Abstrakt

Zvyšujúca sa dostupnosť príbuzných genomických údajov si vyžaduje zmenu paradigmy od bioinformatických analýz založených na jednoduchých lineárnych referenčných genómoch k analýzam využívajúcim komplexné pangenomické referencie. Pangenomické sady dát možno charakterizovať dvoma primárnymi atribútmi: ich obrovskými veľkosťami a vysokou repetitívnosťou. Táto repetitívnosť umožňuje konštrukciu pangenomických referencií znížením veľkostí dát prostredníctvom dvoch vynárajúcich sa prístupov: pangenomických grafov a komprimovaných dátových štruktúr na reťazcoch. Zatiaľ čo pangenomické grafy kombinujú podobné genomické oblasti do jednotlivých vrcholov grafu, stringologické dátové štruktúry využívajú kompresné techniky priamo na textovej reprezentácii pangenómu. Každý z týchto prístupov ponúka odlišné výpočtové, vizualizačné a interpretačné výhody. V tejto práci navrhujeme nové dátové štruktúry, ktoré premosťujú tieto prístupy a umožňujú prechod medzi grafovými a stringologickými reprezentáciami.

Centrálna dátová štruktúra, predstavená v kapitole 3, sa nazýva bezprefixový graf. Tento graf ponúka škálovateľný konštrukčný algoritmus, ktorý sa vyhýba výpočtovo nákladným krokom, ako je viacnásobné zarovnanie sekvencií, pričom je dostatočne flexibilný na vytvorenie grafov podobných ľubovoľným iným pangenomickým grafom. Kapitola 4 predstavuje PFG pozície, ktoré umožňujú iteráciu cez sufixy pangenómu, a tak spája bezprefixové grafy a dátové štruktúry na reťazcoch. Kapitola 5 demonštruje praktickú aplikáciu tohto spojenia zlepšením škálovateľnosti konštrukcie Wheelerových grafov, ktoré predstavujú súčasné zovšeobecnenie Burrows-Wheelerovej transformácie, čo potenciálne odomyká možnosť ich použitia ako pangenomický index pre veľké súbory ako napríklad v projekte 1000 Genomes. Nakoniec kapitola 6 predstavuje pole tagov, novú dátovú štruktúru, ktorá umožňuje preniesť výsledky stringologických metód z komprimovaného priestoru na ľubovoľný graf, pričom sa vyhýba potenciálne veľkému nárastu v počte výsledkov počas dekompresie.

Cieľom týchto nových dátových štruktúr je zjednodušiť integráciu pangenomických reprezentácií, podporiť vývoj efektívnych nástrojov na pangenomickú analýzu a interpretáciu výsledkov a v konečnom dôsledku podporiť širšie prijatie pangenomiky v biologickom výskume.

Kľúčové slová: pangenomika, pangenomické grafy, stringológia, dátové štruktúry, bioinformatika