

Developing SBML Beyond Level 2: Proposals for Development

Andrew Finney

University of Hertfordshire, Hatfield, AL10 9AB, UK
a.finney@herts.ac.uk

Abstract. The Systems Biology Markup Language (SBML) is an XML-based exchange format for computational models of biochemical networks. SBML Level 2, whose definition was established in June 2003, includes several enhancements to the original Level 1. This paper includes a brief overview of Level 2. Several proposals are under development to extend SBML to create Level 3. These include diagrams, 2-D and 3-D spatial characteristics, arrays, model composition and multi-component chemical species. This paper describes the current proposals for the last two features.

1 Introduction

Researchers in systems biology have a wide variety of software tools available to them. The Systems Biology Markup Language (SBML) [1] was initially developed to enable the exchange of models between these tools, as part of the ERATO Kitano project in collaboration with the community of Systems Biology modelers and software developers. A multi-institutional team based at the California Institute of Technology (USA), University of Hertfordshire (UK) and Systems Biology Institute (Japan) provides editorial, organizational and technical support to the SBML development process. Today, over 40 software tools support SBML [2], and in addition, it is the standard model definition language for the DARPA BioSPICE project and the International E. coli Alliance (IECA).

SBML encodes models consisting of biochemical entities (species) linked by reactions to form biochemical networks. SBML is being developed by the community in levels, where each level extends the set of features of the language. The structures that comprise an model in SBML Levels 1 and 2 are described elsewhere [1, 2, 3]. By freezing SBML development at incremental levels, software authors can work with stable standards and gain experience with the standard before further development. The separate levels of SBML are intended to coexist.

2 Proposals for Level 3

The Systems Biology community is now developing SBML Level 3. Although SBML has been successfully adopted by many groups developing systems biology

software, there exist packages that support classes of models which presently cannot be encoded in Level 2. Level 3 is intended to provide support for these tools. The community plans to introduce features in Level 3 that will add support for: (a) composing models from component submodels; (b) describing states and interaction of components of species in terms of rules rather than explicit enumerations of all possible combinations; (c) describing 2-D and 3-D spatial geometries; (d) diagrammatic representations of models; (e) enabling parameter and initial condition values to be defined separately from models; (f) allowing for alternative mathematical representations of reactions; and (g) enabling the association of terms from controlled vocabularies to be associated with SBML elements. Proposals for these features have been described at various levels of detail [2]. The following sections describe proposals for Model Composition and Multi-Component species. These are not yet part of any SBML standard.

2.1 Model Composition

The aim of the Model Composition proposal is to enable a model to be composed hierarchically from a number of submodels. Given the existence of more than one software package supporting hierarchical model composition, (for example Promot/DIVA [4] and VLX Biological Modeler from Teranode Corporation [5]) SBML ought to support the exchange of such models. One of our key aims is enable the multiple instantiation of the same submodel within an enclosing model thus enabling, for example, a model of a cell to be composed from a number of instances of the same mitochondria submodel. This specific capability is not present in CellML despite its support for hierarchical composition. The specific proposal described here [6] builds on previous proposals [7, 8].

The model composition proposal allows for a model to contain a list of submodels and a list of instances of those submodels. Each submodel is a complete SBML model thus allowing for the hierarchical assembly of models. Each instance structure instantiates a submodel (i.e. implies the existence of a complete copy of the submodel) within the immediate containing model. An instance structure refers to a submodel using `XLink` attributes enabling the structure to refer to a submodel that is either internal or external to the containing XML document. (`XLink` [9] is a standard for referring to components of XML documents.)

A system that enables just the assembly of models from submodels as described has little utility. A method for linking model components at various levels of the model instance hierarchy is required. The model composition proposal introduces `ObjectRef` structures which can refer to objects within the instance hierarchy. The proposal has defined `ObjectRef` structures to be recursive thus allowing objects at any level of the instance hierarchy to be referenced.

The `ObjectRef` structure does not use `XLink` attributes because the instance hierarchy does not exist in XML form but is implied by the instance structures in models. For example consider two models *A* and *B*. *A* contains a species *S* and model *B* contains two instances of *A*, *a*₁ and *a*₂. Suppose we wished to refer to the specific species *S* within instance *a*₁. We cannot use an `XLink` attribute because that specific species does not exist in any XML document, it

only exists in the a_1 copy of A . The single XML structure for S in model A does not represent any specific species within instances of A since there can be any number of instances of A .

