Warszawa, dnia 23.08.2013

Recenzja rozprawy doktorskiej pana magistra Jakuba Pasia pt. "Application and implementation of probabilistic profile-profile comparison methods for protein fold recognition" (Wdrożenie i zastosowanie probabilistycznych metod porównawczych profil-profil w rozpoznawaniu pofałdowanych białek)

Rozprawa doktorska pana magistra Jakuba Pasia promowana przez pana doktora habilitowanego Marcina Hoffmanna oraz promotora pomocniczego doktora Krystiana Eitnera z Wydziału Chemii Uniwersytetu im. Adama Mickiewicza w Poznaniu została przedstawiona Radzie Wydziału Chemii UAM.

Format rozprawy

Przewód doktorski magistra Pasia realizowany jest według „nowego trybu" (Ustawa z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki, Ustawa z dnia 18 marca 2011 r. o zmianie ustawy – Prawo o szkolnictwie wyższym, ustawy o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki oraz o zmianie niektórych innych ustaw, a także Rozporządzenie Ministra Nauki i Szkolnictwa Wyższego z dn. 22 września 2011 r.). Przedstawiona rozprawa zawiera spis opublikowanych przez doktoranta prac naukowych będących podstawą rozprawy, krótkie (jedna strona w języku polskim oraz około 25-stronicowy tekst w języku angielskim) streszczenie celów i wyników przedstawionych prac oraz kopii siedemnastu publikacji naukowych doktoranta. Spis tych współautorskich publikacji zawiera dość ogólnikowe oświadczenia doktoranta o jego udziale w opracowaniu wyników prezentowanych badań. Nie przedstawiono oświadczeń współautorów o ich udziałach. Dotychczas recenzowane przeze mnie nieliczne rozprawy doktorskie przedstawiane według „nowego trybu" zawierały takie oświadczenia współautorów. Z drugiej strony Ustawa z dnia 14 marca 2003 r., której artykuł 13 został zacytowany na początku przedstawianej rozprawy, nie stawia takiego wymagania. Prace badawcze wchodzące w zakres prezentowanej rozprawy doktorskiej były realizowane w kilku instytucjach, głównie w stworzonym przez doktora Leszka Rychlewskiego prywatnym instytucie BioinfoBank, ale też w Instytucie Chemii Bioorganicznej Polskiej Akademii Nauk oraz na Wydziale Fizyki Uniwersytetu im. Adama Mickiewicza. Załączone publikacje ukazały się w latach 2002-2011, głównie w renomowanych czasopismach o zasięgu ogólnoświatowym. Poniżej ich spis, zgodny z kolejnością przyjętą w tekście rozprawy.

1. Gould, C.M., Diella, F., Via, A., Puntervoll, P., Gemünd, C., Chabanis-Davidson, S., Michael, S., Sayadi, A., Bryne, J.C., Chica, C., Seiler, M., Davey, N.E., Haslam, N., Weatheritt, R.J., Budd, A., Hughes, T., Paś, J., Rychlewski, L., Travé, G., Aasland, R., Helmer-Citterich, M., Linding, R., Gibson, T.J. "ELM: The status of the 2010 eukaryotic linear motif resource" (2010) Nucleic Acids Research, 38, pp. D167-D180.
2. Ginalski, K., Pas, J., Wyrwicz, L.S., von Grotthuss, M., Bujnicki, J.M., Rychlewski, L. "ORFeus: Detection of distant homology using sequence profiles and predicted secondary

Uniwersytet Warszawski, Wydział Chemii – ul. Pasteura 1, 02-093 Warszawa
tel. 022 822 09 75; centr. 022 822 02 11; fax 022 822 59 96
e-mail: dziekan@chem.uw.edu.pl www: http://www.chem.uw.edu.pl
Bank Millenium S.A. 12 1160 2202 0000 0000 6084 9173 NIP 525-001-12 -66

