Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Toward a Curated CellML Model Repository

David Nickerson, Carey Stevens, Matt Halstead, Peter Hunter and Poul Nielsen
Bioengineering Institute
The University of Auckland
Auckland, New Zealand
Email: d.nickerson@auckland.ac.nz

Abstract— **With the development of standards for the description of experimental data and mathematical models of physiological processes has come the advent of publicly accessible repositories of such descriptions. To make these repositories acceptable to the community, the descriptions contained therein must be curated to provide an indication of how accurately the standard description represents the actual experimental data or mathematical model. We define here a curation method for CellML encoded models and propose a model repository workflow using such curation within the CellML model repository.**

I. Introduction

As the complexity of computational models increases, formal vocabularies are needed to reduce the growing heterogeneity of expressions. Standards are being developed to formalize the description of experimental data and mathematical models of physiological processes [1], [2]. Ontologies that incorporate semantic descriptions of modeling concepts eliminate ambiguities in the modeling environment. To this end, ontologies and representation languages that include these ontologies are being developed under the IUPS Physiome Project (`www.physiomeproject.org`), facilitating communication of biological models through tools for building, sharing, interpreting, and visualizing models [3].

With the development of these languages comes the ability to create and populate repositories of models that are freely available for use by the scientific community. An important issue being addressed by model repositories is that of the community's level of confidence that a given representation or implementation of a model is accurate. Such accuracy can be interpreted as how faithfully the representation or implementation reflects the reference description of the model or perhaps more importantly as how accurately the underlying phenomena represented by the model are reproduced. Standards are being developed that start to address these issues [4] and model repositories are beginning to implement the ideas proposed by such standards (a good example is the BioModels repository: `www.biomodels.net`).

The most advanced of the physiome representation languages is CellML (`www.cellml.org`). Initially designed for application to models of cellular electrophysiology and reaction pathway models, CellML has since been used in a wide range of mathematical models, including constitutive laws for continuum mechanics. The CellML model repository (`www.cellml.org/models`) has its origins as a distribution site for models encoded in CellML to provide examples on the use of CellML as well as to test the features of the language as it developed. It then progressed into a repository of previously published biological models that had been encoded into CellML, a state in which the repository existed for several years, with the number of encoded models growing to over 300. These models had all been encoded into CellML based on the literature, and the models were not tested for consistency or completeness. As tools have evolved we are now in a position to begin curating the models in the repository.

In the following sections we describe recent developments in the CellML model repository which provide a base upon which to develop audited model curation. We then discuss current developments underway to implement the actual model curation.

A. MIRIAM

In a recent article a representative group from the biological modeling community proposed a set of rules for curating quantitative models of biological systems [4]. These rules specify the minimum information requested in the annotation of biochemical models (MIRIAM) and models that pass all the tests and fulfills all the conditions listed in MIRIAM are deemed a MIRIAM-compliant model.

MIRIAM lists six rules that models must adhere to in order to be MIRIAM-compliant. The six rules can be summarised as:

1) the model must be encoded in a public, machine readable format, either standard such as SBML or CellML, or supported by specific software applications;
2) the encoded model must comply with the standard in which it is encoded;
3) the model must be clearly related to a single reference description the describes or references a set of results that one can expect to reproduce using the model;
4) the encoded model structure must reflect the biological processes listed in the reference description;
5) the encoded model must be instantiated in a simulation, meaning that the quantitative attributes of the model have to be defined; and
6) the model, when instantiated within a suitable simulation environment, must be able to reproduce all relevant results given in the reference description that can readily be simulated.

Additionally, MIRIAM specifies annotation that must be included with a quantitative model to achieve MIRIAM compli-

1-4244-0033-3/06/$20.00 ©2006 IEEE.

ance:

1) the preferred name of the model;
2) a citation of the reference description with which the model is associated;
3) name and contact information for the model creators, that is, the people who actually contributed to the encoding of the model in its present form;
4) the date and time of creation, and the date and time of last modification; and
5) a precise statement about the terms of distribution.

With some minor additions, the CellML metadata specification (`www.cellml.org/specifications/metadata/`) will provide a standard for MIRIAM-compliant annotation of models. Models in the CellML model repository will, however, have to be modified to include all annotations required by MIRIAM. The first two rules above are already met by most (if not all) models in the CellML model repository – *i.e.,* they are encoded in valid CellML.

