

SBML Level 3: Proposals for Advanced Model Representations

Andrew Finney

University of Hertfordshire, UK
Molecular Sciences Institute, USA

Michael Hucka

California Institute of Technology, USA

<http://www.sbml.org>

What is SBML?

The Systems Biology Markup Language (SBML) [1-3] is a computer-readable format for representing **models of biochemical reaction networks**. It is applicable to many subject areas:

- metabolic networks,
- cell-signaling pathways,
- genomic regulatory networks, and
- many other modeling problems in systems biology.

SBML is based on XML, a standard medium for representing and transporting data that is widely supported on the Internet as well as in computational biology and bioinformatics.

Because SBML is **completely tool-independent**, it enables

- use of multiple simulation and analysis tools in a single research project without rewriting models for each tool
- publication of models in peer-reviewed journals: other researchers can download and use your model even if they use a different modeling environment
- survival of models: they can outlive the software used to create them, making your work still useful even if a particular simulation package is no longer supported

SBML has been evolving since mid-2000 through the efforts of many collaborators who make up the *SBML Forum*. Today, SBML is supported by over 60 software applications

A Format That Mirrors the Problem Domain

The representation of a system of biochemical reactions expressed in SBML simply consists of formalized statements about the various components of the model:

- **Species** which are pools of chemical entities, each species represents a single pool of a specific type of chemical entity located in a specific compartment
- **Compartments** in which the species are located, each compartment is a well stirred volume
- Chemical **reactions** affecting the species, reactions include transport processes that move chemical entities from one compartment to another
- **Parameters** representing numerical variables
- Definitions of **units** on numerical quantities
- Definitions of mathematical **functions** used in formulas
- Discrete **events** representing changes in the system's state
- Additional mathematical **constraints** on the system

SBML Evolution Through Levels

As SBML evolves the community creates SBML Levels. Each new level adds new features to the language. SBML Level 2 was standardized in 2003. Simple software tools can use SBML Level 1, the first and most basic version of SBML. More sophisticated systems can use SBML Level 2, with its enhanced capabilities. SBML Level 3 is actively being developed through the SBML Forum

The Evolution of SBML Continues

SBML Level 3 is being designed collaboratively by today's **leading developers** of open-source software for computational biology. SBML Level 3 development has been divided into several modules including:

- **Diagrams:** SBML extensions to store the graphical diagrams of models that can be created in many of today's graphical pathway editors.
- **Model Composition:** SBML extensions to support the representation of models that are composed from submodels (see next section below).
- **Multicomponent Species:** SBML extensions to enable the compact representation of species having multiple possible states (e.g., due to phosphorylation) and/or configurations with other species (e.g., protein complexes). (See section 'Requirements of a Multicomponent Species Proposal' and following sections.)
- **Arrays:** SBML data structures to permit arrays of items (such as species, compartments, and others) to be grouped and manipulated en masse. Sparse arrays will be supported and could be used as a way to describe network connection schemes.
- **Spatial Features:** SBML extensions to describe the 2-D and 3-D spatial characteristics of models, including the geometry of compartments, the diffusion properties of species, and the specification of different species concentrations across different regions of a cell.

- **Controlled Vocabularies:** extension of SBML to enable components of a model to be labeled with terms taken from biologically and computationally meaningful controlled vocabularies.

Proposals for Model Composition

To date, there have been several proposals for SBML extensions to support model composition from Martin Ginkel (MPI Magdeburg) [4], Jonathan Webb (BBN) [5] and Andrew Finney [6]. The common idea is support the composition of larger models from smaller ones (submodels). Under these proposals, a model could contain:

