

Bachelor thesis on the topic:

Combination of Mathematical Models of Signaling Pathways

Study course: Bioinformatics

Anna Kosenko

Pflugstr. 22

10115 Berlin

Date of submission: 17.08.2009

First advisor

Dr. Wolfram Liebermeister
Humboldt University Berlin
Department of Biology
Theoretical Biophysics Group
Invalidenstrasse 42
10115 Berlin

Second advisor

Prof. Dr. Alexander Bockmayr
Freie University Berlin
Department of Mathematics and Computer Science
Arnimallee 6
14195 Berlin

Statement of authorship

I hereby certify that this diploma/master thesis has been composed by myself, and describes my own work, unless otherwise acknowledged in the text. All references and verbatim extracts have been quoted, and all sources of information have been specifically acknowledged. It has not been accepted in any previous application for a degree.

1 Introduction

1.1 Systems Biology Fundamentals

First, I would like to introduce you to Systems Biology. Following are a couple of definitions, which can help to comprehend the idea(s) of this comparatively new field of science.

“Systems biology is the study of an organism, viewed as an integrated and interacting network of genes, proteins and biochemical reactions which give rise to life. Instead of analyzing individual components or aspects of the organism, such as sugar metabolism or a cell nucleus, systems biologists focus on all the components and the interactions among them, all as part of one system. These interactions are ultimately responsible for an organism’s form and functions.”[1]

“The science that discovers the principles underlying the emergence of the functional properties of living organisms from interactions between macromolecules”[2]

Accordingly systems biology is a heterogeneous field of science, which includes topics from physics, biology, chemistry and mathematics, combining them for getting a better understanding of the life’s fundamentals. The expression of biological events through mathematical models allows an accurate analysis of the processes and dependencies between single components in this processes. Often mathematical models are represented by ODEs (Ordinary Differential Equation Systems) which can be solved numerically. Other representations are for example stochastic models, which include the stochastic effects of small molecule numbers, or discrete models like boolean methods.

In the following, some basic concepts which are important for systems biology are introduced, more detailed information is given in section 2.1.

SBML

The major part of systems biology deals with model creation. In order to create a computational model one needs a suitable language, which a computer can understand. The Systems Biology Markup Language (SBML) is a computer-readable format for representing models of biological processes. It’s applicable to simulations of metabolism, cell-signaling, and many other topics. [3] The majority of the models found in current databases like BioModels, JWS are represented in SBML. All the models in my bachelor’s thesis are also based on SBML.

MIRIAM

After a model is created, it is a good style to annotate it, that means to give references to certain elements in databases for model components in order to give a biological meaning to the elements in the model. MIRIAM is an effort to standardize the Minimal Information Requested In the Annotation of Models, so that different groups can collaborate on annotating and curating computational models in biology. The goal of the project, initiated by the BioModels.net effort is to produce a set of guidelines suitable for use with any structured format for computational models [4]. If every modeler will annotate her/his models following the MIRIAM Standard, model exchange is going to be much more easy.

BioModels Database

Such an exchange of models written in SBML and annotated with MIRIAM compliant annotations can be found in the BioModels Database [5]. The BioModels Database is a data resource that allows modelers to submit, search and retrieve published mathematical models of biological interest. Models present in the BioModels Database are annotated and linked to relevant data resources, such as publications, databases of compounds and pathways, controlled vocabularies, etc. In this thesis I used various models from the BioModels Database.

Saccharomyces cerevisiae

One of the most examined and therefore modeled organisms in systems biology is the budding yeast or *Saccharomyces cerevisiae*. It is probably the most intensively studied eukaryotic model organisms in molecular and cell biology. Many proteins important in human biology were first discovered by studying their homologs in yeast; these proteins include cell cycle proteins, signaling proteins, and protein-processing enzymes [6]. All models in this bachelor thesis describe processes in *Saccharomyces cerevisiae* cells.

1.2 Aim of this project

Model combination

The merging of mathematical models based on biological processes involves the combination of equal or, sometimes, similar components of these models. Thereby one of the models is expanded by another. Model merging is a useful procedure, which could simplify the modeling process of

large or complex models, because many modelers can work simultaneously on different parts of the model, and finally merge all these parts to complete model. Thereby the modeling can be accomplished in much less time and due to several authors with smaller error rate (as the modelers can correct each other).