structure" (2003) Nucleic Acids Research, 31 (13), pp. 3804-3807.
3. Feder, M., Pas, J., Wyrwicz, L.S., Bujnicki, J.M. "Molecular phylogenetics of the RrmJ/fibrillarin superfamily of ribose 2'-O-methyltransferases" (2003) Gene, 302 (1-2), pp. 129-138.
4. Von Grotthuss, M., Koczyk, G., Pas, J., Wyrwicz, L.S., Rychlewski, L. "Ligand.Info small-molecule meta-database" (2004) Combinatorial Chemistry and High Throughput Screening, 7 (8), pp. 757-761.
5. Von Grotthuss, M., Pas, J., Wyrwicz, L., Ginalski, K., Rychlewski, L. "Application of 3D-Jury, GRDB, and Verify3D in Fold Recognition" (2003) Proteins: Structure, Function and Genetics, 53 (SUPPL. 6), pp. 418-423.
6. Von Grotthuss, M., Pas, J., Rychlewski, L. "Ligand-Info, searching for similar small compounds using index profiles" (2003) Bioinformatics, 19 (8), pp. 1041-1042.
7. Pas, J., Wyszko, E., Rolle, K., Rychlewski, L., Nowak, S., Zukiel, R., Barciszewski, J. "Analysis of structure and function of tenascin-C" (2006) International Journal of Biochemistry and Cell Biology, 38 (9), pp. 1594-1602.
8. Pas, J., Von Grotthuss, M., Wyrwicz, L.S., Rychlewski, L., Barciszewski, J. "Structure prediction, evolution and ligand interaction of CHASE domain" (2004) FEBS Letters, 576 (3), pp. 287-290.
9. Plewczyński, D., Paś, J., von Grotthuss, M., Rychlewski, L. "3D-Hit: fast structural comparison of proteins" (2002) Appl Bioinformatics, 1 (4), pp. 223-225.
10. Barciszewska, M.Z., Szymanski, M., Wyszko, E., Pas, J., Rychlewski, L., Barciszewski, J. "Lead toxicity through the leadzyme" (2005) Mutation Research - Reviews in Mutation Research, 589 (2), pp. 103-110.
11. Plewczynski, D., Pas, J., Von Grotthuss, M., Rychlewski, L. "Comparison of proteins based on segments structural similarity" (2004) Acta Biochimica Polonica, 51 (1), pp. 161-172.
12. Wyrwicz, L.S., Von Grotthuss, M., Pas, J., Rychlewski, L. "How Unique Is the Rice Transcriptome?" (2004) Science (Letters to the Editor), 303 (5655), p. 168.
13. Narozna, D., Paś, J., Schneider, J., Mądrzak, C.J. "Two sequences encoding chalcone synthase in yellow lupin (Lupinus luteus L.) may have evolved by gene duplication" (2004) Cellular and Molecular Biology Letters, 9 (1), pp. 95-105.
14. Von Grotthuss, M., Wyrwicz, L.S., Pas, J., Rychlewski, L. "Predicting protein structures accuratety" (2004) Science (Letters to the Editor), 304 (5677), p. 1597.
15. Fischer, D., Paś, J., Rychlewski, L. "The PDB-Preview database: A repository of in-silico models of 'on-hold' PDB entries" (2004) Bioinformatics, 20 (15), pp. 2482-2484.
16. Wyszko, E., Nowak, M., Pospieszny, H., Szymanski, M., Pas, J., Barciszewska, M.Z., Barciszewski, J. "Leadzyme formed in vivo interferes with tobacco mosaic virus infection in Nicotiana tabacum" (2006) FEBS Journal, 273 (22), pp. 5022-5031.
17. Pas, J., Stępniak, P., Wyrwicz, L.S., Ginalski, K., Rychlewski, L. "GRDB-Gene Relational DataBase" (2011) BioinfoBank Library Acta, 11(1) p.2659

Uniwersytet Warszawski, Wydział Chemii – ul. Pasteura 1, 02-093 Warszawa
tel. 022 822 09 75; centr. 022 822 02 11; fax 022 822 59 96
e-mail: dziekan@chem.uw.edu.pl www: http://www.chem.uw.edu.pl
Bank Millenium S.A. 12 1160 2202 0000 0000 6084 9173 NIP 525-001-12 -66

Tylko dwie z tych publikacji (Numer 9 i 17) nie maja określonego IF. Pozostałe to obszerne prace lub krótkie listy (w Science) w czasopismach o znaczących lub bardzo wysokich współczynnikach oddziaływania (IF).

Streszczenia i podsumowania rozprawy

Jak wspomniałem powyżej, rozprawa zawiera króciusieńkie streszczenie w języku polskim i bardziej obszerne omówienie i podsumowanie uzyskanych wyników w języku angielskim. **Streszczenie w języku polskim** w istocie nie jest streszczeniem rozprawy, a tylko krótkim przedstawieniem postawionego zadania badawczego. To sformułowanie celu jest czytelne, aczkolwiek rażą błędy i niezręczności językowe. Zamiast słowa „fold" powinno się używać polskiego słowa „zwój". Sekwencje mogą być podobne, ale nie homologiczne – homologiczne są białka. A co znaczy zdanie: „Każde nowo odkryte białko ma duże szanse by zostać *sklsyfikowane* jako członek jednej z takich grup"? Podobnych niezręczności językowych i nieścisłości nomenklaturowych jest więcej (na jednej stronie).

