Oponentský posudok na diplomovú prácu Radoslava Kriváka „Algorithms for protein-ligand binding site discovery“

Proteíny sú najdôležitejšie biologicky aktívne makromolekuly. Chemicky sú to lineárne reťazce aminokyselín pospájané kovalentnými väzbami. Vlastnosti takýchto polypeptidov sú dané nielen fyzikálno-chemickými vlastnosťami jednotlivých amínokyselín, ale hlavne tým, ako sa daný proteín "poskladá" v priestore. Cieľom predkladanej práce bol návrh algoritmu na detekciu väzobných miest na proteínoch. Špeciálne sú to tzv. kapsy, do ktorých sa môže naviazať menšia molekula, tzv. ligand. Algoritmy riešiace tento problém sú vyvíjané už takmer 30 rokov. Typickým výstupom známych algoritmov je zoznam možných kapies na povrchu konkrétnej trojrozmernej štruktúry, do ktorej by sa daný proteín mohol poskladať. Hlavným využitím pre takéto algoritmy je návrh nových liekov.

Diplomant najprv urobil stručný prehľad známych metód predikcie kapies. Stručne zhodnotil ich výhody a nevýhody a ukázal, že pri hodnotení týchto algoritmov v literatúre sa používajú rôzne kritériá. Rôzne algoritmy často vydajú viacej kandidátov na kapsy, než je skutočný počet použiteľných kapies, pričom skutočné kapsy nie sú medzi prvými. To vedie zbytočne na negatívne hodnotenie takéhoto algoritmu. Diplomant preto vybral jeden z rýchlych algoritmov nazývaný Fpocket, ktorý hľadá kapsy len na základe geometrie povrchu proteínu pri zadanej trojrozmernej štruktúre proteínu. Pre tento algoritmus autor práce navrhol vziať jeho výstup, pridať k nemu charakteristiky nájdených kandidátov na kapsy, spočítať z nich nové ohodnotenie kandidátov na kapsy a urobiť nové usporiadanie nájdených kandidátov na základe spočítaného ohodnotenia. Výstupom modifikovaného algoritmu je zoznam tých istých kapies ako vydal pôvodný algoritmus, ale v lepšom usporiadaní.

Pri výpočte ohodnotenia kapsy diplomant používa les náhodne generovaných rozhodovacích stromov, ktorý sa strojovo učí z databázy proteínov a kapies, ktorú diplomant vytvoril spojením troch dostupných súborov takýchto dát.

Výsledkom je skutočné zlepšenie oproti výstupu programu Fpocket, čím diplomant splnil zadanie práce. Jeho výsledky sú zaujímavé a bolo by vhodné vyskúšať obdobný prístup k výstupom dokonalejších programov na hľadanie kapies, ktoré sú založené na kombinácii geometrického prístupu (ako je Fpocket) a prístupov založených na počítaní energie väzieb, prípadne evolučnej podobnosti nájdenej kapsy s nejakou známou kapsou.

Na druhú stranu, k práci mám nasledujúce pripomienky:

1. Práca je napísaná anglicky, kde jazyková kvalita prehľadu známych metód je značne vyššia, než jazyková kvalita časti s vlastnými výsledkami. Dokonca v časti s vlastnými výsledkami (na rozdiel od prehľadu) je mnoho preklepov, na ktoré by stačil ľubovoľný spell checker. Okrem toho je tam systematicky zamenený termín "Voronoi diagram" za nesprávny "Vornoi diagram".
2. Popis známych metód hľadania kapies je miestami príliš stručný a spolieha na geometrickú predstavivosť čitateľa, pričom by aj jednoduché obrázky značne pomohli pochopeniu textu.
3. Napriek dobrým výsledkom, ktoré dáva autorom navrhnutý postup, v práci chýba zhodnotenie, či všetky charakteristiky, ktoré použil na natrénovanie skórovacej funkcie sú skutočne potrebné. Týka sa to aj jednoduchých úvah ako napr. či je nutné do vektora charakteristík dávať 3 bity pre rezídua, ktoré môžu byť vo vodíkových mostíkoch donorom (hBondDonor), akceptorom (hBondAcceptor), alebo súčasne donorom i akceptorom (hBondDonorAceptor), alebo by stačilo použiť iba 2 bity (hBondDonor a hBondAcceptor).
4. Ďalej pri počítaní charakteristík kapsy má veľký význam použitá agregačná funkcia (str. 40). Ako autora napadla práve takáto forma agregačnej funkcie (vzorce (3.2) a (3.3))? Nebola by iná agregačná funkcia vhodnejšia?
5. Skórovacia funkcia je počítaná lesom náhodných rozhodovacích stromov. Koľko stromov bolo použitých? Aké mali ďalšie parametre? Čo bola cieľová hodnota, ktorú mala skórovacia

funkcia vydať? Všetky tieto parametre mali byť v práci diskutované. Konštatovanie, že boli vyskúšané rôzne metódy klasifikácie s podobnými výsledkami je nedostatočné.

Toto považujem za najväčší nedostatok práce, pretože to ukazuje, že autor sa uspokojil s vyskúšaním jediného postupu s jediným nastavením parametrov a vôbec sa nezamyslel nad jeho vhodnosťou.

6. Chýba akákoľvek dokumentácia k priloženým programom, čo prípadnému používateľovi znemožní ich použitie. Iba zo zdrojových kódov sa dá vyčítať, čo autor naprogramoval sám a čo sú cudzie knižnice.

Napriek vyššie uvedeným nedostatkom práca splnila stanovené ciele a po doplnení spomínaných častí by bola vhodná i na publikovanie. Preto doporučujem prácu Radoslava Kriváka uznať ako diplomovú prácu.

Praha, 3. 9. 2013

RNDr. František Mráz, CSc.

KSVI MFF UK