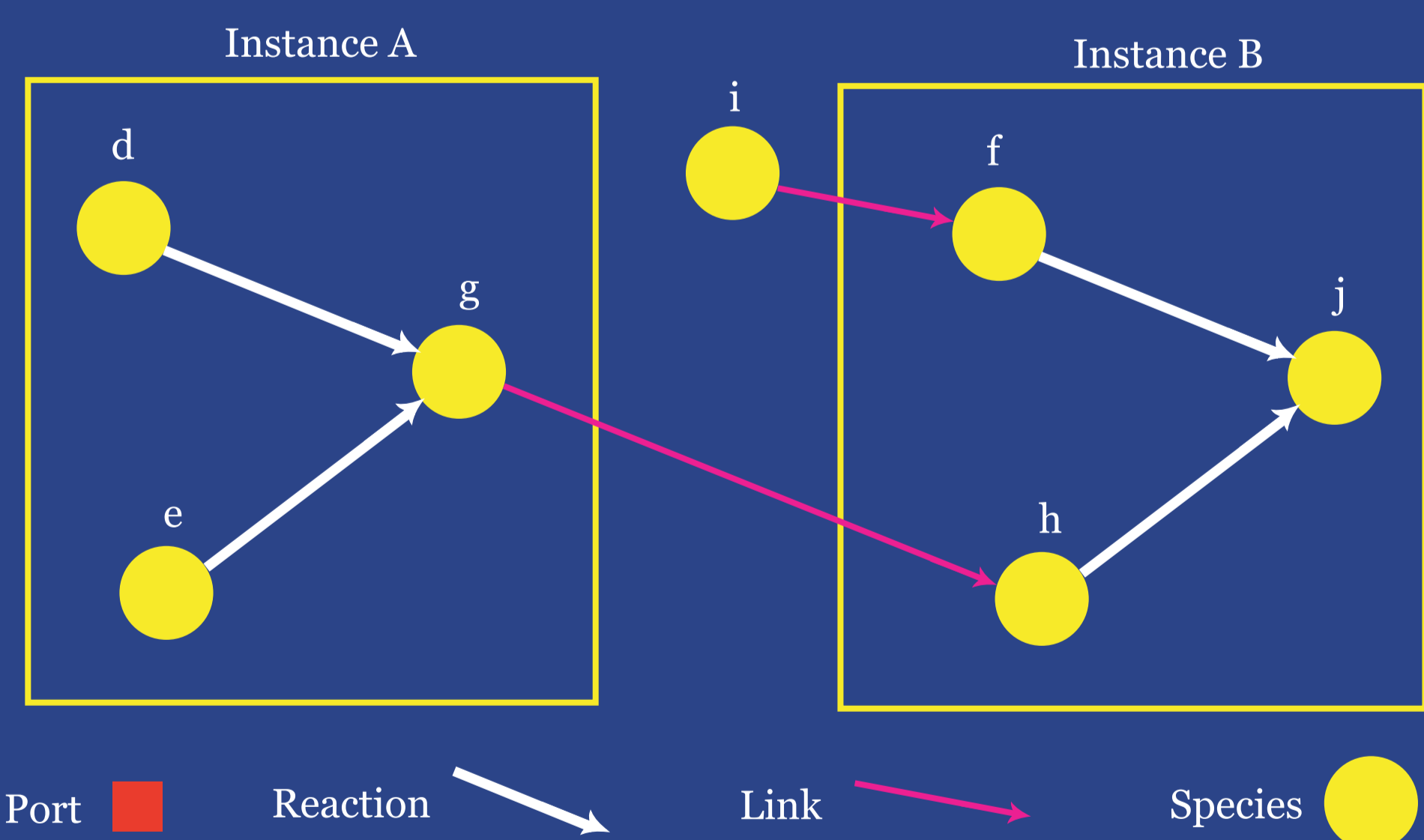
- *Submodel definitions:* Models may be contained within an SBML document or an SBML document can reference external models.
- *Instances of submodels:* Models may contain instances of submodels that are complete copies of the submodels. A model can contain more than one instance of a submodel. A model consists of a heirarchy of instances of submodels.
- *Links between objects:* Models may contain links between objects at arbitrary positions in the instance hierarchy. Such a link indicates that the linked objects are replaced by a single object. The links are directional; the direction indicates which object overloads its attribute values to create the final object.
- *Direct Reference links:* SBML attributes that reference other objects, for example `species` on `speciesReference` can be replaced by elements which enable objects in arbitrary positions in the instance hierarchy to be referenced.

