OPTIMIZATION OF THE NEXT-GENERATION SEQUENCING DATA ALIGNMENT

Vojtěch Šalanda

Master Degree Programme (2), FIT BUT

E-mail: xsalan01@stud.fit.vutbr.cz

Supervised by: Ivan Vogel

E-mail: ivogel@fit.vutbr.cz

Abstract: Next-generation sequencing is nowadays a widely used technology as it allows us to sample genome and transcriptome in a cost efficient manner. This paper is focused on optimizing short reads alignment to the reference sequence. Main targets of the optimization are speed and accuracy of the alignment. There were assembled suitable parameters for two selected alignment tools to be optimized using differential evolution. The experimental results will be validated on both software-generated and real data.

Keywords: next-generation sequencing, alignment tools, sequence read, alignment optimization

1. ÚVOD

Součástí molekulárně genetických metod analýzy biologického materiálu je sekvenování DNA. Tímto pojmem se označuje proces čtení vzorku DNA, jehož výsledkem je posloupnost nukleotidů, které tvoří primární strukturu DNA.

Současnou dominantu v oblasti sekvenačních technik tvoří sekvenace nové generace (NGS) [1]. Zavedením těchto metod výrazně poklesly náklady a vznikly nové možnosti zkoumání nukleotidových sekvencí uvnitř buňky. Společným rysem je masivní paralelismus při sekvenování krátkých sekvenčních čtení (*reads*). Jednotlivé metody se liší přesností a délkou výsledných sekvenčních čtení.

Navazujícím krokem je zarovnání výsledných čtení k referenční sekvenci nebo sestavení původní sekvence *de novo*. Pro samotné zarovnání lze použít některý z 60 různých bioinformatických nástrojů, které se liší použitými algoritmy [2]. Výsledky jsou různě přesné v závislosti na nastavení parametrů běhu.

Za tímto účelem byl navržen optimalizační proces hledání vhodných parametrů, aby bylo výsledné zarovnání co nejpřesnější a nedošlo k výraznému zpomalení běhu nástroje. Při optimalizaci se využívá evolučních algoritmů.

2. VÝBĚR MAPOVACÍCH NÁSTROJŮ

Mapovací nástroje lze podle použitého algoritmu rozdělit do dvou kategorií: založené na hashovací tabulce a založené na sufixových stromech. Z první kategorie byly k optimalizaci vybrány nástroje BLAT [3] a LAST [4]. K jednotlivým nástrojům byly sestaveny seznamy všech parametrů, ze kterých byly vybrány konkrétní parametry pro optimalizaci. Jedná se zejména o parametry ovlivňující přesnost mapování semínka (*seed*) a následně celého sekvenčního čtení. Parametry zaměřené na penalizaci mezer nachází uplatnění zejména při zarovnání nasekvenovaného transkriptomu, kde dochází k výskytu velkých mezer v důsledku sestřihu (*splicing*).

Program LAST je nástrojem v první skupině algoritmů. Princip zarovnání sekvenčního čtení k referenční sekvenci spočívá v prvotním zarovnání malé části sekvence, která se označuje jako semínko. V druhé fázi dochází k rozšiřování zarovnání po celé délce zarovnávaného čtení. Hlavní výhodou tohoto nástroje je možnost použití prostorově rozšířeného semínka. Toto semínko je

reprezentováno binární šablonou, kde „1“ odpovídá přesné shodě v místě zarovnání. V případě „0“ není třeba přesné shody. Autor publikuje optimální binární šablony pro délku sekvenčních čtení max. 50 bází, přičemž dnešní přístroje produkují mnohem delší čtení.

V případě programu BLAT se jedná o stejný princip zarovnání jako u programu LAST. Přístup k prostorově rozšířenému semínku je však odlišný v tom, že není využito přesné binární šablony, ale na délce *N* celého semínka je povoleno *k* neshod při počátečním zarovnání.

3. OPTIMALIZACE PARAMETRŮ

Při výběru parametrů k optimalizaci byly brány v úvahu zejména ty parametry, u kterých se předpokládá největší vliv na rychlost a citlivost zarovnání. Vybrané parametry obou nástrojů LAST a BLAT jsou v následující tabulce:

Nástroj	Parametr	Popis
LAST	-m	Binární šablona prostorově rozšířeného semínka
	-w	Délka kroku při mapování semínka
	-p	Skórovací matice (ohodnocuje každý možný nukleotidový pár)
	-l	Délka počáteční shody (binární šablona se cyklicky opakuje)
BLAT	-tileSize	Celková délka zarovnávaného semínka
	-oneOff	Počet povolených neshod
	-stepSize	Délka kroku při mapování semínka

Tabulka 3.1: Vybrané parametry k optimalizaci.