Nevertheless, since model merging is not a common procedure yet and therefore not perfectly studied, some modelers may have difficulties, for example with the merging of not carefully prepared models.

However, with more interest in the topic of model merging and therefore more studies and more development, the merging software is going to improve, as well as the modelers are going to create new models, considering that these can be used for merging. That is why it is most important to engage in merging process development and in the development of possibilities to improve it.

Global Model Project

In order to analyze model merging, I chose a subproject of the UniCellSys Project [7]. This project concentrates on combining several basic cell process models, aiming to create a global model. The global model would represent a cell, including mostly all metabolic processes as well as cell growth. However, the merging of models from different creators may be very complicated to accomplish for someone who did not create them, due to differing naming conventions and modeling style. For this reason some guidelines were suggested, which should standardize the models, simplifying the merging procedure. Creating such a global model would be a very important step not only for systems biology, but for molecular biology, genetics and other fields of biology and medicine.

Goals

Therefore the major goal of this bachelor thesis is the accomplishment and analysis of a model combination procedure. As merging of some models (e.g. models with few overlapping regions) can be problematic, problems or errors, occurring during the merging process, and their solutions should also be discussed in this thesis. Furthermore, the annotation of the models to merge and guidelines from the global modeling project are to analyze. The expansion of these guidelines and the creation of some examples for whose utilization are more goals of this thesis.

1.3 Overview on topics of this work

The main topics of this thesis are the following:

- Annotation of the models used in this thesis, following the annotation guidelines of the global modeling project and the proposal of extended requirements.
- Analysis of the model merging procedure, including the process of model merging, examination of the resulted merged model and the analysis of the process itself. The handling of occurring problems and difficulties are also to be mentioned.
- The extension of existing annotations and naming guidelines proposed by the global modeling project.
- Implementation of a “Model Validator”, an application for checking models for correspondence with the extended guidelines. Only models that have been proved to correspond to the guidelines are to be included in the global modeling project.
- The special feature of merging of models with variable compartment volume and what one should consider in this case.

1.4 Models

In this section some models are introduced, which were used for the preparation of this thesis or resulted from it. Detailed description of the work on these models can be found in section 3.4.

HOG Pathway

The HOG Pathway model is a mathematical model, which describes the response of yeast (*Saccharomyces cerevisiae*) on osmotic shock. Osmotic shock means an extreme change in the osmolarity outside of the cell, which activates specific receptors, initiating a signaling pathway. As the result of this pathway, glycerol accumulates, causing the opening of the cell membrane channels and the change of the cell volume. In this way the cell adapts the osmolarity on the inside to the changed osmolarity outside the cell. This model has been introduced first by Klipp et al. [8] and then implemented newly in summer 2008 during my internship in Computational Systems Biology Group (since fall 2008 named Group of Theoretical Biophysics). Besides, five submodels, representing five sections of the model, were created. The intention was to test and analyze the merging procedure, by merging the five submodels to one model and compare the result to the original complete model.

Glycolysis model

The Glycolysis model describes the glycolysis pathway in *Saccharomyces cerevisiae* and was introduced by Hynne et al. [9]. For the thesis this model has been reannotated, as an example for how a model should be annotated for merging. The Glycolysis model can be found in the BioModels database under ID “BIOMD0000000061”.

Cell cycle model

This model can be found in the BioModels database under ID “BIOMD0000000056”. The cell cycle model has been introduced by Chen et al. [10] and describes the cell cycle in *Saccharomyces cerevisiae*. This model has been reannotated and also acts as an annotation example for models to merge. The special feature of this model is the use of “events”. The events are the components of the model, which indicates when certain reaction has to start, by modelling instantaneous change in a set of variables (e.g. reactions). For example in the cell cycle models events are used for the indication of the cell division.

Global model

The global model is a relatively small model, which serves as a pool for other models meant to merge for the global modeling project. It was developed by Hans-Michael Kaltenbach and Jörg Stelling. This model is essential to the project since many models have few or no overlapping regions and the merging of them would not take place. But if we merge every model for the project with the global model, it will expand resulting finally in a cell model. I annotated and adjusted the original global model to my project, so that there are two existing versions.

2 Methods

2.1 Basics

As already mentioned in the introduction (section 1.1), there are some important tools and formats in systems biology like SBML, MIRIAM and model databases. As working on models, and in particular working on this thesis requires the use of these concepts, the next section presents them in detail.

