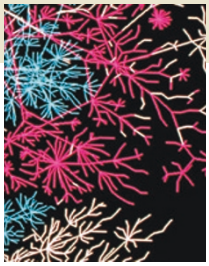


Minimum information requested in the annotation of biochemical models (MIRIAM)

Nicolas Le Novère[1,15], Andrew Finney[2,15], Michael Hucka[3], Upinder S Bhalla[4], Fabien Campagne[5], Julio Collado-Vides[6], Edmund J Crampin[7], Matt Halstead[7], Edda Klipp[8], Pedro Mendes[9], Poul Nielsen[7], Herbert Sauro[10], Bruce Shapiro[11], Jacky L Snoep[12], Hugh D Spence[13] & Barry L Wanner[14]

Most of the published quantitative models in biology are lost for the community because they are either not made available or they are insufficiently characterized to allow them to be reused. The lack of a standard description format, lack of stringent reviewing and authors' carelessness are the main causes for incomplete model descriptions. With today's increased interest in detailed biochemical models, it is necessary to define a minimum quality standard for the encoding of those models. We propose a set of rules for curating quantitative models of biological systems. These rules define procedures for encoding and annotating models represented in machine-readable form. We believe their application will enable users to (i) have confidence that curated models are an accurate reflection of their associated reference descriptions, (ii) search collections of curated models with precision, (iii) quickly identify the biological phenomena that a given curated model or model constituent represents and (iv) facilitate model reuse and composition into large subcellular models.

[1]European Bioinformatics Institute, Hinxton, CB10 1SD, UK. [2]Physiomics PLC, Magdalen Centre, Oxford Science Park, Oxford, OX4 4GA,K. [3]Control and Dynamical Systems, California Institute of Technology,Pasadena, California 91125, USA. [4]National Centre for Biological Sciences, TIFR, UAS-GKVK Campus, Bangalore 560065, India. [5]Institute for Computational Biomedicine, Weill Medical College of Cornell University, New York, New York 10021, USA. [6]Center for Genomic Sciences, Universidad Nacional Autónoma de México, Av. Universidad s/n, Cuernavaca, Morelos, 62100, Mexico. [7]Bioengineering Institute and Department of Engineering Science, The University of Auckland, Private Bag 92019, Auckland, New Zealand. [8]Max-Planck Institute for Molecular Genetics, Berlin Center for Genome based Bioinformatics (BCB), Ihnestr. 73, 14195 Berlin, Germany. [9]Virginia Bioinformatics Institute, Virginia Tech, Washington St., Blacksburg, Virginia 24061-0477, USA. [10]Keck Graduate Institute, 535 Watson Drive, Claremont, California 91711, USA. [11]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. [12]Triple-J Group for Molecular Cell Physiology, Department of Biochemistry, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa. [13]Department of Scientific Computing & Mathematical Modeling, GlaxoSmithKline Research & Development Limited, Medicines Research Centre, Gummels Wood Road, Stevenage, Herts, SG1 2NY, UK. [14]Purdue University, Department of Biological Sciences, Lilly Hall of Life Sciences, 915 W. State Street, West Lafayette, Indiana 47907-2054, USA. [15]These authors have contributed equally to the work. Correspondence should be addressed to N.L.N. (e-mail: lenov@ebi.ac.uk).

Published online 6 December 2005; doi:10.1038/nbt1156

During the genomic era we have witnessed a vast increase in availability of large amounts of quantitative data. This is motivating a shift in the focus of molecular and cellular research from qualitative descriptions of biochemical interactions towards the quantification of such interactions and their dynamics. One of the tenets of systems biology is the use of quantitative models (see **Box 1** for definitions) as a mechanism for capturing precise hypotheses and making predictions[1,2]. Many specialized models exist that attempt to explain aspects of the cellular machinery. However, as has happened with other types of biological information, such as sequences, macromolecular structures or

Box 1 Glossary

Some terms are used in a very specific way throughout the article. We provide here a precise definition of each one.

Quantitative biochemical model. A formal model of a biological system, based on the mathematical description of its molecular and cellular components, and the interactions between those components.

Encoded model. A mathematical model written in a formal machine-readable language, such that it can be systematically parsed and employed by simulation and analysis software without further human translation.

MIRIAM-compliant model. A model that passes all the tests and fulfills all the conditions listed in MIRIAM.