A number of issues remain to be resolved in these proposals for model composition:

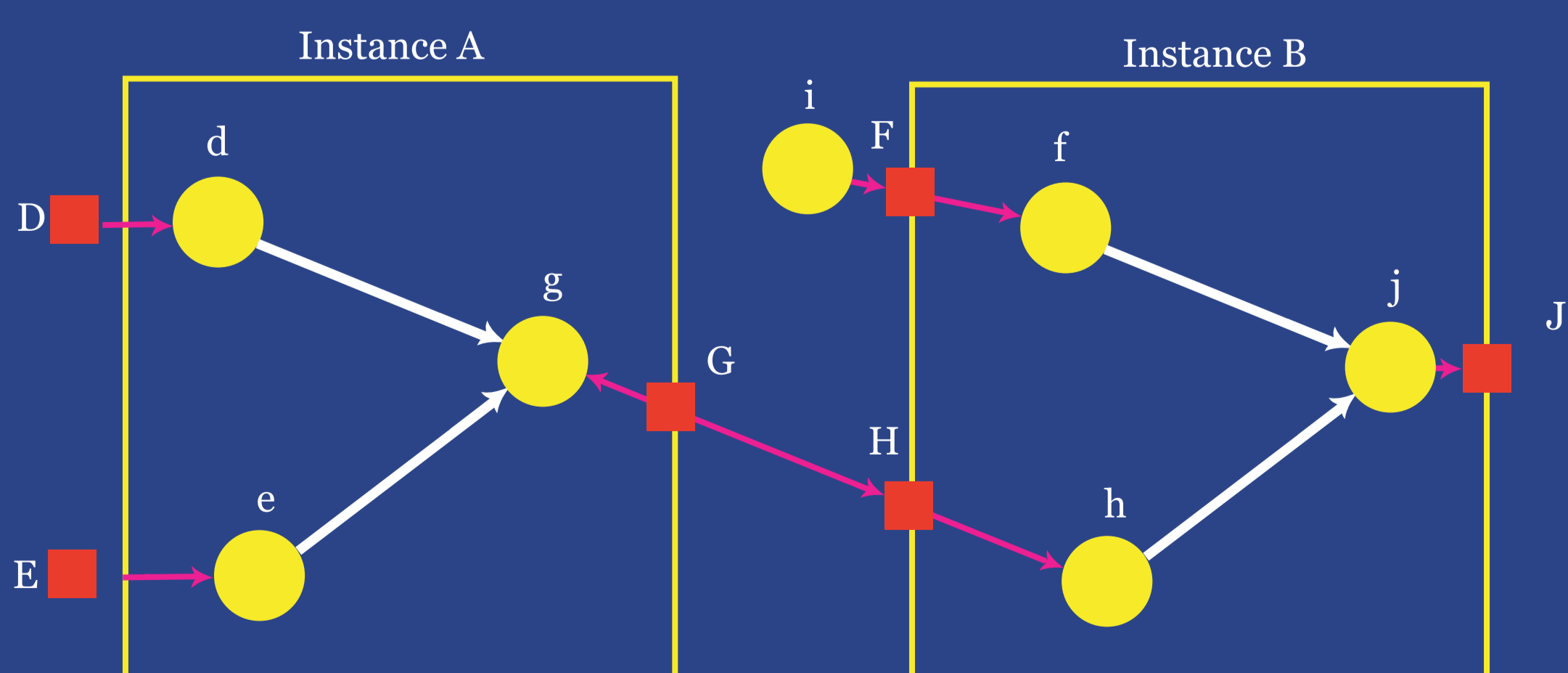
- Should submodels have well-defined and enforced *interfaces* that constrain the links that can be formed between objects?
- To what extent should the standard define the legality of links? Obviously, objects of different types cannot be linked together; however, should links, e.g. between parameters with incompatible unit definitions, be legal?
- Are XML standards, such as XLink, appropriate for the various reference encodings in these proposals?
- Do we need to support linking to arbitrary depths?

