Bioinformatics, 2017, 1–3
doi: 10.1093/bioinformatics/btx767
Advance Access Publication Date: 23 November 2017
Applications Note

OXFORD

Genome analysis

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

Jasper J. Koehorst[1,*], Jesse C. J. van Dam[1], Edoardo Saccenti[1], Vitor A. P. Martins dos Santos[1,2], Maria Suarez-Diez[1] and Peter J. Schaap[1,*]

[1]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen 6708 WE, The Netherlands and [2]LifeGlimmer GmbH, 12163 Berlin, Germany

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on October 10, 2017; revised on November 9, 2017; editorial decision on November 21, 2017; accepted on November 22, 2017

Abstract

Summary: To unlock the full potential of genome data and to enhance data interoperability and reusability of genome annotations we have developed SAPP, a Semantic Annotation Platform with Provenance. SAPP is designed as an infrastructure supporting FAIR *de novo* computational genomics but can also be used to process and analyze existing genome annotations. SAPP automatically predicts, tracks and stores structural and functional annotations and associated dataset- and element-wise provenance in a Linked Data format, thereby enabling information mining and retrieval with Semantic Web technologies. This greatly reduces the administrative burden of handling multiple analysis tools and versions thereof and facilitates multi-level large scale comparative analysis.

Availability and implementation: SAPP is written in JAVA and freely available at https://gitlab.com/sapp and runs on Unix-like operating systems. The documentation, examples and a tutorial are available at https://sapp.gitlab.io.

Contact: jasperkoehorst@gmail.com or peter.schaap@wur.nl

1 Introduction

Managing the genomic data deluge puts specific emphasis on the ability of machines to automatically find and use the data. To meet this demand and to extract maximum benefit from research investments, digital objects should be Findable, Accessible, Interoperable and Reusable (i.e. FAIR) (Wilkinson *et al.*, 2016).

Genome annotation data is usually findable and accessible through public repositories in which the data is linked to metadata providing detailed descriptions of the data acquisition and generation process. Interoperability reflects the potential for seamless integration of data from independent sources. Currently, genome comparisons usually involve a laborious process of data retrieval, modification and standardization (canonicalization). Reusability requires rich metadata with provenance for each annotation. Current standard formats (GenBank, EMBL or GFF3) retain the output of the prediction tools (for example for gene identification) but only when they score better than a predefined, often pragmatic, prediction threshold. Detailed information of the actual prediction scores is lost. This hampers critical re-examination of the results.

Because existing genome annotation data is hard to be made FAIR and managing of FAIR genome annotation data requires a considerable administrative load, we developed SAPP, a semantic framework for large scale comparative functional genomics studies. SAPP can automatically annotate genome sequences using standard tools. The unique characteristic of SAPP is that the annotation results and their provenance are stored in a Linked Data format, thus enabling the deployment of mining capabilities of the Semantic Web. As the automatic annotations are incorporated into a dynamic framework, SAPP supports periodic querying, comparison and linking of diverse annotation sources, resulting in up-to-date genome

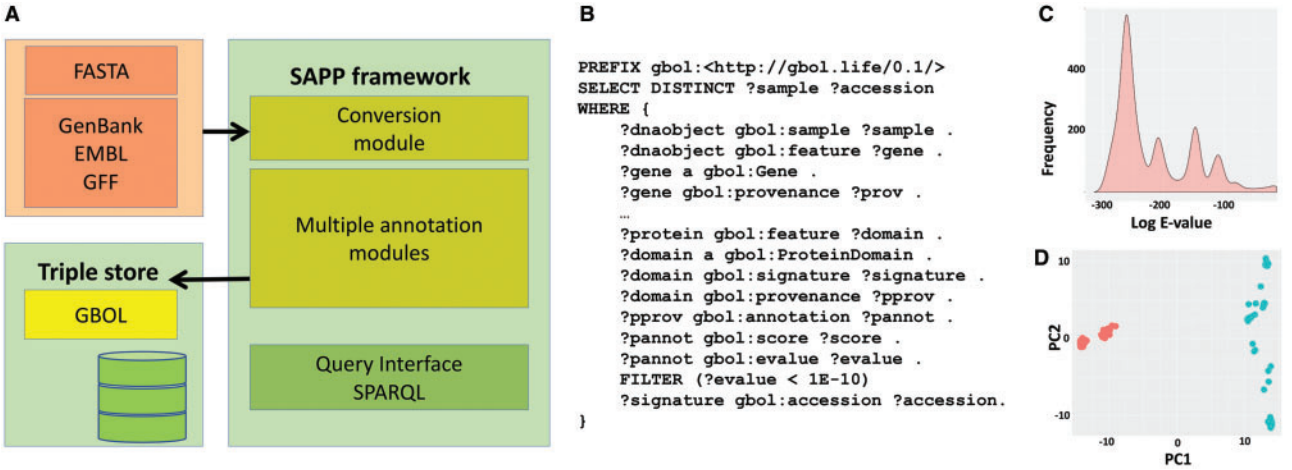

Fig. 1. (**A**) The conversion module imports genome sequences in common formats. Annotation modules perform common tasks such as gene, tRNA, protein and protein domain annotation. Results are stored as Linked Data and consistency is ensured by the GBOL stack. (**B**) SPARQL query to retrieve the E-value score of the instances of the protein domain PF00465 across multiple bacterial genomes. (**C**) Distribution of E-values for protein domain PF00465 across multiple bacterial genomes: note the multimodality of the distribution. (**D**) Principal component analysis of functional similarities of 100 bacterial genomes from the *Streptococcus* (blue) and the *Staphylococcus* (orange) genera. PC1 and PC2 account for 51.4 and 10.1% of the variance in the dataset respectively

annotations. By interrogating metadata as part of a digital annotation object, annotation data becomes interoperable as the extraction procedure requires no additional standardization process.

