Oponentský posudek diplomové práce Bc. Terezy Čalounové

Předkládaná anglicky psaná práce Bc. Terezy Čalounové si klade cíl definovaný názvem, „Mining novel terpene synthases from large-scale repositories / Mining nových terpen syntáz z rozsáhlých databází“. Autorka v diplomovém projektu pracuje se souhrnnou databází charakterizovaných terpensyntáz, která vznikla na pracovišti školitele a na jejímž vzniku se podílela. Analýzou tohoto datového souboru stanovila kritéria pro vyhledání sekvencí kandidátních terpensyntáz v kompilovaném datasetu, který vychází z veřejně dostupných repozitářů a který zahrnuje více než $5*10^9$ aminokyselinových sekvencí. Výstupem, který autorka dále popisuje a diskutuje, je soubor $6*10^5$ kandidátních terpensyntáz.

Literární přehled (kapitoly 1-3) přehledně zpracovává obecné téma terpensyntáz, postupy používané pro „sequence-guided mining“ a shrnuje dosavadní studie, jejichž cílem bylo podobně jako v autorčině případě hledání sekvencí TPS v datasetech různého původu a rozsahu.

Kapitola 4, „Data and methods“ představuje výchozí databázi dosud charakterizovaných terpensyntáz, databázi proteinových domén Pfam a SUPERFAMILY, jejichž prostřednictvím byla mj. definována kritéria vyhledávání, a konečně jednotlivé databáze proteinů, které byly následně použity pro vytvoření prohledáváného datasetu. Přehledně a reprodukovatelně jsou použité postupy shrnuty také v dedikované složce v repozitáři GitHub.

Kapitola 5, „Results“, stručně komentuje proces filtrování výstupu prohledávání, který vedl k získání konečného výstupu a tento dataset analyzuje s ohledem na zastoupení jednotlivých zdrojových databází mezi nalezenými sekvencemi, jejich délku, taxonomické zařazení a přítomnost jednotlivých Pfam a SUPERFAMILY domén. Sekvenční podobnost mezi jednotlivými kandidáty, resp. jejich zástupci v redukovaném datasetu, je znázorněna fylogenetickým stromem a SSN.

Kapitola 6 diskutuje dosažené výsledky, vyzdvihuje potenciálně zajímavé skupiny identifikovaných kandidátů a poctivě upozorňuje na omezení daná vstupními daty i zvoleným postupem a jmenuje případné důsledky. Samostatná kapitola „Conclusions“ pak celý výstup ve třech odstavcích shrnuje a hodnotí.

Práce má běžné členění, prostor věnovaný jednotlivým částem je adekvátní, celkem má text 109 stran. Literární rešerše je dobře strukturovaná a opírá se o dostatečný počet citovaných zdrojů (odhadem >100). Závěry jednotlivých článků jsou dobře interpretovány v kontextu relevantních studií, v tom text působí skutečně věrohodně. Metodický postup i výsledky jsou dobře zdokumentovány a diskutovány. Jazyková úroveň textu je dobrá, přičemž autorka uvádí následující: “AI tools, grammarly.com, and OpenAI's ChatGPT *(were utilized)* for grammar correction and to enhance readability.” Výhrady ohledně jazyka mám pouze k žargonovitému užití slov monocots, dicots, namísto monocotyledons, dicotyledons.
V českém abstraktu působí nepříjemně vazba “kurátorovaná databáze“, nepřeložený výraz „mining“ a chybně je uváděna „terpen syntáza“ (správně terpensyntáza nebo terpenová syntáza). Chybný je formát psaní vyšších taxonomických jednotek (třídy se píšou bez kurzívy).
V kontextu velmi kvalitního a čtivého textu působí rušivě nesprávné užívání závorek v citacích, např. „(Jia et al. 2016) focused on microbial-like TPSs in plants.“, v citacích v textu

je místy nejednotně uváděna i zkratka křestního jména. Nejednotnost panuje v seznamu citací, kde jsou křestní jména rozepsaná i zkracovaná, místy není dodrženo abecední řazení.