Model Composition Example

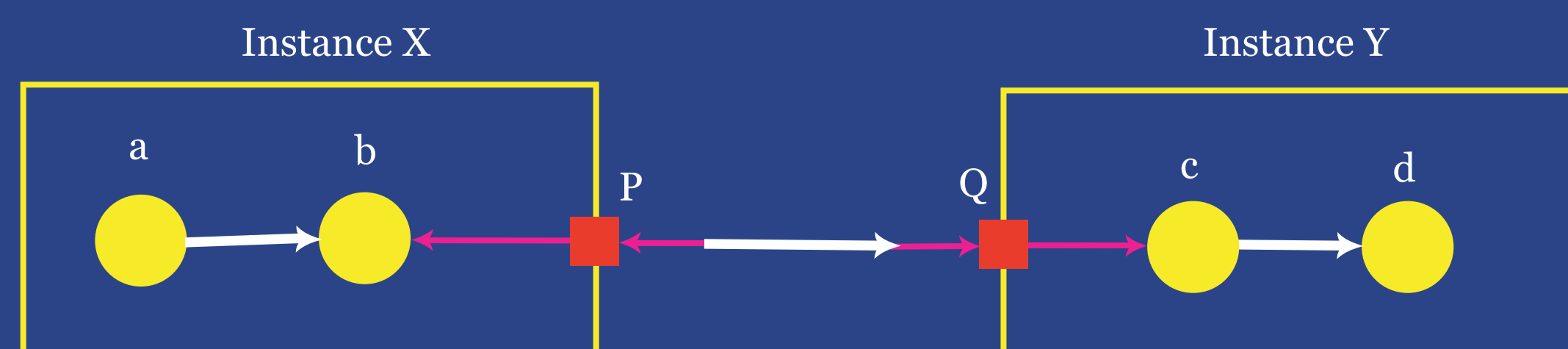
When composing a model, it is often necessary to merge objects from different submodels. The model composition proposals provide mechanism for doing this. Consider the following model, without interfaces, containing two instances each of a different submodel. In this example, we merge species *g* with *h* and *i* with *f*:



The following model is equivalent but has defined interfaces:



Along with merging equivalent entities form a single object, when combining models it is useful to be able to create reactions that link models. The model composition proposals allow reactions to connect species in different instances of submodels. For example, consider the following model containing a reaction between two ports:



Requirements of a Multicomponent Species Proposal

In SBML Level 2 **species** represents a pool of chemical entities all of the same single state in a specific compartment. **species** cannot be composed from components. These limitations have many consequences:

- the different states of a molecule have to be enumerated with no formal relation of those states together;
- species representing the same molecule in different compartments are not formally related; and
- there is no representation of the internal structure of a molecule.

Given that several groups find the SBML Level 2 representation of species and reactions limited a proposal for a multicomponent species extension to SBML has been written [7]. This proposal aims to satisfy the following requirements

- Relate species of the same type that are located in different compartments
- Enable reactions to defined that are generalized across compartments
- Enable species to be defined as composed of components
- Enable reactions to be generalized to apply to sets of species states

These requirements address the near-term needs of modelers of metabolic networks and the longer-term requirements of modelers of signal transduction networks.

Modeling Software Having Multicomponent Species Representations

Several research groups have developed formal computational models of signaling pathways which use more sophisticated representations of species without the limitations of SBML Level 2. The proposal described in this poster is designed to be, among other things, an exchange format for these models. In these models species have a number "bonds" which can be in different states. A Reaction can be generalized to apply to a set of states (instead a single state as in SBML Level 2). The mechanism for translating such a representation into a computational model that might be practically simulated varies between the research groups.

