Posudek diplomové práce

Bc. Vojtěch Tuma: Sumarizace genových expresních čipů z volně žijících druhů

Práce aplikuje znalosti informatiky a strojového učení na konkrétní problém sumarizace genových expresních čipů.

Práce uvádí čtenáře do bioinformatiky, expresních čipů, jejich použití a sumarizace. Autor popisuje metody sumarizace, hypotézu rozdílných výsledků pro volně žijící poddruhy myší a pro laboratorní poddruhy pro které byly čipy vytvořeny. Navrhuje použít znalost o rozdílnosti poddruhů (SNP mutace), vynechat a)zasažené části genové informace b)náhodné části a porovnat výsledky sumarizace.

Nejvýznamnější výhradou je měření vlivu jen na jediné sumární hodnotě (účinku v BEST). Nepovažuji sílu signálu za jediný a nejlepší cíl úpravy sumarizace. Vliv úpravy je třeba měřit na hodnotách použitých při dalším vyhodnocení dat. Např. častým použitím čipu je zjištění exprese xenou (genu, proby). Na úrovni prób by šlo o klasifikaci podle DABG<0.05. Pro konkrétní příklad (core_snp/nosnp, M_KID_ITER_PLIER) mi vychází přes 5% rozdílu v klasifikaci.

Na výsledcích prezentovaných metodou BEST mě zaráží vliv odstranění náhodných prób na účinek, nejspíš i průměrnou hodnotu signálu. Bez jasného vysvětlení budí podezření o chybě v postupu. Také mi není jasný rozdílný počet prób u všech tří metod – snp, nosnp, random. Proč se u nosnp a random liší? Jak byly voleny random?

Kód na CD sice dokumentuje práci studenta, ale není připraven byť jen na testovací spuštění: nenašla jsem datové soubory CLF, PGF, musím přepisovat jména adresářů i souborů (source('exon_assignment_**new**.R') v mcmc.R).

Chybí informace o výpočetním čase, prostoru potřebném v MetaCentru.

Grafy jsou graficky pěkně zpracované, většinou srozumitelné, někdy až příliš zaměřené na ne-odborníky ve statistice/strojovém učení (Obr. 4.1. vysvětlení odlehlých hodnot). U obrázků 6.1,6.2 bych uvítala zřetelněji znázorněný poměr zasažených prób ku ne-zasaženým, uvedení počtu xeonů kde dojde k vyřazení všech prób, kde zbyde jen jedna próba. Na houslových grafech (6.6, 6.7.) by bylo vhodnější zobrazit i 0 na y-ové ose, byl by zřetelnější poměr účinku odstranění zasažených a náhodných prób.

Student naprogramoval potřebné skripty k analýze výstupu expresních čipů v MetaCentru a navrhl metodu úpravy sumarizace pro použití vzorků z nemodelových organizmů. Vzhledem ke složitosti vniknutí do problematiky bioinformatiky jsem ochotna pominout nedotaženost po stránce informatické.

Práci doporučuji uznat jako diplomovou.

Mgr. Marta Vomlelová PhD.
KTIML

V Praze 10.6.2016