Reference description. A unique document that describes, or references the description of the model, the structure of the model, the numerical values necessary to instantiate a simulation from the model, or to perform a mathematical analysis of the model, and the results one expects from such a simulation or analysis.

Curation process. The process by which the compliance of an encoded model with MIRIAM is achieved and/or verified. The curation process may encompass some or all of the following tasks: encoding of the model, verification of the reference correspondence and annotation of the model.

Reference correspondence. The fact that the structure of a model and the results of a simulation or an analysis match the information present in the reference description.

Box 2 Case studies of MIRIAM-compliant models

To provide background to the motivations for MIRIAM, we provide here a number of case studies involving models encoded with the schemes described in this document.

User queries model database

In this scenario, a user wants to design a model of CDC2 function in the human cell-cycle. By interacting with a database consisting of models compliant with MIRIAM, this user searches all the models that contains CDC2 and represent cell cycle. Retrieving models of yeast and amphibian cell cycles, the user then reviews the models by reading the associated documentation and browsing other bioinformatics databases. By following links to databases of biochemical pathways, the user decides which model best describes what he/she knows about the function of CDC2 in the human cell cycle. The user then downloads this model and uses it as a basis for her/his own modeling work.

All of the above is possible if this proposal is applied, ensuring that models correspond to associated reference descriptions and are appropriately annotated.

Journal peer review: JWS Online

In this use case, we describe how a journal peer review process could incorporate MIRIAM, using the example of the procedure carried out by JWS Online[9] with its associated journals. When a manuscript describing a kinetic model is submitted to those journals, the authors are requested to submit the model description in electronic form (encoded in an accessible standard format). A curator parses the model using software that automatically checks its syntax (for instance, SBML and CellML validation tools), and if necessary, corrects the model. The curator then performs the verifications described in the section on reference correspondence. In particular, he/she attempts to reproduce the model results, as shown in the manuscript. If this fails, the curator contacts the authors in an attempt to correct the errors in the description or coding. After the curators and authors reach agreement on model description and simulation results, the model is made available to the reviewers and the authors, in a secure manner. A letter is sent to the reviewers with a set of instructions on how they can test the model remotely, running simulations at JWS Online directly from their web browsers. If the manuscript is accepted by the journal for publication, the model is moved to the public database of JWS Online. Some of the benefits of the procedure are:

- Readers would not have to re-encode models into an accessible format based on the article.
- The reviewers and authors could resolve issues relating to the correspondence between the encoded model and the model described in the article, before publication. Any differences could be eliminated.
- Modelers would be motivated to resolve correspondence issues because the publication of their article would depend on it.

Curation pipeline

The model curation process requires significant effort and this effort will in practice be shared between curators and/or teams of curators, often at different sites. The subdivision of MIRIAM into components is useful for defining the relationships between these individuals and groups. We anticipate that some groups will concentrate on encoding models that comply with the proposal for reference correspondence and the attribution scheme for annotations. Other groups will then continue the curation process by annotating these models so that they comply with the external data resources annotation scheme.

microarray data, quantitative models will be useful only if their access and reuse is made easy for all scientists. Moreover, the next step towards a more synergistic view of living systems is assembling models into larger entities, by module reuse and assembly or modeling across different spatial, temporal or physiological scales. Both model retrieval and model composition require formal descriptions of model structure and semantics. Our separate groups have been active in the development of standards for encoding biological models in machine-readable formats (e.g., CellML[3] and SBML[4,5]) and of public repositories of computational models (such as BioModels Database[6], Sigpath[7], EcoCyc[8], the CellML repository (http://www.cellml.org/examples/repository/), JWS Online[9], RegulonDB[10], DOQCS[11]). We firmly believe in the value of expressing computational models using standardized, structured formats as a means of enabling direct interpretation and manipulation of those models by software tools.