– Alpha Project

- Studying: Yeast Pheromone Pathway
- Software: Molecuizer
- Website: <http://www.molsci.org/>

– LANL T10 Group [8]

- Studying: Signal transduction systems of the immune system
- Software: BioNetGen
- Website: <http://cellsignaling.lanl.gov/>

– Cell Systems Initiative (CSI) [9]

- Studying: T cell activation
- Software: Balsa and SigTran
- Website: <http://www.csi.washington.edu/>

– Carl Firth, Tom Shimizu, Nicolas Le Novère [10]

- Studying: *E. Coli* Chemotaxis Signaling Pathway
- Software: StochSim
- Website: <http://info.anat.cam.ac.uk/groups/comp-cell/StochSim.html>

Basic Facilities Available in Multicomponent Species Proposal

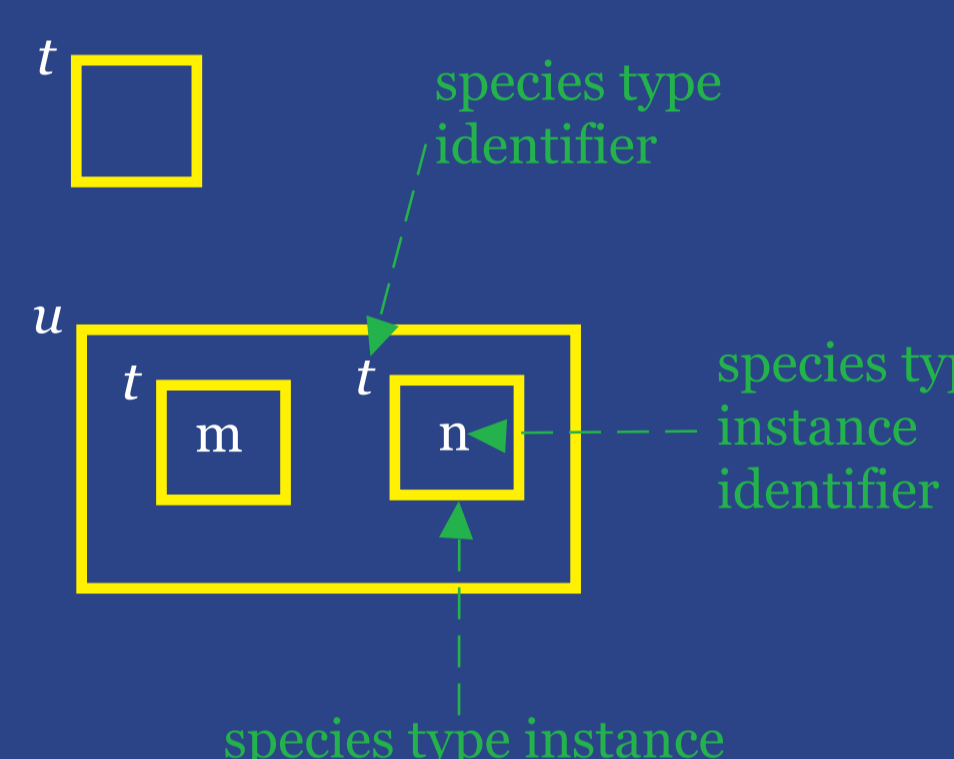
The proposal described here [4] introduces a number of basic facilities that overcome some of the limitations of SBML Level 2. These facilities provide a foundation for a representation scheme that address all the requirements for a multicomponent species proposal.

The proposal introduces a new structure **SpeciesType** which represents the set of all biochemical entities of a given type irrespective of the location of those entities. Species structures can refer to species types which enables species of the same type to be related together when the given species are located in different compartments. Similarly reactions can be generalized to apply to species types instead of species. Such a reaction applies to all compartments in a model.

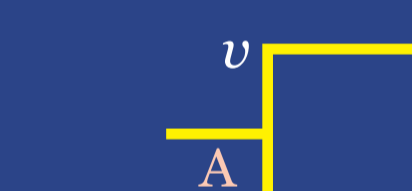
Species Types with Defined Structure

The following diagrams show various cases of how a species type may be defined. Some of these species type structures refer to each other.

