BioPaxCOMP: an efficient system for integrating, compressing, and querying BioPAX

Giuseppe Agapito(1), Andrea Greco(1), and Mario Cannataro(1)
(1)Bioinformatics Laboratory, Department of Medical and Surgical Sciences, University Magna Graecia of Catanzaro, Italy

Biological networks and, in particular, biological pathways are composed of thousands of nodes and edges, posing several challenge regarding analysis and storage. The primary format used to represent pathways data is BioPAX (http://biopax.org.)
BioPAX is a standard language that aims to enable integration, exchange, visualization and analysis of biological pathway data. BioPAX is an open and collaborative effort made by the community of researchers, software developers, and institutions and it specifically supports data exchange between pathway data groups. BioPAX is defined in OWL and is represented in the RDF/XML format. OWL (Web Ontology Language) is a W3C standard and is designed for use by applications that need to process the content of information instead of just presenting information to humans. RDF is a standard model for data interchange on the Web.
Although OWL allows a standard representation of pathways, since it is based on XML, it is a verbose and redundant language, so the storage of pathways may be very huge, preventing an efficient transmission and sharing of this data. The typical size of a pathway is related to the organism, for example, the size of Homo Sapiens pathways (from Reactome database) is near to 200 MB on disk. Moreover, integrating pathways data coming from different data sources may require GBytes of space.
A second problem with pathways is related to the possibility to integrate information coming from different data sources to have updated information in a centralized way. There exist several different databases for pathways data that emphasizes different aspect of the same pathway, thus, it could be useful to integrate and annotate together pathways coming from different databases to obtain a centralized and more informative pathway data. The principal obstacle for integrating, storing and exchanging such data is the extreme size growth when several pathways data are merged together, posing several challenges from the computational and archiving point of view. Pathways data can be easily classified as big data, because they meet all the 5V (Volume, Velocity, Variety, Veracity, Value) characteristics typical of Big Data, thus, the necessity to efficiently integrate and compress pathways data arises.
The methodology for pathways data integration is based on the following steps: i) aggregation and validation locally of data coming from several pathway databases, ii) identification and normalization of compounds and reactions identifier and iii) integration. Integration occurs at the level of physical entities, such as proteins and small molecules. This is accomplished by linking interaction and pathway records together if they use the same physical entities (such as from UniProt for proteins) and by adding annotation data from UniProt or GeneOntology.
The methodology for compressing the pathways data uses a specialized XML compressor named BioPaxCOMP that compresses the data and structure of the pathways. The methodology for compressing the pathways data uses a specialized XML compressor based on [1] that compresses the data and structure of the pathways.
To compress the structure, BioPaxCOMP replaces the XML tags with shorter placeholders to take up less space.
To compress the data, BioPaxCOMP uses a library of different compression algorithms specialized for the different data types contained in a pathway: it analyzes the semantics of the data and applies the most suitable compression algorithm.
The main processing steps of BioPaxCOMP are: 1) reading of pathway; 2) structure analysis including placeholders extraction and compression; 3) semantic analysis of the data, where data are separated by data type and compressed separately choosing the most suitable compression algorithms; 4) bundling of the compressed data consisting in putting together the structure and the compressed data.

Summarizing, pathways data of a specific organism (e.g. homo sapiens) coming from different data sources (e.g. Panther, KEEG, BioCarta and Reactome databases) are first downloaded locally, then they are integrated and finally compressed.

We obtain the BioPaxCOMP prototype for the integration and compression of pathways coming from different data sources, that allows an efficient storage and querying of BioPAX pathways. BioPaxCOMP can integrate and compress pathways data coming from different databases, allowing to save space on disk as well as, to efficiently retrieve specific pathways information without decompressing all the data.

BioPaxCOMP can integrate pathways data coming from different databases, can compress them to save space and finally can query compressed pathway data, without decompressing the entire compressed data.

For example, integrating the homo sapiens pathways information coming from Panther, KEEG, BioCarta and Reactome databases, we obtain a complete pathway that describes the biochemical reactions that happen into the human organism in a broader perspective.

Moreover, BioPaxCOMP makes possible to retrieve the genes/proteins that affected some metabolic functions without to decompress the integrated pathways.

References:

[1] Cannataro, Mario, Carmela Comito, and Andrea Pugliese. SqueezeX: Synthesis and compression of XML data.” Information Technology: Coding and Computing, 2002. Proceedings. International Conference on. IEEE, 2002.