Posudek oponenta diplomové práce

Jméno a příjmení autora posudku:	RNDr. František Mráz, CSc.
Jméno a příjmení autora práce:	Bc. Petr Škoda
Název práce	Efficient molecular representation design

Text posudku

Predložená práca sa zaoberá využitím bitových reťazcov, tzv, fingerprintov, na porovnávanie zložitých organických zlúčenín. Konkrétna úloha, na ktorej diplomant uvedené metódy skúšal, bol tzv. virtuálny screening, kde sa na základe zloženia cieľovej zlúčeniny snažíme nájsť v zozname známych zlúčenín zlúčeniny, ktoré by s danou cieľovou molekulou mohli reagovať. Konkrétne riešenie tohoto značne zložitého problému spočíva v tom, že sa na základe fragmentov, ktoré sa v danej molekule vyskytujú, spočíta bitový reťazec – fingerprint, ktorý ich má reprezentovať. Podobne reagujúce zlúčeniny sa potom určujú ako zlúčeniny s podobným fingerprintom.

Autor v práci popísal vyššie uvedený problém a princípy metód, ktoré sa na jeho riešenie používajú. Potom navrhol vlastný spôsob vytvárania fingerprintov založený na skladaní bitových reťazcov kódujúcich jednoduché fyzikálne-chemické vlastnosti fragmentov, ktoré sa v molekule vyskytujú. Tento postup nie je nový, ale ako ukázali jeho testy, tak aj využitie veľmi jednoduchých atribútov môže dávať výsledky lepšie, než mnohé sofistikovanejšie metódy. Autor svoju metódu potom otestoval s pomocou open-source benchmarkového systému Rinikerovej a Landruma. Implementačná časť práce spočívala v doplnení jeho metódy generovania fingerprintov do uvedeného benchmarkového systému.

Ako som už uviedol, experimenty ukázali, že aj jednoduchá metóda môže dávať výborné výsledky a stálo by za to vyskúšať doplnenie evidovaných parametrov o ďalšie položky, či by to ďalej nezlepšilo kvalitu porovnávania zlúčenín.

Na druhú stranu má práca značné nedostatky:

1. Práca je napísaná po anglicky, ale s ohromným počtom chýb od preklepov po nesprávne koncovky slovies v 3. osobe jednotného čísla, členy atď.

2. Okrem gramatiky text trpí ďalšími nedostatkami po formálnej stránke. Uvediem iba niekoľko príkladov z mnohých:

   - Číslovaný zoznam na str. 25 je nasledovaný totožným textom v jednom odstavci, podobne posledný bod zoznamu na prelome stránok 38 a 39 je hneď doslova zopakovaný.
   - V texte sa vyskytujú odkazy na sekcie, obrázky a tabuľky, ktoré sú však neúplné, alebo úplne nesprávne (napr. začiatok predposledného odstavca na str. 39 tvrdí “As

the results 8.3 8.3 8.3 8.3 9 9 9 show", kde čísla nie sú výsledky, ale mali to byť odkazy na obrázky (podobné odkazy sú na konci prvého odstavca na str. 44 a inde).

- Všetky grafy v práci majú mikroskopické popisy, ktoré sa dajú čítať iba s lupou. Zvolené čiarové grafy sú pre prezentáciu nevhodné. Jednak zobrazované hodnoty by mali byť diskrétne body, pretože medzi testovanými zlúčeninami nie je možné spojite prechádzať, a tiež záleží, ktorá farba bola pri kreslení čiar použitá ako posledná – to potom skresľuje vizuálne porovnanie dosiahnutých výsledkov.
- Diagramy 5.1, 6.1 a 7.1 sú skoro nečitateľné. Obrázok 4.1 (str. 15) nemá popis.
- Premenné v poslednom odstavci na sdtr. 17 sú písané iným písmom, než vzápätí vo vzorcoch.

3. Ďalej prejdeme k obsahovým chybám. Súčasťou práce je implementácia fingerprintov, ktorá je podľa autora vysoko konfigurovateľná, ale čitateľ sa nedozvie, ako by si mohol urobiť vlastný typ fingerprintu. Postup je jednoduchý, ale ani v práci, ani na priloženom CD nie je popísaný a dá sa vyčítať iba zo skromne komentovaných zdrojových súborov. Tiež by bolo rozumné na CD priložiť skripty a návod, ako experimenty, ktoré diplomant urobil, zopakovať. Priložený benchmarkový systém to veľmi jednoducho umožňuje.

