Graph parsing application for bio 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 labeled graph. In order to investigate such data, it may be useful to extract connections with specific constraints. One of natural way to provide constraints is specify language of paths' labels, which can be done by using different classes of grammars. For example, one can use context-free grammars with productions $\{S \to aSb; S \to \varepsilon\}$ to find paths which labels should looks like $ab; aabb; aaabbb; ...$, or, generally, should contains in language $L = \{a^n b^n, n \geq 0\}$. This approach is named *context-free path querying* and can be useful in some bioinformatic applications.

One of examples is an analysis of graphs where vertices correspond to entities and concepts such as gene, phenotype, and edges represent known relationships such as "codes for", "interacts with", etc (UniProt [1] dataset, for example). Paths with special constraints may provide information about links between vertices were unknown before, forming the basis for new hypotheses.

Another example of graph structured data is metagenomic assemblies, and one of 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 finding and classification. There is a big number of research in this area and tools based on this approach, but most of them aimed on linear data processing. Despite the fact of existence tools for metagenomic assemblies analysis, context-free search in graph structured assembly is still a challenge.

Tasks which described above can be solved by using common technique which is 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 algorithms (such as cycles processing problem in [2]), and provide ability to use GPGPU and multi core systems for graph parsing, which can be useful for huge biological data analysis. Now we are working on long subsequences of 16s rRNA reconstruction from metagenomic assembly, and on entities connections detection. We want to present current results and also we want 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.