

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 David Hoksza
Pracoviště KSI MFF UK

Role Vedoucí

Text posudku:

Předložená diplomová práce se zabývá predikcí terciární struktury RNA molekul a jedná se tedy o multidisciplinární práci na pomezí informatiky a molekulární biologie. Ačkoli predikce 3D struktury proteinů je často řešený problém, u predikce terciární struktury RNA tomu tak není. To je dáno především faktem, že relativně donedávna bylo pohlíženo na RNA pouze jako na molekulu, jejíž účelem je přenos genetické informace z jádra buňky. V posledních 20 letech bylo ale ukázáno mnoho dalších rolí, na nichž se RNA podílí, a zájem o výzkum této molekuly tudíž vzrostl. S tím souvisí pak nárůst počtu determinovaných RNA struktur ve strukturních databázích a tedy i počet struktur, které se dají použít pro tzv. homologní modelování/predikci. Při homologním modelování je na vstupu sekvence (tzv. target) molekuly pro níž se snažíme identifikovat 3D strukturu a množina známých struktur (sa tedy i sekvencí), které se používají jako vzory (tzv. template). Tento proces je založen na identifikaci mapování na úrovni target-template sekvence a přenosu částí template struktury v místech s dostatečnou target-template podobností. Zbytek struktury, tj. tam kde nebylo možno použít template, je pak predikován de novo.

Základní metoda homologního modelování pro predikci 3D struktury RNA byla vyvinuta studentem v rámci jeho bakalářské práce. V rámci diplomové práce se pak student pokoušel 1) vylepšit metodu v několika směrech, 2) vyvinout jednoduše použitelný softwarový nástroj, implementující navrženou metodu a 3) vyhodnotit metodu vzhledem ke state of the art přístupům.

Vylepšení metody, které student navrhl, se týkalo dvou směru: využití predikce sekundární struktury a využití více vzorů. Predikce sekundární struktury (ať už u proteinů nebo RNA) je jednodušší problém a proto se autor pokusil nejdříve implementovat metodu pro jednoduchou predikci sekundární struktury a predikovanou sekundární strukturu pak použít jako omezující podmínky při predikci částí terciární struktury, které nebyly přeneseny ze vzoru. Toto je možné, protože sekundární struktura je v podstatě množina dvojic nukleotidů, které jsou spojeny vodíkovými můstky a jsou tedy blízko sebe ve 3D prostoru. Tzn. že se jedná o aproximaci struktury terciární, kdy sekundární struktura dává sadu omezení na cílovou 3D strukturu.

Druhý pokus o vylepšení predikce se pak týkal použití více vzorů. Myšlenka tohoto přístupu byla taková, že části z vybraného vzoru, které nejsou dostatečně podobné nelze přenést, ale potenciálně je možné identifikovat další RNA molekuly, které by byly podobné k target molekule v místech, kde nebyla identifikována dostatečná podobnost s primárním vzorem.

Řešení bylo implementováno v plnohodnotném nástroji Trooper, který byl publikován na Githubu a přístupná široká komunitě. Implementace vyžadovala netriviální množství práce týkající se nutnosti správného zpracování široké škály bioinformatických vstupů, integrace externích nástrojů pro řešení dílčích úloh, i implementaci řešení pro vyhodnocení na virtuální infrastruktuře.

Všechny navržené přístupy pak byly otestovány a srovnány se state of the art metodou ModeRNA pro homologní predikci RNA struktur, jediný nástroj svého druhu. Výsledky srovnání ukazují, že nástroj Trooper je srovnatelný a v mnoha případech dosahuje lepších výsledků než ModeRNA. Na druhou stranu, navržená vylepšení algoritmu vedla ke smíšeným výsledkům, kdy na některých typech struktur tato vylepšení vedla ke zlepšení predikce, nicméně na některých ke zhoršení. Vzhledem k tomu, že se jedná o výzkumnou práci, i tyto typy výsledků jsou jistě přínosem. Zvláště pak protože práce obsahuje i možnou interpretaci negativních výsledků a tedy dobrou výchozí pozici pro další výzkum v této oblasti. Tyto výsledky byly prezentovány jako recenzovaná publikace na IEEE bioinformatické konferenci.

Na závěr bych poznamenal, že předložená práce řeší komplexní a stále otevřený problém z oblasti bioinformatiky, což je zrcadleno i v textu práce a předloženém softwarovém řešení. Jedná se o komplexní systém, při jehož vývoji bylo třeba do detailu pochopit biologické pozadí problému, navrhnout řešení (často vyzkoušet více různých přístupů), toto řešení implementovat a potýkat se s mnoha specialitami v dané doméně. Množství práce, které bylo třeba vykonat a předložené výsledky jsou, dle mého názoru, zcela v souladu s požadavky kladenými na kvalitní diplomovou práci.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

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

Datum 25. 8. 2019

Podpis