Abstrakt

V tejto práci sme študovali dva výpočtové problémy v oblasti štatistickej analýzy genomických dát. Prvý z nich je odhad proporcií variantov koronavirusu SARS-CoV-2 v zmiešanej sekvenačnej vzorke. Takéto odhady umožnujú monitorovať varianty cirkulujúce v danej komunite pomocou sekvenovania odpadových vôd. Navrhli sme zmiešaný model sekvenačného procesu (vrátane sekvenačných chýb), používajúc frekvencie mutácií vybraných variantov. Návrh modelu umožnuje paralelný výpočet odhadov proporcií. Presnosť odhadov sme overili na simulovaných dátach, laboratórnej zmesi a vzorkách z odpadových vôd z Francúzska a Slovenska. Druhý problém je výpočet štatistickej významnosti počtu prekryvov medzi dvomi genómovými anotáciami (tzv. kolokalizačná analýza). Genómová anotácia je množina intervalov na genóme, napr. gény alebo teloméry. Najprv sme ukázali, že jedna z často používaných definícií nulovej hypotézy vedie k $\mathcal{NP}$-ťažkému výpočtovému problému. Ďalej sme navrhli schému na priame vzorkovanie anotácií z rozdelenia, vyplývajúceho z danej definície. Hlavným prínosom je preformulovanie nulovej hypotézy pomocou Markovových reťazcov a návrh plne polynomiálneho algorithmu na výpočet p-hodnôt pre túto nulovú hypotézu. Časová zložitosť daného algorithmu je nezávislá od dĺžky genómu a jednotlivých intervalov. Toto umožňuje analyzovať anotácie s desiatkami tísic prvkami (napr. všetky exóny v ľudskom genóme) na bežnom počítači v rozumnom časovom rozmedzí. Efektivitu a presnosť nášho prístupu sme experimentálne overili na simulovaných aj reálnych dátach.

Kľúčové slová: SARS-CoV-2, varianty, sekvenovanie odpadových vôd, genómová anotácia, kolokalizačná analýza, Markovove reťazce