

Posudek dizertační práce

Autorka dizertační práce:

Ing. Denisa Maděránková

Název práce:

Identifikace organismů pomocí analýzy nukleotidových denzitních vektorů

Předložená dizertační práce se zabývá numerickým mapováním genomických dat a využitím konkrétní metody založené na nukleotidových denzitních vektorech pro molekulární identifikaci organismů.

Téma práce je podnětné a vědecky aktuální – s ohledem na současný rapidní rozvoj oboru bioinformatika. Svým zaměřením tato práce zasahuje do řady oblastí od molekulární biologie přes bioinformatiku až po zpracování a analýzu digitálních signálů, a vykazuje tedy vysokou míru interdisciplinarity. Dle mého názoru dizertační práce Ing. Denisy Maděránkové dobře zapadá do oboru Biomedicínská elektronika a biokybernetika.

Autorka ke zpracovávanému tématu přistupuje systematicky a problematiku jednotlivých kapitol popisuje velmi podrobně a srozumitelně. Úvodní dvě kapitoly jasně vysvětlují teoretické pozadí vzniku genomických dat, popisují princip a využití tzv. DNA barcodingu pro molekulární identifikaci organismů. Ve třetí kapitole se autorka věnuje různým metodám pro převod nukleotidových sekvencí v symbolickém zápisu do numerické reprezentace, přičemž hlavní autorčinou motivací je využít pro následné analýzy metody zpracování digitálních signálů známé na poli biomedicínského inženýrství. Po čtvrté kapitole, která sumarizuje cíle disertační práce, se autorka věnuje výhradně vlastním metodám a algoritmům pro zpracování genomických dat a jejich komparační analýzu s cílem identifikace druhů organismů a vyšších taxonomických kategorií. Výzkum, který autorka v této části popisuje, proběhnul na datech pořízených nikoli laboratorně, ale z veřejně dostupných databází. Tento fakt lze považovat za výhodu předložené disertační práce, neboť umožnil autorce jednak otestovat své metody na velkém počtu souborů dat, které by v lokálních laboratorních podmínkách nebylo možné posbírat, a dále své dosažené výsledky porovnávat s výstupy jiných, v dané oblasti standardně používaných metod.

Za hlavní vědecký přínos předložené práce lze považovat prokázání výhodnosti neobvyklého deterministického přístupu k analýze genomických dat. Autorka volí pro analýzu dat metody jednoduché a přímočaré, což se na poli bioinformatického překotného výzkumu často nevidí. Při zkoumání řádově desítek souborů druhů a čeledí autorka podrobila své komparační analýze řádově tisíce sekvencí, u nichž bylo potřeba provést zarovnání, převod na numerickou reprezentaci pomocí nukleotidových denzitních vektorů s použitím různých délek výpočetního okna, vytvoření referenčních sekvencí s použitím různých metod pomocí průměru denzit, mediánu denzit a průměru denzit purinových nukleotidů. Kromě samotných výsledků identifikace autorka v práci poukazuje na řadu nesrovnalostí ve veřejné databázi BOLD. Práce tak odhalila řadu sekvencí s chybně uvedenou klasifikací, přičemž takovéto případy autorka podrobně diskutuje a poukazuje u nich i na výsledky dalších identifikačních metod (zejména metody BLAST), které pro ověření svých tvrzení používá, prohledává i další databáze, zabývá se i zvláštnostmi morfologické klasifikace u takto problematických případů. Dobrý dojem budí i přehledná grafická znázornění procentuální

úspěšnosti identifikace do vyšších taxonomických kategorií (čeledí a řádů), kde se autorka vypořádává s potřebou ukázat výsledky ve vztahu k řadě parametrů konfigurace její metody.

Práce je napsaná v českém jazyce, její rozsah je 159 stran včetně všech formálních náležitostí. Samotný text práce na 137 stranách je rozdělen do 9 kapitol s přehlednou a logickou hierarchií podkapitol. Práce obsahuje 84 obrázků, 22 tabulek a odkazuje se na 174 literárních pramenů. Po formální stránce nelze autorce vytknout nic závažného – pouze několik nevýznamných gramatických či stylistických chyb (chyba ve slově „standartní“ – str. 94, nedobré skloňování v některých komplikovanějších větách – např. str. 125 ř. 10, překvapivý odklon od trpného rodu k vyprávění v první osobě singuláru – str. 105). Tyto drobné nedostatky však nesnižují celkově velmi pozitivní dojem z formálního zpracování práce i z jejího samotného obsahu.

