BioModels.net, tools and resources to support Computational Systems Biology

Nicolas Le Novère
EMBL-EBI
Wellcome-Trust Genome Campus, Hinxton CB10 1SD, UK

September 15, 2005

Abstract

The field of Computational Systems Biology matured quickly. If one wants it to fulfil its central role in the new Biology, the reuse of quantitative models needs to be facilitated. The community has to develop standards and guidelines in order to maximise the diffusion of its scientific production, but also to render it more trustworthy. We will review the various projects recently launched by the international BioModels.net initiative: MIRIAM is a standard to curate and annotate models, in order to facilitate their reuse. SBO is an ontology aimed to be used within models, in order to characterise their components. BioModels Database is a resource that allows biologists to store, search and retrieve published mathematical models of biological interests. We expect that those resources together with the use of formal languages such as SBML, will support the fruitful exchange and reuse of quantitative models.

1 Introduction

The rising popularity of Systems Biology, and its recognition as a new field of life science, brought forward its computational part, formerly a specific field of theoretical (or mathematical) Biology. As a consequence, what was once the territory of a small population of specialists is now visited by various actors of biomedical research. In parallel, the formal models used in

Biochemistry and Cell Biology are growing, both in size and complexity. A given modeller is therefore less likely to be an expert of all the corners of a quantitative model, whether the biological knowledge or even the mathematical approaches. Finally, the population of modellers can no longer be identified with the tribe of software developers.

This maturity called for a shift of paradigm in the way software tools are developed and used in the community. The design of standard formal languages to encode models, such as SBML [4] or CellML [6], was a first step. Their development actually served modellers in more than one aspect, fostering the creation of an actual community, and helping to shed light on the bottlenecks that precluded the smooth diffusion and reuse of quantitative models. Now that the way has been paved, one needs to walk forward.

First of all, one needs more automated support to handle formal models. Modellers should not have to fiddle with the gritty details of file formats for instance, or to have to dissect-out a model to understand what it is about. Secondly, now that the syntax problems are taken care of, the community shall move to the semantics. Finally, one needs to integrate modelling work with the other sources of knowledge.

The BioModels.net project, decided in 2004, is the next step: an international effort to (1) define agreed-upon standards for model curation, (2) define agreed-upon vocabularies for annotating models with connections to biological data resources, and (3) provide a free, centralised, publicly-accessible database of annotated, computational models in SBML and other structured formats.

2 Minimal Information Requested In the Annotation of Models

The possibility to encode a model under a standard format is far from being sufficient to make it understandable by user. If searching for existing relevant models, a researcher comes after a model *Model1* describing the reactions A and B between the molecular components X and Y, what can he/she makes any use of it? Where does this model come from? What are the components X and Y? It could help to know what process is modelled by A and B. Providing one finally elucidates the origin of the model, and the identity of its components, how can we know that when instantiated, this model provides the correct numerical results?

The aim of MIRIAM [5] is to define processes and schemes that will increase the confidence in model collections and enable the assembly of model collections of high quality. A first part of the guidelines is a standard for reference correspondence dealing with the syntax and semantics of the model. A second part is a proposed annotation scheme that specifies the documentation of the model by external knowledge. The scheme for annotation can itself be further subdivided into two sections. The *attribution* covers the minimum information that is required to associate the model with a reference description and an actual encoding process. The *external data resources* covers information required to relate the components of quantitative models to established data resources or controlled vocabularies.

The aim of standard for reference correspondence is to ensure that the model is properly associated with a reference description and is consistent with that reference description. In order to be declared MIRIAM-compliant, a quantitative model must fulfil the following rules:

1. The model must be encoded in a public, standardised, machine-readable format such as (but not restricted to) SBML or CellML, and it must comply with the standard in which it is encoded.

2. The model must be clearly related to a single reference description. If a model is derived from several initial reference descriptions, there must still be a reference description that describes or references a set of results that one can expect to reproduce when simulating the derived/combined model.

3. The encoded model structure must reflect the biological processes listed in the reference description (a one-to-one correspondence between model components is not required).

4. Quantitative attributes of the model, such as initial conditions and parameters, as well as kinetic expressions for all reactions, have to be defined, in order to allow to instantiate a simulation.

5. The model, when instantiated within a suitable simulation environment, must be able to reproduce all results given in the reference description that can readily be simulated.

