Abstrakt

Dizertačná práca sa venuje problémom na viacerých úrovniach bioinformatickej analýzy, od mapovania čítaní cez genotypizáciu tandemových opakovaní po spracovanie na úrovni vzorky populácie.

Pri riešení problému mapovania čítaní sme sa zameriali na krátke čítania a indexové dátové štruktúry. Navrhli sme dva nové varianty FM-indexu, kde cieľom variantov bolo optimalizovať využitie vyrovnávacej pamäte pri dopytovaní sa indexu. Myšlienka spočívala v preusporiadaní dátových štruktúr FM-indexu do agregovaných nezávislých blokov, kde pamäťovo efektívne uloženie daných štruktúr bolo dosiahnuté vďaka zameraniu sa výlučne na DNA, resp. RNA abecedu.

Pri genotypizácii tandemových opakovaní sme navrhli nový nástroj - WarpSTR - charakterizujúci tandemové opakovanie priamo zo surových signálov získaných nanopórovým sekvenovaním. WarpSTR adresuje výzvy spracovania takýchto dát - vysoká miera šumu a distorzia v časovej oblasti - reprezentáciou štruktúry tandemového opakovania v konečnom stavovom automate a jeho následným zarovnaním so surovým signálom. Pre finálnu charakterizáciu tandemového opakovania sme navrhli heuristiku založenú na Gaussovských zmiešaných modelov. Prínosy nástroja WarpSTR sú dvojaké. Nielenže sme zlepšili presnosť charakterizácie pre jednoduché opakovania, ale umožnili sme spracovať aj komplexné tandemové opakovania.

V prípade bioinformatickej analýzy na úrovni populácie sme sa venovali detekcii epistáz, teda kombinácií variantov, ktorých len vzájomná interakcia má vplyv na fenotyp. Keďže na efektívne prehľadanie priestoru všetkých kombinácií sa väčšinou používajú prírodou inšpirované algoritmy, v práci sme experimentovali s netopierím algoritmom a algoritmom opeľovania kvetov, využívajúc viacúčelovú optimalizáciu.

Kľúčové slová: mapovanie čítaní, zarovnanie, epistázy, tandemové opakovania, nanopórové sekvenovanie