K práci mám několik převážně formálních komentářů a tři otázky:

- Na některé obrázky v textu chybí odkaz (Figure 2, 3, 33), textový odkaz Figure 38 se ve skutečnosti týká Figure 37.
- Ve Figure 2 není vybraný příklad pro TPS II. třídy vhodný, lepší by bylo takové schéma, které končí terpenovým produktem, nikoli pyrofosfátem.
- Figure 3 by zasloužila drobné úpravy místo přímého převzetí: text v obrázku „OILTS this work“ je irelevantní a matoucí.
- Figure 4 je oproti textu, který má ilustrovat, zastaralý, protože je přímo převzatý ze starší publikace. Vhodnější by bylo ve stejném duchu připravit obrázek vlastní, na kterém by byly znázorněny domény a motivy TPS všech aktuálně známých typů.
- Ve Figure 35 v legendě chybí popis barevného kódování větví stromu, zřejmě závisí na tom, zda se jedná o charakterizované TPS, resp. jaký je jejich původ? V anotaci stromu by bylo přehlednější graficky od sebe oddělit poslední mezikruží, které odpovídá Pfam, a první, které odpovídá SUPERFAMILY.
- V kapitole 4.1.2 Protein sequence databases je u Phytozome je uvedena aktuální verze, ze které pocházejí data, u ostatních podobnou specifikaci (nebo datum přístupu?) postrádám.
- Místy má autorka odvážný přístup k fylogenetice, Amoebozoa by neměla být řazena mezi „Animals“, označení kingdom ve vztahu k ruduchám je sporné.
- V grafech ve Figure 9.A a 10 jsou mj. uvedeny taxonomické skupiny Animalia (Coral), Animalia (Insecta), Animalia (Marine Sponge) a Animalia. Které další zástupce zahrnuje skupina Animalia? Podobně u Table 7 – zvlášť uvedeni koráli, hmyz, dále sběrná skupina Animalia.

1. V kapitole 4.1.3 Protein domain databases je popsán způsob ověření vhodnosti vybraných profilových HMM z databází Pfam a SUPERFAMILY („each model's precision was estimated using the SwissProt database protein family annotation“). Databáze SwissProt není na rozdíl od ostatních v textu dříve zmíněna, zejména by se hodilo znát způsob generování jejích anotací. Takto si jako laik kladu otázku, zda nejsou anotace ve SwissProt a databázích propojených s Pfam / SUPERFAMILY do nějaké míry provázané a zda tím není ohrožena relevance kalkulace „precision estimate“. Jak to je?
2. V Table 2 mě zaujalo zařazení Complement components (např. Thio-ester containing domain (TED) from Complement C3, C4adg fragment of complement factor C4a domain) do skupiny Terpenoid cyclases/Protein prenyltransferases. Dá se odhadnout, na základě čeho tyto domény „skončily“ v dané kategorii?
3. V textu zmiňujete jednu z hypotéz o evoluci TPS rostlin, konkrétně Yan et al., 2023 (“…hypothesize that fusion of PF03936 and PF01397 occurred in an ancestral land plant, which likely acquired both domains independently from microbes through horizontal gene transfer.”). Jak se na tuto hypotézu díváte např. v kontextu Vámi také citované práce Jiang et al., 2019 a případnému evolučnímu vztahu k IDS enzymům?

Závěr:

Předložená práce Terezy Čalounové splňuje všechny požadavky kladené na diplomové práce. Autorka si stanovila ambiciózní cíl, který podle mého názoru splnila. Z práce mám radost, představuje užitečný odrazový můstek pro další pátrání po terpensyntázách. Vřele ji tedydoporučuji k obhajobě.

V Praze, 31.5.2024

Mgr. Jitka Štáfková, Ph.D.