

BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems

Nicolas Le Novère*, Benjamin Bornstein¹, Alexander Broicher, Mélanie Courtot, Marco Donizelli, Harish Dharuri², Lu Li, Herbert Sauro², Maria Schilstra³, Bruce Shapiro¹, Jacky L. Snoep⁴ and Michael Hucka⁵

European Bioinformatics Institute, EMBL Wellcome-Trust Genome Campus, Hinxton, CB10 1SD, UK, ¹Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA, ²Keck Graduate Institute, 535 Watson Drive, Claremont, CA 91711, USA, ³STRI, University of Hertfordshire, Hatfield, Herts AL10 9AB, UK, ⁴Department of Biochemistry, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa and ⁵Control and Dynamical Systems, California Institute of Technology, Pasadena, CA 91125, USA

Received July 29, 2005; Revised October 4, 2005; Accepted October 16, 2005

ABSTRACT

BioModels Database (<http://www.ebi.ac.uk/biomodels/>), part of the international initiative **BioModels.net**, provides access to published, peer-reviewed, quantitative models of biochemical and cellular systems. Each model is carefully curated to verify that it corresponds to the reference publication and gives the proper numerical results. Curators also annotate the components of the models with terms from controlled vocabularies and links to other relevant data resources. This allows the users to search accurately for the models they need. The models can currently be retrieved in the SBML format, and import/export facilities are being developed to extend the spectrum of formats supported by the resource.

INTRODUCTION

The number of quantitative models trying to explain various aspects of the cellular machinery is increasing at a steady pace, thanks in part to the rising popularity of systems biology (1). However, as for all types of knowledge, such models will only be as useful as their access and reuse is easy for all scientists. A first step was to define standard descriptions to encode quantitative models in machine-readable formats. Example of such formats are CellML (2) and the Systems Biology Markup Language (SBML) (3,4). The biomedical community now needs public integrated resources, where authors can

deposit, in controlled formats, the models they describe in scientific publications.

Some general repositories of quantitative models have been made available, such as the CellML repository CellML repository [(5), <http://www.cellml.org/examples/repository/index.html>] JWS Online (6) and the former SBML repository. In addition specialist repositories include SenseLab ModelDB (7), the Database of Quantitative Cellular Signalling (DOCQS) (8) and SigPath (9). However no general public resource existed that allowed the user to browse, search and retrieve annotated models

Here we present BioModels Database, developed as part of the BioModels.net initiative (<http://www.biomodels.net/>). BioModels.net is a collaboration between the SBML Team (USA), the EMBL-EBI (UK), the Systems Biology Group of the Keck Graduate Institute (USA), the Systems Biology Institute (Japan) and JWS Online at Stellenbosch University (South Africa). Its aims are as follows: (i) to define agreed-upon standards for model curation, (ii) to define agreed-upon vocabularies for annotating models with connections to biological data resources and (iii) to provide a free, centralized, publicly accessible database of annotated, computational models in SBML and other structured formats.

BioModels Database is an annotated resource of quantitative models of biomedical interest. Models are carefully curated to verify their correspondence to their source articles. They are also extensively annotated, with (i) terms from controlled vocabularies, such as disease codes and Gene Ontology terms and (ii) links to other data resources, such as sequence or pathway databases. Researchers in the biomedical and life science communities can then search and retrieve models related to a particular disease, biological process or molecular complex.

*To whom correspondence should be addressed. Tel: +44 1223 494521; Fax: +44 1223 494468; Email: lenov@ebi.ac.uk

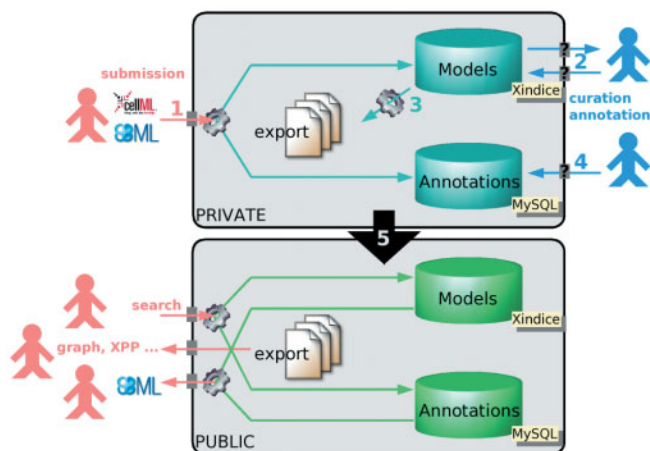


Figure 1. Pipeline describing the structure of BioModels database.

