Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

Validity and Combination of Biochemical Models

Wolfram Liebermeister

Computational Systems Biology, Max-Planck-Institut für molekulare Genetik, Ihnestraße 63 – 73, 14195 Berlin, Germany

E-Mail: lieberme@molgen.mpg.de

Received: 14th April 2008 / Published: 20th August 2008

Abstract

The merging of mathematical models (either manually or assisted by computer programs) is an important requisite for creating large mathematical models of cells. A kinetic model describes biochemical quantities such as concentrations and reaction rates by explicit differential and algebraic equations. We can regard it as a list of model statements, each comprising a biochemical quantity (e. g. a substance concentration), the corresponding mathematical object (e. g. a variable or parameter), and a mathematical equation that makes it possible to compute its numerical value. When two such models are merged, typical conflicts have to be detected and resolved: (i) incompatible names or identifiers; (ii) incompatible physical units; (iii) duplicate elements with contradicting assignments; (iv) conflicting (“semantically dependent”) quantities; (v) cyclic dependencies between model equations. To define and judge whether merging algorithms are trustworthy, we need formal criteria for the validity of models; such criteria can be classified into the categories “syntax”, “computation”, “biochemical semantics”, “physical laws and empirical knowledge” and “model relevance”.

Merging of Biochemical Models

Living cells can be described by mathematical models in order to test biological hypotheses by computer simulations and mathematical analysis. The mathematical elements in biochemical models (e. g. variables and equation terms) refer to chemical substances and processes such as transport, binding and reactions. In publications, models are described verbally and by mathematical formulae, but researchers also publish them in computer-readable formats like SBML [1] (systems biology markup language) and models become increasingly available in databases [2, 3]. One intention behind this is that models should be reusable; model reuse is further facilitated if standards [4] are respected – as put forward in the MIRIAM proposal [5] (''minimal information requested in the annotation of biochemical models'').

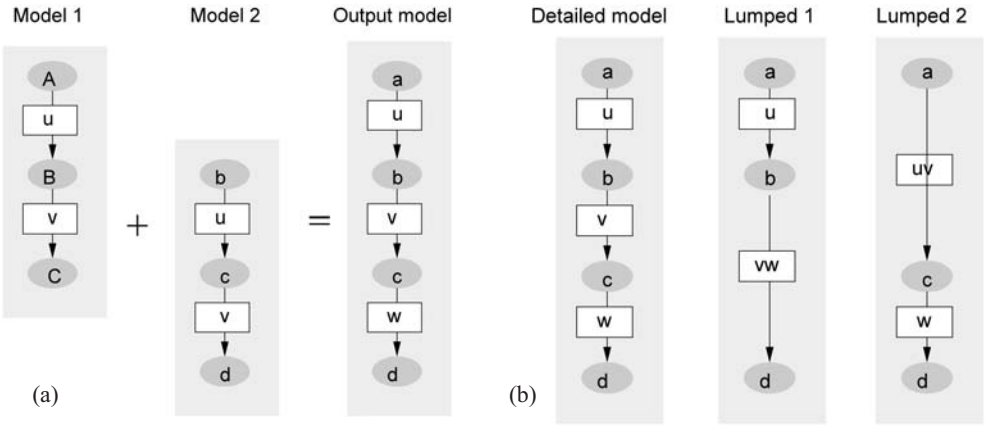

Figure 1. Merging of structural models. (a) Structural models of two metabolic pathways are added by taking the set union of all model elements (metabolites and reactions shown as ellipses and boxes, respectively). The graph topology is coded by stoichiometric coeffcients, contained as additional information in the reaction elements (not shown). The names of model elements can differ from model to model, so elements must be compared by their annotations (not shown) and a consistent set of names has to be chosen for the output model. (b) Models can contain lumped reactions (e. g. the lumped reaction VW in model ''Lumped 1'' contains reactions V and W from the detailed model). If lumped reactions overlap partially (like VW and UV in the models ''Lumped 1'' and ''Lumped 2''), they do not fit to each other and the models cannot be directly combined. Similar conflicts would occur with lumped metabolites (not shown).

With a number of models already available, large dynamic models may be built by combining existing models of biochemical reactions [6] or cellular pathways [7]. Model merging can be straightforward if the input models originate from the same modelling framework, share the same naming conventions, and are based on a common set of non-conflicting biochemical quantities. In general, however, models will originate from different sources, so conflicts may easily occur. Model merging could be facilitated by computer programs that

execute uncritical steps and perform validity checks, but such tools and the theory to support them are still in their infancy. Model combination-whether manually or assisted by computer programs – requires that models are appropriately prepared: experiments and model formats must be standardized [5, 4], and all model elements need to have a clear biochemical meaning.

