Cell Signaling Pathway Annotation and Visualization for Dictybase

Guidance Document

Draft Version 0.1

Authors:

**Matthew Dapas**

Ph.D. Student, Driskill Graduate Program

Northwestern University Biomedical Informatics Center

Table of Contents

I. Introduction & Background 3

About this Document 3

Initial Pathway for Proof-of-Concept 3

Reference Documents 3

Revision History 3

II. Source Data 4

Source for Chemotaxis Pathway 4

SBML Data Format 4

SBML Data Manipulation Using LibSBML 5

SBML Software 5

III. Semantic Data Annotation 6

The Semantic Web 6

RDF in SBML 6

Annotating SBML Using RDF 7

IV. Connecting Pathway Data to Genomic Databases 8

Embedding Annotated Pathway Data in a Genome Browser 8

Integrating Annotated Pathway Data with a Genome Database 8

Formatting Pathway Data for Web Display 8

V. Areas for Further Investigation 9

VI. Appendices 10

Appendix A: *Dictyostelium discoideum* cAMP Chemotaxis Pathway 10

Appendix B: SBML Formatting & Example 11

# Introduction & Background

## About this Document

This document serves as a guide for initializing cell signaling pathway data for Dictybase annotations.

The document was initially created to serve as the knowledge container resulting from a proof-of-concept study during the course of a graduate student rotation for Dictybase. Subsequent versions may update or expand on the initial proof-of-concept findings.

## Initial Pathway for Proof-of-Concept

The pathway chosen for the proof-of-concept analysis was the cyclic adenosine monophosphate (cAMP) chemotaxis pathway for *Dictyostelium discoideum*. Chemotaxis refers to the directed movement of cells in response to external chemical signals, which for this case is cAMP. The specific pathway (see [*Appendix A: Dictyostelium discoideum cAMP Chemotaxis Pathway*](#_Dictyostelium_discoideum_cAMP)) was selected mostly because of its biological importance and the fact that *D*. *discoideum* has served as the primary model organism for chemotaxis study. When considering other factors around usability, however, the pathway also appeared to represent a “Goldilocks option”: the pathway is relatively well understood overall, but there are still sub-processes that aren’t yet clearly defined; and the size of the pathway is manageable, but large enough to warrant automated manipulation.

## Reference Documents

|  |  |
| --- | --- |
| **Document** | **Filename** |
| *Science Signaling* Data Terms of Usage | SciSignal\_Data\_Terms.pdf |
| *Science Signaling* XML Data Cleaning Script | SciSignal\_SBML\_Cleaning\_Script.py |
| Ontological Mapping for D. discoideum cAMP Chemotaxis Pathway | Chemotaxis\_Ontology\_Mapping.txt |
| Dictybase Pathway SBML RDF Tagging Script | SBML\_Dicty\_RDF\_Tagger.py |

## Revision History

|  |  |  |
| --- | --- | --- |
| **Version** | **Author** | **Summary of Changes** |
| *1.0* | Matthew Dapas | Initial Draft |

# Source Data

## Source for Chemotaxis Pathway

The source data for the chemotaxis pathway was obtained with permission from the *Science Signaling* Database of Cell Signaling (**©**2008 AAAS and Stanford University). One-year licenses on *Science Signaling* data are available for academic researchers (*refer to* [*Reference Documents: Science Signaling Data Terms of Usage*](#_Reference_Documents)).

Use of individual pathways from the *Science Signaling* Database of Cell Signaling should be cited in the following format:

Author of the pathway, Title of the pathway. Sci. Signal. (Connections Map in the Database of Cell Signaling, as seen DAY MONTH YEAR), URL of the pathway.

Any use of the chemotaxis pathway in Dictybase should be accompanied by the following citation:

Carole A. Parent, Alan R. Kimmel, Dictyostelium discoideum cAMP Chemotaxis Pathway. *Sci. Signal.* (Connections Map in the Database of Cell Signaling, as seen [DD Month YYYY]), http://stke.sciencemag.org/cgi/cm/stkecm;CMP\_7918.