A simple species type is indivisible

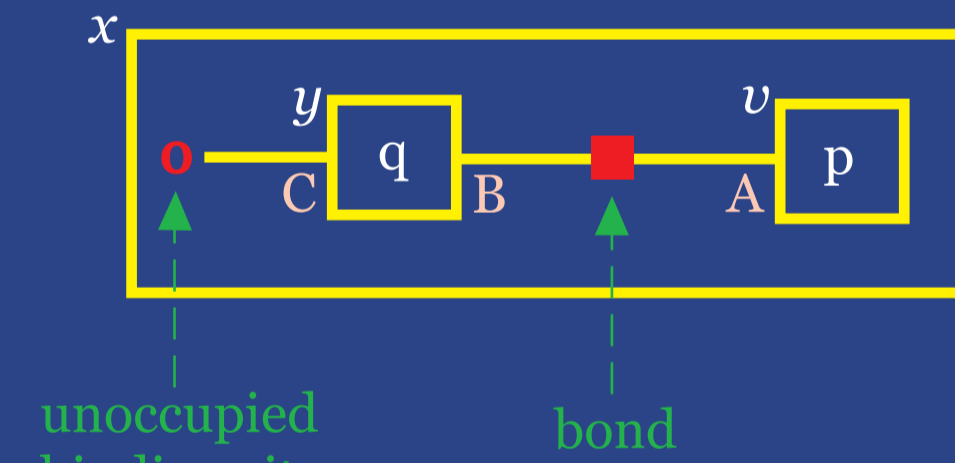


A species type can contain unconnected instances of other species types



A species type can define a number of external labeled binding sites

A species type is a graph of species type instances connected by bonds

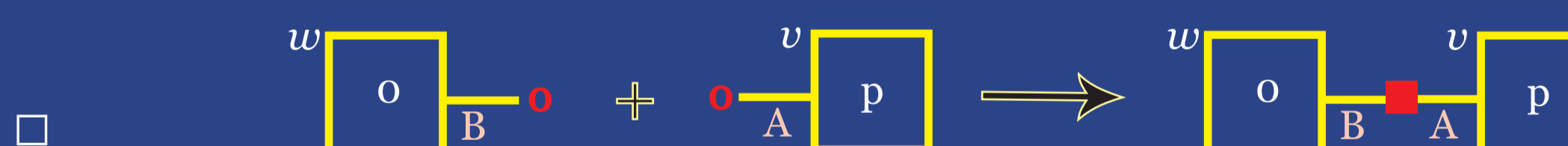


Reactions Operating at Different Levels of Detail

In this section we show examples of two ways in which a reaction can be defined under this proposal. The following diagram shows an example of the first approach. The diagram shows a simple reaction in which two entities of types *t* and *z* are consumed to create an entity of type *s*. The internal structure of *t*, *z* and *s* are not relevant to the reaction.

