Posudek oponenta diplomové práce

Autor práce: **Kateřina Kratochvílová**

Název práce: **Nástroj pro automatickou identifikaci KIR alel**

Obsah práce

Práce se zabývá automatickou identifikací KIR alel na základě analýzy sady nepřesných digitálních záznamů krátkých úseků DNA vzniklých sekvenovacím procesem.

Kvalita řešení a dosažených výsledků

Text předložené diplomové práce je obtížně čitelný. Důvodů je hned několik. Kapitoly 2 a 3 zabíhají zbytečně do detailů genetiky a imunologických procesů, které jednak nejsou příliš relevantní k zadání práce (jedná se např. o popis NK buněk nebo biochemické principy nejrůznějších metod sekvenování) a jednak vyžadují znalosti, kterými běžný čtenář z oblasti informatiky nedisponuje (např. nevím, co je cytoplazmatický ocásek nebo imunoglobulinová doména, nebo jakým způsobem dojde k nastřihání DNA). Definice problému z informatického hlediska není vůbec popsána. Čtenář se jen může domýšlet, co se skrývá za operací „zarovnávání readů". Některé klíčové pojmy nejsou dobře definovány (např. definice na str. 8 zní, že ready jsou „krátké kusy DNA", ale zřejmě se jedná o textová data obsahující informaci o krátkém kusu DNA zapsanou v podobě posloupnosti znaků A, C, G, T) nebo dokonce nejsou vůbec definovány (např. „procentuální šířka zarovnání" – viz str. 40). Větě „Po analyzování všech smazaných alel se ukázalo, že všechny byly smazány kvůli alele, která do genomu také patří." na str. 45 jednoduše nerozumím. Obrázky na str. 45 a dále jsou pro nezasvěceného čtenáře nepochopitelné. Co je na osách x a y? Jaký je rozdíl mezi obrázkem nahoře a dole?

Ačkoliv Kapitola 4 naznačuje, že diplomantka provedla analýzu velkého množství dostupných bioinformatických nástrojů, vlastní popis analýzy v textu práce je strohý a není z něj zřejmé, proč byl pro zarovnávání readů byl vybrán právě nástroj Bowtie2. Zdá se, že jediným rozhodujícím kritériem byla jeho rychlost. Jak si však tento nástroj stojí v porovnání s ostatními s ohledem na kvalitu výsledků? Vyloučení nástroje Segemehl pro jeho paměťovou náročnost považuji za nedomyšlené, protože mnohé stolní počítače v laboratořích mají 64 GB fyzické paměti a rychlý SSD disk, takže požadavek na „až 70GB paměti" mi nepřijde až tak limitující.

Postrádám rovněž zdůvodnění výběru metriky pro měření vzdáleností mezi sekvencemi, což je, spolu s volbou algoritmu pro zarovnávání readů (Bowtie2), klíčovým prvkem všech navrhovaných přístupů pro identifikaci KIR alel. Je skutečně Levenhsteinova vzdálenost nejvhodnější v uvažovaném kontextu? Na str. 28 je tvrzení, že při sekvenování zvolenou metodou (Illumina) dochází typicky pouze k substitučním chybám. Domnívám se proto, že použití např. Hammingovy vzdálenosti by bylo vhodnější. Alternativně lze užít modifikovanou Levenhsteinovu vzdálenost s různou cenou jednotlivých operací (INS, DEL, SUBS).

Diplomantka navrhla celkem 4 různé přístupy pro identifikaci KIR alel, z toho 3 naimplementovala a otestovala. Navržené přístupy považuji za rozumné. V popisu nicméně chybí zdůvodnění voleb nejrůznějších prahových hodnot, např. vzdálenost menší než 100 na str. 44, pokrytí menší než 90 %, nebo vzdálenost 30 pro shlukování na str. 50.