*Science Signaling (*originally *Science’s STKE*) is a scientific research journal focused on cell signaling that is published weekly by the American Association for the Advancement of Science (AAAS). The Database of Cell Signaling organizes data about numerous cellular signaling pathways, both generalized and specific, which can then be visualized as “Connection Maps”. Once database access is granted, the data can be downloaded in XML format.

For more information on *Science Signaling* and the Database of Cell Signaling, refer to the *Science Signaling* website ([stke.sciencemag.org](http://stke.sciencemag.org/)).

## SBML Data Format

The de facto standard format for organizing biochemical and systems biology network data is a specific format of XML known as Systems Biology Markup Language, or [SBML](http://sbml.org/Main_Page). The SBML framework can be used for representing cell signaling pathway models, such as the cAMP chemotaxis pathway in *D. discoideum*.

SBML organizes network nodes and edges into “Species” and “Reactions”, respectively, with associated metadata and/or components for each. For an outline of SBML formatting, refer to [*Appendix B: SBML Formatting & Example*](#_Appendix_B:_SBML).

The advantage of SBML over other standard languages for biological data, such as [BioPAX](http://www.biopax.org/) or [CellML](http://www.cellml.org/), is that SBML enables quantitative analysis. SBML models can be used to compute information on sizes, amounts, and kinetics such that one could run simulations against experimentally derived expression datasets. Because of its potential for quantitative analysis, and because the source data from *Science Signaling* is stored in an early version of SBML, the SBML format was chosen as the master data file type for the proof-of-concept project involving the chemotaxis pathway.

Subsequent visualizations or database writing can stem from SBML data, but the corresponding SBML file(s) should be current, cleaned, and annotated to ensure usability. The source files from *Science Signaling* are outdated and do not contain metadata that can be integrated with genome annotations. The chemotaxis pathway file, therefore, required significant alterations before it could be employed for Dictybase.

An excellent source of model SBML usage is the [BioModels Database](http://www.ebi.ac.uk/biomodels-main/). When upgrading old SBML pathway files, the BioModels Database can serve as a guide. The BioModels Database is a repository of curated quantitative models that can be downloaded in SBML format. Models available on the BioModels Database are annotated with relevant external data, such as publications, ontologies, or other databases.

## SBML Data Manipulation Using LibSBML

LibSBML is an open-source programming library for SBML data manipulation. LibSBML supports a number of different programming languages through different application programming interfaces (APIs), including C++, C#, Java, Matlab, Octave, and Python. LibSBML is comprised of routines, data structures, object classes, and variables designed around SBML and the corresponding operations that are needed when working with SBML. Specific instructions and information regarding how to use LibSBML are contained within the SBML website ([sbml.org/Software/libSBML](http://sbml.org/Software/libSBML)).

For the chemotaxis pathway, the libSBML Python API was utilized for cleaning and updating the source file. The source file, initially created in 2003, employed an earlier version of SBML (Level 1) that is generally no longer accepted by SBML-parsing software. Using the libSBML Python API, a custom program (*see* [*Reference Documents: Science Signaling XML Data Cleaning Script*](#_Reference_Documents_1)) was written and executed to update the source file to the latest version of SBML (Level 3, Version 1), and re-label the network nodes to match either their protein abbreviations (e.g. “RasG”) or regulatory action (‘+’ or ‘–’).

When publishing about software or research in which libSBML is used, the following libSBML paper should be cited:

Bornstein BJ, Keating SM, Jouraku A, Hucka M. LibSBML: An API library for SBML. *Bioinformatics*. 2008;24(6):880-881. doi:10.1093/bioinformatics/btn051.

Once the SBML data is in a usable state, there are a number of software programs that can be run against the SBML data file.

