Prof. dr hab. Włodzimierz Ogryczak
Instytut Automatyki i Informatyki Stosowanej
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

8 luty 2018

*RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY NAUKOWEJ INSTYTUTU PODSTAWOWYCH PROBLEMÓW TECHNIKI PAN*

Tytuł rozprawy: Metody obliczeniowe jedno- i wielokryterialnej optymalizacji rojem cząstek. Zastosowanie w bioinformatyce

Autor rozprawy: mgr Mateusz Banach

1. Jakie zagadnienie naukowe jest rozpatrzone w pracy (teza rozprawy) i czy zostało ono dostatecznie jasno sformułowane przez autora? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, inny)?

   Rozprawa dotyczy rozwiązywania ważnego specyficznego zagadnienia z zakresu bioinformatyki. Przedmiotem rozprawy jest użycie i rozwój metod optymalizacji jedno- i wielokryterialnej do weryfikacji założeń modelu rozmytej kropli oliwy (fuzzy oil drop, FOD) związanych z wpływem opisywanych przez ten model oddziaływań hydrofobowych na proces tworzenia się kompleksów typu białko-białko. W celu sprawdzenia poprawności założeń modelu FOD został opracowany eksperyment *in silico* przewidywania struktury kompleksu 200 białek homodimerycznych wybranych z bazy Protein Data Bank (PDB).

   Tezy rozprawy nie zostały sformułowane, natomiast zostały jasno sformułowane trzy cele rozprawy. Mają one naturę hierarchii w tym sensie, że każdy następny umożliwia realizację wcześniejszego. Zasadniczym celem była weryfikacja założeń modelu rozmytej kropli oliwy (FOD) dotyczących wpływu oddziaływań hydrofobowych na proces tworzenia się kompleksów typu białko-białko w oparciu o eksperyment przewidywania struktury czwartorzędowej 200 białek homodimerycznych wybranych z bazy PDB i porównanie uzyskanych wyników z wynikami uzyskanymi przy pomocy chemicznych pól siłowych (pole ECEPP/3). Drugim celem było użycie metaheurystyki PSO (algorytmów opartych na sposobie działania roju cząstek) do wykonania eksperymentu i sprawdzenie, czy jest możliwe jej skutecznie stosowanie w symulacji procesów związanych z białkami. Trzecim celem było opracowanie algorytmu optymalizacji wielokryterialnej opartego na PSO, umożliwiającego przeprowadzenie symulacji równoczesnego wpływu sił opisywanych przez model FOD i pole ECEPP/3. Dodatkowo cel poboczny, obejmował modyfikacje sposobu obliczeń wykonywanych przez model FOD dla umożliwienia jego efektywnego stosowania jako kryterium procedurach optymalizacji.

2. Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł (w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle) świadczący o dostatecznej wiedzy autora? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?

   Szeroka analiza źródeł zawiera się w dwóch pierwszych rozdziałach rozprawy. Rozdział 1.Wprowadzenie zawiera podrozdział 1.3. Dotychczasowy stan wiedzy omawiający kolejno istotne dla umiejscowienia zadania elementy i zagadnienia: mechaniki molekularnej, chemicznych pól siłowych, modeli wody i hydrofobowości, kompleksowania białek oraz optymalizacji jedno i wielokryterialnej. Rozdział 2.Materiały i metody uściśla wykorzystywane modele i metody. W pierwszej (biologicznej) części jest przedstawiona baza danych białek homodimerycznych oraz pola wewnętrzne (ECEPP/3) i zewnętrzne (model FOD) użyte w eksperymencie przewidywania struktury czwartorzędowej białek. Druga (informatyczna) część omawia dokładnie algorytm optymalizacji rojem cząstek (PSO) oraz pozostałe algorytmy wykorzystane do realizacji zadania. Literatura cytowana w tych rozdziałach jest bardzo bogata i obejmuje ponad 300 pozycji. Ta obszerność wynika z rozległości tematyki pracy, ale także z cytowania historycznych prac inicjujących poszczególne metody a nie współczesnej literatury monograficznej (np. praca Land i Doig dla metody podziału i ograniczeń). Z drugiej srony brak jest najnowszych pozycji literatury np. dotyczących wielokryterialnych metod PSO.