In publications, the elements are usually described in words (e. g. ''cellular concentration of ATP in mM''), while in computer-readable formats like SBML, they can be annotated with references to public databases (e. g. ATP may be represented by the identifier `C00002` in the KEGG database [8]). A chemical reaction can be annotated by an identifier or by specifying its substrates and products.

Merging of model structures

The aim in biochemical model merging is to combine several models describing reactions or biochemical pathways in order to obtain a valid model of the combined system. Before considering dynamic models, let us first have a look at simple structural models as shown in Fig. 1a. A structural model consists of a list of elements representing biochemical entities (e. g. metabolites and chemical reactions specified by annotations); in different models, the same entity can bear different names. Model elements may be linked to further information (e. g. pictures, comments, or mathematical expressions). Figure 1a shows how two overlapping pathway structures are combined: the resulting pathway contains all elements of the original models and pairs of duplicate elements (B = b, v = u and C = c) are merged into single elements, respectively. Merging of structural models involves the following steps:

1. The model elements have to be compared-either by a human expert or automatically-to detect duplicates. For automatic comparison, model elements have to bear annotations (i. e. standardized substance names or links to biological databases) that unambiguously determine their biochemical meaning. A simple string comparison between element names would not suffice because models may follow different naming conventions.
2. If duplicate elements are found, their accompanying information needs to be merged; if the two elements contain contradicting information (e. g. two models assign different concentrations to the same metabolite), some of the information has to be discarded.
3. Severe conflicts can arise if an element in one model (e. g. the lumped reaction VW in Fig. 1b) corresponds to several elements in another model (reactions V and W), or if several elements partially overlap in their meaning (e. g. the lumped reactions UV and VW). Such overlaps are a notorious source of conflict: in particular, if elements in a model are linked to mathematical expressions (e. g. chemical reactions are described by kinetic rate laws), the expressions for over-

lapping entities will probably not fit to each other. Therefore, overlapping elements generally should be avoided in model merging.

In this article, I shall discuss some basic theoretical concepts behind model merging. Merging of model structures will be our starting point: the same scheme also applies, *mutatis mutandis*, to dynamical models from different mathematical frameworks as long as all mathematical equations are given in the form of explicit assignments. Firstly, I shall focus on kinetic models as a special case and discuss the following questions: what are the basic elements in such models, in analogy to the structural elements shown in Fig. 1? How can we compare the biochemical meaning of elements and how can we detect conflict between them? Which additional problems can occur in dynamical models? Secondly, a general merging algorithm for explicit biochemical models is presented; it is applicable both in manual and automatic model merging. Finally, I shall classify some general validity criteria for biochemical models and discuss how models should be prepared to allow for model merging and other kinds of model reuse.

Mathematical Models and their Biochemical Semantics

Explicit biochemical models

Mathematical models allow the simulation of the dynamics of biochemical processes; depending on the system studied and on the questions to be answered, various mathematical frameworks can be used, including kinetic models, reaction–diffusion models, particle-based stochastic models, or constraint-based flux models. Despite their different forms, all such models describe a number of mathematical elements (variables, parameters...,) that are associated with biochemical objects (e. g. molecules) or quantities (e. g. concentrations). In addition, they contain mathematical statements supposed to hold for these quantities (e. g. ordinary differential equations, equality constraints, maximal postulates). The list of statements may either be an *ad hoc* collection (e. g. a number of the constraints used for flux balance analysis) or a complete description that allows for predictive simulations (e. g. a system of rate equations); in the latter case, mathematical solutions of the model should correspond, approximately, to the possible behaviour of the biological system under study.

Kinetic models

As a well-known example, we shall consider kinetic models comprising independent substance concentrations $c_i(t)$, dependent substance concentrations $c_j^{\mathrm{dep}}(t)$, external substance concentrations c_l^{ext} , reaction velocities $v_k(t)$, and kinetic constants p_m. The values are determined by explicit equations:

$$
\begin{aligned}
& p = (p_1, p_2, \ldots)^{\mathrm{T}} && \text{(constant numbers)} \\
& c_{\mathrm{ext}} = (c_1^{\mathrm{ext}}, c_2^{\mathrm{ext}}, \ldots)^{\mathrm{T}} && \text{(constant numbers)} \\
& c(0) = (c_1(0), c_2(0), \ldots)^{\mathrm{T}} && \text{(constant numbers)} \\
& c_j^{\mathrm{dep}}(t) = g_j(c(t)) \\
& \nu k(t) = f_k(c(t), c^{\mathrm{dep}}(t), c^{\mathrm{ext}}, p) \\
& \frac{dc_i}{dt} = \sum_l N_{il} v_l(t)
\end{aligned}
\qquad (1)
$$

for all values of k, i, and j, where N is the stoichiometric matrix, the functions f_k denote kinetic rate laws, and the functions g_j relate the dependent concentrations to the independent concentrations. The variables and parameters represent biochemical quantities and are described by explicit algebraic or differential equations; models with these two properties (e. g. kinetic models, reaction–diffusion models, but also certain stochastic models) will be called "explicit biochemical models".