SBML

SBML (Systems Biology Markup Language) is a machine-readable format for representing models. It's oriented towards describing biological processes that take place over time. An example of this is a network of biochemical reactions. SBML has been created to represent these biochemical reaction networks. [3] SBML is based on XML and the first official version (Level 1 Version 1) was released in March 2001 by a small team of researchers. Levels of SBML are upward-compatible, making it easier to use with every kind of software, even with the older versions.

The main purpose of SBML is to enable modelers to use different tools, without rewriting a model each time and to ensure the survival of models even if the software used for creation is not available anymore. Another unique feature of SBML is that to every entity in the model one can attach a machine-readable annotation, which can be used to link this entity to external databases. Biomodels Database (more see below) is the best example for annotated SBML models.

In this thesis SBML plays a leading part, as model creation and annotation are based on this language.

MIRIAM

In order to standardize computational models, the MIRIAM (Minimal Information Requested In the Annotation of Models) Standard is composed of three parts, these are: reference correspondence, attribution annotation, and external resource annotation. Only models, that satisfy the requirements of these three components are recognized as MIRIAM annotated. More detailed information about these three requirements categories are to be found in [4].

MIRIAM supports only particular databases for model annotation. These databases are listed on [11]. Especially for models describing processes in *Saccharomyces cerevisiae* there are a couple of databases most important. In connection with the global modeling project, the databases CHEBI

(for annotation of metabolites), SGD (for protein annotation) and GO (annotation of compartments, events or parameters) are of use.

The requirements for the global modeling project are based on MIRIAM and every model to be used in the project and in this thesis has to be compliant with the MIRIAM standard.

Databases

The most important database for working on this thesis is the Biomodels Database [5]. As already mentioned, models in this database are fully annotated and linked to relevant data resources. The cell cycle and glycolysis model, introduced in section 1.4 are originally taken from Biomodels Database and have been modified and reannotated for the use in this thesis. Also of importance is the JWS-Online Database, which is another database, where many useful computational models of biological processes are stored. However, models from JWS-Online are not necessarily annotated as in the case of the Biomodels Database.

Annotation in SBML

Model annotations have an important role in this thesis, for annotation is essential for model combination. Annotating a model means attaching a special XML tag to every model component, linking it to special MIRIAM Resource. Since the naming of model components is not standardized even in MIRIAM Standards, annotations are the only way for software (or a person, who combines models manually, but did not create them) to detect the biological meaning of single compartments, species or parameters and merge them if they are equal.

Besides MIRIAM, there is also a possibility of SBO-Term annotation. SBO (Systems Biology Ontology) is a controlled vocabulary created specially for systems biology uses. Attaching an SBO-Id to a model entity can define classes of model elements (e.g. enzymes) and is mostly helpful for further working with the model. More information about SBO classification can be found in [12]. SBO-term annotations are also a mandatory requirement within the global modeling project.

2.2 Software

There are many tools available for free use, which help the user in creating, annotating and merging of models. In the following section two of them, semanticSBML and COPASI are presented because these tools are most important for the thesis.

semanticSBML

semanticSBML is a free software developed in the Theoretical Biophysics Group at the Humboldt University Berlin [13]. This software is used for automatic construction, annotation, merging, and checking of mathematical models. A Web version of semanticSBML also exists, but for models that have not been carefully prepared, it is recommendable to use the full version, for merging conflicts are resolved automatically in web version. For this reason this thesis has been prepared using the full version of semanticSBML. Since the topic of this thesis is the combination of models, which requires whose annotation, this software is the central and the most important tool in this project. More information about semanticSBML and its documentation can be found on [14]

COPASI

COPASI is a software application for creation, simulation and analysis of biochemical network models. The principle of model creation in COPASI is the representation of metabolic reactions as ODEs. Thus models can be simulated and analyzed on a mathematical level.

In this thesis (more precisely, during the internship in the Computational Systems Biology Group, renamed since fall 2008 in Theoretical Biophysics Group) COPASI was used for the implementation of the HOG models (see section 1.4) and it is the central software for model simulation. The special characteristic of COPASI is the use of the special XML format for model creation. This XML is a different one from SBML, so models created in COPASI have to be exported in SBML for further use. For simple models this does not affect the results, but for more complicated ones, some converting issues may occur.