Część w języku angielskim również otwiera krótkie streszczenie (Rozdział I: **Summary**). Treść pierwszej połowy strony jest nieco podobna do streszczenia w języku polskim. Tu też jest trochę niezręczności. Na przykład: pierwszy akapit rozpoczyna zdanie „Fold recognition is a method of fold detection and protein tertiary structure prediction…", a następny akapit zaczyna się od prawie identycznego zdania "Fold recognition methods are useful for protein structure prediction…".

Właściwe omówienie wyników pracy doktorskiej (strony 3-28) poprzedzone jest wprowadzeniem do podstaw teoretycznych rozprawy (Rozdział II: **Introduction**). Wstęp ten omawia problemy związane z porównaniem aminokwasowych sekwencji białek i wynikających z tego relacji ewolucyjnych. W pierwszym paragrafie zatytułowanym **Sequence comparison methods used for fold recognition** dość szczegółowo omówione są klasyczne metody bioinformatyczne porównywania (uliniowienia) sekwencji. Przedstawiono kolejno algorytmy: Needleman-Wunsch, Smith-Waterman i BLAST. W drugim paragrafie (**Sequence-Profile methods**) omówiono nieco bardziej złożone sposoby porównywania sekwencji z tak zwanymi profilami sekwencyjnymi: PSI-BLAST i RPS-BLAST. Trzeci paragraf (**Profile-profile methods**) przedstawia metody wzajemnego porównywania profili sekwencyjnych: BASIC (Bilateral Amplified Sequence Information Comparison), FFAST (Fold & Function Assignemnt) i ORFEUS. W paragrafie czwartym opisane są bardziej ogólne metody identyfikacji zwojów białek (**Other fold recognition methods**). Chyba najczęściej, i z dużym powodzeniem, stosowaną grupą metod jest przewlekanie (**threading**). Przewlekanie polega na przeszukiwaniu różnych sposobów nałożenia sekwencji białka testowego na sekwencje białek o znanych strukturach. Poszczególne nałożenia oceniane są według różnych kryteriów podobieństwa sekwencyjnego oraz preferencji energetycznych wynikających z uzyskiwanych elementów struktury przestrzennej. W drugiej części paragrafu czwartego zdefiniowano metody *ab initio* (chyba częściej używa się sformułowań *de novo* lub free-modeling) przewidywania struktury białek. To chyba najciekawsze wyzwanie teoretyczne w biologii molekularnej białek, dalekie jeszcze od zadawalającego rozwiązania. Rozdział II kończy omówienie niektórych najbardziej znanych doświadczalnych testów różnych metod rozpoznawania zwojów (**Improvement and benchmarking of fold recognition methods**). Pokrótce przedstawione są zasady ogólnoświatowych konkursów CASP (Critical Assesment of protein Structure Prediction), CAFASP (Critical Assesment of

Fully Autoamated Structure Prediction) oraz LIVEBENCH. Pomimo drobnych usterek językowych i bardzo skróconej prezentacji oceniam tę ogólną część wstępu teoretycznego (Rozdział II) bardzo wysoko. Tekst ten świadczy o dobrym zrozumieniu najważniejszych podstawowych problemów badawczych bioinformatyki/biologii obliczeniowej białek.

Następne dwa rozdziały (III. **Applications profile-profile comparison methods** oraz IV. **Implementations of profile-profile comparison methods**) opisują już metody obliczeniowe rozwinięte i zastosowane w trakcie realizacji pracy doktorskiej. W tytule rozdziału trzeciego chyba brakuje „of". Na marginesie, nawiązując do tytułu całej rozprawy i jego tłumaczenia na język polski razi przejście od „Application and implementation.." do „Wdrożenie i zastosowanie...". Chyba jest to inna kolejność użytych terminów, a tym samym mylące tłumaczenie. Cały ten początek tytułu wydaje mi się niezręczny. Wracając do tekstu domyślam się, że rozdziały III-IV powinny chyba być skrótowym podsumowaniem opublikowanych wyników, ale niektóre publikacje włączone do rozprawy nie są cytowane. Omówione w rozdziale III praktyczne zastosowania metod opartych na porównywaniu profili sekwencyjnych są interesujące i dotyczą ważnych zagadnień biologii molekularnej. Pokazano, że tego typu podejście pozwala na wykrywanie bardzo odległych ewolucyjnie, ale homologicznych białek oraz dostarczają dobrych uliniowień sekwencyjnych, co umożliwia tworzenie dokładniejszych modeli strukturalnych.