Zadání k vypracování oponentského posudku vyžaduje, aby se oponent vyjádřil k následujícím bodům:

1) Odpovídá námět práce oboru disertace a je aktuální z hlediska současného stavu vědy?

Ano, jak konstatuji ve druhém odstavci svého posudku.

2) Vykazuje práce původní přínosné části? Konkretizujte prosím, v čem spatřujete originální přínos

Ano, práce vykazuje původní přínosné části, jak uvádím ve čtvrtém odstavci svého posudku.

3) Bylo jádro disertační práce na potřebné úrovni publikováno?

V diskuzi o dosažených výsledcích klasifikace sekvencí do vyšších taxonomických kategorií (čeledí a řádů) – kap. 7.4 autorka uvádí, že dosahuje srovnatelné, ba i lepší procentuální úspěšnosti (byť na jiných datech) než dosud publikované práce, přičemž poukazuje i na to, že se touto problematikou zabývala za posledních pět let jediná práce v impaktovaném časopise. S ohledem na vynikající výsledky úspěšnosti identifikace vyšších taxonomických kategorií prezentované v této dizertační práci je škoda, že tyto dosud autorka ve formě článku v impaktovaném časopise nepublikovala. Z přehledu tvůrčích aktivit doktorandky a z dodaných reprintů konferenčních článků vyplývá, že své výsledky publikovala na renomovaných konferencích a dále výsledky své vědeckovýzkumné činnosti promítla do řady veřejně dostupných softwarových produktů (9 ×, z toho 6× jako první autorka).

4) Vyplývá ze seznamu vědecké činnosti uchazečky, že se jedná o pracovníka s vědeckou erudicí?

Ano, uchazečka publikovala řadu prací na domácích a zahraničních konferencích – nejčastěji jako první autorka. Vytvořila řadu softwarových produktů. Z pozice hlavní řešitelky podala, získala a obhájila grantový projekt inovace výuky v předmětech zaměřených na bioinformační technologie na FEKT VUT v Brně.

V závěru svého oponentského posudku konstatuji, že předložená práce Ing. Denisy Maděránkové splňuje požadavky obvykle kladené na úroveň dizertačních prací. Dizertační práce i přehled tvůrčích aktivit doktorandky vyhovuje obecně uznávaným požadavkům k udělení titulu Ph.D.

Při obhajobě dizertační práce bych rád předložil následující dotazy:

- 1) Ve své práci uvádíte, že numerická reprezentace nukleotidových sekvencí pomocí denzitních vektorů zatím nebyla publikována v relevantní vědecké literatuře a odkazujete se neurčitě na publikace vydané před rokem 1980. Můžete identifikovat práce, ze kterých jste vycházela při popisu výpočtu nukleotidových denzitních vektorů v kapitole 5?

- 2) Jak se liší Vaše vlastní implementace metody nukleotidových denzitních vektorů od dostupné implementace v Bioinformatics Toolbox pro Matlab (např. funkce *ntdensity()*)?
- 3) Vaše komparační analýza spočívá ve vytváření referenčních sekvencí metodou zprůměrování s následným hledáním referenční sekvence, která je od testované sekvence nejbliž ve smyslu Euklidovské vzdálenosti. Tento přístup by se dal označit jako klasifikace pomocí centroidové metody. Jaké výhody či nevýhody byste očekávala od klasifikačního přístupu metodou průměrné vazby?
- 4) Pro vytváření referenčních sekvencí oddělujete ze souboru dat část sekvencí a druhou část používáte pro testování, tj. pro výpočet úspěšnosti identifikace. V práci jsem nevyčetl, podle jakých kritérií soubor rozdělujete? Budou výsledky úspěšnosti identifikace rozdílné, pokud se pro referenční sekvence použije jiná část souboru?

V Brně dne 22.11.2015



Daniel Schwarz



doc. Ing. Daniel Schwarz, Ph.D.

Zástupce ředitele pro vývoj IT
Systémový analytik

+420 549 492 854 | +420 604 996 753

schwarz@iba.muni.cz

Masarykova univerzita, Institut biostatistiky a analýz

Kamenice 126/3, 625 00 Brno

<http://www.iba.muni.cz>