Box 3 Rules for reference correspondence

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. Relevant examples include those aimed at biological modeling (GENESIS[44], XPP[45]) or generic scientific software packages (Mathematica, MatLab, SciLab, Octave)
2. The encoded model must comply with the standard in which it is encoded. The syntax of the language must be respected, and the model has to pass validation at curation time. The form of this validation will depend on the format in which the model is encoded. For the SBML and CellML standards, formal validation software should be used; see http://sbml.org/ and http://www.cellml.org/, respectively. For application-specific formats, the model must be parsed (loaded) successfully by the relevant application.
3. The model must be clearly related to a single reference description that describes or references a set of results that one can expect to reproduce using the model. If the model is associated with only part of a reference description, then that part must be clearly identified (although failure to do so does not preclude MIRIAM compliance). If a model is derived from several initial reference descriptions, there must still be a reference description associated with the derived/combined model.
4. The encoded model structure must reflect the biological processes listed in the reference description. For instance, one should be able to map a reaction network in the encoded form to a reaction graph in the associated description. It is not essential that the constituents of the encoded model correspond one-to-one with the constituents described in the associated reference description. The software used to build

Box 3 Rules for reference correspondence (continued)

the initial model and the standard format used to encode the model may impose constraints on the form of the model. For example, a modeler might have to add reactions to represent the creation or removal of mass. A ligand in excess may be represented either as an independent constituent, or as an event modifying parameters.

5. The encoded model must be instantiated in a simulation. This means that quantitative attributes of the model have to be defined. Therefore, the model must contain, or be associated with, values (or ranges of values) for all initial conditions and parameters, as well as kinetic expressions for all reactions. These values can be provided as a separate file from the model itself. If the model was not submitted as an adjunct to the original description, then one should be able to trace all quantities in the encoded form to quantities enumerated in the reference description. The values of quantitative variables and their units must be equivalent to the values listed in the reference description. Any missing values have to be added (perhaps by contacting the authors) before the model can be claimed to be MIRIAM compliant
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. Not only does the simulation have to provide results qualitatively similar to the reference description, such as oscillation, bistability, chaos, but the quantitative values of variables, and their relationships (e.g., the shape of the phase portrait) must be reproduced within some epsilon, the difference being attributable to the algorithms used to run the simulation, and the roundup errors. Some software exists that can help to compare qualitatively the results of a simulation with a benchmark; see for instance BIOCHAM[46].

Databases of quantitative models are valuable resources only if researchers can trust the quality of their content. Similarly, repositories are not useful unless users can search for specific models and then relate model constituents to other data sets such as bioinformatics databases and controlled vocabularies. To meet these needs, we believe four complementary aspects of the quality of an encoded model must be addressed: (i) the quality of the documentation (e.g., journal article) associated with the encoded model, (ii) the degree of correspondence between the encoded model and the documentation, (iii) the accuracy and extent of the annotations of the encoded model and (iv) whether the model is encoded in a machine-readable format, that is, a format that can be immediately and unambiguously parsed by software to perform simulations and analysis.

Most of the encoded models available in scientific publications or on the Internet are not in a standard format. Of those that are encoded in a standard format, it turns out that most actually fail compliance tests developed for these standards. Failures occur for a variety of reasons, ranging from minor syntactic errors to significant conceptual problems, including the incorrect specification of units. Even deeper semantic inaccuracies can lie in the structure of the model itself. Finally, there is no standard naming scheme for the model constituents, so the precise identification of constituents depends on the associated documentation/annotation. Most models available today are not annotated, and as a result, users are faced with such things as a reaction 'X' between the constituents 'A' and 'B,' producing 'C' and modulated by 'M.' As a consequence, models frequently have to be re-encoded in order to be reused, a process that in practice is often performed by a different person from the original author.

These quality issues must be addressed when curating model collections for public use, just as it is done for other type of biological data. One crucial step is the development of interchange standards[12], such as those developed for microarray data[13], protein interactions[14] or metabolic analyses[15]. By 'curation,' we mean the processes of collecting models, verifying them to some degree and annotating them with metadata.

Box 4 Annotation that must be included with a quantitative model to achieve MIRIAM compliance

1. The preferred name of the model, in order to facilitate discussions about it.
2. A citation of the reference description with which the model is associated. This citation can be a complete bibliographic record, a unique identifier such as a Digital Object Identifier (http://www.doi.org/), a PubMed identifier (http://www.pubmed.gov/) or, in the last resort, an unambiguous URL pointing to the description itself (but not a generic URL, for instance of an archive containing the description). The main point is that the citation should provide access to the complete description of the model and should make possible the identification of the authors of the reference description. These authors should be contacted if there are concerns with the biological basis of the model (such as the presence of an interaction undocumented in the scientific literature).
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. In many cases, there will be many creators who either encoded the model from scratch, or debugged it. For instance, the semantic curators of a database would be creators. The creators should be contacted if there are problems with the structure of the model (initial conditions, kinetics parameters, reaction scheme).
4. The date and time of creation, and the date and time of last modification. This is particularly important in order to know if a model has been modified since its creation, and to compare various versions of the same model. A history of the modifications could be useful, but is not required for MIRIAM compliance. A checksum could be useful to identify a specific version of a model, but is not required for MIRIAM compliance.
5. A precise statement about the terms of distribution. The statement can be anywhere from 'public domain' to 'copyrighted' and 'freely distributable' to 'confidential.' It is important to note that MIRIAM itself does not require free distribution, whether in the sense of 'freedom of use' or 'no cost.' However, MIRIAM is intended to allow models to be communicated better, and stipulating the terms of distribution are essential for that purpose.

