POSUDEK DISERTAČNÍ PRÁCE

Autor: Ing. Janka Puterová
Název práce: Detekce repetitivních sekvencí v genomech

Recenzent: doc. Ing. Jiří Kléma, Ph.D.
Katedra počítačů
Fakulta elektrotechnická, České vysoké učení technické v Praze
tel.: +420 224 357 608, email: klema@fel.cvut.cz

Disertační práce Janky Puterové se soustředí na problematiku detekce repetitivních sekvencí v již sestavených genomech i přímo z NGS dat. Práce má v zásadě dvě části. Jádrem disertace je soubor 4 impaktovaných časopiseckých publikací, které se k tématu detekce repetitivních sekvencí přímo vztahují a prezentují autorkou dosažené výsledky v dané oblasti. Toto jádro je doplněno o vysvětlující kontext, který je vzhledem k interdisciplinární povaze této práce vhodným vodítkem pro četbu uvedených článků.

Konkrétně, kapitola 2 poskytuje biologické pozadí ústředního námětu disertační práce, kapitola 3 je zaměřena na detekční metody vycházející z již sestaveného referenčního genomu. Kapitola 4 vysvětluje jak technologie sekvenování ovlivňuje detekci repetic bez předchozího sestavení genomu. Využití detekce repetitivních sekvencí v biologii je obecně diskutováno v kapitole 5. Kapitola 6 prezentuje dosažené publikační výsledky scientometricky, plné texty publikací jsou k dispozici v přílohách. Kapitola 7 celou práci shrnuje a diskutuje možnosti budoucího výzkumu.

Už samotná struktura práce určuje, že jádro disertace bylo publikováno na potřebné úrovni. Pro danou fázi kariéry uchazečky považuji její publikační aktivitu za velmi dobrou a beru ji za významný příznak její vědecké erudice. Z důvodu svého informatického zaměření a také z důvodu podílu uchazečky na jednotlivých publikacích jsem se při posouzení významu dosažených výsledků soustředil především na publikaci v Genome Biology and Evolution mířící na analýzu satelitní DNA a na publikaci v BMC Bioinformatics navrhující nástroj digIS pro detekci inzerčních sekcencí. Z pohledu bioinformatického je přínos obou publikací spíše integrační a aplikační. Spočívá zejména v navržení pracovního toku, který využívá stávajících komplexnějších algoritmů. Oceňuji ale schopnost autorky navržené nástroje správně motivovat a také odpovídajícím způsobem použít a vyhodnotit. To je u mezioborové práce velmi důležité. Je evidentní, že díky spolupráci s biology a botaniky má práce Janky Puterový významný biologický přesah, zhodnocení jeho významu ale ponechávám na povolanějších.

Pokud se vrátím k informatické podstatě obou výše zmíněných článků, v prvním případě šlo především o pokročilejší post-processing výstupu z nástroje RepeatExplorer. Ten s pomocí shlukování založeného na grafech vyhledává repetitivní DNA v NGS datech. V článku je navržen postup, jak identifikovat opakující se úseky satelitní DNA zvané monomery, systematicky posoudit jejich homogenitu a využít vzájemné podobnosti k predikci jejich lokalizace na pohlavních chromozomech. V druhém případě je navržen nástroj digIS pro detekci inzerčních sekvencí. digIS vychází z profilových skrytých Markovových modelů soustředících se na katalytické domény

transpozáz představující nejkonzervovanější část inzerčních sekvencí zachovávající sekundární strukturu v rámci jedné rodiny. Návrh tohoto pracovního toku pokládám za originální a netriviální. Na některé dílčí nejasnosti při jeho vyhodnocení se odkazuji v otázkách v závěru posudku. Zodpovězení těchto otázek může osvětlit potenciál nástroje pro jeho širší použití, není ale kritické z pohledu posouzení splnění cílů předložené práce.

Ing. Janka Puterová ve své disertaci prokázala schopnost samostatné vědecké práce. Soustředila se na prakticky významnou a vědecky aktuální problematiku detekce repetitivních sekvencí a splněním svých cílů dosáhla v této oblasti zřejmého pokroku. Disertační práce uchazečky odpovídá obecně uznávaným požadavkům k udělení titulu PhD a práci **doporučuji k obhajobě**.

Otázky:

1. digIS je pracovní tok s řadou hyperparametrů. Můžete tyto hyperparametry pojmenovat a vysvětlit, jak bylo jejich nastavení odděleno od dat, která sloužila k vyhodnocení detekčních schopností digIS?

2. Pokud se na digIS podívám jako na nástroj získávání informací, lze jej hodnotit i pomocí dobře známé F1 míry. Ta je harmonickým průměrem precision (přesnost, 1-FDR vzhledem k metrice použité v článku) a recall (výtěžnost, identická se senzitivitou použitou k hodnocení v článku). digIS není z pohledu F1 nejvýkonnějším detektorem manuálně kuratovaných IS datasetů. Můžete vysvětlit principiální důvody proč digIS může překonat ostatní detektory repetic z pohledu nových, dosud neznámých inserčních sekvencí?

V Praze, 1. prosince 2021 doc. Ing. Jiří Kléma, Ph.D., recenzent