GBOL a Genetic Biology Ontology Language

Jasper J. Koehorst[1], Jesse C.J. van Dam[1], Jon Olav Vik[2], Maria Suarez Diez[1] and Peter J. Schaap[1]

[1]Systems and Synthetic biology, Wageningen University & Research, Wageningen, The Netherlands
[2]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences (IHA), Faculty of Life Sciences (BIOVIT), Norwegian University of Life Sciences (NMBU), PO Box 5003, Ås, Norway

E-mail: jasper.koehorst@wur.nl, peter.schaap@wur.nl

1. Introduction

Sequence data is growing exponentially and so the number of tools to analyze these data are growing. Each tool generates its own type of data and combining these results is often needed to gain the best insights. Such an integration of data and computational predictions requires data provenance to included so that the quality of the results can be assessed.

Currently, the GenBank, EMBL and GFF formats are used to store genomic DNA sequences and associated annotations. However, these formats provide limited support for storing provenance. Moreover, in these formats there exist multiple options to store annotation resulting in limited interoperability which hampers querying at large scale. To overcome these problems, we use solutions based on semantic technologies which enable data interlinking to store genetic data and annotations. To this end we have defined the Genetic Biology Ontology Language (GBOL).

2. Results

GBOL includes an ontology that defines relationships between concepts that are typically found in GenBank, EBML and GFF formats. GBOL provides a non-ambiguous description of annotation and includes support for data provenance. Core to the GBOL ontology is consistent representation of the annotation of genomic DNA, genes, transcripts, proteins and functional annotation for prokaryotic and eukaryotic organisms, see figure 1. We used the Faldo [1] ontology to store genetic loci information. We used and extended ProvO [2] to store Provenance information. Whenever applicable terms are linked to the Sequence Ontology and SBOL ontologies. Associated to the ontology, we developed a schema validation based on the associated ShEx definition, which is included into a Java and R API.

GBOL data is stored in any of the linked data formats (RDF), such as Turtle. Afterwards, JSON-LD framing API [3] can be used to serialize the GBOL linked data as JSON, which can be subsequently serialized as YAML [4]. This format mimics the indentation structure as seen in GenBank and EMBL formats, however has integrated support to add additional information.

GBOL and its associated API have been used to develop SAGeR-P. SAGeR-P provides a user-friendly interface allowing users to browse and compare genomic information. Currently it contains functional annotation data of more than 5000 prokaryotic organisms. For each organism GBOL stores and integrates annotation from InterProScan, SignalP, TMHMM, Wolf PSORT, BLAST, EnzDP, Prodigal and RNAmmer. We applied these results to perform large scale function comparison among bacteria [5,6].

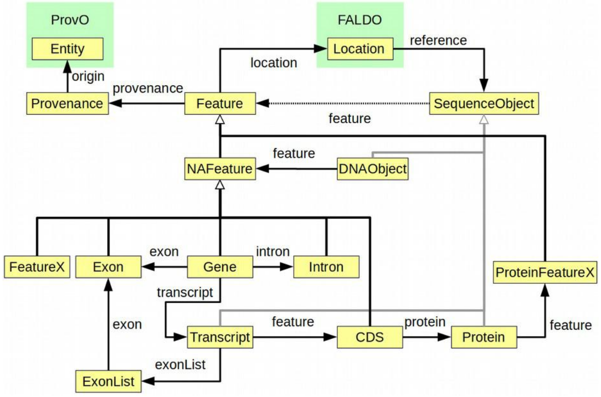

Figure 1. Simplified overview of the core elements in the GBOL ontology

3. Discussion

Large scale analysis of often heterogeneous biological data is hampered by a lack of interoperability. The main rationale of applying formalized information models is to provide semantic standards to improve the exchange of information. Applying formal ontologies to biodata thus will make this data inter-operable and FAIR while also providing support for automatic reasoning and processing. GBOL provides a formal representation of genomic entities, along with their properties and relations.

References

1. J.T. Bolleman, et al. “FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation”. *Journal of Biomedical Semantics*, 2016.
2. PROV-O: The PROV Ontology. https://www.w3.org/TR/prov-o/
3. JSON-LD 1.0 Processing Algorithms and API. https://www.w3.org/TR/json-ld-api/
4. YAML 1.2. http://yaml.org/
5. J. J. Koehorst, et al. "Comparison of 432 Pseudomonas Strains through Integration of Genomic, Functional, Metabolic and Expression Data." *Scientific Reports*, 2016
6. J.J. Koehorst, et al. "Protein Domain Architectures Provide a Fast, Efficient and Scalable Alternative to Sequence-based Methods for Comparative Functional Genomics." *F1000Research*, 2016, PMID: 27703668