Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Rastislav Galvánek
Název práce	Predikce terciární struktury RNA s využitím více vzorů
Rok odevzdání	2019
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	Mgr. Jan Jelínek **Role** Oponent
Pracoviště	KSI MFF UK

Text posudku:

Cílem diplomové práce bylo rozšířit algoritmus Trooper na predikci terciární struktury RNA pomocí jedné vzorové RNA se známou strukturou navržený v bakalářské práci. Algoritmus měl být rozšířen jednak o predikci pomocí více vzorových struktur a jednak o mezikrok s predikcí sekundární struktury. V práci se autor dále věnoval odhadu asymptotické časové složitosti navržených algoritmů, zvýšení uživatelské přívětivosti jejich implementace a porovnání s vybraným konkurenčním algoritmem.

Co se týče využití sekundární struktury byl autor relativně úspěšný a povedlo se mu výrazně zlepšit na sekvencích střední délky (RMSD 6,91Å vs. 3,46Å), ovšem za cenu zhoršení kvality predikce krátkých sekvencí (RMSD 5,80Å vs. 8,23Å). V obou kategoriích je kvalita predikce lepší než kvalita predikce konkurenčního algoritmu ModeRNA (3,76Å, resp. 8,53Å), při tomto srovnání je ovšem potřeba vzít v do úvahy, že zatímco navrženému algoritmu trvá predikce celý den (resp. dny), algoritmu ModeRNA trvá predikce pouze minuty. Navržený algoritmus tak je použitelný spíše ve specifické situaci, kdy nezáleží na času běhu, resp. predikujeme pouze pár sekvencí a jde nám o maximální možnou kvalitu predikce.

V případě využití více zdrojových struktur autor příliš úspěšný nebyl. Autor sice navrhl dva přístupy k tomuto problému, ale jeden způsob predikce nedokázal zprovoznit a druhý způsob nevedl ke zlepšení kvality predikce (RMSD 10,72Å místo původních 6,16Å).

Z hlediska zvýšení uživatelské přívětivosti nastalo oproti bakalářské práci jisté zlepšení. Již například není nutné spouštět jednotlivé části na strojích s různým operačním systémem a kopírovat soubory mezi nimi. Pro reálné použití bych však autorovi doporučil využít vybraný systém pro správu balíků, který by potenciálním uživatelům usnadnil instalaci prerekvizit a zprovoznění algoritmu. V prostředí bioinformatiky se pro toto využívá například Conda. Dále by bylo vhodné program doplnit o podrobnou dokumentaci.

V práci je také analyzována asymptotická časová složitosti navržených algoritmů. Tato část nemá příliš velkou vypovídající hodnotu, neboť části algoritmu s nejhorší asymptotickou složitostí běží pouze sekundy až minuty (`Align` a `MapConservedParts`, resp. `PredictUnconservedSecStr` či `FindTemplateForGap` pro rozšíření) zatímco výpočetně nejnáročnější část algoritmu (`PredictUnconservedPairs`) je shora omezena časovým limitem a tedy v podstatě konstantní, přestože reálně běží mnohonásobně déle (limit je nastaven na 24 hodin). Navíc mám z popisu algoritmu jisté pochyby, zda je limitem omezena celá metoda `PredictUnconservedPairs`, nebo pouze jednotlivá volání FARFAR pro jednotlivé nekonzervované úseky. V druhém případě by totiž byla časová

složitost $O(1)$ pouze při paralelním volání FARFAR a při sériovém běhu by složitost byla $O(l)$ (minimální délka nekonzervaných úseků je zdola omezena svou délkou a tedy maximální počet nepřekrývajících se úseků je délkou sekvence shora omezen a závisí na ní lineárně). Tudíž by mě více zajímala analýza reálné časové náročnosti, například charakter sekvencí, které se do zvoleného časového limitu stihnou napredikovat, kolik struktur se obvykle stíhá do limitu namodelovat atd. případné se i na základě experimentů pokusit odhadnout složitost algoritmu FARFAR. Popis by navíc není dostatečně formální – l je definována pouze jako délka molekuly, resp. struktury v databázi, ale není již řečeno, zda délka vzoru/ dotazu, resp. maximum/ průměr délek. Navíc g (počet nekonzervovaných úseků) lze omezit pomocí l (viz výše).

