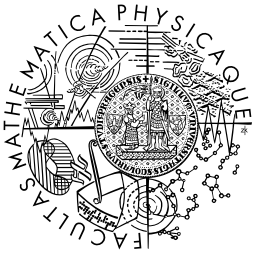

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Richard Eliáš

Analýza datových toků ve databázových systémech

Katedra distribuovaných a spolehlivých systémů

Vedoucí diplomové práce:	RNDr. Pavel Parízek, Ph.D.
Studijní program:	Informatika
Studijní obor:	Umělá inteligence

Praha 2019

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů (dále jen "autorský zákon"), především § 35 a § 60 autorského zákona upravující školní dílo. V případě počítačových programů, jež jsou součástí mojí práce či její přílohou, a veškeré související dokumentace k počítačovým programům (dále jen "software"), uděluji tzv. MIT Licenci.

Obsahem MIT Licence je oprávnění bezúplatně užít software. Toto oprávnění uděluji každé osobě, která má o užití software zájem.

Každá osoba je oprávněna pořídit si kopii software (včetně související dokumentace) bez jakéhokoli omezení a dále je oprávněna bez jakéhokoli omezení software zejména užívat, kopírovat, upravovat, sloučit, publikovat, distribuovat, poskytovat podlicence a / nebo prodávat kopie software a umožnit výkon těchto práv osobám, kterým bude dále software poskytnut. Způsoby užití software ani rozsahu tohoto užití nejsou jakkoli omezeny.

Osoba, která má o užití software zájem je povinna připojit text licenčních podmínek následujícího znění:

V ..... dne ..... Podpis autora

Název práce: Analýza datových toků ve databázových systémech

Autor: Richard Eliáš

Katedra: Katedra distribuovaných a spolehlivých systémů

Vedoucí bakalářské práce: RNDr. Pavel Parízek, Ph.D., Katedra distribuovaných a spolehlivých systémů

Abstrakt: TODO Abstrakt

Klíčová slova: TODO Klucove slova

Title: Analyzing Data Lineage in Database Frameworks

Author: Richard Eliáš

Department: Department of Distributed and Dependable Systems

Supervisor: RNDr. Pavel Parízek, Ph.D., Department of Distributed and Dependable Systems

Abstract: TODO Abstrakt

Keywords: TODO Klucove slova

TODO Podakovanie

Obsah

Úvod

TODO Nejaky uvod...

Donedávna sa myslelo, že úloha ribonukleovej kyseliny, RNA, je obmedzená iba na syntézu bielkovín, buď ako nositeľka genetickej informácie (mRNA), alebo ako prenášač aminokyselín pri ich tvorbe (tRNA). Avšak existuje mnoho ďalších druhov, od relatívne malých molekúl majúcich iba desiatky nukleotidov, ktoré ovplyvňujú expresiu génov (miRNA, siRNA, snRNA, snoRNA a ďalšie) (Carthew a Sontheimer (2009), Kiss (2002)), až po veľké molekuly s tisíckami nukleotidov, ktoré sa podieľajú na tvorbe ribozómu (rRNA).

Spolu s objavmi ďalších funkcií RNA molekúl rastie záujem o nástroje dovoľujúce študovať ich štruktúru. Primárna štruktúra je určená poradím nukleotidov v reťazci RNA. Terciárnu štruktúru získame ich priestorovým usporiadaním. Poslednou a v tejto práci pre nás najdôležitejšou bude sekundárna štruktúra. Tú reprezentuje zoznam nukleotidov ktoré sú spojené väzbou. Spárované nukleotidy sú blízko seba v priestore a tak nám sekundárna relatívne dobre aproximuje terciárnu štruktúru. Predpovedanie terciárnej štruktúry nieje veľmi spoľahlivé ani pre kratšie molekuly. Naopak, pre menšie molekuly a ich sekundárnu štruktúru existujú spoľahlivejšie metódy. Ich prehľad a porovnanie nájdeme v článku od autorov Puton a kol. (2013). Príbuzné štruktúry nám vedia poslúžiť k predpovedaniu konzervovaných častí, dokonca aj veľkých rRNA molekúl (Pánek a kol. (2014)).

Vizualizácia sekundárnej štruktúry RNA sa dá previesť na kreslenie grafu, ktorého vrcholy tvoria nukleotidy a hrany reprezentujú páry medzi nimi. Kreslenie grafov je značne preskúmanou témou, keďže nachádza uplatnenie vo veľa doménach, ako napríklad analýza sociálnych sietí (Scott (1988)) alebo vo všeobecnej analýze dát (Tamassia (2007)).