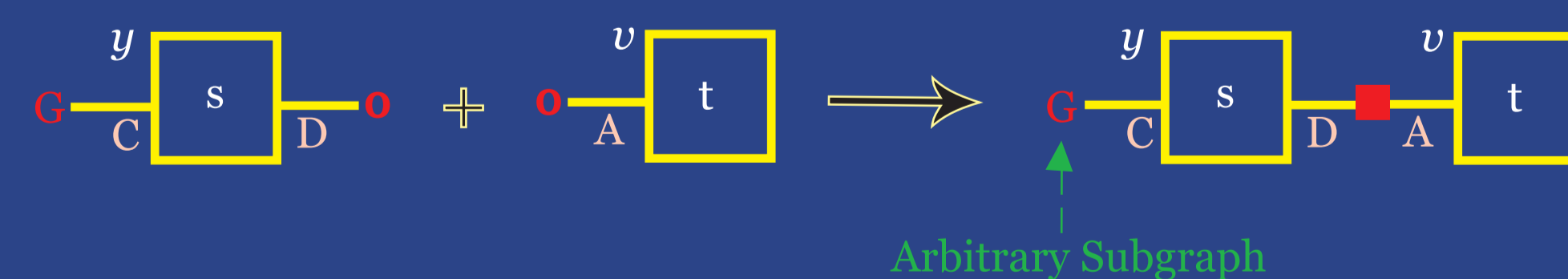


The following diagram shows the second more complex approach in which the reactants and products of a reaction are defined as graphs of species instances. The diagram shows a reaction in which two entities come together to form a larger molecule. The instances of species types are identified so that the transformational details of the reaction are captured.



Reactions Generalized to Apply to Classes of Species Types

The complex reaction scheme described above is extended so that reactions can be applied to a class of species states rather than individual species states. Without this extension, all species states and the reactions that apply to them would have to be enumerated. A reaction can be generalized to cover all states of one or more binding sites. In the following example diagram, species type *y* has 2 binding sites C and D. This reaction shows that an entity *t* of type *v* binds to an entity *s* of type *y* irrespective of the state of the C binding site on *s*. The state of the C binding site on *s* is captured by the variable *G* which is mapped from the reactants to the product.



In the general case a reaction may move the subgraph represented by a variable such as *G* from one binding site to another. However in many cases the variable components will not move. For this case the proposal includes syntactic 'sugar'. Binding sites can be omitted from a reaction if those binding sites are entirely unchanged by the reaction and the reaction is generalized to cover all states of that binding site. (The verbose form can be reconstructed by referring to the species type definitions for the species type instances in the reaction.) As an example, the reaction shown in the previous diagram can be recast as shown in the following diagram.



Acknowledgments

Support for the development of SBML and associated software and activities comes from the National Human Genome Research Institute (USA), the National Institute of General Medical Sciences (USA), the International Joint Research Program of NEDO (Japan), the ERATO-SORST Program of the Japan Science and Technology Agency (Japan), the Ministry of Agriculture (Japan), the Ministry of Education, Culture, Sports, Science and Technology (Japan), the BBSRC e-Science Initiative (UK), the DARPA IPTO Bio-Computation Program (USA), and the Air Force Office of Scientific Research (USA).

References

- [1] M. Hucka *et al* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models *Bioinformatics*, Vol 19, 524-531
- [2] A. Finney and M. Hucka Systems Biology Markup Language: Level 2 and Beyond *Biochem. Soc. Trans.*, Vol 31, 1472-1473
- [3] M. Hucka *et al* Evolving a Lingua Franca and Associated Software Infrastructure for Computational Systems Biology: The Systems Biology Markup Language (SBML) Project *IEE Systems Biology*, In Press
- [4] M. Ginkel Modular SBML, Proposal for an Extension of SBML towards level 2 *Proceedings of the 5th Workshop on Software Platforms for Systems Biology*, <http://sbml.org/workshops/ffth/sbml-modular.pdf>
- [5] J. Webb BioSpice MDL Model Composition and Libraries <http://bio.bbn.com/biospice/mdl/design/compose.html>
- [6] A. Finney Systems Biology Markup Language (SBML) Level 3 Proposal: Model Composition Features <http://www.cds.caltech.edu/~afinney/model-composition.pdf>
- [7] A. Finney Systems Biology Markup Language (SBML) Level 3 Proposal: Multicomponent Species Features <http://www.cds.caltech.edu/~afinney/multi-component-species.pdf>
- [8] W.S. Hlavacek *et al* The complexity of complexes in signal transduction *Biotechnol. Bioeng.*, Vol 84, 783-794
- [9] P. Loriaux *et al* An Integrated Modeling Environment for Cell Systems Biology *BISTIC Symposium 2003*
- [10] Le Novère, N. and Shimizu, T. S. StochSim: modeling of stochastic biomolecular processes *Bioinformatics*, Vol 17, 575-576