Table 1 Possible sources of annotation for different model constituents[a]

Constituent	Resources
Model	Digital Object Identifier, Medline, PubMed, Gene Ontology[30] (BP, MF, CC), International Classification of Disease, Online Mendelian Inheritance in Man[31] (OMIM), Taxonomy[32,33]
Physical compartment	Gene Ontology (CC), Taxonomy
Reacting entity	BIND complex, Chemical Entities of Biological Interest (ChEBI), Ensembl[29], Gene Ontology (MF, CC), InterPro[34], KEGG[35] compound, OMIM, Protein DataBank (PDB), PIRSF[36], Reactome[37], UniProt[28] ...
Reaction	BIND interaction[38], EC code, Gene Ontology (BP, MF), KEGG reaction, IntAct[39]

[a]This list is by no means exhaustive, but rather represents the diversity of available resources. BP, biological process; MF, molecular function; CC, cellular component.

We propose to standardize an approach to the curation of model collections and the encoding of models using a framework of rules we call MIRIAM, the Minimum Information Requested In the Annotation of Models. MIRIAM aims to define processes and schemes that will instill confidence in model collections, enable the assembly of meta-collections of models at the same high level of quality and allow the curation process to be shared among teams at different sites and institutions. The standard we propose is designed to cover encoding processes that may be conducted either up front by the model author or *post hoc* by a curator. However, we do not believe that the *post hoc* approach is particularly efficient, and prefer modelers to make their models available in standard formats. **Box 2** describes some uses of MIRIAM.

Scope of MIRIAM

MIRIAM applies only to models linked to a unique reference description. MIRIAM does not address directly issues of quality of documentation (although sufficiently poor documentation can make a model impossible to curate). The assessment of the quality of documentation is well established in the scientific community. We expect that, by assessing the documentation describing quantitative models, peer reviewers (not the model curators) will assess the models' ability to represent and predict the quantitative behavior of biological systems and/or make an important theoretical contribution. Instead, MIRIAM focuses on the correspondence of an encoded model to its associated description and how the encoded model is annotated. In other words, even if it is MIRIAM compliant, a model may not necessarily make sense in biological terms. Conversely, many models that cannot be declared MIRIAM compliant may still be of high scientific interest.

We expect MIRIAM to apply mainly to quantitative models that can be simulated over a range of parameter values and provide numerical results. This encompasses not only models that can be integrated or iterated forwards in time, such as ordinary and partial differential equation models and differential algebraic equation models, but also other quantitative approaches such as steady-state models (e.g., Metabolic Control Analysis[16], Flux Balance Analysis[17]). Discrete approaches, such as logical modeling[18–20] or stochastic and hybrid Petri Net[21], can also be considered when they can lead to specific numerical results. Although we are aware that this means we can cover only part of the modeling field, we make this our initial focus because only these models can lead to quantitative numerical results providing refutable predictions. The comparison of these predictions with the reference description of the model is a crucial test of MIRIAM compliance.

Overview of the proposal

MIRIAM is divided into two parts. The first is a proposed standard for reference correspondence dealing with the syntax and semantics of the model, whereas the second is a proposed annotation scheme that specifies the documentation of the model by external knowledge.

Standard for reference correspondence

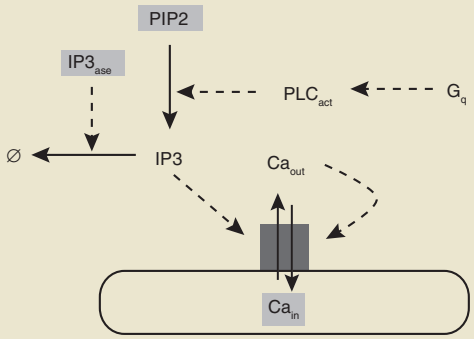
The aim of this proposal is to ensure that the model is properly associated with a reference description and is consistent with that reference description. To be declared MIRIAM compliant, a quantitative model must fulfill a set of rules dealing with its encoding, its structure and the results it should provide when instantiated in simulations. These rules are detailed in **Box 3**.