There is also an important characteristic of COPASI as a model creating software, which is the possibility to implement models with variable compartment volumes (e.g. in the cell cycle model the volume of cytosol is changing during cell division). COPASI is the only application so far supporting this implementation. Unfortunately, a software tool which creates the models in SBML and supports variable volumes does not yet exist. That is why models supposed to be exported to SBML can't use the advantage of COPASI's volume modeling.

3 Results

3.1 Requirements on models

As first result of this thesis the new specified requirements for model creation and annotation are to be named. These requirements evolved from already defined requirements listed in Deliverable D7.1.

Deliverable D7.1

Deliverable D7.1 is a document created by Hans-Michael Kaltenbach and Jörg Stelling, that specify a list of requirements for models to be merged within the global modeling project. It contains requirements for five areas of modeling:

Modeling framework and file formats The model has to be encoded using SBML file format and be modeled via ODE. In exceptions discrete models (event-based) can be considered.

Global non-technical information included in models Information about the model, the author(s), the date of the creation and the last modification as well as a reference to a publication or report containing further information on the model have to be added to each model.

Model annotation and completeness Correct SBML annotations has to be added to each model according to MIRIAM Standard. Further, more naming conventions and annotation standards are required, like the sbo term annotation, the databases to use and the guidelines for the annotation of the protein complexes and the post-translational modifications. Information about the annotation of the parameters and the units is given as well.

SBML version, validation and storage The model should be a correct SBML and be available on public databases like the BioModels Database or JWS Online.

Global variables To each model some global species (according to the pathway, which the model describe) have to be added. For example the concentration of ATP or the cytosol volume.

This document is designed only for internal use by the UniCellSys project, and was kindly provided to me by Hans-Michael Kaltenbach.

New requirements

During analyzing, preparing for merging and annotation of several models, a list of new requirements, more strict and specifying than is Deliverable D7.1 was made. Especially annotation and

naming conventions of model components, proposed in Deliverable D7.1, were sharpened. There are as well some mandatory and some non-binding requirements, but for better standardization of models non-binding requirements should be satisfied as well.

Mandatory

In the following, new mandatory requirements are listed. One has to consider that these requirements are just additions to basic requirements presented in Deliverable D7.1.

- **SBML Version:** since Level 2 Version 1/2 do not support SBO-term annotations, models have to be written in SBML Level 2 Version 3 or higher
- **Units:** models to be merged must use exclusively time units in seconds (s), concentration units in milimole (mM) and volume units in liters (l)
- **Annotation:** each compartment, species and event must contain at least one MIRIAM annotation (further see D7.1)
- **SemanticSBML check:** must not report any errors (warnings are allowed)

Good style

These are some non-binding requirements, resulting from analyzing and merging of several models. Though these specification are not mandatory, it is nevertheless advisable for all modelers to satisfy them, when creating a model.

- Naming of species should be specified more clearly than in Deliverable D7.1 requirements (intuitive, clear description of the biochemical meaning).
- In more complicated models some species represent multiple reactants. In such case either that species has to include each reactant in whose name or model supplementary should indicate every single reactant belonging to this species. The annotation of such species should contain annotations for all reactants, ensuring that the model merging software can detect the possible match to equal species. For example APC (anaphase promoting complex) is one of the species in the cell cycle model. It contains many single proteins, that have to be annotated individually. The naming of every protein should be described in the supplementaries of the model, but the name of the species can remain "APC".
- As many modelers use a "total" amount of species as an independent species, new naming convention for such case could be consistent. For example of protein Tip1: e.g. "Tip1p total".

- For naming of metabolites consistent names should be taken. The proposal would be to take full names, e.g. “Glucose”; abbreviations can be used as IDs, e.g. “Glc”.
- SBO-term annotation should be more specific than in D7.1. Guideline: always choose the most specific sbo term possible.

3.2 General Annotation Guidelines

The most important matter to take into account when merging models for the global modeling project is the use of databases only listed in D7.1. These are:

- SGD (Saccharomyces Genome Database) for proteins and genes annotations
- GO (Gene Ontology) for compartment, events, specific species and parameter annotations
- ChEBI (Chemical Entities of Biological Interest) for the annotation of metabolites or additional chemical groups in modified proteins, e.g. phosphorylation

However, by creating models of other organisms or for different purposes, more databases from the MIRIAM Resources can be used for model annotation. But one has to make sure that if the model is intended to be merged with another one, the data types used for model annotation must be identical, because e.g. the same substance in different databases won't be recognized as equal by merging software (e.g. semanticSBML).