Největší slabinou práce je její textová část, v níž se kromě množství chyb také opakovaně vyskytují zavádějící či dokonce nepravdivé informace. Například:

- Podle specifikace fasta formátu je neznámý nukleotid reprezentován pouze znakem N. X je pouze pro aminokyseliny, takže je sice možné, že ho občas někdo chybně využije i pro DNA/RNA, ale dle mých zkušeností tato chyba není obvyklá. Takže ani autorem generované soubory (kapitola 4.8) neodpovídají specifikaci.
- U template-based predikce lze jako vzor použít nejen experimentálně určenou strukturu, ale i předpovězenou. Byť se samozřejmě chyba predikce bude kumulovat.
- U semiglobálního zarovnání se nemusí nutně zarovnávat kratší sekvence na delší, může to být i naopak. Což dokonce není pravda ani v této práci, kde se zarovnávají fragmenty na vzory velikosti alespoň 80% délky příslušného fragmentu.

Práce dále obsahuje množství chyb, kterým se dalo předejít použitím automatické kontroly pravopisu. Například hned v obsahu jsou minimálně tři (existujúcixh, pôvdného a tempalte) a jedna je i v pěti klíčových slovech (modelovnie).
Čitelnost práce dále snižuje přílišné používání synonym bez zjevného důvodu, takže si čtenář není jist, zda jsou i autorem považovány za ekvivalentní, nebo zda jsou využity pro upozornění na nějaký drobný rozdíl. Například v tabulce 2.1 jsou dvě metody označeny jako komparativní modelování, ačkoliv v textu je zmíněno pouze homologní modelování a u obrázku 2.4 je pro změnu označení template based; nebo u jedné metody je zdůrazněno, že jde o metodu typu Monte Carlo, zatímco u druhé metody se to nezmíní.
Dále kostra algoritmu v kapitole 3.1 obsahuje o krok více, než jeho popis v kapitole 3.2, což snižuje přehlednost a komplikuje zpětné reference v pozdějším textu.
Matoucí je též použití pojmu „sekundární struktura“ ve dvou významech – jednak ve smyslu úrovně přiblížení při pohledu na strukturu a jednak ve smyslu doplňkové struktury pro rozšíření hlavní struktury. V druhém případě bych tak doporučil zvolit jiné označení, které by nejednoznačnostem předešlo.

Další chyby:

- "Očakávaná presnosť predikcie je priamo úmerná dĺžke neznámeho predikovaného úseku." – naopak čím větší predikovaný úsek, tím nižší přesnost predikce.
- Velikost přepočítávané sféry v pseudokódu v kapitole 3.5 neodpovídá popisu v kapitole 3.2 (odvozené z počtu reziduí v mezeře vs. od vzdálenosti krajních reziduí).
- Kapitola 4.5 a tabulka 4.2: V textu je psáno, že ModeRNA byla úspěšnější při predikci delších struktur, zatímco Trooper u kratších. Údaje v tabulce ale svědčí o opaku (nižší odchylku na kratších má ModeRNA a u delších je tomu naopak).
- Autor přiznává, že `MapConservedParts` je implementována "dosť nešikovne". Dle mého názoru takovéto vyjadřování do diplomové práce nepatří. Vhodnější by bylo např. napsat že metoda není z hlediska složitosti optimální, ale že není úzkým hrdlem programu (podobně jako v případě ukládání DB v kapitole 6) nebo že reálná/

očekávaná složitost je lepší. Za zavádějící považuji také formulaci "...sme popísali aj v článku" v případě článku, jehož autor práce není autorem.

