Standards in Computational Systems Biology

Edda Klipp[1], Wolfram Liebermeister[1], Anselm Helbig[1], Axel Kowald[1,2], Jörg Schaber[1]

[1]Max Planck Institute for Molecular Genetics, Berlin, Germany

[2]Ruhr University Bochum, Bochum, Germany

Abstract

Systems biology aims at modeling and quantitative simulation of complex biological systems. This endeavor demands close collaboration and communication between modeler and experimenter, which can be facilitated by standards concerning workflows and data formats. We have conducted a survey to find out what the community thinks about the requirements and benefits of standardization efforts; we also evaluated which modeling methods and software tools are being used in the systems biology community. Free availability and flexibility are the top criteria that govern the choice of software tools. SBML is widely recognized as a standard format for systems biology models. Eighty percent of the 125 respondents favor standardization, but there is also consensus that standards should not be enforced at all costs. The most significant demands are standardized formats for experimental data and mathematical models, standardized names for metabolites, reactions and enzymes, and standardized graphical representation of networks.

1 Introduction

1.1 Standards in Systems Biology

Currently, we observe intense debates about standards in the systems biology scientific community. A series of articles have discussed various aspects of standardization, such as minimal requirements in the annotation of biochemical models (MIRIAM[1]), the compatibility of tools for kinetic modeling[2], or the emerging standards in high-throughput technologies and related ways to establish other standards in biology[3].

Standards are agreements between people in order to enhance information exchange and mutual understanding. They may concern state-of-the-art methods and workflows for experiments and modeling, data formats for experimental results and mathematical models, or agreed nomenclature and graphical representation for biochemical systems. Standards can arise as informal agreements between researchers, they can result from the use of software tools, or they may be enforced by journals and funding organizations.

Standardization also plays a major role on the level of research politics: almost all systems biology projects that are currently funded by the European Commission promise to develop or to define some standards. The hype for standards originates in the facts that the field is quite new and needs to be structured and that it involves expertise from diverse scientific backgrounds. On the one hand, the research object is life itself with all its complexity and diversity, so the definition of standards is not straightforward; on the other hand, standardization efforts in other fields have shown that standards can greatly help to avoid misunderstanding and duplication of work.

To start this process in systems biology, we should know about de facto standards that are already established in the community. We should also know if the existing tools and methods fulfill the researchers' needs and whether scientists would appreciate the development or enforcement of further standards for modeling, data exchange, and model publication.

1.2 The Questionnaire

To answer these questions, we developed a questionnaire[4] that asked the interested colleagues about their modeling habits, the software tools they use for modeling, and their opinions about different aspects of standardization. It was originally addressed to the members of the EU-funded Yeast Systems Biology Network (ysbn.org), but since we obtained the response that these problems are of more general interest, we spread the information as widely as possible, using a variety of formal and informal means. The questionnaire consisted of multiple choice questions and fields for free-text comments. Details about the questionnaire and the statistical evaluation are given in the methods section.

Eventually, 125 persons filled the questionnaire until a deadline (August 29, 2006). The respondents cover all areas of systems biology and describe themselves as modelers (75%), experimentalists (4%), or both (21%). Their research areas include

- modeling of individual pathways such as glycolysis or specific signaling pathways,
- investigation of complex processes such as aging, cell cycle regulation, cancer, robustness, disease dynamics,
- development and application of methods such as metabolic engineering, statistics, model identification, or machine learning, and
- development of software tools for modeling and simulation.

The studied model organisms range from various prokaryotes, yeast, and worms to mammals such as mice and humans. As identified by the ending of email addresses (*at, au, be, ch, cn, de, es, fi, fr, gr, hu, in, it, jp, nl, pl, pt, ro, ru, uk, se, su, tr, za* as well as *com, edu, gov, net*), the origins of respondents cover many countries and all continents.

The responses indicate to what extent standardization in systems biology and especially in modeling is desired by the researchers, which standards are already common, and for which aspects standards are requested. In this article, we will summarize the comments about general aspects of standardization, i.e. (i) if standards are considered necessary at all, (ii) which standards have already emerged, and (iii) where further standardization is requested (section 2.1). We also give an overview about modeling approaches used to study different biological problems (section 2.2) and about the usage of software tools (section 2.3). All percentages refer to a total number of 125 responses to the questionnaire.

Different lessons can be drawn from the results of our survey: (1) experimentalists who are diving into theory will learn about modeling and analysis methods that are suitable for particular biological problems, (2) modelers are provided with a list of exchange formats and advantages or disadvantages of the various tools, and (3) tool developers can learn about exchange formats and functionality requested by the users.

2 Results

2.1 Opinions about Standards

In the following, we shall summarize opinions and statements of the respondents; we assume that they express a consensus that is shared by the majority of modelers in systems biology.

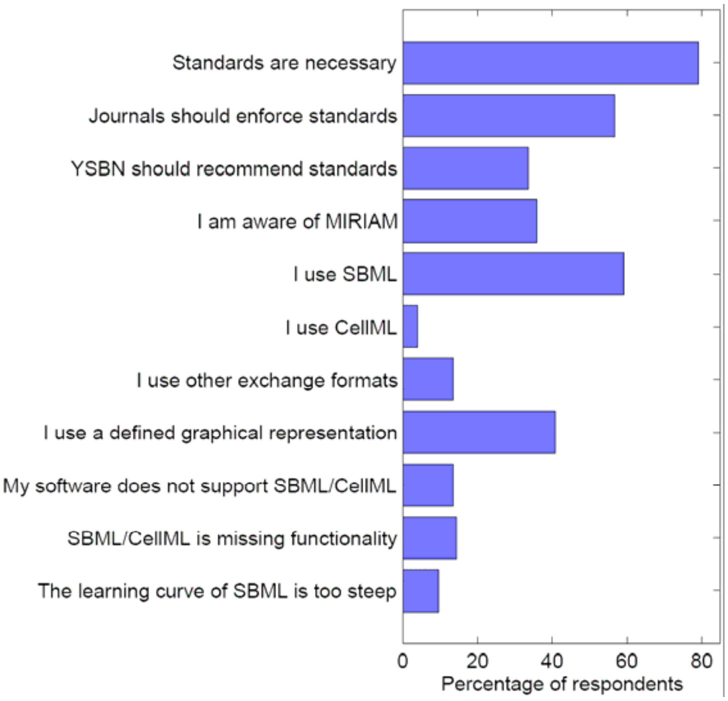


Figure 1

Opinions about standards

Percentages of researchers (from a total number of 125) who marked the respective response in the online questionnaire.