SUBMISSION, CURATION AND ANNOTATION

Models can be submitted by anyone to the curation pipeline of the database (Figure 1). At present, BioModels Database aims to store and annotate models that can be encoded with SBML. CellML models are also accepted. These model formats are synonymous with models that can be integrated or iterated forwards in time, such as ordinary differential equation models. Although we are aware that this means we can cover only a restricted part of the modeling field, we make this our initial focus for the following reason: (i) since a crucial part of the curation process is the verification that the models produce numerical results similar to the ones described in the reference article, iterative simulations over ranges of parameter values and perturbation of simulations at equilibrium are mandatory and (ii) a very large number of such models have already been published, and the pace of their publication is increasing steadily. As a consequence, they are sufficient to consume all the curation workforce we have, and we can envision to gather in the near future.

To be accepted in BioModels Database, a model must be compliant with MIRIAM, the Minimal Information Requested in the Annotation of Models (10). One of the requirements of MIRIAM is that a model has to be associated with a reference description that provides directly, or through references, the structure of the model, the necessary quantitative parameters and presents the results of numerical analysis of the model. BioModels Database further refines the notion of reference description, by 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. Curators do not normally challenge the biological relevance of the models, and assume the peer-review process already filtered out unsuitable contributions. However, in specific cases, curators can spot mistakes in an article and, with the agreement of the authors, modify the model accordingly. 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.

In order to be confident in reusing an encoded model, one should be able to trace its origin, and the people who were involved in its inception. The following information is therefore added to the model: (i) either a PubMed identifier (<http://www.pubmed.gov>) or a DOI (<http://www.doi.org>) or an URL that permits identifying the peer-review article describing the model; (ii) name and contact details of the individuals who actually contributed to the encoding of the model in its present form; (iii) name and contact of the person who finally entered the model in the production database and who should be contacted if there is a problem with the encoding of the model or the annotation.

In addition, model components are annotated with references to relevant 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 in that it permits the unambiguous identification of molecular species or reactions and enables effective search strategies.

SEARCH AND RETRIEVAL

The thorough annotation of models allows a triple search strategy to be run in order to retrieve models of interest (Figure 2).

The models converted to SBML are stored directly in an XML native database (Xindice, <http://xml.apache.org/xindice/>), enabling those models and/or their components to be retrieved based on the content of their elements and attributes (using XPath, <http://www.w3.org/TR/xpath>). For instance, the user can search for a given string of characters in the id, name and notes elements of each model component.

Models can be retrieved by searching the annotation database directly, using SQL. Although this search is quick, it requires knowing the exact identifiers used by curators to annotate a model and relate it to third party resources, such as UniProt accession, Gene Ontology Term ID, etc.

We, therefore, implemented a more advanced search system. A user can actually search third party resources directly, such as PubMed, Gene Ontology and UniProt, for instance with literal text matching. The search system retrieves the relevant identifiers and then searches 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'.

Several searches of any of the three types can also be run in parallel, the results being thereafter combined with boolean operators.

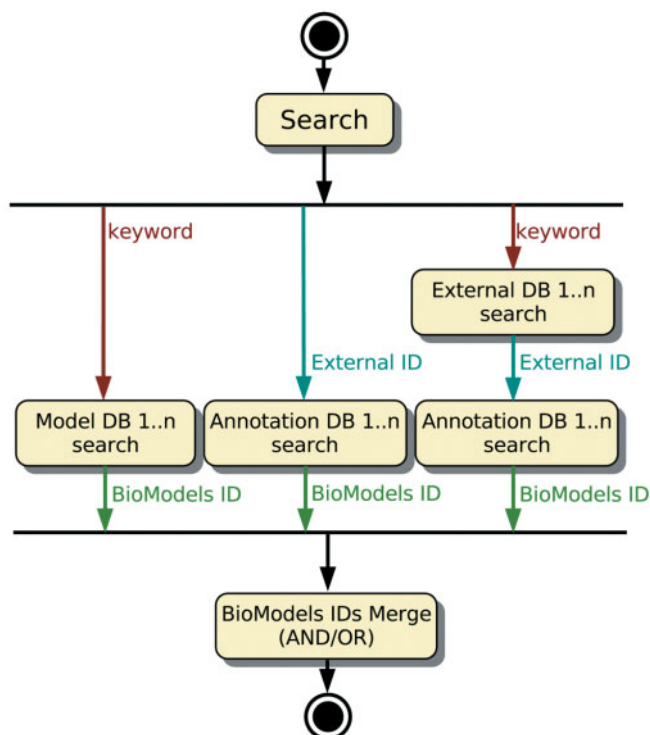


Figure 2. Schema representing the cascading search strategy. The result is a list of BioModels entries.