Table 2 Examples of different physical locations related to the same URIs expressed as a URL or a LSID

URI	Example of alternative physical locations
Taxonomy	
http://www.ncbi.nlm.nih.gov/Taxonomy/#9606	http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9606 (ref. 32)
urn:lsid:ncbi.nlm.nih.gov:Taxonomy:9606	http://www.ebi.ac.uk/newt/display?search=9606 (ref. 33)
Gene Ontology	
http://www.geneontology.org/#GO:0045202	http://www.ebi.ac.uk/ego/DisplayGoTerm?id=GO:0045202
urn:lsid:geneontology.org:GO:0045202	http://www.godatabase.org/cgi-bin/amigo/go.cgi?view=details&query=GO:0045202
UniProt	
http://www.uniprot.org/#P62158	http://www.ebi.uniprot.org/entry/P62158 (ref. 28)
urn:lsid:uniprot.org:P62158	http://us.expasy.org/uniprot/P62158 (ref. 40)
	http://www.pir.uniprot.org/cgi-bin/upEntry?id=P62158 (ref. 41)
EC code	
http://www.ebi.ac.uk/intenz/EC 1.1.1.1	http://www.ebi.ac.uk/intenz/query?cmd=SearchEC&ec=1.1.1.1 (ref. 42)
urn:lsid:ebi.ac.uk:intenz:EC 1.1.1.1	http://www.genome.jp/dbget-bin/www_bget?ec:1.1.1.1 (ref. 35)
	http://www.chem.qmul.ac.uk/iubmb/enzyme/EC1/1/1/1.html
	http://us.expasy.org/cgi-bin/nicezyme.pl?1.1.1.1 (ref. 43)

Table 3 Example of a small curated and annotated model

Creators	Joe User (juser@eden.com), Anne Other (aother@eden.com)
Creation date	01 January 2000
Last modification	31 May 2005



Constituent	Data type	Identifier	Qualifier	Meaning
Model	http://www.pubmed.gov/	0000000		
	http://www.ncbi.nlm.nih.gov/Taxonomy/	9606		*Homo sapiens*
	http://www.geneontology.org/	GO:0007204	IsVersionOf	Positive regulation of cytosolic [Ca^{2+}]
	http://www.geneontology.org/	GO:0051279	IsVersionOf	Regulation of release of sequestered Ca^{2+} into cytoplasm
	http://www.genome.jp/kegg/pathway/	hsa04020	IsPartOf	Calcium signaling pathway, *H. sapiens*
	http://www.genome.jp/kegg/pathway/	hsa04070	IsPartOf	Phosphatidylinositol signaling system, *H sapiens*
Compartment ER	http://www.geneontology.org/	GO:0005790		Smooth endoplasmic reticulum
Reactant Ca_{in}	http://www.ebi.ac.uk/chebi/	CHEBI:29108		$Calcium^{2+}$
Cytoplasm	http://www.geneontology.org/	GO:0005737		Cytoplasm
Reactant Ca_{out}	http://www.ebi.ac.uk/chebi/	CHEBI:29108		$Calcium^{2+}$
Reactant IP3	http://www.ebi.ac.uk/chebi/	CHEBI:16595		1D-myo-inositol 1,4,5-tris (dihydrogen phosphate)
Reactant PIP2	http://www.ebi.ac.uk/chebi/	CHEBI:18348		1-phosphatidyl-1D -myo-inositol 4,5-bisphosphate
Reactant IP3R	http://www.uniprot.org/	Q14643	HasVersion	Inositol 1,4,5-trisphosphate receptor type 1
	http://www.uniprot.org/	Q14571	HasVersion	Inositol 1,4,5-trisphosphate receptor type 2
	http://www.uniprot.org/	Q14573	HasVersion	Inositol 1,4,5-trisphosphate receptor type 3
Reactant PLC_{act}	http://www.uniprot.org/	Q9NQ66	IsVersionOf	PIP2 phosphodiesterase β-1
Reactant PLC_{tot}	http://www.uniprot.org/	Q9NQ66		PIP2 phosphodiesterase β-1
Reactant $IP3_{ase}$	http://www.uniprot.org/	Q14642		Type I inositol-1,4,5-trisphosphate 5-phosphatase
Reactant G_q	http://www.uniprot.org/	Q6NT27		Guanine nucleotide binding protein Gq
Reaction $Ca_{release}$	http://www.geneontology.org/	GO:0005220		IP3-sensitive calcium-release channel activity
	http://www.geneontology.org/	GO:0008095	IsVersionOf	IP3 receptor activity
Reaction $IP3_{production}$	http://www.geneontology.org/	GO:0004435	IsVersionOf	Phosphoinositide phospholipase C activity
	http://www.ebi.ac.uk/intenz/	3.1.4.11	IsVersionOf	Phosphoinositide phospholipase C
Reaction $IP3_{degradation}$	http://www.ebi.ac.uk/intenz/	3.1.3.56	IsVersionOf	Inositol-polyphosphate 5-phosphatase
Reaction $PLC_{activation}$	http://www.geneontology.org/	GO:0007200		G-protein signaling coupled to IP3 second messenger