2.1.1 Do we need Standards at all?

About eighty percent of the respondents stated that they consider the creation of standards necessary or desirable (see Figure 1). Although about 20% disagreed, this is a clear vote for standards in general.

Most arguments for standardization were connected to the problems that appear if standards are missing. It was mentioned that many weak models are published; respondents observed that it is often difficult to reproduce and to check the simulation results from computational models; important modeling details are hidden in the paper or supplement or are not mentioned at all. Standards are expected to improve model reuse, expandability, and integration. Respondents expected that the adoption of standards enables modelers to reproduce each other's results, to collaborate more productively, and to increase the efficiency. In addition, theoreticians and software developers need standards, test data sets, and benchmark data sets as experimentally verified gold standards in order to apply and to improve their methods.

Standards are expected to direct modelers to the appropriate software for a given problem and to allow communication between different tools with different functionalities. They are useful for educational purposes; they allow free exchange of information and comparison between different studies. Reimplementation of models becomes easier or dispensable, which reduces duplication of work and the possibility of implementation errors.

Along with positive statements about standardization, it was also stated that standards should be flexible, not become too restrictive, and not prevent alternatives or new developments. Standards must be developed in parallel with both tool development and modeling projects.

Two main arguments against standardization were put forward: first, biology is considered to be too complex to be standardized: all solutions lead to new problems as the field keeps changing. Secondly, specific problems might be solved more easily with specific approaches; obeying standards can be too time-consuming and restrictive. In general, the arguments against standardization were less definitive than the arguments for standardization.

2.1.2 Which Standards have Already Emerged?

In the free-text fields of the questionnaire, Systems Biology Markup Language[5] (SBML) was often mentioned as an emerging standard, sometimes with very positive emphasis, sometimes combined with suggestions for improvement. MIRIAM was mentioned several times as an important guideline for model publication.

About 60% of the respondents actually use SBML, while 4% use CellML[6]. Respondents who do not use such exchange formats explained that they do not know them, see no benefit, never needed them, or do not consider exchanging their models. Besides SBML and CellML, also other model formats are used: most respondents use the storage format that comes with the modeling tool that they use, such as mat-files for MATLAB. In addition, models are stored as text files, as textual description, or in “home-made” formats.

Next we asked about graphical representation of (i) data and information, (ii) interaction networks or biochemical reaction networks, and (iii) model results. Standards for graphical notation have experienced great attention in response to this question, but also in other sections of the questionnaire. About 40% of the respondents claimed to draw graphical representations (wiring schemes, network graphs) according to a defined nomenclature. Many respondents also use the functionality that comes with the modeling tools. CellDesigner[7] received many positive comments regarding the convenience of drawing networks with this tool. Systems Biology Graphical Notation (SBGN) was acknowledged as an attempt to develop standards for graphical notation. Other comments referred positively to the Molecular Interaction Map[8].

2.1.3 Where is Further Standardization Requested?

Finally, we asked about future perspectives of standardization. The answers to this question concerned graphical representation of networks, experimental procedures and data, model encoding, model analysis, spatial modeling, and model publication.

Graphical representation

Many respondents requested a standardized graphical representation of biochemical networks. In particular, there is a need for a graphical formalism that covers fundamental biochemical processes and that can be uniquely mapped (i) to mathematical objects such as ordinary differential equations (ODE) or stochastic simulation schemes, and (ii) to a textual description. Standardization of graphical representation is expected to stimulate the development of new computational tools supporting the interaction between graphical and mathematical modeling, as well as analysis tools. SBGN and SBML are viewed as promising approaches towards a connection between graphics and models. A standardized graphical representation of experimental data was also considered important.

Experimental procedures and for experimental data encoding

It was recognized that in order to support modeling, experimental biology should establish standards that ensure reproducibility of the experiments; for instance, experimental settings such as specific strains or cell lines, experimental protocols, perturbation techniques, or quantitative assays. It was emphasized that standardized experimental conditions are necessary to produce data sets that can be used for model fitting.

In addition, standards are needed to ensure that quantitative experimental results can be encoded electronically. Related aspects are data import and export protocols as well as protocols for retrieving and storing data to databases. Further standardization was requested for diagrams and for validation suites as well as statistical and numerical analysis of experimental results.

Model encoding and model exchange

The mathematical representation of models is perceived as an important issue that deserves to be standardized. In this context, many respondents mentioned model description, accessibility, and exchange formats. It was also emphasized that published models should be electronically available.

SBML was clearly perceived as an upcoming standard, but also drawbacks or further needs were mentioned: (i) The SBML language is becoming more and more complex and therefore difficult to be implemented completely. If different tools support different subsets of features, model exchange is hampered. The conclusion is to demand definition of minimal subsets that all SBML enabled tools must support in order to be called SBML enabled. (ii) The interchangeability of SBML-models by the (putatively) SBML-compatible programs is not satisfying and must be improved (see also [2])

Without reference to SBML, respondents expressed the following needs: transfer of a model from one mathematical framework to another must be enabled. Efforts in standardization are necessary for the exchange between the different tools and methods. Hence, standardization of data input and output and of the running environment is crucial. This would enable the linking of different models as "modules".

Another issue is the relation between model entities and their biological counterparts and their appropriate annotation. It was stressed several times that names of enzymes and metabolites should be unified across species such that one can easily compare reaction networks and presence of pathways. There is need for a unified nomenclature with respect to molecule IDs, sub-cellular localization, and kinetic parameters.

Computational analysis and modeling evaluation

On the computational side, a need was expressed for standards (i) in computational encoding of numerical analysis procedures, (ii) in numerical analysis results, and (iii) in the validation procedures. For models, there is a demand for model validation tests and for protocols to test the quality of models. It is also necessary to establish clear definitions of computational methods. On the other hand, the need for established data sets for performance evaluation of modeling tools was pointed out. Benchmarking of numerical analysis tools was also suggested.