Another important feature is the use of qualifiers. Qualifiers are additions to annotations, which can express the relation the model component stands to its annotation. There are ten qualifiers available in the current SBML version. These are: **encodes**, **hasPart**, **hasVersion**, **is**, **isDescribedBy**, **isEncodedBy**, **isHomologTo**, **isPartOf**, **isVersionOf**, **occursIn**. More detailed information for each type of qualifier can be found in [15]. Using qualifiers for model annotation improves it a lot and is highly helpful for the merging.

Events

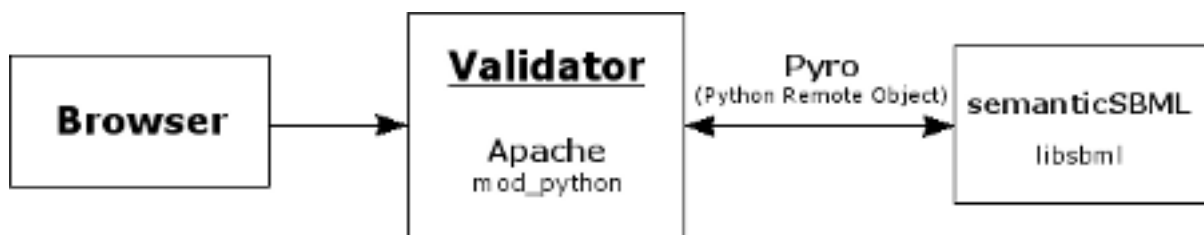
Besides compartments, species, reactions and parameters, some models include “events”, which could for example be a change in the concentration of some reactant above a predetermined threshold. An event triggers specific reactions. As events are not commonly used model components like compartments or species, their annotation is not quite clear. Therefore my proposal for the annotation of events is the use of GO (Gene Ontology) IDs, since GO is the most extensive database and therefore has the biggest chance of having a suitable entry for almost every possible event.

3.3 Model Validator

The requirements for model creation, naming and annotation presented in section 3.1 can certainly be checked manually, but in case of large and complex models the checking can become error-prone and time-consuming. That is why a Model Validator is needed. The task of the Model Validator is to verify that a model fulfils the requirements. The Model Validator lists all requirements and whether the requirements are fulfilled. Moreover every species is listed in the output and one can see if the annotation of every species is correct. Currently the Model Validator checks for:

- Correct unit definitions, ensuring that time units are defined in seconds, volume units in liters and substance units in milimole
- Correct level and version of SBML used for model creation, which is Level 2 Version 3 or higher, ensuring that sbo term annotation is possible
- Correct annotations (MIRIAM Standard) of all compartments, species and events in the model

The Model Validator was implemented as a web application in python. For the implementation of the view (html output of the results) I used code developed by Jannis Uhlendorf as a template. The realization of this web site is made with Pyro (PYthon Remote Object), a distributed object system for Python. In a distributed object system clients send requests to distant servers, which host the remote objects. Pyro simplifies the creation of clients and servers a lot [16]. As the Model Validator uses a function from semanticSBML, the use of Pyro eases the loading of the internal databases from semanticSBML a lot. The following schema is a simple demonstration of how the Model Validator web site is working:



The API for the class validator can be found in the appendix, as well as screenshots of an example of how to use the Model Validator. For now, the Model Validator is not yet available for use in web.

3.4 Models

During the preparation of this thesis, I worked on the models introduced in section 1.4. My results are presented in this section. Furthermore, I added all these models to the model repository of the Theoretical Biophysics Group, using SVN.

HOG Pathway

The HOG pathway models are to mention first. These were created during the internship in the summer 2008 based on the implementation by [8]. Because the merging process was first analyzed using them, they are important for this thesis. During the work on this thesis the HOG pathway models have been adjusted to the new model requirements, which include the adjustment of the time, substance and volume units, renaming of some species and the annotation of the model components with sbo terms. The models have also been reannotated, due to the changed MIRIAM annotation tags since summer 2008.