Cieľom vizualizácie sekundárnej štruktúry RNA je zachytit párovanie nukleotidov v molekule a ideálne všetky ďalšie motívy, ktoré sa v štruktúre vyskytujú, ako napríklad hairpin, bulge, interior a multi-branch loop. Existujúce nástroje na vizualizáciu zväčša volia medzi tromi tipmi reprezentácie štruktúry (Wiese a kol. (2005)): spojnicový graf (linked graph), kruhový graf (circular graph) alebo štandardná štruktúra (classical structure). Aj keď spojnicový a kruhový graf podporujú vizualizáciu párovania báz, motívy sa v nej dajú nájsť len veľmi ťažko, ak vôbec. Preto nám na hľadanie motívov v RNA ostala štandardná reprezentácia štruktúry RNA. Bolo vymyslených množstvo riešení - RNAfold z balíka ViennaRNA (Lorenz a kol. (2011)), VARNA (Darty a kol. (2009)), RnaViz (De Rijk a kol. (2003)), jViz.RNA (Wiese a kol. (2005)), mfold (Zuker (2003)), PseudoViewer (Han a kol. (2002)) alebo RNAView (Yang a kol. (2003)). Avšak iba niektoré z týchto nástrojov a algoritmov sú použiteľné pre vizualizáciu veľkých štruktúr, akou sú napríklad podjednotky rRNA (RNAfold, RnaViz a RNAView). Porovnanie väčšiny z nich nájdeme v článku autorov Ponty a Leclerc (2015).

Vzhľadom k tomu, že je nekonečne mnoho možností ako rozložiť sekundárnu štruktúru, potrebujeme zistiť aké kritéria by malo nakreslenie RNA spĺňať. Nanešťastie tieto kritéria niesú formalizované, avšak niektoré vlastnosti ako napríklad rovinnosť nakreslenia, kreslenie loopov (hairpin, bulge, interior a multi-branch) na kružnice a vrcholy stemu ležiace na jednej priamke sú spoločné pre väčšinu vizualizácii používaných vo vedeckej komunite, ako popisujú Auber a kol. (2006). Ostatné sa prispôsobujú oblasti štúdia, kvôli čomu každý algoritmus nebude vyhovovať veľkej časti používateľov. Dôvod môžeme ilustrovať na ribozomálnej RNA. Štruktúry týchto molekúl majú veľké konzervované časti, ktoré biológovia očakávajú na rovnakom mieste na obrázku, vďaka čomu sa vedia orientovať aj v relatívne veľkej molekule a môžu skúmať a nachádzať tie menej konzervované časti, ktoré sa líšia medzi organizmami. To znamená, že ak chceme štruktúru vizualizovať, potrebujeme ukladať konzervované časti vždy na rovnaké miesto v obrázku.

Potreba vizualizácie, ktorá by zachovávala čo najviac spomínaných vlastností nás viedla k vytvoreniu nového vizualizačného algoritmu založeného na použití šablónovej (vzorovej) molekuly (Eliáš a Hokza (2016)). Algoritmus na vstupe vezme cieľovú štruktúru, ktorú chceme vizualizovať a inú, podobnú štruktúru u ktorej poznáme jej nakreslenie. Túto podobnú molekulu nazývame šablónou. Obe štruktúry sú prevedené do ich stromovej reprezentácie. Následne nájdeme najkratšiu postupnosť editačných operácií, ktoré prevedú strom molekuly šablónovej na vizualizovanú molekulu a rovnako menia aj nakreslenie šablóny na nakreslenie cieľovej molekuly. Vzhľadom k tomu, že editačné operácie ktoré menia nakreslenie odpovedajú minimálnej editačnej vzdialeností medzi stromami, nakreslenie spoločných častí sa nemení. Táto metóda je

teda schopná vizualizovať sekundárnu štruktúru RNA molekuly presne podľa zvyku biológov - podľa poskytnutého vzorového nakreslenia.

Záver

TODO Nejaky zaver

V rámci práce sme vytvorili program TRAVeLer umožňujúci vizualizáciu sekundárnej štruktúry RNA pomocou existujúceho obrázka molekuly, ktorý nám slúži ako predloha.