Parameter estimation methods received special attention. Obtaining reasonable parameter values is often considered the hardest task in model development; it was pointed out that the same conceptual model can give very different results, if different kinetic descriptions or different mathematical formalisms are used. The combination of all the available information is an important challenge. Also fitting procedures should be standardized to allow compatible results.

Spatial modeling

Various respondents claimed a lack of standardization efforts in the field of three-dimensional or spatial modeling. Respondents asked for an SBML extension that considers spatial aspects beyond compartmentalization and diffusion, e.g. PDEs with active transport terms or spatial heterogeneities of parameters. PDE models have no real standard yet. The description of three-dimensional geometry also deserves attention.

Model publication

Publication of models and the issue of reproducible model results have been mentioned several times and connected with a hint to MIRIAM as an attempt to set guidelines. 57% of the respondents answered Yes to the question “Should scientific journals support development and establishment of standards, e.g. by demanding authors to submit models in certain format?”. Comments against enforcement by journals stress that the field is still developing, standards do not cover all types of models, and that the researchers are responsible to establish standards. Further suggestions towards the contribution of journals are that (i) journals (should) ask for the software tool that generated the results, and (ii) model publication could be more concise, concentrating on the model and relevant aspects.

To summarize, strong standards for the graphical representation of interaction networks as well as for the presentation of experimental data and modeling results are crucial. A significant demand is to have standardized names for metabolites, reactions, and enzymes. Standardization of experimental protocols and model exchange formats is also considered extremely important.

2.2 Model Types and Analysis Techniques

The participants were asked to state the biological problems they investigate, the kinds of models they developed to solve these problems, and the analysis techniques they apply (see Fig. 1).

Most respondents use deterministic and continuous models, in particular ordinary differential equations. Among the analysis methods (Figure 2B), general methods such as steady state determination, parameter estimation, bifurcation analysis, or optimization, are more popular than specific methods for metabolic systems (stoichiometric analysis, MCA, FBA, pathway analysis).

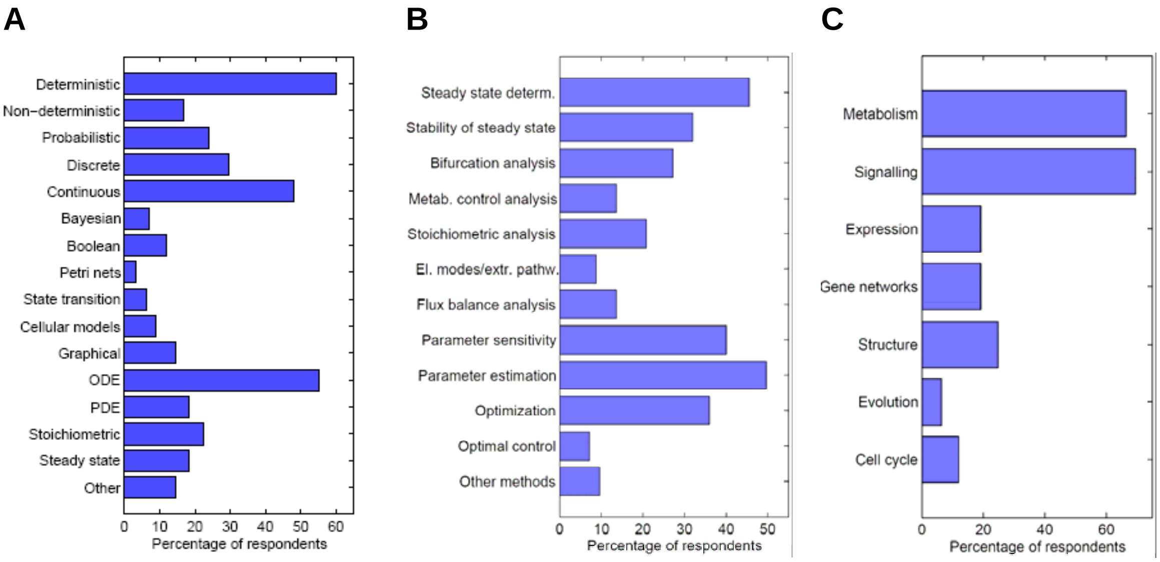


Figure 2: Popularity of model types and analysis methods.

A: Model types employed by the respondents. Bars denote percentages among the 125 respondents. B: Usage of analysis methods. Possible answers for A and B were provided as a list in the questionnaire; multiple choices were possible. C: Biological problems addressed by the models were grouped into seven categories.