Computational cycles

For explicit biochemical models, computations become much more simple if all equations can be evaluated one after the other. To illustrate this point, let us consider an equation system with parameters p_1, p_2,..., differential equations of the form $\mathrm{d}x_i/\mathrm{d}t = f_i(x,y,p)$, and algebraic equations of the form $y_k(t) = g_k(x,y,p)$. If the mathematical formula for $g_k(x,y,p)$, corresponding to variable y_k, contains another variable y_l, then y_k is said to be computed from y_l. If each variable y_k in the model is only computed from variables y_l with smaller $l < k$, then the model is sequentially computable: the equations can be directly evaluated one after the other in each integration step. To check whether a model is sequentially computable, we can build a graph with nodes corresponding to the variables y_k and directed edges representing the "computed-from" relation. Cycles in the graph are called "computational cycles". If this graph is acyclic, the variables can be ordered such that the model is sequentially computable. If a model contains computational cycles (e. g. $y_1 = g_1(y_2)$, $y_2 = g_2(f_1)$), computations can become difficult and solutions may be non-unique or they may not even exist. Cycles between the differential equations, on the other hand, will not cause such problems.

Models as statement lists

In the following, we shall only consider explicit biochemical models consisting of algebraic equations $x = f(...)$ or differential equations $\mathrm{d}x/\mathrm{d}t = f(...)$. In such models, each quantity (or its time derivative) can be directly computed if the values of other quantities are known. Other

kinds of mathematical statements, such as inequalities $x < f(...)$, maximal requirements $x \stackrel{!}{=} \arg\max_y f(y, ...)$, or probabilistic assignments will not be considered here. Formally, an explicit biochemical model can be regarded as a list of model elements (also called "model statements"), each consisting of a biochemical quantity, a mathematical object and a mathematical assignment:

- The biochemical quantity (e. g. a concentration, reaction rate, compartment volume, or kinetic constant) is defined by a type (e. g. `concentration`), a unit (e. g. mM), a biochemical entity (e. g. a certain metabolite), and possibly, a location (e. g. a certain cell compartment). A quantity can also be related to several entities (e. g. a Michaelis constant refers to both an enzyme and a substrate metabolite).
- The corresponding mathematical object (e. g. a variable or a parameter) has a name or a unique identifier and a certain type (e. g. non-negative real number, time-dependent function $c(t)$, field $c(x,t)$).

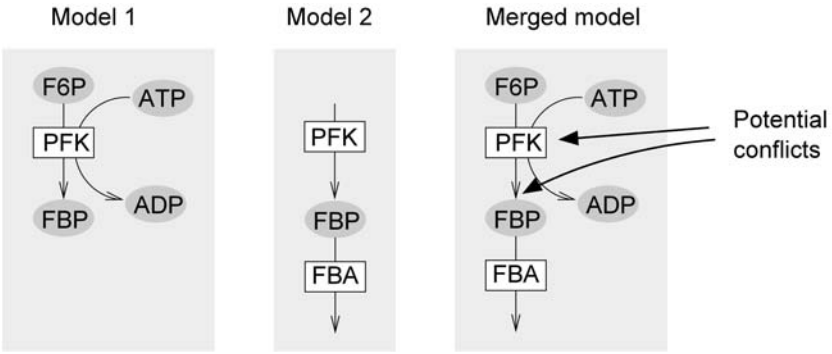

Figure 2. Merging of two small example models (equations see Table 1). Model 1 describes the PFK reaction rate at given substrate and product levels. Model 2 describes the mass balance of FBP resulting from production and degradation. The two models make different statements about the quantities representing PFK and FBP (thick arrows), so concatenating the models leads to conflicts. Abbreviations: ATP: adenosine triphosphate; ADP: adenosine diphosphate; F6P: fructose-6-phosphate, FBP: fructose-1,6-bisphosphate, PFK: phosphofructokinase, FBA: fructose-bisphosphate aldolase.

