Graph parsing application for bioinformatics problems

Semyon Grigorev, Artem Gorokhov
Saint Petersburg State University
7/9 Universitetskaya nab.
St. Petersburg, 199034 Russia
semen.grigorev@jetbrains.com, gorohov.art@gmail.com

Biomedical databases contain huge amounts of rich data which can be represented as a labelled graph. In order to investigate such data, it may be useful to extract connections with specific constraints. One natural way to provide constraints is to specify the language of paths labels which can be done by using of grammars. For example, one can use context-free grammars with productions $\{S \to aSb; S \to \varepsilon\}$ to query paths which labels should take form of $ab; aabb; aaabbb; ...$, or, generally, should belong to the language $L = \{a^n b^n, n \geq 0\}$. This approach is named *context-free path querying* and can be applied to some problems in bioinformatics.

One of the examples is an analysis of graphs where vertices correspond to entities and concepts such as gene or phenotype while edges represent known relationships such as "codes for", "interacts with", etc (UniProt dataset [1]). Querying paths with special constraints may shed light upon unknown before links between vertices, forming the basis for new hypotheses.

Another example of graph structured data is metagenomic assemblies, and the problem is long subsequences detection and reconstruction. Some sequences have specific secondary structure, which can be described in terms of context-free grammar, and this grammar can be used for searching and classification. A lot of research in this area and tools are based on this approach, but most of them are only aimed at linear data processing. Despite

the existence of tools for metagenomic assemblies analysis, context-free search in graph-structured assembly is still a challenge.

Tasks described above can be solved by using common technique named graph parsing — application of classical parsing techniques for graphs; and we have some experience in this field [3, 4]. Our results solve some problems of existing approaches (such as cycles processing problem in [2]), and provide an ability to use GPGPU and multi-core systems for graph parsing which can be useful for huge biological data analysis. Currently we are working on long subsequences of 16s rRNA reconstruction from metagenomic assembly and on entities connections detection. Our aim is to present current results and also to find other applications for this techniques.

References

[1] UniProt Consortium et al. "UniProt: a hub for protein information." *Nucleic acids research.* (2014).

[2] Sevon, Petteri, and Lauri Eronen. "Subgraph queries by context-free grammars." *Journal of Integrative Bioinformatics (JIB)* 5.2 (2008): 157-172.

[3] Grigorev, Semyon, and Anastasiya Ragozina. "Context-Free Path Querying with Structural Representation of Result." *arXiv preprint arXiv:1612.08872* (2016).

[4] Verbitskaia, Ekaterina, Semyon Grigorev, and Dmitry Avdyukhin. "Relaxed Parsing of Regular Approximations of String-Embedded Languages." *International Andrei Ershov Memorial Conference on Perspectives of System Informatics.* Springer International Publishing, 2015.