Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Nikola Kalábová
Název práce	Evolutionary Algorithms for Multi-Stage Transcriptomic Data Analysis
Rok odevzdání	2024
Studijní program	Informatika **Studijní obor** Teoretická informatika

Autor posudku	Michal Kolář **Role** Oponent
Pracoviště	Ústav molekulární genetiky AV ČR

Text posudku:

Zadání:
Cílem diplomové práce bylo zjistit, zda odstranění závislosti TAI (transcriptome age index) na vývojovém stádiu pomocí odebrání malého počtu genů z výpočtu TAI lze použít k identifikaci genů, které jsou důležité v ontogenezi. Tento přístup vychází z následujících myšlenek:

i) Závislost TAI na vývojovém stádiu odráží vztah mezi evolučním stářím genu a vývojovým stádiem (ontogeneze opakuje fylogenezi).
ii) Tuto závislost lze vyrušit odstraněním genů, které jsou spojeny s vývojem, z výpočtu TAI.

K identifikaci takových genů byl vyvinut a použit algoritmus, který se snaží minimalizovat varianci hodnot TAI v průběhu ontogeneze a zároveň minimalizovat počet odebraných genů. Algoritmus je založený na evolučním programování.

Hodnocení:
Myšlenka využití TAI k identifikaci genů důležitých pro vývoj organismu je originální a mohla by pomoci nalézt zatím nepopsané geny spojené s vývojem jedince. Studentka prokázala dobré porozumění komplexnímu problému a vytvořila funkční algoritmus. Implementace algoritmu byla provedena na vysoké úrovni a základy, na kterých je založen, jsou v práci dobře vysvětleny. Bohužel, při aplikaci algoritmu a diskuzi výsledků již pravděpodobně nezbýval čas a práce proto jako celek působí uspěchaně a výsledky nejsou přesvědčivé.

Práce je napsána velmi pěknou angličtinou, což považuji za jednoznačné pozitivum. Je standardně strukturována, i když s nedostatky, které uvádím níže. Literární rešerše je plně dostačující. Úvod práce je dobře napsán a pokrývá základní myšlenky a teoretický kontext. Drobnými nedostatky jsou překlepy a nepřesnosti, například použití termínu „filtering“ místo „selection“ pro výběr specifických typů RNA. Úvod je však promíchán s metodikou, což znesnadňuje čtenáři rozlišit mezi stávajícím poznáním a popisem metod použitých v práci. Některé technické detaily jsou také nejasné, například rovnice (1.1) která uvádí součet přes "gene set", který ale není nikde definován. Trochu nešťastně je obrázek 1.4 ukázán pro D. rerio a 1.5 pro A. thaliana.

Návrh a implementace algoritmu jsou silnou stránkou této práce. Studentka správně identifikovala potřebu minimalizovat varianci TAI během ontogeneze a vyvinula metodu, která se snaží minimalizovat počet odebraných genů. Překlad biologického problému do

algoritmického přístupu je proveden pečlivě, i když rovnice (4.1) je zbytečně komplikovaná, chybí v ní vysvětlení proměnné S (množina vývojových stádií?). Lépe by bylo definovat pouze TAI_{s,I}, které je vzhledem k rovnici (4.2) potřebnější.

V kapitole „Data“ je sice specifikován typ vstupních dat, ale popis výpočtu phylostrat a jejich transformace je velmi obecný. Metodika pro normalizaci transkriptomických dat je uvedena pouze pro RNA-seq (TPM), ale není zmíněna pro čipové experimenty. Navíc vzorová data na GitHubu (D. melanogaster) nevypadají jako TPM hodnoty. Odstavec věnující se předzpracování dat je velmi obecný, i když to v transkriptomice bývá jeden z nejdůležitějších kroků a to především vzhledem k semikvantitativní povaze dat. Při testování softwaru jsem zjistil, že vhodné předzpracování dat může výrazně zlepšit výsledky. Bez správné normalizace dat dochází k preferenci genů s vysokou expresí na úkor transkripčních faktorů, které hrají klíčovou roli ve vývoji.

Výsledková část je promíchána s diskuzí a postrádá jasné oddělení interpretace od popisu výsledků. Algoritmus byl aplikován na dvanáct datasetů, avšak chybí zlatý standard nebo syntetický dataset, na kterém by bylo jasné, jakou mají výsledky kvalitu. Diskuze se soustředí především na stabilitu řešení a jeho významnost oproti náhodně zvoleným genům, zatímco biologická významnost je zanedbána. Funkční anotace identifikovaných genů byla provedena pouze povrchně a nebyly použity metody jako gene set enrichment analysis, které by mohly výrazně pomoci při interpretaci výsledků.

Mezi drobné nedostatky považuji použití částí pythonovských skriptů v osmé kapitole. Vhodnější by bylo použít pseudokódy. V různých částech textu se zmiňuje nastavení parametrů na specifické hodnoty. Nejsem si jistý jaká kritéria byla použita pro jejich volbu. Odstavec 8.5 o postprocesingu je nadbytečný; vhodný preprocessing dat by dokázal více. V příloze jsou uvedeny identifikátory nalezených genů. Zde by bylo mnohem lepší dodat do tabulky i jména a popis genů.

Při aplikaci na experimentální data mi přijde nevhodný výběr míry exprese genů, což negativně ovlivňuje výsledky. Diskuze vybraných genů kladně hodnotí, že mnoho z nich nemá funkční anotaci. To mi přijde zavádějící. Nepřítomnost známých transkripčních faktorů je spíš signálem, že výsledky nejsou správné. Vhodná normalizace dat může tento problém napravit.

V případě single cell dat není zřejmé, co se myslí diferenciačním stavem anebo vývojovým stádiem. Je to populace buněk v scRNA-seq experimentu? Nebo jednotlivá buňka? V prvním případě není uvedeno, jak jsou populace identifikovány, v druhém případě budou různě abundantní populace buněk reprezentovány různým počtem replikátů. Diskuze se zde redukuje pouze na komentování výpočetní komplexity algoritmu.

Závěr:
Tato diplomová práce obsahuje zajímavé myšlenky a slibný algoritmus, ale její aplikace a diskuze výsledků nenaplňují očekávání. Jako celek působí uspěchaně a výstupy nejsou dostatečně přesvědčivé. Přestože studentka odvedla dobrou práci při návrhu a implementaci algoritmu, nedostatky ve zpracování dat a interpretaci výsledků snižují celkovou hodnotu textu. Ten i tak jednoznačně splňuje nároky na diplomovou práci a rád jej doporučuji k obhajobě.

Otázky k obhajobě:
1) Jaké kroky v předzpracování dat by podle vás byly nejvhodnější pro zlepšení výsledků algoritmu (například filtr genů, normalizace na hloubku sekvenování, stabilizace variance, kvantilová normalizace mezi vývojovými stádii, škálování hodnot)? Můžete prosím komentovat jejich vliv na výsledky?

2) Jak se váš přístup vypořádává s paralogními geny (například homeoboxové geny)? Jak se v tomto případě vyhodnocuje evoluční stáří genů?

3) Závislost TAI na vývojovém stádiu není monotónní. Můžete prosím tuto skutečnost komentovat a popsat, jaké to má důsledky pro váš algoritmus?

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 30.8.2024 **Podpis**