacją oraz mutacją. Na wstępie przedstawia się kilka, wykorzystywanych dalej definicji i faktów dotyczących łańcuchów Markowa, półgrup Markowa, stanów stacjonarnych łańcuchów Markowa. Następnie zacytowany jest literaturowy wynik (praca 108) wykorzystujący idee Morana do modelowania oddziaływania mutacji, dryfu genetycznego i rekombinacji dwóch lokusów w DNA. Pozostała część rozdziału ma charakter oryginalny, poświęcona jest uogólnieniom modelu z pracy [108] na większą liczbę lokusów w DNA. Opisuje się strukturę opracowanych modeli, bada się ich stany stacjonarne oraz opisuje się algorytmy implementujące opracowane modele. Wykazane w rozdziale 4 twierdzenie 4.1 rozstrzyga także problem identyfikowalności rekombinacji, a zatem dowodzi pierwszej tezy pracy. Przedstawiona jest analiza złożoności obliczeniowej i pamięciowej opracowanego algorytmu. Przedstawiony jest przypadek trzech lokusów oraz wzory pozwalające na efektywne obliczeniowe analizy tego przypadku. Przedstawione są także przykłady numerycznych wyników analiz dla modelu z pięcioma lokusami (rysunki 4.1-4.5), a także wyniki dla innych liczb lokusów. Dokonuje się także porównania z algorytmem Hudsona wykorzystujących metodę koalescencji.

W rozdziale 5 przedstawione są modele matematyczne dryfu genetycznego dla populacji o zmiennej w czasie liczebności. Przedstawia się, w (5.1), podstawowy wzór opisujący wielowymiarowy rozkład czasów koalescencji w ewolucji zgodnej z modelem Fishera – Wrighta ze zmienną w czasie liczebnością populacji. Opisuje się następnie algorytm, pochodzący z pracy [18], wyliczania rozkładów prawdopodobieństwa czasów do MRCA (Najbliższego wspólnego przodka, Most Recent Common Ancestor). Wskazuje się następnie na ograniczenia obliczeniowe tego algorytmu oraz przedstawia się oryginalny, rekurencyjny algorytm (algorytm 5.1) stanowiący znaczące usprawnienie metody z pracy [18] oraz jego wyprowadzenie. Wykonując wielokrotnie opisany algorytm dokonuje się oszacowania rozkładu prawdopodobieństwa czasu do MRCA (rysunki 5.10-5.13) dla ludzkiej globalnej populacji. Porównuje się osiągnięty wynik z literaturą. Czas do MRCA dla globalnej ludzkiej populacji jest intensywnie studiowany w szeregu prac w literaturze i dokonanie weryfikacji tego parametru z użyciem rozwiniętego warsztatu jest bardzo interesującym oryginalnym osiągnięciem pracy.

W rozdziale 6 przedstawiony jest oryginalny model matematyczny sieci demograficznej (demographic network model). Przez sieć demograficzną rozumie się scenariusz ewolucyjny obejmujący oddziałujące ze sobą populacje, o zmiennej w czasie liczebności. Jest to dość realistyczny model, który może być zastosowany do wielu oddziaływujących populacji, dla których istnieją badania pozwalające oszacować rzeczywisty scenariusz ewolucji. W rozdziale 6 opisuje się zastosowanie opracowanego modelu do badania ewolucji Bałtów Słowian i Finów, zgodnie z rysunkiem 6.2. W ostatniej części rozdziału 6 przedstawia się także zjawisko obciążenia ocen opisywane angielskim terminem „ascertainment bias". Zjawisko to zostało zaobserwowane przy porównywaniu lokusów powtórzeń tandemowych pomiędzy populacjami ludzką i szympansów. Przedstawione jest zastosowanie opracowanego algorytmu do oceny opisanego zjawiska obciążenia. Dzięki zamodelowaniu obciążenia oceny możliwa jest jego kompensacja.

Ostatni rozdział 7 stanowi dyskusję wyników i podsumowanie pracy.