Table 1. The models from Fig. 2 are shown as statement lists. Each row represents a model statement (i. e. a model element). The numerical values in the example have been chosen arbitrarily. Simple concatenation of statements would lead to conflict because of duplicate biochemical quantities (marked by stars). When merging the models, one of the statements for PFK (and one for FBP) has to be chosen.

Model 1

Quantitiy	Math. Object	Assignment	Conflict
ATP concentration [mM]	c_{ATP}	$c_{ATP} = 1$	
ADP concentration [mM]	c_{ADP}	$c_{ADP} = 0.2$	
F6P concentration [mM]	c_{F6P}	$c_{F6P} = 0.5$	
FBP concentration [mM]	c_{FBP}	$c_{FBP} = 0.5$	*
PFK velocity [mM/s]	v_{PFK}	$v_{PFK} = f_{PFK}\ (c_{ATP}, c_{ADP}, c_{F6P}, c_{FBP})$	*

Model 2

Quantitiy	Math. Object	Assignment	Conflict
PFK velocity [mM/s]	v_{PFK}	$v_{PFK} = 0.1$	*
FBA velocity [mM/s]	v_{FBA}	$v_{FBA} = f_{FBA}\ (c_{FBP})$	
FBP concentration [mM]	c_{FBP}	$dc_{FBP}/dt = v_{PFK} - v_{FBA}$, $c_{FBP}(0) = 0.2$	*

- The numerical values of the quantity are determined by mathematical assignments. In our terminology, all mathematical assignments for a quantity are regarded as a combined assignment. A differential equation, for instance, needs to be accompanied by an initial condition; both equations form a combined assignment and appear in the same model element.

Kinetic models like Equation (1) can be written in the form of statement lists: two small example models are shown graphically in Fig. 2 and as statement lists in Table 1.

Relations between biochemical quantities

In general, two biochemical quantities can be (i) identical, (ii) equivalent, i. e. identical up to conversion (e. g. concentration versus amount, same quantity measured in different units), (iii) semantically dependent (e. g. ATP concentration in cytoplasm and entire cell; concentration of glucose and hexoses; lumped metabolic pathway or single reaction in this pathway), or (iv) semantically independent (e. g. a concentration and a reaction velocity, concentrations of two unrelated substances).

Table 2. Possible relations between biochemical quantities. Quantities are specified by four characteristics: type, unit, entity and location. The rows describe conditions for the four possible relations. Several rows for the same relation denote alternative possibilities; bars (-) denote arbitrary entries.

Status	Type	Unit	Entity	Location
(i) Identical	Identical	Identical	Identical	Identical
(ii) Convertible	Identical or related	Different	Identical	Identical
(iii) Dependent	Identical or related	-	Identical or overlapping	Overlapping
	Identical or related	-	Overlapping	Identical or overlapping
(iv) Semantically Independent	Unrelated	-	-	-
		-	No overlap	-
		-	-	No overlap

Conflict within or between models can arise if the same biochemical quantity appears twice or if different quantities are semantically dependent. Two quantities are semantically dependent if their mere definition implies mathematical constraints or dependencies between their numerical values. For instance, the ATP amount in the cell and the ATP amount in the mitochondria are semantically dependent because the ATP amount in the mitochondria can never be larger than the total ATP amount. Another example is the velocities of the lumped reactions in Fig. 1b, which must have identical values. Besides semantic dependence, there may be other dependencies due to empirical laws (e. g. thermodynamic dependencies between rate constants or physiologically required concentration ratios).

To compare two given biochemical quantities, we need to describe them in a formal way: as previously stated, a biochemical quantity is specified by its type, a unit, one or more biochemical entities, and possibly, a location. Different quantity types are related if they refer to the same information: for instance, `concentration` and `amount` may be linked via the definition `concentration = amount/volume`; in different models, a compartment size may be described by a length, an area, or a volume, so these three types are related. Quantities referring to different kinds of objects (e. g. concentrations and reaction velocities) are unrelated.

Entities and locations have to be specified by annotations (e. g. a link to a database entry representing a biochemical substance). The biochemical entities and locations can be seen as notions: the entity `glucose`, for instance, comprises all glucose molecules in a system under consideration. If two different entities (e. g. `hexose` and `glucose`) share, by definition, common instances (in this case, specific glucose molecules), they are called *overlapping*. If instances of one entity (e. g. `ATP`) necessarily contain instances of another entity (e. g. `phosphate group`) as physical parts, the entities are also overlapping. A similar criterion holds for locations: locations are overlapping if they include common spatial regions (e. g. `cell` and `mitochondria`). Based on these relationships, one can compare different biochemical quantities by comparing their four characteristics, as shown in Table 2.