W ostatnim (IV) rozdziale przeglądu wyników rozprawy omówiono dwa serwery internetowe, w opracowaniu których doktorant odegrał znaczną role. Pierwszy to PDB Preview, serwer który raz na tydzień analizuje sekwencje białek pojawiające się w PDB (Protein Data Bank) z zapowiedzią spodziewanych struktur doświadczalnych. Serwer PDB Preview identyfikuje sekwencje o odległym lub niemożliwym do identyfikacji podobieństwie do znanych struktur. Dla białek tych przewidywane są prawdopodobne struktury teoretyczne. Ma to potencjalnie duże znaczenie dla wielu działów biologii molekularnej, w tym dla interpretacji danych strukturalnych. Rozwinięta przez doktoranta metoda profil-profil została też wykorzystana w stworzonym przez BioinfoBank serwisie internetowym GRDB (Gene Relation DataBase). Serwer pozwala na wykrywanie nowych powiązań (homologii) rodzin białek, rozszerzając możliwości innych metod bioinformatycznych.

W podsumowaniu rozprawy (Rozdział V) autor w przekonywujący sposób wymienia najważniejsze osiągnięcia opublikowanych badań, a także wskazuje ciekawe kierunki łączenia zaproponowanej metody porównywania sekwencji białkowych z innymi metodami teoretycznymi, w szczególności z dynamiką molekularną. Biorąc pod uwagę rosnące możliwości współczesnych komputerów, takie połączenie może być bardzo atrakcyjnym sposobem dokładnego przewidywania struktur białek.

Całe streszczenie rozprawy, pomimo drobnych usterek językowych, napisane jest interesująco i świadczy o bardzo dobrym zrozumieniu stawianych zadań badawczych. Doktorant pokazał też, że zdaje sobie sprawę ze znaczenia opracowanych metod obliczeniowych dla wielu kluczowych zagadnień biologii molekularnej.

Publikacje naukowe doktoranta

Magister Jakub Paś jest współautorem 17 publikacji naukowych składających się na przedstawioną rozprawę doktorską. To bardzo duży dorobek, istotnie powyżej średniej dla prac doktorskich. Większość z tych publikacji ukazała się w renomowanych czasopismach naukowych, w tym dwie prace w NAR (IF=8.3), dwa krótkie listy w SCINCE (IF=31.2), praca przeglądowa w Mutations Research/Reviews in Mutation Research (IF=8.2), itd. Tylko

dwie z prezentowanych publikacji nie mają określonego IF: praca w Applied Bioinformatics i tekst w dokumentach BioinfoBank. Zapewne stosunkowo nowemu tytułowi Applied Bioinformatics wkrótce zostanie przyporządkowany IF. Już teraz czasopismo to jest stosunkowo często cytowane. Główne prace wchodzące w zakres rozprawy są bardzo dobrze cytowane, według Web of Knowledge (z dnia 23.08.2013) to 424 cytowania, przy czym praca z NAR ("ELM: The status of the 2010 eukaryotic linear motif resource") ma już około 100 cytowań. Jak na trzy lata od ukazania się tej publikacji to doskonały wynik. Doktorant podaje nieco inne, ale bardzo podobne, liczby cytowań. Prawdopodobnie zostały one pobrane z jego strony w Google Scholar. Google Scholar zwykle pokazuje nieco większe liczby cytowań, prawdopodobnie uwzględniając cytowania w materiałach zjazdowych, czy innych tego typu źródłach ignorowanych przez Web of Knowledge (czy Web of Science). Prace z NAR, z roku 2003 i 2010 to chyba najważniejsze publikacje doktoranta. Pierwsza z nich opisuje metodę obliczeniową i działanie opracowanego przez Gianlskiego, Pasia, Wyrwicza, Grotthusa, Bujnickiego i Rychlewskiego serwera internetowego ORFeus do przewidywania odległych relacji ewolucyjnych pomiędzy białkami globularnymi. Metoda wykorzystuje dobrze pomyślaną kombinację profili sekwencyjnych oraz przewidywanych teoretycznie struktur drugorzędowych. Pokazano, że ORFeus dostarcza istotnie dokładniejszych wyników niż otrzymywane przez inne istniejące metody przewidywania odległych homologii białek. Nieco późniejsza praca w NAR została wykonana przez większą liczbę autorów z kilku laboratoriów, głównie spoza Polski. To też opis serwera internetowego - ELM (Eukaryotic Linear Motifs). Jako „motywy liniowe" (LM) określone są krótkie fragmenty białek wielodomenowych, mające ważne funkcje regulacyjne, na ogół niezależne od struktury trzeciorzędowej. Takie fragmenty białek są bardzo trudne do przewidywania metodami informatycznym. Serwer ELM nastawiony jest na gromadzenie i opracowywanie danych doświadczalnych. Nie ulega wątpliwości, że przedstawiono ciekawe wyniki, a zaproponowany schemat analizy „motywów liniowych" ma istotne znaczenie dla projektowania i interpretacji doświadczalnych badań biologów molekularnych. Duża liczba cytowań też wskazuje na użyteczność prezentowanego opracowania. Doktorant nie tylko brał udział w bioinformatycznych projektach mających na celu analizę białkowych baz danych, ale również wykazał się bardzo dobrą znajomością metod teoretycznego przewidywania struktur białek. Prace nr 7 i nr 8 są przykładami dobrego zastosowania różnych metod biologii strukturalnej do ciekawych zagadnień biologii molekularnej. Magister Paś jest pierwszym autorem tych publikacji.