While MIRIAM provides a useful standard for model curation, it does not fully encompass the requirements we foresee for model curation in the CellML model repository. In the following sections we refer to various aspects of MIRIAM compliance as well as further curation we see as useful for users of the CellML model repository.

II. CellML Model Curation

Currently all models in the CellML model repository have either been encoded directly from previously published peer-reviewed journal articles or been developed as part of the publication process. Such peer-review is an essential requirement for any model to be accepted by the community. Thus, CellML model curation is designed to complement the peer-review process.

The basic measure of curation in a CellML model is defined by the curation level of the model document. We have defined four levels of curation, loosely summarized as:

- **Level 0**: not curated;
- **Level 1**: the encoded model accurately represents the published model;
- **Level 2**: the encoded model works; and
- **Level 3**: the actual model satisfies domain specific constraints.

Levels 1, 2, and 3 are seen as progressive steps forward in the evolution of a CellML model without restrictions on maintaining the requirements of a lower level curation. For example, as described below, modifying a level 1 model to meet the requirements for level 2 curation will generally break level 1 curation. Similarly, changes for level 3 curation may break level 2 and/or level 1 curation.

A. Level 0 curation

A CellML model with a curation level of 0 has received no curation, other than checks on the validity of the CellML document. This is the default initial curation level assigned to a model when it first enters the CellML model repository workflow. Most models imported into the model repository from the previous repository are at this curation level.

B. Level 1 curation

With a curation level of 1, the CellML model is assured to be an accurate representation of the original published model, including typographical, mathematical, and dimensional errors. This level of curation primarily exists to provide a historical reference of published models. Typically models are published containing some errors in the mathematics, or the units for a variable are missing or incorrect, or some initial conditions or parameter values might be missing. In these cases, creating a CellML representation of the model which is capable of reproducing the results given in the original paper requires modifications to the mathematics, variables, units, *etc.* which results in a CellML version of the model which no longer matches the original published model.

Curation level 1 is also useful when dealing with primarily qualitative models encoded in CellML. For such models there is usually not enough information to perform the quantitative simulations required to generate the results required to achieve level 2 curation, thus a level 1 curated version of the model is most valuable.

C. Level 2 curation

To achieve a curation level of 2, the CellML encoded model has been checked for (i) typographical errors, (ii) consistency of units, (iii) that all parameters and initial conditions are defined, (iv) that the model is not over-constrained, in the sense that it contains equations or initial values which are either redundant or inconsistent, and (v) that running the model in an appropriate simulation environment reproduces the results published in the original paper.

D. Level 3 curation

Level 3 curation provides assurance that the actual mathematical model correctly represents the underlying phenomena encompassed by the model. A model with a curation level of 3 has been checked for the extent to which it satisfies domain specific biophysical constraints (*e.g.,* conservation of mass, momentum, energy) and the model metadata must describe these extents. This level of model curation is highly domain specific and needs to be conducted by domain experts.

The metadata for a level 3 curated model must also specify the range of applicability of the model – *i.e.,* explicit statements about assumptions made as well as specie, age, disease state, tissue type, *etc.* dependencies.

E. Model development

It is the hope of the CellML group that as new models develop the CellML representation of the model is concurrently developed. Thus enabling a single CellML representation of a model to have level 1, 2 and 3 curation status. Historically, however, published models contain sufficient error to prohibit the use of a level 1 CellML version of the model to pass all the requirements of level 2 curation. Changes made to the level

1 version of the model to pass level 2 curation results in a level 2 version of the model which no longer satisfies level 1 curation.

While a higher curation level version of a model will typically be more useful in practice it is useful to provide versions of a model at each curation level. The new model repository workflow (Section III) provides the facility to store multiple versions of a model allowing the level 0, 1, 2, and 3 versions to coexist in the repository. We are able, therefore, to serve from the model repository both a usable version of the model (level 2 or 3 curation) and an accurate representation of model as it was originally published (level 1 curation) and perhaps a development version of the model as it was initially uploaded into the repository (level 0). An essential feature of the repository is providing sufficient annotation to allow repository users to determine which version of a model is most suitable for their application.

III. CellML Model Repository Workflow

A key feature of the CellML model repository is the ability to simultaneously store multiple versions of a given model. Versions are distinguished via the model's URI and the document's metadata. This feature allows different versions of a model at various curation levels to be easily stored and used in the CellML model repository.

Fig. 1 provides an illustration of the proposed model repository workflow to be implemented in the CellML model repository. The current CellML model repository implements the steps up to and including level 0 curation. This workflow begins with a new model (or a new version of an existing model) submitted to the repository. The CellML document must first pass some basic validation checks to ensure the encoding of the model is valid.

Once the document passes validation it is checked for the metadata required in a CellML document for entry into the model repository. There is currently no required metadata except for models with a curation level of 3 (as described above in Section II-D). For all models, however, an interface is presented for the entry of a minimal set of information about both the model and its CellML encoding. This minimal metadata, if provided, enhance the document's usefulness in the model repository (*i.e.,* model categorization and repository searching). As model and domain specific ontologies develop the current interface will be developed to allow more detailed ontological annotation. Similarly, metadata required by MIRIAM (Section I-A) can be marked compulsory and edited via this interface.

When a CellML model submitted to the repository has passed validity and metadata checks, it is annotated as being curated to the level 0 requirements and saved in the repository. Saving the model into the repository can either be as a new model, a new version of an existing model, or the replacement of an existing version of a model.

Level 0 curated models may then be reviewed by repository members with sufficient privileges in order to progress to level 1, 2, or 3 curation. Following a successful review, the reviewed version of the model is annotated with the new curation level and entered into the repository either replacing the level 0 curated version of the model or as a new version of the model. If a model fails to meet the requirements for the new curation level the model and/or the model's CellML encoding can be corrected and a new version of the model submitted to the model repository. The new version of the model can then progress through the repository workflow and be reviewed again, hopefully with more favorable results.

As models progress through review and resubmission cycles there is the possibility of either accumulating numerous versions of the same model or continually replacing a single version of the model in the repository. Model authors and developers are free to choose which of these approaches they deem most suitable. Some advantage is seen in a annotated history of model versions as a model is revised to meet the requirements of higher curation levels, an advantage which can similarly be achieved through a suitable modification history associated with a single version of a model.

IV. Conclusion

We have presented a proposed workflow that, once fully implemented, will provide curation of models in the CellML model repository. Model curation is based on the idea of increasing levels of curation with a higher level curation indicating increased assurance in the CellML encoding of a model and the model itself.

The CellML model repository defines four levels of model curation. The first three curation levels define how accurately the CellML encoding of a model reflects the peer-reviewed published model. The fourth curation level indicates that the actual model has been checked by a domain expert and found to accurately represent the physical phenomena encompassed by the model.

To date, the proposed model curation workflow has been partially implemented and work is currently underway to complete the implementation. While we discuss only the CellML model repository here, the proposed repository workflow and model curation is more generally applicable. As such, future work will see similar developments in other model repositories as part of the IUPS Physiome Project (www.physiomeproject.org).

Acknowledgment

This work was supported by Wellcome Trust and the Royal Society of New Zealand (Centre for Molecular Biodiscovery).

References

[1] A. A. Cuellar, C. M. Lloyd, P. F. Nielsen, D. P. Bullivant, D. P. Nickerson, and P. J. Hunter, "An overview of CellML 1.1, a biological model description language," *Simulation*, vol. 79, no. 12, pp. 740–747, Dec. 2003.

