Scientific Symposium FAIR Data Sciences for Green Life Sciences
12 December 2018 | Wageningen | the Netherlands

Conference Proceedings | DOI: https://doi.org/10.18174/FAIRdata2018.16288

SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles

Schaap, P.J.[1], J.J. Koehorst[1], J.C.J. van Dam[1], E. Saccenti[1], V.A.P. Martins dos Santos[1], M. Suarez-Diez[1]

[1] Laboratory of Systems and Synthetic biology, Wageningen University & Research, Stippeneng 4, 6708WE Wageningen, the Netherlands
Corresponding author's e-mail: peter.schaap@wur.nl

There are currently more than 150.000 sequenced genomes available from which considerable amounts of information can be extracted. However, annotation information is often not interoperable, static, lacks provenance and is quickly outdated. Keeping these datasets up-to-date, and interoperable is a challenging and a computational intensive task. Using a Semantic Annotation Platform with Provenance (SAPP) we have made a significant step towards obtaining FAIR annotated genome data with linked provenance. Adding a high level of interoperability to functional genome annotations enables bottom-up and top-down in-depth analysis and comparative genomics at unlimited scales. Using open semantic data frameworks, phenotypic and other heterogeneous data sources can be incorporated making the data increasingly more valuable.

SAPP accepts (non-) annotated sequence files which are converted into an RDF data structure using the GBOL ontology. Within SAPP, structural and functional annotation is performed using a modular approach incorporating existing annotation tools for the detection of genetic elements, CRISPRs and protein functions. The resulting annotation data and associated metadata are stored in a compressed graph database making the data directly interoperable.

Discussion: Large scale genomics requires a management system that links genomic data with provenance. SAPP functionalities are unique since none of the existing de novo annotation pipelines implement Semantic Web technologies. By incorporating GBOL, a strictly defined ontology ensures the interoperability and reusability of the data. SAPP thereby fulfils the applicable requirements for data FAIRness. Through SPARQL complex questions across multiple domains combined with external resources, such as UniProt can easily be answered. Additionally, likelihood values can be integrated. For instance, study of E-value distribution on instances of a protein domain across multiple genomes can inform dynamic optimal threshold selection for functional annotation. SAPP reduces the complexity of genome analysis through the incorporation of semantic web technologies while capturing all information obtained through various annotation modules. It ensures that all data is consistent and accessible through a uniform and simplified approach enabling high throughput and large-scale research.

References:

SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles. J.J. Koehorst, Jesse C.J. van Dam, Edoardo Saccenti, Vitor A.P. Martins dos Santos, Maria Suarez-Diez, Peter J. Schaap. *Bioinformatics,* 34 (8), 1401-1403.

Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining. Jesse C.J. van Dam, Jan J. Koehorst, Olav Vik, Peter J. Schaap, Maria Suarez-Diez https://doi.org/10.1101/184747