Reprezentovať sekundárnu štruktúru RNA sme sa rozhodli pomocou stromov. To nám umožnilo použiť *tree-edit-distance* metriku podobnosti dvoch štruktúr. Algoritmus TED nám nedáva len informáciu o tom, s ako vzdialenými štruktúrami pracujeme, ale dáva nám aj návod, ako transformovať šablónovú molekulu na tu cieľovú. U dostatočne podobných štruktúr nám už namapovaná štruktúra dá predstavu, ako bude výsledná vizualizácia vyzerať a ukáže, ktoré časti sú v molekulách spoločné.

Výsledky ukazujú, že ak použijeme štruktúrne dostatočne blízku molekulu ako šablónu, výsledok vizualizácie bude uspokojujúci. S väčšou vzdialenosťou ale počet prekryvov vo výsledných obrázkoch stúpa.

V budúcnosti bude vhodné upraviť naše kresliace algoritmy. Takýmto vylepšením by bola implementácia otáčania vetiev RNA stromov, v prípade, že sme našli prekryv. Druhou možnosťou by bolo pridanie interaktívneho nástroja na úpravu vzniknutých obrázkov, aby bolo možné výsledné vizualizácie ručne upraviť. Tým by sa vyriešil problém s prekryvmi, ktoré by si užívateľ sám odstránil.

V našej práci sme sa úplne vyhli pseudouzlom. To nám umožnilo reprezentovať sekundárnu štruktúru RNA ako usporiadaný, zakorenený strom. Avšak, pseudouzly sú jej dôležitou súčasťou. Možným vylepšením by preto bolo, zohľadniť ich existenciu pri mapovaní. To si ale bude vyžadovať hlbšiu analýzu tohoto problému.

Program bol uvolnený pre používanie biológmi. Budeme očakávať ich reakcie, na základe ktorých budeme implementovať ďalšie vylepšenia programu.

Zoznam použitej literatúry

Auber, D., Delest, M., Domenger, J.-P. a Dulucq, S. (2006). Efficient drawing of rna secondary structure. *Journal of Graph Algorithms and Applications*, **10**(2), 329–351. URL `http://eudml.org/doc/55423`.

Carthew, R. W. a Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, **136**(4), 642–655.

Darty, K., Denise, A. a Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**(15), 1974–1975.

De Rijk, P., Wuyts, J. a De Wachter, R. (2003). RnaViz 2: an improved representation of RNA secondary structure. *Bioinformatics*, **19**(2), 299–300.

Eliáš, R. a Hokza, D. (2016). RNA Secondary Structure Visualization Using Tree Edit Distance. *International Journal of Bioscience, Biochemistry and Bioinformatics*, **6**(1), 9–17. doi: 10.17706/ijbbb.2016.6.1.9-17.

Han, K., Lee, Y. a Kim, W. (2002). PseudoViewer: automatic visualization of RNA pseudoknots. *Bioinformatics*, **18 Suppl 1**, S321–328.

Kiss, T. (2002). Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **109**(2), 145–148.

Lorenz, R., Bernhart, S. H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F. a Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26.

Ponty, Y. a Leclerc, F. (2015). Drawing and editing the secondary structure(s) of RNA. *Methods Mol. Biol.*, **1269**, 63–100.

Puton, T., Kozlowski, L. P., Rother, K. M. a Bujnicki, J. M. (2013). CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res.*, **41**(7), 4307–4323.

Pánek, J., Hajič, J. a Hoksza, D. (2014). Template-based prediction of ribosomal rna secondary structure. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 18–20. doi: 10.1109/BIBM.2014.6999394.

Scott, J. (1988). Social network analysis. *Sociology*, **22**(1), 109–127. doi: 10.1177/0038038588022001007. URL `http://soc.sagepub.com/content/22/1/109.abstract`.

Tamassia, R. (2007). *Handbook of Graph Drawing and Visualization (Discrete Mathematics and Its Applications)*. Chapman & Hall/CRC. ISBN 1584884126.

Wiese, K. C., Glen, E. a Vasudevan, A. (2005). jviz.rna -a java tool for rna secondary structure visualization. *IEEE Transactions on NanoBioscience*, **4**(3), 212–218. ISSN 1536-1241. doi: 10.1109/TNB.2005.853646.

Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H. a Westhof, E. (2003). Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**(13), 3450–3460.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**(13), 3406–3415.

Zoznam obrázkov