## SBML Software

The [SBML website](http://sbml.org/Main_Page) maintains a [guide](http://sbml.org/SBML_Software_Guide) of known SBML-compatible software packages. As of November 16, 2013, there were 257 software packages listed in the SBML software guide. The online guide provides a matrix in which the software packages are listed against various capabilities, frameworks, platforms, etc.

The software platform chosen for viewing visual representations of the chemotaxis pathway data was [Cytoscape](http://cytoscape.org/). Cytoscape is a versatile, open source platform for visually rendering and analyzing network data. Cytoscape also supports data conversion between SBML and other formats, such as BioPAX or XGMML, which may be preferred depending on the research intention. The Cytoscape 3.x series represents the latest version of Cytoscape, but there may be useful applications available for Cytoscape 2.x that have not yet been ported to Cytoscape 3.x.

When publishing software or research in which Cytoscape was utilized, the following citation should be included:

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T. A travel guide to Cytoscape plugins. *Nature Methods.* 2012 Nov;9(11):1069-76. doi: 10.1038/nmeth.2212. Epub 2012 Nov 6.

Although not chosen for the Dictybase proof-of-concept analysis, another program that can be used for visualizing and editing SBML is [CellDesigner](http://celldesigner.org/), which is also designed as a modeling tool of biochemical networks. Another user-friendly tool is [semanticSBML](http://semanticsbml.org/semanticSBML/simple/index), which facilitates adding ontological tags to SBML data, as explained later in [*Annotating SBML Using RDF*](#_Annotating_SBML_Using).

# Semantic Data Annotation

## The Semantic Web

In an effort to improve the way information is stored and accessed online, a collaborative movement, led by the [Word Wide Web Consortium](http://www.w3.org/) (W3C) has slowly been nudging the World Wide Web towards the vision of a “Semantic Web”, in which computers can easily assemble decentralized and widely distributed information from across the Web using structured metadata components. The basic idea behind the Semantic Web is to store information in a way that makes information semantics computationally accessible, thereby improving the way knowledge is encoded on the Web. The [Resource Description Framework](http://www.w3.org/TR/rdf-mt/) (RDF) is the Semantic Web standard for encoding metadata and other knowledge. RDF essentially provides a triplet structure (subject, predicate, object) for defining statements about resources and relationships among them. By inserting RDF elements that reference corresponding ontological entities into an SBML pathway model, the door is opened for automated annotation of genomic databases with cellular pathway data.

## RDF in SBML

SBML defines a relatively simple format for annotating models with references to controlled vocabulary terms and database identifiers for describing biological/biochemical entities. The format uses RDF and conforms to the guidelines of [MIRIAM](http://www.ebi.ac.uk/miriam/main/mdb?section=intro)[[1]](#footnote-1) (Minimum Information Requested in the Annotation of biochemical Models). References to external resources are expressed using a Uniform Resource Identifier (URI) and a qualifier, e.g. “is\_a” or “is\_version\_of”.

RDF-based content in SBML is placed within the <annotation> element of an SBML component. The RDF tag structure is illustrated in the following template:

|  |
| --- |
| <SBML\_ELEMENT ... metaid="SBML\_META\_ID" ... >  ...  <annotation>  ...  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:bqbiol="http://biomodels.net/biology-qualifiers/" xmlns:bqmodel="http://biomodels.net/model-qualifiers/">  <rdf:Description rdf:about="#SBML\_META\_ID">  <QUALIFIER>  <rdf:Bag>  <rdf:li rdf:resource="URI" />  ...  </rdf:Bag>  </QUALIFIER>  ...  </rdf:Description>  ...  </rdf:RDF>  ...  </annotation>  ...  </SBML\_ELEMENT> |

SBML\_ELEMENT: Stands for the XML tag name of the SBML model component (e.g. model, species, reaction, etc.).

SBML\_META\_ID: The SBML element’s meta-identifier, unique within the file. Optional, but required for RDF annotation.

URI: Identifies associated resource data (e.g. “http://identifiers.org/dictybase.gene/DDB\_G0276269”)

QUALIFIER: Describes the relationship between the SBML model component and the associated resource data. In this example the qualifier name must refer to an element in either the XML namespace “http://biomodels.net/biology-qualifiers/" or in “http://biomodels.net/model-qualifiers/".