In order to be confident in re-using an encoded model, one should be able to trace its origin, and the people who were involved in its inception. The following information should always be joined with an encoded model:

- The preferred name of the model, in order to facilitate discussions about it.
- A citation of the reference description with which the model is associated, either as a complete bibliographic record, or as a unique identifier, Digital Object Identifier (`http://www.doi.org`), PubMed identifier (`http://www.pubmed.gov`), unambiguous URL [12] pointing to the description itself etc.
- Name and contact information for the creators who actually contributed to the encoding of the model in its present form.
- The date and time of creation, and the date and time of last modification.
- A precise statement about the terms of distribution. The statement can be anything from “freely distributable” to “confidential”. MIRIAM being intended to allow models to be communicated better, terms of distribution are essential for that purpose.

The aim of the external data resources annotation scheme is to link model constituents to corresponding structures in existing and future open access bioinformatics resources. Such data resources can be, for instance, database or controlled vocabularies. This will permit the identification of model constituents and the comparison of model constituents between different models, but also the search for specific constituents in models.

This annotation must permit to unambiguously relate a piece of knowledge to a model constituent. The referenced information should be described using a triplet {“data-type”, “identifier”, “qualifier”}. The “data-type” is a unique, controlled, description of the type of data, written as a Unique Resource Identifier [11] (whether a Uniform Resource Locator [12] or a Uniform Resource Name [13]) . The “identifier”, within the context of the “data-type”, points to a specific piece of knowledge. The “qualifier” is a string that serves to refine the relation between the referenced piece of knowledge and the described constituent. Example of qualifiers are “has a”, “is version of”, “is homolog to”, etc. To enable interoperability, the community will have to agree on a set of standard valid URIs. and an API should be created so that a tool can automatically retrieve valid URL(s) corresponding to a given URI. The list should be able to evolve with the evolution of data resources.

Whilst many controlled vocabularies exist that can be used to annotate quantitative models, several additional small controlled vocabularies are required to enable the systematic capture of information in those models. This is why BioModels.net partners started to develop their own ontology.

3 Systems Biology Ontology

An ontology is defined here in its information science meaning, as a hierarchical structuring of knowledge. In our case, it is a set of relational vocabularies, that is a set of terms linked together. Each term has a definition and a unique identifier. The most famous ontology in life-science is Gene Ontology (GO) [1].

One of the goals of the Systems Biology Ontology (SBO) is to facilitate the immediate identification of the relation between a model component and the model structure. SBO is currently made up of three different vocabularies.

1. A classification of rate laws. This CV will be a taxonomy of kinetic rate equations. Examples of potential terms in this CV are “Mass action”, “Henri-Michaelis-Menten”, “Hill” etc. Note that although taking the same mathematical form, the rate-laws “Henri-Michaelis-Menten” “Van Slyke” and “Briggs-Haldane”, being based on different assumptions, will be represented by different terms. This will help a user to choose the adequate conversion to elementary steps if needed.
2. A taxonomy of the roles of reaction participants, including the following potential terms: “substrate”, “catalyst”, “inhibitor”, “competitive inhibitor”, “non-competitive inhibitor” etc.
3. A CV for parameter roles in quantitative models. This CV will include terms like “Hill coefficient”, “Michaelis constant” etc.

Within a vocabulary, the terms are related by "is a" inheritances, which represent sub-classing. However, contrary to GO, and most of the similar ontologies, the links in SBO will cross the vocabulary barriers. For instance a term defining a rate-law will have children representing the relevant reacting species parameters.

The annotation of model components with SO terms will be an essential step to reach MIRIAM-compliance. Not only such an annotation will be important to understand and to programmatically analyse models, it will also power the search strategies used by the databases of models, and in particular BioModels Database.

4 BioModels Database

As for all types of knowledge, quantitative models will be only as useful as their access and reuse is easy for all scientists. Some general repositories of quantitative models have been set up, such as the CellML repository [3], JWS Online [9] and the former SBML repository, and of more restricted focus, e.g. SenseLab ModelDB [8], the Database of Quantitative Cellular Signalling [10] and SigPath [2]. However no general public resource existed that offers complete database services, in terms of browsing, searching and retrieval, of annotated models

BioModels Database is an annotated resource of quantitative models of biomedical interest developed in collaboration by the SBML Team (USA), the EMBL-EBI (United-Kingdom), the Systems Biology Group of the Keck Graduate Institute (USA), and JWS Online at the Stellenbosch University (South Africa). Models can be submitted by anyone to the curation pipeline of the database. At present, BioModels Database aims to store and annotate models that can be encoded with SBML and CellML. BioModels Database goes further than MIRIAM, requiring not only the existence of a reference description, but considering only models described in the peer-reviewed scientific literature.

A series of automated tasks are performed by the pipeline prior to human intervention (see Materials and Methods for details):

- Verification that the file is well-formed XML.
- If necessary, conversion to the latest version of SBML.
- Verification of the syntax of SBML.
- Series of consistency checks, enforcing the validity of the model.

