Radoslav Krivák: Algorithms for protein-ligand binding site discovery

posudek vedoucího diplomové práce

Předkládaná práce se věnuje využití metod strojového učení pro těžký problém identifikace míst interakce proteinu a ligandu, tzv. kapes v proteinové struktuře. Jde o zásadní problém v oblasti bio-informatiky, který slouží např. ke studiu možností proteinu pro strukturální vývoj léků s danými vlastnostmi. Práce přistupuje k řešení problému metodami založenými na identifikaci těchto míst pomocí lokálních geometrických a chemických vlastností Vlastní řešení spočívá ve vylepšení de facto standardního algoritmu Fpocket z roku 2009 a ověření kvality výsledků na používaných datových množinách.

Struktura textu práce je následující: První kapitola uvádí do problematiky proteinů z hlediska chemických vlastností a postupů molekulární biologie. Součástí je i přehled oblastí a metod řešení strukturální bio-informatiky, která je v současnosti velmi bouřlivě se rozvíjejícím oborem. Závěrem této kapitoly je definice problému, který autor v práci řeší. Druhá kapitola popisuje stav dané problematiky na základě rešerše literatury a autorových rozsáhlých zkušeností s těmito metodami a software. Třetí kapitola je klíčová a popisuje autorův původní výsledek, novou funkci predikce kapes, kterou lze použít jako vylepšení některých existujících algoritmů. Čtvrtá kapitola ověřuje navržené řešení na reálných datech a pátá kapitola obsahuje shrnutí a výhled do budoucí práce.

Za hlavní přínosy práce považuji:

Rešeršní část práce je napsána přehledně a erudovaně. Přehled problematiky i vlastní zkušenosti autorovi pak slouží jako východisko pro původní práci.

Vlastní návrh ranking funkce, která určuje pořadí nadějných oblastí proteinu, kde by se mohla vyskytovat vazba protein-ligand. Tento návrh zahrnuje vhodnou extrakci lokálních chemicko-fyzikálních dat a dalších dat o proteinu, jejich agregaci a návrh vhodné klasifikační metody.

Zhodnocení výsledků experimentů, které na jednu stranu ukazuje úspěšnost autorovy metody, ale na druhou stranu hovoří i o vlastnostech datových množin běžně používaných v této oblasti. Výsledky ukazují, že různé testovací množiny jsou navzájem jen těžko porovnatelné, pravděpodobně proto, že obsahují heterogenní data.

Považuji předkládanou práci za velmi kvalitní, o čemž svědčí i to, že se autor této oblasti hodlá věnovat dále i ve svém doktorském studiu, a s ohledem na všechny zmíněné skutečnosti ji doporučuji uznat jako diplomovou práci.

V Praze dne 30. Srpna 2013 Roman Neruda