As shown in the above example, the most relevant set of qualifiers to use for annotating biological pathways are defined in the [BioModels.net Qualifiers](http://co.mbine.org/standards/qualifiers) namespaces. The RDF qualifiers describe the relationship between the SBML component and the associated resource data. The current set of BioModels qualifiers are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model qualifiers:** |  | **Biology qualifiers:** | | |
| * is: * isDerivedFrom * isDescribedBy: |  | * encodes: * hasPart: * hasProperty: * hasVersion: * is: * isDescribedBy: * isEncodedBy: | * isHomologTo: * isPartof: * isPropertyOf: * isVersionOf: * occursIn: * hasTaxon: |

## Annotating SBML Using RDF

Because RDF and SBML both employ standard syntaxes, adding RDF annotations to SBML is relatively easy, given the relationships between the SBML components and corresponding resource data are known. In the case of the chemotaxis pathway for *D. discoideum*, the source data did not contain computer-readable metadata that could be used to connect with genomic database annotations. Therefore, RDF tags that assigned relationships between pathway elements and *D. discoideum* genes were added to the source data file.

A number of software programs exist for adding RDF tags to SBML. One easy-to-use program for annotating SBML is [SemanticSBML](http://semanticsbml.org/semanticSBML/simple/index). With SemanticSBML, one can view, annotate, merge, and compare SBML files using a simple interface without needing to interact directly with XML or a programming language. This would make it convenient for an annotator to add semantic annotations to a SBML file given the network has already been created.

For the chemotaxis pathway, the LibSBML Python API was again used to automatically insert RDF tags for the pathway species (*see* [*Reference Documents: Dictybase Pathway SBML RDF Tagging Script*](#_Reference_Documents_2)). The resultant custom script uses a tab-delimited mapping file (*see* [*Reference Documents: Ontological Mapping for D. discoideum cAMP Chemotaxis Pathway*](#_Reference_Documents_3)) to establish connections to [Uniprot](http://www.uniprot.org/) and Dictybase Gene identifiers (DDB\_G#######).

It’s important to note that while the “metaid” attribute is optional for SBML components, it’s required for RDF annotation for a given SBML component.

# Connecting Pathway Data to Genomic Databases

**NOTE:** This section cover approaches for connecting annotated pathway data to a model organism database, such as Dictybase, thereby enabling targeted queries to return pathway data associated with certain genes. A number of approaches can be taken to integrate RDF data with a more conventional genomic data store, but as of Version 1.0 of this document, the information and proposed methods presented herein remain untested.

## Embedding Annotated Pathway Data in a Genome Browser

One method of integrating pathway metadata into a genome browser is to embed RDF metadata into an image or other compatible HTML element. Pathway model RDF metadata could be inserted into a raster image, or pathway node RDF metadata could be embedded within a vector image. Various, interactive functions can then be performed based on the data. Two JavaScript-based tools that can be used for creating such RDF driven applications are [RForms](https://code.google.com/p/rforms/) and [Callimachus](http://callimachusproject.org/). The enabling technology for embedding semantic data directly in HTML or XHTML is known as [RDFa](http://www.w3.org/TR/xhtml-rdfa-primer/).

## Integrating Annotated Pathway Data with a Genome Database

Another way to include semantic metadata in a genome browser would be to integrate RDF data directly into the Generic Model Organism Database (GMOD). To achieve this, one may first transform the GMOD data into RDF. This can be achieved using the [SADI](http://sadiframework.org/content/about-sadi/) (Semantic Automated Discovery and Integration) framework. The technology [SADI for GMOD](https://code.google.com/p/sadi/wiki/SADIforGMOD), for example, was designed as an add-on for the GMOD project to enable access to sequence feature data in RDF form. Once the genomic data is available in RDF, it can be merged with RDF pathway annotations or returned with pathway data using [SPARQLE](http://www.w3.org/TR/rdf-sparql-query/) queries. A similar approach was taken by the OpenFlyData[[2]](#footnote-2) project on the FlyBase Chado database schema, in which they created relational-to-RDF [mappings](http://www.w3.org/TR/2011/WD-rdb-direct-mapping-20110324/) between the relational database (RDB) and an RDF triple store.

