Oponentský posudok na dizertačnú prácu Ing. Jiřího Hona na tému Mining of soluble enzymes from genomic databases

Jadrom dizertačnej práce sú dva nové bioinformatické softvérové nástroje. Webový nástroj EnzymeMiner umožňuje používateľom interaktívne vyberať proteíny na další výskum, pričom o nich zobrazuje údaje získané z veľkého množstva zdrojov. Nástroj SoluProt využíva techniky strojového učenia na určenie, či daný proteín bude produkovateľný v baktérii E.coli. Téma dizertačnej práce teda spadá do odboru Výpočetní technika a informatika a je tiež veľmi aktuálna, nakoľko pokroky v biotechnológiách umožňujú získavať sekvencie veľkého množstva proteínov a nástroje ako tieto sú nevyhnutné na prioritizáciu ich ďalšieho skúmania.

V obidvoch hlavných častiach práce autor preukázal originálny prínos vyvinutím nových informatických metód. V nástroji EnzymeMiner je originálny celkový prístup k hľadaniu viacerých vhodných kandidátskych proteínov na základe veľkého množstva údajov zhromaždených k proteínom získaným z databáz. V nástroji SoluProt bol originálnym prínosom hlavne dôraz na lepšie spracovanie trénovacej a testovacej množiny než tomu bolo v predchádzajúcich prácach na tú istú tému. To autorom umožnilo mierne zlepšiť dosiahnutú presnosť a tiež lepšie vyhodnotiť presnosť aj predchádzajúcich nástrojov.

Jadro dizertačnej práce bolo publikované na potrebnej úrovni, články o spomínaných nástrojoch boli publikované v časopisoch Nucleic Acids Research a Bioinformatics, ktoré patria medzi popredné časopisy v tejto oblasti. EnzymeMiner bol už viackrát citovaný a obidva nástroje sú využívané vedeckou komunitou. Okrem toho je Jiří Hon spoluautorom aj ďalších štyroch publikácií v oblasti bioinformatiky. Autor významnou mierou prispel k uvedeným publikáciám, čím preukázal vedeckú erudíciu a je pripravený na samostatnú výzkumnú činnosť.

Hlavný text dizertačnej práce má 40 strán a tvorí ju úvod do problematiky a prehľad dosiahnutých výsledkov. Prílohou práce sú tri z autorových článkov, v ktorých je možné nájsť ďalšie detaily. Hlavný text práce je písaný prehľadne, oceňujem špecifikovanie prínosu autora k jednotlivým publikáciám. K textu mám len menšie pripomienky, ktoré spolu s otázkami k obhajobe prikladám v prílohe k posudku.

Celkovo konštatujem, že dizertačná práca spĺňa požiadavky na udelenie akademického titulu PhD.

V Bratislave 9.1.2022

Doc. Mgr. Bronislava Brejová, PhD.
Katedra informatiky
Fakulta matematiky, fyziky a informatiky
Univerzita Komenského v Bratislave
Mlynská dolina, 842 48 Bratislava, Slovensko

Príloha 1: pripomienky k textu práce

- V niektorých častiach práce, napríklad v úvode časti 2.3, by sa hodilo uvedené tvrdenia podporiť citáciami.

- V prehľade databáz by bolo okrem databázy BRENDA spomenúť napríklad aj databázy KEGG a GO, ktoré sa tiež týkajú funkcie proteínov.

- V úvode časti 4.1 sa solubility definuje čisto chemicky, ale neskôr sa za nerozpustné označujú napríklad aj proteíny toxické pre hostiteľský organizmus, ktoré ale z chemického hľadiska môžu byť dokonale rozpustné. Viac sa mi páči definícia z článku o SoluProt: "Specifically, by solubility, we mean the probability of soluble protein (over)expression in Escherichia coli cells." Škoda, že nebola podobná definícia uvedená aj v práci.

Príloha 2: otázky k obhajobe:

- Ak tomu dobre rozumiem, EnzymeMiner zobrazuje identitu iba k proteínom, ktoré boli súčasťou dotazu. Ak však používateľ vyberá väčšiu skupinu kandidátov, nebolo by vhodné zobrazovať aj identitu k už zvoleným proteínom? Viete si predstaviť ako by sa takýto typ interaktivity zakomponoval do vášho systému?

- V budúcej práci navrhujete v softvéri EnzymeMiner využiť predikované štruktúry všetkých zobrazených proteínov. Ako by sa informácia z veľkého množstva štruktúr dala prehľadne sumarizovať a prezentovať používateľovi?

- V softvéri SoluProt dosahujete vyššiu presnosť na trénovacej časti než na validačnej časti trénovacích dát, čo naznačuje určitú mieru preučenia (overfitting). Viete navrhnúť nejaké techniky ako tento jav redukovať?