4. V rešeršnej časti sa vyskytujú atribúty, ktoré nie sú vôbec popísané. Príklady: logP na str. 9. Absolútna väčšina vzorcov v práci obsahuje premenné, ktoré nikde v práci nie sú popísané, prípadne je ich popis nedostatočný:

   - Vzorec pre plochu pod krivkou ROC (AUC na str. 19), ktorý sa používa ako hlavné kritérium správnosti porovnania molekúl, obsahuje n a N, ktoré nie sú zavedené, namiesto A_i a I_i je popisované A a I. O AUC sa tvrdí, že môže byť iba 0 alebo 1, správne malo byť, že môže nadobúdať hodnoty z intervalu $\langle 0, 1\rangle$ (odstavec pod vzorcom).
   - Vzorce a pojmy pre "enrichment factor", "robust initial enhancement" a "Boltzmann-enhanced discrimination of ROC" nie sú dostatočne vysvetlené okrem iného aj preto, že atribúty použité vo vzorcoch nie sú popísané (str. 20–21). Čo je parameter R na riadku pred oddielom 5.2 (str. 21)?
   - Vzorce pre počítanie podobnosti medzi bitovými reťazcami (str. 17) sú len vymenované, bez žiadneho komentára. Prečo tam sú? V experimentoch sa používa iba druhý z nich.

5. Počítanie fingerprintov je kľúčové miesto práce. Konkrétny postup použitý autorom je popísaný na stranách 30 a 31. Výsledný fingerprint vznikne zložením fingerprintov pre jednotlivé fragmenty detekované v molekule. O kľúčovej metóde ako sa fingerprinty fragmentov zostavia do výsledného bitového reťazca je napísané "In our implementation we utilize existing method

that comes with the RDKit". To je všetko, čo sa čitateľ dozvie. Ostatné si má nájsť asi v dokumentácii pre RDKit?

6. Pri návrhu experimentov sa autor obmedzil na 5 atribútov z niekoľkých stoviek, ktoré mu spočítal použitý program `PaDel`. Podľa zadania mal diplomant navrhnúť zobecnenie fingerprintov, ktoré by umožňovalo vo fragmentoch uchovávať a vo fingerprintoch reprezentovať ľubovoľný súbor vlastností atómov. Zvažoval diplomant takéto možnosti?

7. V experimentoch najprv autor porovnával svoje fingerprinty s riešením "BASE", ktoré každý fragment reprezentovalo bitom s hodnotou 1. Dáva takéto porovnanie zmysel v použitom benchmarkovom systéme? Tam sa pri testovaní používajú dve fázy. Najprv sa natrénuje klasifikátor a potom sa testuje. Autor o tom v práci nič nepíše. Pri použití jediného deskriptoru sa osvedčil atribút nHeavyAtom (počet ťažkých atómov vo fragmente). Keď autor testuje dvojice deskriptorov, tak skúša predovšetkým iné dvojice a jedinú dvojicu obsahujúcu deskriptor nHeavyAtom. Prečo?

8. V záverečnom teste, jedinom, kde autor porovnáva svoj 4-bitový fingerprint s vybranými fingerprintami z benchmarku, uvádza, že jedným z nich je Extended connectivity fingerprint (ECFP4), ale hodnoty uvádzané v tabuľke pre tento fingerprint sú odlišné od hodnôt uvedených v článku o bechmarkovom systéme. Myslím, že porovnávaným fingerprintom bol iný fingerprint FCFP4.

Vyššie uvedené poznámky ukazujú, že práca má nízku úroveň a asi nie je tým najlepším vyvrcholením autorovho štúdia na MFF UK. Prácu doporučujem k obhajobe s tým, že finálne rozhodnutie o jej prijatí ako diplomovej práci nechám na komisii.

Doporučení k obhajobě

Z výše uvedených důvodů práci *doporučuji* k obhajobě.

Soutěž studentských prací

V Praze dne 22. 5. 2014

Podpis: