B Proposal: Encoding Gene Protein Associations in SBML

This specification drafted by *Brett G. Olivier* and *Frank T. Bergmann* (2013) with contributions by members of the *FBC working group* as well as *FBC* and *SBML* communities. It builds on and supersedes the proposal included in Section 6 of the Flux Balance Constraints package version 1 specification and as such implemented as an annotation in libSBML. (November 19, 2013).

B.1 Introduction and motivation

Genome scale reconstructions (GSR) are widely used for biotechnological purposes for example creating or increasing the yield of a specific product, rerouting metabolic pathways and increasing biomass yield on specific substrates. In order to facilitate this GSR's are currently annotated with gene-protein relationships which relates particular metabolic genes (implicitly their product) with a protein (typically a protein, enzyme or transporter) that carries a specific metabolic flux. Using this knowledge it it is e.g. possible to perform 'in silico' gene knockout experiments or otherwise relate particular fluxes to measured gene expression levels.

The problem that this proposal seeks to address is that currently there is no standard way of encoding these gene-protein relationships (GPR's) in an unambiguous and interoperable way. Currently this information is encoded using free form annotations (typically stored as **Notes** in SBML encoded GSR models).

```
<notes>
 <html:p>GENE ASSOCIATION: (b2234 and b2235) or (b2675 and b2676)</html:p>
 <html:p>GENE_ASSOCIATION: ( Rv2243 or Rv0649 ) and Rv2244 </html:p>
 <html:p>GENE_ASSOCIATION:( peg.1405 ) or ( peg.565  and  peg.567 )</html:p>
 <p>GENE_ASSOCIATION: (3074.1) or (3073.1) or (Rv0649 and 3074.1)</p>
 <p>GENE_ASSOCIATION: (CRv4_Au5.s1.g1623.t1 or CRv4_Au5.s2.g9153.t1)</p>
</notes>
```

Each of these GPR annotations was taken from a different SBML encoded model and uses a slightly different syntax either in the 'key' or the 'value' (bracketing, syntax, attribute definition, etc.). Note that these expressions contain either one or more than one identifier related in a logical expression. Currently there is no way in SBML Level 3 Version 1 Core or the Flux Balance Constraints package to encode such model properties – a situation this proposal seeks to address.

B.2 Proposed syntax and semantics

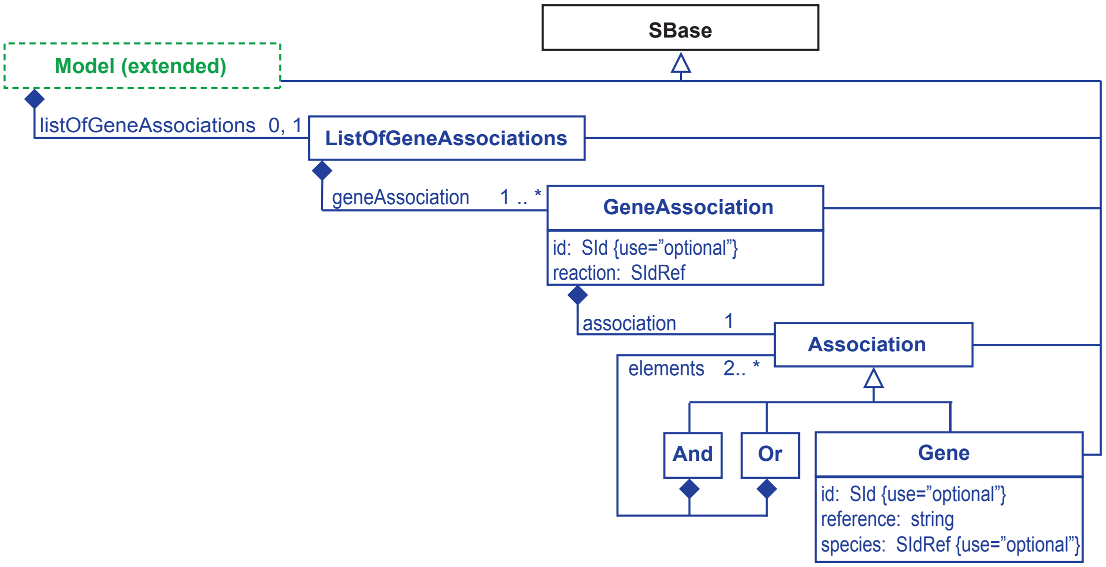

*Figure 7: A UML representation of the Flux Balance Constraints package. Derived from **SBase**, the FBC classes inherit support for constructs such as SBML **Notes** and **Annotation**'s. See Section 1.4 for conventions related to this figure. The individual classes are further discussed in the text.*

B.3 The extended Model class

The FBC listOfGeneAssociations

The **ListOfGeneAssociations** extends SBML Level 3 Version 1 Core, is derived from ***SBase*** and inherits the attributes `metaid` and `sboTerm` as well as the subcomponents for **Annotation** and **Notes** (as shown in Figure 7). If defined **ListOfGeneAssociations** must contain at least one **GeneAssociation** (as defined below in Section B.5).

B.4 The extended Species class

The Flux Balance Constraints package Gene Association Proposal (this document) extends the SBML Level 3 Version 1 Core **Species** class (in addition to `charge` and `chemicalFormula`) with the addition of an attribute `isGene`.

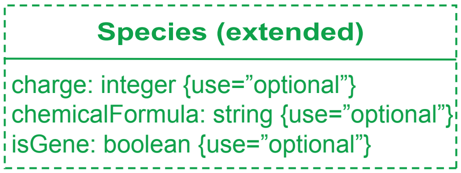
Species (extended)
charge: integer {use="optional"}
chemicalFormula: string {use="optional"}
isGene: boolean {use="optional"}

*Figure 8: A UML representation of the extended **SBML Species** class used in the Flux Balance Constraints package. See Section 1.4 for conventions related to this figure.*

The `isGene` *attribute*

The optional attribute `isGene` contains a `boolean` referring to the fact that the **Species** is not a metabolite that should be included in the reaction network but rather represents a **Gene** product that participates in an **Association** . In addition for a **Species** where `isGene="true"` there should be at least one **Gene** that refers to it via its `species` attribute (for more details see Section B.7).

```
<species metaid="meta_b2013" id="b2013" compartment="Cytosol" boundaryCondition="true"
 hasOnlySubstanceUnits="false" fbc:isGene="true" />
```

B.5 The FBC GeneAssociation class

The Flux Balance Constraints package defines a **GeneAssociation** class that derives from ***SBase*** and inherits the attributes `metaid` and `sboTerm`, as well as the subcomponents for **Annotation** and **Notes**. As shown in Fig. Figure 7 the **GeneAssociation** class connects a **Reaction** (flux) with one or more gene (or gene products) in some form of association.

The `id` *attribute*

The **GeneAssociation** class defines an optional attribute: `id` of type `SId`

The `reaction` *attribute*

The required `reaction` attribute of type `SIdRef`. This attribute must refer to a **Reaction** element defined within the enclosing model.

The `association` *element*

Each **GeneAssociation** must contain a single **Association** , however, as described in Section B.6 an **Association** is an abstract class such that `association` should contain an instance of one of its sub-classes **And** , **Or** or **Gene** .

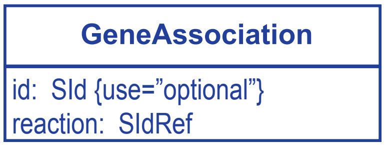

Figure 9: *A UML representation of the Flux Balance Constraints package* ***GeneAssociation*** *class. See Section 1.4 for conventions related to this figure.*

Encoding the ***GeneAssociation***

As described in Section B.5 the **GeneAssociation** is simply a container that relates a **Reaction** to one of three types of **Association** , either holding a single **Gene** or two or more genes in an **And** or **Or** relationship. For example the following typical gene–protein association

```
((B3670 and B3671) or (B0077 and B0078) or (B3768 and B3769 and B3767))
```

from the BiGG database *E. coli* reconstruction (iJR904) Reed et al. (2003); Schellenberger et al. (2010) is now encoded in XML as:

```
<fbc:geneAssociation fbc:id="ga_29" fbc:reaction="R_ACHBS">
 <fbc:or>
  <fbc:and>
   <fbc:gene fbc:reference="b3670"/>
   <fbc:gene fbc:reference="b3671"/>
  </fbc:and>
  <fbc:and>
   <fbc:gene fbc:reference="b0077"/>
   <fbc:gene fbc:reference="b0078"/>
  </fbc:and>
  <fbc:and>
   <fbc:gene fbc:reference="b3768"/>
   <fbc:gene fbc:reference="b3769"/>
   <fbc:gene fbc:reference="b3767"/>
  </fbc:and>
 </fbc:or>
</fbc:geneAssociation>
```

B.6 The FBC Association class

The Flux Balance Constraints package defines an abstract **Association** class that is derived from ***SBase*** and inherits the attributes `metaid` and `sboTerm`, as well as the subcomponents for **Annotation** and **Notes**. It represents either a single gene or a collection of genes and should only ever be instantiated as one of its subclasses: **Gene** (Section B.7), **And** (Section B.8) and **Or** (Section B.9).