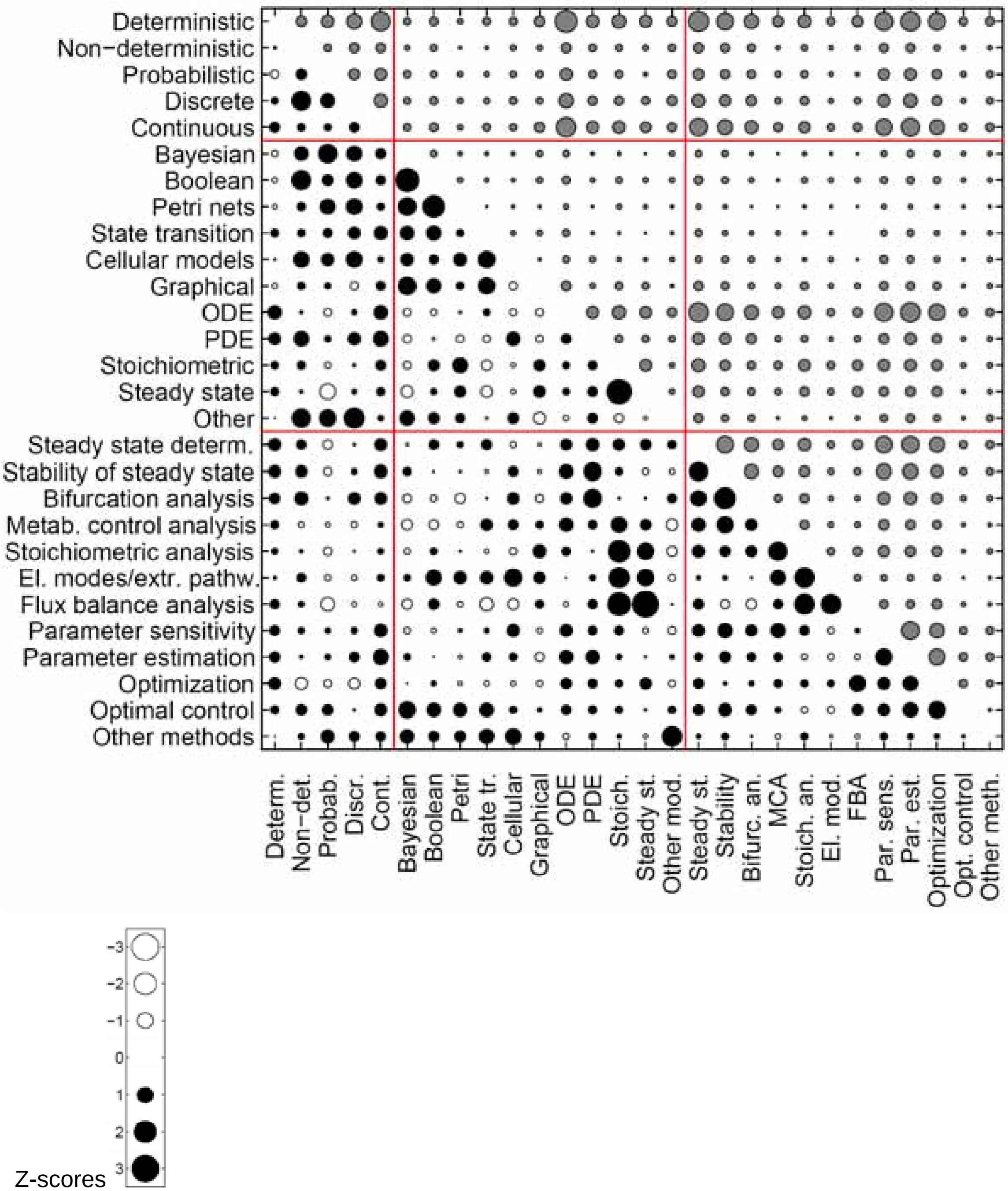


Figure 3: Relation between model types and analysis methods.

Which pairs of methods are preferably used by the same person? Rows and column refer to different methods: lines separate general model types, specific model techniques, and analysis methods. The gray circle areas (upper right triangle) denote the number of persons that marked both methods. Lower left triangle: frequent and rare pairs. The z-scores (circle areas) compare the co-occurrence frequency of methods X and Y to the frequency that would arise under the assumption of mutual independence (see methods). Frequent and rare pairs are denoted, respectively, by black (positive z-scores) and white circles (negative z-scores). Z-scores are given in units of standard deviations.

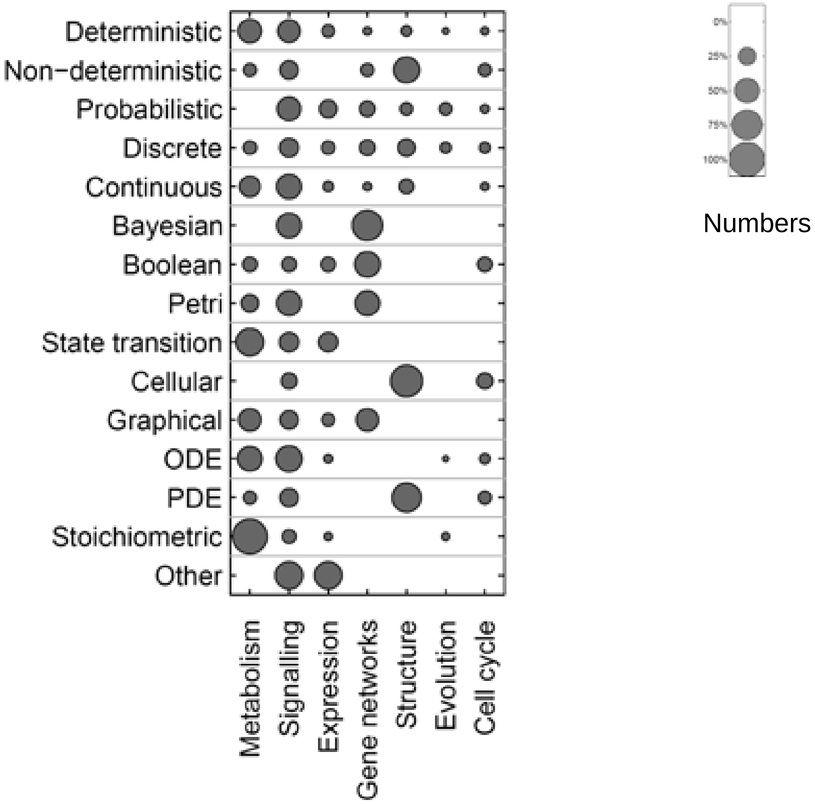


Figure 4: Relation between model types and biological subjects.

Model types and biological problems addressed. Each line refers to the respondents that use a certain model type: circle areas indicate which percentage of these respondents use this model type to address certain types of biological problems (columns).

Each type of mathematical model allows for or requires specific analysis methods; accordingly, certain model types and analysis methods were preferably used/marked together. Figure 3 depicts two different representations of such co-usage; upper triangle: frequency of co-occurrence, lower triangle: z-values indicating the deviation from expected frequencies (see methods). Absolute pair frequencies indicate that most people employ deterministic continuous ODE models to which they fit parameters and whose sensitivity and steady state they analyze. The z-scores, that is, the deviations from expectation values, indicate three groups of methods that tend to be used by the same persons: (i) kinetics-based steady state analysis (steady state determination, stability of steady states, bifurcation analysis and MCA), (ii) pathway analysis (MCA, stoichiometric analysis, elementary modes/pathway analysis, flux balance analysis), and (iii) methods focusing on system parameters (parameter estimation, parameter sensitivity, optimization, optimal control).