Once retrieved, the models of interest can be downloaded in SBML Level 2 format. A number of export filters are under development to provide the models in a wider range of formats.

BioModels Database is copyrighted by The BioModels Team, i.e. the set of individuals developing the resource. However, the copyright on the database does not imply copyright of the original models in BioModels Database. Each individual model retains the copyright assigned by both the creator(s) of the model and the author(s) of the reference publication. Users may distribute verbatim copies of the entire content of BioModels Database, including the models and their annotations, or a subset of the models. Users may also modify any of the models in any way, provided that at least one of the following condition is fulfilled:

- The modified model is used only within the user's organization.
- The modifications are placed in the Public Domain, or otherwise made Freely Available by allowing the Copyright Holders of the model to include the modifications in the standard version of the model.
- The modified model is renamed, and both BioModels Database identifier and any mention of the Copyright Holders of the model is removed.
- Other distribution arrangements are made directly with the Copyright Holders of the model(s) in question.

This restricted license has been rendered necessary by the specific nature of the data distributed by BioModels Database. If a user of BioModels Database downloads a kinetics model

and modifies it, the resulting model could be meaningless, or even worse, exhibits a behaviour completely different of what was initially meant by the authors and the creators. Therefore, we thought that the best compromise was to let complete freedom of reuse and modification, providing that BioModels Database is not associated with any modification.

PERSPECTIVE

Although BioModels Database is a very recent resource, it has already gained momentum thanks to the support of the SBML community, which has started to submit models, and major scientific publishing actors such as Nature Publishing Group, which has publicized the launch of the database. The growth of BioModels Database is currently limited, by the size of the curation workforce, to only a dozen new models a month. We expect that the existence of this public resource will contribute to an improvement in the quality of the models published by establishing an additional process for evaluating those models. The increase in quality and the continuously improved support of SBML by modelling tools should increase the speed of curation. Meanwhile, we will continue to improve the search and retrieval facilities, and support more export formats, so that users can directly use the models contained in the database even in non-SBML compliant tools.

ACKNOWLEDGEMENTS

Authors thank G. Bard Ermentrout, Sarah Keating, Joanne Matthews and Nicolas Rodriguez for sharing their code. Funding to pay the Open Access publication charges for this article was provided by EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Kitano, H. (2005) International alliances for quantitative modeling in systems biology. *Mol. Syst. Biol.* doi: 10.1038/msb4100011.
2. Lloyd, C., Halstead, M.D. and Nielsen, P.F. (2004) CellML: its future, present and past. *Prog. Biophys. Mol. Biol.*, **85**, 433–450.
3. Hucka, M., Bolouri, H., Finney, A., Sauro, H.M., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
4. Finney, A. and Hucka, M. (2003) Systems biology markup language: level 2 and beyond. *Biochem. Soc. Trans.*, **31**, 1472–1473.
5. Lloyd, C. The CellML repository.
6. Olivier, B.G. and Snoep, J.L. (2004) Web-based kinetic modelling using JWS online. *Bioinformatics*, **20**, 2143–2144.
7. Migliore, M., Morse, T.M., Davison, A.P., Marengo, L., Shepherd, G.M. and Hines, M.L. (2003) ModelDB: making models publicly accessible to support computational neuroscience. *Neuroinformatics*, **1**, 135–139.
8. 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.
9. Campagne, F., Neves, S., Chang, C.W., Skrabanek, L., Ram, P.T., Iyengar, R. and Weinstein, H. (2004) Quantitative information management for the biochemical computation of cellular networks. *Sci. STKE*, **248**, PL11.
10. Le Novère, N., Finney, A., Hucka, M., Bhalla, U., Campagne, F., Collado-Vides, J., Crampin, E., Halstead, M., Klipp, E. *et al.* (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.*, **23**, in press.