

Posudok vedúceho na diplomovú prácu Jana Hudečka

„Improving and extending the multiple sequence alignment suite PRALINE“

Alignment – zarovnávanie biologických sekvencií (DNA, RNA a proteínov) je jednou z kľúčových úloh bioinformatiky. Pre zisťovanie funkcie enzýmov je nutné zarovnávať viacej sekvencií naraz – tzv. násobné zarovnávanie (multiple alignment). PRALINE je jedným zo súborov programov umožňujúcich počítať násobné zarovnávanie. Diplomant navrhol, implementoval a na príkladoch overil rozšírenie súboru programov PRALINE o zarovnávanie užívateľom zadaných profilov sekvencií a dve nové metódy násobného zarovnávanie sekvencií založené na skrytých Markovských modeloch (SMM).

Autor v práci veľmi stručne popísal problém zarovnávanie a základné prístupy k jeho riešeniu spolu s ich dostupnými implementáciami. Ďalej popísal využitie skrytých Markovských modelov pri analýze biologických sekvencií. Vlastným výsledkom autora sú dve vylepšenia systému PRALINE:

- A. Zarovnávanie dvoch alebo viacerých profilov sekvencií namiesto priameho zarovnávanie sekvencií, navyše s využitím predikcie sekundárnej štruktúry. Jeden profil sekvencie v skutočnosti popisuje skupinu príbuzných sekvencií. Často sú to starostlivo ručne upravované profily, ktorých využitie zlepšuje kvalitu výsledného zarovnanie.
- B. Nová reprezentácia jadra zarovnanie pomocou SMM, ktorá zlepšuje citlivosť zarovnávanie. Jedná sa o dve metódy. Jednoduchšia metóda HMMGUIDE využíva SMM na počítanie podobnosti sekvencií a riadenie postupného pridávanie sekvencií k výslednému násobnému zarovnanie. Zložitejšia metóda PRCALIGN navyše môže využiť predikovanú sekundárnu štruktúru a výsledné zarovnanie buduje postupne porovnávaním skrytých Markovských modelov profilov namiesto tradičného využitia dynamického programovania.

Autor navrhnuté metódy implementoval a sú dostupné cez webové rozhranie systému PRALINE. Pre vyladenie navrhnutých metód bolo nutné pomocou série experimentov nájsť parametre využívaných štatistických metód. Výsledný program v rôznych nastaveniach autor porovnal s pôvodným programom PRALINE na benchmarkových problémoch. Tieto výsledky potvrdili očakávané zlepšenie oproti pôvodnej verzii programu. Ukázalo sa však, že pre niektoré násobné zarovnanie, o ktorých sa vie, že sú ťažké, metóda PRCALIGN zlyháva.

Diplomová práca Jana Hudečka vznikla počas jeho študijného pobytu na univerzite Vrije v Amsterdame (Holandsko) pod odborným vedením prof. Heringa, ktorý je spoluautorom systému PRALINE. Práca je napísaná slušnou angličtinou a veľmi fundovane s rozsiahlymi citáciami. Miestami by si zaslúžila podrobnejší popis, čo by však viedlo k neúmernému predĺženiu textovej časti práce. Pri jej vypracovaní musel autor zvládnuť rozsiahlu oblasť z moderných metód bioinformatiky, skombinovať a upraviť rozsiahle knižnice programov. Taktiež urobil rozsiahle experimenty na benchmarkových úlohách. Po doladení parametrov metódy PRCALIGN je veľká šanca, že táto metóda nahradí staršie metódy implementované v PRALINE ako kvalitnejšia náhrada.

Celkovo považujem prácu pána Jana Hudečka za výbornú a doporučujem ju k publikácii. Myslím, že je to dobrý základ na pokračovanie výskumu v tejto oblasti. Preto doporučujem, aby práca pána Jana Hudečka bola uznaná, ako diplomová práca.

Praha, 19.5.2010

RNDr. František Mráz, CSc.

KSVI MFF UK