$$k_1 = k_2 = k_3 = 1\ s^{-1}$$

$$Km = 10^{-7}\,M, Km_2 = 10^{-8}\,\mathrm{M}, Km_3 = 2.10^{-6}\,M$$

$$K_A = 10^{-11}, m = 4, n = 3, \alpha = 0.001$$

$$[Ca_{in}] = [IP3R] = [PLC_{tot}] = [PIP2] = [IP3_{ase}] = 0.001\,M$$

$$[G_q] = 0.01\,M,\ [Ca_{out}] = [IP3] = [PLC_{act}] = 0\,M$$

$$\frac{d[Ca_{out}]}{dt} = \frac{k_1[IP3R] * ([Ca_{in}] - [Ca_{out}])}{Km_1 + |[Ca_{in}] - [Ca_{out}]|} * \frac{[IP3]^m}{K_A + [IP3]^m}$$

$$\frac{d[IP3]}{dt} = \frac{k_2[PLC_{act}] * [PIP2]}{Km_2 + [PIP2]} - \frac{k_3[IP3_{ase}] * [IP3]}{Km_3 + [IP3]}$$

$$\frac{d[PLC_{act}]}{dt} = \frac{[G_q]^n}{\alpha + [G_q]^n} * [PLC_{tot}]$$

The model describes the release of calcium from the endoplasmic reticulum, regulated by cytoplasmic calcium and the Inositol 1,4,5-trisphosphate produced in response to G-protein–coupled receptor activation. Note that although working, this model is only meant to provide a large number of example annotations.

To pass the various tests, and in particular the reproduction of described results, a modeler could be required to make minor changes to a model until it is truly consistent with the results given in the associated reference description. If the modeler is not one of the authors, ideally he/she should perform these modifications in collaboration with the authors. Examples include changing a few parameter and/or initial condition values.

When the model given in the text of the reference description is significantly different from the encoded model used to generate the results given in this text, the model cannot be curated and MIRIAM cannot be applied. For example, MIRIAM cannot be applied if a significant number of parameter values are different between the two models (the significance being judged by the curators). The original authors of the model should be encouraged to publish an erratum detailing the correct values.

Annotation schemes

The scheme for annotation is composed of two complementary components: attribution, covering the absolute minimum information that is required to associate the model with both a reference description and an encoding process, and external data resources, covering information required to relate the constituents of quantitative models to established data resources or controlled vocabularies.

The annotations must always be transferred with the encoded model. The ideal case is incorporating these annotations in the same file as the model itself, in a structured form such as the CellML metadata[22] or the SBML simple annotation scheme[23]. However, annotations could also be joined in another form, such as one or several accompanying files, in various formats, textual or graphical.

Attribution annotation

To be confident in being able to reuse an encoded model, one must be able to trace its origin and the people who were involved in its creation. In particular, the reference description has to be identified, as well as the authors and creators of the model. The information that must always be joined with an encoded model is listed in **Box 4**.

External data resources annotation

The aim of this 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 execution of queries on models to recover specific constituents in models. Possible sources of annotation for various types of constituents are listed in **Table 1**.