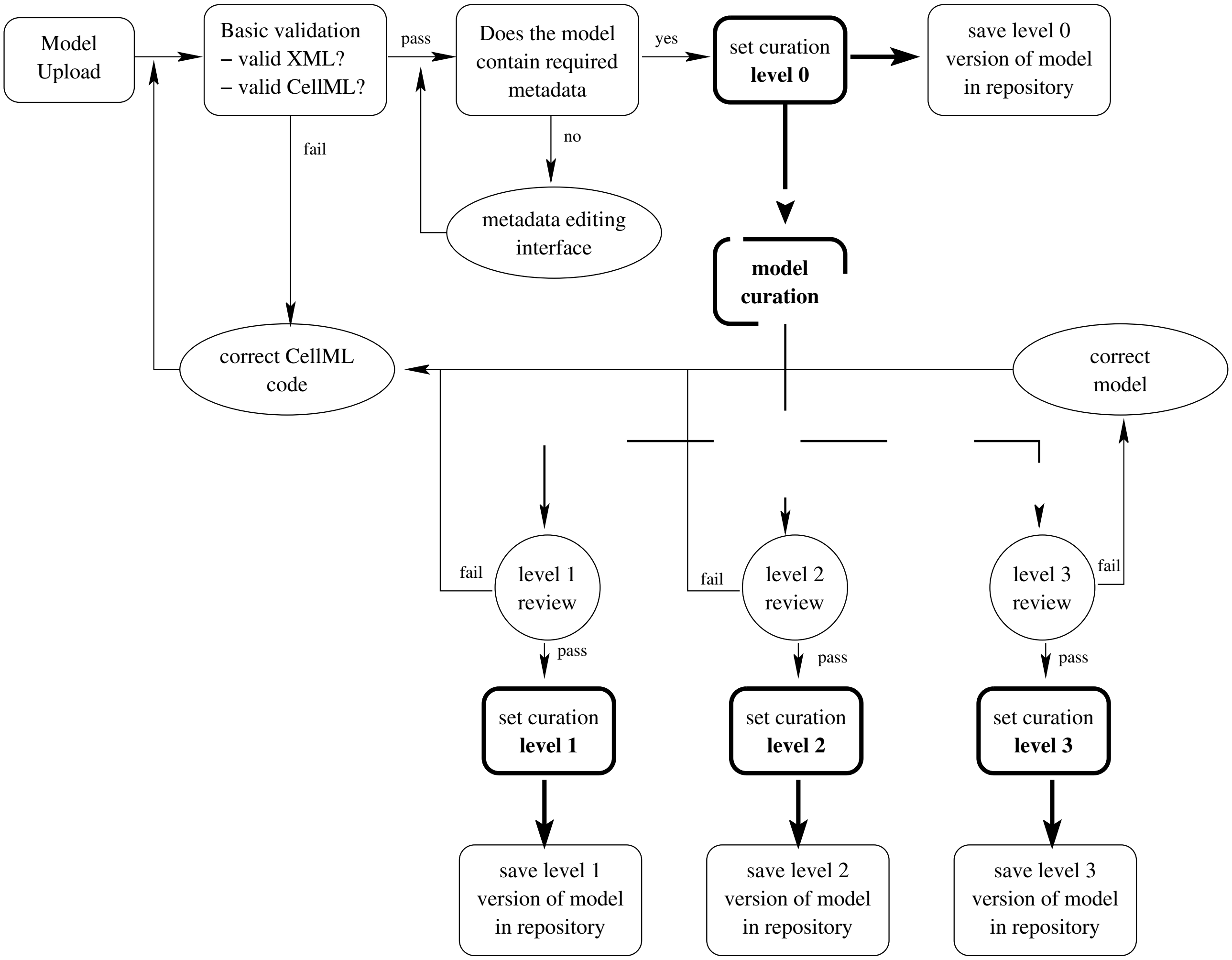

Fig. 1. Diagram representing the proposed CellML model repository workflow. To be stored in the repository a CellML model must reach at least level 0 curation.

[2] M. Hucka, H. Bolouri, A. Finney, H. M. Sauro, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cuellar, S. Dronov, M. Ginkel, V. Gor, I. I. Goryanin, W. Hedley, T. C. Hodgman, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. LeNovere, L. M. Loew, D. Lucio, P. Mendes, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang, "The systems biology markup language (sbml): A medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, pp. 524–531, 2003.

[3] G. R. Christie, D. Bullivant, S. Blackett, and P. J. Hunter, "Modelling and visualising the heart," *Comput. Visual. Sci.*, vol. 4, pp. 227–235, 2002.

[4] N. Le Novere, A. Finney, M. Hucka, U. Bhalla, F. Campagne, J. Collado-Vides, E. J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J. L. Snoep, H. D. Spence, and B. L. Wanner, "Minimum information requested in the annotation of biological models (MIRIAM)," *Nat. Biotechnol.*, vol. 23, no. 12, pp. 1509–1515, 2005.