- Metoda `SelectTemplates` na str. 35 tak jak je napsán pseudokód nevrátí po projití všech struktur požadovaný počet nejlepších, jak je tvrzeno v textu, ale méně. Osobně vidím jako nevýhodu i to, že u této metody záleží na pořadí průchodu sekvencí, takže např. pokud mám limit `returnFirstSuitableX=3` a procházím vzory s podobnostmi 60, 61, 62, 90 (pro jednoduchost navzájem dostatečně nepodobné), tak budou vráceny pouze první tři s podobnostmi 60, 61 a 62; pokud budu procházet vzory v pořadí 60, 90, 89, tak vrátím pouze ty s podobností 60 a 90, zatímco když budou projity v pořadí 89, 90, 60, tak budou vráceny ty s podobnostmi 89 a 90. Validace vzorů by mohla být předpočítaná, stejně jako vzájemná podobnost vzorů. Drobností je chyba ve jméně proměnné v těle metody a zdvojená podmínka podobnosti vzorů ve slovním popisu.
- Kapitola 5.2 pseudokód: Předpokládám, že `CopyConservedPartsSecStr` autor kopíroval z `CopyConservedParts` a zapomněl přepsat názvy některých proměnných. `CleanConservedPartsSecStr` pravděpodobně testuje totéž, co už je otestováno v `CopyConservedPartsSecStr` – tuto domněnku podporuje fakt, že v kapitole 6.2 již metoda není uvedena. Chybí modifikovaná metoda `ProcessShortUnconservedParts`. U metody `ProcessLongUnconservedParts` nesedí argumenty (to se týká i prvního pseudokódu v kapitole 3.5).
- Výsledky v tabulce 5.1 se liší od výsledků v tabulce 4.2, přestože je explicitně řečeno, že jde stejné testování nad stejnými daty.
- Závěry z obrázku 5.5 nemusí být nutně chybné, vzhledem k nedostatečné analýze dat (nebo jejich popisu) mám však pochyby o statistické průkaznosti. Například by mohla shluková analýza ukázat, že stejnou chybu mají RNA stejného typu; případně by mohlo jít o vedlejší efekt odlišné délky sekvencí. Vhodnější by bylo vzít korektní sekundární struktury a do nich náhodně vnášet konkrétní počty chyb.
- FindTemplateForGap v pseudokódu kapitoly 6.2 neodpovídá textům v kapitolách 6.1 a 6.2 (vzor s nejlepším zarovnáním vs. první vzor s dostatečně kvalitním zarovnáním).
- Kapitoly 6.2 a 6.3: Metoda `ProcessLongUnconservedParts` by neměla být vypuštěna, neboť se může stát, že metoda `MultipleTemplatesModule` (alespoň tak jak je popsána) nenajde vhodný doplňkový vzor pro všechny dlouhé nekonzervované úseky. Ostatně, z pseudokódu není tato metoda odstraněna. Dále by měly být zpracované úseky opět analyzovány pro nalezení nekonzervovaných úseků.

Dotazy k obhajobě/ co mohlo být rozebráno důkladněji:

- Jaká je prostorová složitost vašeho algoritmu? Ideálně i porovnejte s konkurenčním ModeRNA.
- Obrázek 3.4: bylo by možné přidat i zarovnání predikce a vzoru?
- Kapitola 4.1: kolik bylo struktur v jednotlivých přihrádkách? Jaký byl důvod zvolit přihrádky zrovna takto? (Velikost přihrádek/ charakter RNA/ podobnost kvality predikce…) Odstraňoval jste duplicitní sekvence? Např. v případě u proteinů je v PDB každá sekvence v průměru 4x, jaká je situace u RNA? Pokud by se některé sekvence vyskytovaly násobně vícekrát než jiné, může to ovlivnit výsledky např. v případě obrázku 5.5. Jaká je RMSD zarovnání struktur stejných sekvencí (resp. i vyřazených příliš podobných), nenastala situace, kdy by stejná/ podobná sekvence měla výrazně odlišnou strukturu?
- Kapitola 4.5: Jaká byla míra neúspěšnosti jednotlivých algoritmů? Čím si

vysvětlujete, že je váš algoritmus lepší v jedné kategorii a ModeRNA v druhé?