Meaning and reference of models

To merge models in a plausible way, we have to consider their biochemical interpretation. To do so, we have to link mathematical objects to biochemical quantities. But moreover, we also need to specify what are the basic statements that constitute a model. To illustrate this, let us consider two models containing the assignments $a = f(b)$, $b = g(c)$ (in model 1) and $a = h(c)$, $b = g(c)$ (in model 2, where the function h is defined by $h(x) := f(g(x))$ for all x). The mathematical relationship between c and a is identical in both models, so the models are mathematically equivalent; semantically, however, they make different statements (a depends on b in model 1, while it depends on c in model b).

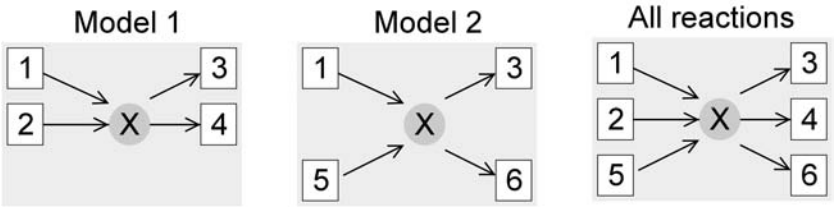

Figure 3. What is a basic model statement? According to model 1, a metabolite X participates in reactions 1, 2, 3 and 4. In model 2, it participates in reactions 1, 3, 5 and 6. In a merged model, one could either accept one of the two rate equations (regarding a rate equation as a basic statement), or one could assume a new rate equation comprising all six reactions (regarding each single term as a basic statement).

Following Frege's distinction between sense ("Sinn") and reference ("Bedeutung") [9] used in theological analysis of phrases, one may say that models 1 and 2 have the same reference (the same overall relation between numerical values), but a different sense (i. e. presumed direct relations between quantities). This difference does not play a role as long as the models are considered in their original form; however, it becomes apparent if the equation for b is changed during model merging: in this case, the mathematical behaviour in model 1 will change, while in model 2, it will not be affected.

For another example, let us consider two models containing contradictory statements for the same metabolite concentration,

$$\text{Model 1: } \mathrm{d}c/\mathrm{d}t = \nu_1 + \nu_2 - \nu_3 - \nu_4 \tag{2}$$

$$\text{Model 2: } \mathrm{d}c/\mathrm{d}t = \nu_1 + \nu_5 - \nu_3 - \nu_6. \tag{3}$$

The variable names are assumed to be non-conflicting and ν_1, ν_2, ν_3, ν_4, ν_5 and ν_6 denote the rates of different reactions (see Fig. 3). When merging the two models, we could accept one of the two rate equations (2) or (3) as our assignment for $c(t)$. This would imply that we

regard an entire rate equation as a basic model statement, which makes sense if we fit a model globally to concentration time series. Alternatively, we could merge the two rate equations and use:

$$dc/dt = \nu_1 + \nu_2 + \nu_5 - \nu_3 - \nu_4 - \nu_6. \quad (4)$$

With this choice, we assume implicitly that each of the terms on the right-hand side represents a basic statement, the fact that the metabolite is involved in a certain reaction. This point of view makes sense if models are built by combining individual reactions, possibly measured *in vitro*. It is also the rationale behind the structure of SBML.

Model Merging

Conflicts between statements

A naive way to merge explicit biochemical models would be to concatenate their statement lists (if necessary, after adjusting the variable names); the concatenated model would cover all quantities and statements from both input models. If all statements in the input models are true, then the concatenated model will be true as well, because correctness of a basic statement does not depend on the other statements around it. On the other hand, if models describe completely unrelated quantities, merging them should not create any conflict either. But if the two models contain identical, equivalent or semantically dependent quantities, the concatenated model may contain contradictions – especially if the original models are inaccurate or fitted to different experimental situations. Typical possible conflicts are as follows:

1. *The concatenated model contains different statements for the same quantity.* For instance, the two models in Fig. 2 make different statements about the PFK reaction rate: in model 1, the value depends on other quantities, while in model 2, the value is fixed. The concatenated statement list would be logically inconsistent because only one of the statements can be correct. Accordingly, the combined model would have no mathematical solution (except for rare cases in which both statements yield the same numerical value). Thus for each duplicate pair, one of the two statements has to be omitted. The combined resulting model will still be complete (no variable is missing), but it may contain computational cycles.
2. *The concatenated model contains semantically dependent quantities*. A model with semantically dependent quantities may be valid, but the corresponding mathematical assignments need to be fine-tuned to satisfy certain restrictions. If two semantically dependent quantities originate from different models, these restrictions will not be stated in either of the models, and it is likely that they will be violated after merging. At the same time, none of the quantities can be omitted

because both may be needed to compute other quantities, so the conflict cannot be resolved. Thus, models with semantically dependent elements should not be directly merged.