This annotation must permit a piece of knowledge to be unambiguously related to a model constituent. The structure of an atomic element of the annotation is similar to the relationshipXref element of BioPAX (http://www.biopax.org/). 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. 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." The qualifier is optional, and its absence does not preclude MIRIAM compliance. When a qualifier is absent, one assumes the relation to be "is."

The "data-type" should be written as a Unique Resource Identifier[24]. This URI can be a Uniform Resource Locator[25] or a Uniform Resource Name[26]. The URL or URN does not have to describe an actual physical location. It is up to the software tool reading the model to decide what to do with this URI. This software can, for instance, use the "identifier" with a search engine built on a database mirroring the "data-type." Alternatively, a reading tool translating the model can build a hyperlink using the "identifier" and another URL related to the "data-type."

The "data-type" and the "identifier" can be combined into a single URL, such as http://www.myResource.org/#myIdentifier or as a URN, for instance using the LSID scheme[27] of urn:lsid:myResource.org:myIdentifier.

To enable interoperability, the community will have to agree on a set of standard, valid URIs. An online resource will be established to catalog the URIs and the corresponding physical URLs of the agreed-upon "data-types," whether these are controlled vocabularies or databases. This catalog will simply list the URIs and for each one, provide a corresponding summary of the syntax for the "identifier." An application programming interface (API) can be created so that software tools can retrieve valid URL(s) corresponding to a given URI. **Table 2** shows a small subset of this forthcoming list. Note that although MIRIAM compliance does not require such a list to exist, it is considered crucial to actually enforce MIRIAM usage, and to make it truly useful. The list will also have to evolve with the data resources.

It is important that model constituents be annotated with perennial identifiers. For example, the "entry name" field of UniProt[28] is not perennial but is modified on a regular basis to reflect the classification of the protein. However, the "accession" field of UniProt is perennial. Consider a model with an entity representing the protein calmodulin. An annotation of this entity referring to the UniProt record for calmodulin should therefore use a URI containing the "accession" field value for calmodulin "P62158" rather than the "entry name" field value "CALM_HUMAN."

Quite often, several identified biological entities, physical components or reactions are lumped in a single constituent of the model. For instance, successive reactions of a pathway may be merged into one reaction, or a set of different molecules is represented by one pool. The annotation must reflect this situation, either by enumerating the biological entities, or with a carefully chosen term from a controlled vocabulary (an example of a curated and annotated model is presented in **Table 3**).

Conclusions

We believe that through the standardization of the model curation process, it will be possible to create resources that are as significant to systems biology as resources like Ensembl[29] are to genomics. Pursuing this proposal will in the short term allow us to establish collections of models of sufficient quality to gain the confidence of the systems biology community. To pave the way, the resources handled by the authors of this manuscript (BioModels Database, CellML repository, DOQCS, SigPath) endorse the standard, and will undertake efforts to make them MIRIAM compliant. In the longer term, the application of MIRIAM will enable the peer review process to become more efficient and its products more accessible. We also hope the standard will be adopted by publishers of scientific literature, as was the case with other standards such as MIAME[13].

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Kitano, H. Computational systems biology. *Nature* **420**, 206–210 (2002).
2. Crampin, E. *et al.* Computational physiology and the physiome project. *Exp. Physiol.* **89**, 1–26 (2004).