V případě programu BLAT jsou parametry podobné, ale prostorově rozšířené semínko je specifikováno počtem povolených neshod.

Tyto parametry budou zakódovány do chromozomů a optimalizovány pomocí diferenciální evoluce.

4. DIFERENCIÁLNÍ EVOLUCE

Diferenciální evoluce je globální optimalizační algoritmus, který je součástí evolučních technik. Princip spočívá v zachování populace kandidátních řešení, která je iterativně vystavena změnám v důsledku selekce, evaluace a rekombinace [5]. Pseudokód algoritmu je následující:

Vstup: Population_{size}, Problem_{size}, $\text{Weighting}_{factor}$, $\text{Rekombination}_{rate}$

Výstup: S_{best}

Postup:
```
Population = InitPopulation(Population_size, Problem_size)
Evaluate(Population)
S_best = GetTheBestOne(Population)
while(!StopCondition())
    NewPopulation()
    for (P_i in Population)
        S_i = NewSample(P_i,Population, Problem_size, Weighting_factor, Rekombination_rate)
        NewPopulation.Insert(Better(S_i, P_i))
    end
    Population = NewPopulation
    EvaluatePopulation(Population)
    S_best = GetTheBestOne(Population)
end
return S_best
```

Vznik nového kandidátního řešení z jedinců populace se nazývá perturbace. Perturbovaný jedinec vznikne ze třech náhodně vybraných jedinců podle následujícího vzorce:

$$S_i = P_3 + (P_1 - P_2) * \text{Weighting}_{factor} \quad (1)$$

Do nové populace je podle algoritmu vybráno vždy lepší z původního a perturbovaného kandidátního řešení. Ukončující podmínkou je stav, kdy je celá populace složená ze stejných jedinců, nebo je počet generací evolučního algoritmu předem omezen.

Fitness funkce je vytvořena pomocí programu RABEMA, který porovnává zkoumané zarovnání vůči referenčnímu. Zdrojová data jsou generována pomocí simulátoru sekvenačního přístroje Illumina a výsledky z experimentů budou porovnány s výsledky dosaženými na reálných datech.

5. EXPERIMENTY

V případě programu BLAT byla pro experimenty použita referenční sekvence délky 100 kilobází, ze které bylo vygenerováno desetinásobné pokrytí sekvenčními čteními. Přesnost zarovnání, která byla s implicitním nastavením parametrů 69,7 %, vzrostla díky použití evolučně získaných parametrů na 83,2 %. Díky distribuované výpočetní infrastruktuře MetaCentra lze paralelně provádět ohodnocení kandidátních řešení pomocí spuštění instance programu BLAT se zkoumanými parametry. V poslední 20. populaci 15 kandidátních řešení, které se liší pouze hodnotou posledního parametru –oneOff, je z kandidátů vybrána nejrychlejší varianta.

Při experimentech jsou používána jednoduchá (*single-end*) sekvenční čtení. Součástí dalších postupů bude aplikace párových (*pair-end*) sekvenčních čtení, která nesou informaci o vzdálenosti dvou čtení v páru, optimalizace parametrů programu LAST a aplikace na reálných datech.

6. ZÁVĚR

Na základě provedených experimentů lze očekávat, že použitá evoluce bude konvergovat k optimálnímu nastavení parametrů, které bude srovnáno s jejich implicitním nastavením. Hlavním kritériem je přesnost zarovnání a maximální rychlost při zachování této přesnosti. Optimální řešení může při aplikaci na reálná data vykazovat mírné odlišnosti.

PODĚKOVÁNÍ

Pro provádění experimentů byla využita distribuovaná výpočetní infrastruktura MetaCentra (projekt LM2010005). Tato práce byla podpořena projektem Výzkum pokročilých metod ICT a jejich aplikace (FIT-S-14-2299).

REFERENCE

[1] Metzker, M. L.: Sequencing technologies – the next generation. *Nature Reviews Genetics*, ročník 11, č. 1, 2009: s. 31-46

[2] Li, H.; Homer, N.: A survey of sequence alignment algorithms for next-generation sequencing. *Bioinformatics*, ročník 11, č. 5, 2010: s. 473-483

[3] Kent, W. J.: BLAT-the BLAST-like alignment tool. *Genome research*, ročník 12, č. 4, 2002: s. 656-664

[4] Kielbasa S. M. a kol.: Adaptive seeds tame genomic sequence comparison. *Genome research*, ročník 21, č. 3, 2011: s. 487-493

[5] Brownlee, J.: Clever Algorithms: Nature-Inspired Programming Recipes. Lulu, 2012, ISBN 978-1-4467-8506-5