Tools for genome annotation
SAPP/GBOL/Empusa

Jasper Koehorst
jasper.koehorst@wur.nl

Laboratory of Systems and Synthetic Biology
Fairbydesign.nl

POSTER + DEMO 1 & 2
Example questions e.g.:

- What (genes) distinguish species that have a desired trait?
- Which enzymes are there that can catalyze reaction X (maybe with different cofactors?)

Consistently annotated genomes that can be mined
Requirements for genome mining

• A semantic annotation platform that incorporates common tools and stores the results in “proper” format. SAPP

• A definition of the “proper format”: definitions of biological terms and their relationships: GBOL ontology

• Interface to use the ontology: GBOL stack

• Tools to develop all of these: Empusa: code generator
GBOL: Genome Biology Ontology Language

<table>
<thead>
<tr>
<th>Sub domain</th>
<th>Classes</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomic locations</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Genes transcripts and features</td>
<td>114</td>
<td>133</td>
</tr>
<tr>
<td>Document structure</td>
<td>27</td>
<td>107</td>
</tr>
<tr>
<td>Dataset-wise provenance</td>
<td>22</td>
<td>54</td>
</tr>
<tr>
<td>Element-wise provenance</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>BIBO</td>
<td>59</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 1. Overview of domains, classes and properties described by GBOL ontology. Note that some properties might be in multiple subdomains.

To meet the requirements and to ensure interoperability, an extendable ontology developed for the application of semantic web technologies in genome annotation and mining. As such GBOL was developed focusing on its function as a file format and as a database schema and has the capability to describe complete synthetic constructs and the interactions between ontology (OWL), the API and the Shape Expression Language (SBOL) (Van Dam et al. 2015). In this respect, ontologies already exist for various aspects of biology (FALDO) only purpose is to unambiguously store genetic locations on a sequence. The Synthetic Biology Open Language (SBOL) (Foster et al. 2012) was designed to consistently store feature predictions with evidence provenance and therefore none of these tools provides scale functional analyses. Modules have been developed to serialize the linked data (RDF) and to generate a plain text format files.

To further simplify the ontology, every attribute is defined as a direct property within the class that links to either a string, an integer, another object or a class in an enumeration set. For example, the object 'Gene' has an attribute 'geneName' that can be a value of 'GeneName'. To further simplify the ontology, every attribute is defined as a direct property within the class that links to either a string, an integer, another object or a class in an enumeration set. For example, the object 'Gene' has an attribute 'geneName' that can be a value of 'GeneName'.

The ontology contains 251 classes that can be categorized into 6 broad domains (Table 1). The ontology contains 251 classes that can be categorized into 6 broad domains (Table 1). In GBOL, sequences have been split into two main elements: Firstly, the Genome Biology Ontology (GBOL), which directly integrates evidence provenance and therefore none of these tools provides scale functional analyses. Modules have been developed to serialize the linked data (RDF) and to generate a plain text format files. To meet the requirements and to ensure interoperability, an extendable ontology developed for the application of semantic web technologies in genome annotation and mining. As such GBOL was developed focusing on its function as a file format and as a database schema and has the capability to describe complete synthetic constructs and the interactions between ontology (OWL), the API and the Shape Expression Language (SBOL) (Van Dam et al. 2015).

To further simplify the ontology, every attribute is defined as a direct property within the class that links to either a string, an integer, another object or a class in an enumeration set. For example, the object 'Gene' has an attribute 'geneName' that can be a value of 'GeneName'. To further simplify the ontology, every attribute is defined as a direct property within the class that links to either a string, an integer, another object or a class in an enumeration set. For example, the object 'Gene' has an attribute 'geneName' that can be a value of 'GeneName'.

To further simplify the ontology, every attribute is defined as a direct property within the class that links to either a string, an integer, another object or a class in an enumeration set. For example, the object 'Gene' has an attribute 'geneName' that can be a value of 'GeneName'. To further simplify the ontology, every attribute is defined as a direct property within the class that links to either a string, an integer, another object or a class in an enumeration set. For example, the object 'Gene' has an attribute 'geneName' that can be a value of 'GeneName'.