Glycolysis model and Cell cycle model

Another result of this thesis are models, that are fully annotated. They also serve as an example for new additional specifications. These models are the glycolysis model [9] and the cell cycle model [10]. As these models were created for other purposes than merging them into a global model, they did not satisfy my proposed requirements. First of all, the models were downloaded from BioModels database in the SBML Level 2 Version 3 and not in the original version, since SBML under Version 3 does not support sbo term annotation. Further, the models were reannotated, using databases listed in section 3.2 and then the units were adjusted to the requirements. Some model components even had to be renamed in order to satisfy non-binding specifications. This procedure finally resulted in two models, which can serve as an example for model creation for all modelers who are interested in the global modeling project. Nevertheless, some difficulties occurred during model adjustment, these are described in section 4.2.

Global model

At last the global model, which serves as a pool for many single pathway models to merge, was edited. It contains such common species as ATP/ADP/AMP and the most common compartments, like cytosol, nucleus, mitochondrion and the extracellular space. The original version of the global model, kindly provided by Hans-Michael Kaltenbach, was modified in preparation for merging it with other models. More precisely, the model was annotated and some statements were altered. Among them are the volume statements of the cytosol [17] and the nucleus [18]. The merging of the global model with the Glycolysis model illustrates the merging procedure and serves as an example of how the merging of the global modeling project is meant to be processed (more in section 3.6).

3.5 Volume modeling

Another aspect of the model creation to be mentioned in this thesis is the modeling of variable compartment volumes. Some pathways, like cell cycle or HOG pathway, describe processes that cause a change in the volume of compartments. In cell cycle models the volume of cytosol is changing during the cell division and in the HOG pathway the volume of the cytosol adapts to the changing osmotic pressure. By default in model creation, modelers usually use concentrations of species (i.e. the amount of species divided by the volume of the compartment) and not the amounts. It is clear therefore, that a change of volume causes the concentration of all species in this particular compartment to change.

Guidance

After implementing and working on the model with variable compartment volume (HOG Pathway), I figured out some hints for modelers intending to create such models. In order to model such a case they can use the software COPASI. It provides an unique function for the compartment volume modeling, allowing to set the compartment volume as an assignment, depending on one or many other model components, or as a differential equation. The most important feature is that COPASI automatically adjusts the concentration of every species located in the compartment, with a changing volume. In this way the model shows realistic behavior during simulation.

Unfortunately COPASI creates models in proprietary format, which is based on XML but differs from SBML. Although a possibility exists to export SBML from COPASI, the information about compartments and associated volume equations get lost during the export process, because of the software tools for the implementation of the SBML models, that don't support volume modeling. Therefore the modeling of the variable compartment volume in COPASI is not suitable for every modeler.

Nevertheless, there is a possibility to create such a model in SBML. The idea is simple: create a global variable (e.g. "cyt_vol"), define an assignment rule or a rate rule, which describe the behavior of this global variable (respectively the change in the compartment volume). Thus the simulation of this global variable represents the simulation of the compartment volume.

Unfortunately I am not yet aware of a possibility to implement the compartment with the changing volume in SBML, which influence the concentration of the affected species. Until SBML software tools support such a feature, COPASI is probably the most useful tool for analyzing volume behavior in mathematical models in systems biology.

3.6 Model Combination

The next major part in this thesis is the issue of the model merging or the model combination. The rationale of the whole process is the assembly of two or more models to one single model, in which the equal model components are merged into one new component. At this point the annotation of the models is of primary importance.

Internship

During my internship in the Computational Systems Biology Group (new name since fall 2008: Theoretical Biophysics Group) in summer 2008 I had a task to test the new software, semanticSBML, which was developed in the group. For this reason I reimplemented the HOG pathway model [8] and five separate models, each representing one submodel of the complete HOG model. The aim of the test was to figure out if the simulation of the original model was equal to the simulation of the model, created by merging the five submodels. The test was successful as the equal model components were merged, but the simulations were regrettably not equal. However, the reason for unequal simulations was found and the bug in semanticSBML was fixed after my internship. This thesis is based on my work and insights during summer 2008.

Global Module

As an example for model merging, a Global model and the Glycolysis model, both introduced in section 1.4, were merged using semanticSBML. On this part, two topics of this thesis intersect, namely the merging issue and the global modeling project. The goal of merging these two models is to analyze the merging process and to create an example of how the global modeling project may perform.