If any of those steps is not completed, a member of the distributed team of curators can reject the model, or instead correct it and resubmit it to the pipeline. The last, and most important step, of the curation process, is verifying that when instantiated in a simulation, the model provides results corresponding to the reference scientific article. Once the model is verified to be valid SBML, and to correspond well to the article, it is accepted in the production database for annotation.

Model components are annotated with references to adequate resources, such as terms from controlled vocabularies (Taxonomy, Gene ontology, ChEBI etc.) and links to other databases (UniProt, KEGG, Reactome etc.). This annotation is a crucial feature of BioModels Database that permits the unambiguous identification of molecular species or reactions and is used in search strategies.

The thorough annotation of models allows a triple search strategy to be run in order to retrieve models of interest. Since the models encoded in SBML are stored directly in an XML native database, those models can be retrieved based on the content of their elements and attributes, using XPath. Models can be retrieved by searching directly the annotation database, using SQL. Although this search is quick, it requires the knowledge of the exact identifiers used by curators to annotate the model. A more advanced search system has therefore been implemented, using direct string search of the third party resources, retrieval of the relevant identifiers, and then search BioModels database for the models annotated with those identifiers. As a consequence, the user can retrieve all the models dealing with “cell cycle” or “MAPK”, without having to type “GO:0007049” or “P27361”. Once retrieved, the models of interest can be downloaded in SBML Level2. A number of export filters are under development to provide the models in a wider range of formats.

Although BioModels database is a very recent resource, it has already gained momentum thanks to the support of the SBML community, but also of major scientific actors such as Nature Publishing Group, who publicised its launching and started to submit models. The growth of BioModels Database is currently limited by the curation workforce, to only a dozen models a month. It is expected that the existence of a public resource will contribute to improve the quality of the models produced, by putting peer-pressure on the modellers.

5 acknowledgements

Michael Hucka initiated the whole BioModels.net idea, and is a major force behind its various projects. Thanks to all the authors of the MIRIAM standard [5]. The computing infrastructure of the BioModels Database has been almost entirely developed by Marco Donizelli. Thanks also to the other curators of the database, Harish Dharuri and Bruce Shapiro, and to all the model contributors.

References

[1] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., *et al.*, Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.* **25**, 25–29 (2000).

[2] Campagne, F., Neves, S., Chang, C., Skrabanek, L., Ram, P., Iyengar, R., and Weinstein, H. (2004) Quantitative information management for the biochemical computation of cellular networks. *Science STKE,* **248**, PL11.

[3] Lloyd, C. The CellML repository. Available via the World Wide Web at `http://www.cellml.org/examples/repository/index.html`.

[4] Hucka, M., Bolouri, H., Finney, A., Sauro, H., Doyle, J., *et al.*, The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).

[5] Le Novère, N., Finney, A., Hucka, M., Bhalla, U., Campagne, F., Collado-Vides, J., Crampin, E., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J., Spence, H., Wanner, B. , Minimal information requested in the annotation of models (MIRIAM). *Nature Biotechnol* submitted.

[6] Lloyd, C., Halstead, M., & Nielsen, P., CellML: its future, present and past. *Prog. Biophys. Mol. Biol.* **85**, 433–450 (2004).

[7] Martin, S., Niemi, M., & Senger, M., Life sciences identifiers rfp response. Available via the World Wide Web at `http://www.omg.org/docs/lifesci/03-12-02.txt`.

[8] Migliore, M., Morse, T., Davison, A., Marenco, L., Shepherd, G., and Hines, M. (2003) Modeldb making models publicly accessible to support computational neuroscience. *Neuroinformatics,* **1**, 135–139.

[9] Olivier, B. and Snoep, J. (2004) Web-based kinetic modelling using JWS Online. *Bioinformatics,* **20**, 2143–2144.

[10] Sivakumaran, S., Hariharaputran, S., Mishra, J., and Bhalla, U. (2003) The database of quantitative cellular signaling: management and analysis of chemical kinetic models of signaling networks. *Bioinformatics,* **19**, 408–415.

[11] Berners-Lee, T., Fielding, R., & Masinter, L., Uniform resource identifier (uri): Generic syntax. Available via the World Wide Web at `http://www.gbiv.com/protocols/uri/rfc/rfc3986.html`.

[12] Berners-Lee, T., Uniform resource locators (url). a syntax for the expression of access information of objects on the network. Available via the World Wide Web at `http://www.w3.org/Addressing/URL/url-spec.txt`.

[13] Moats, R., Urn syntax. Available via the World Wide Web at `http://www.ietf.org/rfc/rfc2141.txt`.