- Kapitola 4.6, resp. tabulka 4.3: jak byly vybrány tyto páry, jde o náhodné páry/ jediné páry/ páry vybrané cíleně pro demonstraci problémů ModeRNA/ ...?
- Kapitola 4.8 Mohl byste rozvést, proč je generování fasta sekvencí z pdb souborů komplikováno nekompletností těchto pdb souborů, když jste problém nekompletnosti pdb souborů chtěl eliminovat právě generováním fasta sekvencí z těchto pdb souborů? Mimochodem, z textu práce si nejsem zcela jist, jako indexy nukleotidů používáte sekvenční číslo rezidua `resSeq`? Pokud ano, tak bych doporučoval používat spíše pořadí nukleotidů v souboru - dle specifikace mají být nukleotidy v souboru setříděny od 5' konce ke 3' konci; a i kompletní pdb soubor může mít `resSeq` nesouvislé kvůli konzistenci očíslování mezi homologními sekvencemi (nebo naopak mít více nukleotidů se stejným `resSeq` rozlišeným pomocí různého `iCode`). Je problémem pouze zcela chybějící reziduum, nebo i nekompletní reziduum? Rezidua chybějící v pdb souboru často mohou znamenat, že daná část struktury nemá stabilní strukturu, neuvažoval jste o možnosti zahrnout i tuto informaci do procesu predikce?
- Čím si vysvětlujete, že kvalita predikce je na kratších sekvencích výrazně horší, než na delších sekvencích? Je zajímavé, že při použití sekundární struktury jsou výsledky natolik blízké výsledkům ModeRNA, čím si to vysvětlujete?
- Kapitola 6.1: Proč se bere první vyhovující doplňkový vzor a nikoliv nejlepší, je to z časových důvodů, nebo jste otestoval, že se to výsledkově příliš neliší?
- Kapitola 6.2: Proč by se údaje o databázi musely přepočítávat při každé změně, nestačilo by přepočítat změněná data? Rozšíření nekonzervovaných úseků se dělá již v metodě `ProcessGaps`, protože by nekonzervovaný úsek mohl ovlivnit okolní strukturu; neměly by se i ze zarovnání doplňkové struktury krajní rezidua ignorovat, protože i tam už může být změna struktury vynucená nepodobnými částmi sekvence navazujícími na zkoumaný úsek? Jaké má následky, když se překryjí dva nekonzervované úseky? Sekundární struktura doplňována dalšími vzory není?
- Kapitola 6.5: Čím si vysvětlujete, že se kvalita predikce většinou zhoršila? Zkoušel jste analyzovat samostatné doplňkové vzory, zda je špatné už zarovnání na nich nebo zda je problém s jejich napojením na hlavní vzor?
- Závěr: Když srovnáváte váš algoritmus s ModeRNA, máte lepší průměrné RMSD. Jaký je rozptyl RMSD, resp. jaké je rozložení rozdílů kvality predikce vašeho algoritmu a ModeRNA na jednotlivých strukturách – jste lepší ve výrazné většině případů, nebo je lepší výsledek způsoben malou skupinou výrazně lepších výsledků (v tom případě byl byste je schopen algoritmicky rozpoznat)? Zkoumal jste, zda jste schopen algoritmicky rozpoznat, kdy má smysl použít k predikci sekundární strukturu/ více vzorů a kdy by to vedlo naopak ke zhoršení predikce?
- Je možné při predikci přímo využít vlastní sekundární strukturu, nebo jsou potřeba úpravy v implementaci?

Práci nedoporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 26. 8. 2019 **Podpis**