3. Lloyd, C., Halstead, M. & Nielsen, P. CellML: its future, present and past. *Prog. Biophys. Mol. Biol.* **85**, 433–450 (2004).
4. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
5. Finney, A. & Hucka, M. Systems biology markup language: level 2 and beyond. *Biochem. Soc. Trans.* **31**, 1472–1473 (2003).
6. Le Novère, N., *et al.* BioModels Database: A free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.* **34**, (2006).
7. Campagne, F. *et al.* Quantitative information management for the biochemical computation of cellular networks. *Sci. STKE* **248**, PL11 (2004).
8. Keseler, I. *et al.* EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.* **33**, D334–D337 (2005).
9. Olivier, B. & Snoep, J. Web-based kinetic modelling using JWS Online. *Bioinformatics* **20**, 2143–2144 (2004).
10. Salgado, H. *et al.* RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* **32**, D303–D306 (2004).
11. Sivakumaran, S., Hariharaputran, S., Mishra, J. & Bhalla, U. The database of quantitative cellular signaling: management and analysis of chemical kinetic models of signaling networks. *Bioinformatics* **19**, 408–415 (2003).
12. Quackenbush, J. Data standards for 'omic' science. *Nat. Biotechnol.* **22**, 613–614 (2004).
13. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
14. Hermjakob, H. *et al.* The HUPO PSI's molecular interaction format-a community standard for the representation of protein interaction data. *Nat. Biotechnol.* **22**, 177–183 (2004).
15. Lindon, J. *et al.* Summary recommendations for standardization and reporting of metabolic analyses. *Nat. Biotechnol.* **23**, 833–838 (2005).
16. Kacser, H. & Burns, J. The control of flux. *Symp. Soc. Exp. Biol.* **27**, 65–104 (1973).
17. Savinell, J. & Palsson, B. Optimal selection of metabolic fluxes for *in vivo* measurement. I. Development of mathematical methods. *J. Theor. Biol.* **155**, 201–214 (1992).
18. Thomas, R. Boolean formalisation of genetic control circuits. *J. Theor. Biol.* **42**, 565–583 (1973).
19. Sánchez, L. & Thieffry, D. Segmenting the fly embryo: a logical analysis of the pair-rule cross-regulatory module. *J. Theor. Biol.* **224**, 517–537 (2003).
20. Laubenbacher, R. & Stigler, B. A computational algebra approach to the reverse engineering of gene regulatory networks. *J. Theor. Biol.* **229**, 523–537 (2004).
21. Doi, A., Fujita, S., Matsuno, H., Nagasaki, M. & Miyano, S. Constructing biological pathway models with hybrid functional petri nets. *In Silico Biol.* **4**, 271–291 (2003).
22. Cuellar, A., Nelson, M. & Hedley, W. The CellML metadata 1.0 specification. http://www.cellml.org/specifications/metadata/.
23. Le Novère, N. & Finney, A. A simple scheme for annotating SBML with references to controlled vocabularies and database entries. http://www.ebi.ac.uk/compneur-srv/sbml/proposals/AnnotationURI.pdf.
24. Berners-Lee, T., Fielding, R. & Masinter, L. Uniform resource identifier (URI): Generic syntax. http://www.gbiv.com/protocols/uri/rfc/rfc3986.html.
25. Berners-Lee, T. Uniform resource locators (URL): a syntax for the expression of access information of objects on the network. http://www.w3.org/Addressing/URL/url-spec.txt.
26. Moats, R. URN syntax. http://www.ietf.org/rfc/rfc2141.txt.
27. Martin, S., Niemi, M. & Senger, M. Life sciences identifiers RFP response. http://www.omg.org/technology/documents/formal/life_sciences.htm
28. Apweiler, R. *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **32**, D115–D119 (2004).
29. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
30. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
31. Hamosh, A., Scott, A., Amberger, J., Bocchini, C. & McKusick, V. Online mendelian inheritance in man ([OMIM]), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
32. Wheeler, D. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **28**, 10–14 (2000).
33. Phan, I., Pilbout, S., Fleischmann, W. & Bairoch, A. NEWT, a new taxonomy portal. *Nucleic Acids Res.* **31**, 3822–3823 (2003).
34. Mulder, N.J. *et al.* InterPro, progress and status in 2005. *Nucleic Acids Res.* **33**, 201–205 (2005).
35. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
36. Wu, C. *et al.* PIRSF: family classification system at the protein information resource. *Nucleic Acids Res.* **32**, D112–D114 (2004).
37. Joshi-Tope, G. *et al.* The genome knowledgebase: A resource for biologists and bioinformaticists. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 237–243 (2003).
38. Bader, G. & Hogue, C. BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16**, 465–477 (2000).
39. Hermjakob, H. *et al.* IntAct—an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–D455 (2004).
40. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
41. Wu, C. *et al.* Update on genome completion and annotations: protein information resource. *Nucleic Acids Res.* **31**, 345–347 (2003).
42. Fleischmann, A. *et al.* IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.* **32**, D434–D437 (2004).
43. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304–305 (2000).
44. Bower, J. & Beeman, D. *The Book of GENESIS* (Springer-Verlag, New York, 1998).
45. Ermentrout, B. *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students* (Society for Industrial & Applied Math, Philadelphia, PA, 2002).
46. Chabrier, N. & Fages, F. Symbolic model checking of biochemical networks. in *International Workshop on Computational Methods in Systems Biology* (Springer-Verlag, New York, 2003).