The merging of the models was successful and uncomplicated, because both models were prepared accordingly to the new specifications and annotated as the new specifications require. If some conflicts between the models occur during merging in semanticSBML, e.g. the different concentrations of the species, which is equal in both models, the conflict resolution menu suggests how the conflict may be resolved (the screenshot of the conflict resolution menu can be found in the annex). In the case of these two models, almost every conflict was resolved by the user in favor of the global model, as the latter model serves as a pool for other models.

The merging identified three components in the distinct models that were identical. These are cytosol, extracellular space and ATP. Three may appear too few, but the aim of the global model project

is to expand the initial small global model to the complete network. The complete network includes many different pathways, which intersect almost only in such common species like ATP/ADP/AMP. That is why the merging of only one model with the global model does not result in a merged model with many intersection points.

4 Discussion

4.1 Results Summary

During the preparation for this thesis, I worked on several topics regarding the combination of mathematical models of signaling pathways. Among them are annotation issues, the analysis of the model merging process and the issue of the modeling of the compartment volumes. In this section I would like to present a brief summary of the achieved results.

1. New requirements for the models are presented. These ensure that the models, that satisfy these specifications are best prepared for the merging process.
2. A proposal for the annotation of the events was made. This proposal is not meant to be mandatory for the modelers, but merely advisory.
3. Model Validator was implemented, which is also available as a web site. The aim of the Model Validator is to check the models for their adherence to the new guidelines proposed in this thesis.
4. Two models, annotated to satisfy the new requirements, serve as an example for other modelers interested in this topic.
5. A guidance for the creation of the models with a variable compartment volume was proposed. It may help other modelers, who are interested in this issue.
6. A merged model, composed of global model and the glycolysis model, was presented and described. This model shows how the global modeling project may proceed.

4.2 Model Requirements

The new requirements, proposed for the use for all modelers, interested in the merging of their models, are most helpful. These requirements sharpened the basic requirements from D7.1 in order to make the merging process more smooth and easy. It is important to understand that with every satisfied requirement a model becomes more comprehensible for a machine (in order to find equal model components and merge them respectively) as well as for the person working with this model. In my opinion, these requirements can definitely help to improve the merging procedure, but there are also some other aspects to consider. For example, these requirements were established after the examination of only a couple of models, which are the cell cycle model [10], the glycolysis model [9] and the HOG pathway models [8] (examined in summer 2008). That is why these requirements are by no means complete or final and have to be improved further. For this purpose, more models

should be examined and merged in the future. In order to expand the specifications, the corresponding conclusions have to be made and new requirements established, after the models have been merged.

4.3 Merging

During the preparations for this thesis, two models were merged (as described in section 3.6). Furthermore, more models (HOG Pathway) were merged during my internship in 2008. All these merging processes were made using semanticSBML and were successful. The reason for the success was, in my opinion, the careful preparation of the models, and that every model to merge satisfied the guidelines from D7.1 and the new requirements, proposed in this thesis. That is why the merging process with prepared models is uncomplicated. Nevertheless, several difficulties may occur during the merging of the models, depending on their specific characteristics. In particular models with different units are no longer correct after merging, as a user has to select one of the several units, which is then applied to the new merged model. The solution for such issues can be the standardization of the models or the reinitialization of the values in the new merged model. But the last proposition is rather unrealistic for large or complex models, as this procedure may become very time consuming. Another possibility could be the conversion of units before merging or the use of unit conversion software, that is however not yet available for common use.

Events

Another issue in the merging procedure is how the events are to be merged if the models contain these. Even though this kind of merging did not occur in this thesis, it is anyway an important topic to discuss here. One of the difficulties with the “events” is their annotation. I have already proposed in section 3.2 how the events may be annotated, but in those cases, where the GO Database does not have suitable annotation, other databases will be used. In this way, the events can not be merged automatically. But that is not the major problem, because of the possibility in semanticSBML to merge some model components manually. Another question to be answered is how exactly two or more events can be combined. One possibility may be just to combine the conditions (bound with the logical “and”) of the events to one condition, as well as the actions, which come about if the condition is fulfilled.

4.4 Annotation

The annotation of the models is the most important precondition for the successful merging of the models. During the work on this thesis two models (the cell cycle model [10] and the glycolysis model [9]) were reannotated in order to fulfill the requirements for the models to merge. As both models were implemented in 2004, and 2001 respectively, neither the model merging nor the annotation and model creation standards were an issue. For this reason these models were quite difficult to reannotate. In particular the use of species, which represent multiple substances (but not the complex of these), made the annotation unclear. The annotation of such species with references to every single substance, make it equal to the complex of these species and therefore is incorrect. For such a case I would propose for all the modelers to use one species for one substance, even if they behave equally.