3. *The combined model may violate a physical law.* An example is the Wegscheider condition (see, e. g. [10]), which constrains the kinetic parameters along a circle in a metabolic network. If the merging of two models leads to a new circle and if the models were not especially prepared, the newly arising Wegscheider condition will probably not be satisfied. In the case of Wegscheider conditions, safe merging could be ensured by an appropriate parametrization of the kinetic rate laws [11, 12, 10].

A simple merging algorithm

In principle, merging of acyclic explicit biochemical models resembles the merging of structural models shown in Fig. 1. As before, we need to match elements and find identical and conflicting pairs: but now the elements (boxes and ellipses) represent model statements, they are compared according to their biochemical quantities, and the mathematical assignments are treated as additional information.

In the present context, explicit biochemical models are called valid if they satisfy the following criteria: (V1) correct syntax including consistent use of variable names; (V2) all elements are properly annotated; (V3) for each quantity, there is an assignment that allows it (or its time derivative) to be computed from the other quantities; (V4) the model can be sequentially computed, i. e. it does not contain computational cycles; (V5) each assignment agrees with the definition of the biochemical quantity described (e. g. only positive values for concentrations); (V6) all quantities are semantically independent (i. e. there are no pairs of identical, convertible, or semantically dependent quantities).

For the merging of two models (called ‘‘model 1’’ and ‘‘model 2’’), we assume that they are both valid. Ideally, the merged model should contain all statements from the input models; if this is not possible, some of the statements may be left out. In any case, the merged model has to be valid. This can be achieved by the following algorithm [13], which is actually quite similar to the merging scheme for structural models:

1. Convert all element names and physical units to a common set of names and standard units.
2. Compare all pairs of quantities from both models (see Fig. 4, left). Because of the previous conversion, pairs of quantities will either be identical (i. e. duplicates), semantically dependent, or semantically independent.

3. If semantically dependent quantities have been detected, stop the merging process and raise a warning message.
4. For each pair of duplicate quantities, choose one of the alternative model statements. The choice can be made either by the user or automatically (e. g. according to a rule like "always choose assignments from model 1").
5. Certain combinations of choices may lead to computational cycles (see Fig. 4, right), but it is always possible to avoid them by an appropriate choice of statements (e. g. by always choosing the assignments from model 1). In the algorithm, cycles are detected (by analysing the graph of dependencies between algebraic equations) and removed by revising some of the earlier choices.

This algorithm will return either a valid output model or stop with a warning.

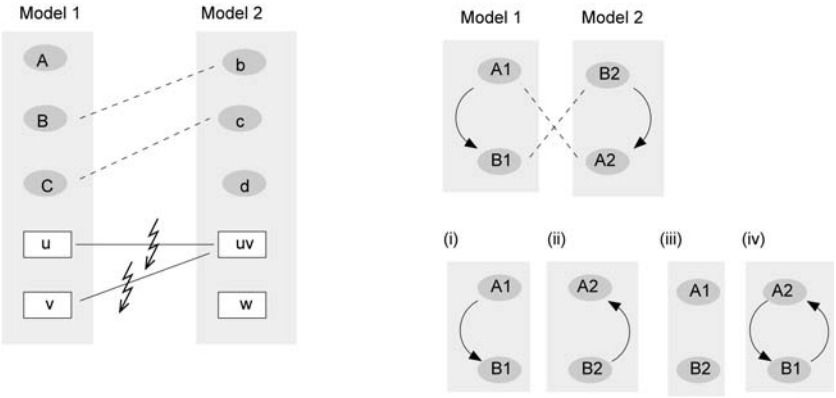

Figure 4. Merging of annotated biochemical models. Left: result of the pair-wise comparison between model 1 and model "lumped 2" from Fig. 1. Some elements are found to be identical (dotted lines) or semantically dependent (solid lines). Due to the semantic dependencies, merging should be abandoned in this case. Right: removal of computational cycles. Solid arrows within the models show that quantities are computed from each other; e. g. in model 1, quantity B is computed from quantity A. After matching the duplicate pairs (A1 = A2, B1 = B2), there are four possible choices: (i) keeping both elements from model 1; (ii) keeping both elements from model 2; (iii) keeping the two independent elements; (iv) keeping the dependent elements. The last choice creates a computational cycle and should therefore be avoided.