Respondents could state the research areas and biological problems they address: metabolism and signaling were mentioned most often, comprising 60% of all answers (Figure 4). ODE models are mostly used for modeling metabolism, cell signaling, and cell cycle. Not surprisingly, stoichiometric models are mainly used for metabolic systems, and partial differential equations for spatially structured problems. Genetic networks are mostly described by discrete models such as Bayesian, Boolean, and Petri networks, and by graphical models.

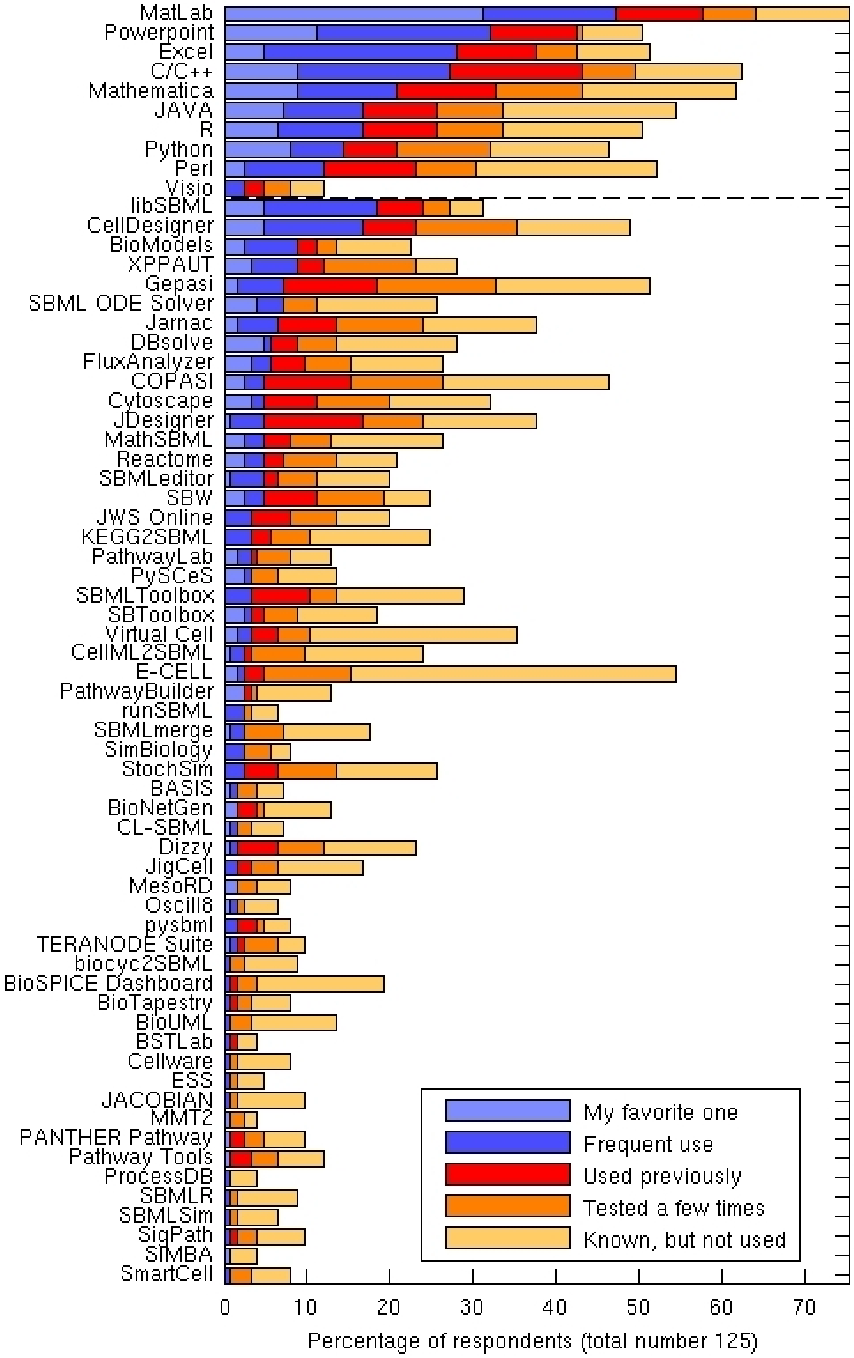


Figure 5: Popularity of software tools.

Usage of 10 general purpose tools (upper part) and 95 specialized tools (lower part, separated by a line). Bars denote the percentage of positive answers in different categories of usage frequency. The tools are sorted according to the sum of the frequencies for “My favorite one” and “Frequent use”. Software tools without positive answer are not shown.

2.3 Modeling Tools

The participants could indicate their usage of tools out of a list of 105 computer programs and could also add programs that were not listed. Each tool could be characterized by categories ranging from “My favorite one” to “Never heard of” (Figure 5).

Almost half of the respondents regularly use MATLAB, while Mathematica is used less than half as often. About a third of the respondents regularly use Powerpoint and Excel. Among the programming languages, C/C++ is most popular, followed by Java, R, Python and Perl. CellDesigner[7] is the most popular stand-alone application for systems biology. Gepasi[9], COPASI[10] and E-CELL[11], have existed for a longer time and are known by many researchers, but are only used sporadically. Certain tools are preferably used in conjunction (Figure 6A). We also noticed certain typical combinations of software tools and analysis methods (Figure 6B). We list the full set of comments to specific software tools in the **Supplementary Material – Tools**, online.