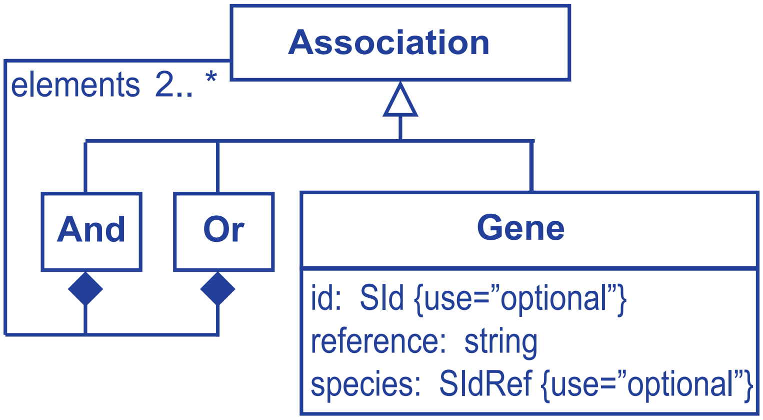

Figure 10: *A UML representation of the Flux Balance Constraints package* ***Association*** *and derived classes. See Section 1.4 for conventions related to this figure.*

B.7 The FBC Gene class

The Flux Balance Constraints package defines a **Gene** class that represents a gene (or gene product) is derived from an **Association** and thereby inherits the ***SBase*** attributes `metaid` and `sboTerm`, as well as the subcomponents for **Annotation** and **Notes** as described in Figure 10.

It is highly recommended that for future interoperability genes should be annotated using the inherited MIRIAM compliant SBML **Annotation** mechanism thus reducing the ambiguity of using an overloaded, semantically meaningful `reference`.

The `id` *attribute*

The **GeneAssociation** class defines an optional attribute `id` of type `SId`.

The `species` *attribute*

The optional attribute `species` attribute of type `SIdRef` can refer to a **Species** element defined within the enclosing model. The intention here is to allow gene–protein associations to be linked to **Species** which may represent them in the model thus bridging two conceptually different (yet equally valid) ways of representing such relations. This attribute should be used in conjunction with the extended **Species** attribute `isGene` (see Section B.4 for details).

The `reference` *attribute*

The `reference` attribute holds a string referring to the nameof a gene (or gene product). However, as seen in the example shown in Section B.1 there is no set format for this identifier e.g. `Rv0649`, `3074.1` or `CRv4_Au5.s2.g9153.t1`. While ideally some form of restriction should be placed on the value of `reference` at this point it is left as a best practice suggestion that this attribute's value conform to the definition of an `SId`. As an example take an existing GPR annotation:

```
<p>GENE_ASSOCIATION: (Rv0649)</p>
```

this can now be formally (and unambiguously) encoded as:

```
<fbc:geneAssociation fbc:id="synth1" fbc:reaction="R_HMTH" fbc:species="s_Rv0649">
 <fbc:gene metaid="meta_gene_1" fbc:reference="Rv0649">
  <annotation>
   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:vCard="http://www.w3.org/2001/vcard-rdf/3.0#"
    xmlns:bqbiol="http://biomodels.net/biology-qualifiers/"
    xmlns:bqmodel="http://biomodels.net/model-qualifiers/">
    <rdf:Description rdf:about="#meta_gene_1">
     <bqbiol:is>
      <rdf:Bag>
       <rdf:li rdf:resource="http://identifiers.org/kegg.genes/mtu:Rv0649"/>
      </rdf:Bag>
     </bqbiol:is>
    </rdf:Description>
   </rdf:RDF>
  </annotation>
 </fbc:gene>
</fbc:geneAssociation>

<species id="s_Rv0649" fbc:isGene=true compartment="Cytosol" hasOnlySubstanceUnits="false"
                                       boundaryCondition="true" constant="true"/>
```

B.8 The FBC And class

The Flux Balance Constraints package defines an **And** class that represents a gene (or gene product) and is derived from and **Association** and thereby inherits the ***SBase*** attributes `metaid` and `sboTerm`, as well as the subcomponents for **Annotation** and **Notes** as described in Figure 10. This class represents a set of two or more associations that are related in an order independent *'and'* relationship.

The `elements` *element*

Each **And** must contain two or more instances (not necessarily of the same type) of any **Association** subclass (**And** , **Or** , **Gene**).

```
<fbc:geneAssociation fbc:id="ga_18" fbc:reaction="R_ACACCT">
 <fbc:and>
  <fbc:gene fbc:reference="b2222"/>
  <fbc:gene fbc:reference="b2221"/>
 </fbc:and>
```

```
</fbc:geneAssociation>
```

B.9 The FBC Or class

The Flux Balance Constraints package defines an **Or** class that represents a gene (or gene product) and is derived from and **Association** and thereby inherits the ***SBase*** attributes `metaid` and `sboTerm`, as well as the subcomponents for **Annotation** and **Notes** as described in Figure 10. This class represents a set of two or more associations that are related in an order independent *'or'* relationship.

The `elements` *element*

Each **Or** must contain two or more instances (not necessarily of the same type) of any **Association** subclass (**And** , **Or** , **Gene**).

```
<fbc:geneAssociation fbc:id="ga_16" fbc:reaction="R_ABTA">
 <fbc:or>
  <fbc:gene fbc:reference="b2662"/>
  <fbc:gene fbc:reference="b1302"/>
 </fbc:or>
</fbc:geneAssociation>
```