Posudek oponenta bakalářské práce

Autor práce: **David STANÍČEK**

Název práce: **Automatické měření metrik NGS zarovnávacích nástrojů**

Jazyková a grafická úprava

Průměrné

Formální a obsahová stránka práce

Průměrné

Vhodnost použitých metod

Průměrné

Způsob zpracování a vyhodnocení

Průměrné

Správnost získaných výsledků

Průměrné

Vlastní přínos

Průměrné

Doplnění hodnocení, připomínky:

BP se zabývá definicí metrik pro porovnání NGS (Next-Generation Sequencing) zarovnávacích nástrojů, implementaci nástroje pro automatické hodnocení definovaných metrik a zhodnocením vybraných nástrojů. V úvodní části práce se autor věnuje popisu DNA dat a metodám jejich získávání. Jsou zde popsány přístupy pro získávání DNA dat 1. generace a primárně přístupy pro získávání NGS dat. Dále je zde seznámení s formáty dat sloužících k uložení DNA dat a obecný postup zpracování dat. Další část je věnována zarovnávání NGS dat. Je zde popsán problém zarovnání dat na referenční sekvenci a jednotlivé zarovnávací nástroje. V následující části se autor zabývá výběrem metrik pro hodnocení zarovnávacích nástrojů. Jsou zde popsány jednotlivé metriky a zdůvodnění jejich výběru. Následuje část návrhu nástroje pro automatické hodnocení. Autor pro implementaci zvolil programovací jazyk Python. Popisuje zde generování syntetických dat, knihovny použité pro práci s DNA daty a nástroj pro předzpracování vstupních dat. Dále je zde popsán postup měření metrik a vizualizace výsledků. Závěr práce je věnován zhodnocení vybraných nástrojů. V této části autor zdůvodňuje výběr zarovnávacích nástrojů, volbu testovacích dat a provedené experimenty pro účely zhodnocení nástrojů. K práci mám několik připomínek:

1. V práci se vyskytují obrázky, které nikde nejsou z textu odkazovány ani vysvětleny – např.: obrázek 2 a 3 v sekci 2.2 nebo obrázek 8 v sekci 3.3.1.
2. Dále jsou v práci použité pojmy, které nejsou v textu vysvětleny – např.: „read“, „seedy“, „full-text minute index“.
3. Některé sekce postrádají zdůvodnění / závěr, např.: sekce 4.1 Hammingova vzdálenost, kde je napsáno: „Tato metrika není příliš vhodná pro účely porovnávání genomických sekvencí, což lze demonstrovat na následujícím příkladu porovnání dvou sekvencí“, je zde uveden příklad, ale nikde není vysvětleno, co je špatně a proč metrika není vhodná.
4. Každý experiment byl proveden pouze dvakrát a byl hodnocen jako samostatná iterace. Chybí zde statistické zhodnocení přes více iterací.
5. Výsledný nástroj pro automatické hodnocení spouští každý zarovnávací nástroj s jiným počtem výpočetních vláken (Bowtie: 2, BWA: 1, Gem3: všechna dostupná, Kart: 4). Výsledné časové náročnosti tak neodpovídají skutečnosti.

Dotazy

1. Z jakého důvodu nebyl v testovaných nástrojích zahrnut BWA-MEM2, ale pouze jeho starší verze?
2. Proč byly nástroje spouštěné s různým počtem výpočetních vláken? Jak by výsledky ovlivnilo spuštění se stejným počtem vláken?

Splnění bodů zadání

úplně

Doporučení k obhajobě

ANO

Hodnocení: 2 - Velmi dobře

V ---------- dne ---------- Ing. Filip Jani