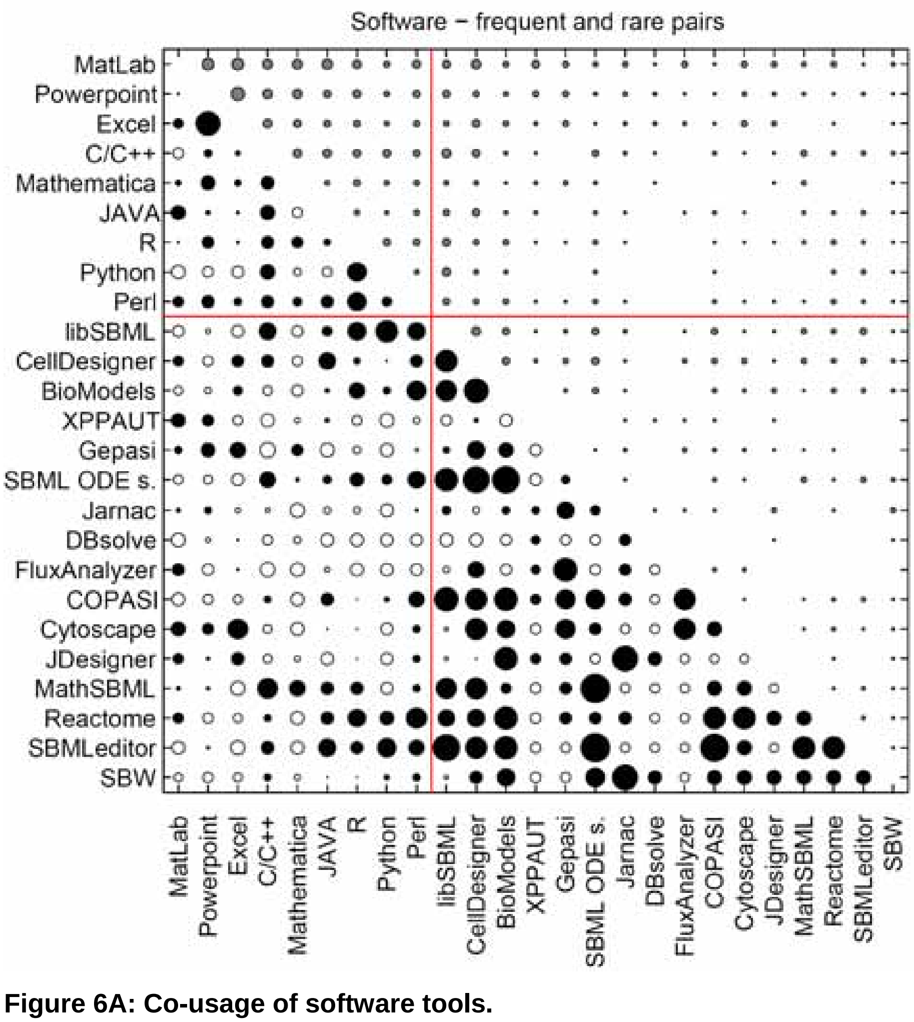


Figure 6A: Co-usage of software tools.

Pairs of tools used by the same person. Absolute numbers (upper triangular part) and z-values (lower triangular part) are shown as in figure 2.