   Wyjątkowo obszerna bibliografia obejmuje 428 pozycje. Jest ona uporządkowana według kolejności pierwszego cytowania. Brak jest standardowego stylu pozycji bibliograficznych i w wielu przypadkach to samo czasopismo bywa różnie podawane. Należy podkreślić staranność w zakresie zapewnienia adresów DOI dla prawie wszystkich pozycji, co zapewnia szybki dostęp do literatury w przypadku korzystania z elektronicznej wersji rozprawy. Przeprowadzona przez autora analiza źródeł świadczy o głębokiej wiedzy autora w danym zakresie. Wnioski są formułowane w sposób jasny i przekonywujący.

3. Czy autor rozwiązał postawione zagadnienia, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione?

   W celu sprawdzenia poprawności założeń modelu FOD został opracowany eksperyment *in silico* przewidywania struktury kompleksu 200 białek homodimerycznych wybranych z bazy Protein Data Bank. Do wykonania tego eksperymentu zastosowano dwa algorytmy: algorytm optymalizacji rojem cząstek (PSO) oraz specjalnie opracowany algorytm wielokryterialnych rodzin rojów (MOSF). Za pomocą algorytmu PSO została wykonana optymalizacja globalna konformacji kompleksu według osobnych kryteriów pola zewntrznego i wewntrznego, czyli sprawdzono efekty osobnego wpływu oddziaływań hydrofobowych (opisywanych przez model FOD) oraz oddziaływań niekowalencyjnych (opisywanych przez pole ECEPP/3) na układy par łańcuchów polipeptydowych. Natomiast za pomocą specjalnie rozwiniętego wielokryterialnego algorytmu MOSF wykonano optymalizację dwukryterialną, czyli sprawdzono efekty równoczesnego wpływu oddziaływań hydrofobowych (opisywanych przez model FOD) oraz oddziaływań niekowalencyjnych (opisywanych przez pole ECEPP/3) na układy par łańcuchów polipeptydowych. Ocena zgodności

uzyskanych w tym eksperymencie kompleksów białkowych z ich strukturami natywnymi została wykonana przy pomocy miary pierwiastka średniej kwadratów różnicy (RMSD) i w przestrzeni krzywych ROC porównania map kontaktów niewiążących.

Efektywne wykorzystanie metaheurystyk optymalizacyjnych do weryfikacji modelu FOD wymagało modyfikacji modelu FOD dotyczącej sposobu układania atomów efektywnych białka. Opracowanie sposobu zastąpienia algorytmu opartego na średnicach analizą składowych głównych do układania atomów efektywnych białka zgodnie z osiami układu współrzędnych pozwoliło na zmniejszenie złożoności obliczeniowej procedury wyznaczania rozkładu hydrofobowości teoretycznej modelu FOD ze względu na liczbę reszt z kwadratowej w najgorszym przypadku (liniowo-logarytmicznej oczekiwanej) do liniowej.

Algorytm MOSF został zaprojektowany dla uzyskania możliwości wykonywania w trakcie optymalizacji wielokryterialnej analizy skupień odnalezionych wektorów niezdominowanych, osiągania jednorodnej reprezentacji zawartości zbioru Pareto-optymalnego oraz skalowalności dzięki liniowej złożoności obliczeniowej ze względu na liczbę optymalizowanych kryteriów i cząstek. Efektywność algorytmu MOSF, do rozwiązywania różnorakich wielokryterialnych problemw optymalizacyjnych, została zademonstrowana przez porównanie ze standardowymi algorytmami: NSGA-II i NSPSO, na zbiorze wybranych funkcji testowych oraz generowanych losowo.

Rozprawa składa się formalnie z pięciu rozdziałów. Dwa z nich mają charakter wprowadzający (1.Wprowadzenie oraz 2.Materiały i metody) oraz dwa podsumowujący (4.Dyskusja i wnioski oraz 5.Podsumowanie). Cała zasadnicza merytoryczna treść rozprawy jest zawarta w jednym rozdziale (4.Wyniki) liczącym 100 stron. Rozdział ten zarówno przedstawia zaproponowany algorytm optymalizacji wielokryterialnej MOSF, jego porównanie z innymi metodami z tej dziedziny, jak również modyfikację modelu FOD, analizę białek z bazy danych, a także opis oraz wyniki eksperymentu *in silico* kompleksowania typu białko-białko. Jego uzupełnieniem są dodatki zawierające rysunki i tabele. Cała rozprawa liczy 300 stron.

4. Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stany wiedzy czy poziomu techniki reprezentowanych przez literaturę światową?

   Oryginalność rozprawy polega na wprowadzeniu nowych modeli optymalizacyjnych i technik algorytmicznych.

   Opracowano nowy algorytm optymalizacji wielokryterialnej oparty na zasadzie metaheurystyk PSO. Algorytm o nazwie MOSF (od multi objective swarm families), daje możliwości wykonywania w trakcie optymalizacji wielokryterialnej analizy skupień wyznaczonych wektorów niezdominowanych. Algorytm MOSF zapewnia skalowalność dzięki liniowej złożoności obliczeniowej ze względu na liczbę optymalizowanych kryteriów i cząstek, a także może być wykonywany równolegle z zyskiem czasowym proporcjonalnym do liczby jednostek obliczeniowych. Eksperymentalne porównanie algorytmu MOSF ze standardowymi algorytmami NSGA-II i NSPSO wykazuje wyższą dokładność w przybliżaniu zbioru Pareto-optymalnego i frontu Pareto osiąganą w porównywalnym lub krótszym czasie.

Opracowano eksperyment *in silico* przewidywania struktury czwartorzędowej białek polegający na optymalizacji globalnej i wielokryterialnej kryteriów pól zewnętrznego i wewnętrznego przy pomocy algorytmów typu PSO. Dla jego realizacji zaproponowano modyfikację modelu FOD w zakresie sposobu układania atomów efektywnych białka zgodnie z osiami ukadu współrzędnych opartą na wykorzystaniu technik analizy składowych głównych. Obniżyło to znacznie złożoność obliczeniową procedury wyznaczania rozkładów hydrofobowości. Wykazana została skuteczność stosowania algorytmów optymalizacji opartych na zasadzie PSO (jedno- i wielokryterialnych) w opracowanym eksperymencie. Pokazano przy tym, że modelowanie jednoczesnego wpływu pola zewnętrznego i wewnętrznego przy pomocy optymalizacji wielokryterialnej prowadzi do lepszych wyników (większej liczby struktur kompleksów zbliżonych do natywnych) niż w przypadku jednokryterialnej optymalizacji globalnej. Są to nowe wartościowe wyniki. Do tej pory niepodejmowano podejścia do tematyki optymalizacji wielokryterialnej w zastosowaniu do badań nad wpływem oddziaływań hydrofobowych na procesy związane z białkami.

5. Czy autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?

   Zasadniczo autor wykazał umiejętność poprawnego i przekonywującego przedstawiania uzyskanych przez siebie wyników. Skomplikowane zależnosci są bogato ilustrowane odpowiednimi przykładami. Rozprawa zachowuje przy tym zwięzłość i jasność przy formułowaniu i uzasadnianiu tez oraz prezentacji wyników. Tym niemniej rozprawa jest bardzo obszerna obejmując blisko 300 stron, w tym ok. 60 stron dodatki z rysunkami i tabelami, a ponad 20 bibliografia. Redakcja roprawy jako całości nie jest najlepsza, ze względu na przyjętą strukturę pracy, gdzie poza częścią wprowadzającą (rozdz. 1 i 2) cała zasadnicza treść rozprawy jest zawarta w jednym 100 stronicowym rozdziale o ogólnym tytule Wyniki.

6. Do której z następujących kategorii Recenzent zalicza rozprawę:

   (a) nie spełniająca wymagań stawianych rozprawom doktorskim przez obowiązujące przepisy

   (b) wymagająca wprowadzenia poprawek i ponownego recenzowania

   (c) spełniająca wymagania

   (d) spełniająca wymagania z wyraźnym nadmiarem — **TAK**

   (e) wybitnie dobra, zasługująca na wyróżnienie

   Uważam, że rozprawa mgr. Mateusza Banacha spełnia z wyraźnym nadmiarem wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy. Należy przy tym podkreślić, że część wyników rozprawy została już opublikowana w literaturze międzynarodowej, a w szczególności w wielu współautorskich artykułach w czasopismach indeksowanych w JCR.