POSUDEK VEDOUCÍHO DIPLOMOVÉ PRÁCE

Autor práce: Bc. Marek Zvara
Název práce: Automatická detekce intronů hub pomocí pravděpodobnostních modelů
Vedoucí: Doc. Ing. Jiří Kléma, Ph.D.

Automatická segmentace genomických sekvencí je tradiční úlohou. I když v ní bylo dosaženo významných úspěchů, automatické oddělení jednotlivých typů segmentů (např. intergenomická oblast, exony, introny) stále nedosahuje dokonalosti, se kterou s DNA pracuje buňka sama. V této diplomové práci se pan Zvara soustředil na specifickou podúlohu výše zmíněné obecné úlohy, konkrétně detekci intronů v genomech hub. Práce má velký aplikační potenciál, po odstranění intronů z neznámé DNA lze mnohem spolehlivěji rozhodnout o biologickém druhu, ze kterého DNA pochází. Introny jsou totiž mnohem méně evolučně konzervovány než exony a po jejich odstranění lze dosáhnout nadprahové shody při zarovnání sekvencí evolučně blízkých i vzdálenějších druhů. To je v případě hub velmi důležité. Odhaduje se, že existuje více než milion druhů hub, sekvenován přitom byl pouze zlomek z tohoto počtu.

Na tématu pracovali současně dva studenti, jejich práce byly rozděleny podle typu použitých metod učení. V případě této práce šlo o pravděpodobnostní metody. Diplomant v souladu se zadáním provedl rešerši literatury a na jejím základě se podle očekávání rozhodl pro využití zobecněných skrytých markovských modelů (GHMM). Vyšel ze dvou popsaných a hojně citovaných nástrojů, jeden je nezávislý na typu organismu (Augustus), druhý se zaměřuje přímo na houby (CodingQuarry). Inspiroval se obecnými postupy použitými v těchto nástrojích, pozměnil ale jak strukturu modelů (množinu vnitřních stavů), tak i konfiguraci hyperparametrů ovlivňující průchod sekvence modelem a rychlost tohoto průchodu. Ve svém řešení totiž musel zohlednit i otázky efektivity, kvadratická složitost Viterbiho algoritmu pro zobecněný HMM může být pro dané délky vstupních sekvencí příliš vysoká. Pro informaci, trénovací data měla desítky GB, šlo ale samozřejmě o velký počet kontigů a hub. Při použití očekáváme současné použití několika modelů na (meta)genomech v délce jednotek GB.

Diplomant řešil své téma dlouhodobě. Pravidelně jsme se scházeli přibližně po dobu jednoho akademického roku, za tu dobu nenastaly v řešení žádné větší prodlevy. Naše komunikace byla bezproblémová. I když postup prací nebyl vyloženě rychlý a občas se diplomant zasekl na drobných implementačních, či logických chybách, všechny zásadní problémy se nakonec podařilo vyřešit a cíle určené v zadání byly dosaženy. Musím ocenit samostatnost s jakou si pan Zvara doplnil nutné biologické vědomosti. I když použitelnost v cílové aplikaci zmíněné v prvním odstavci zatím nedokážeme odhadnout, nemám k řešení věcné výhrady. Některé překážky, jako je například časová složitost, jsou objektivní a lze je minimalizovat pouze heuristickými postupy. O to se pan Zvara pokusil. Práci nehodnotím nejlepší známkou víceméně pro dílčí nepřesnosti, ať už v implementaci, nebo v textu diplomové práce. Popis je výborný pokud jde o obecnou úroveň, není ale příliš detailní, což může ztížit další využití výsledků práce.

Souhrnně lze říci, že práce splňuje všechny cíle, které jsme si na počátku vytkli. Byl vytvořen a otestován konkrétní pravděpodobnostní algoritmus detekce intronů v genomech hub. Práci doporučuji k obhajobě a hodnotím ji známkou

B — velmi dobře.

V Praze, dne 10. června 2019 Doc. Ing. Jiří Kléma, Ph.D., vedoucí