The ontology contains 251 classes that can be categorized into 6 broad domains (Table 1). In GBOL, sequences have been split into two main elements: Firstly, the Genome Biology Ontology (GBOL), which directly integrates evidence provenance and therefore none of these tools provides scale functional analyses. Modules have been developed to serialize the linked data (RDF) and to generate a plain text format files. To meet the requirements and to ensure interoperability, an extendable ontology developed for the application of semantic web technologies in genome annotation and mining. As such GBOL was developed focusing on its function as a file format and as a database schema and has the capability to describe complete synthetic constructs and the interactions between ontology (OWL), the API and the Shape Expression Language (SBOL) (Van Dam et al. 2015).

The ontology contains 251 classes that can be categorized into 6 broad domains (Table 1). In GBOL, sequences have been split into two main elements: Firstly, the Genome Biology Ontology (GBOL), which directly integrates evidence provenance and therefore none of these tools provides scale functional analyses. Modules have been developed to serialize the linked data (RDF) and to generate a plain text format files. To meet the requirements and to ensure interoperability, an extendable ontology developed for the application of semantic web technologies in genome annotation and mining. As such GBOL was developed focusing on its function as a file format and as a database schema and has the capability to describe complete synthetic constructs and the interactions between ontology (OWL), the API and the Shape Expression Language (SBOL) (Van Dam et al. 2015).

The ontology contains 251 classes that can be categorized into 6 broad domains (Table 1). In GBOL, sequences have been split into two main elements: Firstly, the Genome Biology Ontology (GBOL), which directly integrates evidence provenance and therefore none of these tools provides scale functional analyses. Modules have been developed to serialize the linked data (RDF) and to generate a plain text format files. To meet the requirements and to ensure interoperability, an extendable ontology developed for the application of semantic web technologies in genome annotation and mining. As such GBOL was developed focusing on its function as a file format and as a database schema and has the capability to describe complete synthetic constructs and the interactions between ontology (OWL), the API and the Shape Expression Language (SBOL) (Van Dam et al. 2015).
SAPP: Annotation information storage

- Wrapper to commonly used annotation tools (prokaryotes and eukaryotes) that generates FAIR data
- Examples:
  - Uniform annotation of over 80,000 bacterial species.
  - Uniform annotation of eukaryotes e.g. (human, fish, insects, plants, ...)

Diagram: Diagram showing genome assembly, GeneMark, and various annotation steps including Exon 1, Exon 2, Exon 3, Transcript 1, 5' UTR, CDS, 3' UTR, Domain X, etc.
Modular design of SAPP

Conversion types
- EMBL / GenBank
- FASTA
- GFF
- QTL
- VCF
- ...

Genetic elements
- Gene prediction
- tRNA/rRNA
- Crispr
- ...

Functional annotation
- BLAST
- Enzyme predictions
- Domain annotation
- Signal peptides
- Transmembrane
- Localization
- ...

Conversion types
- EMBL / GenBank
- FASTA
- GFF
- QTL
- VCF
- ...

Genetic elements
- Gene prediction
- tRNA/rRNA
- Crispr
- ...

Functional annotation
- BLAST
- Enzyme predictions
- Domain annotation
- Signal peptides
- Transmembrane
- Localization
- ...
Tool development for FAIR genome annotation

- **SAPP**: an annotation platform

**Developer:**

- **GBOL stack:**
  - GBOL ontology (backbone)
  - Java/R Api
  - Owl/ShEx
  - Interface gate keeper

- Code generator: **Empusa** useful for developers
Availability

- **SAPP** Koehorst et al Bioinformatics 2017 [https://sapp.gitlab.io](https://sapp.gitlab.io)
- **Empusa**: [https://gitlab.com/Empusa](https://gitlab.com/Empusa)
- **GBOL**: Documentation & namespace: [http://gbol.life/0.1/](http://gbol.life/0.1/)