Nie będę szczegółowo oceniał jakości naukowej wszystkich publikacji doktoranta. Większość z dołączonych prac była zapewne poddana szczegółowej ocenie przez recenzentów wybranych przez edytorów poszczególnych czasopism. W renomowanych czasopismach, o wysokim współczynniku oddziaływania (IF) są to zwykle bardzo rygorystyczne oceny. Tak więc włączony do rozprawy dorobek naukowy pana magistra Pasia został już drobiazgowo oceniony przez wielu ekspertów. Warto też pokreślić, że badacze z którymi współpracował doktorant są rozpoznanymi ekspertami w ważnych dziedzinach biologii molekularnej i bioinformatyki. Prace naukowe profesora Barciszewskiego, profesora Bujnickiego, doktora Ginalskiego i doktora Rychlewskiego w sumie były cytowane ponad 10 tysięcy razy. Doktorant brał więc udział w ważnych i rozpoznawanych w światowej nauce zadaniach badawczych. Jestem przekonany, że miało to duży, pozytywny wpływ na jego rozwój naukowy. Publikacje włączone do przedstawianej rozprawy przedstawiają nowe i bardzo wydajne metody bioinformatyczne, przydatne w analizie ewolucyjnego pokrewieństwa białek, przewidywaniu ich struktury oraz funkcji biologicznych. Co ważne,

nowe metody bioinformatyczne zostały udostępnione środowiskom naukowym w formie serwerów internetowych. Ciekaw jestem, jaki jest zakres ich wykorzystywania przez innych badaczy. Spodziewam się, że serwery te gromadzą dane statystyczne na ten temat. Rozprawa nie zawiera takich informacji.

Ocena końcowa

W podsumowaniu mojej oceny rozprawy doktorskiej pana magistra Jakuba Pasia chcę przede wszystkim stwierdzić, że prezentowany dorobek naukowy oceniam bardzo wysoko. Wyniki prezentowane w 17 publikacjach włączonych do rozprawy są ważne, ciekawe i mogą mieć duże znaczenie dla wielu zadań badawczych bioinformatyki i biologii molekularnej. Uważam, że rozprawa doktorska pana magistra spełnia wymagania ustawowe i zwyczajowe stawiane przewodom doktorskim. Wnoszę zatem do Rady Wydziału Chemii Uniwersytetu im. Adama Mickiewicza w Poznaniu o dopuszczenie pana magistra Jakuba Pasia do dalszych etapów przewodu doktorskiego.

Andrzej Koliński

Prof. dr hab. Andrzej Koliński

Uniwersytet Warszawski, Wydział Chemii – ul. Pasteura 1, 02-093 Warszawa
tel. 022 822 09 75; centr. 022 822 02 11; fax 022 822 59 96
e-mail: dziekan@chem.uw.edu.pl www: http://www.chem.uw.edu.pl
Bank Millenium S.A. 12 1160 2202 0000 0000 6084 9173 NIP 525-001-12 -66