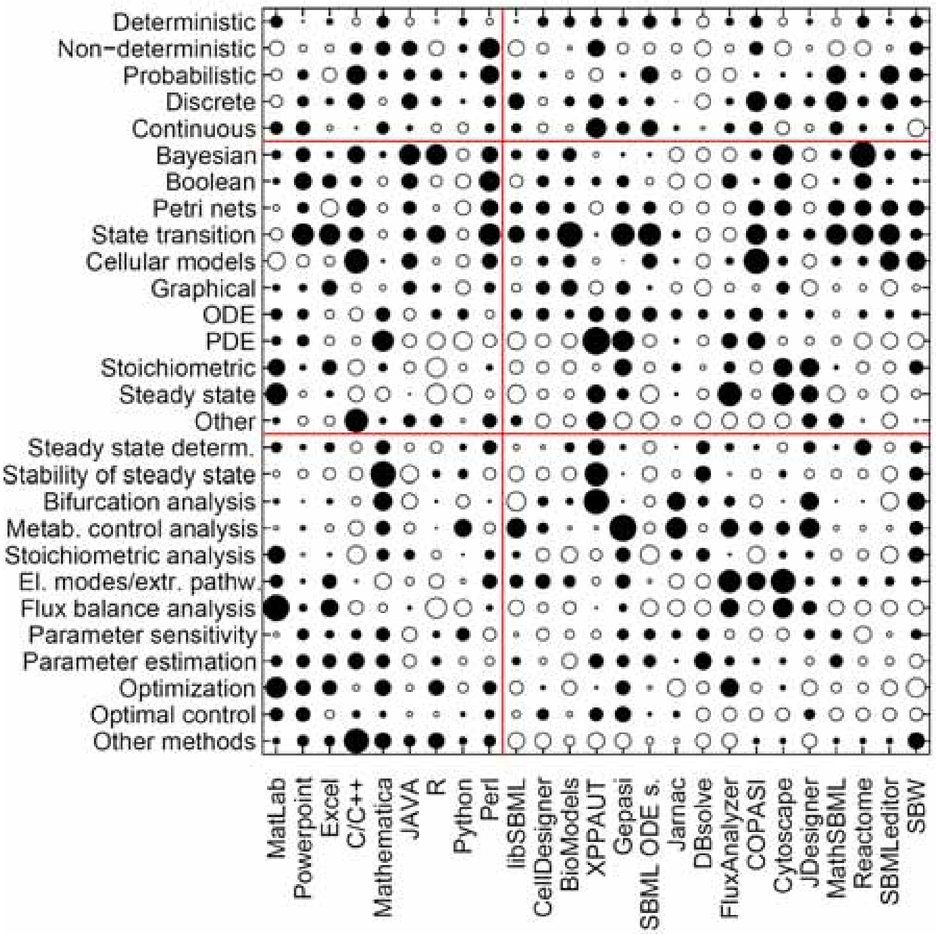


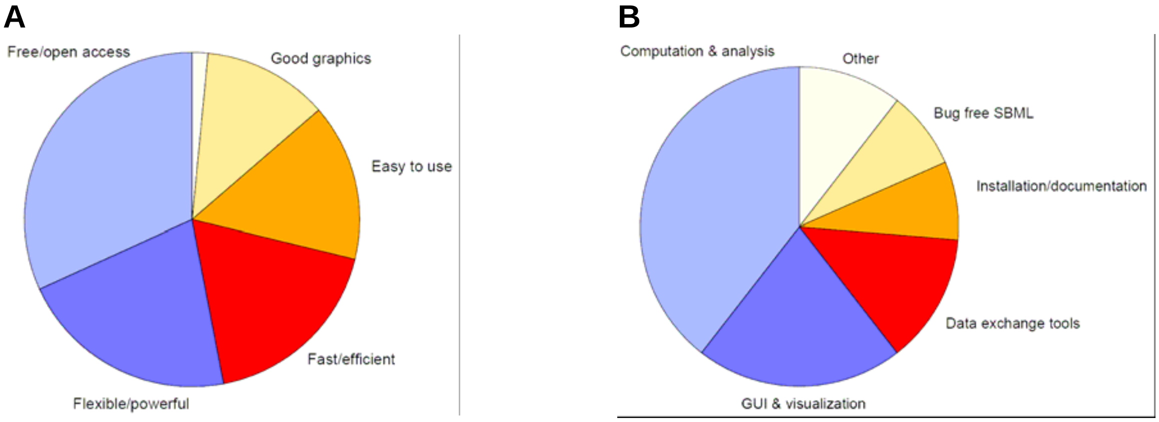
Figure 6B: Pairs of model types/analysis methods and tools.

The z-values indicate co-occurrence of method X (row) and usage of software Y (column). In A and B, only tools frequently used by five or more respondents are shown.

Researchers were also asked to describe the advantages of their favorite tool and to state what makes a tool superior to others. Most answers fell into one of the following categories (see Figure 7):

1. The software is open source or can be obtained for free.
2. The software is flexible (can be applied to many problems) or has powerful analysis methods.
3. The software is fast and efficient.
4. The software is easy to learn and use.
5. The software has good graphics capabilities.
6. There exists good support for the software.

For many respondents, the price and free availability of a tool is of utmost importance, followed by the requirement that it should be flexible and applicable to many different problem types. The main reason is the time and effort necessary to become acquainted with a new tool: it seems that people prefer to learn one complex general-purpose tool instead of several simpler specialized ones. The widespread use of MATLAB indicates that flexibility is seemingly more important than free availability. Note that many respondents regard MATLAB as free software, probably because many universities have campus licenses.



"Please list advantages (or disadvantages, if any) of your favorite tools. What are they good for? Why are they for you superior to other tools"

"What type of functionality is missing – in your favorite tool or in general?"

Figure 7: Advantages and missing features of software tools.

A: Advantages/disadvantages of favorite tool. Distribution of the 66 comments (from free-text fields) over six categories. B: Features missing in software tools. Distribution of the 38 comments (free-text) over six categories.

We also asked about features that are missing in the current tools. The answers were much more diverse than the comments regarding the advantages; eventually, we established the following groups:

1. The software should have better computation and analysis capabilities.
2. Either the GUI of the tool itself or the graphical representation of the generated output should be improved.
3. There should be tools for the exchange of data between different tools.
4. The installation procedure and the documentation should be improved.
5. Tools should implement a complete and bug free SBML support.

6. Requirements that do not fit any of the categories ("others").

The addition of further computation and analysis methods was of the highest importance. However, which specific method or algorithm was asked for heavily depended on the specific tool and can thus not be generalized. The relatively large category "others" that contains requests that cannot be assigned to more specific groups shows the diversity of replies. Not surprisingly, different favorite tools were described to have different advantages and missing functionality (see Supplementary Material – Tools, online).

Finally, researchers were asked: “Looking into the future: Which functionality or tool would you like to have (find for free download on a website)?” Besides wishes for specific functionality (details see Supplementary Material – Tools), respondents stated general positive properties of tools: they should provide proper documentation, their installation and use should be easy, and they should be flexible and extensible. Again, it was stated several times that certain existing tools should be freely available, or existing freely available tools should provide service and performance of existing commercial tools. More specific wishes: their design should be user-targeted (e.g., different for biologists and computer scientists), and the user should be warned if parts of models couldn’t be interpreted. Finally, users asked for a robust integration of interfaces between existing tools and a central, well documented repository for the download of tools. Detailed information is given in the Supplementary Material – Tools.

3 Discussion

From our own experience, modeling is a research area between art and craft. It is an art because modeling means abstraction: choosing the right level of description, finding appropriate system boundaries, and applying useful mathematical concepts are tasks that can hardly be standardized. To demand strict application of standards can strongly restrict scientific imagination and its prospect contribution to cognition. However, with the massive attempts to model larger and larger parts of the cell, the crafting aspect of modeling comes into play, together with an urgent need for standardization: scientists have to be able to repeat, reuse, and improve what their colleagues have done. Some standards, such as SBML, are already emerging. Altogether, we expect that the further development of systems biology will be accompanied by the development of standards.

The large number of respondents and their elaborate and well-advised comments clearly indicate that many researchers in systems biology consider standardization an urgent issue. Respondents expressed a graded feeling of need for standardizing the various aspects of modeling: most of them consider standards necessary or desirable; about half of the respondents state that standards should be enforced by journals. We found that SBML is accepted and appreciated by the majority of modelers; model and data exchange formats are considered important.

Commonly used software tools contribute to the standards. Certain programming tools like MATLAB or C++, or specialized tools like CellDesigner, are already used by a large fraction of modelers. Concerning the further development of software, main demands were flexibility of use and free availability of the software as well as compatibility between specialized tools.

We conclude from our survey that the system biology community would welcome standardization efforts, and possibly the enforcement of standards, at least in the following fields:

1. Graphical representation of biochemical networks
2. Experimental conditions and scenarios
3. Exchange formats for computational models
4. Nomenclature of cellular compounds and molecules

The scientific community has already made efforts to develop standards for these topics: (1) The Systems Biology Graphical Notation initiative fosters standards in network representation. Users consider CellDesigner and the Molecular Interaction Map as important contributions to implement standards. (2) Defined sets of experimental conditions and scenarios can ensure the production of data useful for modeling; deciding on such sets needs strong interaction between experimentalists and modelers and will be a main field of further development. This task is, for example, an important topic within the Yeast Systems Biology Network (YSBN). (3) SBML has become an important exchange format for models. It is supported by specialized and general tools. Advantages and drawbacks are discussed in Alves et al.[2]. (4) A unique nomenclature of proteins among species remains an unsolved problem. The very positive exception is the EC nomenclature for enzymes.