2 Implementation

SAPP accepts annotated and non-annotated sequence files which are converted into an RDF data structure using the GBOL ontology (van Dam *et al.*, 2017). Within SAPP, structural and functional annotation is performed using add-on modules incorporating existing standard annotation tools such as Prodigal and Augustus (Hyatt *et al.*, 2010; Stanke and Morgenstern, 2005). Modules for tRNA, tmRNA, rRNAs, protein domain and CRISPR repeats annotation are also available. New modules can be added. Annotation data and metadata are stored in a compressed graph database (Fernández *et al.*, 2013), as shown in Figure 1A.

Genome annotations can be exported to standard formats. All data can be directly queried and compared using the SPARQL endpoint or via the GBOL API (Java/R). Complex queries can be performed on multiple genomes while simultaneously taking meta-data into account. A SPARQL query example is provided in Figure 1B. Examples to query SAPP from R, Java or Python, a tutorial and a list of publications in which SAPP was used can be found at http://sapp.gitlab.io.

3 Results and discussion

Reproducible computational research requires a management system that links data with data provenance. Interoperability requires a strictly defined ontology. Using and sharing Linked Data based on controlled vocabularies and ontologies ensures the interoperability and reusability of the data. SAPP functionalities are unique since none of the existing *de novo* annotation pipelines implement Semantic Web technologies. SAPP generated data fulfil the applicable requirements for data FAIRness proposed by Wilkinson *et al.* (2016).

For input and output, these tools interact directly with the database thereby forcing automatic linkage of data and provenance. In this way there is no need to work with predefined thresholds on the parameters controlling the annotation output. SAPP uses a controlled vocabulary to describe genome annotations. Consistency is ensured through the GBOL Stack (van Dam *et al.*, 2017).

The GBOL ontology enables consistent genome annotation while integrating dataset-wise and element-wise provenance. The element-wise provenance is the statistical basis or score of each individual annotation, whereas the dataset-wise provenance refers to the programs, versions thereof and parameters used for the complete annotation of the (set of) sequences under study.

GBOL makes use of existing ontologies: PROV-O for activity capturing (Lebo *et al.*, 2013); FOAF for agent information (Brickley and Miller, 2007); BIBO for article information stored within the annotation files (Giasson and D'arcus, 2008); SO for sequence information (Eilbeck *et al.*, 2005); FALDO for genomic location (Bolleman *et al.*, 2016), among many others. We refer the reader to van Dam *et al.* (2017) for detailed information on the integrated ontologies and the data model.

Annotations can be evaluated through critical examination of the provenance. The use of SPARQL allows complex queries across data annotated with SAPP and in direct comparison of these annotations with external resources, such as UniProt. Additionally for specific questions, likelihood values can be integrated, normalized or corrected for multiple testing. For instance, study of E-value distribution on instances of a protein domain across multiple genomes can inform optimal threshold selection, as shown in Figure 1C. SAPP implements existing tools: consistency of SAPP annotation and a comparison with deposited annotations is shown and discussed in Koehorst *et al.* (2016).

By querying multiple consistently annotated genomes simultaneously, large scale functional comparisons can be performed without additional conversion steps [see Fig. 1D and Koehorst *et al.* (2017)].

These examples demonstrate that by adopting FAIR principles to genome annotation, knowledge discovery is facilitated.

Funding

This work has received funding from the Research Council of Norway, No. 248792 (DigiSal) and from the European Union FP7 and H2020 under grant agreements No. 305340 (INFECT), No. 635536 (EmPowerPutida), Synthetic Biology Investment Theme (KB-32) from Wageningen University & Research, and No. 634940 (MycoSynVac).

Conflict of Interest: none declared.

References

Bolleman,J. *et al*. (2016) FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation. *J. Biomed. Seman*., **7**, 1–19.

Brickley,D. and Miller,L. (2007) Foaf vocabulary specification 0.91.

Eilbeck,K. *et al*. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol*., **6**, R44.

Fernández,J.D. *et al*. (2013) Binary RDF representation for publication and exchange (HDT). *Web Semant. Sci. Serv. Agents World Wide Web*, **19**, 22–41.

Giasson,F. and D'arcus,B. (2008) Bibliographic ontology. Technical report, Technical report.

Hyatt,D. *et al*. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

Koehorst,J.J. *et al*. (2016) Comparison of 432 pseudomonas strains through integration of genomic, functional, metabolic and expression data. *Sci. Rep*., **6**.

Koehorst,J.J. *et al*. (2017) Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics. *F1000Research*, **5**, 1987.

Lebo,T. *et al*. (2013) Prov-o: The prov ontology. Technical report, W3C Recommendation.

Stanke,M. and Morgenstern,B. (2005) Augustus: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res*., **33**, W465–W467.

van Dam,J.C.J. *et al*. (2017) Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining. *bioRxiv*, **184747**, 1–9.

Wilkinson,M.D. *et al*. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.