Návrh a implementace celé procesní pipeline v jazyce Python je adekvátní vzhledem k úloze. Zdrojové kódy jsou vhodně rozděleny do jednotlivých souborů a metod a vhodně okomentovány. Příručka pro uživatele obsahuje dostatek informací, aby i neproškolený uživatel mohl nástroj nakonfigurovat (dle potřeby) a spustit.

Z příloh je patrné, že diplomantka provedla značné množství experimentů. Je škoda, že v textu práce není jejich kritické zhodnocení provedeno na úrovni, kterou bych očekával.

Formální úroveň

Text práce je vhodně strukturován do kapitol a podkapitol, nicméně pořadí předkládaných informací není vždy logické (např. Kapitola 3 obsahuje v prvním odstavci informaci, jejíž znalost je apriori již vyžadována v kapitole 1) a jednotlivé podkapitoly na sebe často plynule nenavazují (viz např. podkapitolu 4.2.1 a 4.2.2 nebo také podkapitolu 5.5 a kapitolu 6). Připomínku bych měl rovněž k abstraktu, který není příliš informačně zajímavý.

V textu práce jsou dále hojně využívány nejrůznější zkratky, přičemž mnohé nejsou ani předem definovány. Domnívám se, že by bývalo pomohlo předsunout seznam zkratek uvedený na samém konci textu práce na jeho začátek nebo alespoň na existenci upozornit čtenáře hned v úvodu.

Počet překlepů a obdobných chyb je velmi nízký (např. „Následujcí" na str. 55, chybějící tečka na konci věty na str. 17, nebo chybný odkaz v popisu Obrázku 2.6). Za drobný prohřešek považuji používání anglických slov namísto běžně používaného českého ekvivalentu – viz zejména „clustery" namísto „shluky" na str. 50.

Po typografické stránce lze práci vytknout umístění popisků k tabulkám na nové stránce namísto toho, aby byl popisek spolu s tabulkou (viz str. 55) a použití příliš malého fontu u obrázků v kapitole 5.2 jsou příliš malé.

Práce s literaturou

Práci s literaturou lze stručně zhodnotit za přiměřenou a odpovídající typu práce. Diplomantka čerpala informace z celkem 41 pro práci relevantních zdrojů. Poněkud neobvyklý však je formát jejich citací a způsob, jakým jsou v textu odkazovány – na mnoha místech jsou odkazy samostatně mimo jakoukoliv větu až na konci odstavce (viz např. str. 8), takže není vždy jednoznačné, k čemu se vlastně vztahují. Tvrzení na str. 19 „podle některých studií …" postrádá odkaz na tyto citace. Obdobně na str. 30 z věty „proto byl algoritmus rozšířen, jak je popsáno dále" není zřejmé, kým byl rozšířen a lze jen odhadovat, že autory článku [27].

Splnění zadání

Zadání práce považuji za splněné s výhradami ke kvalitě řešení (viz výše).

Dotazy k práci

Diplomantka by se měla v průběhu obhajoby vyjádřit k následujícím otázkám:

- *Jaký vliv by mělo použití jiného nástroje pro zarovnávání „readů" a jiné metriky vzdálenosti na přesnost identifikace KIR alel?*
- *Na str. 57 se nachází tvrzení: „Jak je vidět z obrázku, u genu 2DS1 je možné určit coding region tzn. prvních pět čísel. Oproti tomu v případě genu 2DL4 to možné není". Nic takového tam nevidím. Co se na obrázku vizualizuje? Jak to správně interpretovat?*
- *Jaký nástroj/algoritmus byste doporučila pro bezpečnou identifikaci KIR alel?*

Závěrečné shrnutí

Protože diplomantka ve své práci jednoznačně prokázala, že je schopna provést samostatně návrh, realizaci a ověření inženýrského řešení, ***práci doporučuji k obhajobě****. S ohledem na kvalitu řešení a výsledků práce ji, s přihlédnutím k rozsahu, navrhuji hodnotit známkou* ***dobře****.*

V Plzni dne 20. 8. 2020

Doc. Ing. Josef Kohout, PhD.
KIV-FAV-ZČU