The modelers should also consider to document the implemented model completely, describing every component in the model and its role in the pathway. Detailed documentation helps a lot while annotating. Other issues are the qualifiers, already mentioned in section 3.2. Even though in this stage of development the merging software semanticSBML can not yet distinguish between qualifiers for model merging, it is still important to make use of them in order to create fully annotated models with all advantages SBML offers. Beyond that, the qualifiers can improve the quality of the automatic merging, for example the species ATP is not going to be merged with another species, which is just containing ATP.

4.5 Global Modeling Project

As the conclusion of this thesis, I would like to present an overview of the advices (regarding the creation of the models) for the modelers, the curator of the databases (like BioModels or JWS) and the software user (who is commonly not aware of the background of the model, at least, not in detail), in order to assure that the model is well prepared for the merging procedure.

Modeler

In the first place, the success of the model merging depends on the modeler, who implements the model and is the one responsible for the model to fulfill the requirements. Here is a brief summary of the specifications for the models to be merged:

- Model Validator (see section 3.3), as well as semanticSBML check, must not report any errors.

- Every condition of D7.1 (see section 3.1) has to be fulfilled.
- The implementation of the model and all the components has to be documented in detail.

Curator

If the modeler adheres to the advices, there is not much work for the curator to do. As the curator can not be aware of the details of every single modeled pathway, he/she should pay more attention to the model documentation and make sure that the technical requirements, like the correct version of SBML are fulfilled.

Software User

The software user, who uses the merging software and who is not the creator of the models to be merged, has to make sure that the models consist of at least some overlapping parts. Few overlapping regions does not mean that the merged model is of low quality. One should consider that model parts are only merged if they are identical. If at least one identical element was identified the merging process should be considered successful.

Another hint for software users is not to trust automatical merging, because yet conflicts have to be resolved by a human. Furthermore as merging is done based on the annotations and not yet on the qualifiers, each merging step has to be supervised.

References

- [1] http://www.systemsbiology.org/Intro_to_ISB_and_Systems_Biology/Systems_Biology_-_the_21st_Century_Science
- [2] Lilia Alberghina and Hans V. Westerhoff. *Systems Biology: Definitions and Perspectives*. Springer, 1 edition, November 2005.
- [3] http://sbml.org/Main_Page
- [4] <http://www.ebi.ac.uk/miriam/main/mdb?section=standard>
- [5] <http://www.ebi.ac.uk/biomodels-main/>
- [6] http://en.wikipedia.org/wiki/Saccharomyces_cerevisiae
- [7] <http://www.unicellsys.eu/>
- [8] Edda Klipp, Bodil Nordlander, Roland Kruger, Peter Gennemark, and Stefan Hohmann. Integrative model of the response of yeast to osmotic shock. *Nat Biotech*, 23(8):975-982, 2005.
- [9] F. Hynne, S. Danø, and P. G. Sørensen. Full-scale model of glycolysis in *saccharomyces cerevisiae*. *Biophysical Chemistry*, 94(1-2):121-163, December 2001.
- [10] Katherine C. Chen, Laurence Calzone, Attila Csikasz-Nagy, Frederick R. Cross, Bela Novak, and John J. Tyson. Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell*, 15(8):3841-3862, August 2004.
- [11] <http://www.ebi.ac.uk/miriam/main/mdb?section=browse&offset=0&nb=76>
- [12] <http://www.ebi.ac.uk/sbo/>
- [13] Liebermeister W., Krause F., and Klipp E. Merging of systems biology models with semantics-bml. *5th Workshop on Computation of Biochemical Pathways and Genetic Networks*, 2008.
- [14] <http://www.semanticsbml.org/>
- [15] <http://www.ebi.ac.uk/miriam/main/mdb?section=qualifiers>
- [16] <http://packages.ubuntu.com/de/hardy/pyro>
- [17] http://yeastpheromonemodel.org/wiki/Main_Page
- [18] <http://bionumbers.hms.harvard.edu>

Annex

On the following few pages, some additional material is presented:

- The API of the class “validator”, which is the main part of the Model Validator web application.
- The example of the use of Model Validator: the home view and the output view.
- The conflict resolution menu of semanticSBML