Conversely, a more traditional technique would involve reading RDF data from the SBML file and writing it to an RDB model. The simplest way to do this would be to reproduce the RDF triple-store relation with three columns (Node, PropertyType, Value) in the RDB, but other methods exist for storing RDF data in a relational model[[3]](#footnote-3).

## Formatting Pathway Data for Web Display

If the goal is only to display a visualization of the desired pathway, then this can be accomplished by either embedding a dynamic plug-in or by uploading a static image. For a dynamic visualization, one could export the pathway from Cytoscape or an equivalent application as a graph file (e.g. XGMML), and then use a program like [Cytoscape Web](http://cytoscapeweb.cytoscape.org/) to embed the image. A raster image can be output from any software that’s used locally to edit the pathway.

# Areas for Further Investigation

* Updating *D. discoideum* cAMP Chemotaxis pathway data based on recent publications
* Actualizing connection of cell signaling pathway data to gene annotations
* Integrating pathway data in Dictybase, either as RDF or in relational database
* Using SBML models to quantitatively analyze cell signaling pathways against expression data

# Appendices

## Appendix A: *Dictyostelium discoideum* cAMP Chemotaxis Pathway[[4]](#footnote-4)

## 

## Appendix B: SBML Formatting & Example[[5]](#footnote-5)

The complete SBML XML Schema can be viewed on [sbml.org](http://sbml.org/Special/xml-schemas/sbml-l2v4-schema/sbml.xsd).

Below is a scrubbed snippet of the cAMP chemotaxis SBML source data:

No metaid in source data components. Must be created to incorporate RDF metadata.

|  |
| --- |
| <listOfSpecies>  <species id="species\_ID\_12345" compartment="compartment0">  <annotation>  <node pathID="CMP\_1234" componentID="CMC\_12345" xPos="732.0" yPos="292.0" type="CV\_1234" localization="\_CV\_12701" abbreviation="ACA">  <auditing authorityID="curator1" createdBy="curator1" creationDate="2002-01-01 12:34:56.789" lastModifiedBy="2003-01-01 09:08:07.56" lastModifiedDate="curator1" />  <listOfCitations>  <citation type="journal" pmid="1234567" authors="Doe, J., Smith, J." docType="Journal Article" title="Article about ACA in Dictyostelium" journal="Biochemistry" volume="123" pages="12345-6" issue="12" date="1999 June 12" />  </listOfCitations>  </node>  </annotation>  Name is hidden in Annotation. Must be moved to Species tag to be read by visualization software.  </species>  ... |

1. Le Novère N, et al. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology*. 2005;23:1509-1515. [↑](#footnote-ref-1)
2. Miles A, Zhao J, Klyne G, et al. OpenFlyData: an exemplar data web integrating gene expression data on the fruit fly Drosophila melanogaster. *J Biomed Inform*. 2010;43(5):752-61. [↑](#footnote-ref-2)
3. Bornea MA, Dolby J, Kementsietsidis A, et al. Building an efficient RDF store over a relational database. SIGMOD’13. June 22-27, 2013. New York, NY, USA. [↑](#footnote-ref-3)
4. Carole A. Parent, Alan R. Kimmel, Dictyostelium discoideum cAMP Chemotaxis Pathway. *Sci. Signal.* (Connections Map in the Database of Cell Signaling, as seen [11 December 2013]), http://stke.sciencemag.org/cgi/cm/stkecm;CMP\_7918. [↑](#footnote-ref-4)
5. <http://sbml.org/More_Detailed_Summary_of_SBML> [↑](#footnote-ref-5)