The results of the questionnaire suggest the following guidelines for SB researchers.

For software developers it is important to ensure that the tools are freely available, flexible, compatible, and that they are well documented.

To enable reuse and exchange of models, a standardized description of the model structure and the parameters is necessary, which must be developed by tool developers and tool users together. In addition, modelers are demanded to make their models available in a format that allows the exchange and testing by others. Considering the frequent use of SBML, this would be preferably SBML format. If this is not possible, then the program code should be made available. Those demands have also been formulated previously by MIRIAM [1]. The further processing of model-derived data, such as simulation results, is still an open problem.

The frequency of simultaneous marking of model types or techniques and analysis methods by the respondents gives hints for modelers starting new projects or newcomers to the field, which techniques or methods could be appropriate for their subject of research.

An important conclusion for biological experimentalists is to store and publish the experimental data in a digital format, not only in illustrative, but barely quantifiable figures. Unique representation of data would foster data exchange and comparison and would support the communication with modelers.

Development of standards must be performed by the scientists. In addition, science related institutions such as scientific journals and funding organizations and also initiatives like the SBML consortium can support the acceptance and the necessary progression of standardization.

4 Methods

4.1 Construction and Distribution of the Questionnaire

The background for conducting our survey was to get an overview about currently used standards within the frame of the EU project Yeast Systems Biology Network (YSBN), as a basis for further work on that subject. The construction of the questionnaire was based on our own experience, incorporating comments from the YSBN consortium. Due to our involvement in various modeling projects, we cannot exclude that the construction of the questionnaire is biased towards SBML. The list of tools is partially taken from the website sbml.org, extended by a series of tools that we knew of.

The questionnaire was made available at http://www.molgen.mpg.de/~ag_klipp/questionnaire/ from March 11, 2006 until August 29, 2006. The document is still available, but in the present reporting, we only consider the 125 forms filled within this period. The awareness about the questionnaire was spread by the ysbn.org web-site, and by available email-lists comprising researchers active in the field of systems biology.

4.2 Categorization of free-text fields

After collecting the free-text comments concerning biological problems, the advantages of software tools, and the missing functionality of tools, we decided to group them into categories of similar comments.

From 125 participants, 88 (70%) made use of the opportunity to state the research area or biological problem. We tried to categorize the biological problems that people described and came up with five groups that correspond to five areas of actual systems biology research. These groups were metabolism, signaling, transcription, genetic networks, structure, evolution and cell cycle. The research areas of metabolism, signaling, genetic networks and cell cycle are rather well defined because they were either stated as such by the people themselves or were easily assigned. The category “transcription” comprises all answers that dealt with modeling of gene transcription or gene expression. The category “structure” is the most diverse one because it comprises problems related to the structure of cells, such as volume, structure of proteins as well as structure of populations. Using these categories, we could group 70 (85%) biological problems into one of them.

4.3 Z-score for detecting frequent and rare pairs of answers

We used a z-score to detect frequent and rare pairs of answers. Let $A = \{a_1, a_2, ...\}$ and $B = \{b_1, b_2, ...\}$ denote (disjoint) sets of questions, each with a choice between the answers “yes” and “no”. Let n_{ik} denote the number of persons who gave positive answers to both a_i and b_k. Let $n_{i\bullet}$ and $n_{\bullet k}$ denote the number of persons who gave a positive answer to a_i and b_k, respectively, and let n denote the total number of respondents. We estimate the percentage of positive answers to a_i by $p_i = n_{i\bullet} / n$ and the percentage of positive answers to b_k by $p_k = n_{\bullet k} / n$. If all answers were uncorrelated, we would expect that a percentage $p_{ik} = p_i p_k$ of all respondents would give a positive answer to both questions; this corresponds to an expected number of $\bar{n}_{ik} = np_{ik}$. To detect the numbers n_{ik} that are unexpectedly larger or smaller than the expected value $\bar{n}_{ik}$ we compute and plot the Z-score $\left(n_{ik} - \bar{n}_{ik}\right) / \sigma_{ik}$ with $\sigma_{ik} = \sqrt{\bar{n}_{ik}}$. The standard deviation σ_{ik} reflects our assumption that n_{ik} follows a Poisson distribution with mean value $\bar{n}_{ik}$. When rows and columns on the matrix refer to the same set of questions, the matrix is symmetric and the values on the diagonal are meaningless.

5 Acknowledgements

We are very grateful to all researchers who filled in a questionnaire and thereby ensured its success. We thank Cornelia Kunz for helpful discussions about questionnaire analysis.

This work was supported by grants of the European Community to E.K.: the Yeast Systems Biology Network (YSBN, EU-Project No FP6-2004-LSH-5 # 018942), EMICD (EU-Project No 503269), QUASI (EU-Project No LSHG-CT2003-530203) and ENFIN (EU-Project No LSHG-CT-2005-518254).

6 References

1. Novere, N.L. et al. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* **23**, 1509-1515 (2005).
2. Alves, R., Antunes, F. & Salvador, A. Tools for kinetic modeling of biochemical networks. *Nat Biotechnol* **24**, 667-672 (2006).
3. Brazma, A., Krestyaninova, M. & Sarkans, U. Standards for systems biology. *Nat Rev Genet* **7**, 593-605 (2006).
4. http://www.molgen.mpg.de/~ag_klipp/questionnaire/.
5. 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).
6. Lloyd, C.M., Halstead, M.D. & Nielsen, P.F. CellML: its future, present and past. *Prog Biophys Mol Biol* **85**, 433-450 (2004).
7. Funahashi, A., Tanimura, N., Morohashi, M. & Kitano, H. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico* **1**, 159-162 (2003).
8. Kohn, K.W., Aladjem, M.I., Weinstein, J.N. & Pommier, Y. Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. *Mol Biol Cell* **17**, 1-13 (2006).
9. Mendes, P. Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem Sci* **22**, 361-363 (1997).
10. Hoops, S. et al. COPASI - a COmplex PAthway SImulator. *Bioinformatics* (2006).
11. Tomita, M. et al. E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15**, 72-84 (1999).