SBML documents have many attributes which link the various structures forming the reaction network and placing species into compartments. In the model composition proposal these links can optionally be replaced by `ObjectRef` structures thus enabling links to form between models thus enabling, for example, a reaction to be created between species in two submodel instances or for a species to be placed inside a compartment within a submodel instance. An `ObjectRef` can only refer to an object within the instance hierarchy created by the immediate containing model: since an `ObjectRef` does not contain `XLink` attributes it cannot create arbitrary linkages. The `ObjectRef` structures and the Level 2 attributes they replace must refer to the same object types. These types are restricted as described in the SBML Level 2 specification. For example the `speciesLink` element, an instance of the `ObjectLink` type, which replaces the `species` attribute on `SpeciesReference` structures, can only refer to `Species` objects.

Potentially when submodels are combined to form a model one biochemical entity will be represented separately in more than one of the submodels. In many bioinformatics systems this is resolved using a synonym dictionary. To enable the construction and reuse of abstract or generic models a structure is required that enable the parametrization of models. In the general case the parameter and synonym requirements can be viewed as the requirement to support one object overloading another object. Under the proposal a model contains a list of `Link` structures which could be viewed either as a directed synonym dictionary or a set of parameter assignments. Each `Link` structure consists of two `ObjectRef` structures which refer to the overloading and overloaded objects. The overloading object replaces the attribute values of the overloaded object. If the objects are species then they form a single node in the reaction network. If the objects are compartments then they form a single compartment. The types of the two objects referred to by a `Link` structure must have the same type. For example a `Species` object cannot overload a `Compartment`. There is no consensus in the community of modelers concerning the ideal form of composed models and thus the proposal does not describe any further restrictions on the linkages that can be formed between objects in the instance hierarchy. However SBML has features, such as a units system, which are appropriate for further validation of model composition.

2.2 Multi-component Species

Although SBML Level 2 can encode biochemical reaction networks the following concepts are not easily represented: (a) the hierarchical description of biochemical entities through the composition of other biochemical entities or (b) the description of generalized biochemical reactions that avoid the enumeration of many species states and reactions. Biochemical entities, depending on the description, can be composed either as simple aggregation or through graphs of

other biochemical entities where arcs represent kinds of bonding. SBML requires new features to enable the representation of, for example, proteins which can contain many phosphorylation states, complexes of these proteins and models of signalling pathways which contain these proteins. Several laboratories have data sets and/or software which make explicit some or all of these features of biochemical networks often in generalized form. For example the Genome KnowledgeBase [10] captures the hierarchical assembly of complexes in the context of pathways; the LANL T10 group have developed models of generalized reaction systems from which software automatically generates an ODE based representation containing very large sets of species and reactions [11]; and the StochSim environment [12] stochastically simulates the interaction of individual chemical entities each of which has a state vector where reactions form a set of complex state transitions.

This section contains proposals for a set of new structures for addition to SBML to implement the above requirements. This proposal builds on previous proposals for representing multi-component species [13, 14]. These structures, whilst primarily aimed at supporting the representation of phosphorylated protein complexes, will be capable of capturing both the detailed structure of, and the processes acting on, all types of macromolecules.

Under this proposal a model would optionally contain a set of species type structures. A given species type simply represents all biochemical entities with the same biochemical structure (that is having identical structure for the purposes of the model). In SBML Level 2 the species structure represents a pool of entities of the same type located in a specific compartment. In this proposal the type of a species structure is made explicit via a new `speciesType` attribute which refers to a species type structure. A reaction can be generalized to occur in any compartment by referring to reactants, products and modifiers by species type rather than by compartment specific species.

Under this proposal a species type can optionally contain a set of instances of other species type structures which define the composition of the containing species type. A model can be described using such a system of hierarchically contained components however, under this proposal, the species type instances can be explicitly connected i.e. a species type can describe a graph where the nodes are species type instances and the arcs are bonds. In this scheme a species type has a set of binding site structures each of which is a potential end point for a bond. A bond is simply a pair of references to binding sites on species type instances. Just as the bond structures are optional in species type structures it is not proposed here that SBML specify the level of decomposition at which a given model will operate, for example, a protein could be described as a single indivisible object or as a sequence of amino acids.

Whilst the structures described above capture a significant amount of information that cannot be made explicit in SBML Level 2 they do not provide any facilities for representing reactions generalized to apply to classes of species types. With just these structures an accurate model would still have to contain an enumeration of all the species type structures that could occur in the mod-

elled system. In many modelled systems the number of structures required is so large that both this representation scheme and SBML Level 2 become impractical. To solve this problem, under this proposal, reactions can be generalized to apply to classes of species types. The complete set of species and species type structures are then implied from the reactions rather than fully enumerated. In this context the species structures contained in an SBML document define the initial state and boundary conditions of the system. The species type structures define a set of types that enable the definition of reactions and species. The reactions are applied to the biochemical entities in the modelled system that match the reactions' reactants and construct new entities as defined by the reactions' products. Thus a generalized reaction is a template for manipulating graphs of biochemical entities and contains structures which enable a reactant to match with species from a range of species types.