Merging of SBML models

SBML [1] is a widely used, XML-based format for biochemical models. SBML is tailored for kinetic models and describes simultaneously the mathematical form and the biochemical interpretation of a model. The main elements of a model represent compartments, substances and reactions including stoichiometries and kinetic laws and parameters. For simulations, the

`species` tags, which refer to substance amounts or concentrations, are translated into mathematical variables. By default, amounts and concentrations are assumed to follow a kinetic rate equation, but it is possible also to specify algebraic and differential equations for them.

Model elements (substances, compartments, reactions etc.) in SBML are not denoted by standard names, but by identifiers defined *ad hoc* within each model. However, elements can also be annotated by references to databases; a recommendable format is the MIRIAM-compliant RDF syntax with BioModels qualifiers [5, 14, 15]. It allows the annotation of model elements (e. g. a `species` tag describing a substance) with biological entities listed in databases. Besides exact equality, the qualifiers make it possible to specify different kinds of relationship: the `isVersionOf` qualifier, for instance, indicates that a substance described in a model (e. g. glucose) belongs to a substance class (e. g. hexoses) listed in the database.

Syntactically, SBML does not have the form of a statement list; however, a valid and properly annotated SBML file corresponds to an explicit biochemical model, so the above merging algorithm is in principle applicable to SBML models; we have implemented similar merging algorithms in the tools SBML merge [13] and SemanticSBML [16]. SemanticSBML allows the annotation, checking and merging of SBML models. It helps the user to annotate model elements with unique identifiers from various databases including KEGG [8], Reactome [17] and ChEBI [18]. These annotations are used for comparing the elements in model merging: the tool aligns presumably identical model elements and indicates conflicts between them; the user can then revise the alignment and resolve the conflicts. Following the structure of SBML, the chemical reactions (corresponding to individual terms in the rate equations) and not the entire rate equations are treated as basic model statements.

Validity Criteria for Biochemical Models

A main task in model merging is to ensure or to check the correctness of the merged model. Correctness, however, is a matter of definition: according to George Box [19], ‘‘essentially, all models are wrong, but some are useful’’, so even the best cell model is only a rough approximation of reality. Thus, instead of requiring correctness in an absolute sense, we shall ask whether a model satisfies certain *validity criteria*; which of the criteria is relevant in a specific case depends on the type and the purpose of the model. Even if a validity criterion is almost trivial, it may become an issue when models are merged automatically. The criteria can be grouped into five categories:

1. **Syntax**. Syntactical correctness ensures that a model can be read and processed, which is a basic requirement for all further validity checks and for model reuse in general. Syntactic problems, such as typos or missing tags in an SBML file, can be detected automatically from the model alone without any reference to a math-

ematical or biochemical interpretation; an automatic validation tool for SBML files can be found at [20]. In a broader sense, we can also regard verbal descriptions in a paper as syntactically incorrect if they are unclear or incomplete.

2. **Mathematics and calculation**. Depending on the intended sorts of calculations, a model should have certain mathematical properties, in particular, existence or uniqueness of mathematical solutions. For kinetic models, for instance, one may require that (i) there is one explicit equation per variable, (ii) the right-hand sides are defined for all allowed values of the function arguments (e. g., non-negative values for all concentration variables; real values for all flux variables) and (iii) there are no computational cycles.
3. **Biochemical semantics**. In this category, we consider the biochemical meaning of model elements, but only regarding simple ontological facts (''glucose is a hexose'', ''mitochondria are part of the cell'', ''reaction VW contains reactions V and W as parts''). Possible validity requirements are: (i) all elements must be correctly annotated; (ii) individual model statements must agree with their semantics (e. g. a variable representing a concentration must be non-negative); (iii) statements must agree with each other (or, more strictly, all quantities must be semantically independent).
4. **Empirical facts**. In addition, one may require that a model respects certain laws of physics (e. g. second law of thermodynamics), chemistry (e. g. conservation of atom numbers), or biochemistry (e. g. realistic values for concentrations). Testing these criteria may require semantic annotations and additional information (for instance, about molecule structures, energies etc.).
5. **Relevance**. Even if a model is free from conflict, it will not automatically be useful; in fact, a model should be based on plausible assumptions, represent a biological system of interest, bring out its basic mechanism, contain only relevant processes and agree with available data. These requirements do not concern the model alone, but also its relationship to available data and to other competing models. It is hard to test them automatically, and they are possibly beyond the realm of automatic checking and merging.

In my point of view, a model is wrong if it fails to fulfil a validity criterion *that it should fulfil*. A didactic model (e. g. a prototypic oscillator model) must have a mathematical solution, but it need not refer to a specific system, so criteria regarding biochemical semantics and realistic numerical values do not play a role. On the other hand, a model that describes a specific pathway should meet these requirements. For automatic model checking, it would be helpful to state explicitly the scope of a model, i. e. which cell types and experimental situations are described, which validity criteria should be fulfilled, or which calculations should be possible; to date, however, there is no formal way to state such requirements in SBML files.

