

Posudek vedoucího diplomové práce

Vedoucí práce: Petr Daněček

Autor práce: Martin Dráb

Název práce: Variant calling using local reference-helped assemblies

Určování genetických variant (variant calling) z NGS sekvenačních dat se provádí pomocí namapování a zarovnání mnoha krátkých čtení (short reads) oproti sekvenci referenčního genomu a následného vyhodnocení rozdílů pomocí statistického modelu. Na první pohled snadná úloha je komplikovaná tím, že se vinou nedokonalých experimentálních technologií v datech vyskytuje velké množství náhodných i systematických chyb, které se mohou jevit jako falešné jedno- či vícebázové varianty (SNPs a indels).

Náhodné sekvenační chyby lze do značné míry potlačit již na úrovni experimentu hlubším prosekvencováním, tj. redundantním čtením při kterém je každá oblast genomu přečtena opakováně z více nezávislých kopií. Větší problém představují systematické chyby způsobené nesprávným mapováním a zarovnáním. Mapovací algoritmy mohou umístit krátká čtení na nesprávné místo a/nebo je nesprávně zarovnat, v datech následně vzniká falešný signál obtížně odlišitelný od skutečných genetických variant. Předložená diplomová práce se zabývá právě tímto aspektem a novou metodou, založenou na De Bruijnových grafech, se v problematických oblastech snaží rekonstruovat původní sekvenci.

V práci bylo nutné proniknout do problematiky variant calling a assemblies, porozumět sekvenačním technologiím a používaným datovým formátům; těmto aspektům se věnuje první část. Druhá část se podrobně věnuje nově navrženému algoritmu, včetně popisu parametrů nezbytných pro reprodukování dosažených výsledků. Je popsána úsporná reprezentace datových struktur, konstrukce modifikovaného De Bruijnova grafu a jeho následná optimalizace, cílená na odstranění podgrafů generovaných sekvenačními chybami. V závěrečném kroku algoritmus detekuje varianty a na základě pravděpodobnostního modelu stanovuje odhad spolehlivosti. Problematické korekce chyb na základě rozboru četnosti k-merů je věnovaná samostatná kapitola.

Úskalí zadané práce spočívalo v principiální neřešitelnosti problému - hledané řešení není exaktní ale spadá do oblasti heuristických přístupů. Bylo nutné analyzovat značné množství skutečných biologických dat a navrhnut dostatečně robustní metodu při zachování rozumné citlivosti. V porovnání s naivním přístupem (mpileup) i s alternativním mikro-assembly přístupem (fermikit) dosáhla nově vypracovaná metoda velice dobrých výsledků. Práce byla zhotovena zcela samostatně a z konzultací i práce samotné je zřejmé, že jí bylo věnováno mnoho úsilí.

Doporučení k obhajobě:

Z výše uvedených důvodu práci doporučuji k obhajobě.