3. Ocena rozprawy

Silne strony rozprawy

Większość wyników przedstawionych w recenzowanej rozprawie ma charakter nowy i oryginalny, stanowią one istotne rozwinięcia i rozszerzenia cytowanych prac z literatury. Oryginalność zaprezentowanych metod wynika z jednej strony ze zrozumienia matematycznych metod modelowania procesów ewolucji, modeli Markowa, teorii koalescencji, modelu Fishera – Wrighta, Modelu Morana, a z drugiej strony z bardzo dobrego opanowania warsztatu informatycznego. Połączenie tych elementów pozwoliło na rozwinięcie metodologii, którą efektywnie zastosowano do modelowania oddziaływania dryfu, mutacji i rekombinacji dla wielu lokusów, dla oceny rozkładów prawdopodobieństwa czasów do MRCA, do badania sieci demograficznych, a także do badania zjawiska obciążenia (ascertainment bias) w pewnych modelach ewolucji.

Tezy rozprawy zostały przekonująco wykazane.

Zakres zastosowań opracowanych algorytmów w genetyce statystycznej i matematycznej teorii ewolucji jest bardzo szeroki. Zagadnienia badane w pracy najintensywniej studiowanych problemów modelowania matematycznego w genetyce i genomice.

Opracowane algorytmy zostały zaimplementowane w postaci programów komputerowych i zastosowane do analiz rzeczywistych danych. Doktorant dotarł do bardzo interesujących, najnowszych danych opisujących scenariusze ewolucyjne, które modelował (np. publikacje 100, 112, 113, 116, 117, 121, 136, 142, 159).

Autor opisuje i stosuje bardzo szeroki i zaawansowany warsztat matematyczny.

Praca jest odniesiona do bardzo szerokiego i współczesnego wykazu źródeł literaturowych.

Praca jest starannie zredagowana. Jakość ilustracji nie budzi zastrzeżeń.

Uwagi krytyczne i dyskusyjne

Ponieważ wynikiem pracy są konkretne obliczenia i oszacowania, niektóre ze stosowanych pojęć i metod, pochodzących z literatury mogłoby być dokładniej opisane, np. metoda oceny haplotypów (str. 89), odległość Slatkina (str. 89), mechanizmy obciążenia ocen (ascertainment bias) (str. 91).

W pracy bardzo wiele uwagi poświęca się implementacji modeli. Interesujące byłoby także dokładniejsze przedyskutowanie problemu oceny parametrów modeli, intensywności mutacji, rekombinacji, liczebności populacji.

Krytyka algorytmu wyprowadzonego przez prof. Bobrowskiego (wzór 5.7) przedstawiona na str. 54 jest dość nieprecyzyjna. Najlepiej byłoby wyznaczyć złożoność obliczeniową tego algorytmu.

Na str. 56 stwierdza się, że stosowana jest metoda programowania dynamicznego. Dobrze byłoby to przedstawić dokładniej. Powinien zostać przedstawiony odpowiedni problem optymalizacji dynamicznej, oraz wyprowadzone równanie Bellmana.

Str. 59 – Jest: „Module GW performs experiments" – powinno się napisać: „... experiments performed by using module GW ..."

Str. 55 – Jest: „as follow" – powinno być „ as follows ".

4. Podsumowanie

Uwagi krytyczne mają charakter bądź dyskusyjny bądź marginalny, zdecydowanie przeważają pozytywne elementy oceny.

Stwierdzam, że recenzowana rozprawa spełnia wymogi odpowiedniej ustawy i wnioskuję o dopuszczenie jej do publicznej obrony.

Dodatkowo, biorąc pod uwagę bardzo wysoki poziom naukowy rozprawy, szereg oryginalnych wyników zamieszczonych w pracy oraz opublikowanie przez autora dwóch oryginalnych artykułów:

Adam Bobrowski, Tomasz Wojdyła, Marek Kimmel, (2010), Asymptotic behavior of a Moran model with mutations, drift and recombination among multiple loci, J. Math. Biol. (2010) 61:455–473.

Tomasz Wojdyła, Marek Kimmel, Adam Bobrowski, (2011), Time to the MRCA of a sample in a Wright–Fisher model with variable population size, Theoretical Population Biology.

w bardzo renomowanych czasopismach naukowych, wnioskuję o wyróżnienie pracy.