How to prepare reusable models

Besides model merging, there are also other situations in which models are reused: models may be refitted to new data, expanded, simplified, or used as examples to build models for other cell types. Modellers should bear in mind that their models might be reused later, possibly by other people, and should ensure reusability of models right from the beginning (a strategy that could be termed "sustainable model development"). So when constructing a model (and even when designing the experiments that will lead to a model), one should think of typical problems that might occur later, for instance: (i) if the experimental conditions (e. g. the microbial strain used) are not standardized or not well documented, the resulting models may not be compatible; (ii) lumped reactions and metabolites may cause problems and should be avoided, used in a systematic manner, or at least be described unambiguously; (iii) globally fitted parameters may become meaningless after merging and will have to be estimated again.

To avoid such problems, experimentalists and modellers should support standardization efforts (e. g. SBML, MIRIAM, and STRENDA [21]); models should be published (as required in MIRIAM) with all information necessary to reproduce the simulations and model fitting; they should be accessible in a standard (preferably free) format like SBML and be submitted to repositories such as BioModels [2] or JWS online [3]. The meaning of model elements has to be specified unambiguously: in publications, standardized identifiers or names should be used to describe the model elements.

Acknowledgments

I would to thank Falko Krause and Jannis Uhlendorf for their comments on this manuscript. This work was funded by the European integrated project BaSysBio.

References

[1] Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgemn, T.C., Hofmeyer, J.H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novère, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J. (2003) The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* **19**(4):524 – 531.

[2] http://www.ebi.ac.uk/biomodels.

[3] Olivier, B., Snoep, J. (2004) Web-based kinetic modelling using JWS online. *Bioinformatics*, **20**(13):2143 – 2144.

[4] Klipp, E., Liebermeister, W., Helbig, A., Kowald, A., Schaber, J. (2007) Systems biology standards-the community speaks. *Nat. Biotechnol*. **25**:390 – 391.

[5] Le Novère, N., Finney, A., Hucka, M., Bhalla, U.S., Campagne, F., Collado-Vides, J., Crampin, E.J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J.L., Spence, H.D., Wanner, B.L. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* **23**(12):1509 – 1515.

[6] Teusink, B., Passarge, J., Reijenga, C.A., Esgalhado, E., vanderWeijden, C.C., Schepper, M., Walsh, M.C., Bakker, B.M., van Dam, K., Westerhoff, H.V., Snoep, J.L. (2000) Can yeast glycolysis be understood in terms of *in vitro* kinetics of the constituent enzymes? Testing biochemistry. *Eur. J. Biochem.* **267**:5313 – 5329.

[7] Snoep, J.L., Bruggeman, F., Olivier, B.G., Westerhoff, H.V. (2006) Towards building the silicon cell: A modular approach. *Biosystems* **83**:207 – 216.

[8] Kanehisa, M., Goto, S., Kawashima, S., Nakaya., A. (2002) The KEGG databases at genomenet. *Nucleic Acids Res.* **30**:42 – 46.

[9] Frege, G. (1892) Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* **100**:25 – 50.

[10] Ederer, M., Gilles, E.D. (2007) Thermodynamically feasible kinetic models of reaction networks. *Biophys. J.* **92**:1846 – 1857.

[11] Liebermeister, W., Klipp, E. (2005) Biochemical networks with uncertain parameters. *IEE Proc Systems Biology* **152**(3):97 – 107.

[12] Liebermeister, W., Klipp, E. (2006) Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor. Biol. Med. Mod.* **3**:41.

[13] Schulz, M., Uhlendorf, J., Klipp, E., Liebermeister, W. (2006) SBMLmerge, a system for combining biochemical network models. *Genome Informatics Series* **17**(1):62 – 71.

[14] RDF/XML syntax specification (revised) (2004) http://www.w3.org/TR/rdf-syntax-grammar/.

[15] http://www.ebi.ac.uk/compneur-srv/miriam-main/mdb?section=qualifiers.

[16] http://sysbio.molgen.mpg.de/semanticsbml/.

[17] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio,P., Schmidt, E., deBono, B., Jas-sal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E., Stein, L. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*. **33** Database Issue:D428 –D432.

[18] http://www.ebi.ac.uk/chebi/.

[19] Box, G.E.P., Draper, N.R. *Empirical Model-Building and Response Surfaces*. Wiley, New York, 1987.

[20] http://www.sbml.org/Facilities/Validator.

[21] http://www.strenda.org/.