In a generalized reaction the reactant, product and modifier structures are *generic graphs* of a form similar to those graphs contained by species type structures. Whereas a species type can contain concrete bonds, which refer to pairs of binding sites, a generic graph can additionally contain generic bond structures, which refer to only one binding site. A generic bond simply represents a portion of a matching species graph which the reaction does not directly transform. A generic bond is identified by an associated symbol. The same symbol typically will occur in both the reactants and products indicating that a matching component is transferred from reactant to product.

3 Conclusions

SBML is a defacto standard for the exchange of biochemical reaction networks. The development of SBML is ongoing to ensure that it meets the full requirements of the Systems Biology community. The proposals described for model composition and multi-component species will ensure that SBML can represent highly complex biological systems such as signal transduction systems.

Acknowledgements

The development of SBML was originally funded by the JST (Japan). Support for the continued development of SBML today comes from the following sources: NHGRI/NIH (USA), NIGMS/NIH (USA), NEDO (Japan), BBSRC (UK), DARPA (USA), and the Air Force OSR (USA). The SBML development community includes members of the BioSPICE Model Definition Language, sys-bio and sbml-discuss mailing lists. I thank Michael Blinov, Roger Brent, Fabian Campagne, Michael Hucka, Sarah Keating, Larry Lok, Nicolas Le Novère, Robert Phair and Jeremy Zucker for helpful comments.

References

1. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novère, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J.: The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* **19** (2003) 524–531
2. Hucka, M., Kovitz, B., Matthews, J., Schilstra, M., Finney, A., Bornstein, B., Shapiro, B., Keating, S., Funahashi, A.: The SBML website. Available via the World Wide Web at <http://www.sbml.org/>. (2003)
3. Finney, A., Hucka, M.: Systems Biology Markup Language: Level 2 and Beyond. *Biochemical Society Transactions* **31** (2003) 1472–1473
4. Ginkel, M., Kremling, A., Tränkle, F., Gilles, E.D., Zeitz, M.: Application of the process modeling tool ProMot to the modeling of metabolic networks. In Troch, I., Breitenecker, F., eds.: *Proceedings of the 3rd MATHMOD*. (2000) 525–528
5. Duncan, J., Arnstein, L., Li, Z.: Teranode corporation launches first industrial-strength research design tools for the life sciences at demo 2004. Available via the World Wide Web at http://www.teranode.com/about/pr_2004021601.php (2004)
6. Finney, A.: Systems Biology Markup Language (SBML) Level 3 Proposal: Model composition features. Available via the World Wide Web at <http://www.cds.caltech.edu/~afinney/model-composition.pdf> (2003)
7. Ginkel, M.: Modular SBML Proposal for an Extension of SBML towards level 2. In: *Proceedings of 5th Forum on Software Platforms for Systems Biology (SBML Forum)*. (2002) Available via the World Wide Web at <http://www.sbml.org/workshops/fifth/sbml-modular.pdf>.
8. Webb, J.: BioSpice MDL Model Composition and Libraries. Available via the World Wide Web at <http://bio.bbn.com/biospice/mdl/design/compose.html> (2003)
9. DeRose, S., Maler, E., Orchard, D.: XML Linking Language (XLink) Version 1.0 W3C Recommendation. Available via the World Wide Web at <http://www.w3.org/TR/2000/REC-xlink-20010627/> (2001)
10. Hodge, R.: Linking the levels of life from genes to cellular processes with the Genome Knowledge Base. Available via the World Wide Web at http://www.ebi.ac.uk/Information/News/ensembl_040203.pdf. (2003)
11. Goldstein, B., Faeder, J.R., Hlavacek, W.S., Blinov, M.L., Redondoc, A., Wofsy, C.: Modeling the early signaling events mediated by Fc ϵ RI. *Molecular Immunology* **38** (2001) 1213–1219
12. Le Novère, N., Shimizu, T.S.: StochSim: Modelling of stochastic biomolecular processes. *Bioinformatics* **17** (2001) 575–576
13. Le Novère, N., Shimizu, T.S., Finney, A.: Systems Biology Markup Language (SBML) Level 3 Proposal: Multistate Features. Available via the World Wide Web at <http://sbml.org/multistates.pdf> (2003)
14. Finney, A.: Internal discussion document possible extension to the Systems Biology Markup Language Complex Species and Species Graphs. Available via the World Wide Web at <http://www.cds.caltech.edu/~afinney/CplxSpecies.pdf> (2001)